Learning pathological representations in neuroimaging : Predicting psychiatric diagnosis by integrating heterogeneity constraints

Robin Louiset

université PARIS-SACLAY

Learning pathological representations in neuroimaging:

Predicting psychiatric diagnosis by integrating heterogeneity constraints.

Apprentissage de représentations pathologiques en neuroimagerie:

prédiction du diagnostic psychiatrique en intégrant des contraintes d'hétérogénéité.

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°575 : electrical, optical, bio : physics and engineering (EOBE)
Spécialité de doctorat: Physique et imagerie médicale
Graduate School : GS Sciences de l'ingénierie et des systèmes.
Référent : Faculté des sciences d'Orsay.

Thèse préparée dans l'unité de recherche **BAOBAB** (Université Paris-Saclay, CEA, CNRS), sous la direction d'**Edouard DUCHESNAY**, Directeur de Recherche, la co-direction d'**Antoine GRIGIS**, Habilité à Diriger des Recherches, et le co-encadrement de **Pietro GORI**, Maître de conférences

Thèse soutenue à Paris-Saclay, le 19 juin 2024, par

Robin LOUISET

Composition du jury

Membres du jury avec voix délibérative

Bertrand Thirion Professeur, Inria Paris-Saclay	Président
Andre F. Marquand Professeur, Donders Institue	Rapporteur & Examinateur
Jean-Philippe Thiran Professeur, Ecole Polytechnique Fédérale de Lausanne	Rapporteur & Examinateur
Marco Lorenzi Professeur, Université Côte d'Azur, Inria	Examinateur
Maria Vakalopoulou Professeur Associé, CentraleSupelec	Examinatrice

THESE DE DOCTORAT
NNT : 2024UPAST044

université PARIS-SACLAY
ÉCOLE DOCTORALE
Physique et ingénierie : Electrons, Photons, Sciences du vivant (EOBE)

Titre: Apprentissage de représentations pathologiques en neuroimagerie : prédiction du diagnostic psychiatrique en intégrant des contraintes d'hétérogénéité.
Mots clés: Apprentissage Profond, Neuro-imagerie, Apprentissage de Représentation, Découverte de Sous-groupes, Analyse Contrastive, Troubles Psychiatriques.

Résumé: Les mécanismes biologiques qui sous-tendent les symptômes des maladies psychiatriques sont encore mal compris à de nombreux égards. Une hypothèse qui justifierait la difficulté des chercheurs à identifier des marqueurs biologiques associés spécifiquement avec les troubles schizophrènes, bipolaires ou autistique serait qu'il existe une hétérogénéité neurobiologique importante au sein de chacune de ses maladies, qui rend plus difficile l'analyse de ces affections. Dans cette thèse, notre objectif est de développer des techniques d'apprentissage statistique automatique pour stratifier les maladies psychiatriques en sous-groupes ou en spectre de dimensions indépendantes sur la base de marqueurs biologiques objectifs acquis à l'aide de technique d'imagerie IRM neuroanatomique. Dans un premier temps, nous nous sommes concentrés sur le développement de méthodes de regroupement («clustering») visant à stratifier une maladie en sous-groupes homogènes. Un obstacle majeur fût d'observer que la prédominance de marqueurs neuroanatomiques sous-tendus par des facteurs de variabilité «généraux» (age, sexe, site d'acquisition, ethnicité) rendait difficile l'identification de marqueurs pathologiques distincts (associé à des échelles cliniques telles que la paranoïa, l'anxiété, ou la dépression par exemple). Notre première contribution fût donc de développer un algorithme linéaire de découverte de sous-groupes: UCSL qui se base uniquement sur des facteurs de variabilité existant spécifiquement et uniquement dans la population malade et non dans la population saine. Afin d'étendre ce travail à des algorithmes d'analyse profond non-linéaire, potentiellement plus puissant, susceptibles de reconnaître des signatures pathologiques plus complexes, nous avons étendu l'algorithme UCSL à un algorithme de Deep Learning intitulé Deep UCSL. Deep UCSL est capable d'extraire directement des caractéristiques directement dans l'image IRM anatomique et démontre des capacités de généralisation à d'autres domaines d'imagerie médicale (pathologies de l'œil et des poumons). Puis, afin d'illustrer l'utilité de ces méthodes de découverte de sous-groupes, nous avons utilisé la méthode linéaire UCSL dans une cohorte de personnes atteintes de schizophrénie pour identifier plusieurs sous-types et analyser leur pertinence clinique. Une autre ligne de recherche intéressante consiste à estimer les facteurs génératifs latents distincts et interprétables qui sous-tendent l'hétérogénéité neurobiologique propre à la maladie psychiatrique. Ainsi, nous nous sommes donc intéressés à une classe de méthodes d'apprentissage de représentations pathologiques (qui capture des motifs de variabilité qui n'existent que dans la maladie) intitulée «l'analyse contrastante». Ce champ de l'apprentissage statistique vise à séparer les facteurs de variabilité «communs» et «cibles», étant donné un ensemble de données «source» et un ensemble de données «cible». Dans notre cas, l'objectif serait d'identifier d'un côté la projection permettant d'identifier les motifs de variabilité sains et de l'autre côté la projection permettant d'identifier les «signatures pathologiques» qui n'existe que dans la classe des malades et non dans la classe des gens sains. Nous avons développé une méthode d'auto-encodeur variationnelle: SepVAE, qui contribue à des méthodes concurrentes via l'ajout de deux fonctions de coût à minimiser.. Enfin, nous avons développé une nouvelle méthode d'analyse contrastante: SepCLR, qui étend le cadre des méthodes d'Analyse Contrastive à une autre classe de méthodes prometteuse d'apprentissage de représentations: l'apprentissage par maximisation d'information mutuelle. Nous avons validé ces deux contributions méthodologiques sur des ensembles de données de vision, médicaux et de neuroimagerie.

Title: Learning pathological representations in neuroimaging :
Predicting psychiatric diagnosis by integrating heterogeneity constraints.
Keywords: Deep Learning, Neuroimaging, Representation Learning, Subgroup Discovery, Contrastive Analysis, Psychiatric disorders.

Abstract: The biological mechanisms that underlie the symptoms of psychiatric diseases, such as schizophrenia, bipolar, or autistic disorders, are still poorly understood in many regards. One of the main reasons is the neurobiological heterogeneity associated with these diseases. Furthermore, healthy subjects usually share common but irrelevant factors of variation with the patients, such as age, sex, acquisition site, and ethnicity. This hampers the identification of clear and interpretable biological markers associated with these diseases. In this thesis, our goal is to develop machine learning techniques to automatically stratify psychiatric diseases into homogeneous subgroups or to automatically identify the pathological latent distinct and interpretable generative factors, based on objective biological markers acquired through neuroanatomical MRI imaging techniques. At first, this thesis focused on developing clustering methods to stratify a disorder into homogeneous subgroups. Our first contribution was a *linear* subgroup discovery algorithm, called UCSL (Unsupervised Clustering driven by Supervised Learning), which identifies subgroups that stem only from the pathological variability specific to the disorder while disregarding the common variability shared with the healthy population. As a second contribution, this was then extended with a non-linear deep features extractor, potentially more powerful in recognizing complex pathological signatures. This new deep learning method entitled Deep UCSL, can directly extract features from anatomical MRI images, showed state-of-the-art results in neuro-psychiatric subgroup identification, and demonstrated generalization capabilities to other medical imaging domains (eye and lung pathologies). Ultimately, to illustrate the usefulness of such Subgroup Discovery methods, the linear method UCSL was leveraged to identify subtypes in a cohort of individuals with schizophrenia and to analyze their clinical relevance.
Another line of research investigated in this thesis consisted of estimating the latent distinct and interpretable generative factors that underpin the neurobiological heterogeneity proper to the psychiatric disorder. To address this objective, this thesis investigated a class of representation learning methods that enable separating pathological patterns from healthy patterns of variability: contrastive analysis methods. These methods do not require assuming the existence of homogeneous subgroups. This field of statistical learning aims at separating "common" and "target" variability factors given a "source" dataset and a "target" dataset. In our case, the goal is to identify, on the one hand, the projection that allows identifying healthy variability patterns and, on the other hand, the projection that allows identifying "pathological signatures" that exist only in the class of patients and not in the class of healthy people. A contrastive variational autoencoder method entitled SepVAE was developed and contributed to competing methods in two ways: by adding a classification task in the pathological space and by adding a cost function based on mutual information to minimize information redundancy between the common space and the pathological space. Eventually, to provide a rich methodological perspective, a novel contrastive analysis strategy was developed. This method extends the framework of contrastive analysis methods to another promising class of representation learning methods: mutual information maximization learning. These methodological contributions were then validated on vision, medical, and neuroimaging datasets.

Remerciements (Acknowledgements)

Tout d'abord je tiens à remercier le jury de thèse qui a accepté d'assister à ma soutenance de thèse malgré un agenda serré et un long manuscript. Les rapports de thèse ont été élogieux et j'ai pu compter sur votre expertise pendant la session question, une expérience qui m'enrichit personnellement et professionnellement. Ensuite je tiens, bien sûr à remercier mon équipe de travail, je remercie Antoine, qui a toujours aidé à corriger les derniers détails de mes travaux. Je remercie Benoit, avec qui j'ai pu avoir des échanges passionnants pendant la thèse qui m'ont inspiré dans mon travail. Benoit tu m'as éprouvé intellectuellement, mais aussi physiquement parsque tu es vraiment beaucoup trop fort en course à pied !

Je remercie également Pietro, qui a endossé le rôle d'encadrant de thèse peut-être encore mieux que son rôle de professeur à Télécom. Tu m'as fait aimé le côté rigoureux et technique de la recherche. Tes retours étaient toujours honnêtes, pertinents et bienveillants. J'ai beaucoup progressé pendant nos réunions et nos sessions d'écriture, parfois jusqu'à minuit. Je n'aurais pas fait la même thèse sans tes retours et sans tes directives, merci pour ça.

Je remercie Edouard, tu as été un directeur de thèse très inspirant, qui m'as fait aimé le côté créatif de la recherche, qui m'a introduit au monde de la psychiatrie et de la neuroimagerie, un monde que je n'ai plus envie de quitter, même si c'est un monde difficile. Tu as été un très bon pédagogue et tu as aussi su m'enseigner l'importance de communiquer des messages scientifiques simplement. Au-delà de ton statut de directeur de thèse, tu as aussi été un bon collègue, toujours prêt à prendre du temps pour une course à pied, une discussion scientifique, technique ou politique. Tu étais également toujours prêt à prendre du temps pour un gin tonic au bar pendant une conférence, mais retenons plutôt les discussions scientifiques et techniques !

Je remercie l'ensemble de mes collègues et amis de NeuroSpin avec qui j'ai passé de très bons moments, l'équipe course à pied et escalade qui a su faire monter mon cardio avec Coco, Benoit, Raph, Thibaut, Julien (les deux), Joel, Nico, François, Aymeric, Loic et Chloé. Et je remercie aussi globalement l'équipe groupe de lecture qui a su faire monter mes compétences à décrypter des équations. Je remercie l'équipe Glasgow, Montréal et Grenoble avec Sara, Wenqi, Pierre, Cyril, Julie, Loic et Clément, qui eux, ont su faire progresser ma tolérance à la bière. Avec Wenqi, on trouvait qu'il faudrait une section anti-remerciements dans la thèse pour désigner

les collègues avec qui il est agréable de faire une pause café, de s'accorder une course à pied ou un ping-pong. Donc voilà, je tiens à anti-remercier quelques collègues qui se reconnaîtront.

Je remercie aussi bien sûr les membres de l'équipe image de Télécom dont mes amis Chloé, Raph, Erwan, Antoine, Ines et Marie. On a passé des bons moments à Télécom même si je ne m'y suis pas assez rendu !

Au-delà du travail et des collègues, même si ce sont bien souvent des amis, il y a aussi un équilibre très important avec la vie personnelle, dans laquelle j'ai eu la chance d'avoir des amis proches exceptionnelles qui m'ont gratifié de bons moments de vie. Alors j'aimerai remercier mes amis d'école Ramine, Antoine, Paul, Geert-Jan, Adrien, Matthias, Raph, Soso, Adriano. Je sais que je peux compter sur vous pour me changer les idées. Et je ne peux m'empêcher d'avoir une pensée pour Yaya, avec qui j'aurais aimé pouvoir me réjouir de ce diplôme. On pense fort à elle.

Il est impossible pour moi maintenant d'évoquer mes amis sans parler de ceux que j'ai rencontré pendant mon enfance et mon lycée. Alors j'envoie beaucoup d'amour à Yohann et olo, mes deux plus proches troubles-fêtes, et bien sûr à Nino, Chloé, Vico, Testi, Moumousse, Tony, Damdam, Meumeu, Teddy, Agathe et tout les autres. Les moments passés avec vous sont parmi les plus heureux dans ma vie.

J'envoie aussi mes pensées en Thaïlande à mon amie éternelle, Sabine, ça fait plus de 20 ans qu'on se connaît et rien n'a changé, tu comptes beaucoup pour moi.

J'aimerai maintenant exprimer une certaine tendresse à celle qui partage ma vie, Jeanne, qui pas une seule fois n'a exprimé sa frustration alors que je travaillais un week-end, un jour férié, ou un soir. On a vécu des belles aventure ensemble, on s'est soutenu, à travers les canicules, les feux de forêts ou les pluies diluviennes. Saches que tu as été une personne essentielle dans cette aventure là qui est ma thèse. Alors merci pour ça.

Et pour finir, je remercie ma famille, mes parents, Théo, Morgane et les autres, vous m'avez épaulé pendant cette thèse, vous m'avez écouté avec beaucoup intérêt, et j'ai aimé partager cet aspect important de ma vie avec vous. Papa, Maman, vous êtes toujours en vadrouille, toujours prêt à vous intéresser à ce que je raconte, toujours prêt à aborder le monde avec curiosité, c'est

vous qui m'avez montré l'exemple. Théo, quand j'ai commencé ma thèse et qu'on était en collocation, tu m'as dit "*La thèse c'est un marathon, il faut bien gérer ses efforts pour pas exploser en plein vol*". Et bien grâce à toi Théo, j'ai passé la ligne d'arrivée. Et au moment où j'écris ces lignes, j'ai encore de l'énergie pour m'attaquer à un authentique marathon en septembre alors j'espère que toi, les parents, Yohann et Jeanne vous serez encore là pour me soutenir.

Contents

Chapter 1

Résumé en français / French Summary

Contents

1.1 Contexte et objectifs

Lors des dernières décennies, la recherche en psychiatrie a été sujette à des investissements et des efforts de la part de la communauté scientifique. Cet intérêt est justifié par le besoin de mieux diagnostiquer et comprendre les maladies psychiatriques. Selon l'organisation mondiale de la santé (OMS), en 2019, environ 970 millions d'individus dans le monde souffraient d'une maladie mentale, la dépression et l'anxiéte étant parmis les affections les plus courantes. Ces affections sont d'une part catastrophiques pour les individus concernés mais également pour la société dans laquelle ils vivent puisqu'elles représentent un coût non-négligeables pour la communauté comme le prouve une étude de l'URC-Eco pour la fondation FondaMental [1] (2018).

1.1.1 Diagnostic et classification des maladies psychiatriques

Dans la routine psychiatrique actuelle, diagnostiquer une maladie psychiatrique s'effectue sur la base de méthodes d'évaluations psychométriques, cognitives, ou encore comportementaux, via des questionnaires, entretiens ou observations menés par le cliniciens ou reportés par le patient, ou ses proches. Généralement, ses évaluations sont mises en parrallèle avec un système de classification comme le DSM-5, la dernière version du Manuel du Diagnostic des Maladies Psychiatriques [69]. Cependant, de récents travaux ont démontré que cette manière de diagnostiquer produisait des résultats variables, qui dépendent du clinicien, de son choix de de méthodes d'observations (comportementale, cognitive, ou encore émotionelle), ou de selon si la méthode choisie est menée par le clinicien, le patient, ou ses proches [302, 187]. Par ailleurs, la classification des maladies mentales et plus généralement leur étiologie est encore aujourd'hui régulièrement remise en question régulièrement.

1.1.2 Comprendre les maladies psychiatriques en observant le cerveau

Afin de mieux comprendre les maladies psychiatriques et de redéfinir de manière plus objective et concrète leur étiologie, plusieurs initiatives comme celle de Thomas R. Insel, and Remi Quirion en 2005 [133] ont appelé à considérer la psychiatrie comme une discipline de neurosciences, dont la compréhension doit être alimentée par des observations du cerveau. Cette idée permettrait potentiellement d'identifier des lésions ou des marqueurs biologiques dans le cerveau associés avec une ou plusieurs maladies psychiatriques et de guider le développement de nouveaux traitements pharmaceutiques [1, 2, 96, 135, 304]. Par exemple, plusieurs travaux

[1] https://www.fondation-fondamental.org/

ont motivé la recherche de sous-groupes de malades sous-tendus par des lésions objectives et associés à des symptômes distincts mais potentiellement intersectables [89, 57]. Ces contributions sont de grands pas en avant afin de simplifier et raffiner l'étiologie psychiatrique à l'aide d'observations biologiques.

Pour observer des lésions potentielles, plusieurs modalités d'acquisitions se distinguent. Parmis celles-ci, nous pouvons lister le génotypage, des techniques d'identifications de molécules, l'EEG (éléctro-encéphalographie), l'imagerie PET (Positron Emission Tomographie), l'imagerie IRM neuro-anatomique (sMRI) ou encore l'imerie IRM fonctionelle (fMRI). Parmis toutes ces modalités, nous déciderons d'étudier l'imagerie neuro-anatomique car c'est une technique d'acquisition non-invasive, moins bruité que d'autres techniques comme l'IRM fonctionelle ou l'EEG. Par ailleurs, l'IRM neuro-anatomique permet d'observer directement le résultat des interactions entre les facteurs génétiques et les facteurs de stress environnementaux, ce qui n'est pas le cas des techniques de génotypage.

1.2 Dépasser le paradigme de la classification cas / contrôle

1.2.1 Prédire le diagnostic individuellement avec l'apprentissage machine

L'espoir de trouver des marqueurs neuro-anatomiques reproductibles et robustes est motivé par de nombreux travaux de recherche qui ont pu identifier des motifs de déviations morphologiques dans plusieurs maladies psychiatriques tel que la schizophrénie, la bipolarité, ou encore l'autisme, en comparant à des groupes de contrôles sains, [138, 177, 141, 255, 244, 163, 126, 282, 114, 198, 120, 121, 172, 54, 204, 235, 290, 80, 81]. Bien que ces études suggèrent des marqueurs neuro-anatomiques pour les maladies psychiatriques, elles échouent à respecter les critères qui permettent d'en faire des bio-marqueurs de pronostic ou de diagnostic. Cette limitation vient du fait que ces bio-marqueurs sont prédictifs à l'échelle de la comparaison statistique groupe-à-groupe, mais qu'ils ne sont pas satisfaisant pour la prédiction de diagnostic à l'échelle individuelle comme l'explique Bzdok et al. [37] en 2017.

Afin de développer des méthodes statistiques performantes à l'échelle individuelle, de nombreuses recherches se sont tournés vers les techniques de Machine Learning qui, après s'être entraîné à prédire correctement un statut clinique à partir de données (d'imagerie par exemple), peuvent extrapoler et généraliser à de nouvelles entrées individuelles. Néanmoins, bien que ces méthodes soient performantes, elles échouent tout de même à obtenir des scores de

précision satisfaisants sur des données indépendantes d'évaluation. Cela est justifié par le fait que, généralement, ces méthodes restent entraînées dans un paradigme de classification binaire cas/contrôle.

1.2.2 Stratifier l'hétérogénité au sein des maladies psychiatriques

En effet, comme pour les symptômes, les marqueurs biologiques associés spécifiquement avec les troubles schizophrènes, bipolaires et autistiques pourraient être potentiellement très hétérogènes, ce qui rendrait plus difficile l'analyse statistique de ces affections dans un paradigme de comparaison cas/contrôle. Pour cette raison, de nombreux travaux visent à développer des techniques d'apprentissage statistique automatique pour stratifier les maladies psychiatriques en sous-groupes ou en spectre de dimensions indépendantes sur la base de marqueurs biologiques objectifs acquis à l'aide de l'imagerie IRM anatomique.
Pour analyser la variabilité au sein des maladies psychiatriques, de nombreux travaux ont noté un obstacle majeur: la prédominance de marqueurs sous-tendus par des facteurs de variabilité «normaux» (age, sexe, site d'acquisition, ethnicité) rend difficile l'identification de marqueurs pathologiques distincts (associé à des échelles cliniques telles que la paranoïa, l'anxiété, ou la dépression par exemple). Ainsi, pour surmonter ce problème, des techniques pour ignorer ces facteurs, ou ajuster l'estimation de sous-groupes ou d'un espace de dimensions pathologiques par rapport à ces facteurs ont émergé. Parmis, ces techniques, se trouvent des techniques de résidualisation, de modèles normatifs, qui sont des méthodes de pré-traitement pour ajuster les données d'entrées par rapport à des facteurs de variabilité connus. Il existe aussi des techniques de découverte de sous-groupes ou encore de dimensions pathologiques latentes (des motifs de variabilité dans les données d'entrée qui n'existent que dans la population malade). Ces techniques visent à estimer les facteurs latents statistiques (sous formes de sous-groupes ou de dimensions) qui sous-tendraient l'hétérogénéité neuroanatomiques propre à la population malade, en contrastant avec les facteurs de variabilité qui existent dans la population saine. Ces techniques sont complémentaires à celles de résidualisation et de modèles normatifs et seront celles qui nous intéresserons tout au long du manuscript. Les objectifs de cette thèse peuvent se diviser en deux contributions distinctes:

1. la recherche et le développement d'une méthode linéaire et d'une méthode de réseau de neurones profonds capables d'identifier des sous-groupes de malades à partir d'une variabilité qui serait propre à la cohorte pathologique. En effet, de précédents travaux ont montré que des méthodes de regroupements simples estimés sur une cohorte malade

ou saine avaient tendance à discriminer selon une variabilité "générale", ou "commune", c-à-d qui existe dans la population saine et dans la population malade. De récents travaux ont proposé des méthodes linéaires pour la découverte de sous-groupes [279, 280, 127] mais requiert plusieurs hypothèses majeures à propos de la topologie des données pathologiques et de la linéarité des lésions pathologiques. En développant un cadre statistique général, permettant de développer un méthode linéaire et profonde pour la découverte de sous-groupe, cette thèse donne de nouvelles méthodologies pour stratifer des maladies mentales (et pas que) tout en ignorant les facteurs de variabilité généraux comme le vieillisement dans les applications neuropsychiatrique par exemple. Nous avons ensuite investigué notre méthode de découverte de sous-groupe sur une problématique concrète: "Quels sont les sous-types biologiques que nous pouvons identifier dans une cohorte de patients atteints de troubles schizophrènes ?". "Quelles observations pouvons-nous faire et quelles corrélations pouvons-nous identifier entre les bio-marqueurs et les échelles cliniques et cognitives ?". "Les sous-groupes sont-ils homogènes et bien cloisonnés ?" Cette analyse approfondie des troubles schizophrènes nous donne un aperçu de la variabilité biologique qui existe au sein de ce trouble mental et des hypothèses à faire lors du développement de méthodes d'analyse de l'hétérogénéité neuro-anatomique de la schizophrénie.

2. la recherche, le développement et l'évaluation d'une méthodolgie robuste et reproductible d'Analyse Contrastive en neuroimagerie pour la recherche en psychiatrie. Ces techniques visent à estimer les facteurs latents statistiques (sous formes de sous-groupes ou de dimensions) qui sous-tendraient l'hétérogénéité neuroanatomiques propre à la population malade, en contrastant avec les facteurs de variabilité qui existent dans la population saine. Ainsi, elles ont le potentiel de séparer les motifs de variabilités neuroanatomiques que les patients malades partagent avec les contrôles sains des motifs de variabilité neuroanatomque qui sont uniquement spécifiques à la pathologie. Elles permettent également de trouver des dimensions pathologiques qui organisent la population malade en un spectre continu, plutôt qu'en sous-groupes homogènes. Dans cette thèse, nous développons des méthodologies pour ce genre d'applications avec diverses techniques d'apprentissage de la représentation, telles que les autoencodeur variationnels et les méthodes d'apprentissage de représentation contrastive. A l'aide de la technique de l'encodeur variationnel, nous avons mis en évidence deux fonctions de coûts importante pour ce genre de méthode: 1) une fonction de classification non-linéaire dans l'espace salient, 2) une fonction de minimisation de l'information mutuelle entre l'espace de représentation commun et pathologique. Puis, nous avons validé notre méthode sur plusieurs jeux de données, dont deux de neu-

ropsychiatrie acquis à l'aide d'IRM neuro-anatomique. En ce qui concerne la technique d'apprentissage de représentation contrastives, nous avons formulé un nouvel objectif à maximiser à l'aide de l'information mutuelle. Puis, nous avons estimé les quantités statistiques d'intérêt à l'aide de fonctions de coût inspitré de l'apprentissage de représentation contrastif. Nous avons ensuite validé notre méthode sur plusieurs jeux de données, dont un jeu de neuroimagerie en psychiatrie.

1.3 Développement de méthodes de découvertes de sous-groupes

Dans un premier temps, nous nous sommes concentrés sur le développement de méthodes de regroupement («clustering») visant à stratifier une maladie en sous-groupes homogènes. Un obstacle majeur à cette tâche fût d'observer que la prédominance de marqueurs sous-tendus par des facteurs de variabilité «normaux» (age, sexe, site d'acquisition, ethnicité) rendait difficile l'identification de marqueurs pathologiques distincts (associé à des échelles cliniques telles que la paranoïa, l'anxiété, ou la dépression par exemple). Notre première contribution fût donc de développer un algorithme linéaire de découverte de sous-groupes: UCSL qui se base uniquement sur des facteurs de variabilité existant spécifiquement et uniquement dans la population malade et non dans la population saine.
Dans le paradigme de la découverte de sous-types [20, 279, 280], les cliniciens (*par exemple : les dermatologues*) s'intéressent à découvrir des sous-groupes interprétables (*par exemple : les mélanomes*) au sein d'un groupe de patients partageant des motifs spécifiques à la maladie (*par exemple : texture, couleur, asymétrie*). Ce sous-domaine de l'analyse des données doit être différencié du regroupement (*clustering*), qui vise à découvrir des groupes d'échantillons sémantiquement similaires de manière non supervisée. Dans un contexte médical, les regroupements peuvent être guidés par des motifs sains (*par exemple : couleur de la peau, présence de cheveux, modèles de vieillissement*) communs aux sujets sains et aux patients. Ces facteurs sont donc sans pertinence à des fins de stratification.

1.3.1 UCSL: regroupement non-supervisé piloté par un signal supervisé pour la découverte de sous-groupe.

Pour exposer spécifiquement les sources pathologiques de variabilité, nous avons développé une méthode linéaire intitulée UCSL : Regroupement non supervisé guidé par l'apprentissage su-

pervisé (https://github.com/neurospin-projects/2021_rlouiset_ucsl/). UCSL effectue un regroupement dans un espace qui capture uniquement la variabilité utile pour la tâche de classification. Ainsi, les sous-types identifiés ne dépendent pas de la variabilité générale, telle que l'âge, le sexe ou le site d'acquisition. Mais plutôt de la variabilité spécifique, *c'est-à-dire* propre à la population pathologique. D'un point de vue mathématique, cette méthode peut être définie comme une tâche de regroupement guidée par l'apprentissage supervisé pour découvrir des sous-groupes en ligne avec la prédiction supervisée. Concernant le schéma d'optimisation, nous proposons un cadre d'ensemble général d'attente-maximisation. Nous proposons de construire un modèle non linéaire en fusionnant plusieurs estimateurs linéaires, un par cluster. Chaque hyperplan est estimé de manière à discriminer - ou prédire - correctement un seul cluster. De plus, pour effectuer une analyse de cluster dans un espace plus adapté, nous avons également proposé un algorithme de réduction de dimension qui projette les données sur un espace orthonormal pertinent pour la tâche supervisée. Plus de détails méthodologiques peuvent être trouvés dans la publication originale [179], acceptée à la conférence ECML-PKDD 2021.

1.3.2 Deep UCSL: extension de l'algorithme UCSL à l'aide un encodeur non-linéaire.

Afin d'étendre ce travail à des algorithmes d'analyse profond non-linéaire, potentiellement plus puissant, susceptibles de reconnaître des signatures pathologiques plus complexes, nous avons étendu l'algorithme UCSL à un algorithme de Deep Learning intitulé Deep UCSL. Deep UCSL est capable d'extraire directement des caractéristiques directement dans l'image IRM anatomique et démontre des capacités de généralisation à d'autres domaines d'imagerie médicale (pathologies de l'œil et des poumons).

Comme dans UCSL, nous sommes partis du constat que les méthodes de regroupement profond [40, 41, 169] donnent généralement des regroupements basés sur les facteurs de variation généraux (communs avec les sujets sains). Nous avons proposé de réutiliser le cadre mathématique et le processus d'optimisation d'UCSL (Expectation-Maximisation). Cependant, l'utilisation d'un extracteur de caractéristiques profondes plutôt que de modèles linéaires a complexifié le nombre de solutions possibles auxquelles nous pourrions converger (minima locaux). Nous avons donc jugé nécessaire d'ajouter plusieurs régularisations pour négliger la variabilité saine dans le processus d'entraînement. Ce travail méthodologique a été soumie au journal IEEE TMI au début de l'année 2024.

1.3.3 Identification de deux sous-groupes neuroanatomiques de troubles schizophrènes

Étant donné ces algorithmes, nous avons utilisé la méthode linéaire UCSL pour tenter d'identifier plusieurs sous-groupes dans une cohorte de personnes atteintes de trouble de la schizophrénie [292]. Les caractéristiques anatomiques ont été obtenues puis utilisées grâce au pipeline de traitement des IRM neuro-anatomiques CAT12 [98, 99]. De manière similaire à [45], deux sous-types de schizophrénie ont été identifiés, représentant respectivement 87 % et 13 % de la population schizophrène. Ces sous-types étaient répartis de manière égale en termes de distribution d'âge et de répartition des sexes. Le sous-groupe A fût determiné comme étant plus atrophié que le sous-groupe B en terme de quantité de matière grise. Des analyses statistiques ont été réalisées par rapport aux composantes, chaque composante étant associée à un sous-groupe. La composante A montre un déclin cognitif global similaire, et la composante B a révélé un déclin dans selon deux scores parmis neuf. Concernant les échelles cliniques psychiatriques, nous n'avons pas noté de différences entre les deux sous-groupes, si ce n'est que le cluster B était moins affecté sur l'échelle des symptômes généraux.

Il fût noté que les sous-groupes identifiés n'étaient pas nécessairement associés à des échelles cliniques psychatriques indépendantes. Cette limite est intéressante à noter car elle permet de réfléchir à des perspectives de travail qui permettraient d'intégrer ces propriétés, notamment l'identification de facteurs neurobiologiques sous-jacents distincts et interprétables, corrélant avec des échelles cliniques psychatriques indépendantes.

1.4 Développement et validation de méthodes d'analyse contrastante

Dans un second temps, nous nous sommes intéressés à des méthodes d'apprentissage de représentations ne nécessitant pas nécessairement de supposer l'existence de sous-groupes homogènes: l'analyse contrastante. Ce champ de l'apprentissage statistique vise à séparer les facteurs de variabilité «communs» et «cibles», étant donné un ensemble de données «source» et un ensemble de données «cibles». Dans notre cas, l'objectif serait d'identifier d'un côté la projection permettant d'identifier les motifs de variabilité sains et de l'autre côté la projection permettant d'identifier les «signatures pathologiques» qui n'existe que dans la classe des malades et non dans la classe des gens sains.

1.4.1 SepVAE: un auto-encodeur variationelle contrastant pour séparer les motifs de variabilité pathologiques et sains.

Les autoencodeurs variationnels (VAE) [150] ont le potentiel d'identifier des motifs de variabilité au sein d'un ensemble d'images unique. Cependant, séparer les facteurs de variabilité communs des facteurs pathologiques reste difficile. Pour cette raison, des VAE d'analyse contrastive ont été développés pour identifier les motifs uniques à un ensemble de données cible (TG) (*c'est-à-dire :* pathologique) par rapport à un ensemble de données de fond (BG) (*c'est-à-dire :* population en bonne santé). Inspirés par des idées précédentes, nous avons conçu un modèle qui est un cas particulier d'Auto-Encodeur Variationnel. Il comporte deux encodeurs (commun et saillant) et un seul décodeur. Des expériences sur des maladies mentales telles que la schizophrénie et le trouble autistique montrent que les vecteurs saillants de SepVAE prédisent mieux les variables spécifiques à la maladie (*c'est-à-dire :* SAPS (échelle des symptômes positifs), SANS (échelle des symptômes négatifs), ADOS (calendrier du diagnostic de l'autisme d'observation), ADI (calendrier du diagnostic de l'autisme par interview) et diagnostic). En revanche, les vecteurs saillants ne prédisent pas bien les variables démographiques : l'âge, le sexe et le site d'acquisition. Notre travail a été accepté à l'atelier ICML 2023 - Interpretable Machine Learning in Healthcare (IMLH 2023), à Hawaï, Honolulu. Ce travail a également été accepté à OHBM (Organization of Human Brain Mapping) 2023, à Montréal, Canada.

1.4.2 SepCLR: une méthode d'apprentissage contrastif de représentations pour séparer les motifs de variabilité pathologiques et sains.

L'Analyse Contrastive (CA) est un domaine de l'Apprentissage de Représentations qui vise à séparer les facteurs de variation communs entre un ensemble de données de fond (BG) (c'est-à-dire, des sujets sains) et un ensemble de données cible (TG) (c'est-à-dire, des patients) des facteurs propres à l'ensemble de données cible. Malgré sa pertinence, peu de modèles ont démontré des performances compétitives dans l'apprentissage de représentations sémantiquement expressives et dans la qualité de génération. En effet, actuellement, les méthodes de CA sont ou des modèles linéaires [327, 100], ou des Auto-Encodeurs Variationnels [3, 299, 328]. D'autres classes de méthodes génératives ou d'apprentissage de représentations existent et ont surpassé les méthodes linéaires et variationnelles en terme de performance dont de nombreux cas d'applications de traitemetnt d'images naturelles ou médicales. Notamment, les méthodes

génératives (GANs, modèles de diffusion) et les méthodes d'apprentissage de représentations (modèles d'apprentissage contrastif) ont produit des performances prometteuses récemment. Pour cette raison, nous avons proposé de tirer parti de la capacité de l'Apprentissage Contrastif (CL) à apprendre des représentations sémantiquement expressives pour effectuer une Analyse Contrastive (CA). Tout d'abord, nous avons détaillé un nouveau cadre théorique d'Information Mutuelle inspiré de [293] qui permet de récupérer et d'étendre des pertes contrastives récentes (InfoNCE [50], SupCon [153]). Ensuite, nous avons utilisé ce cadre pour développer une méthode d'Analyse Contrastive en distillant des hypothèses pertinentes. Enfin, nous avons introduit une nouvelle stratégie de minimisation de l'Information Mutuelle pour éviter les fuites d'informations entre les distributions communes et saillantes. Nous avons estimé les quantités statistiques d'intérêt à l'aide de fonctions de coût inspitré de l'apprentissage de représentation contrastif, en utilisant notamment des techniques d'Estimation de Densité par Noyaux. Nous avons ensuite validé notre méthode sur plusieurs jeux de données, dont un jeu de neuroimagerie en psychiatrie. Ce travail a été accepté à ICLR (International Conference on Learning Representations) 2024 - Viennes, Autriche, et au colloque IABM (Colloque Français d'Intelligence Artificielle en Imagerie Biomédicale) 2024 à Grenoble.

Chapter 2

Introduction & background

Summary of the chapter. This chapter describes the clinical heterogeneity of psychiatric diseases and emphasizes the need for biological insights to better understand these conditions. Despite efforts in recent years to unveil the biological underpinnings of psychiatric symptoms, significant inter-individual variability sources have been observed in the neuroanatomical readouts of both healthy and pathological cohorts, and have complicated the identification of consistent and interpretable biological markers.

This chapter describes several machine learning techniques aiming at parsing psychiatric disorder heterogeneity based on neuroimaging observations. Foremost, it details methods that mitigate the impact of known inter-individual sources of variability by adjusting observations per confounding factors (such as age, sex, or scanner type for ex.). Then, it motivates and describes methods that model the pathological variability with subgroups, or latent factors by contrasting with the general population.

Eventually, several arguments are given to motivate the development of Deep Learning methods for parsing the neuroanatomical heterogeneity in mental disorders.

Contents

Over the past century, research in psychiatric care has gained significant efforts from the scientific community. At the core of this interest lies the need for better recognition and characterization of mental disorders. Mental disorders encompass various conditions that are manifested by anomalous deviations in terms of behavior and cognitive patterns often associated with substantial distress or impairment for the concerned individual. Importantly, being diagnosed with a mental disorder induces an increased risk of developing further mental disorders as well as other comorbidity [205]. As a non-exhaustive list, comorbidity associated with a mental disorder encompasses premature death, suicide, substance abuse, and chronic physical illness.

According to a recent survey [306] led by the World Health Organization (WHO), in 2019, approximately 970 million people worldwide (1 in every 8 individuals) were experiencing a mental disorder, with depressive and anxiety disorders being prevalent compared to the other disorders. Besides affecting people's well-being, these conditions represent a non-neglectable cost for the community. In 2018, a URC-Eco study for the FondaMental Foundation [1] revealed that mental illnesses cost France approximately 160 billion euros. As a reason, about 12 million people were diagnosed with affection, such as severe depression, bipolar and schizophrenia disorders, obsessive-compulsive disorders, and anxiety disorders, among other mental health conditions.

2.1 Mental disorders organized as clinical categories

In the modern psychiatric routine, the diagnosis of mental disorders is conducted through interviews, questionnaires, and observations. In this setup, psychiatrists play a key role in performing assessments, utilizing various methods, such as cognitive and psychometric examinations [187]. These assessments aim to determine the presence, severity, frequency, and duration of diverse psychiatric symptoms. The assessment questionnaires are generally based on classification systems such as the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) [69], or International Classification of Diseases (ICD-11) [131] where a spectrum of symptoms are described and associated with typical mental health disorders. This approach has led to the organization of disorders into different clinical categories, citing: Attention-Deficit/Hyperactivity Disorder (ADHD), Autism Spectrum Disorder (ASD), schizophrenia (SZ), Bipolar Disorder (BD), and Major Depressive Disorders (MDD). These screening questionnaires and interviews permit clinicians to evaluate the psycho-pathological state of the patients and permit deter-

[1] https://www.fondation-fondamental.org/

mining a diagnosis and then a therapy or a treatment regime. However, multiple assessment tools exist, and using different assessment tools by different clinicians can potentially introduce variability and inconsistency in determining the diagnosis [302]. Besides, depending on the clinician's favored choice of assessment tool(s), or depending on whether the questionnaires or interviews are self-rated, parents-rated, or clinician-led, biases toward behavioral, emotional, cognitive, or physical manifestation can pollute the diagnosis assessment [187]. Therefore, the diagnosis is likely to be influenced by individual standpoints, potentially leading to variations in the categorization of the disorder.

2.2 The clinical heterogeneity of mental disorders

While the assessment of a psychiatric disorder diagnosis appears like a non-completely trivial task, the development of a precise nosology of psychiatric diseases is also a challenging task. From Kraepelin in 1883, the pioneer in this domain, until recently, psychiatric nosology has undergone successive modifications and has been marked by multiple refinements of the disorder's etiology. The latest version of the DSM (Diagnostic and Statistical Manual of Mental Disorders), the DSM 5-TR [70], released in 2022, provides a classification system that separates mental illnesses into diagnostic categories based on descriptions of symptoms and the course of the illness. Likewise, the latest International Classification of Disease-11 (ICD-11) [131], released in 2019, describes diagnostic categories similar to the DSM-5-TR. These successive modifications demonstrate how difficult, yet central, it is for psychiatrist experts to define mental disorder distinct classes.

Despite these successive refinements of their etiology, mental disorders such as Major Depressive Disorder, Bipolar Disorder, Post-Traumatic Stress Disorder (PTSD), and Schizophrenia still exhibit significant heterogeneity in their symptom presentations. For instance, depression is clinically diagnosed when a patient reports a minimum of five out of nine symptoms. This criterion potentially results in 256 distinct symptom combinations of alterations in mood, appetite, sleep, energy, and cognition, that meet the criteria for MDD, highlighting the multifaceted nature of depression [34], as also demonstrated in [216]. Similarly, Bipolar Disorder, which already generally admits two or three subtypes according to common knowledge, exhibits a potential diversity of symptom combinations in DSM-IV-TR, revealing a humongous number of possibilities, such as over 5 billion combinations [170]. PTSD, as outlined in DSM-5, presents over 600,000 ways in which symptoms can be combined, emphasizing the intricate variations within this disorder [137, 314]. In the case of schizophrenia, recognized for its clinical

heterogeneity [277], attempts have been made to stratify different behavioral phenotypes into subgroups [86], but this task remains challenging and unanswered.

2.3 Redefinition of disorders based on neurobiological roots of dimensional constructs

These observations raise concerns about the reliability of a nosology based on the assessment and the observation of exterior cognitive, behavioral, emotional, and physical symptoms. In 2005, Thomas R. Insel and Remi Quirion encouraged researchers to consider psychiatry as a clinical neuroscience discipline driven by brain-based observations [133]. Overall, several lines of research have provided considerable efforts to identify physiological biomarkers to better understand the underlying etiologies, aid the diagnosis assessment of mental health disorders, and guide the discovery of pharmacological intervention [1, 2, 96, 135, 304]. For example, several works motivated the seek for consistent disorder subtypes underpinned by objective and tangible biomarkers and associated with distinct sets of co-occurring symptoms [89, 57]. Another line of works, such as initiatives like the National Institute of Mental Health's Research Domain Criteria (RDoC) [134] and the European ROAdmap for MEntal Health Research (ROAMER) [250] encourage researchers to link symptom dimensions with biological systems to find *"new ways of classifying psychiatric diseases based on multiple dimensions of biology and behavior"*. These contributions aim to simplify and refine the psychiatric etiology based on brain-based biological readouts.

Searching for diagnostic biomarkers. Overall, the pursuit of biomarkers in psychiatry is driven by the critical need for objective measures to confirm the diagnosis and prognosis and improve treatment decisions in mental disorders [1, 2, 96, 135, 304]. In 2016, the FDA-NIH Biomarker Working Group [88] converged to a general definition of a "biomarker" as "a defined characteristic that is measured as an indicator of normal biological processes, pathogenic processes or responses to an exposure or intervention", a definition also introduced in [92], and [232]. Moreover, to guarantee the use of a biomarker in a clinical case, a relevant biomarker measure should be reproducible, consistent, coherently, and reliably evolve as the clinical condition progresses. This definition delineates the frame of what a biomarker is and paves the way toward the seeking of a characteristic that helps improve outcomes with a medical approach tailored to each individual [185, 188]. The US Food and Drug Administration (FDA) categorizes biomarkers based on their applications and how they could impact the clinical care of mental disorders. In practice, biomarkers can be classified into several categories [38, 96],

among which stand *prognostic biomarkers*, which anticipates the potential development of an illness for preventive interventions; *predictive biomarkers*, used to identify individuals that are likely to respond positively to a given therapeutic strategy, *diagnostic biomarkers*, which aims to detect or confirm the presence of a disease or medical condition. From a research stance, the latter category of biomarkers is particularly interesting as it could help identify subtypes of diseases as explained in [96]. Thus, the development of precision medicine could benefit from discovering diagnostic biomarkers to detect patients with a disease and refine its etiology. Identifying a precise etiology in psychiatry is critical since numerous diseases exhibit subtypes with distinct prognoses or responses to treatment. Therefore, diagnostic biomarkers could potentially be used in mental health clinical care as prognostic markers for early detection or as predictive markers to anticipate the individual response when exposed to a particular therapeutic strategy, as in depression, for example, [262].

The choice of a suitable modality. Numerous acquisition modalities emerge as good candidates to identify biomarkers, including various neuroimaging techniques [193, 165, 212], genomics technologies [83], molecule identification techniques [2], etc. However, to date, no measures have proven sufficiently reliable, valid, and useful to be adopted clinically. Nevertheless, neuroimaging emerges as a robust and promising modality for identifying biomarkers [1, 135, 304] to delineate psychiatric biotypes. These techniques, such as magnetic resonance imaging (MRI), electroencephalography (EEG), positron emission tomography (PET), and functional MRI (fMRI), provide a non-invasive means of investigating the results of the complex interactions between genetic predispositions and environmental factors [231, 111] via the brain's structural and functional changes. The information captured by these acquisition strategies allows researchers to discern subtle yet significant patterns, providing a basis for identifying distinct biotypes ("biologically distinctive phenotypes") within psychiatric populations to help understand the heterogeneous nature of psychiatric disorders. This thesis focuses on structural MRIs when parsing the biological heterogeneity of psychiatric disorder with statistical methods.

2.4 Individual-level prediction with Machine Learning (ML)

Group-level statistical methods are limited. Across recent years, numerous group-level statistical studies have been conducted to identify deviating neuroimaging patterns in brain disorders, including schizophrenia (SZ) [141, 255, 244, 163, 126, 282], Bipolar Disorder (BD) [114, 198, 120, 121, 172, 54], Autism Spectrum Disorder [204, 235, 290, 80, 81], Attention-

Deficit/Hyperactivity Disorder (ADHD) [143, 85, 283, 58], Major Depressive Disorders [177], and mild cognitive impairment (MCI), Alzheimer's disease (AD) [138]. These works have pointed out structural discriminative features (*i.e.*, grey matter atrophy) within patient groups compared to healthy cohorts. While these studies produce insights into identifying disease biomarkers, they generally fail to respect the clinical diagnosis or prognosis criteria in practice. This limitation arises because many findings, though statistically significant at the group level, generally fail at the individual level.

Individual-level inference with Machine Learning. Machine learning emerges as a compelling solution to produce individual-level predictions from brain imaging data [14, 195]. Whereas traditional statistical approaches focus on identifying statistically significant effects at the group level, machine learning models are statistical methods trained to predict specific clinical statuses (such as diagnosis, prognosis, or other phenotype variables) from individual entries. Once trained, these models can extrapolate and generalize predictions for novel entries, providing a personalized prediction based on the brain imaging features. This distinction is important, as explained in Bzdok et al. [37], which emphasizes that statistical significance at the group level does not necessarily enable a high prediction accuracy in independent data:"machine learning and classical statistics do not judge data on the same aspects of evidence: an observed effect assessed to be statistically significant by a p-value does not in all cases yield a high prediction accuracy in new, independent data, and vice versa".

Training Machine Learning methods beyond the case-control paradigm. However, even though Machine Learning classification methods can produce individual-level predictions, they can still fail when they are trained in a case/control paradigm [14]. As a reason, Marquand et al [195] explain "the case-control paradigm induces an artificial symmetry such that both cases and controls are assumed to be well-defined entities", even though psychiatric disorders are diagnosed based on overlapping and heterogeneous symptoms that overlap between disorders. Several strategies can be employed to reduce the neurobiological heterogeneity of clinical cohorts.

2.5 Parsing disorder heterogeneity with Machine Learning (ML)

To date, several families of Machine Learning methods to parse heterogeneity in mental disorders have emerged. Among these methods, neural basis component analysis and clustering methods emerged in the early 2010s.

Clustering methods. Data-driven clustering methods have emerged from 2010 to 2015 to stratify clinical groups based on neuroanatomical or neuro-functional measures in attention-deficit/hyperactivity disorder [143, 85, 283, 58], mood disorders [162, 284], and schizophrenia [32, 36]. These approaches are useful to stratify the disorder population. However, experiments have shown that mere clustering or components-identification algorithms applied to patients or healthy controls usually discriminate between young and older subjects or male and female subjects. This result suggests that the neuroanatomical variability is dominated by *general*, *non-pathological* factors such as aging [**?**]. Varol et al. [280] also identify this concern: "Such an approach aims to cluster brain anatomies instead of pathological patterns. Thus, it has the potential risk of estimating clusters that reflect normal inter-individual variability, some of which is due to sex, age, and other confounds, instead of highlighting disorder heterogeneity". To investigate the heterogeneity of mental disorders, the researcher's efforts focused on parsing *pathological variability* of the neuro-anatomical features, *i.e.* patterns of variability that only exist in the pathological population. To this end, several methods have been proposed to disregard *general variability* factors.

Component discovery methods. Another line of research, entitled "component discovery methods" in this thesis, assumes that mental disorders are underpinned by a mixture of neural bases (*i.e.* dimensions) associated with clinical dimensions. Several methods have attempted to discover pathological components via latent factor analysis [143, 242, 218]. These approaches are aligned with the dimensional approach (*i.e.*, assuming a continuous spectrum across brain disorders, potentially sharing common dimensions) as promoted by the RDoC initiative [134]. In practice, these works may still seek to discover disorder subtypes from the estimated low-dimensional factors [143, 242, 218].

2.6 Identification of the disorder's specific variability by adjusting for confounding factors or covariates

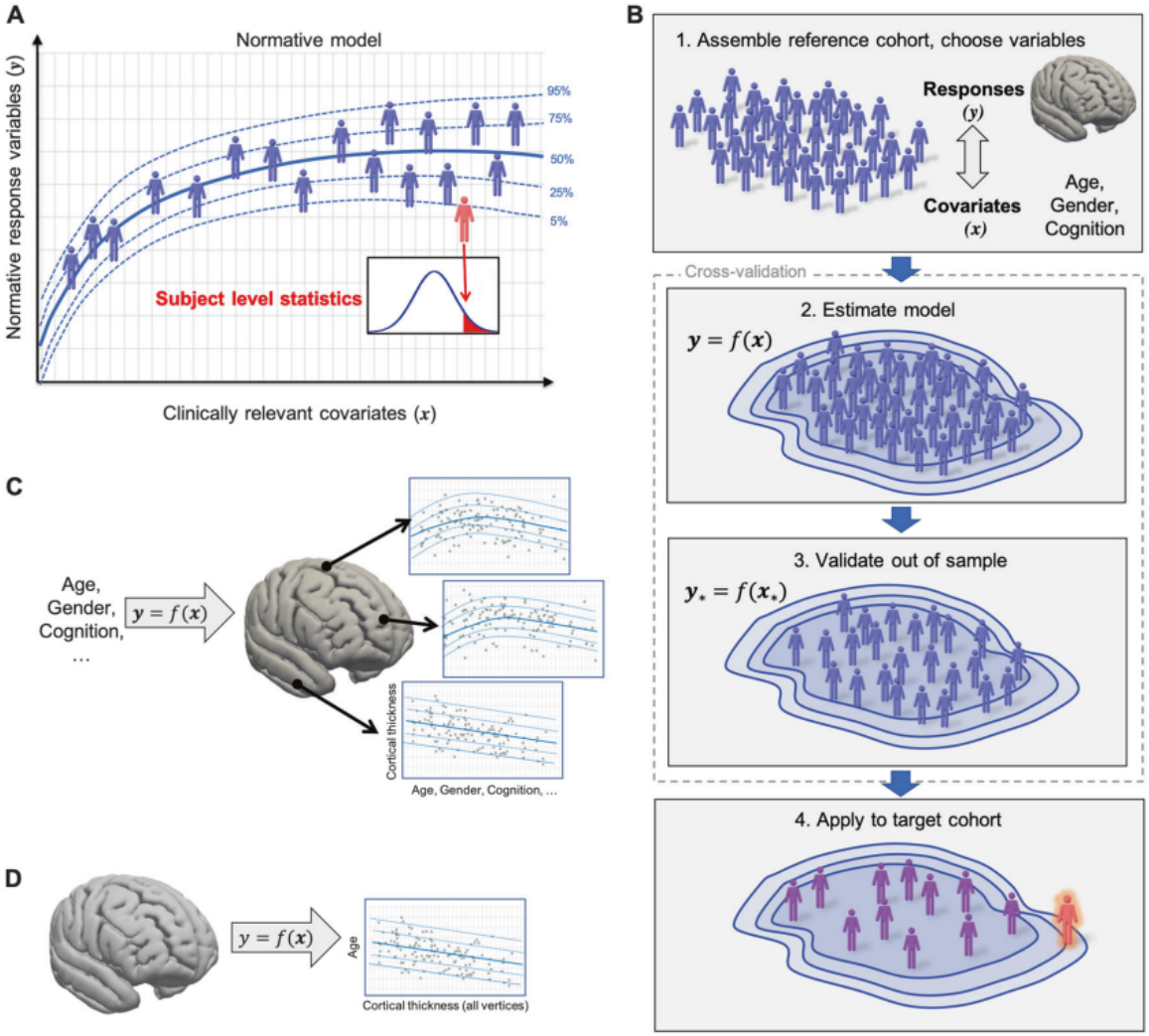
Residualization adjusted methods: Residualization of phenotypes aims to remove (or adjust for) confounding factors or covariates [289, 94, 95, 103]. These approaches are particularly relevant to produce neuro-anatomical features that are not driven by aging, sex, or acquisition site for example. However, they generally require the practitioner to assume that the confounding factors are either linearly related to the input features or additive offsets. Besides, residualization-adjusted features may remain rooted in non-specific factors which may undermine their relevancy. This issue was observed in 2021 by Iftimovici et al. [**?**], where pathological subtypes discovered based on covariate-adjusted neuroanatomical features (residualization-adjusted on age, sex, acquisition site, medication, and substance use) remained driven by a physiological variability that also exists in the healthy cohort. This result suggests that irrelevant confounding factors may not be strictly limited to known covariates such as age, sex, or acquisition sites as other unknown irrelevant general variability factors may come into play to undermine the relevancy of the heterogeneity analysis. Aside from unknown confounding factors, other known covariates can be considered (*e.g.*, education, ethnicity, urbanicity, etc.), but unavailable for training or inference.

Normative modeling: Normative modeling [195, 194, 196] provides a statistical framework to establish connections between demographic or clinical characteristics and quantitative biological measures, offering estimates of variation centiles within the population as demonstrated in Fig 2.1. A conceptual overview of normative modeling is provided in Fig 2.1.

Normative modeling closely relates to the use of growth charts in pediatric medicine. In the computational psychiatry application, the chosen measure (*e.g.* height or weight) is generally a quantitative brain-based biological readout (*e.g.* local volume of gray matter volume of multiple Regions of Interest). The adjustment covariable is generally chosen among demographic variables (*e.g.* age, or sex for ex.). Once estimated on a healthy cohort, a normative model estimates the normative distribution of a biological readout given demographic attributes. Ultimately, the estimated model can be applied to a target cohort (*e.g.*. pathological clinical cohort) to observe biological deviations given known covariates (*i.e.* given age and sex) at the individual level (red figure). Mathematically, normative models can be described by a set of functions ($y = f(x)$) that predicts the neurobiological response variables y from the clinical covariates x.

This approach enables examining and quantifying individual differences and emerges as a relevant candidate to parse individual-level heterogeneity across cohorts. By delivering statistical inferences at the individual level, normative modeling quantifies the extent to which each participant deviates from the normative pattern. A common use for normative modeling in psychiatry

Figure 2.1: Scheme of the concepts behind Normative Modeling from the foundation article of Marquand et al. [194]. A. Normative modeling provides statistical inference at the level of each subject with respect to the normative model. B. 4 steps training and evaluation process of a normative model. C. Mathematically, normative models can be described by a set of functions ($y = f(x)$) that predicts the neurobiological response variables y (*e.g.:* cortical thickness measures, local gray matter volumes, gyrification indices...) from the clinical covariates x (*e.g.:* age, gender, cognition...). D. Alternatively, the neurobiological response variables can be considered as the covariate and vice-versa, which establish a link with brain age estimation methods.



[303, 317] utilizes age, sex, and/or acquisition site as covariates to predict a quantitative biological readout (*e.g.*, local gray matter volume per region-of-interest). This setup enables learning a healthy normative pattern with respect to age and sex covariates. Then, applied to a clinical cohort, it enables determining where patients lie on the healthy continuum given their age and sex. In this application, normative modeling is particularly relevant to compute deviation indices for each input feature (*e.g.*, local gray matter volume of each region-of-interest). These deviation indices (or z-scores) can be further used as features to analyze the inter-individual heterogeneity that is not driven by the covariates, *i.e.,* age and sex. Such deviation indices can further serve as features for the identification of clusters (subtypes) or components (dimensions), which makes normative modeling a complementary approach to these techniques.
As for residualisation methods, normative models require the practitioner to have the covariate information both during training and inference. Interestingly, normative models do not require a particular assumption on the form of the dependency (linear dependency for ex.) between the covariates and the biological readouts (*i.e.* the input features). However, while normative modeling successfully reduces the *general variability* variance, it does not reduce the dimensionality of input features. Eventually, normative models are complementary to other analysis tools. Given covariate-adjusted deviation scores, various analysis tools can be used (such as component identification or subgroup discovery methods) to parse the heterogeneity of a population (by estimating components, physiological biotypes, or disorder subtypes, for example).

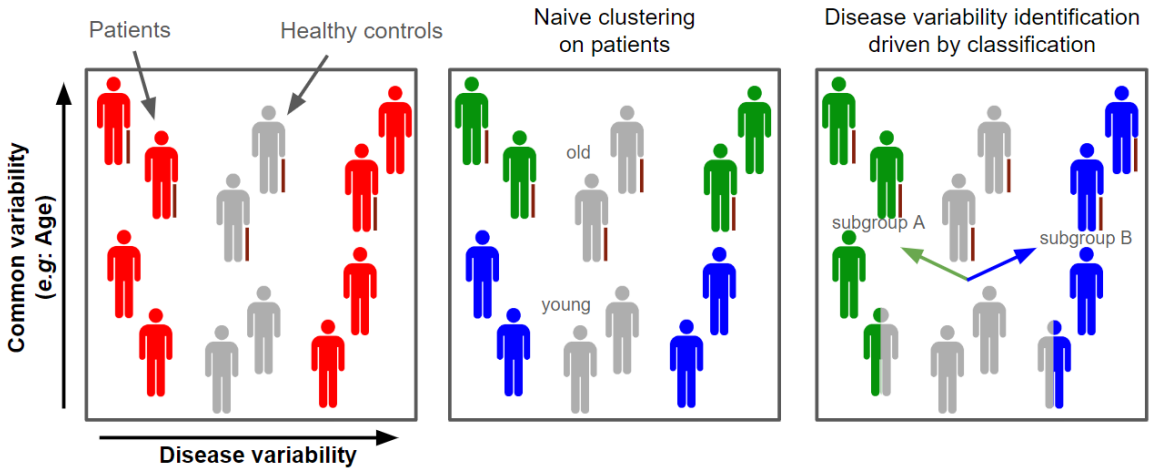
2.7 Identification of the disorder's specific variability by contrasting with the general population

2.7.1 Discovering disorder subgroups by contrasting with the general population

As an alternative to disregarding confounding covariates that are not necessarily known or available, methods to produce dimensions or clusters by contrasting with the *general population* have emerged. As an example, Subgroup Discovery methods [127, 279, 280, 300, 79, 91], extend clustering methods as they consist of finding consistent subgroups (and directions) within a population or a class of objects that are also relevant to a certain supervised upstream task. This means that the discovery of subtypes should not be fully unsupervised, as in standard clustering, but it should also be driven by a supervised task. For example, in clinical research,

it is essential to identify subtypes of patients with a given disorder (red icons in Fig. 2.2). The problem is that the general variability (that stems from age or sex) is observed in both healthy controls (grey icons in Fig. 2.2) and disorder patients. Therefore it will probably drive the clustering of patients toward a non-specific solution (second plot in Fig. 2.2). Adding a supervised task (healthy controls vs patients) can be used to find direction(s) (horizontal arrows) that discards non-specific variability to emphasize more disorder-related differences (right plot in Fig. 2.2).

Figure 2.2: Subtype discovery in clinical research. Given a healthy population (black) and a pathological population (red) (left plot), several homogeneous subgroups are assumed to exist within the disorder. However, the general variability (which stems from age or sex, for ex.) is observed in both healthy controls and patients. Therefore, a naive clustering of patients often yields a non-specific solution (middle plot). Nevertheless, the use of a classification task (healthy controls vs patients) helps to find direction(s) (horizontal arrow) that discards non-specific variability to emphasize more disorder-related differences (right plot).

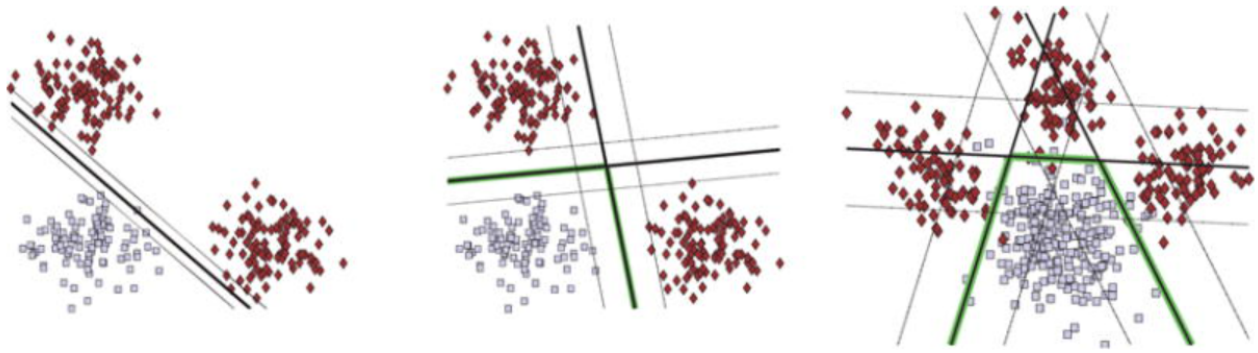


In 2016, Dong et al. [62, 127] proposed CHIMERA, a subtype (and dimension) discovery algorithm driven by supervised classification between healthy and pathological samples. It assumes that the pathological clusters can be modeled as a convex combination of linear transformations from the reference set of healthy subjects to the patient distribution, where each transformation corresponds to one pathological subtype. disorder subgroups are assumed to have the same variance. And the use of linear transformation enables providing pathological direction (or dimension) along with the disorder subgroups. However, this design choice is a strong assumption and limits the discoverable patterns to linear transformations only.

In 2017, Varol et al [279, 280] proposed a general Machine Learning algorithm that alternates between supervised learning and unsupervised cluster analysis where each step influences the

other until it reaches a stable configuration. The algorithm simultaneously solves binary classification and intra-class clustering in a hybrid fashion thanks to a maximum margin framework. The method discriminates healthy controls from pathological patients by optimizing the best convex polytope that is formed by combining several linear hyperplanes. The clustering ability is drawn by assigning patients to their best-discriminating hyperplane. Each cluster corresponds to one face of the piece-wise linear polytope and heterogeneity is implicitly captured by harnessing the classification boundary non-linearity. The efficiency of this method heavily relies on the prior hypothesis that healthy samples lie inside the convex discrimination polyhedron. This strong assumption may not hold for a given data set. Another hypothesis is that relevant psychiatric subtypes are equally severely affected. This prior implies that clusters should be along the classification boundary. Even though it may help circumvent general variability issues, this strongly limits the method's applicability to a specific variety of subgroups.

Figure 2.3: Scheme of HYDRA from the original contribution [280]. HYDRA parses the heterogeneity of the disorder dataset while classifying it from the healthy cohort. The healthy class is represented by the gray squares, it is separated from the heterogeneous disorder class denoted by red icons. (left) A Linear hyperplane separates the healthy class from a heterogeneous disorder population (two subtypes) by a small margin. (middle) HYDRA classifies each cluster independently, which enables a more confident inference associated with a larger margin. (right) The three subtypes model distinct deviations from the healthy population. A different face of the piece-wise linear convex polytope captures each deviation. Solid green lines correspond to the estimated discriminative polytope.



In 2020, Wen et al. [300] further used this algorithm to develop MAGIC: "Multi-scAle heterogeneity analysis and Clustering" that builds onto HYDRA. MAGIC addresses a limitation of both clustering and subgroup discovery linear machine learning methods. The authors argue that these methods generally model, learn, or infer from data using user-defined features from a standardized anatomical atlas, which limits the complexity and possibility of the discovered patterns. To be able to capture multi-scale patterns when parsing the disorder heterogeneity,

MAGIC [300] builds on HYDRA to derive clinically interpretable feature representations. To do so, the authors exploit a double-cyclic optimization procedure to enable the identification of inter-scale-consistent disorder subtypes. This work is interesting as it demonstrates the need for models that can potentially identify complex multi-scale, potentially non-linear patterns. However, such a method requires a complex optimization procedure with no convergence guarantees. Given its satisfactory results on case/control classification tasks, Deep Learning emerges as a more general and consistent candidate than such methods to identify non-linear complex biological patterns when parsing psychiatric disorder heterogeneity. Even though deep neural networks still suffer from their lack of interpretability.

2.7.2 Identifying the latent pathological factors using supervised strategy

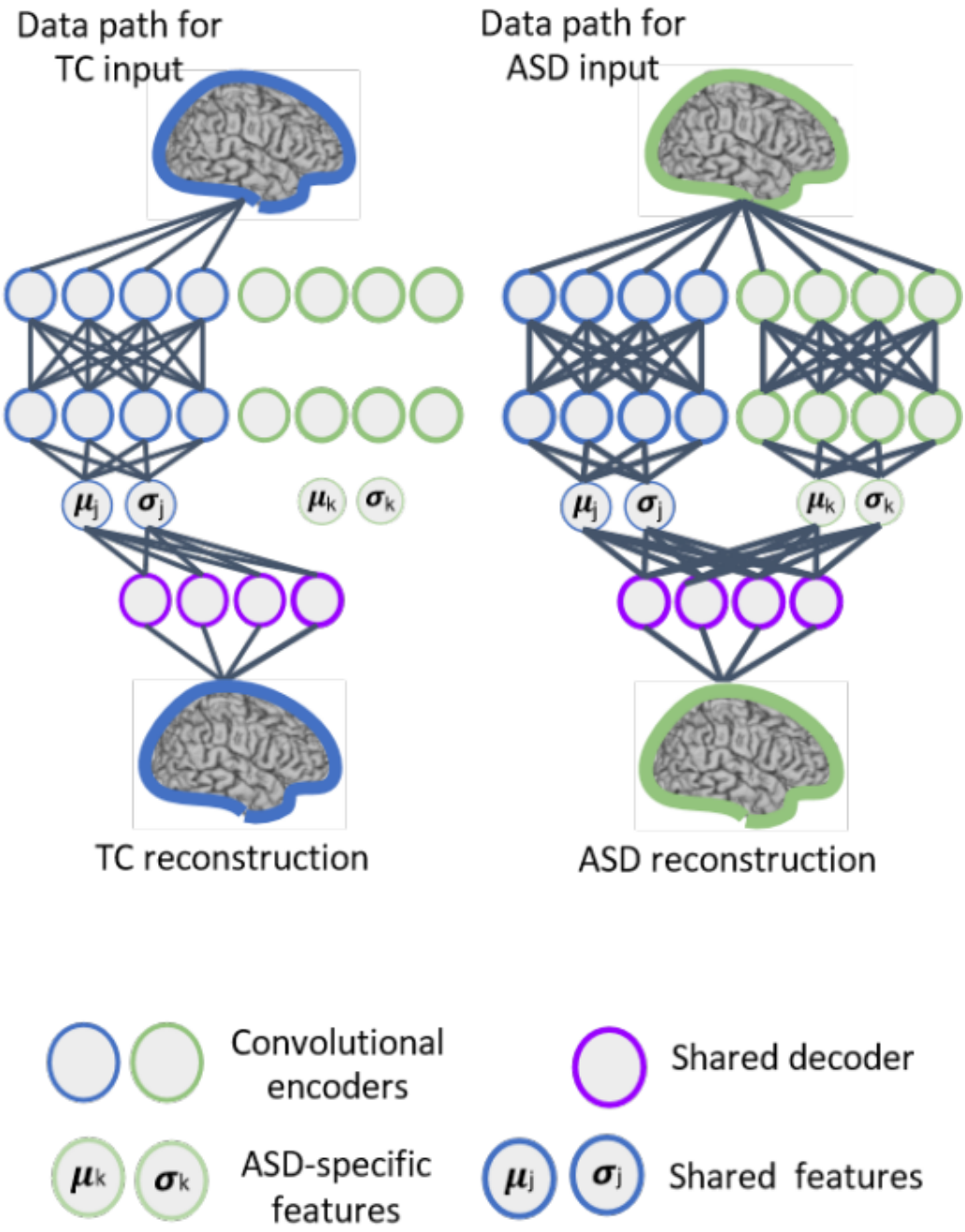
In 2017, Drysdale et al. [71], used functional magnetic resonance imaging (fMRI) in a large multisite dataset of 1,188 samples to discover two dimensions associated with depression and anxiety in depression. The authors used Canonical Correlation Analysis (CCA) in a supervised manner to define a low-dimensional representation of connectivity features that only relate to the clinical traits. Interestingly, this approach is a hybrid approach between clinical supervision and latent structure discovery as encouraged in [89]. Differently from HYDRA [280], it requires the supervision signal from distinct clinical traits of interest, *i.e.* anhedonia and anxiety clinical scale measures to identify two distinct and interpretable dimensions, one per clinical scale. Then, based on these two clinically supervised dimensions, the authors further estimated depression biotypes defined by distinct patterns of dysfunctional connectivity. However, these subtypes could not be further reproduced [66], suggesting that the connectivity heterogeneity in depression may be underpinned by a continuous spectrum rather than by homogeneous subgroups.

2.7.3 Identifying the latent pathological factors using contrastive analysis with the general population

Similarly, in 2022, Aglinskas et al. [6] applied a deep learning dimension discovery method to parse the heterogeneity of a cohort with autism disorder. The method they propose uses Contrastive Analysis Variational Auto-Encoders (CA-VAE). This class of methods requires two datasets, a healthy (or *"background"*) set of images, and a pathological (or *"target"*) set of images. It aims at separating the latent generation factors of the images into two parts, the healthy (or

"common") latent space, which captures the patterns of variability that exist in both datasets, and the pathological (or *"salient"*) latent space which captures the patterns of variability that exist in the pathological dataset only. The Fig. 2.4 depicts a scheme of the method employed in Aglinskas et al. [6]. In the original contribution, the authors show that their common vectors correlate with shared demographic variables such as age, sex, and scanner type but not with clinical scale measures. They also show that their salient vectors correlate with clinical scale measures such as ADOS, DSM IV, and Vineland. This class of methods is particularly relevant as it enables investigating the pathological factors of variability that drive the disorder heterogeneity without being flooded with *"general"* patterns of variability that may also exist in a healthy cohort. Indeed, the *"general"* patterns of variability appear irrelevant for parsing the disorder's heterogeneity. This kind of method does not require assuming that the heterogeneity of the disorder stems from homogeneous subgroups. **This thesis studies the potential of these methods, their performances on several neuroimaging applications, and their potential pitfalls and drawbacks**.

Figure 2.4: Scheme of CVAE, an instance of a Variational Autor-Encoder used for Contrastive Analysis. Scheme taken from the original contribution ([6]). The Contrastive Variational Autoencoder separates autism-specific variations from variations shared with healthy participants. The architecture employs a two-stream convolutional neural network built to disentangle shared and disorder-specific features. Healthy images are reconstructed using shared features, while disorder images are reconstructed using both shared and disorder-specific features.



2.8 Toward non-linear deep learning methods

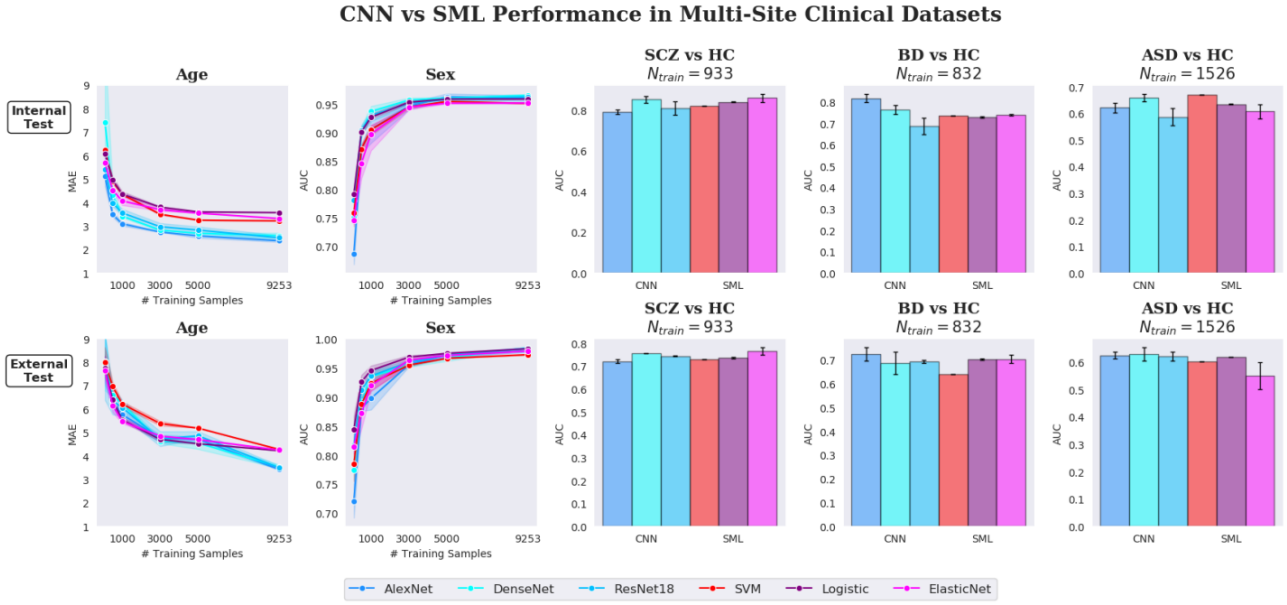
In the previous sections, various methods were detailed, linear, and deep, for parsing the heterogeneity within a pathological cohort. This section motivates the need for developing **deep learning methods for parsing the pathological heterogeneity**.

2.8.1 Comparing deep learning and standard machine learning

Machine Learning methods appear promising in neuroimaging, as they can produce performant individual-level predictions. Let us provide an overview of the capacity of various Machine Learning methods trained to predict demographic variables and psychiatric disorders diagnosis (in a simple case/control group classification paradigm). In Fig. 2.5, Benoit Dufumier (2022) [76] compared the performances of supervised Deep Learning and linear and kernel Machine Learning methods in the prediction of demographic variables and psychiatric disorder diagnosis on multi-site datasets. For Machine Learning methods, linear models with l1 (Lasso) and l1+l2 (ElasticNet) regularizations are evaluated, as well as kernel SVM with a Radial Basis Function (RBF) kernel. Concerning Deep Learning methods, the convolutional architecture [5] is compared with residual and dense architectures ResNet-18 and DenseNet-121 [73]. Both Deep Learning (DL) and Standard Machine Learning (SML) algorithms are trained on whole-brain Voxel-Base Morphometry [17] (VBM) preprocessed 3D neuro-anatomical MRIs and evaluated on two different test sets: an internal test set (with the same acquisition sites as in the training set) and an external test set (with different acquisition sites than the training set).
In Fig. 2.5, the authors observed that Machine Learning (Deep, Linear, and Kernel) models perform similarly in the sex prediction tasks with training samples up to 9253 and a binary AUC up to 95%. Also, the performances on the diagnosis prediction tasks are significantly different from the average random chance, with an AUC of 85% on SCZ vs. HC, 76% AUC on BD vs. HC, and 65% AUC on ASD vs. HC on the internal test set. These results are particularly relevant to developing this thesis's motivations and objectives as they show to which extent Machine Learning models can predict demographic variables and psychiatric diagnosis.

Deep Learning case/control classification fails to perform better than linear models. Nevertheless, in Fig. 2.5's results, the authors observed on heterogeneous classification setups (psychiatric diseases classification setups) that Deep Learning models generally perform as well as Linear or Kernel models, which suggests that they do not succeed in capturing additional highly non-linear patterns. This can be explained by the presence of noise in the input data

Figure 2.5: Overview of the performances of several Deep Convolutional Learning methods and Linear Machine Learning methods on the prediction of demographic variables: Age and Sex, and the prediction of common psychiatric disorders: HC/ASD (Healthy Controls vs Autism Spectrum Disorders), HC/BD (HC vs Bipolar Disorder), and HC/SZ (HC vs Schizophrenia Disorder) given Voxel-Base Morphometry [17] (VBM) preprocessed 3D neuro-anatomical MRIs. The internal validation and test sets for demographic variables comprise 662, and 655 samples, while the external test set comprises 640 samples. For psychiatric disorders, the internal validation and test sets of the SZ cohort respectively comprise 118, and 116 samples, while the external test set comprises 133 samples. The BD cohort's internal validation and test sets respectively comprise 107, and 103 samples, while the external test set comprises 131 samples. The internal validation and test sets of the ASD cohort respectively comprise 188, and 184 samples, while the external test set comprises 207 samples. For demographic variables (age, and sex), the author performed a 5-fold (resp. 3-fold) Monte Carlo Cross-Validation sub-sampling procedure for $N_{\text{train}} \in \{100, 500\}$ (resp. $N_{\text{train}} \in \{1000, 3000, 5000, 9253\}$. As for diagnosis classification tasks, each model is trained 3 times with different random initializations, and average and standard deviations are reported. Mean Absolute Error (MAE) is the reference measure for age prediction and Area Under the Curve (AUC) is for binary classification tasks. Credits to Dufumier et al. [76].



[247] and by the existence of high inter-individual heterogeneity in neuroanatomical images [189, 303, 317]. This assumption was notably formulated in 2020 by Schulz et al. [247, 248] who wrote: "High levels of noise in neuroimaging data may effectively linearize decision boundaries, potentially leaving little nonlinear structure for machine learning models to exploit". This

observation further motivates the need for more individual-level machine learning to parse the heterogeneity of the disorder while disregarding high inter-individual general variability.

Deep Learning performances increase as training samples increase. In Fig. 2.5, on the age regression task, the authors observed that Deep Learning models outperform Linear and Kernel Machine Learning models as the number of training samples increases. This behavior is promising, as it provides an encouraging result about the capacity of deep learning models to be better predictive models than traditional linear and kernel methods as the number of samples increases.

2.8.2 Leveraging large cohorts with transfer learning and deep neural networks

The crucial result pinpointed above motivates the use and development of deep learning methods as increasingly large neuroimaging datasets become progressively available. Several international initiatives (SchizConnect [292], ABIDE [67, 68], ENIGMA) furnished endeavors to retrospectively aggregate cohorts for each specific disorder. However, the considerable heterogeneity of these datasets still limits the potential of such initiatives.

Transfer Learning. A promising methodological perspective is to leverage large heterogeneous cohorts to enhance the performances of models tuned on more relevant but smaller pathological cohorts. This goal can be addressed using transfer learning strategies. In principle, Transfer Learning [44, 33, 313] consists of pre-training of a deep model on a source domain D_S with a source task T_S to produce a suitable representation to solve a target task T_T on a target domain D_T. As a source task, transfer learning strategies generally employ self-supervised learning, as well as multi-task learning (*e.g.* age regression and sex classification). In the context of transfer learning in neuroimaging, Dufumier et al. [77] proposed to learn a deep encoder from the healthy brain dataset and re-use this encoder as a weight initialization to better discriminate patients with brain disorders from healthy controls.
Recently, Edouard Duchesnay [72] has suggested two practical strategies for transferring deep learning models from a big cohort to a small pathological cohort. These ideas are illustrated in Fig. 2.6. In the upper part of the figure, step (1), a pre-training encoder E_u is trained in either i) a self-supervised paradigm (*e.g.*, by reconstructing the input with an auto-encoder or by producing a perturbation-invariant latent space that classifies instances with a contrastive learning method), or ii) performs multiple pretext tasks at the same time (*e.g.*, sex classification

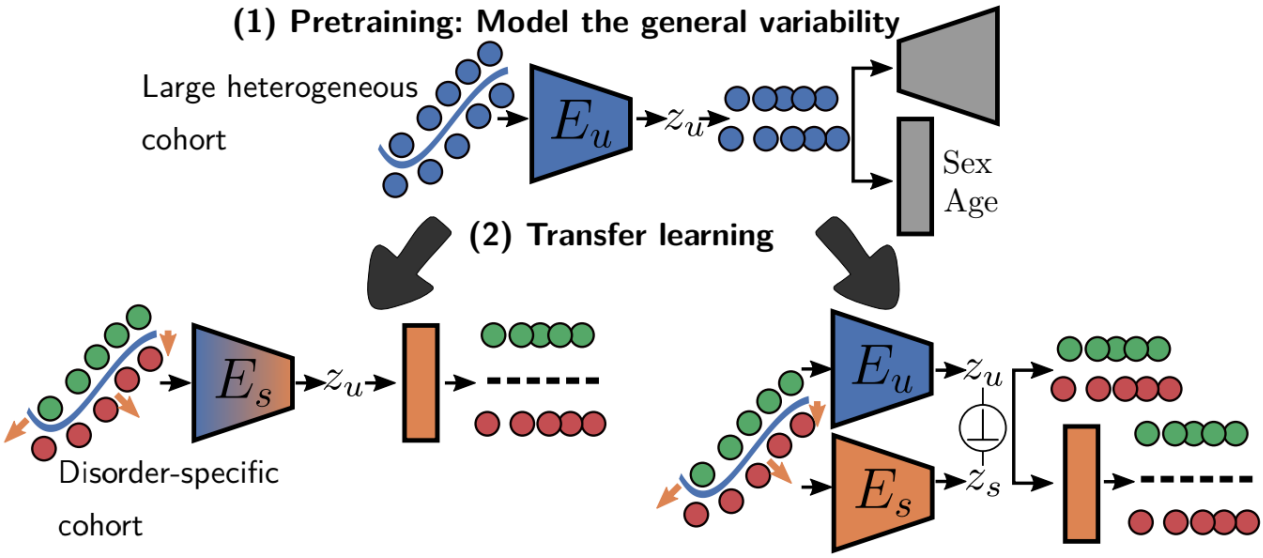


Figure 2.6: Scheme depicting two strategies of Transfer Learning from large general variability healthy cohorts to small pathological cohorts. In step (1), a model learns from a large heterogeneous cohort on a self-supervised task, such as a reconstruction task or a demographic attributes supervised task. In step (2), the learned representation is transferred to a disorder-specific cohort on either a classification task (left) or a contrastive analysis task (for ex.) (right). It is generally assumed that features learned during pre-training will be re-used during fine-tuning on the target task, either by enhancing the classification performance or by providing a representation that captures the general variability, which is useful for a contrastive analysis task (for ex.). Credits to [72].

and age regression). In the lower part of the figure, step (2), the pre-trained encoder is either i) (left) fine-tuned to predict a pathological diagnosis (in general, the encoder is expected to re-use already learned low-level features to generalize better on a low-data regime task), or ii) (right) frozen while a second and independent *disorder-specific* encoder E_s learns a latent space z_s that focuses on the patterns of variability that dominate the pathological cohort (*e.g.* with Contrastive Analysis methods).

In Fig. 2.6, the lower left transfer learning strategy has been investigated in depth by Dufumier Benoit [76]. The strategies he has developed rely on self-supervised learning techniques, from the Contrastive Learning paradigm, and also relate to multi-task learning, since attributes, such as age and sex, have been used as supervision signals in the training process [73]. After fine-tuning, performance gains on the classification of several brain disorders (Schizophrenia, Bipolar Disorder, Autism, and Alzheimer's disease) have been observed when using these transfer learning strategies. These results justify and motivate the use of Deep Learning in neuroimaging

data analysis, as deep representation encoders are particularly suited for Transfer Learning. Nevertheless, the fine-tuning strategy employed by Dufumier merely remains a case/control discrimination setup, which appears limited in the case of heterogeneous psychiatric disorders, [14, 37, 195]. Still, these observations motivate the need for developing **a deep learning method for parsing the pathological heterogeneity**, which could serve as an efficient fine-tuning strategy.

2.9 Thesis objectives and contributions

Identifying precise and subtle diagnostic signatures for psychiatric diseases relies on establishing robust correlations between objective biomarkers and observable symptoms. This thesis focuses on identifying neuroanatomical patterns extracted from structural T1w-MRI data. However, understanding the complex latent mechanisms contributing to neuroanatomical variability in psychiatric disorders is challenging due to inter-individual differences and intra-disorder heterogeneity. The primary objective of this Ph.D. project is to implement a neuroanatomical biomarker-based method to parse the heterogeneity of mental disorders, aiming to uncover correlations between disorder subtypes and symptoms or treatment responses. Related lines of research have revealed that commonly used data-driven stratification algorithms often distinguish between young and older subjects, indicating a significant influence of inter-individual non-specific factors such as age, sex, acquisition site, or other confounding factors in neuroanatomical variability. By developing novel machine learning methodologies, this thesis aims to discern neuroanatomical patterns characterizing pathological biotypes or driving independent variability factors within each psychiatric disorder. This approach aims at enhancing our understanding of disorder mechanisms and provides novel insights into understanding the neuroanatomical heterogeneity of each mental disorder.
In a nutshell, the research objectives in this thesis are:

1. Exploring deep and linear Subtype Discovery methods producing interpretable subgroups within a group of patients that share disorder-specific patterns. Indeed, previous works have shown that mere clustering algorithms applied to patients or healthy controls usually discriminate between young and older subjects. Some other works [279, 280, 62, 127] proposed subgroup discovery methods but required constraining assumptions about the topology of the pathological population, the size of the subgroups, their pathological severity, or the linearity of the pathological patterns. This thesis proposed novel linear [179] and deep models (Deep UCSL, submitted at IEEE TMI) for subtype discovery. These models aim to stratify each mental disorder while disregarding confounding factors such as aging. The linear model was then applied to experimental datasets of patients with schizophrenia to investigate potential neuro-anatomical biotypes of patients (paper in preparation for a computational psychiatry research journal).

2. Exploring robust and reproducible Contrastive Analysis methodologies in neuroimaging for psychiatric research. Indeed, Contrastive Analysis techniques can potentially separate the neuro-anatomical patterns that patients share with healthy controls from those that

are proper to the pathology. And allows for the finding of pathological dimensions that organize the disorder population in a continuous spectrum, rather than in homogeneous subgroups. In this thesis, methodologies have been developed for such methods, empowered with various representation learning techniques, such as Variational Auto-Encoder [180], and Contrastive Representation Learning methods [181]. Using the variational encoder technique, two important cost functions for this type of method were highlighted: 1) a non-linear classification cost function in the pathological representation space and 2) a function to minimize mutual information between the common and pathological representation spaces. Then, this method was validated on several datasets, including two from neuropsychiatry acquired using neuro-anatomical MRI. Regarding the contrastive representation learning technique, a novel objective to maximize using mutual information was formulated. Then, statistical quantities of interest were estimated using cost functions inspired by contrastive representation learning. This method was then evaluated on several datasets, including one in neuroimaging in psychiatry.

2.10 Thesis organization

Chapter 2. This chapter provides the context, motivation, methodological background, and the thesis objectives. Additionaly, Chapter A of the Appendix, pinpoints several key points required to understand this thesis, such as the data acquisition process, the pre-processing, the features extraction methods, and the description of some relevant Machine Learning algorithms used in this thesis.

Chapter 3. This chapter provides the thesis contributions in Subgroup Discovery:

1. Sec. 3.1 corresponds to the paper "UCSL: A Machine Learning Expectation-Maximization Framework for Unsupervised Clustering Driven by Supervised Learning" [179], published in the proceedings of ECML-PKDD 2021.

2. Sec. 3.2 corresponds to the paper "Deep UCSL: Unsupervised Discovery of Disease Subtypes by Contrasting with Healthy Controls" submitted to the journal IEEE TMI in February 2024.

3. Sec. 3.3 corresponds to an article entitled "The discovery of two schizophrenia biotypes suggested by Machine Learning." in preparation for a computational psychiatry research journal.

Chapter 4. This chapter provides the thesis contributions in Contrastive Analysis:

1. Sec. 4.1 corresponds to the paper "SepVAE: A contrastive VAE to separate pathological patterns from healthy ones." [180], accepted to ICML 2023 IMLH Workshop and in rebuttal for MIDL 2024.

2. Sec. 4.2 corresponds to the paper "SepCLR: Separating common from salient patterns with Contrastive Representation Learning." [181] published in the proceedings of ICLR 2024.

Chapter 5. This chapter concludes on how this thesis contributes to the actual state of knowledge described in the literature and offers perspectives and openings about future works and ideas to fulfill the objectives laid out and described by the scientific community.
In this manuscript, I propose the reader to browse each of my articles in chronological order. Luckily, how my supervisors and I organized my Ph.D. project is suitable for stacking articles because each paper naturally flows from the previous contribution by extending it, either with applied or methodological works. Of note, Section 3.3 corresponds to an unpublished article yet, but will eventually be submitted to a journal in computational psychiatry research.

2.11 Publications

This PhD has led to several publications in peer-reviewed journals and international conferences.
Published papers:

1. Auriau, Pierre and Grigis, Antoine and Dufumier, Benoit and Louiset, Robin and Gori, Pietro and Mangin, Jean-François and Duchesnay, Edouard, **Supervised diagnosis prediction from cortical sulci: toward the discovery of neurodevelopmental biomarkers in mental disorders.** under review. for *IEEE International Symposium on Biomedical Imaging 2024 (IEEE ISBI 2024).*

2. Louiset, Robin and Gori, Pietro and Grigis, Antoine and Duchesnay, Edouard, **SepCLR - Separating common from salient patterns with Contrastive Representation Learning**. In *International Conference on Learning Representations (ICLR 2024)*, https://arxiv.org/abs/2402.11928.

3. Carton, Florence and Louiset, Robin and Gori, Pietro, **Double InfoGAN for Contrastive Analysis**. In *Artificial Intelligence and Statistics (AISTATS 2024)*, https://arxiv.org/abs/2401.17776.

4. Louiset, Robin and Gori, Pietro and Dufumier, Benoit and Grigis, Antoine and Duchesnay, Edouard, **A contrastive VAE to separate pathological patterns from healthy ones.**, in *Workshop on Interpretable ML in Healthcare at International Conference on Machine Learning (ICML 2023).*
https://arxiv.org/abs/2307.06206

5. Dufumier, Benoit and Carlo Alberto Barbano and Louiset, Robin and Grigis, Antoine and Duchesnay, Edouard and Gori, Pietro, **Integrating Prior Knowledge in Contrastive Learning with Kernel.** in *International Conference on Machine Learning (ICML 2023)*,
https://arxiv.org/abs/2206.01646

6. Louiset, Robin and Gori, Pietro and Dufumier, Benoit and Houenou, Josselin and Grigis, Antoine and Duchesnay, Edouard, **UCSL: A Machine Learning Expectation-Maximization Framework for Unsupervised Clustering Driven by Supervised Learning.** in *European Conference on Machine Learning and Principles and Practices of Knowledge Discovery in Data (ECML-PKDD 2021)*,
https://arxiv.org/abs/2107.01988

Under review:

1. Louiset, Robin and Gori, Pietro and Dufumier, Benoit and Grigis, Antoine and Duchesnay, Edouard, **Deep UCSL: Unsupervised Discovery of Disease Subtypes by Contrasting with Healthy Controls.**, submitted to *IEEE Transactions on Medical Imaging (IEEE TMI).*

2. Dufumier, Benoit and Gori, Pietro, and Victor, Julie and Louiset, Robin, and Mangin, Jean-François, and Grigis, Antoine, and Duchesnay Edouard, "Deep Learning Improvement over Standard Machine Learning in Anatomical Neuroimaging comes from Transfer Learning", submitted to *Nature Machine Intelligence.*

In preparation:

1. Louiset, Robin, and Iftimovici, Anton and Gori, Pietro and Grigis, Antoine and Duchesnay, Edouard, **The discovery of two schizophrenia biotypes suggested by Machine Learning.**, in prep. for a journal in computational psychiatry research.

Chapter 3

Subgroup identification in medical imaging and neuroimaging

Chapter summary. This chapter focuses on subtype discovery methods that stratify diseases into homogeneous subgroups while discriminating with healthy controls.
The first section described in this chapter is a *linear* subgroup discovery algorithm, called UCSL (Unsupervised Clustering driven by Supervised Learning). This algorithm is based on a general subgroup discovery statistical framework and enables identifying subgroups that stem only from the pathological variability specific to the disorder while disregarding the common variability shared with the healthy population. This algorithm is then validated on a synthetic experiment, a controlled digit subgroups discovery task on MNIST, and a neuropsychiatric subgroup discovery task.
The second section of the chapter introduces Deep UCSL, which extends UCSL with a non-linear deep features extractor, potentially more powerful in recognizing complex pathological signatures. This novel deep learning method can directly extract features from anatomical MRI images, showed state-of-the-art results in neuro-psychiatric subgroup identification, and demonstrated generalization capabilities to other medical imaging domains (eye and lung pathologies).
Eventually, to illustrate the usefulness of such Subgroup Discovery methods, the linear method UCSL was leveraged to identify biological subtypes ("biotypes") in a cohort of individuals with schizophrenia, and correlations are computed to analyze the clinical relevance of the two discovered subtypes.

Contents

3.1 UCSL: an Unsupervised Clustering driven by Supervised Learning framework

3.1.1 Abstract

Subtype Discovery consists in finding interpretable and consistent sub-parts of a dataset, which are also relevant to a certain supervised task. From a mathematical point of view, this can be defined as a clustering task driven by supervised learning in order to uncover subgroups in line with the supervised prediction. In this paper, we propose a general Expectation-Maximization ensemble framework entitled UCSL (Unsupervised Clustering driven by Supervised Learning). Our method is generic, it can integrate any clustering method and can be driven by both binary classification and regression. We propose constructing a non-linear model by merging multiple linear estimators, one per cluster. Each hyperplane is estimated so that it correctly discriminates - or predicts - only one cluster. We use SVC or Logistic Regression for classification and SVR for regression. Furthermore, to perform cluster analysis within a more suitable space, we also propose a dimension-reduction algorithm that projects the data onto an orthonormal space relevant to the supervised task. We analyze our algorithm's robustness and generalization capability using synthetic and experimental datasets. In particular, we validate its ability to identify suitable consistent sub-types by conducting a psychiatric disorder cluster analysis with known ground-truth labels. The proposed method's gain over previous state-of-the-art techniques is about +1.9 points in terms of balanced accuracy. Finally, we make codes and examples available in a scikit-learn-compatible Python package: https://github.com/neurospin-projects/2021_rlouiset_ucsl/

3.1.2 Introduction

Subtype discovery is the task of finding consistent subgroups within a population or a class of objects that are also relevant to a certain supervised upstream task. This means that the definition of homogeneity of subtypes should not be fully unsupervised, as in standard clustering, but it should also be driven by a supervised task. For instance, when identifying flowers, one may want to find different varieties or subtypes within each species. Standard clustering algorithms are driven by features that explain most of the general variability, such as the height or the thickness. Subtype identification aims at discovering subgroups describing the specific heterogeneity within each flower species and not the general variability of flowers. To disentangle these sources of variability, a supervised task can identify a more relevant feature

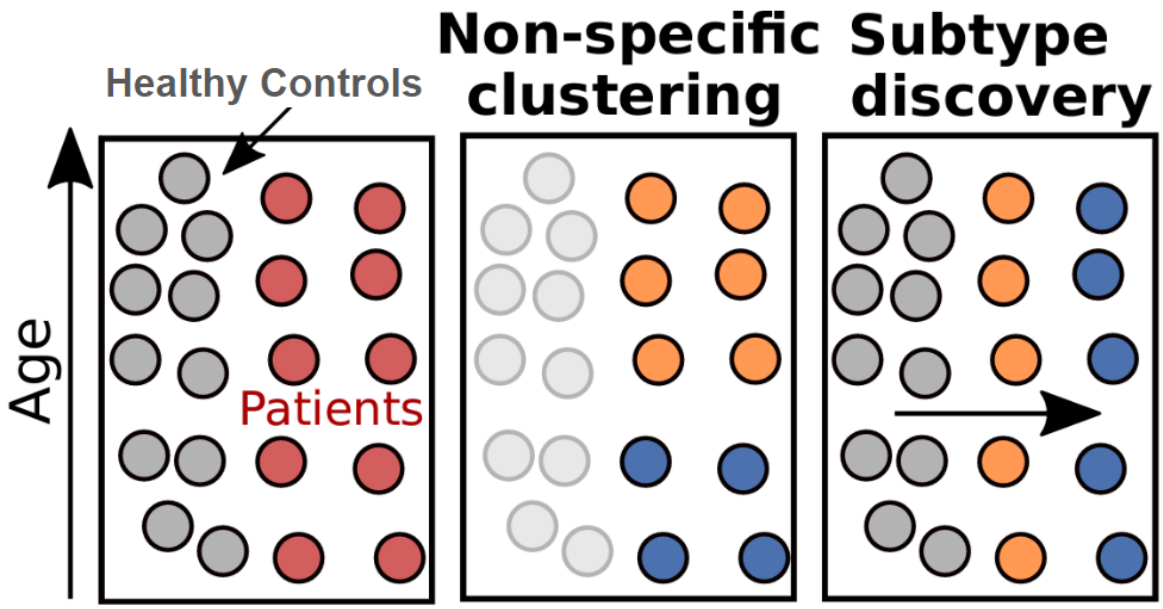


Figure 3.1: Subtype discovery in clinical research.

space to drive the intra-species clustering problem. Depending on the domain, finding relevant subgroups may turn out to be a relatively hard task. Indeed, most of the time, boundaries between different patterns are fuzzy and may covariate with other factors. Hence, ensuring that resulting predictions are not collapsed clusters or biased by an irrelevant confound factor is a key step in the development of such analysis. For example, in clinical research, it is essential to identify subtypes of patients with a given disorder (red dots in Fig. 3.1). The problem is that the general variability (that stems from age or sex) is observed in both healthy controls (grey dots in Fig. 3.1) and disorder patients. Therefore it will probably drive the clustering of patients toward a non-specific solution (second plot in Fig. 3.1). Adding a supervised task (healthy controls vs patients) can be used to find direction(s) (horizontal arrow Fig. 3.1) that discards non-specific variability to emphasize more disorder-related differences (subtype discovery in Fig. 3.1). This is a fundamental difference between unsupervised clustering analysis and subtype identification.

Subgroup identification is highly relevant in various fields such as in clinical research where disorder subtype discovery can lead to better-personalized drug treatment and prognosis [307] or to better anticipate at-risk profiles [296]. Particularly, given the extreme variability of cancer, identifying subtypes enables the development of precision medicine [43, 307, 207, 197, 225]. In psychiatry and neurology, different behavior, anatomical and physiological patterns point out variants of mental disorders [194] such as for bipolar disorder [310], schizophrenia [127, 45], autism, [316, 268], attention-deficit hyperactivity disorder [291], Alzheimer's disease [90, 309,

280, 300] or Parkinson's disease [84]. In bioinformatics, DNA subfolds analysis is a key field for the understanding of gene functions and regulations, cellular processes, and cell subtyping [261]. In the field of data mining, crawling different consistent subgroups of written data enables enhanced applications [229].

3.1.3 Related works

Early works [43, 84] proposed traditional clustering methods to find relevant subgroups for clinical research in cancer and neurology. However, they were very sensitive to high-dimensional data and noise, making them hardly reproducible [215, 225]. To overcome these limits, [261] and [225] evaluated custom consensus methods to fuse multiple clustering estimates to obtain more robust and reproducible results. Additionally, [261] also proposed to select the most important features to overcome the curse of dimensionality. Even if all these methods provide relevant strategies to identify stable clusters in high-dimensional space, they do not allow the identification of disease-specific subtypes when the dominant variability in patients corresponds to the variability in the general population. To select disease-specific variability, recent contributions propose hybrid approaches integrating a supervised task (patient vs. controls) to the clustering problem. In [249], authors propose a hybrid method for disorder subtyping in precision medicine. Their implementation consists of training a Random Forest supervised classifier (healthy vs. diseased) and then applying SHAP algorithm [183, 182] to get explanation values from Random Forest classifiers. This yields promising results even though it is computationally expensive, especially when the dataset size increases.

Differently, a wide range of Deep Learning methods propose to learn better representations via deep encoders and adapt the clustering method on compressed latent space or directly within the minimizing loss. In this case, encoders must be trained with at least one non-clustering loss, to enhance the representations [243] and avoid collapsing clusters [308]. [40] proposes a Deep Clustering framework that alternates between latent cluster estimation and likelihood maximization through pseudo-label classification. Yet, its training remains unstable and is designed for large-scale datasets only. Prototypical Contrastive Learning [169], SeLA [16], SwAV [41] propose contrastive learning frameworks that alternatively maximize 1- the mutual information between the input samples and their latent representations and 2- the clustering estimation. These works have proven to be very efficient and stable on large-scale datasets. They compress inputs into denser and richer representations and successfully eliminate unnecessary noisy dimensions. Nevertheless, they still do not propose a representation aligned with the supervised task at-hand. To ensure that resulting clusters identify relevant subgroups for the supervised

task, one could first train for the supervised task and then run clustering on the latent space. This would emphasize important features for the supervised task but it may also regress out intra-class specific heterogeneity, hence the need of an iterative process where clustering and classification tasks influence each other.
CHIMERA [127], proposes an Alzheimer's subtype discovery algorithm driven by supervised classification between healthy and pathological samples. It assumes that the pathological heterogeneity can be modeled as a set of linear transformations from the reference set of healthy subjects to the patient distribution, where each transformation corresponds to one pathological subtype. This is a strong a priori that limits its application to (healthy reference)/(pathological case) only. [280, 300] propose an alternate algorithm between supervised learning and unsupervised cluster analysis where each step influences the other until it reaches a stable configuration. The algorithm simultaneously solves binary classification and intra-class clustering in a hybrid fashion thanks to a maximum margin framework. The method discriminates healthy controls from pathological patients by optimizing the best convex polytope that is formed by combining several linear hyperplanes. The clustering ability is drawn by assigning patients to their best discriminating hyperplane. Each cluster corresponds to one face of the piece-wise linear polytope and heterogeneity is implicitly captured by harnessing the classification boundary non-linearity. The efficiency of this method heavily relies on the prior hypothesis that negative samples (not being clustered) lie inside the convex discrimination polyhedron. This may be a limitation when it does not hold for a given data-set (left examples of Fig. 3.2). Another hypothesis is that relevant psychiatric subtypes should not be based on the disorder severity. This a priori implies that clusters should be along the classification boundary (upper examples of Fig. 3.2). Even though it may help circumvent general variability issues, this strongly limits the method's applicability to a specific variety of subgroups.

Contributions:

Here, we propose a general framework for Unsupervised Clustering driven by Supervised Learning (UCSL) for relevant subtypes discovery. The estimate of the latent subtypes is tied to the supervision task (regression or classification). Furthermore, we also propose to use an ensembling method in order to avoid trivial local minima or collapsed clusters.
We demonstrate the relevance of the UCSL framework on several data-sets. The quality of the obtained results, the high versatility, and the computational efficiency of the proposed framework make it a good choice for many subtype discovery applications in various domains. Additionally, the proposed method needs very few parameters compared to other state-of-the-

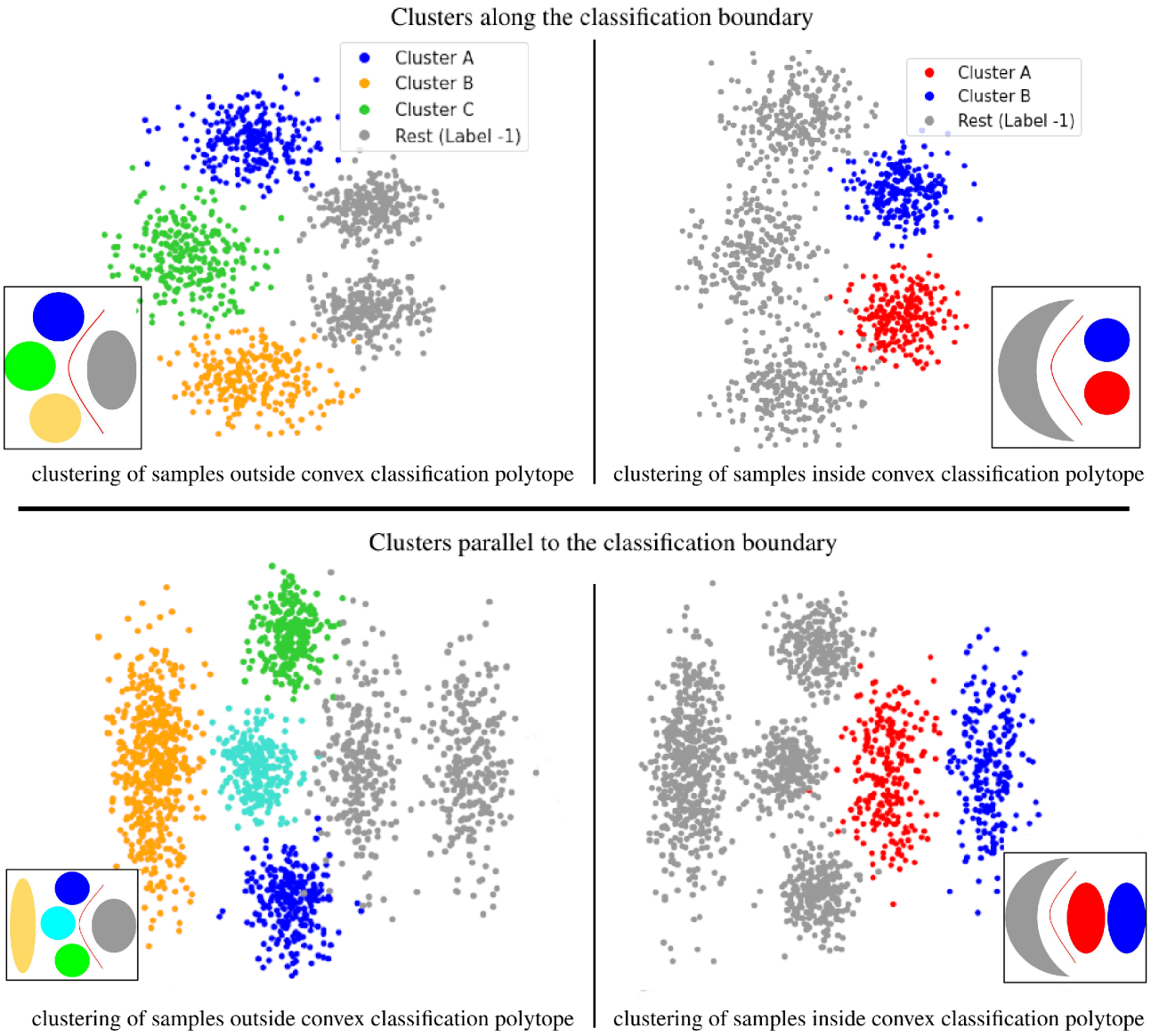


Figure 3.2: Toy Datasets - Different configurations we want to address. Grey points represent negative samples. The upstream task is to classify negative (grey) samples from all positive (colored) samples, while the final goal is to cluster positive samples. The upper plots show 3 and 2 clusters respectively along the classification boundary. The lower plot shows 4 and 2 clusters parallel (and along on the left) to the classification boundary. Furthermore, plots on the left and right show clusters outside and inside the convex classification polytope respectively.

art (SOTA) techniques, making it more relevant for a large number of medical applications where the number of training samples is usually limited. Our three main contributions are :

1. A generic mathematical formulation for subtype discovery that is robust to samples inside

and outside the classification polytope (see Fig.3.2).

2. An Expectation-Maximization (EM) algorithm with an efficient dimensionality reduction technique during the E step for estimating latent subtypes more relevant to the supervised task.

3. A thoughtful evaluation of our subtype discovery method and a fair comparison with several other SOTA techniques on both synthetic and real data-sets. In particular, a neuroimaging data-set for psychiatric subtype discovery.

3.1.4 Unsupervised Clustering driven by Supervised Learning

A statistical formulation for Subtype Discovery

Let $(X, Y) = \{(x_i, y_i)\}_{i=1}^n$ be a labeled data-set composed of n samples. Here, we will restrict to regression, $y_i \in \mathcal{R}$, or binary classification, $y_i \in \{-1, +1\}$. We assume that all samples, or only positive samples ($y_i = +1$), can be subdivided into latent subgroups for regression and binary classification, respectively.
The membership of each sample i to latent clusters is modeled via a latent variable $c_i \in C = \{C_1, ..., C_K\}$, where K is the number of assumed subgroups. We look for a discriminative model that maximizes the joint conditional likelihood:

$$\sum_{i=1}^{n} \log \sum_{c \in C} p(y_i, c_i | x_i) \tag{3.1}$$

Directly maximizing this equation is hard, and it would not explicitly make the supervised task and the clustering depend on each other, namely we would like to optimize both $p(c_i|x_i, y_i)$ (the clustering task) and $p(y_i|x_i, c_i)$ (the upstream supervised task) and not only one of them. To this end, we introduce Q, a probability distribution over C, so that $\sum_{c_i \in C} Q(c_i) = 1$.

$$\sum_{i=1}^{n} \log \sum_{c \in C} p(y_i, c_i | x_i) = \sum_{i=1}^{n} \log \left(\sum_{c \in C} Q(c_i) \frac{p(y_i, c_i | x_i)}{Q(c_i)} \right). \tag{3.2}$$

By applying the Jensen inequality, we then obtain the following lower-bound:

$$\sum_{i=1}^{n} \log \left(\sum_{c \in C} Q(c_i) \frac{p(y_i, c_i | x_i)}{Q(c_i)} \right) \geq \sum_{i=1}^{n} \sum_{c \in C} Q(c_i) \log \left(\frac{p(y_i, c_i | x_i)}{Q(c_i)} \right), \tag{3.3}$$

It can be shown that equality holds when:

$$Q(c_i) = \frac{p(y_i, c_i|x_i)}{\sum_{c \in C} p(y_i, c_i|x_i)} = \frac{p(y_i, c_i|x_i)}{p(y_i|x_i)} = p(c_i|y_i, x_i). \tag{3.4}$$

The right term of Eq. 3.3 can be re-written as:

$$\sum_{i=1}^{n} \sum_{c \in C} \Big(Q(c_i) \log \Big(p(y_i|c_i, x_i) p(c_i|x_i) \Big) - Q(c_i) \log Q(c_i) \Big). \tag{3.5}$$

We address the maximization of Eq. 3.5 with an EM optimization scheme (algo. 2) that exploits linear models to drive the clustering until we obtain a stable solution. First, during the Expectation step, we tighten the lower bound in Eq. 3.3 by estimating Q as the latent clusters conditional probability distribution $p(c_i|y_i, x_i)$ as in Eq. 3.4. Then, we fix Q, and maximize the supervised conditional probability distribution $p(y_i|c_i, x_i)$ weighted by the conditional cluster distribution $p(c_i|x_i)$ as in Eq. 3.5.

Expectation step

In this step, we want to estimate Q as $p(c_i|y_i, x_i), \forall i \in [\![1, n]\!], \forall c \in C$ in order to tighten the lower bound in Eq. 3.3. We remind here that latent clusters c are defined only for the positive samples ($y = +1$), when dealing with a binary classification, and for all samples in case of regression. Let us focus here on the binary classification task. Depending on the problem one wants to solve, different solutions are possible. On the one hand, if ground truth labels for classification are *not* available at inference time, Q should be computed using the classification prediction. For example, one could use a clustering algorithm only on the samples predicted as positive. However, this would bring a new source of uncertainty and error in the subgroups discovery due to possible classification errors. On the other hand, if ground truth labels for classification are available at inference time, one would compute the clustering using only the samples associated with ground-truth positive labels $\tilde{y}_i = +1$, and use the classification directions to guide the clustering. Here, we will focus on the latter situation since it interests many medical applications.

Now, different choices are again possible. To influence the resulting clustering with the label prediction estimation, HYDRA [280] proposes to assign each positive sample to the hyperplane that best separates it from negative samples (i.e. the furthest one). This is a simple way to align resulting clustering with estimated classification while implicitly leveraging classification

boundary non-linearity. Yet, we argue that this formulation does not work in the case where clusters are disposed parallel to the piece-wise boundary as described in Fig. 3.3. We propose to project input samples onto a supervision-relevant subspace to overcome this limit before applying a general clustering algorithm.

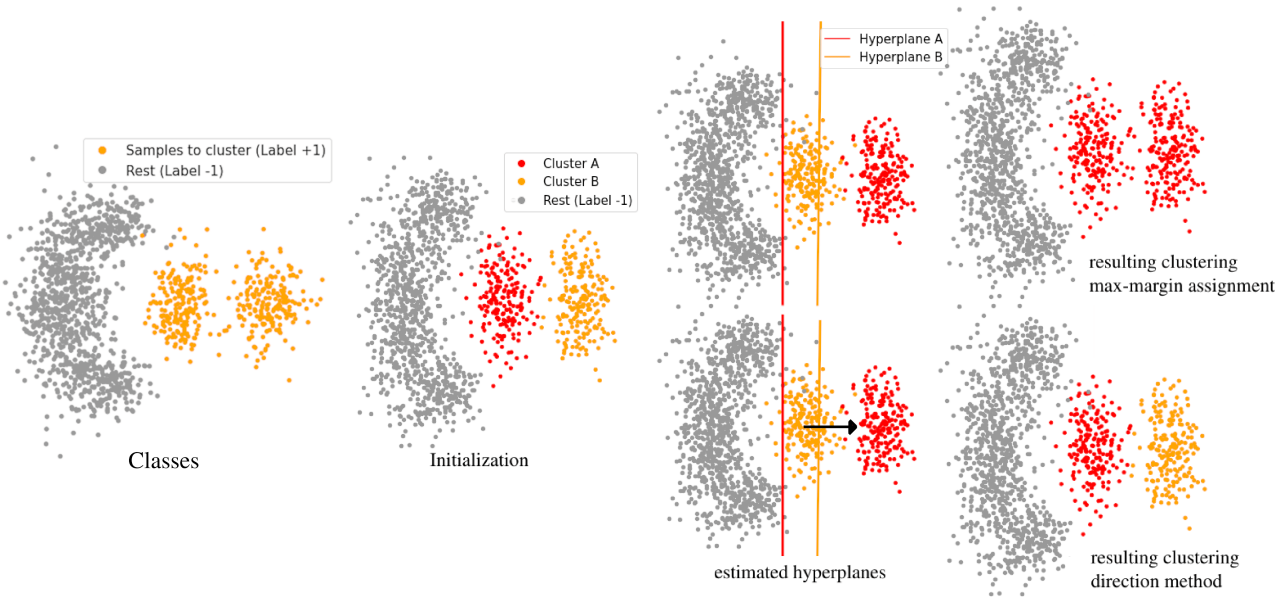


Figure 3.3: Limit of maximum-margin-based clustering starting from an optimal cluster initialization. When the separation of clusters to discover is co-linear to the supervised classification boundary, the maximum margin cluster assignment (as in [280]) converges towards a degenerate solution (upper figures). Instead, with our direction method (lower figures), the Graam-Schmidt algorithm returns one direction where input points are projected to and perfectly clustered.

Dimension reduction method based on discriminative directions: Our goal is a clustering that best aligns with the upstream task. In other words, in a classification example, the discovery of subtypes should focus on the same features that best discriminate classes, and not on the ones characterizing the general variability. In regression, subgroups should be found by exploiting features that are relevant to the prediction task. In order to do that, we rely on the linear models estimated from the maximization step. More specifically, we propose first creating a relevant orthonormal sub-space by applying the Graam-Schmidt algorithm to all discriminant directions, namely the normal directions of estimated hyperplanes. Then, we project input features onto this new linear subspace to reduce the dimension and perform cluster analysis on a more suitable space. Clustering can be conducted with any algorithm such as Gaussian Mixture Models (GMM), K-Means (KM) or DBSCAN for example.

Algorithm 1 Dimension reduction method based on discriminative directions

Input : $X \in \mathbf{R}^{n\text{x}d}$, training data with n samples and d features.
Output : $X' \in \mathbf{R}^{n\text{x}K}$, training data projected onto relevant orthonormal subspace.

1: Given K estimated hyperplanes, concatenate normal vectors in $D \in \mathbf{R}^{K\text{x}d}$.
2: Ortho-normalize the direction basis D with Graam-Schmidt obtaining $D^{\perp} \in \mathbf{R}^{K\text{x}d}$.
3: Project training data onto the orthonormal subspace, $X' = X(D^{\perp})^T$.

Maximization step

After the expectation step, we fix Q and then maximize the conditional likelihood. The lower bound in Eq. 3.5 thus becomes:

$$\sum_{i=1}^{n} \sum_{c \in C} Q(c_i) \log p(y_i|c_i, x_i) + \sum_{i=1}^{n} \sum_{c \in C} Q(c_i) \log p(c_i|x_i) \tag{3.6}$$

Here, we need to estimate $p(c_i|x_i)$. A possible solution, inspired by HYDRA [280], would be to use the previously estimated distribution $p(c_i|y_i, x_i)$ for the positive samples and a fixed weight for the negative samples, namely:

$$p(c_i|x_i) = \begin{cases} p(c_i|x_i, y_i) & \text{if } \tilde{y}_i = +1 \\ \frac{1}{K} & \text{if } \tilde{y}_i = -1 \end{cases} \tag{3.7}$$

However, as illustrated in Fig. 3.4, this approach does not work well when negative samples lie outside of the convex classification polytope since discriminative directions (or hyperplanes) may become collinear. This collinearity hinders the retrieving of informative directions and consequently degrades the resulting clustering.

To overcome such a shortcoming, we propose to approximate $p(c_i|x_i)$ using $p(c_i|x_i, y_i)$ for both negative and positive samples or, in other words, to extend the estimated clustering distribution to all samples, regardless their label y. In this way, samples from the negative class ($y_i = -1$), that are closer to a certain positive cluster, will have a higher weight during classification. As shown in Fig. 3.4, this results in classification hyperplanes that correctly separate each cluster from the closer samples of the negative class, entailing better clustering results. From a practical point of view, since we estimate $Q(c_i)$ as $p(c_i|y_i, x_i)$, it means that $p(c_i|x_i)$ can be approximated by $Q(c_i)$. $Q(c_i)$ being fixed during the M step, only the left term in Eq. 3.6 is maximized.

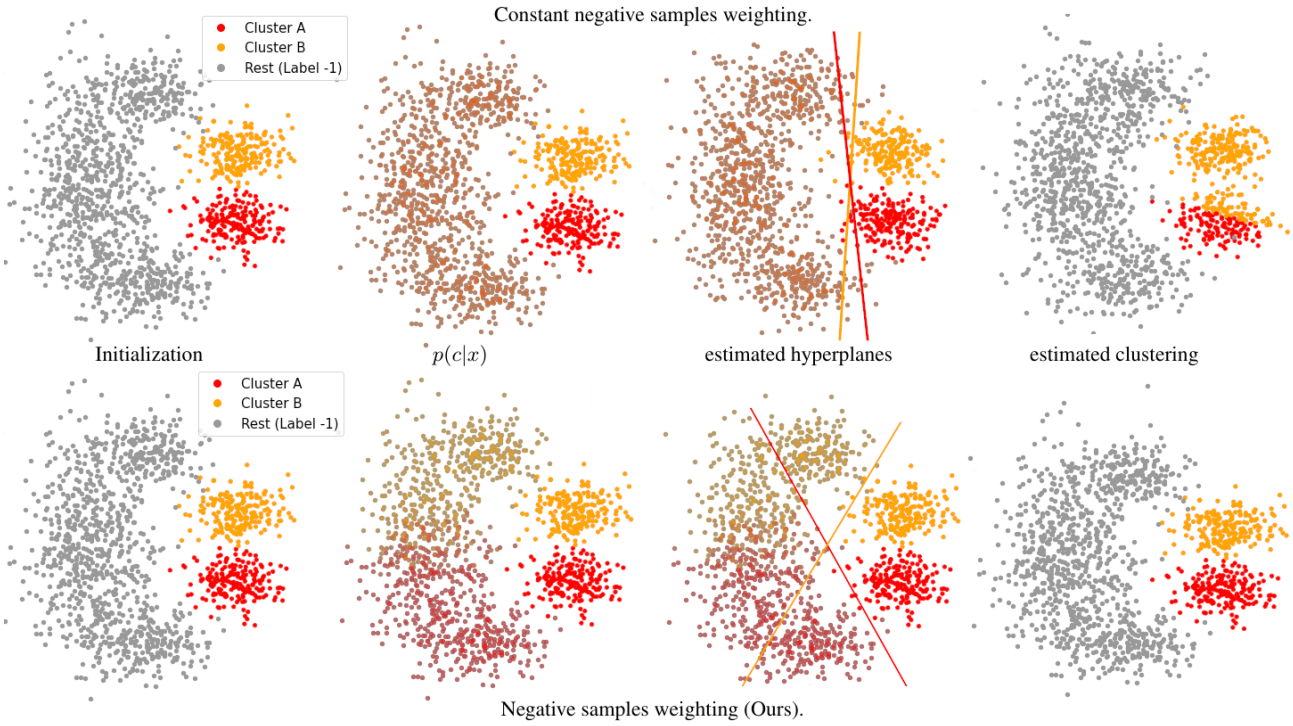


Figure 3.4: Starting with an optimal initialization of clusters to discover, constant negative samples weighting (top row) may lead to co-linear discriminative hyperplanes and thus errors in clustering. Conversely, our negative samples weighting enforces non-colinearity between discriminative hyperplanes resulting in higher quality clustering.

Supervised predictions

Once trained the proposed model, we compute the label y_j for each test sample x_j using the estimated conditional distributions $p(y_j|c_j, x_j)$ and $p(c_j|x_j)$ as:

$$p(y_j|x_j) = \sum_{c_j \in C} p(y_j, c_j|x_j) = \sum_{c_j \in C} p(y_j|x_j, c_j)p(c_j|x_j) \tag{3.8}$$

We thus obtain a non-linear estimator based on linear hyperplanes, one for each cluster.

Application

Multiclass case: In the case of classification, we handle the binary case in the same way as [280] does. We consider one label as positive $\tilde{y}_i = 1$ and cluster it with respect to the other one $\tilde{y}_i = -1$. Using the one-vs-rest strategy, we can cast it as several binary problems in the multi-class case.

Ensembling with Spectral clustering: The consensus step enables the merging of several different clustering propositions to obtain an aggregate clustering. After running the EM iterations N times, the consensus clustering is computed by grouping samples assigned to the same cluster across different runs. In practice, we compute a co-occurrence matrix between all samples. And then we use co-occurrence values as a similarity measure to perform spectral clustering. Hence, for example, given two samples i and j and 10 different runs, if samples i and j ended up 4 times in the same cluster, the similarity measure between those 2 samples will be $\frac{4}{10}$. Given an affinity matrix between all samples, we can then use the spectral clustering algorithm to obtain a consensus clustering.

UCSL's Pseudo-code

The pseudo-code of the proposed method UCSL (Alg. 2) can be subdivided into distinct steps:

1. **Initialization:** First, we have to initialize the clustering. There are several possibilities here, we can use traditional ML methods such as KM or GMM. For most of our experiments, we used GMM.

2. **Maximization:** The Maximization step consists in training several linear models to solve the supervised upstream problem. It can be either a classification or a regression. We opted for well-known ML linear methods such as logistic regression or max-margin linear classification method as in [280].

3. **Expectation:** The Expectation step uses the supervised learning estimates to produce a relevant clustering. In our case, we exploit the directions exhibited by the linear supervised models. We project samples onto a subspace spanned by those directions to perform the unsupervised clustering with positive samples.

4. **Convergence:** To check the convergence, we compute successive clustering Adjusted Rand Score (ARI). The closer this metric is to 1, the more similar both clustering assignments are.

5. **Ensembling:** Initialization and EM iterations are performed until convergence N times, and an average clustering is computed with a Spectral Clustering algorithm [280], [225] that proposes the best consensus. This part enables us to have more robust and stable solutions avoiding trivial or degenerate clusters.

Algorithm 2 UCSL general framework pseudo-code

Input : $X \in \mathbf{R}^{nxd}$, $y \in \{-1, 1\}^n$, K number of clusters.
Output : $p(c|x, y) = Q(c)$, $p(y|x, c)$ (linear sub-classifiers).

for ensemble in n_ensembles **do**
  Initialization: Estimate $Q^{(0)}$ for all samples ($y = \pm 1$) with a clustering algorithm (e.g. GMM) trained with positive samples only ($y = +1$).
  while not converged **do**
    M step (supervised step) :
      Freeze $Q^{(t)}$
      for k in $[1, K]$:
        Fit linear sub-classifier k weighted by $Q^{(t)}[:, k]$ (Eq. 3.6).
      end for
    E step (unsupervised step) :
      Use Alg.1 to obtain $X' \in \mathbf{R}^{nxK}$ from sub-classifiers normal vectors $D \in \mathbf{R}^{Kxd}$.
      Estimate $Q^{(t+1)} = p(c|x, y)$ (Eq. 3.4) for all samples with a clustering algorithm trained on X' with positive samples only.
  end while
end for
Ensembling: Compute average clustering with the ensembling method (Sec. 3.1.4).
Last EM: Perform EM iterations from ensembled latent clusters until convergence.

3.1.5 Results

We validated our framework on four synthetic data-sets and two experimental ones both qualitatively and quantitatively.

Implementation details: The stopping criteria in Alg. 2 is defined using the ARI index between two successive clusterings (at iteration t and iteration $t+1$). The algorithm stops when it reaches the value of 0.85. In the MNIST experiment, convolutional generator and encoder networks have a similar structure to the generator and discriminator in DCGAN [227]. We trained it during 20 epochs, with a batch size of 128, a learning rate of 0.001 and with no data augmentation and a SmoothL1 loss. More information can be found in the Supplementary material. Standard deviations are obtained by running 5 times the experiments with different initializations (synthetic and MNIST examples) or using a 5-fold cross-validation (psychiatric dataset experiment). MNIST and synthetic examples were run on Google Colaboratory Pro, whose hardware equipments are PNY Tesla P100 with 28Gb of RAM.

Synthetic dataset: First, we generated a set of synthetic examples that sum up the different configurations on which we wish our method to be robust: subtypes along the supervised boundary or parallel to it. We designed configurations with various number of clusters, outside or inside the convex classification polytope. UCSL was run with Logistic Regression and GMM. To make our problem more difficult we decided to add noisy unnecessary features to the original 2-D toy examples. For each example and algorithm, we performed 10 runs with a different initialization each time (GMM with only one initialization) and we did not perform the ensembling step for fair comparison with the other methods. We compared with other traditional ML methods such as KM GMM, DBSCAN, and Agglomerative Clustering. Results are displayed in Fig. 3.5. For readability, we divided the standard deviation hull by 2. Compared with the other methods, UCSL appears to be robust to unnecessary noisy features. Furthermore, it performs well in all configurations we addressed.

MNIST dataset: To further demonstrate what intra-class clustering could be used for, let us make an example from MNIST. We decided to analyze the digit 7 looking for subtypes. To perform this experiment, we trained on 20,000 MNIST digits and considered the digit 7 as the positive class. We use a one-vs-rest strategy for classification with flattened images as inputs.

Visually, digit 7 examples have two different subtypes: with or without the middle-cross bar. In order to quantitatively evaluate our method, we labeled 400 test images in two classes, 7 with a middle-cross bar, and those with none. We ran UCSL with GMM as a clustering method, logistic regression as classification method and compared with clustering methods coupled with deep learning models or dimension reduction algorithms. We use the metrics V-Measure, Adjusted Rand Index (ARI) and balanced accuracy (B-ACC), since we know the expected clustering result.

As noticeable on Table 3.1, UCSL outperforms other clustering and subtypes ML methods. We also compared our algorithm with DL methods, a pre-trained convolutional network and a simple convolutional encoder-decoder. Only the convolutional autoencoder network along with a GMM on its latent space of dimension 32 slightly outperforms UCSL. However, it uses a definitely higher number of parameters (7500 times more!) and takes twice the time for training. Our model is thus more relevant to smaller data sets, which are common in medical applications. Please note that UCSL could also be adapted in order to use convolutional autoencoders or contrastive methods such as in [169] and [40], when dealing with large data sets. This is left as future work.

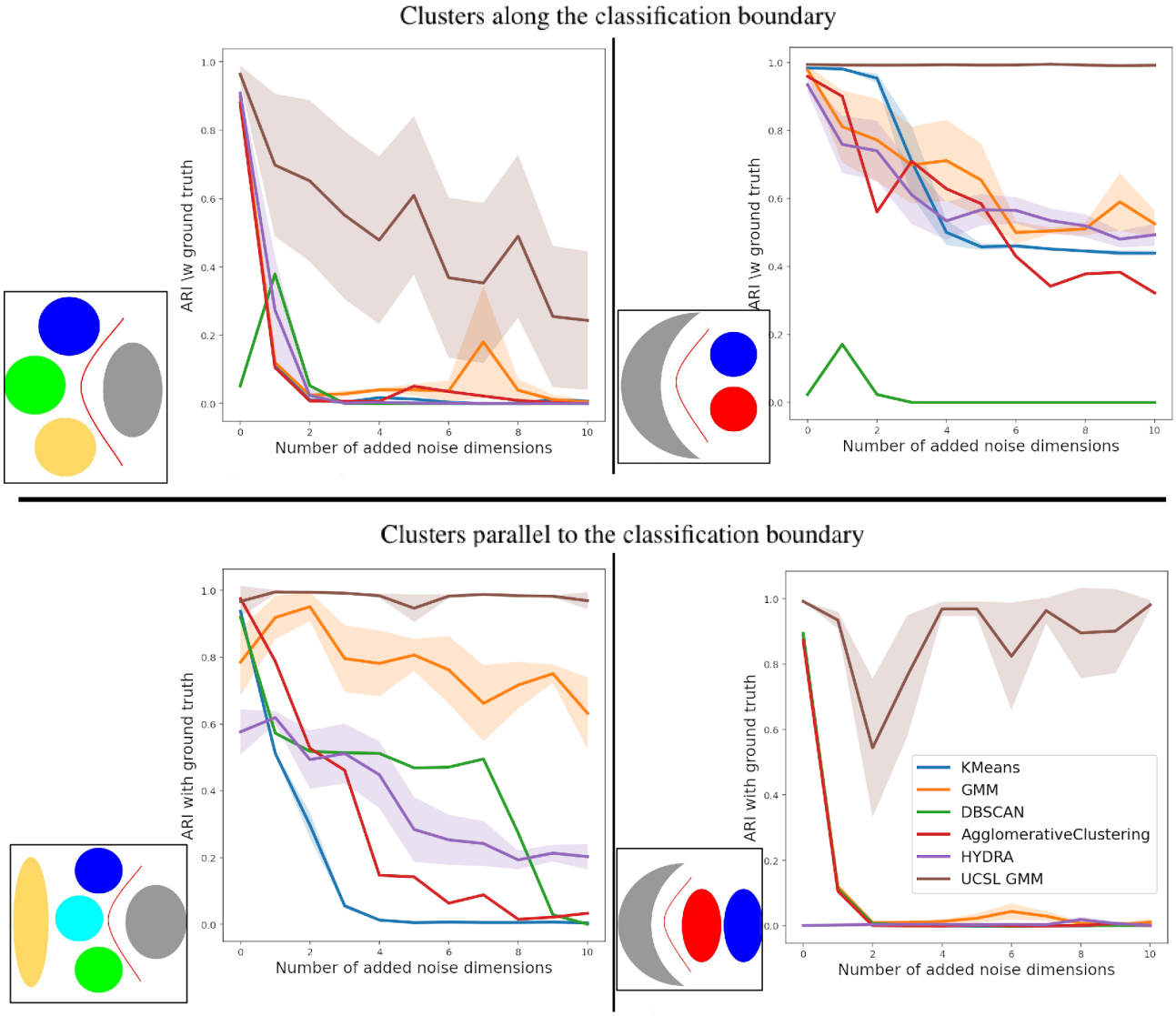


Figure 3.5: Comparison of performances of different algorithms on the four configurations presented in Fig.3.2. Noisy features are added to the original 2D data. For each example, all algorithms are run 10 times with different initialization.

Psychiatric dataset: The ultimate goal of the development of subtype discovery methods is to identify homogeneous subgroups of patients that are associated with different disorder mechanisms and lead to patient-specific treatments. With brain imaging data, the variability specific to the disorder is mixed up or hidden to non-specific variability. Classical clustering algorithms produce clusters that correspond to subgroups of the general population: old participants with brain atrophy versus young participants without atrophy, for instance.

Methods	Latent Size	Nb params	Avg Exec Time	V-measure	ARI	B-ACC
AE + GMM	32	3M	21m40s	0.323±0.013	0.217±0.025	0.823±0.009
UCSL (our)	2	406	12m31s	0.239±0.001	0.330±0.001	**0.808±0.001**
PT VGG11 + KM	1000	143M	32m44s	0.036±0.001	0.087±0.001	0.616±0.001
AE + GMM	2	3M	13m34s	0.031±0.015	0.033±0.021	0.607±0.025
t-sne* [190] + KM	2	4	2m04s	0.029±0.020	0.049±0.056	0.568±0.033
t-sne* [190] + GMM	2	14	2m04s	0.023±0.021	0.020±0.048	0.566±0.033
umap* [206] + KM	2	4*	24s	0.050±0.015	0.078±0.015	0.555±0.022
umap* [206] + GMM	2	14*	24s	0.025±0.006	0.080±0.010	0.547±0.005
SHAP [182]* + KM	196	392*	1h02	0.012±0.007	-0.014±0.035	0.540±0.016
KM	196	392	0.32ms	0.006±0.000	0.010±0.000	0.552±0.000
HYDRA	196	394	9m45s	0.005+/-0.006	0.024±0.031	0.520±0.018
GMM	196	77K	0.32ms	0.0002±0.000	-0.001±0.000	0.510±0.000

Table 3.1: MNIST dataset, comparison of performances of different algorithms for the discovery of digit 7 subgroups. AE : convolutional AutoEncoder; PT VGG11: VGG11 model pre-trained on imagenet; GMM: Gaussian Mixture Model; KM: K-Means. Latent size: dimension of space where clustering is computed. * : to limit confusion, we assign no parameters for t-sne, umap and SHAP. We use default values (15,30,100) for perplexity, neighbours and n estimators in t-sne, umap and SHAP respectively.

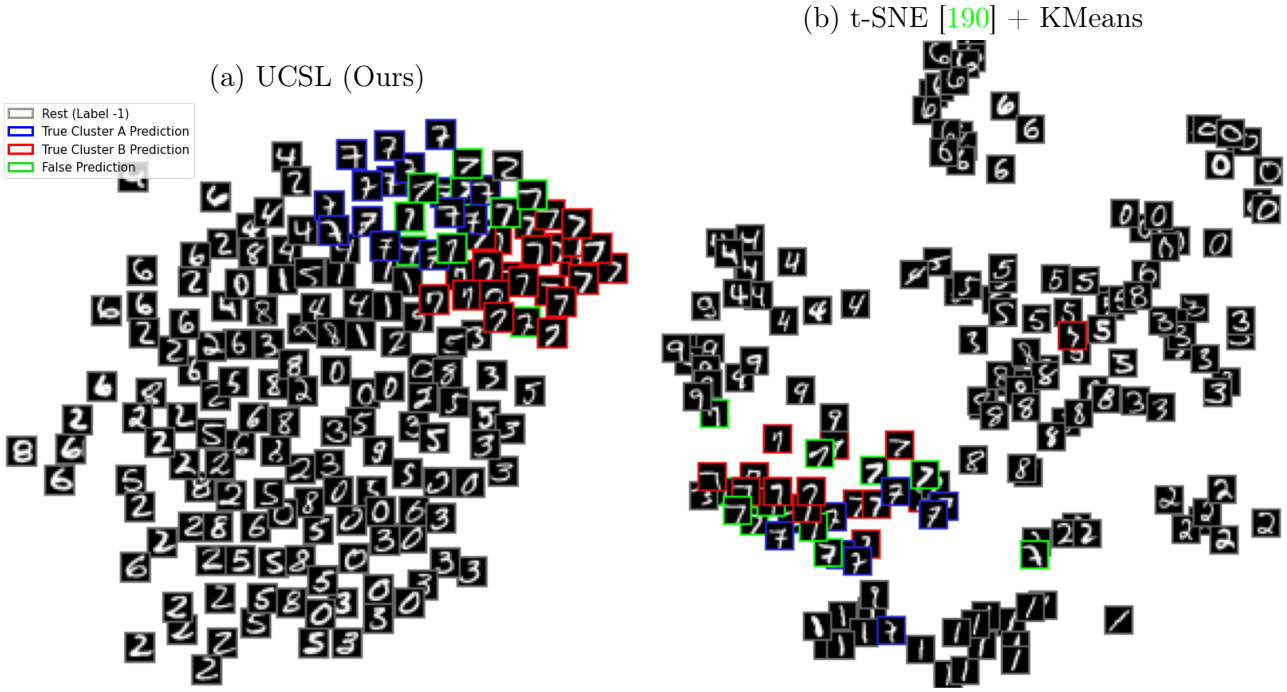


Figure 3.6: Comparison of latent space visualization in the context of MNIST digit “7" subtype discovery. Differently from t-SNE, our method does not focus on the general digits variability but only on the variability of the “7". For this reason, subtypes of “7" are better highlighted with our method.

To validate the proposed method we pooled neuroimaging data from patients with two psychiatric disorders, (Bipolar Disorder (BD) and Schizophrenia (SZ)), with data from healthy controls (HC). The supervised upstream task aims at classifying HC from patients (of both disorders) using neuroimaging features related to the local volumes of brain grey matter measured in 142 regions of interest (identified using CAT12 software [98, 99] from the SPM toolbox). Here, we used a linear SVM for classification. The clustering task is expected to retrieve the known clinical disorder (BD or SZ). Training set was composed of 686 HC and 275 SZ, 307 BP patients.

We measured the correspondence (Tab. 3.2) between the clusters found by the unsupervised methods with the true clinical labels on an independant TEST set (199 HC, 190 SZ, 116 BP) coming from a different acquisition site. As before, we used the metrics V-Measure, Adjusted Rand Index (ARI) and balanced accuracy (B-ACC). Please note that the classification of SZ vs BD is a very difficult problem due to the continuum between BP and SZ. Therefore, performances should be compared with the best expected result provided by a purely supervised model (here a SVM) that produces only 61% of accuracy (last row of Tab. 3.2).

Algorithm	V-measure	ARI	B-ACC
GMM	0.002±0.001	0.003±0.008	0.491±0.024
KMeans	0.008±0.001	-0.01±0.001	0.499±0.029
umap* [206] + GMM	0.001±0.002	0.000±0.007	0.497±0.013
umap* [206] + KM	0.000±0.002	0.001±0.005	0.502±0.006
t-sne* [190] + GMM	0.002±0.0024	-0.00±0.005	0.498±0.028
t-sne* [190] + KM	0.004±0.004	0.003±0.005	0.505±0.041
HYDRA [280]	0.018±0.009	-0.01±0.004	0.556±0.019
SHAP [249] + GMM	0.004±0.005	0.000±0.006	0.527±0.027
SHAP [249] + KMeans	0.016±0.005	0.017±0.012	0.575±0.011
UCSL + GMM	**0.024±0.006**	**0.042±0.016**	**0.587±0.009**
UCSL + KMeans	**0.030±0.012**	0.004±0.006	**0.594±0.015**
Supervised SVM	*0.041±0.007*	*0.030±0.008*	*0.617±0.010*

Table 3.2: Results of the different algorithms on the subtype discovery task BP / SZ. The last row provides the best-expected result obtained with a supervised SVM.

As expected, mere clustering methods (KMeans, GMM) provide clustering at the chance level. Detailed inspection showed that they retrieved old patients with brain atrophy vs younger patients without atrophy. Only clustering driven by supervised upstream tasks (HYDRA, SHAP+KMeans, and all UCSL) can disentangle the variability related to the disorders to provide results that are significantly better than chance (59% of B-ACC). Models based on

USCL significantly outperformed all other models approaching the best expected result that would provide a purely supervised model.

3.1.6 Conclusion

We proposed in this article a Machine Learning (ML) Subtype Discovery (SD) method that aims at finding relevant homogeneous subgroups with significant statistical differences in a given class or cohort. To address this problem, we introduce a general Subtype Discovery (SD) Expectation-Maximization (EM) ensembled framework. We call it UCSL: Unsupervised Clustering driven by Supervised Learning. Within the proposed framework, we also propose a dimension reduction method based on discriminative directions to project the input data onto an upstream-task relevant linear subspace. UCSL is adaptable to both classification and regression tasks and can be used with any clustering method. Finally, we validated our method on synthetic toy examples, MNIST, and a neuro-psychiatric data set on which we outperformed previous state-of-the-art methods by about +1.9 points in terms of balanced accuracy.

3.2 Deep UCSL: Automatic Discovery of Disease Subgroups by Contrasting with Healthy Controls

Context: In several medical domains, researchers are interested in discovering interpretable and homogeneous subgroups within a pathology. Identifying pathology subtypes would enable refining its nosology and push forward the understanding of its underlying biological process and the research for more personalized treatments. However, conventional clustering algorithms fail in this task when the pathological variability (*i.e.*, the set of diseased biomarkers) is dominated by the variability in the general population (*i.e.*, in healthy samples). In this case, the clusters identified generally reflect the general variability (*e.g.*, old vs. young people, or along any other general physiological variability) and are thus irrelevant when characterizing more precise phenotypes of pathologies. To address this problem, we have previously proposed a statistical framework (UCSL) [179], implemented with linear machine learning algorithms, that clusters patients while contrasting with a dataset of control subjects to ensure that the resulting clustering is specific to the pathology of interest.

Motivation: However, in practice, UCSL is a limited Subgroup Discovery method because it generally requires a user-defined feature extraction step and is constrained to linear predictors only. To overcome this shortcoming, we propose Deep UCSL: a general Subgroup Discovery method that does not depend on user-defined features generalizes to several medical imaging domains, and leverages the representation quality of non-linear deep learning methods.

3.2.1 Abstract

In Subgroup Discovery, practitioners are interested in discovering interpretable and homogeneous subgroups within a group of patients. In this paper, assuming that healthy subjects (i.e., controls) share common but irrelevant factors of variation with the patients, we motivate and develop a Contrastive Subgroup Discovery method, entitled Deep UCSL. By contrasting controls with patients, we identify subgroups that stem only from the pathological variability specific to the disorder, while disregarding the common variability shared with the controls. To this end, we propose a Deep Learning framework that learns a discriminative and expressive representation space through a deep features extractor. From a mathematical standpoint, we derive a novel loss from the conditional joint likelihood between latent clusters (disorder subgroups) and patient/control labels. The optimization procedure is based on an Expectation-Maximization strategy alternating between a) subgroups inference and b) features encoder parameters update. Furthermore, we introduce a novel regularization term that promotes the representation space to emphasize disorder-specific variability while disregarding the common variability shared with the controls. Compared to previous related works, our approach quantitatively improves the quality of the estimated subgroups, as demonstrated on an MNIST-based toy example and four distinct real medical imaging datasets. Code and datasets are available at https://github.com/neurospin-projects/2023_rlouiset_deep_ucsl.

3.2.2 Introduction

In the past decades, unsupervised and self-supervised learning techniques have proven to be particularly effective at identifying relevant patterns and factors of variation within a dataset. Combined with powerful Neural Networks (NNs), these methods can produce semantically rich representations [50, 52, 118, 326]. Notably, unsupervised Deep Clustering (DC) methods [16, 40, 41, 169, 278] seek to produce a suitable representation space for identifying homogeneous latent clusters based on the *general variability* of the entire dataset (i.e., imaging patterns common to all samples).

With a different perspective, Subgroup Discovery (SD) in medical applications [20, 154, 311] aims at identifying relevant latent subtypes/subgroups that arise from the *pathological variability* of the diseased population and not from the irrelevant common variability that may exist in both healthy subjects (i.e., controls) and diseased patients. For instance, in dermatology, practitioners are interested in identifying a stratification within a dataset of malignant melanomas. In this case, the objective would be to retrieve relevant dermatological subgroups,

from a medical standpoint (*e.g.,* nodular/lentigo melanoma), for a more targeted drug or treatment delivery. In this paper, we argue that the discovery of these subgroups should be based on disorder-specific patterns (e.g.: texture, color, asymmetry) rather than irrelevant patterns shared with healthy skin samples (*e.g.*: skin color, hair, moles, or nevi).

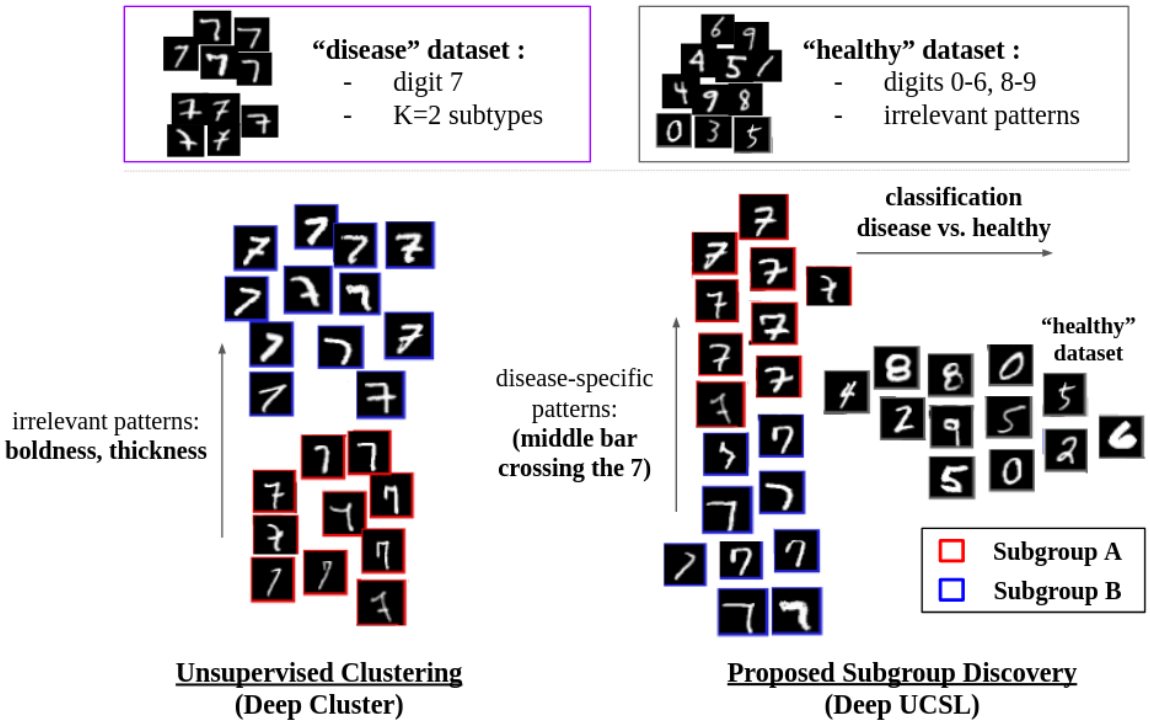


Figure 3.7: Comparison between a Deep Clustering method (Deep Cluster [40]) and the proposed Subgroup Discovery method (Deep UCSL) on a subtype discovery task within the digit 7. The relevant subtypes stem from the digit-7's specific variability, *i.e.,* 7 with a middle-cross bar (red) and without (blue). We show two 2D PCA plots of the representation spaces learned by the two methods. Deep Cluster is driven by the general variability of the digits, particularly the boldness. Differently, Deep UCSL uses a supervised classification task to contrast with the "healthy" class (*i.e.,* the rest of the digits). In this way, it encourages the "disorder" subtype identification to discard the shared factors of variability, focusing only on the relevant variability of the 7 (*i.e.,* the middle crossbar).

In Fig. 3.7, we use an intuitive toy example based on the MNIST dataset to better clarify the differences between Deep Clustering and Subtype Discovery. We consider the digit "7" as the pathological group and all the other digits as the healthy group. Results show how Deep Cluster's subgroups [40] of the digit "7" are only defined by the most predominant characteristics (i.e.: boldness of the digit) common to all digits. Instead, the proposed Subtype Discovery method, called Deep UCSL, disregards these common characteristics and uses only the specific

patterns of the digit "7" (i.e., the presence of the crossing middle bar) to define the subgroups. In many medical domains, the discovery of new subgroups is left to human experts. However, this task may be difficult and observer-dependent, with potentially high-dimensional images and subtle imaging patterns. Furthermore, there are often no precise guidelines to define the subgroups, and there is also considerable disagreement among pathologists, with high intra- and inter-rater variability [311]. This motivates the need for new machine-learning methods to help human experts discover or validate subgroups, while relying on reproducible, data-driven, and objective imaging patterns.

3.2.3 Related works

Subgroup Discovery applications. In the last decade, Subgroup Discovery works have been proposed across various topics. In medicine, many works have been conducted to refine the characterization of cancer variants [43, 197, 225, 307]. In psychiatry, mental diseases are known as being extremely heterogeneous [194], and multiple works have tried to refine the disorder categorization into homogeneous subgroups [45, 90, 127, 312, 300]. However, most of these subgroup discovery methods either rely on priors dependent on the domain (*e.g.,* the aspect of pathological patterns of neurodegenerative disorders), or are based on user-defined features and linear predictors. Notably, emerging machine learning methods, such as HYDRA [280] and our preliminary work UCSL [179], have proposed a framework for Subgroup Discovery, where the supervised classification (healthy vs. patients) and the subgroup identification (clustering of patients only) depends on each other. This framework respects two properties: 1) subgroups should be identified from *diseased* samples *only*, and 2) pathological subgroups should be correctly discriminated from the healthy class in the estimated representation space. However, these works are limited because they require a user-defined feature extraction step and are constrained to linear predictors only. Indeed, each inferred pathological subgroup is being discriminated from the healthy class with a linear classifier. To sum up, actual subgroup discovery works are either not generalizable to other domains or depend on user-defined feature choices and generally do not leverage the representation quality of non-linear deep learning methods. In Deep Representation Learning, two main lines of works have been used to identify clusters or group structure: unsupervised Deep Clustering and Self-Supervised learning, and in particular, Contrastive Learning.

Unsupervised Deep Clustering Recent works [16, 40, 41, 184, 319], proposed to uncover semantically relevant groups of samples in a dataset without using other auxiliary tasks. Notably, Deep Cluster [40] alternates between the pseudo-label estimation phase (i.e.: the clustering es-

timation step) and the update of encoder parameters. At each epoch, previous clustering assignments produced by a K-Means method are used as pseudo-labels for minimizing a Cross-Entropy loss with the network output logits. This strategy is simple, but it requires a few tricks in order to converge. The first problem is the so-called "clustering degeneracy". It happens when the estimated clusters are slightly imbalanced, and therefore the network training will naturally orientate samples towards the most represented clusters. After several iterations, this phenomenon may lead toward empty clusters or, in the worst case, to a trivial solution with only one not-empty cluster. Several solutions have been proposed in the literature. Deep Cluster [40] and ODC [319] proposed to weigh samples' importance with the inverse of the associated cluster size. Another work, SwAV [41], regularizes the prototype centroids so that they respect the equipartition constraint (i.e.: samples get equally distributed across the clusters). Another common issue in Deep Clustering methods is the cluster shuffle across epochs. It occurs when the optimization strategy alternates between a) the estimation of pseudo-labels by fitting a clustering method (*e.g.* K-Means), and b) a gradient descent step of a classification head trained to predict the clustering pseudo-labels. In this setting, step a) is re-computed at each epoch and therefore the cluster's indices might not be consistent across epochs (i.e., sample x has pseudo-label a at epoch t and pseudo-label b at epoch $t+1$). Due to this inconsistency across epochs, step b)'s classifier weights are outdated at each new epoch. As a solution, [40] has proposed to re-initialize the classification head's parameters right after step a). However, several works, such as [319], and [16], argue that this re-initialization disrupts the network training and that cluster centroids should be updated steadily along with the classifier's parameters, which is not trivial.

Self-Supervised Learning. Another relevant line of work comprises self-supervised, and particularly contrastive, learning methods [50, 52, 109, 118, 326], that learn representations potentially suitable for downstream tasks, such as clustering or classification. A pivotal method of this literature is SimCLR [50], which encourages the encoder to be invariant to user-defined image augmentation (*i.e.,* it discards data-augmentation variability). Specifically, it encourages two views of an image to be aligned in the representation space while constraining the dataset representations to follow a uniform hyperspherical distribution, as explained in [293]. Contrastive learning methods only align representations of the same image, which implies that several images with the same semantic content can be pulled apart. This flaw is known as the Class Collision problem [326] and may be harmful to both clustering and classification downstream tasks. To solve this problem, two lines of works have emerged. In the first one, annotated labels are leveraged to explicitly capture healthy/pathological discriminative pat-

terns. For instance, in SupCon [153], authors use a loss where positive and negative samples are defined based on their class (healthy or patient).
The second direction comprises methods that include a contrastive loss within a Deep Clustering or a Deep Nearest Neighbour framework [169, 41, 275, 278]. For example, PCL [169] proposes to use cluster assignments to align positive views from the same cluster rather than from the same image. Other works, such as [60, 278], propose to align nearest neighbors instead. These works either leverage the disorder/healthy label information (first group) or encourage clustering structures (second group). However, a Subgroup Discovery method should combine both properties to effectively define homogeneous groups *only* among patients and *only* based on the pathological patterns/variability of the disorder.

3.2.4 Contributions

To overcome these shortcomings, this paper proposes a Deep Unsupervised Clustering driven by Supervised Learning (Deep UCSL), where we use a Deep Neural Network as an automatic feature extractor to produce representations suitable for identifying relevant disorder subgroups that are contrasted from the healthy class. As in UCSL [179], we derive our objective by seeking the parameters that maximize the conditional joint likelihood between latent subgroups and supervised labels. However, as a difference, the feature space is estimated with a trainable deep encoder that allows an automatic non-linear feature extraction. Then, as in UCSL [179], we propose the use of an Expectation-Maximization optimization process to alternate between the conditional joint likelihood parameters update and the subgroups pseudo-labels estimation. However, here we propose a new regularization loss that improves the separation between healthy and diseased patterns and, differently from UCSL, it guarantees the monotonical convergence of the model parameters. In our approach, the optimization procedure consists of alternatively 1) estimating the subgroups pseudo-labels (*i.e.*: pathological subgroups within the disorder classes) and 2) updating the features encoder. Our objective is to: a) correctly discover the pathological subgroups, b) encourage healthy samples not to belong to a pathological subgroup, and c) accurately discriminate each subgroup from the healthy class (Mixture-of-Classifying Experts).
To demonstrate the superiority of our method, we compare it with other state-of-the-art representation learning methods (*i.e.*, Deep Clustering and Contrastive learning methods) on three different tasks: 1) 7-digit subgroup identification, 2) psychiatric application, 3) pneumonia subgroups identification, and 4) eye pathological subgroups identification. Deep UCSL outperforms all other methods in these tasks. In a nutshell, our contribution is three-fold:

1. To the best of our knowledge, we propose the first Deep Learning method for Subgroup Discovery that performs disorder/healthy classification while identifying subgroups in the diseased class.

2. We motivate and design a clustering regularization loss that forces the learned representation to disregard the healthy population variability focusing only on the disorder-specific variability.

3. A fair and careful evaluation of our method, as well as a comparison with recent state-of-the-art methods.

3.2.5 Methodology

Mathematical formulation

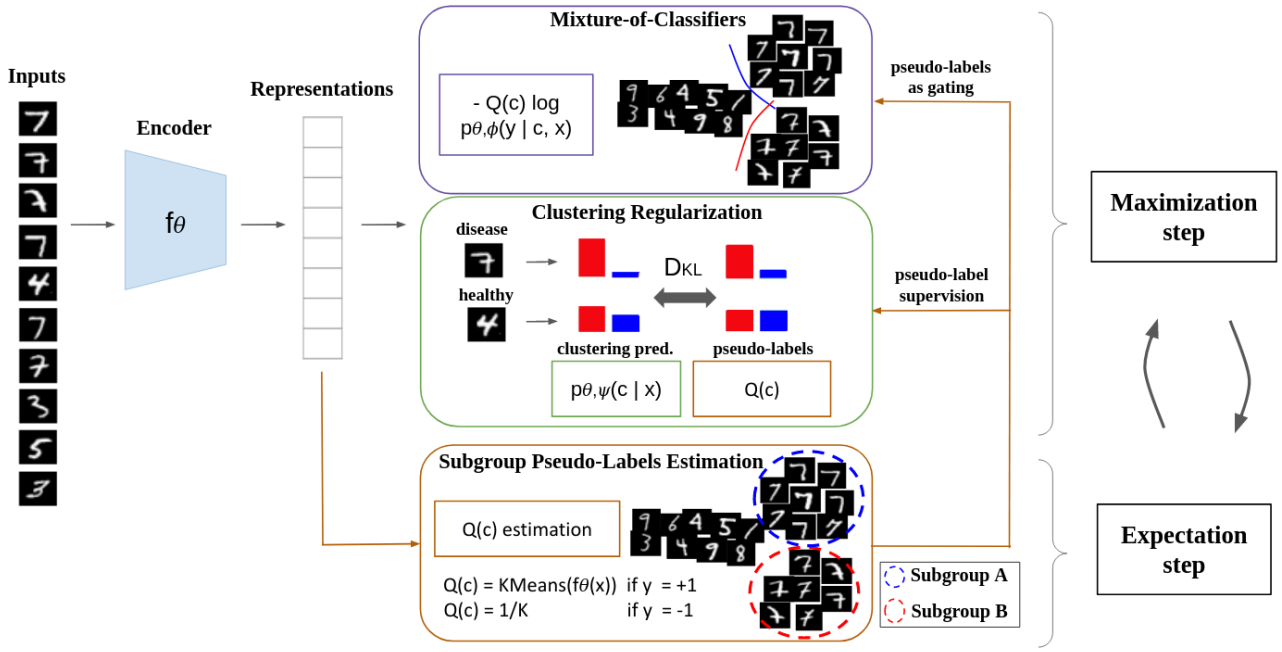


Figure 3.8: A schematic diagram of Deep UCSL with $K = 2$ subgroups (red and blue). At each epoch, K-Means produces subgroup pseudo-labels during the Expectation step (in brown). These pseudo-labels are then used to weight a classification Mixture-of-Experts (in purple) between the "healthy" class (digits 0-6, 8-9) and the "disorder" class (digit 7). Additionally, the pseudo-labels are also used for the clustering regularization (in green), where uniform pseudo-labels (*i.e.*: $\frac{1}{K}$) are used to regularize the healthy class distribution, so that healthy samples are equidistant from all the diseased subgroups. This forces the learned representation to disregard the general variability, common to both healthy and patients.

Let $(X, Y) = \{(x_i, y_i)\}_{i=1}^{N}$ be a labeled dataset composed of N samples. We will restrict to the binary (e.g., patient/control) classification paradigm, $y_i \in \{-1, +1\}$, which is very common in medical imaging. We will denote with N^+ and N^- ($N = N^+ + N^-$) the number of positive and negative samples, respectively. Our objective is to estimate the latent pseudo-labels[1] of subgroups within disorder samples ($y_i = +1$). The membership of each sample i to latent subgroups is modeled via a latent categorical variable $c_i \in C = \{1, ..., K\}$, where K is the number of subgroups. We look for a discriminating model that maximizes the joint conditional likelihood:

$$\sum_{i=1}^{n} \log p(y_i|x_i) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} p(y_i, c_i = k|x_i) \tag{3.9}$$

To attain the three objectives *(a), b), c))* described in Sec. 3.2.4, we need to optimize Eq. 3.9 with respect to both $p(c_i|x_i, y_i)$ and $p(y_i|x_i, c_i)$. Indeed, we need to identify the subgroups only within the patients (thus knowing y) and to accurately discriminate the healthy class from each subgroups (thus knowing c). However, developing the joint conditional likelihood in Eq. 3.9 would result in either $p(c_i|x_i, y_i)$ or $p(y_i|x_i, c_i)$, but not in both. To solve that, as in UCSL [179], we introduce a probability distribution Q over the subgroups C, so that $\sum_{k=1}^{K} Q(c_i = k) = 1 \quad \forall i$, and use the Jensen inequality to obtain a tractable, lower bound of Eq.3.9:

$$\sum_{i=1}^{n} \log \sum_{k=1}^{K} Q(c_i = k) \frac{p(y_i, c_i = k|x_i)}{Q(c_i = k)} \geq \sum_{i=1}^{n} \sum_{k=1}^{K} Q(c_i = k) \log \left(\frac{p(y_i, c_i = k|x_i)}{Q(c_i = k)} \right) \tag{3.10}$$

where equality holds when:

$$Q(c_i = k) = \frac{p(y_i, c_i = k|x_i)}{\sum_{k=1}^{K} p(y_i, c_i = k|x_i)} = \mathbf{p(c_i = k|x_i, y_i)} \tag{3.11}$$

Then, Eq. 3.10 can be rewritten with respect to both $p(y_i|x_i, c_i)$ and $Q(c_i)$ (estimated to approximate $p(c_i|x_i, y_i)$):

$$\underbrace{\sum_{i=1}^{n} \sum_{k=1}^{K} Q(c_i = k) \log \mathbf{p(y_i|x_i, c_i = k)}}_{\text{Mixture-of-Classifying Experts term}} \underbrace{-D_{KL}(Q(c)||p(c|x))}_{\substack{\text{Clustering} \\ \text{Regularization term}}} \tag{3.12}$$

Our goal is to learn a single representation space where both the classifying experts $p(y_i|x_i, c_i =$

[1] we call them pseudo-labels, since we assume that subgroups labels are not known at training

k) and the disorder subgroup $p(c_i = k|x_i, y_i = +1)$ can be accurately estimated. To this end, we propose using a deep encoder f_θ with parameters θ for feature extraction and two neural networks with parameters ϕ and ψ for the classifying experts $p_{\theta,\phi}(y_i|c_i = k, x_i)$ and the unsupervised clustering head $p_{\theta,\psi}(c_i = k|x_i)$, respectively. An overview of the proposed method can be seen in Fig.3.8. To optimize the proposed cost function (Eq. 3.12), we use an EM algorithm that alternatively:

1. estimates Q as $p(c_i = k|x_i, y_i)$ (E-step, Eq. 3.11) at the end of each epoch, freezing the encoder f_θ

2. estimates $p(y_i|x_i, c_i = k)$ and $p(c_i = k|x_i)$ batch-wise by maximizing Eq. 3.12 (M-step) at the beginning of each epoch, freezing Q

Comparison with UCSL

This mathematical framework is similar to the one of our preliminary work UCSL [179], but with significant differences.
First, Deep UCSL uses a deep feature encoder, instead of user-defined features, and two neural-networks for classification and subgroup estimate, instead of linear models.
Second, we do not assume that $p_\theta(c|x) = Q(c)$, as in UCSL, but we force it by explicitly introducing and minimizing the clustering regularization term $KL(Q(c)||p_{\theta,\psi}(c|x_i))$. This guarantees the monotonic convergence of the optimization procedure, which was not the case in UCSL.
Third, since we want to estimate subgroups only within positive/patients, all negative/healthy samples are assigned a uniform probability for all subgroups. This strategy, also not proposed in UCSL, encourages the features encoder f_θ to produce a representation space where negative samples do not belong to (positive) subgroups. These new contributions imply that $p_\theta(c_i|x_i)$, the estimated clustering distribution, is not simply extended to all samples regardless their label y, as in UCSL, but the representation space is estimated so that the unsupervised clustering $p_\theta(c_i|x_i)$ gives the same result as the "supervised" subgroups estimation $p(c_i = k|y_i, x_i)$, namely knowing the label y. As explained in Sec.3.2.5, this entails an encoder f and a representation space where the general variability, common to both positive and negative samples, is discarded and the subtype estimation only depends on the specific variability of the positive class.

Expectation step

During the Expectation step, we freeze the current estimate of the encoder parameters $\theta^{(t)}$, at epoch t, and estimate $Q^{(t)}$ as $p_{\theta^{(t)}}(c_i = k|y_i, x_i), \forall i \in |[1, n]|, \forall k \in |[1, K]|$, see Eq. 3.11. In order to do that, since we assume that only the positive class ($y_i = +1$) contains subgroups, we first compute $p_\theta(c_i = k|x_i, y_i = +1)$ using a regularized K-mean. Please note that any clustering algorithm could be used here to compute $p_\theta(c_i = k|x_i, y_i = +1)$ depending on prior knowledge about the subgroups' number, size, distribution, or density. Here, we assume that the number of subgroups K is known, and thus K-means seems to be a reasonable and simple choice. Concerning samples from the healthy class ($y_i = -1$), as we aim to estimate sub-groups within patients only, we propose to use a uniform clustering probability distribution:

$$p_\theta(c_i = k|y_i = -1, x_i) = \frac{1}{K} \quad \forall i \in |[1, n]|, \forall k \in |[1, K]| \tag{3.13}$$

This means that healthy samples have an equal probability of belonging to the subgroups and, as detailed in Sec. 3.2.5, this will be used to regularize the representation so that samples from the healthy class are equidistant, in the representation space, from the subgroups centroids. By applying this strategy, the general variability, common to both healthy and diseased classes, should be disregarded in the representation space, which should be structured only by the pathological variability of the diseased class. As in Deep Cluster [40], during the subgroup re-estimation, the subgroup membership c_i is re-estimated at each epoch for each sample. This entails two potential issues that have to be dealt with: subgroup degeneracy and subgroup re-identification.

Subgroup degeneracy: In DeepCluster [40], authors have observed that K-Means may yield imbalanced or empty clusters. In order to avoid that, we use a Sinkhorn-Knopp (SK) [59] regularization embedded into soft K-Means. Our method combines thus the clustering expressivity of soft K-Means with the equipartition regularization of the SK algorithm, as in [41]. Furthermore, one can keep these soft pseudo-labels ($0 \leq Q(c) \in \mathcal{R} \leq 1$) or make them hard ($Q(c) \in \{0, 1\}$) using the OneHot encoding function. Similarly to [266], we propose to linearly interpolate $Q(c)$ from soft to hard probabilities along the epochs. In this way, if the initialization is not good, the use of soft pseudo-labels at the beginning of the training avoids fitting unreliable hard pseudo-labels. Instead, toward the end of the training, hard pseudo-labels are preferred to avoid under-fitting. To justify this choice, we conducted an ablation study on MNIST in Tab 3.3. In our algorithm, we first initialize the subgroups centroids $\mu = \{\mu^k\}_{k \in |[1,K]|}$ with the K-Means ++ algorithm. Then, we compute the soft probabilities $Q(c_i = k) = \frac{1/||f_\theta(x_i) - \mu^k||_2^2}{\sum_{a=1}^{K}(1/||f_\theta(x_i) - \mu^a||_2^2)}$

only for positive samples and encourage their equipartition across the subgroups via the SK regularization: $Q' = SK(Q, \epsilon)$, as in [41], where $Q' \in (\mathbf{0}, \mathbf{1})^{N^+ \mathrm{x} K}$. We then interpolate between soft and hard pseudo-labels using: $\hat{Q} = \frac{T-t}{T}\mathrm{OneHot}(Q') + \frac{t}{T}Q'$, where t is the current epoch and T is the total number of epochs, and finally update the centroids. The hyper-parameter ϵ controls the strength of the SK regularization. Typically, $\epsilon = 0$ means no regularization and $\epsilon \gg 0$ brings to subgroups of similar size (i.e.: of a similar number of samples). This is desirable in the medical subgroup discovery context where one often has *a priori* knowledge of the relative size of the subgroups. At inference time, we compute the subgroup probabilities with the same formula as above.

Subgroup re-identification : At epoch $t+1$, since we re-estimate the centroids of each subgroup, we need to identify which was the corresponding centroid at epoch t in order not to disrupt the training process. To guarantee that, we introduce a new permutation operation $\sigma(.)$ from the labels (i.e., $1, .., K$) of the centroids at epoch t to the ones at epoch $t+1$. For instance, $\sigma(1) = 2$ means that the second centroid at epoch $t+1$ corresponds to the first centroid at epoch t. Being a permutation, we would like it to be a bijective mapping in order not to merge two subgroups into one and thus create empty subgroups (e.g., $\sigma(k) = \varnothing$). This issue could occur by naively assigning to each updated centroid (epoch $t+1$) the label of its most similar previous centroid (epoch t). To avoid that, we first compute the similarity matrix (i.e., normalized dot product) of size $K \times K$ between the K previous centroids (epoch t) and the K updated centroids (epoch $t+1$). Then, we apply the optimal transport algorithm using the similarity matrix and the Sinkhorn-Knopp regularization, so that we assign to each cluster at epoch $t+1$ the label of its most similar cluster at epoch t, respecting at the same time the equipartition constraint and thus the bijective mapping. Indeed, since we have K centroids for K labels, the equipartition property is respected if and only if each centroid at epoch $t+1$ is mapped to a single label, which is equivalent to having a bijective mapping between centroids at epoch t and $t+1$. The implementation is straightforward and requires no extra hyperparameter tuning, see code and pseudo-code on GitHub.

Maximization step

Here, we fix the current estimate of $Q^{(t)} = p_{\theta^{(t)}}(c_i|x_i, y_i)$ and maximize Eq. 3.12 to estimate the classifying experts $p_{\theta,\phi}(y_i|x_i, c_i = k)$ and the clustering head $p_{\theta,\psi}(c_i = k|x_i)$.

Mixture-of-Classifying Experts loss: The classifying experts $p_{\theta,\phi}(y_i|x_i, c_i = k)$ are modeled as a single neural network with K outputs, one per subgroup. Let $p_{\theta,\phi_k}(y_i = +1|x_i, c_i = k) = S(h_{\phi_k}(f_\theta(x_i)))$ be the output prediction associated to the subgroup k, where S is a sigmoid

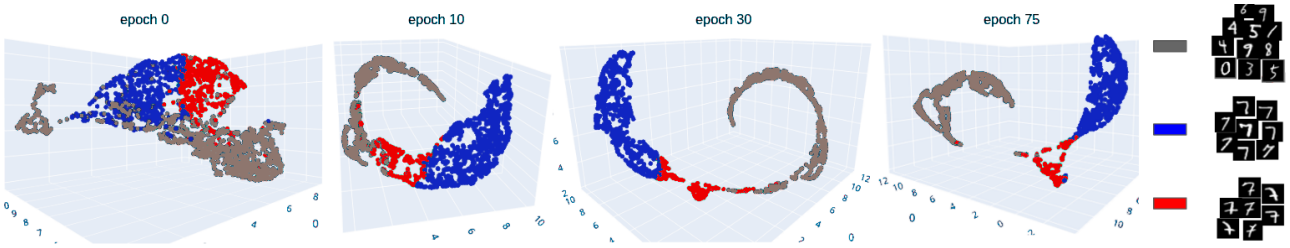


Figure 3.9: 3D UMAP plots of the representation space of Deep UCSL along the epochs on the subgroup identification of the MNIST digit 7. Red and blue points represent the actual (manually labeled) subgroups of the positive class, i.e.: the digit "7" with and without the crossing middle bar. Gray points are samples from the negative class, i.e., other digits than "7". At epoch 1, both classes are fused, there is no clear distinction between subgroups (B-ACC=56%). During training, classes are progressively separated. At the end of the training (B-ACC=97%), the two subgroups (red and blue) are well separated and the negative representations (gray) are equidistant from both subgroups. The ratio of distances between healthy and subgroups centroids ($\frac{d(\text{healthy centroid,subgroup 1 centroid})}{d(\text{healthy centroid,subgroup 0 centroid})}$) gets close to 1 (equidistant case) during the training (epoch 1: 3.68; epoch 75: 1.35).

activation function. Equivalently, we have $p(y_i = -1|x_i, c_i = k) = 1 - S(h_{\phi_k}(f_\theta(x_i)))$. For each subgroup k, we can thus interpret the label y_i as a Bernoulli random variable with $\Pr(y_i = +1) = p_k = S(h_{\phi_k}(f_\theta(x_i)))$ and rewrite the first term of Eq. 3.12 as: $Q(c_i = k) \log\left(\frac{p_k^{y_i}}{(1-p_k)^{y_i}}\right) = Q(c_i = k)\big(y_i \log p_k - y_i \log(1 - p_k)\big)$, which is a weighted binary cross entropy between ground truth labels y_i and output predictions p_k. The maximization of this term is equivalent to the estimate of K sub-classifiers, one per subgroup. This approach is known as a Mixture-of-Experts (MoE) [275], [332] where an expert k is specialized in discriminating the subgroup k samples from the negative class. Even if all the expert's classifiers are trained with the same objective function (i.e., a weighted binary cross-entropy), they are expected to converge towards different solutions since they are differently weighted by $Q(c_i = k)$. Please note that the proposed method is different from the standard MoE routing mechanism where the gating function seeks clusters *across all classes* ($Q(c) = p(c|x)$) [332], whereas our gating function seeks subgroups *only within the positive class* ($Q(c) = p(c|x, y)$). This brings a different mathematical formulation and optimization procedure.

Clustering head loss: The clustering head $p_{\theta,\psi}(c_i|x_i) = \sigma(h_\psi(f_\theta(x_i)))$ is modeled as a neural network with a softmax function σ as output activation function since it predicts cluster probabilities. The parameters θ of the encoder and the parameters ψ of the clustering network are updated through the maximization of the second term of Eq. 3.12 (i.e., clustering regularization term). It represents the Kullback-Leiber divergence between the subgroup pseudo-labels

$Q(c_i)$ (estimated to approximate $p_\theta(c_i|y_i, x_i)$) and the clustering head predictions $p_{\theta,\psi}(c_i|x_i)$. This loss aims at minimizing the discrepancy between the two distributions $p_\theta(C|Y, X)$ and $p_{\theta,\psi}(C|X)$, producing a representation space more suited for subgroup discovery. Indeed, negative samples should be encoded in the representation space as points equidistant from the subgroup centroids since their membership probability should be the same for all subgroups (i.e., $1/K$). Furthermore, positive samples should be clustered as in $Q(c_i)$, namely as if the "unsupervised" clustering algorithm was only considering the pathological/positive variations. This regularization promotes a representation space where the general variability (common to both negative and positive classes) is discarded for the identification of subgroups. Using the MNIST example, we plot a 3D visualization of the representation space in Fig. 3.9. We observe that both bold and thin digits of the negative class (i.e., all digits but the "7") are encouraged to be equidistant from the subgroups of the digit "7" in the representation space across the epochs. Boldness is thus considered as an irrelevant source of variability for subgroup discovery.

Mixture-of-experts: At inference, one can compute the classification label y_j for a new test sample x_j using the Mixture-of-Experts prediction, defined as:
$p(y_j|x_j) = \sum_{k=1}^{K} p_{\theta,\phi_k}(y_j|x_j, c_j = k)p_{\theta,\psi}(c_j = k|x_j)$.

3.2.6 Experiments

This section evaluates Deep UCSL and compares it with several SOTA (State-Of-The-Art) methods on a synthetic dataset (Digit 7 MNIST) (with an Nvidia K80), three medical image applications (with an NVIDIA V100) and a neuro-psychiatric application (with an NVIDIA RTX3800). Results uncertainty (*i.e.,* $\pm$) are obtained with 3 different initialization evaluated on the same independent test set. Only for the neuro-psychiatric case, the variability is instead obtained from 5 different TRAIN/VAL splits (0.9, 0.1) whose models are evaluated on the same external TEST set.

Implementation choices.

For each dataset, we use different and appropriate architectures, with relative hyper-parameters, like the batch size, that performed well in previous works (more details in each section). The only hyper-parameter proper to Deep UCSL is the Sinkhorn-Knopp strength $\epsilon \in \mathcal{R}^+$. We use the same GPU implementation as in [41], where $\epsilon = 0.05$ by default.

Evaluation criteria. We train all methods using only the class label y (healthy vs disorder), but **not** the subgroup labels c. Then, to quantitatively evaluate performance, we use test

Algorithm 3 Deep UCSL pseudo-code

Input: $X \in \mathbf{R}^{N\text{x}(C*W*H)}$, $y \in \{-1, 1\}^N$, K: # subgroups, ϵ: SK temperature, T: # epochs
Output:
    Features encoder: f_θ
    Clustering head: $p_{\theta,\psi}(c_i|x_i)$
    Classifying Experts: $p_{\theta,\phi}(y_i|x_i, c_i = k), \forall k$
    Fitted K-Means: $p_\theta(c_i|x_i, y_i)$
Initialization step: Estimate $Q^{(0)}$:
    Compute probability matrix $Q_+^{(0)} \in (\mathbf{0}, \mathbf{1})^{N^+\text{x}K}$ with
    soft K-Means only on positive samples (i.e., $y = +1$)
    Regularize $Q_+^{(0)}$ with Sinkhorn-Knopp (SK).
    Apply a Soft→Hard linear interpolation:
        $Q_+^{(0)} = \frac{T-t}{T}\text{OneHot}(Q_+^{(0)}) + \frac{t}{T}Q_+^{(0)}$
    Set $Q_-^{(0)} = \frac{1}{K}$ for background samples ($y = -1$).
    Concatenate $Q_-^{(0)}$ and $Q_+^{(0)}$ to get $Q^{(0)} \in (\mathbf{0}, \mathbf{1})^{N\text{x}K}$.
for t in T epochs:
    M step (supervised step):
    Freeze $Q^{(t)}$
    for expert k in K, estimate:
        $p_{\theta,\phi_k}(y_i = +1|x_i, c_i = k) = S(h_{\phi_k}(f_\theta(x_i)))$
    Estimate $p_{\theta,\psi}(c_i|x_i) = \sigma(h_\psi(f_\theta(x_i)))$.
    Compute L_{experts} and $L_{\text{clustering}}$ (Eq. 3.12).
    E step (unsupervised step):
    Freeze $\theta^{(t)}$.
    Estimate $Q^{(t+1)}$ as in initialization step.

sets where we know both the class label y and the subgroup label c. About the representation/contrastive learning methods that do *not* have a classification head (e.g., Deep Cluster, SimCLR), we test their performance only in subgroups identification with a K-means algorithm fitted only on target samples (as if they had a perfect classification head). We use three different metrics: 1) *Class Balanced Accuracy (Class B-ACC):* which is the binary Balanced Accuracy between true labels y_j and class predictions $p(y_j|x_j)$. 2) *Subgroup Balanced Accuracy (Subgroup B-ACC):* Balanced Accuracy between true subgroups c_j and inferred ones $p(c_j|x_j)$.
3) *Overall B-ACC*: takes into account both class and subgroup prediction errors: $\frac{1}{2}\frac{TP}{TP+FN} + \frac{1}{2}\frac{TN}{TN+FP}$, where TN and FN are the class true and false negatives, namely the number of healthy and disorder samples classified as healthy, respectively. TP is the number of disorder samples correctly classified *AND* assigned to the right subgroup. FP is the number of healthy samples classified as disorder *OR* disorder samples correctly classified but assigned to the wrong subgroup.

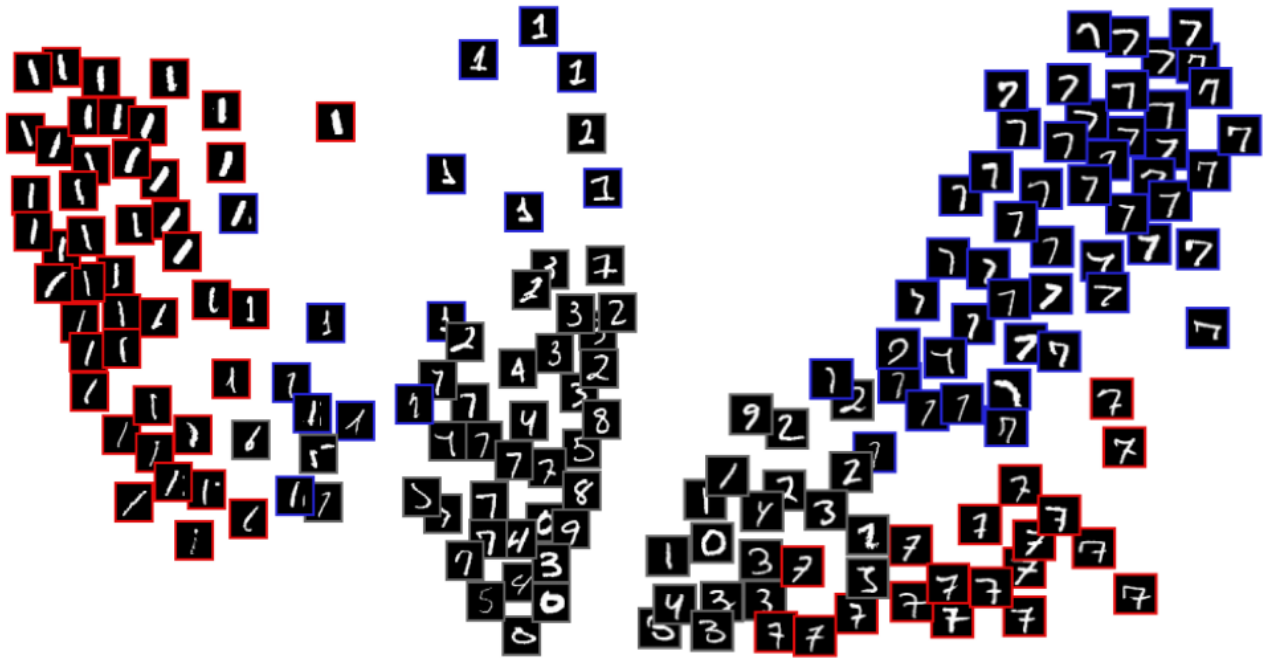


Figure 3.10: MNIST qualitative results. 2D PCA plots of the representation spaces learned by Deep UCSL when seeking two subgroups for digits 1 (left) and 7 (right). PCA is trained on the TRAIN set and then used on the TEST set for visualization purposes.

Digit MNIST dataset

We create two datasets based on MNIST, specifically conceived for subgroup identification. We consider either the digit 7 or 1 as the positive class and all other MNIST digits as the negative class. See Fig.3.10 for visual examples. Digit 7 can be divided into two subgroups: with (red) or without (blue) the middle crossbar. Digit 1 can also be divided into two subgroups: without (red) or with (blue), an additional top diagonal segment. For each dataset, we first train on 12,530 MNIST digits (half positive digit (e.g.: 7 or 1) and half the other digits, equivalently distributed, contrarily to [179]). Importantly, note that even if the subgroups are strongly imbalanced within the positive class (digit 7: 20-80 %, digit 1: 5-95 %), Deep UCSL still manages to correctly identify them. Concerning the 7-digit, we created a test set containing 400 digit-7 samples hand-labeled with the corresponding subgroup (with or without cross middle bar). We present qualitative results in Fig.3.10 and quantitative results in Table 3.3, where Deep UCSL outperforms unsupervised clustering, representation learning, and supervised methods in identifying the two subgroups. Contrastive Learning methods (e.g.: SimCLR, SupCon) produce representations invariant to user-defined image augmentations. The choice of these augmentations is crucial and limited. In practice, given a pathological dataset, a practitioner seeking to discover subgroups may not know all sources of irrelevant variations. Furthermore, even if all irrelevant sources of variations are known, they might not always be easily implemented in practice. Deep UCSL does not need user-defined augmentations to produce representations that are invariant to irrelevant sources of variability. It can automatically disregard the *general* variability, common to both healthy and pathological datasets, and focus only on the *specific* variability of the pathological samples.
In the MNIST experiment, we can notice that boldness is an important general source of variation (see Fig.3.10), common among all MNIST digits and thus irrelevant for subgroup discovery. In order to encourage boldness invariance, we propose to simulate morphological augmentations (erosion and dilation) during the training of contrastive learning methods (*Morpho* SimCLR and *Morpho* SupCon). This design choice improves the representation quality for subgroup identification, demonstrating that a representation space invariant to general, irrelevant sources of variations provides better features for subgroup discovery. However, this design choice is probably not enough, and we can observe that *Morpho* SimCLR and *Morpho* SupCon's performances are inferior to Deep UCSL's. Please note that we also use geometric transformations for every method (Deep UCSL included): a RandomRotation with ± 25 degrees, a RandomAffine with translate parameters of (0.1, 0.1), and a shear parameter set to 0.1.
By contrasting with healthy (negative) samples, the representation space of Deep UCSL cap-

tures more discriminative and fine-grained patterns related to the pathological (positive) variability than its Unsupervised Deep Clustering counterparts. Similarly, SupCon performs better than SimCLR and BCE+K-Means (Classification with clustering on the representations).
We also compare with the best linear method on this experiment: UCSL [179], which directly processes the pixels of the input image, which, by design, makes it sensitive to translation variability. On the contrary, due to its deep convolutional features extractor, (*i.e:* 4 convolutional layers with 7×7 kernel, padding of 3, batch norms between each layer, numbers of channels: 16, 32, 64, 128, average poling layer onto a representation of size 128), Deep UCSL (and the other deep neural network methods) is translation-invariant and can non-linearly process the input pixels. Deep learning methods were trained with an Adam optimizer, a learning rate of 1e-5, trained during 75 epochs, and a batch size of 256.

Table 3.3: MNIST experiment about subgroup discovery of digit 7. Top (unsupervised) methods are trained on 7 digits only. "*Morpho*" methods have morphological data augmentations that simulate digit boldness. For Representation Learning methods, clusters were fitted and inferred using K-Means on the representations of positive samples only. We do not report the Class Accuracy (7 vs rest), as it is always 100%.

Algorithm	Subgroup B-ACC
Deep Cluster-v2 [40], [41]	0.540±0.032
PCL [169]	0.555±0.018
Morpho PCL	0.732±0.027
BYOL [109]	0.552±0.013
SCAN [278]	0.597±0.039
SwAV [41]	0.601±0.009
SimCLR [50]	0.634±0.007
Morpho SimCLR	0.721±0.021
AE + K-Means	0.776±0.007
MoE [332]	0.5±0.0
BCE + K-Means	0.633±0.020
SupCon [153]	0.669±0.066
Morpho SupCon	0.808±0.018
UCSL [179]	0.815±0.009
Deep UCSL with soft pseudo-labels	0.816±0.017
Deep UCSL with hard pseudo-labels	0.895±0.014
Deep UCSL without SK	0.908±0.008
Deep UCSL	**0.920±0.015**

Neuro-psychiatry application.

We create another dataset for subgroup identification comprising 3D MRI T1 weighted images of the brain. The healthy class contains Healthy Controls (HC=686), and the disorder class, Mental Disorders (MD), comprises two subgroups: 1) patients with Schizophrenia (SZ=275), from SCHIZCONNECT [292], and 2) patients with Bipolar Disorder (BD=307), from BIOBD dataset [245]. Voxel-based morphometry (VBM) is performed with CAT12 [98, 99] to preprocess the images. The analysis pipeline includes non-linear spatial registration on a standard template (MNI), Gray Matter (GM), White Matter (WM), and Cerebrospinal Fluid (CSF) tissue segmentation, bias correction and segmentations modulation. VBM images are made isotropic with 1.5mm3 spatial resolution, and the output dimension is $121 \times 145 \times 121$. From there, images are cropped to $121 \times 121 \times 121$ and padded to reach a dimension of $128 \times 128 \times 128$. Voxel values are centered on a unit-gaussian distribution per image (*i.e:* mean of voxels of 0, the standard deviation of voxels of 1). With CAT12, we further compute GM volumes averaged on the Neuromorphometrics atlas that includes 142 brain cortical regions of interest (ROIs). For Deep Learning methods, we use the pre-processed GM-only images as inputs of a 3D-DenseNet deep encoder, as in [73]. For the linear method UCSL, we consider the GM ROIs features.
In Table 3.4, we show the subgroup identification capability of Deep UCSL compared with related works. All evaluation criteria are computed on an independent TEST set (199 HC, 190 SZ, 116 BP), coming from the BSNIP cohort [271], with different acquisition sites. Controls and patients share common (thus irrelevant) sources of variations (e.g.: age, sex, acquisition site). Please note that this is a challenging subgroup discovery problem given the: high dimensionality of 3D brain images, the subtle differences between healthy controls and patients, the few training data, the continuum between both diseases, and the different acquisition machines/protocols. To compare with an upper bound, we train a Deep Neural Network to classify between SZ and BD in a fully supervised manner with a Binary Cross-Entropy (BCE). For all neuro-psychiatric deep methods, including Deep UCSL, we chose a batch size of 8, and the data augmentation strategy is similar to [73, 77]. Deep learning methods were trained with an Adam optimizer, learning rate 1e-5 during 100 epochs.
Interestingly, it seems that the UCSL method's performance highly depends on the feature extraction step. In particular, when using as features the latent vectors of a Variational AutoEncoder (VAE)[150] (with an architecture similar to [180]), the performances decrease. On the other hand, when using highly specific features obtained from more than 20 years of research (GM ROI features with age confound effect correction), performances are among the best. We argue that Deep UCSL provides an end-to-end subgroup discovery method that needs no prior

Table 3.4: Results on Neuro-psychiatry task (BP/SZ) on an independent TEST set. Upper methods are trained on [SZ+BP] only.

Algorithm	Subgroup B-ACC	Class B-ACC	Overall B-ACC
Deep Cluster - v2 [40], [41]	0.517±0.010	×	×
PCL [169]	0.542±0.030	×	×
SwAV [41]	0.522±0.008	×	×
SCAN [278]	0.509±0.008	×	×
SimCLR [50]	0.571±0.017	×	×
BYOL [109]	0.508±0.006	×	×
VAE [150] + UCSL [179]	0.5348±0.016	0.588±0.013	0.459±0.018
BCE + K-Means	0.507±0.005	0.653±0.025	0.428±0.038
SupCon [153]	0.550±0.014	0.656±0.017	0.458±0.017
GM ROI features [99] + UCSL	**0.590±0.016**	0.653±0.012	0.525±0.011
Deep UCSL	**0.589±0.011**	**0.671±0.018**	**0.543±0.014**
CE (upper bound)	0.615±0.007	×	×

knowledge about the feature extraction step and leads to better or similar performances. In particular, in the following two applications, the manual feature extraction step is highly complex and less performing since, compared to neuro-imaging, less research has been conducted. In that case, it is thus highly beneficial to have an end-to-end method with a trainable feature extractor, as Deep UCSL.

Pneumonia subgroup identification.

Here, we propose to address the identification of two subgroups in pediatric medicine: viral pneumonia and bacterial pneumonia. We use the same training and testing datasets as in [146]. For the training set, we choose a balanced subset of 1341 viral samples, 1341 bacterial samples, and 1341 healthy samples. The testing set contains 234 healthy samples, 242 bacterial samples, and 148 viral samples, see Sec.B.5.3 in the Appendix for more details about the dataset. For this application, we use the same architecture and image sizes for every method (except for VAE, where we chose an experimental setup similar to [142]), namely, a ResNet-18 (pre-trained on ImageNet) and 224^2 pixels images. We trained the methods for 50 epochs, a batch size of 256, and an Adam optimizer with a learning rate 1e-5. We present quantitative results in Table 3.5 where Deep UCSL is the best performing method in the subgroup identification task. In Fig. 3.11, we display the nearest images from each subgroup centroid. We observe distinct pathological patterns that are specific to bacterial and viral pneumonia. These results illustrate how practitioners can leverage a Subgroup Discovery method to stratify a pathology.

Table 3.5: Comparison of different methods for the viral/bacterial subgroup identification along with diagnosis classification. Upper methods are trained on disorder samples only.

Algorithm	Class B-ACC	Subgroup B-ACC	Overall B-ACC
DeepCluster-v2 [40], [41]	×	0.814±0.008	×
PCL [169]	×	0.773±0.055	×
SwAV [41]	×	0.815±0.006	×
SCAN [278]	×	0.576±0.059	×
SimCLR [50]	×	0.741±0.027	×
BYOL [109]	×	0.748±0.034	×
VAE [150] + UCSL [179]	0.734±0.020	0.731±0.004	0.646±0.007
BCE + K-Means	**0.917±0.012**	0.560±0.014	0.752±0.007
SupCon [153]	0.895±0.004	0.576±0.036	0.744±0.008
Deep UCSL without SK	0.880±0.019	**0.847±0.037**	**0.812±0.017**
Deep UCSL	0.886±0.010	**0.835±0.007**	**0.820±0.012**
CE (upper bound)	×	0.891±0.005	×

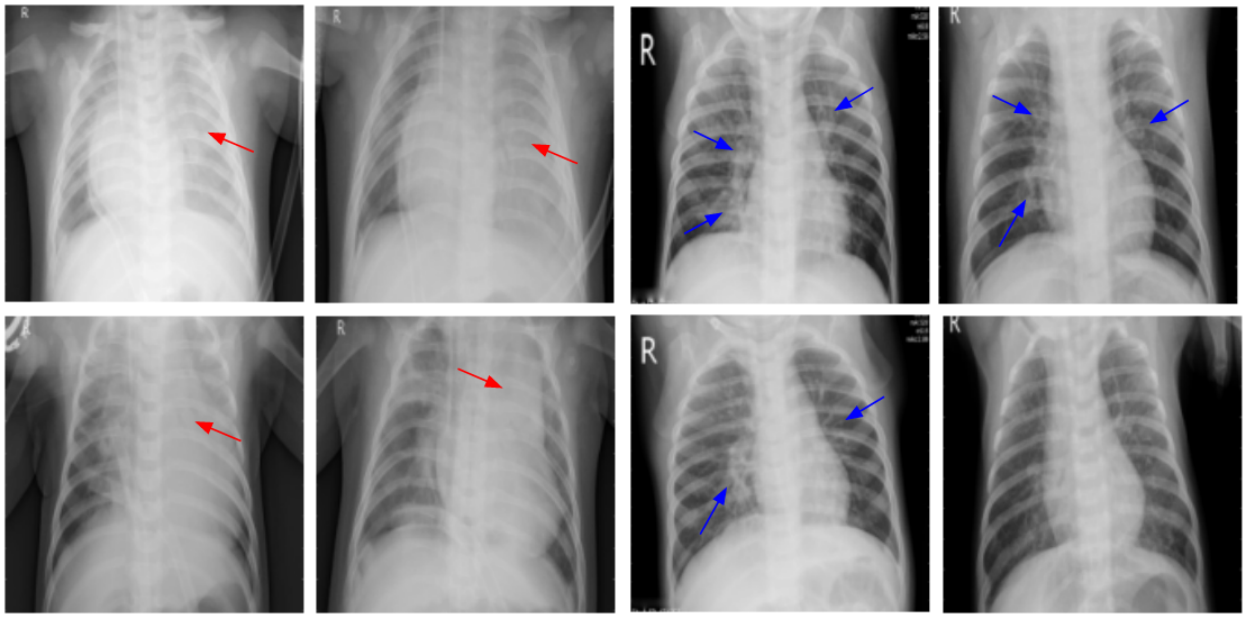


Figure 3.11: A qualitative figure of 4 pneumonia images associated with the most certain subgroup prediction inferred by the clustering head $p_{\theta,\phi}(c|x)$. We can observe pathological patterns used by experts in practice: bacterial pneumonia typically shows a lobar consolidation in one of the lungs (red arrows), whereas viral pneumonia exhibits diffuse interstitial patterns in both lungs [146] (blue arrows).

Retinal pathology applications

We ultimately validate our method on the discovery of disorder subgroups in retinal pathologies. We evaluate our performance on two different data sets. For both experiments, we use the

same setup for every method (except for VAE, where we chose an experimental setup similar to [142]), namely, a ResNet-18 (pre-trained on ImageNet) and 224^2 pixels images. First, we use the Retinal Pathology OCT (Optical Coherence Tomography) dataset introduced in [146]. The train set comprises 3,000 healthy and 3,000 diseased eye images divided homogeneously into 3 disease subgroups: Choroidal Neo-Vascularization (CNV), Diabetic Macular Edema (DME) and Drusens in age-related macular degeneration. The test set comprises 242 healthy samples and 242 samples for each pathological subgroup. For this experiment, quantitative results are shown in Table 3.6. As in the previous section, qualitative results in Fig 3.12 show that the closest TEST images to subgroup centroids show distinct pathological patterns that are specific to each retinal pathology. Again, this satisfactory result illustrates how practitioners can leverage a Subtype Discovery method to stratify and better interpret pathological images. We also use the Ocular Disease Intelligent Recognition (ODIR) dataset[2], which contains 1,890 healthy and 1,391 diseased patients, divided heterogeneously into 5 disease subgroups: Diabetes, Glaucoma, Cataracts, Age-related macular degeneration, and pathological Myopia. The test set contains 210 healthy and 155 pathological samples,divided heterogeneously across disease subgroups. Table 3.7 displays quantitative results. Deep UCSL performs better than all other SOTA methods on the subgroup (and Overall) identification task. For both experiments, all methods were trained with an Adam optimizer, learning rate 1e-5, 50 epochs, batch size of 256.

Table 3.6: Experiments on retinal OCT dataset [146] (3 subgroups).
Upper methods are trained on patients only.

Algorithm	Class B-ACC	Subgroup B-ACC	Overall B-ACC
DeepCluster-v2 [40], [41]	×	0.593±0.052	×
PCL [169]	×	0.428±0.022	×
SwAV [41]	×	0.563±0.073	×
SCAN [278]	×	0.449±0.045	×
SimCLR [50]	×	0.578±0.009	×
BYOL [109]	×	0.337±0.002	×
VAE [150] + UCSL [179]	0.350±0.0017	0.585±0.026	0.369±0.016
BCE + K-Means	0.996±0.003	0.371±0.009	0.672±0.001
SupCon [153] + K-Means	0.999±0.001	0.618±0.002	0.732±0.001
Deep UCSL without SK	0.999±0.001	0.345±0.003	0.668±0.001
Deep UCSL	**1.0±0.0**	**0.626±0.010**	**0.735±0.003**
CE (upper bound)	×	0.971±0.004	×

[2] https://odir2019.grand-challenge.org/

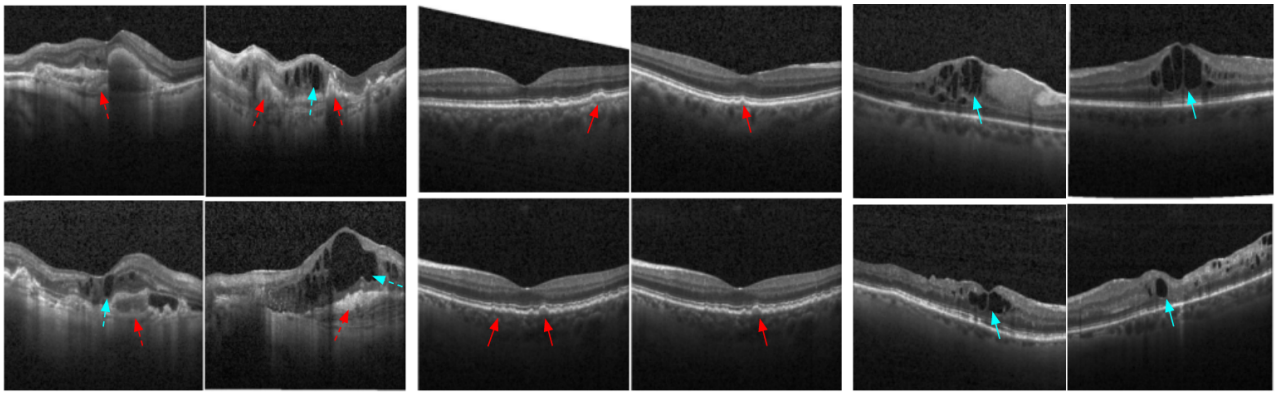


Figure 3.12: A qualitative figure of the retinal pathology four images associated with the most certain predictions inferred by the clustering head $p_{\theta,\phi}(c|x)$. The most certain predictions of cluster A (left) show choroidal neovascularization (CNV) (discontinued red arrows show neovascularization evidence, discontinued blue arrows show subretinal fluid pouches). (middle) Cluster B's most certain predictions show drusens present in early age-related macular dementia (AMD) (red arrows show extracellular deposits). (right) Cluster C's most certain predictions show diabetic macular edemas (DME) (blue arrows show intraretinal fluid).

Table 3.7: Experiments on ODIR dataset [1] (5 subgroups).
Upper methods are trained on patients only.

Algorithm	Class B-ACC	Subgroup B-ACC	Overall B-ACC
DeepCluster-v2 [40], [41]	×	0.533±0.023	×
PCL [169]	×	0.308±0.009	×
SwAV [50]	×	0.415±0.026	×
SCAN [278]	×	0.467±0.043	×
SimCLR [50]	×	0.452±0.033	×
BYOL [109]	×	0.287±0.026	×
VAE [150] + UCSL [179]	0.532±0.030	0.309±0.047	0.469±0.023
BCE + K-Means	0.728±0.006	0.424±0.038	0.548±0.027
SupCon [153] + K-Means	0.716±0.002	0.524±0.021	0.575±0.012
Deep UCSL without SK	**0.736±0.001**	0.371±0.061	0.522±0.024
Deep UCSL	**0.732±0.003**	**0.560±0.020**	**0.619±0.006**
CE (upper bound)	×	0.737±0.003	×

3.2.7 Discussion and Conclusion

In this work, we proposed, to the best of our knowledge, the first deep-learning method for Disease Subgroup Discovery that contrasts with healthy controls. Our work is motivated by the failure of linear methods, such as UCSL and HYDRA, when latent subgroups stem from

non-linear patterns in input images, which are difficult to capture with manual feature engineering. To do so, we took inspiration from UCSL to derive our objective and motivate the use of a deep encoder network as a feature extractor. As in UCSL, we use an Expectation-Maximization optimization process to alternate between the subgroup pseudo-labels estimation and the classification of each subgroup from the healthy class. Differently from UCSL, we motivate the need for a clustering regularization to update the encoder's representation so that we can correctly discriminate the pathological subgroups and encourages healthy samples not to belong to a pathological subgroup. Furthermore, this regularization guarantees the monotonical convergence of the optimization procedure.

Importantly, the addition of deep neural networks, compared to UCSL, was not trivial. Notably, the pseudo-label supervision raises two problems already described in the Deep Clustering literature: clustering degeneration and clustering re-identification (across epochs). Inspired by the Optimal Transport algorithm Sinkhorn-Knopp, we successfully designed two distinct strategies to tackle these issues. Concerning clustering re-identification, we develop a manner to ensure a bijective mapping between K former centroids and K updated centroids. As for clustering degeneracy, we took inspiration from existing methods to develop a Soft K-Means algorithm regularized to respect (up to a certain threshold ϵ) the equipartition of training samples across the subgroups. Intriguingly, we show in the experiments that our method remains robust to highly imbalanced Subgroup Discovery cases.

In medical imaging research, subgroups are usually not known in advance, and practitioners lack unsupervised proxy measures to assess the relevance of the inferred subgroups. In order to evaluate our method, we conceived datasets where the subgroup evaluation can be quantitatively measured. We thus demonstrated, through quantitative and qualitative evaluations, that our method could help practitioners identify subgroups within pathological samples in real-world scenarios. Still, in our method, the number of subgroups K needs to be chosen by the practitioners before the model training, even though it could be unknown *a priori*. The choice of this hyper-parameter might thus be difficult and it points out the need for an unsupervised proxy measure to evaluate the quality of inferred subgroups. Another interesting perspective would be the extension of the proposed model to regression and multi-classification applications, which may be also relevant to medical imaging.

3.3 The discovery of two schizophrenia biotypes with UCSL

3.3.1 Abstract

Background: Psychiatric diseases are still poorly understood and researchers have had trouble finding consistent biomarkers that may explain such pathologies. One key to overcoming this situation would be to successfully parse the neurobiological heterogeneity in the population suffering from psychiatric disorders.

Study design: In this article, a cohort with schizophrenia disorder is analyzed with the help of a novel subtype discovery Machine Learning method. The cohort encompasses samples coming from multiple sites of acquisition, either healthy or suffering from schizophrenia. The Subtype discovery method is called UCSL and was specifically designed to find subtypes that stem from disorder-specific factors of variability rather than factors in common with the healthy population. Anatomical variations captured thanks to structural T1w MRIs were studied. These variations were summed up in local volumetric measures thanks to the CAT12 MRIs processing pipeline. Two schizophrenia subtypes were found, each representing 87% and 13% of the schizophrenia population, and were equally distributed in terms of age distribution and sex repartition.

Study results: Statistical analyses with cognitive and clinical scales were performed concerning subgroups' associated components. Component A shows an overall cognitive decline on 8 out of 9 cognitive scales, and component B revealed a cognitive decline on 2 out of 9 cognitive scales. In terms of clinical psychiatric scales that were also analyzed, component A correlated with 5 different clinical scales, indicating a global disorder severity along that dimension. Component B correlated with 4 clinical scales and suggested a mildly less severe global disorder severity. Regarding gray matter atrophies, more global overall atrophy in around 90 regions of interest along component A was found and a set of 5 local gray matter atrophies along component B was identified.

Conclusions: Our analysis shows the importance of analyzing dimensions with respect to subgroups as psychiatric traits likely stem from continuous alterations from healthy populations toward disorder biotypes. Yielding statistical correlations that consider the heterogeneity of the disorder and its subtypes with respect to the healthy population seems to be a promising idea to enhance key findings in schizophrenia mechanism understanding. Interestingly, both components, though neuroanatomically independent, led to the same clinical affections, but slightly different cognitive issues. These findings enrich the field of possibilities to develop clinical trials and precision-medicine drugs that are tailored on biological insights rather than

clinical and cognitive observations.

3.3.2 Introduction

The identification of consistent biomarkers in schizophrenia is crucial to understanding the mechanisms that underpin schizophrenia disorder and to develop personalized treatments [2, 96, 135, 304]. However, this disorder appears extremely heterogeneous in terms of clinical [64, 86, 277], genetic [231, 260, 15] and neuroanatomic [57, 186, 89] patterns. Therefore, recent research has attempted to stratify the disorder into homogeneous biotypes [127, 78, 57] which may enhance diagnosis detection and drive the search for relevant biomarkers by maximizing the signal-to-noise ratio [78].

Currently, the nosology and the diagnosis assessment in the psychiatric clinical routine are based on clinical observations and questionnaires. Notably, some works have attempted to refine the schizophrenia disorders' nosology based on cognitive [301, 298, 124], or clinical observations [186, 287, 156]. However, several initiatives motivated the need for identifying biological biomarkers rather than behavioral markers, to build a nosology that bridges the heterogeneous behavioral observations with the underlying biological processes. Notably, recent efforts have focused on studying the correlations between neuro-anatomical variations and behavioral biotypes [301, 320, 305, 186, 156]. Therefore, neuroanatomical patterns can be observed in a non-invasive manner and tangibly result from the interaction between polygenic risk factors and environmental stress factors [231, 111].

Prior research attempts have used an unsupervised clustering algorithm to group patients with similar anatomical patterns. For instance, one method [301] using the k-means algorithm identified two clusters, a cognitively impaired and a preserved subtype, which correlated with atrophies in basal ganglia and cerebellum areas. Another unsupervised clustering algorithm to find two subgroups with differences in cognitive performance and illness duration. However, these differences may take root from differences in terms of age, sex, and other factors of variability that also exist in the healthy population.

Prior research attempts have been undermined by confounding variables such as age, sex, and acquisition site. Although methods [289, 94, 95, 103] have been developed to evict these confounding factors, neuroanatomical variability remains dominated by non-specific factors that may underpin the neuro-anatomical-based subtype discovery. All the more so as these confounding factors are not strictly limited to known covariates such as age, sex, or acquisition sites as other factors may come into play such as IQ, education, ethnicity, urbanicity, etc...

Another range of methods, entitled normative models [195, 194, 196, 149, 148] proposed to

chart the natural variability of a brain modality with respect to covariates of interest. These methods enable overcoming the confounding issues raised by variables such as age, sex, or acquisition site. However, given age, sex, and site, for instance, it requires estimating the healthy variability among individuals of the same age, and sex and coming from the same acquisition site.

Another line of research investigated hybrid methods [127, 179, 280, 270, 62, 45] designed to spot the bio-markers that are useful to discriminate healthy controls from patients. Then, by focusing on contrasting biomarkers, these methods are designed to identify differences among the disorder effects. Therefore, it enables the production of clusters based on variability that is not driven by common characteristics (age, sex, or education for example). Instead, it focuses on neuro-anatomical deviations that correlate with disorder-related traits (positive symptoms, negative symptoms, cognitive decline, speech disorganization, psychosis, etc...).

For a clear understanding, we propose to entitle the common factors between healthy and diseased populations: general variability. Concerning the traits that characterize the heterogeneity within the disorder, we call it: specific variability. The general and specific variability differences are illustrated in Fig. 3.13.

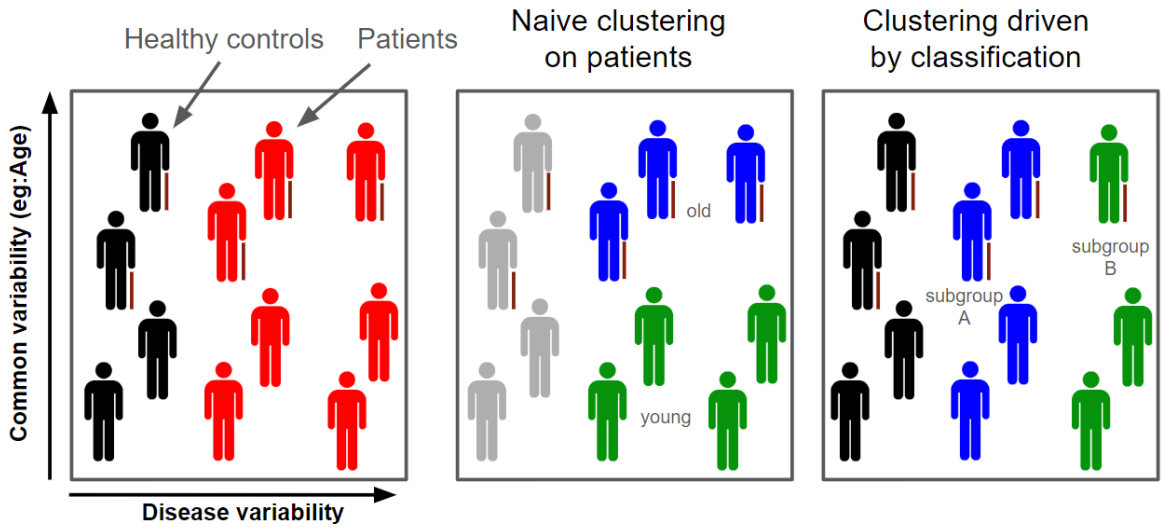


Figure 3.13: Subtype discovery in clinical research. Given a healthy population (black) and a pathological population (red) (left plot), we assume the existence of homogeneous subtypes within the disorder. However, the general variability (which stems from age or sex, for ex.) is observed in both healthy controls and patients. Therefore, a naive clustering of patients often yields a non-specific solution (middle plot). Nevertheless, the use of a classification task (healthy controls vs patients) helps to find direction(s) (horizontal arrow) that discards non-specific variability to emphasize more disorder-related differences (right plot).

3.3.3 Materials and methods

Study sample and image acquisition Neuroimaging data from several studies were pooled together to produce this analysis. The different studies contain Healthy Control (HC) and patients with Schizophrenia (SZ). In detail, we used SCHIZCONNECT [292] (368 HC and 275 SZ) and BSNIP [271] (199 HC, 190 SZ). SCHIZCONNECT-VIP encompasses 4 publicly available cohorts of controls and patients with schizophrenia. These cohorts encompass heterogeneous acquisition scanners and geographical sites. Regarding the cohort Bipolar and Schizophrenia Network for Phenotype Analysis (BSNIP), images were acquired at 5 different centers with 3T scanners across the United States of America. We chose neuro-imaging features related to the local volumes of brain gray matter measured in 142 regions of interest (identified using to CAT12 software [98, 99] from the SPM toolbox).

	Patients			Controls		
Study	**Age (avg +/- std)**	**Sex (%F)**	**Site**	**Age (avg +/- std)**	**Sex (%F)**	**Site**
BSNIP	34.33+/-12.28	31.0	Baltimore : 83 Boston : 24 Dallas : 22 Detroit : 6 Hartford : 55	38.64+/-12.46	58.2	Baltimore : 58 Boston : 25 Dallas : 44 Detroit : 21 Hartford : 51
ICAAR-START	21.95+/-2.60	31.8	Sainte-Anne : 16	21.31+/-2.25	43.7	Sainte-Anne : 22
PRAGUE	N/A	N/A	N/A	27.74 +/- 6.70	55.5	Prague : 90
SCHIZCONNECT-VIP	34.50+/-11.97	27.6	MRN : 77 NU : 42 WUSTL : 117 vip : 39	32.35+/-12.51	472	MRN : 87 NU : 38 WUSTL : 152 vip : 53

Figure 3.14: Demographic and acquisition statistics of the dataset.

Subtype Discovery within a pathology with UCSL In this work, we propose to use a recent machine learning algorithm for subtype discovery, called UCSL [179], to identify subtypes in schizophrenia. We identify subtypes based on structural neuro-biomarkers rather than symptoms. This Subtype discovery method aims at identifying subtypes describing the specific variability that is proper to a given pathology.

More specifically, UCSL proposes to identify the specific variability by jointly optimizing a set of linear supervised classifiers and a clustering method. Iteratively, the linear models are trained to classify each inferred disorder subtype from the healthy samples. Then, the clustering

method refines its subtypes inference by leveraging the linear model coefficient to put more emphasis on the features likely to be important in the discrimination task. Therefore, to identify the sources of variability within the disorder, UCSL performs a clustering that only captures the components useful for the classification task. Thus, the identified subtypes do not rely on general variability, such as age, sex, or education for example, but rather on the specific variability that is proper to the pathological population. Moreover, as it consistently ignores the features that do not help the classification task, UCSL does not need the use of any additional strategy to disentangle confounding variables. UCSL also points out components of interest, *i.e.*: independent groups of features that separate the healthy population from each subgroup. These components enable us to perform statistical analyses that take into account a potential continuum between the general healthy population and each of the subgroups. We depict the training scheme in Fig. 3.15.

Reproducibility analysis and hyper-parameters The optimal hyper-parameters were determined using a reproducibility analysis. Several hyper-parameters are of particular interest, some others were chosen by default, as proposed in the original method UCSL. Of interest, the number of subtypes is crucial and its choice will determine the tone of the following analysis. Additionally, the choice of the linear classification method and the choice of the clustering method embedded within UCSL have a severe impact on the subtype inference. In our study, we assume that the subtypes discovered are the most relevant when their inference is the most stable and reproducible. A measure of reproducibility and stability was performed as proposed in similar methods [280].

In detail, an average measure with a standard error was produced by successively splitting (10 times) the dataset into train and validation sets (respectively 66.66% and 33.33%). Given a set of parameters, for each split, we fit an instance of UCSL on the training set. Then, we produce a clustering inference on SZ samples of the validation set. Finally, we compute the similarities between each clustering inference of a given set of parameters. The similarity measure we chose was the Adjusted Rand Index Score (ARI).

We found that a solution with 2 clusters was the most reproducible and stable (Adjusted Rand Index Score: 0.661 ± 0.324). Other hyperparameters were chosen thanks to this reproducibility analysis. Notably, the type of covariance matrix used in the Gaussian mixture method embedded in UCSL was chosen between a full covariance matrix and a spherical covariance matrix. The latter was shown to be the most reproducible. The linear model embedded in UCSL was picked between a Support Vector Classifier and a Logistic Regression. The latter, with a tol-

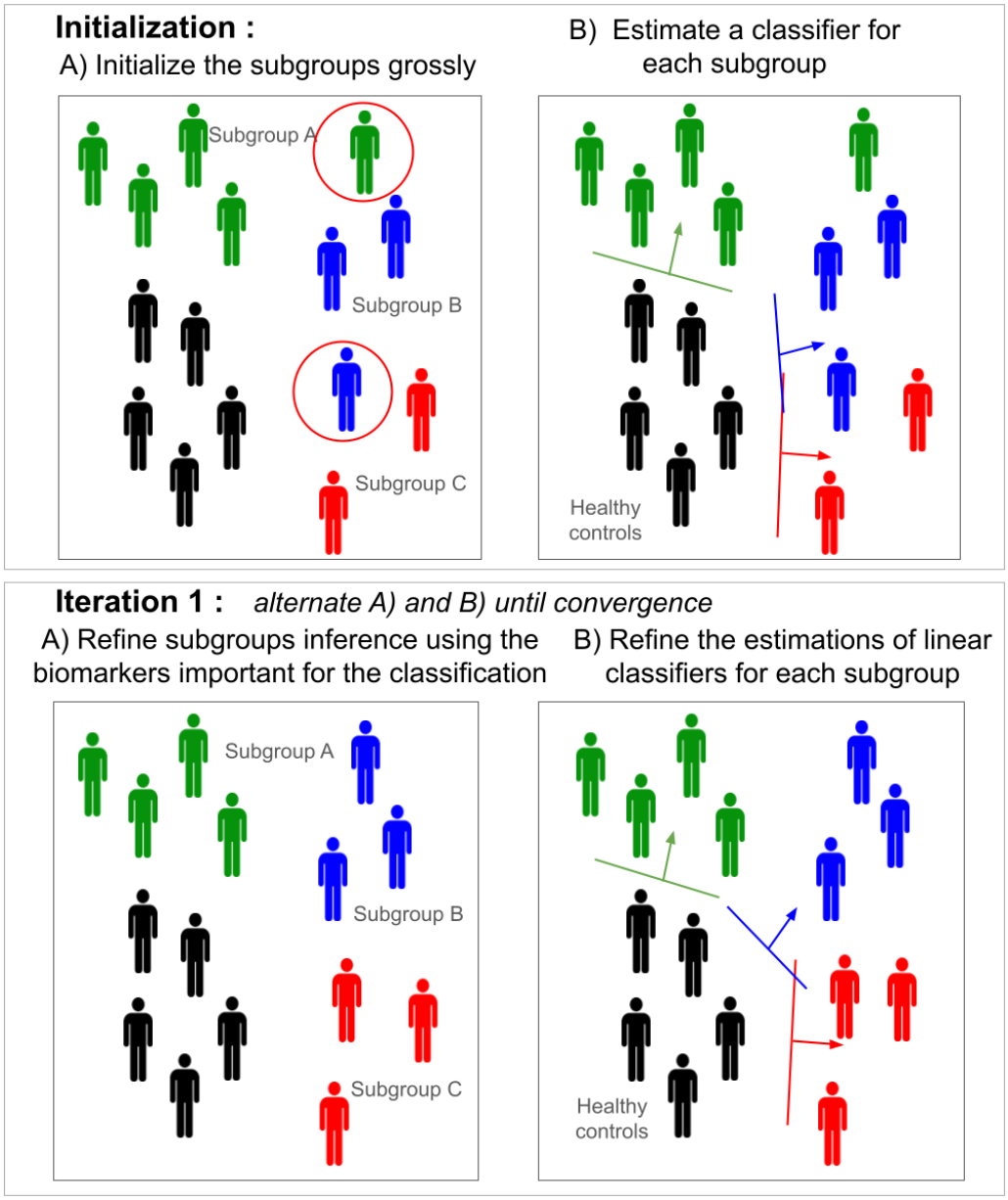


Figure 3.15: A scheme of the method UCSL. The Machine Learning method UCSL iterates between a clustering step and a classification step until the discovered subtypes are stable. Given gross clusters, a classification boundary is estimated for each subtype (vs Healthy controls). This classification step is illustrated in the left plot. Given a set of linear classifiers, we extract a set of components (arrows) that discriminate healthy controls from pathological subtypes. Then, these components are leveraged to guide the clustering step. The clustering step refines the subtype inference by weighting the features according to their importance in the classification task. Hence, it ensures that the subtypes stem from variability that is specific to the disorder.

erance parameter C=1, seemed to be the most reproducible. Also, the best stability threshold in order to decide when to stop UCSL Expectation-Maximization iterative optimization was found to be ARI= 0.85 (among [0.85, 0.9, 0.95]).

Furthermore, in order to guarantee reproducibility with respect to the initialization we propose a model selection process. First, the UCSL algorithm was trained 100 times, with a different initialization each time. For a given run i, we calculated the average Rand Index Score metric between the clustering i and all of the other runs clustering (so that $j \neq i$). Then, the method that produces the clustering that is the most similar to the others was chosen. Therefore, this easy process enables us to obtain the fitted method which is the most similar to the others. Notably, it guarantees that we do not tune the random seed to obtain our results and it ensures the exact reproducibility of the method. We wish to encourage subtype discovery relative works to employ this process or to discuss this point.

A measure of the disorder-specific of identified subgroups In its description, we guaranteed that UCSL embeds the disorder-specific variability exhibition in its core (with respect to the common variability with the healthy population). Nevertheless, we still propose a simple and reproducible process to measure how disorder-specific the inferred clustering is.

As assumed earlier, an ill-posed clustering method is likely to yield clusters based on common variability (e.g.: young vs old patients). Hence, the clustering inference of patient samples should be similar whether the method was fitted on healthy controls or pathological patients. On the contrary, it is unlikely for a method like UCSL because it stems its clustering rule from the disorder-specific variability only. Therefore, a quantitative measure of how a clustering method exhibits disorder-specific variability can be made. For that, it simply needs to estimate the correlation between the clustering yielded by the method fitted on the healthy population and the method fitted on the diseased population.

The Adjusted Rand Score metric was chosen as a similarity measure between two clustering inferences. This similarity score is standard. Anyway, we observed the same results when choosing the Adjusted Mutual Information Score or the Chi-squared test. An average measure with a standard error was produced by successively splitting (10 times) the dataset into train and validation sets (respectively 66.66% and 33.33%). Given a split, we fit an instance of UCSL to discover subtypes within the SZ population while contrasting with the healthy control population. Then, we fit another instance to cluster the healthy samples while contrasting them with the population of patients. Finally, we infer two clusterings (one per instance) on the validation set SZ samples and estimate the similarity between both inferences. To compare

with, we propose to fit a K-Means instance to discover clusters within the SZ population. Then, we fit another instance to cluster the healthy population. Hence, we infer two clusterings (one per instance) on the validation SZ samples and estimate the similarity between both clusterings. The disorder-specificity score found with the fully unsupervised methods K-Means was ARI= 0.367 ± 0.061. The results show that a naive clustering method yields a similar clustering if fitted on a healthy or diseased population. This demonstrates the need to use methods that explicitly search patterns that are proper to the pathological samples. On the other hand, UCSL yields a dissimilar clustering (ARI= 0.0 ± 0.0) when looking for subtypes in the pathological (resp. healthy) population while contrasting with the healthy (resp. pathological) population.

3.3.4 Results

Two subtypes, entitled subtypes A and B, were identified along corresponding components thanks to the previously described process. These subtypes respectively represent 427 samples (87.68%) and 60 samples (12.32%) among the SZ patients. Subtypes are equally distributed in terms of age (average values of respectively 34.1 ± 0.61 and 32.2 ± 1.48) and sex (average percentage of females of respectively 0.28 ± 0.02.

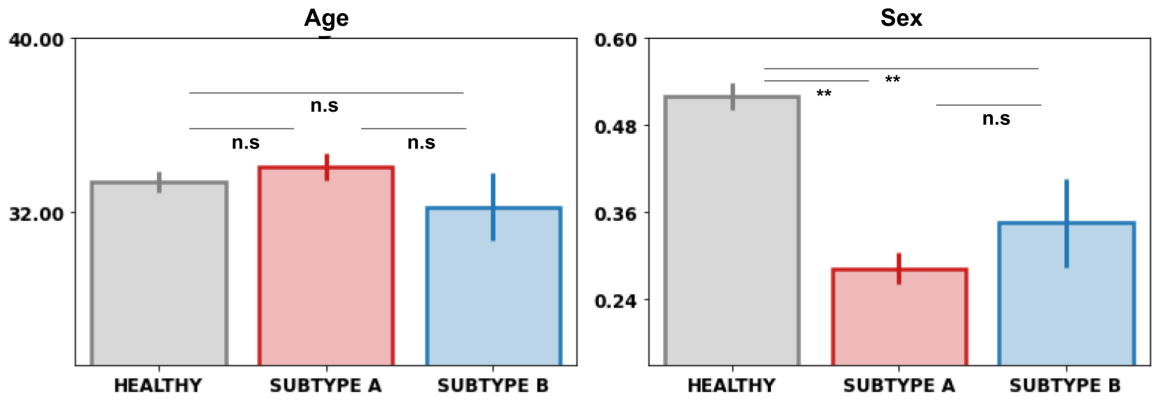


Figure 3.16: Distribution of metadata across the subtypes. (upper left plot) average population age for a healthy population and per subtype. (upper right plot) average population gender for a healthy population and per subtype. No significant differences were found when comparing the distributions of age and sex between the subtypes.

Fig. 3.16 illustrates the differences in terms of age and sex distributions between the subtypes and the healthy population. In terms of global atrophy, the patients within subgroup A show a reduced volume of white matter (WM) and gray matter (GM) compared to healthy samples and to subtype B, this is illustrated in Fig. 3.17. Concerning the volumes of cerebrospinal fluid

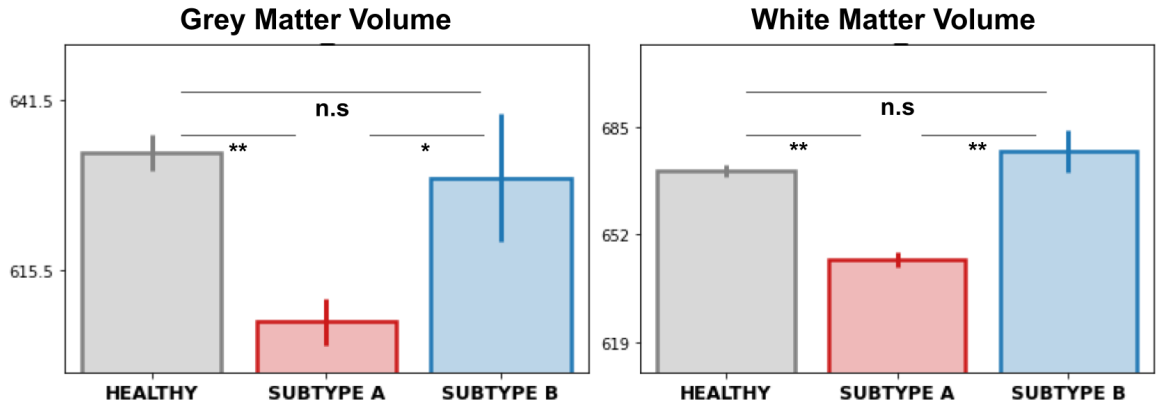


Figure 3.17: Differences in terms of average grey and white matter volumes. (upper left plot) average population age for a healthy population and per subtype. (upper right plot) average population gender for a healthy population and per subtype. No significant differences were found when comparing the distributions of age and sex between the subtypes.

(CSF), no significant differences were observed between the subtypes A and B.

3.3.5 Compute cognitive and clinical measures associations

In order to produce an analysis that is more suited to this assumption, we searched for correlations between phenotype scores and directions (ie: dimensions) rather than with homogeneous subtypes. In practice, for each inferred subtype, UCSL provides an associated component for each subgroup in the original features space. These dimensions can be decomposed into an orthonormal basis in order to project inputs onto a plot which illustrates the discovery of subtypes and their associated components, Fig. 3.18.

Linear correlations adjusted for age and sex were run in order to assess the correlation between our components and the clinical and cognitive scores. Given a clinical or cognitive scale, we assess how component A (resp. B) correlates with the clinical or cognitive measures of healthy and subgroup A's (resp. subgroup B's) patients with a regression model implemented with the library statsmodel in Python. The psychiatric scores we considered were the following: SIPSD (Structured Interview of Disorganization Symptoms), SIPSG (General Symptoms), SANS (Scale for Assessment of Negative Symptoms), SAPS (Scale for Assessment of Positive Symptoms), and MADRS (Montgomery Asberg Depression Rating Scale). Component A shows a correlation between multiple of the whole spectrum of psychiatric scales and a severe global cognitive decline (except for TMT B - A). Component B exhibits correlations with all clinical scales except for the general symptoms scale, and with two cognitive scales: the WMS logical

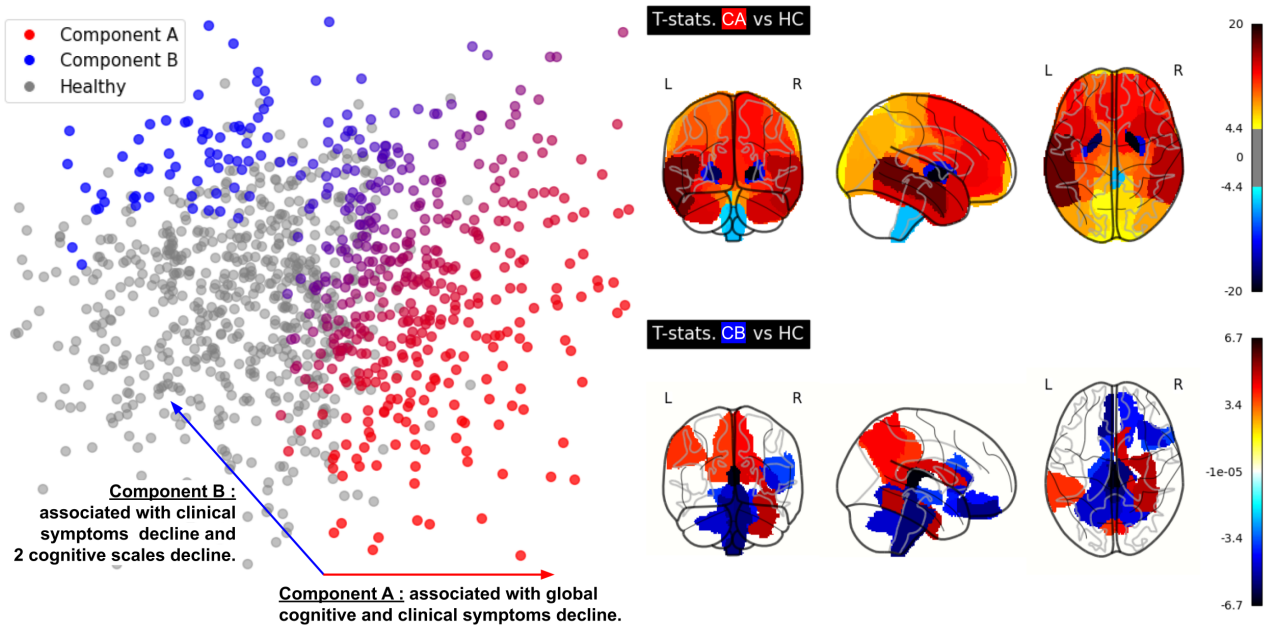


Figure 3.18: Projections of the brain gray matter volumes onto UCSL's 2D orthonormalized latent space and atlases of gray matter deviations along pathological components, red zones: atrophies, blue zones: hypertrophies. Red and blue axes show the discriminative dimensions between healthy controls and schizophrenia subtypes. Discriminative dimensions were found with the UCSL algorithm. (left plot) To take into account the continuum between the subtypes, statistical correlations between the discriminative components and the cognitive and clinical scores were estimated. The component A was found to be associated with a global cognitive decline and more severe overall symptoms manifestation. Component B also revealed an overall association with clinical scales but less severely, however, it only correlates with two cognitive scales that measure verbal episodic memory, and goal maintenance in working memory. (right plot) In terms of atrophy patterns, component A revealed a global gray matter atrophy over around 90 regions of interest and bilateral hypertrophies in the putamen, palladium, caudate, and brain stem, while component B revealed a set of local gray matter atrophies in the precuneus (bilaterally), in the peripheral area of the lateral right ventricle, left supramarginal gyrus, and right fusiform.

memory measure and the AX-CPT scale that measures goal maintenance in working memory. A summary of these associations with the clinical variables of interest is displayed in Fig .3.19.

We also compared the distributions of the cognitive and clinical measures between the subgroups with a statistical correlation computed at the group level. We found no significant differences in terms of clinical scales, and only two significant differences in terms of cognitive scales as subgroup A's patients appeared more affected on language skills tests (WAIS vocabulary), and the verbal, short-term, and working, memory test (WMS digit span) than subgroup B's

patients.

	Component A vs HC			Component B vs HC			Subgroup A vs Subgroup B		
Cognitive tests	t-stat	p-value	sign.	t-stat	p-value	sign.	t-stat	p-value	sign.
A-X Context Performance Test	**-3.75**	**0.002**	**	**-3.74**	**0.002**	**	-0.36	1.0	n.s
WMS Logical Memory	**-8.12**	**2.0e-13**	***	**-3.50**	**0.005**	**	-0.37	1.0	n.s
WMS Family Picture	**-7.32**	**3.2e-11**	***	-2.50	0.12	n.s	-0.049	1.0	n.s
WAIS Matrix Reasoning	**-5.05**	**7.7e-06**	***	-1.31	1.0	n.s	-1.93	0.49	n.s
Card Sorting Test	**5.51**	**8.0e-07**	***	2.70	0.06	n.s	0.184	1.0	n.s
WAIS vocabulary	**-6.35**	**9.2e-09**	***	-2.57	0.099	n.s	**-2.81**	**0.05**	*
WMS Digit Span	**-4.09**	**0.0005**	***	-1.52	1.0	n.s	**-3.01**	**0.028**	*
WMS Spatial Span	**-5.89**	**1.1e-07**	***	-1.84	0.60	n.s	-1.98	0.44	n.s
TMT B - A	0.40	1.0	n.s	0.17	1.0	n.s	1.42	1.0	n.s
SIPSD	**5.96**	**4.2e-8**	***	**3.52**	**0.003**	***	-1.25	1.0	n.s
SIPSG	**3.46**	**0.003**	***	2.49	0.069	n.s	1.51	0.66	n.s
SANS	**9.46**	**1.4e-17**	***	**3.77**	**0.001**	***	0.76	1.0	n.s
SAPS	**7.03**	**1.0e-10**	***	**4.57**	**4.9e-05**	***	-0.08	1.0	n.s
MADRS	**3.12**	**0.010**	**	**2.83**	**0.026**	*	0.76	1.0	n.s

Figure 3.19: Correlations between the clinical, and cognitive scales and the components inferred by UCSL. In detail, a linear regression adjusted for age and sex covariable was fitted to regress the cognitive scores from two components (associated with Subgroup A or with Subgroup B). This table reports the t-test and the p-value associated with the components variable. Besides, we also provide the statistical tests between the subgroups distributions of cognitive and clinical scales.

3.3.6 Identifying and analyzing neuro-anatomical deviations

To identify an underlying set of bio-markers that code for each subtype, we computed correlations for each region of interest between gray matter volumetric measures and each of the subgroup's components. Precisely, we assess how component A (resp. B) correlates with each

region of interest gray matter volumetric measures of healthy and subgroup A's (resp. subgroup B's) patients with a regression model adjusted for age and sex implemented with the library statsmodel in Python. We thus compute a set of correlations between biological readouts and the subgroup attribution score (projection of each individual's data point on the subgroup's component). Component A revealed a global gray matter atrophy over around 90 regions of interest and bilateral hypertrophies in the putamen, palladium, caudate, and brain stem, while component B revealed a set of local gray matter atrophies in the precuneus (bilaterally), in the peripheral area of the lateral right ventricle, left supramarginal gyrus, and right fusiform.

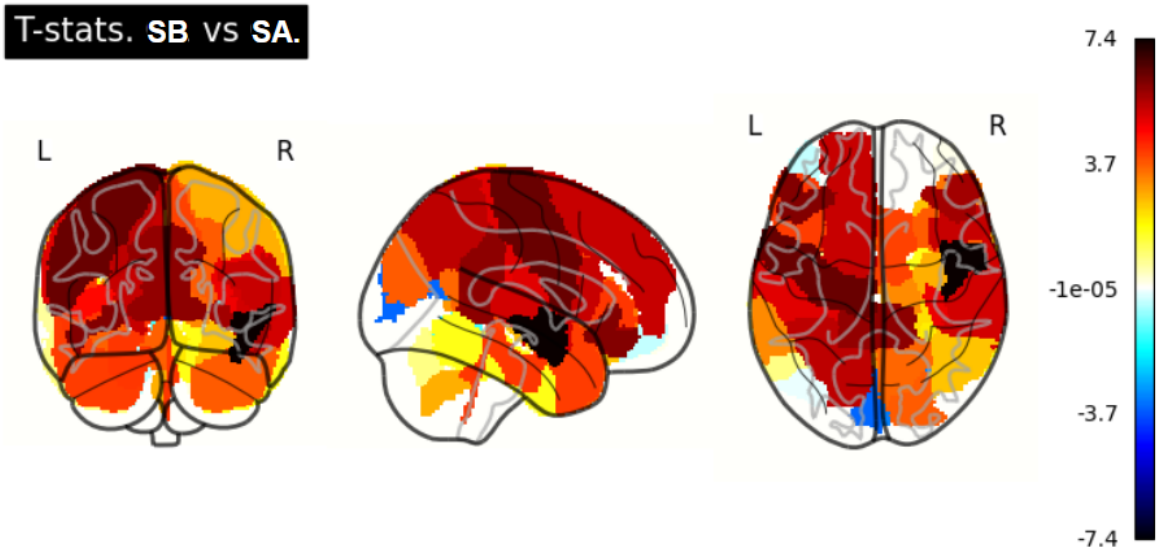


Figure 3.20: Atrophies map when comparing Subgroup A and Subgroup B (adjusted on age and sex covariates). Red zones show atrophies of subgroup B compared to Subgroup A. Blue zones show atrophies of subgroup B compared to Subgroup A. We observe a general atrophy of the Subgroup A compared to the Subgroup A.

3.3.7 Conclusion and Discussion

In this article, we proposed to identify consistent subtypes within a population suffering from schizophrenia disorder. Two subtypes were obtained by using UCSL, a subtype discovery method that also extracts a set of independent components of interest. This method ensures that the discovered subtypes and components stem from disorder-specific features by contrasting them with a healthy population. A disease-specificity measure was proposed, and coherent results were found when comparing UCSL with a naive k-means algorithm. The stability and reproducibility of the results were ensured for reproducibility.

Our results indicate the discovery of two subtypes, evenly distributed in terms of age and sex repartition. Subgroup B represents 13%, while subgroup A represents 87% of the schizophrenia cohort. Differently from subgroup B, subgroup A exhibits a clear gray and white matter overall loss compared to the healthy population. Our analysis revealed that both subgroups do not differ clinically, even though different biological abnormalities underpin them. Subgroups are yet different in terms of cognitive abilities, with subgroup B being more preserved than subgroup A. A set of only five gray matter local atrophies was identified in subgroup B, while subgroup A demonstrates clear generalized gray matter atrophies on around 90 regions of interest.
Although we assumed the existence of homogeneous sub-groups among schizophrenia (e.g: paranoic, psychotic, etc...) patients, we performed statistical correlations with respect to components (i.e.: directions) rather than per subgroups. This allows us to analyze how each subgroup differs from the healthy general distribution while considering a potential severity gradient along the subgroup's associated component. To justify this choice, we refer to the DSM 5: it argues that each individual suffering from a psychiatric disorder may lie on a continuum rather than in dichotomous categories [69].
Interestingly, the subgroups we identified were not significantly different in terms of clinical phenotype, yet they stem from drastically different neuro-anatomical atrophy patterns. Interestingly, this result suggests that different biological processes can underpin the same clinical phenotype, which would explain why clinical-based stratifications have failed to produce relevant biotypes. Nevertheless, the subgroups we discovered still differ regarding cognitive abilities, suggesting that distinct neuro-anatomical abnormalities lead to distinct cognitive and behavioral phenotypes.
We believe that his work has discovered a clear, biologically deviating subgroup in schizophrenia on behalf of subgroup B. This subgroup represents only 13% of the cohort and clearly differs from the rest of the patients while remaining statistically distinguishable from healthy patients. In future works, our research team will attempt to reproduce these findings in independent cohorts and search for associations with fine-grained clinical symptoms (*e.g.*: loss of self, anhedonia, visual or auditory delusions, etc...), genetic patterns, and medication response. Nevertheless, the subgroups remain highly heterogeneous in terms of clinical symptoms and our method has not successfully disentangled each biomarkers-clinical scale's associations. A step forward would be to identify interpretable pathological distinct latent generative factors in each of the subgroups we found to further parse the biological causes that underpin each of the distinct clinical and cognitive scores measured in the clinical routine.

Chapter 4

Contrastive Analysis in medical imaging and neuroimaging

Chapter summary. This chapter investigates Contrastive Analysis methods, that aim at estimating the latent distinct and interpretable generative factors that underpin the neurobiological heterogeneity proper to the psychiatric disorder.
This field of statistical learning aims at separating "common" and "target" variability factors given a "source" dataset and a "target" dataset. In this thesis, the goal is to estimate, on the one hand, the projection that identifies healthy variability patterns and, on the other hand, the projection that identifies the "pathological signatures" that are exclusive to patients with psychiatric disorders.
Initially, the chapter introduces a novel contrastive variational autoencoder method called SepVAE. This method enhances existing approaches by incorporating a classification task within the pathological space and integrating a cost function based on mutual information to minimize redundancy between the common space and the pathological space. SepVAE's effectiveness is demonstrated through validation across various datasets, including vision, medical, and neuroimaging data.
Eventually, to provide a novel methodological perspective, a novel contrastive analysis strategy entitled SepCLR was developed. This method extends the framework of contrastive analysis methods to another promising family of methods: contrastive representation learning. These methodological contributions were then validated on vision, medical, and neuroimaging datasets.

Contents

4.1 SepVAE: a contrastive VAE to separate pathological from healthy patterns

Context: The previous chapter produced insights about the modeling of the pathological heterogeneity within schizophrenia. Notably, two homogeneous subgroups were identified driven by distinct neuroanatomical deviations, with comparable clinical profiles, but different cognitive affections. These results suggested that distinct neuroanatomical deviations lead to distinct cognitive and behavioral phenotypes, which justified the research and use of subgroup discovery methods in psychiatric research. Nevertheless, the dominant subgroup we discovered (comprising around 87% of the diseased cohort) remains largely heterogeneous in terms of neuroanatomical deviations (around 100 regions of interest concerned by gray matter atrophies) and clinical profiles (with a decline on every clinical and cognitive scale). This observation suggested that a large part of the schizophrenia cohort still encompasses several pathological profiles. Besides, one could assume that these profiles may not be underpinned by homogeneous subgroups, but rather by a sum of distinct pathological factors (a restricted set of neuroanatomical deviations associated with a restricted set of clinical or cognitive scales). This change of paradigm is illustrated in Fig. 4.1.

Motivation: At this point of the manuscript, the objective is to develop a method that identifies the distinct and interpretable generative factors that are specific to the pathology (*i.e.* that do not exist in the healthy population) without necessarily assum ing the presence of homogeneous pathological subgroups. To address this objective, endeavors have been furnished to investigate and develop Contrastive Analysis methods, and particularly Contrastive Analysis Variational Auto-Encoders.

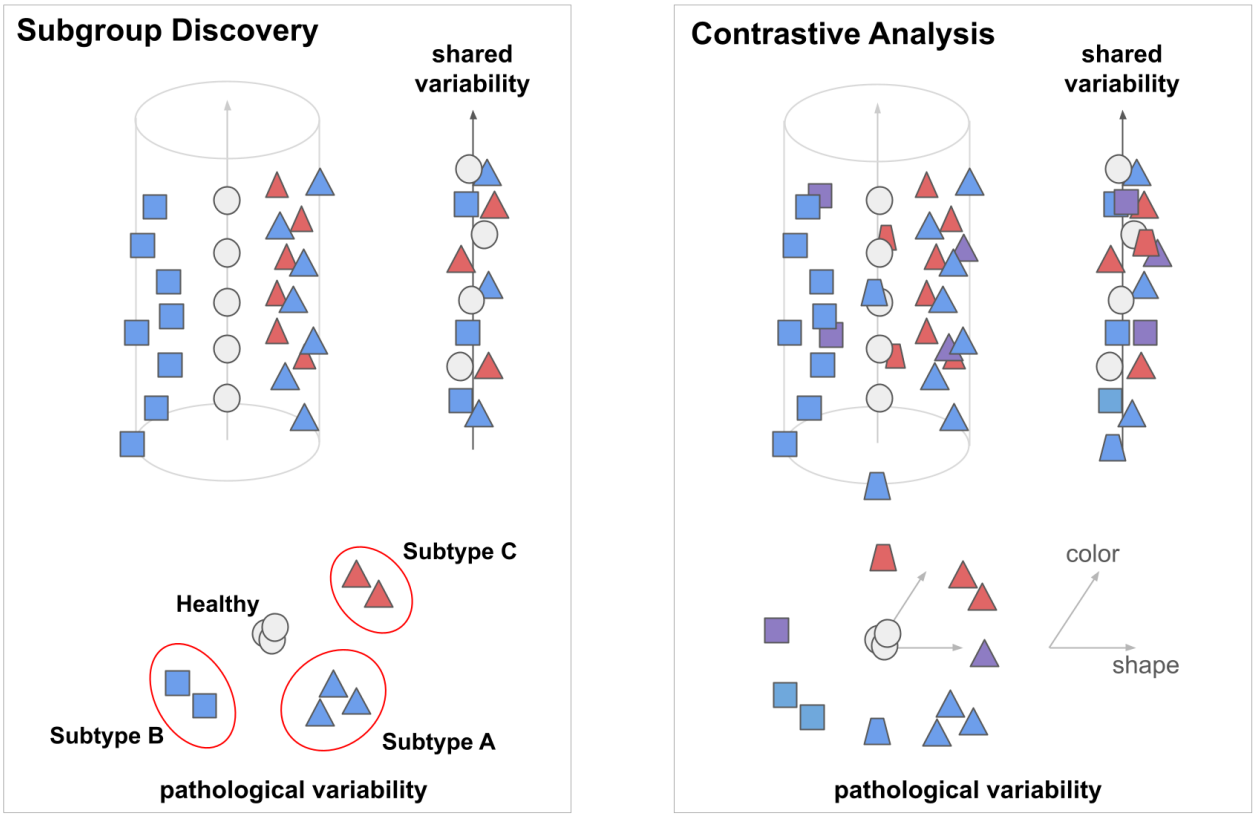


Figure 4.1: Illustration of two different paradigms when modeling the pathological heterogeneity. (left) The Subtype Discovery paradigm assumes 1) that patients share common variability (vertical axis) with healthy controls (white circles) and 2) that homogeneous subgroups can be distinguished from disorder-specific variability (horizontal hyperplane), *e.g.* blue squares, blue triangles, and red triangles. In the Subgroup Discovery paradigm, clusters are assumed to be homogeneous and separable (separated with low-density zones). Subgroup Discovery methods generally estimate subgroups by focusing on the projections that enable discriminating healthy from patients (in that case, it disregards the vertical axis). (right) The Contrastive Analysis paradigm assumes 1) that patients share common variability (vertical) with healthy controls (white circles), and 2) that distinct and interpretable generative factors can be identified from disorder-specific variability (horizontal hyperplane), *e.g.* color and shape continuous factors of generation. Contrastive Analysis methods generally separate shared variability factors (vertical projection) from disorder-specific variability factors (horizontal projections). In the contrastive analysis setting, it is assumed that the data points are continuously distributed along the latent factors of variability, there is thus a continuum in terms of shape (with squared-triangle data points) and color (with purple data points).

4.1.1 Abstract

Contrastive Analysis VAE (CA-VAEs) is a family of Variational auto-encoders (VAEs) that aims at separating the common factors of variation between a *background* dataset (BG) (*i.e.,* healthy subjects) and a *target* dataset (TG) (*i.e.,* patients) from the ones that only exist in the target dataset. To do so, these methods separate the latent space into a set of **salient** features (*i.e.,* proper to the target dataset) and a set of **common** features (*i.e.,* exist in both datasets). Currently, all models fail to prevent the sharing of information between latent spaces effectively and to capture all salient factors of variation. To this end, we introduce two crucial regularization losses: a disentangling term between common and salient representations and a classification term between background and target samples in the salient space. We show improved performance than previous CA-VAEs methods on three medical applications and a natural images dataset (CelebA). Code and datasets are available on GitHub: https://github.com/neurospin-projects/2023_rlouiset_sepvae.

4.1.2 Introduction

One of the goals of unsupervised learning is to learn a compact, latent representation of a dataset, capturing the underlying factors of variation. Furthermore, the estimated latent dimensions should describe distinct, noticeable, and semantically meaningful variations. One way to achieve that is to use a generative model, like Variational Auto-Encoders (VAEs) [150, 122] and disentangling methods [122, 35, 256, 325, 49, 9, 167]. Differently from these methods, which use a *single* dataset, in Contrastive Analysis (CA), researchers attempt to distinguish the latent factors that generate a *target* (TG) and a *background* (BG) dataset. Usually, it is assumed that target samples comprise additional (or modified) patterns with respect to background data. The goal is thus to estimate the **common** generative factors and the ones that are **target-specific** (or **salient**). This means that background data are fully encoded by some generative factors that are also **common** with the target data. On the other hand, target samples are assumed to be partly generated from strictly proper factors of variability, which we entitle **target-specific** or **salient** factors of variability. This formulation is particularly useful in medical applications where clinicians are interested in separating common (i.e., healthy) patterns from the salient (i.e., pathological) ones in an *intepretable* way.

For instance, consider two sets of data: 1) healthy neuro-anatomical MRIs (BG=*background dataset*) and 2) Alzheimer-affected patients' MRIs (TG=*target dataset*). As in [139, 12, 178], given these two datasets, neuroscientists would be interested in distinguishing common factors

of variations (*e.g.:* effects of aging, education or gender) from Alzheimer's specific markers (*e.g.:* temporal lobe atrophy, an increase of beta-amyloid plaques). Until recently, separating the various latent mechanisms that drive neuro-anatomical variability in neuro-degenerative disorders was considered hardly feasible. This can be attributed to the intertwining between the variability due to natural aging and the variability due to neurodegenerative disease development. The combined effects of both processes make hardly interpretable the potential discovery of novel bio-markers.

The objective of developing such a Contrastive Analysis method would be to help separate these processes. And thus identifying correlations between neuro-biological markers and pathological symptoms. In the **common features** space, aging patterns should correlate with normal cognitive decline, while **salient features** (*i.e.:* Alzheimer-specific patterns) should correlate with pathological cognitive decline.

Besides medical imaging, Contrastive Analysis (CA) methods cover various kinds of applications, like in pharmacology (placebo versus medicated populations), biology (pre-intervention vs. post-intervention cohorts) [324], and genetics (healthy vs. disorder population [140], [112]).

4.1.3 Related works

Variational Auto-Encoders (VAEs) [150] have advanced the field of unsupervised learning by generating new samples and capturing the underlying structure of the data onto a lower-dimensional data manifold. Compared to linear methods (e.g., PCA, ICA), VAEs make use of deep non-linear encoders to capture non-linear relationships in the data, leading to better performance on a variety of tasks.

Disentangling methods [122, 35, 256] enable learning the underlying factors of variation in the data. While disentangling [325, 49] is a desirable property for improving the control of the image generation process and the interpretation of the latent space [9, 167], these methods are usually based on a *single* dataset, and they do not explicitly use labels or multiple datasets to effectively estimate and separate the common and salient factors of variation.

Semi and weakly-supervised VAEs [199, 151, 192, 142] have proposed to integrate class labels in their training. However, these methods solely allow conditional generalization and better semantic expressivity rather than addressing the separation of the factors of variation between distinct datasets.

Contrastive Analysis (CA) works are explicitly designed to identify patterns that are unique to a target dataset compared to a background dataset. First attempts [327, 4, 100] employed linear methods in order to identify a projection that captures the variance of the target dataset

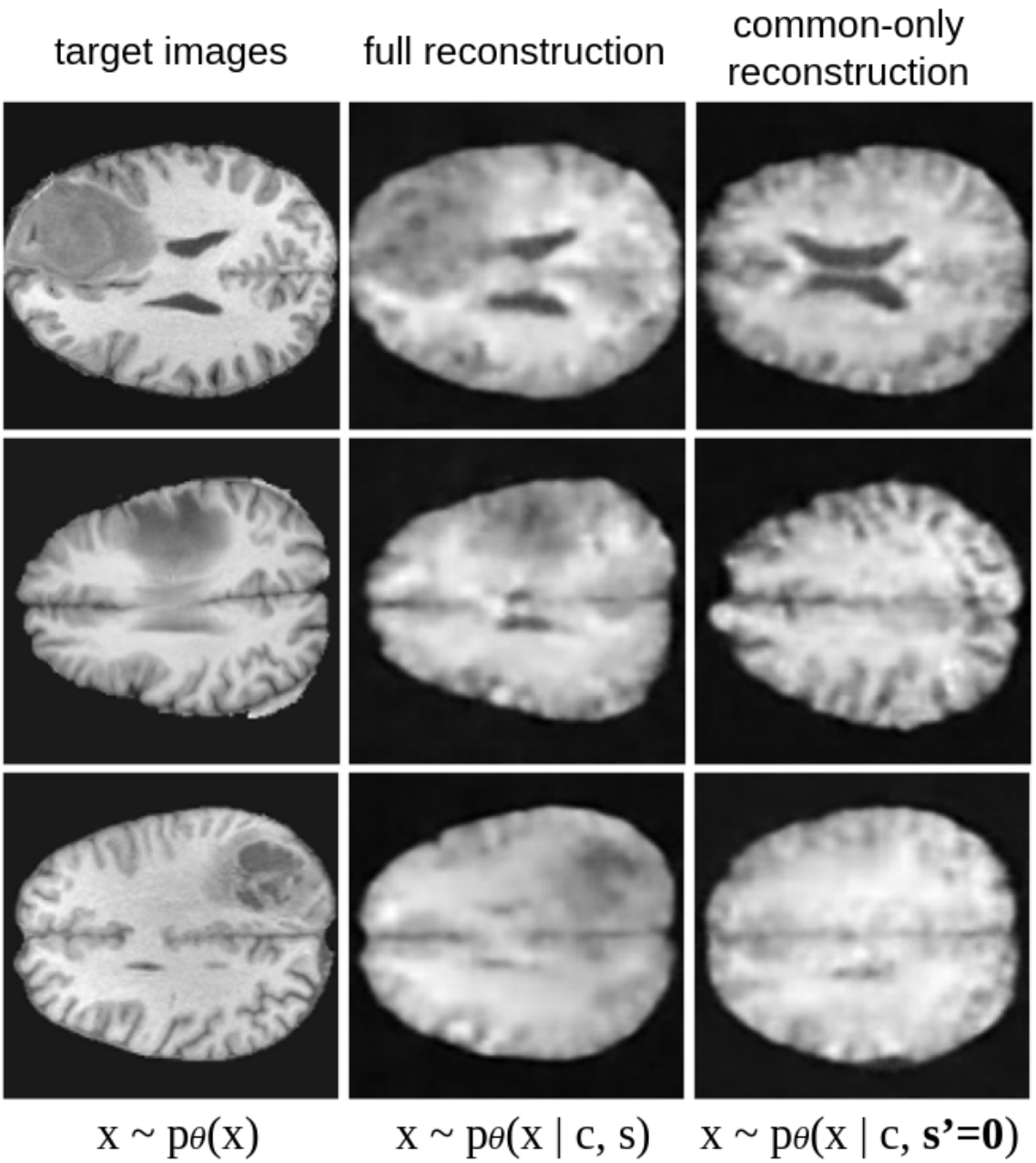


Figure 4.2: SepVAE reconstructions on BRATS 2021 dataset [208]. (Middle) full reconstructions using the estimated common and salient latent vectors. (Right) common-only reconstructions using the estimated common latent vectors and fixing the salient factors to s'. The common latent variables encode the healthy factors of variability (*e.g. :* brain shape and aspect), while the salient factors encode the pathological patterns (*e.g. :* tumors), which are not visible in the right columns (common-only).

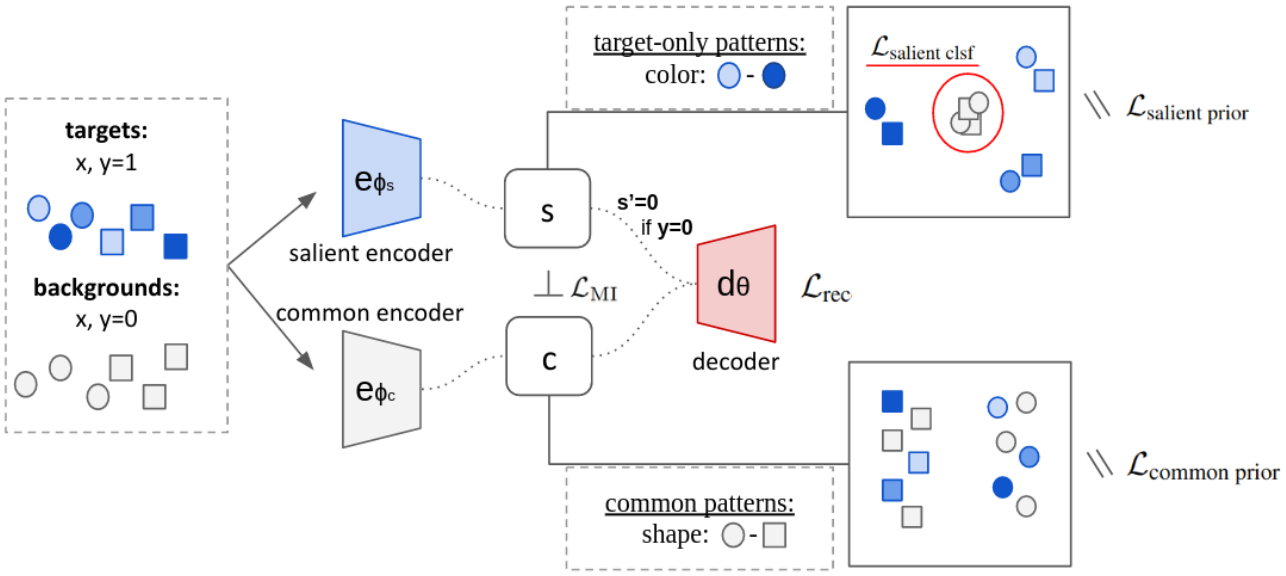


Figure 4.3: Illustration of SepVAE training. Target and background images are encoded with the same encoders e_{ϕ_s} and e_{ϕ_c}. The first encoder e_{ϕ_s} estimates the salient factors of variation s of the target samples ($y = 1$). Background samples ($y = 0$) salient space is set to an informationless value $s' = 0$. The second encoder e_{ϕ_c} estimates the common factors c. Images are reconstructed using a single decoder d_θ fed with the concatenation of c and s.

while minimizing the background information expressivity. However, due to their linearity, these methods had reduced learning expressivity and were also unable to produce satisfactory generation. Contrastive VAE [3, 299, 254, 240, 328, 55] have employed deep encoders to capture higher-level semantics. They usually rely on a latent space split into two parts, a common and a salient, produced by two different encoders. First methods, such as [254], employed two decoders (common and salient) and directly sum the common and salient reconstructions in the input space. This seems to be a very strong assumption, probably wrong when working with high-dimensional and complex images. For this reason, subsequent works used a single decoder, which takes as input the concatenation of both latent spaces. Importantly, when seeking to reconstruct background inputs, the decoder is fed with the concatenation of the common part and an informationless reference vector **s'**. This is usually chosen to be a null vector in order to reconstruct a null (i.e., empty) image by setting the decoder's biases to 0. Furthermore, to fully enforce the constraints and assumptions of the underlying CA generative model, previous methods have proposed different regularizations. Here, we analyze the most important ones with their advantages and shortcomings:

Minimizing background's variance in the salient space Pioneer works [254, 3] have

shown inconsistency between the encoding and the decoding task. While background samples are reconstructed from **s'**, the salient encoder does not encourage the background salient latents to be equal to **s'**. To fix this inconsistency, posterior works [299, 328, 55] have shown that explicitly nullifying the background variance in the salient space was beneficial. This regularization is necessary to avoid salient features explaining the background variability but not sufficient to prevent information leakage between common and salient spaces, as shown in [299].

Independence between common and salient spaces Only one work [3] proposed to prevent information leakage between the common and salient space by minimizing the total correlation (TC) between $q_{\phi_c,\phi_s}(c, s|x)$ and $q_{\phi_c}(c|x) \times q_{\phi_s}(s|x)$, in the same fashion as in FactorVAE [152]. This requires to independently train a discriminator $D_\lambda(.)$ that aims at approximating the ratio between the joint distribution $q(x) = q_{\phi_c,\phi_s}(c, s|x)$ and the marginal of the posteriors $\bar{q}(x) = q_{\phi_c}(c|x) \times q_{\phi_s}(s|x)$ via the density-ratio trick [211, 265]. In practice, [3]'s code does not use an independent optimizer for λ, which undermines the original contribution. Moreover, when incorrectly estimated, the TC can become negative, and its minimization can be harmful to the model's training.

Matching background and target common patterns Another work [299], has proposed to encourage the distribution in the common space to be the same across target samples and background samples. Mathematically, it is equivalent to minimizing the KL between $q_{\phi_c}(c|y = 0)$ and $q_{\phi_c}(c|y = 1)$ (or between $q_{\phi_c}(c)$ and $q_{\phi_c}(c|y)$). In practice, we argue that it may encourage undesirable *biases* to be captured by salient factors rather than common factors. For example, let's suppose that we have healthy subjects (*background* dataset) and patients (*target* dataset) and that patients are composed of both young and old individuals, whereas healthy subjects are only old. We would expect the CA method to capture the normal aging patterns (*i.e.:* the bias) in the common space. However, forcing both $q_{\phi_c}(c|x, y = 0)$ and $q_{\phi_c}(c|x, y = 1)$ to follow the same distribution in the common space would probably bring to a biased distribution and thus to leakage of information between salient and common factors (i.e., aging could be considered as a salient factor of the patient dataset).This behavior is not desirable, and we believe that the statistical independence between common and salient space is a more robust property.

Contributions Our contributions are three-fold:

• We develop a new Contrastive Analysis method: SepVAE, which is supported by a sound and versatile Evidence Lower BOund maximization framework.

• We identify and implement two properties: the salient space discriminability and the salient / common independence, that have not been successfully addressed by previous Contrastive

VAE methods.

• We provide a fair comparison with other SOTA CA-VAE methods on 3 medical applications and a natural image experiment.

4.1.4 Contrastive Variational Autoencoders

Let $(X, Y) = \{(x_i, y_i)\}_{i=1}^N$ be a data-set of images x_i associated with labels $y_i \in \{0, 1\}$, 0 for background and 1 for target. Both background and target samples are assumed to be i.i.d. from two different and unknown distributions that depend on two latent variables: $c_i \in \mathbf{R}^{D_c}$ and $s_i \in \mathbf{R}^{D_s}$. Our objective is to have a generative model $x_i \sim p_\theta(x|y_i, c_i, s_i)$ so that: 1- the **common** latent vectors $C = \{c_i\}_{i=1}^N$ should capture the common generative factors of variation between the background and target distributions and fully encode the background samples and 2- the **salient** latent vectors $S = \{s_i\}_{i=1}^N$ should capture the distinct and interpretable generative factors of variation of the target set (*i.e.,* patterns that are only present in the target dataset and not in the background dataset). Similar to previous works [3, 299, 328], we assume the generative process: $p_\theta(x, y, c, s) = p_\theta(x|c, s, y)p_\theta(c)p_\theta(s|y)p(y)$. Since $p_\theta(c, s|x, y)$ is hard to compute in practice, we approximate it using an auxiliary parametric distribution $q_\phi(c, s|x, y)$ and directly derive the Evidence Lower Bound of $\log p(x, y)$. Based on this generative latent variable model, one can derive the ELBO of the marginal log-likelihood $\log p(x, y)$,

$$-\log p_\theta(x, y) \leq \mathbf{E}_{c,s\sim q_{\phi_c,\phi_s}(c,s|x,y)} \log \frac{q_{\phi_c,\phi_s}(c, s|x, y)}{p_\theta(x, y, c, s)} \tag{4.1}$$

where we have introduced an auxiliary parametric distribution $q_\phi(c, s|x, y)$ to approximate $p_\theta(c, s|x, y)$.

From there, we can develop the lower bound into three terms, a conditional reconstruction term, a common space prior regularization, and a salient space prior regularization:

$$-\log p_\theta(x, y) \leq \underbrace{-\mathbf{E}_{c,s\sim q_{\phi_c,\phi_s}(c,s|x,y)} \log p_\theta(x|y, c, s)}_{\textbf{Conditional Reconstruction}} + \underbrace{KL(q_{\phi_c}(c|x)||p_\theta(c))}_{\textbf{b) Common prior}} + \underbrace{KL(q_{\phi_s}(s|x, y)||p_\theta(s|y))}_{\textbf{c) Salient prior}} \tag{4.2}$$

Here, we assume the independence of the auxiliary distributions (*i.e.:* $q_{\phi_c,\phi_s}(c, s|x, y) = q_{\phi_c}(c|x)q_{\phi_s}(s|x, y)$) and prior distributions (*i.e.:* $p_\theta(c, s) = p_\theta(c)p_\theta(s)$). Both $p_\theta(x|y_i, c_i, s_i)$ (i.e., single decoder) and $q_{\phi_c}(c|x)q_{\phi_s}(s|x, y)$ (i.e., two encoders) are assumed to follow a Gaussian distribution parametrized by a neural network. To reinforce the independence assumption between c and s, we introduce a Mutual Information regularization term $KL(q(c, s)||q(c)q(s))$.

Theoretically, this term is similar to the one in [3]. This property is desirable in order to ensure that the information is well separated between the latent spaces. However, in [3], the Mutual Information estimation and minimization are done simultaneously [1]. In this paper, we argue that the estimation of the Mutual Information requires the introduction of an independent optimizer, see Sec. 4.1.4. To further reduce the overlap of target and common distributions on the salient space, we also introduce a salient classification loss defined as $\mathbf{E}_{s\sim q_{\phi_s}(s|x,y)} \log p(y|s)$. By combining all these losses together, we obtain the final loss $\mathcal{L}$:

$$\mathcal{L} = \underbrace{-\mathbf{E}_{c,s\sim q_{\phi_c,\phi_s}(c,s|x,y)} \log p_\theta(x|c,s,y)}_{\textbf{a) Conditional Reconstruction}} + \underbrace{KL(q_{\phi_c}(c|x)||p_\theta(c))}_{\textbf{b) Common Prior}} + \underbrace{KL(q_{\phi_s}(s|x,y)||p_\theta(s|y))}_{\textbf{c) Salient Prior}} + \underbrace{KL(q(c,s)||q(c)q(s))}_{\textbf{e) Mutual Information}} - \underbrace{\mathbf{E}_{s\sim q_{\phi_s}(s|x,y)} \log p_\theta(y|s)}_{\textbf{d) Salient Classification}} \tag{4.3}$$

Conditional reconstruction:

The reconstruction loss term is given by $-\mathbf{E}_{c,s\sim q_{\phi_c,\phi_s}(c,s|x,y)} \log p_\theta(x|c,s,y)$. Given an image x (and a label y), a common and a salient latent vector can be drawn from q_{ϕ_c,ϕ_s} with the help of the reparameterization trick.
We assume that $p(x|c,s,y) \sim \mathcal{N}(d_\theta([c, ys + (1-y)s'], I)$, *i.e:* $p_\theta(x|c,s,y)$ follows a Gaussian distribution parameterized by θ, centered on $\mu_{\hat{x}} = d_\theta([c, ys+(1-y)s'])$ with identity covariance matrix, and d_θ is the decoder and $[.,.]$ denotes a concatenation.
Therefore, by developing the reconstruction loss term, we obtain the mean squared error between the input and the reconstruction: $\mathcal{L}_{\text{rec}} = \sum_{i=1}^{N} ||x - d_\theta([c, ys+(1-y)s'])||_2^2$. Importantly, for background samples, we set the salient latent vectors to **s'** $= 0$. This choice enables isolating the background factors of variability in the common space only.

Common prior

By assuming $p(c) \sim \mathcal{N}(0, I)$ and $q_{\phi_c}(c|x) \sim \mathcal{N}(\mu_\phi(x), \sigma_\phi(x,y))$, the KL loss has a closed form solution, as in standard VAEs. Here, both $\mu_\phi(x)$ and $\sigma_\phi(x,y)$ are the encoder outputs e_{ϕ_c}.

[1] In [3], Algorithm 1 suggests that the Mutual Information estimation and minimization depend on two distinct parameters update. However, in practice, in their code, a single optimizer is used. This is also confirmed in Sec. 3, where authors write: "discriminator is trained simultaneously with the encoder and decoder neural networks".

Salient prior:

To compute this regularization, we first need to develop $p_\theta(s) = \sum_y p(y)p_\theta(s|y)$, where we assume that $p(y)$ follows a Bernoulli distribution with probability equal to 0.5. Thus, the salient prior reduces to a formula that only depends on $p_\theta(s|y)$, which is conditioned by the knowledge of the label (0: background, 1: target). This allows us to distinguish between the salient priors of background samples ($p(s|y = 0)$) and target samples ($p(s|y = 1)$).
Similar to other CA-VAE methods, we assume that $p(s|y = 1) \sim \mathcal{N}(0, I)$ and , as in [328], that $p(s|x, y = 0) \sim \mathcal{N}(s', \sqrt{\sigma_p}I)$, with $s' = 0$ and $\sqrt{\sigma_p} < 1$, namely a Gaussian distribution centered on an informationless reference s' with a small constant variance σ_p. We preferred it to a Delta function $\delta(s = s')$ (as in [299]) because it eases the computation of the KL divergence (i.e., closed form) and it also means that we tolerate a small salient variation in the background (healthy) samples. In real applications, in particular medical ones, diagnosis labels can be noisy, and mild pathological patterns may exist in some healthy control subjects. Using such a prior, we tolerate these possible (erroneous) sources of variation.
Furthermore, one could also extend the proposed method to a continuous y, for instance, between 0 and 1, describing the severity of the disorder. Indeed, practitioners could define a function $\sigma_p(y)$ that would map the severity score y to a salient prior standard deviation (*e.g.,* $\sigma_p(y) = y$). In this way, we could extend our framework to the case where pathological variations would follow a continuum from no (or mild) to severe patterns.

Salient space classification;

In the salient prior regularization, as in previous works, we encourage background and target salient factors to match two different Gaussian distributions, both centered in 0 (we assume $s' = 0$) but with different covariance. However, we argue that target salient factors should be further encouraged to differ from the background ones in order to reduce the overlap of target and common distributions on the salient space and enhance the expressivity of the salient space.

To encourage target and background salient factors to be generated from different distributions, we propose to minimize a Binary Cross Entropy loss to distinguish the target from background samples in the salient space. Assuming that $p(y|s)$ follows a Bernoulli distribution parameterized by $f_\xi(s)$, a 2-layers classification Neural Network, we obtain a Binary Cross Entropy (BCE) loss between true labels y and predicted labels $\hat{y} = f_\xi(s)$.

Mutual Information minimization:

To promote independence between c and s, we minimize their mutual information, defined as the KL divergence between the joint distribution $q(c, s)$ and the marginals product $q(c)q(s)$. However, computing this quantity is not trivial, and it requires a few tricks in order to correctly estimate and minimize it. As in [3], it is possible to take inspiration from FactorVAE [152], which proposes to estimate the density-ratio between a joint distribution and the product of the marginals. In our case, we seek to enforce the independence between two sets of latent variables rather than between each latent variable of a set. The density-ratio trick [211, 265] allows us to estimate the quantity inside the log in Eq.4.4. First, we sample from $q(c, s)$ by randomly choosing a batch of images (x_i, y_i) and drawing their latent factors $[c_i, s_i]$ from the encoders e_{ϕ_c} and e_{ϕ_s}. Then, we sample from $q(c)q(s)$ by using the same batch of images where we shuffle the latent codes among images (*e.g.*, $[c_1, s_2]$, $[c_2, s_3]$, etc.). Once we obtained samples from both distributions, we trained an **independent** classifier $D_\lambda([c, s])$ to discriminate the samples drawn from the two distributions by minimizing a BCE loss. The classifier is then used to approximate the ratio in the KL divergence, and we can train the encoders e_{ϕ_c} and e_{ϕ_s} to minimize the resulting loss:

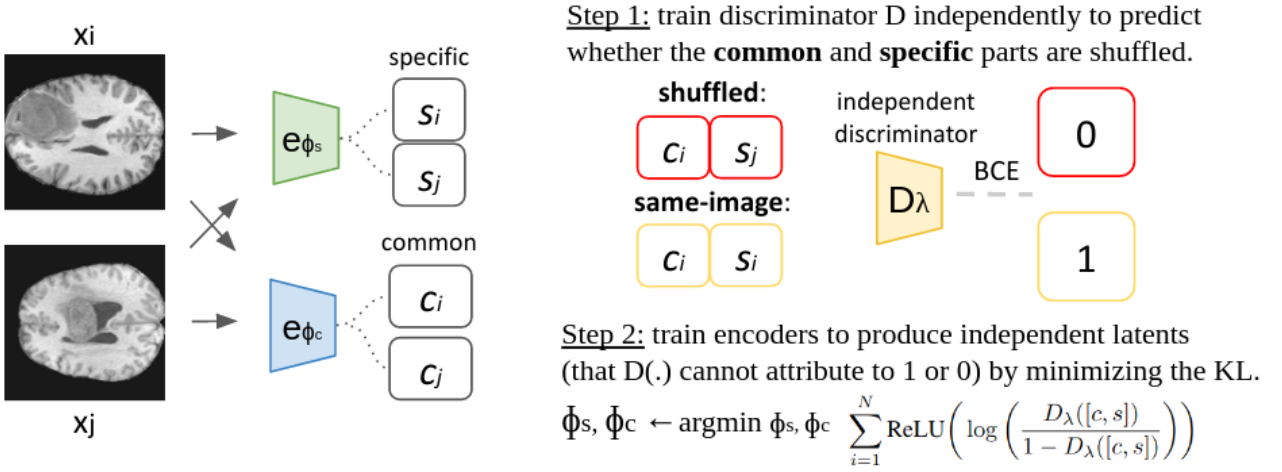
$$\mathcal{L}_{\text{MI}} = \mathbb{E}_{q(c,s)} \log \left(\frac{q(c, s)}{q(c)q(s)} \right) \approx \sum_i \text{ReLU} \left(\log \left(\frac{D_\lambda([c_i, s_i])}{1 - D_\lambda([c_i, s_i])} \right) \right) \tag{4.4}$$

where the ReLU function forces the estimate of the KL divergence to be positive, thus avoiding to back-propagate wrong estimates of the density ratio due to the simultaneous training of $D_\lambda([c, s])$. In [3], while Alg.1 of the original paper describes two distinct gradient updates, it is written that "This discriminator is trained simultaneously with the encoder and decoder neural networks". In practice, a single optimizer is used in their training code. In our work, we use an independent optimizer for D_λ, in order to ensure that the density ratio is well estimated. Furthermore, we freeze D_λ's parameters when minimizing the Mutual Information estimate. The pseudo-code is available in Alg. 4, and a visual explanation is shown in Fig.4.4.

4.1.5 Experiments

Evaluation details:

Here, we evaluate the ability of SepVAE to separate common from target-specific patterns on three medical and one natural (CelebA) imaging datasets. We compare it with the only SOTA



$$\phi_s, \phi_c \leftarrow \text{argmin}_{\phi_s, \phi_c} \sum_{i=1}^{N} \text{ReLU}\left(\log\left(\frac{D_\lambda([c,s])}{1 - D_\lambda([c,s])}\right)\right)$$

Figure 4.4: Illustration of Mutual Information loss between the common and the salient space. Given two images x_a and x_b, 4 sets of latents are computed: c_a and s_a latents of the image a, c_b and s_b latents of the image b. A non-linear MLP is independently trained with a binary cross-entropy loss to classify shuffled concatenations (i.e., from different images) with the label 0 and concatenations of latents coming from the same image with the label 1. Then, during training, encoders should not to be able to identify whether a concatenation of latents belong to class 0 (shuffled common and salient spaces) or class 1 (common and salient spaces coming from the same image). We encourage that by minimizing $D_{KL}(p_{\phi_s,\phi_c}(c,s)||p_{\phi_c}(c) \times p_{\phi_s}(s))$.

CA-VAE methods whose code is available: MM-cVAE [299] and ConVAE [2] [3].

For quantitative evaluation, we use the fact that the information about attributes, clinical variables, or subtypes (e.g. glasses/hats in CelebA) should be present either in the common or in the salient space. Once the encoders/decoder are trained, we evaluate the quality of the representations in two steps. First, we train a Logistic (resp. Linear) Regression on the estimated salient and common factors of the training set to predict the attribute presence (resp. attribute value). Then, we evaluate the classification/regression model on the salient and common factors estimated from a test set. By evaluating the performance of the model, we can understand whether the information about the attributes/variables/subtype has been put in the common or salient latent space by the method. Furthermore, we report the background (BG) vs target (TG) classification accuracy. To do so, a 2 layers MLPs is independently trained, except for SepVAE, where salient space predictions are directly estimated by the classifier.

In all Tables, for categorical variables, we compute (Balanced) Accuracy scores (=(B-)ACC), or Area-under Curve scores (=AUC) if the target is binary. For continuous variables, we

[2]ConVAE implemented with correct Mutual Information minimization, *i.e.:* with independently trained discriminator.

Algorithm 4 Minimizing the Mutual Information between common and salient spaces.

Input: $X \in \mathbf{R}^{B \times C \times W \times H}$

For t in epochs :

Discriminator training :

Sample $z = [c, s]$ from q_{ϕ_c, ϕ_s}.

Sample $\bar{z} = [c, \bar{s}]$ from $q_{\phi_c} \times q_{\phi_s}$ by shuffling s along the batch dimension.

Compute $\mathcal{L}_{BCE} = -\log(D(z)) - \log(1 - D(\bar{z}))$

Freeze ϕ_c and ϕ_s. Update D parameters only.

Encoders training :

Sample $z = [e_{\phi_c}(x), e_{\phi_s}(x)]$ from q_{ϕ_c, ϕ_s}.

Compute $\mathcal{L}_{MI} = \sum_{i=1}^{B} \text{ReLU}\left(\log \frac{D(z_i)}{1 - D(z_i)} \right)$

Freeze D parameters. Update ϕ_c and ϕ_s.

EndFor

use Mean Average Error (=MAE). Best results are highlighted in bold, second best results are underlined. For CelebA and Pneumonia experiments, mean, and standard deviations are computed on the results of 5 different runs in order to account for model initializations. For neuro-psychiatric experiments, mean and standard deviations are computed using a 5-fold cross-validation evaluation scheme.

Qualitatively, the model can be evaluated by looking at the full image reconstruction (common+salient factors) and by fixing the salient factors to s' for target images. Comparing full reconstructions with common-only reconstructions allows the user to interpret the patterns encoded in the salient factors s (see Fig.4.2 and Fig.4.5).

CelebA - glasses or hats identification:

To compare with [299], we evaluated our performances on the CelebA with attributes dataset. It contains two sets, target and background, from a subset of CelebA [174], one with images of celebrities wearing glasses or hats (target) and the other with images of celebrities not wearing these accessories (background), see Sec.C.2.1 in the Appendix for more details. The discriminative information allowing the classification of glasses vs. hats should only be present in the salient latent space. We demonstrate that we successfully encode these attributes in the salient space with quantitative results in Tab. 4.1, and with reconstruction results in Fig. 4.5. Furthermore, in Fig. 4.6, we show that we effectively minimize the background dataset variance

Figure 4.5: SepVAE qualitative example on the CelebA with accessories dataset (BG = no accessories, TG = hats and glasses). (Middle, common+salient): Full reconstructions using the estimated common and salient factors. (Right, common only): Reconstruction using only the estimated common factors fixing the salient to s'. The salient latent variables capture the accessories (hats and glasses), which are target-specific patterns. The common latents capture the common attributes (e.g., identity, skin color).

Table 4.1: CA-VAE methods performance on CelebA with accessories dataset. Accessories (glasses/hat) information should only be present in the salient space, not in the common.

	Glss/Hats Acc salient ↑	Glss/Hats Acc common ↓	Bg vs Tg AUC salient ↑	Bg vs Tg AUC common ↓
ConVAE	82.32±1.17	75.01±2.526	82.46±0.586	78.39±0.41
MM-cVAE	85.17±0.60	73.938±1.66	88.536±0.39	78.036±0.35
SepVAE	**87.62±0.75**	**72.16±2.02**	**93.15±1.65**	**77.604±0.20**

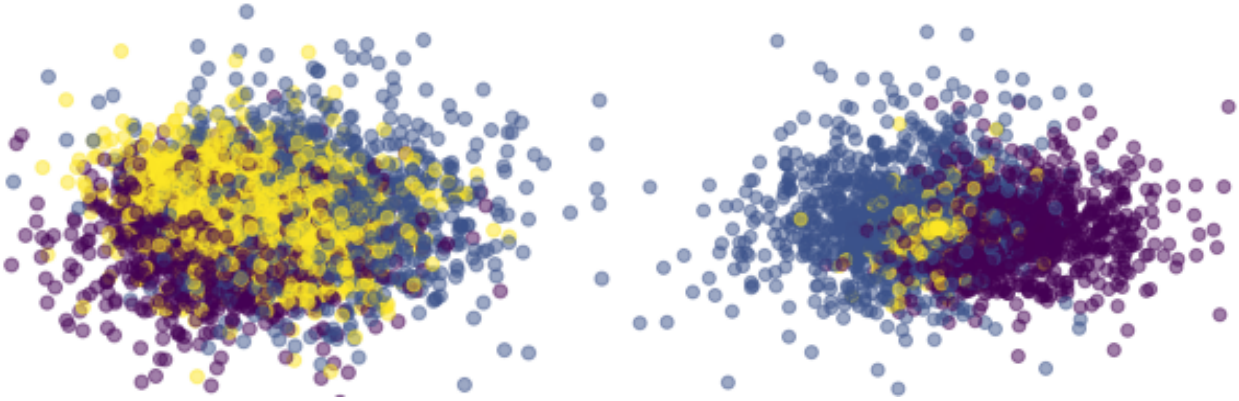


Figure 4.6: PCA projections of MM-c-VAE (left) and SepVAE (right) salient space on CelebA TEST set. Yellow: no accessories. Dark Blue: glasses. Purple: hats. We can clearly observe that our method maximizes the target variance while reducing the background variance. We attribute this different behavior to our salient classification loss, which reduces the overlap between background and target salient distributions.

in the salient space compared to MM-cVAE[3].

Identify pneumonia subgroups:

As in Chap.3.2, we gathered 1342 healthy X-ray radiographies (*background* dataset), and 2684 radiographies of pneumonia radiographies (*target* dataset) from [146]. Two different sub-types of pneumonia constitute this set, viral (1342 samples) and bacterial (1342 samples), see Sec.B.5.3. In Tab. 4.2, we demonstrate that our method is able to produce a salient space that captures the pathological variability as it allows distinguishing the two subtypes: viral and bacterial pneumonia.

Ablation study: In the lower part of Tab. 4.2, we propose to disable different components of

[3]Our evaluation process is different from [299] as their TEST set has been used during the model training. Indeed, the TRAIN / TEST split used for training Logistic Regression is performed after the model fitting on the set TRAIN+TEST set. Besides, we were not able to reproduce their results.

Table 4.2: CA-VAE methods performance on the Healthy vs Pneumonia X-Ray dataset. Accuracy scores are obtained with linear probes fitted on common c or salient s latent vectors of the images of the target dataset. Pneumonia subtypes information should only be present in the salient space. The lower part shows an ablation study of regularization losses.

	Subgrp Acc salient ↑	Subgrp Acc common ↓	Bg vs Tg Acc salient ↑	Bg vs Tg Acc common ↓
ConVAE	82.30±1.53	73.58±1.84	67.80±5.93	58.05±7.17
MM-cVAE	82.86±1.87	74.35±3.19	70.44±2.69	59.94±5.88
SepVAE	**84.78±0.42**	**70.92±1.39**	**78.13±3.03**	57.52±4.14
SepVAE no **MI**	84.10±0.48	71.792±2.94	75.186±5.69	60.35±4.73
SepVAE no **CLSF**	84.71±1.19	73.58±2.19	71.91±4.65	**55.79±5.41**
SepVAE no **REG**	83.98±0.85	72.61±2.05	73.03±2.97	61.43±2.25

the model to show that the full model SepVAE is always better on average. no **MI** means that we disabled the Mutual Information minimization loss (no Mutual Information Minimization). no **CLSF** means that we disabled the classification loss on the salient space (no Salient Classification). no **REG** means that we disabled the regularization loss that forces the background samples to align with an informationless vector $\mathbf{s'} = 0$ (no Salient Prior).

Table 4.3: CA-VAE methods performance on the prediction of schizophrenia-specific variables (SANS, SAPS, Diag) and common variables (Age, Sex, Site) using only salient factors reconstructed by test images of the target (MD) dataset.

	Age MAE ↑	Sex B-Acc ↓	Site B-Acc ↓	SANS MAE ↓	SAPS MAE ↓	Diag AUC ↑
ConVAE	7.46±0.18	72.72±1.32	54.46±2.46	**3.95±0.28**	2.76±0.18	58.53±4.87
MM-cVAE	7.10±0.34	**72.15±2.47**	56.69±9.84	4.52±0.33	3.16±0.05	70.94±4.08
SepVAE	**7.98±0.25**	72.61±2.19	**44.10±5.78**	4.14±0.39	**2.60±0.27**	**79.15±3.39**

Table 4.4: CA-VAE methods performance on the prediction of autism-specific variables (ADOS [10], ADI-social, Diag) and common variables (Age, Sex, Site) using only salient factors reconstructed by test images of the target (MD) dataset.

	Age MAE ↑	Sex B-Acc ↓	Site B-Acc ↓	ADOS MAE ↓	ADI-s MAE ↓	Diag AUC ↑
ConVAE	3.97±0.19	66.67±1.12	40.97±2.06	10.1±1.27	5.14±0.17	54.93±2.04
MM-cVAE	3.74±0.12	64.07±2.58	40.93±2.66	10.5±2.47	5.09±0.16	54.88±2.76
SepVAE	**4.38±0.09**	**59.61±1.78**	**33.58±1.86**	**8.55±1.68**	**4.91±0.17**	**59.73±1.78**

Parsing neuro-anatomical variability in psychiatric diseases:

The task of identifying consistent correlations between neuro-anatomical biomarkers and observed symptoms in psychiatric diseases is important for developing more precise treatment options. Separating the different latent mechanisms that drive neuro-anatomical variability in psychiatric disorders is a challenging task. Contrastive Analysis (CA) methods such as ours have the potential to identify and separate healthy from pathological neuro-anatomical patterns in structural MRIs. This ability could be a key component to push forward the understanding of the mechanisms that underlie the development of psychiatric diseases.
Given a background population of Healthy Controls (HC) and a target population suffering from a Mental Disorder (MD), the objective is to capture the pathological factors of variability in the salient space, such as psychiatric and cognitive clinical scores, while isolating the patterns related to demographic variables, such as age and sex, or acquisition sites to the common space. For each experiment, we gather T1w anatomical VBM [17] pre-processed images resized to $128 \times 128 \times 128$ of HC and MD subjects. We divide them into 5 TRAIN, VAL splits (0.75, 0.25) and evaluate in a cross-validation scheme the performance of SepVAE and the other SOTA CA-VAE methods. Please note that this is a challenging problem, especially due to the high dimensionality of the input and the scarcity of the data. Notably, the measures of psychiatric and cognitive clinical scores are only available for some patients, making it scarce and precious information.

Schizophrenia: We merged images of schizophrenic patients (TG) and healthy controls (BG) from the datasets SCHIZCONNECT-VIP [292] and BSNIP [271]. Results in Tab. 4.7 show that the salient factors estimated using our method better predict schizophrenia-specific variables of interest: SAPS (Scale of Positive Symptoms), SANS (Scale of Negative Symptoms), and diagnosis. On the other hand, salient features are shown to be poorly predictive of demographic variables: age, sex, and acquisition site. It paves the way toward a better understanding of schizophrenia disorder by capturing neuro-anatomical patterns that are predictive of the psychiatric scales while not being biased by confounding variables.

Autism: Second, we combine patients with autism from ABIDE1 [67, 68] and ABIDE2 [119] (TG) with healthy controls (BG). In Tab. 4.4, SepVAE's salient latents better predict the diagnosis and the clinical variables, such as ADOS (Autism Diagnosis Observation Schedule) and ADI Social (Autism Diagnosis Interview Social) which quantifies the social interaction abilities. On the other hand, salient latents poorly infer irrelevant demographic variables (age, sex, and acquisition site), which is a desirable feature for unbiased diagnosis tools.

4.1.6 Reproducing the results of Aglinskas et al, 2022

In 2022 Aglinskas et al. [6] applied a Contrastive Analysis VAE entitled cVAE to parse the heterogeneity of a cohort with autism disorder. In their original contribution, they show that their common vectors correlate with shared demographic variables such as age, sex, and scanner type, but not with clinical scale measures. They also show that their salient vectors correlate with clinical scale measures such as ADOS, DSM IV, and Vineland. Interestingly, the IQ variable is expected to correlate with the shared latent space (as there is a natural IQ variability in the healthy population, a Gaussian distribution around a score of 100 and with a standard deviation of 15), but also with the pathological space (as there exist a high IQ variability within the autism spectrum, with heterogeneous autism subtypes ranging from Asperger syndrome, Low-Functioning Autism (LFA) to Classic Autism (CA)). In practice, our attempts to reproduce their results were unsuccessful. We pinpoint our results compared to theirs in Fig. 4.7. For a fair comparison, we trained the traditional VAE with $\beta = 1$, while the original code trained it with $\beta = \frac{1}{64}$ [4]. Interestingly, we found that the VAE's latents correlate with demographic (age, sex, IQ, and scanner type) and clinical variables (notably, IQ, DSM IV, and Vineland) but, by design, does not successfully separate shared from pathological patterns.

Given their open-source code and processed datasets, we observed highly variable results depending on the initialization of the models, which suggests a high epistemic error with such a model. All the more so these results were computed within the train set, as proposed in the original contribution, even though a cross-validation scheme would have been expected [5]. This result remains informative on the potential pitfalls and drawbacks of this Contrastive Analysis strategy. Thus, we only compared our method with CA-VAEs enhanced with thoughtful strategies aiming at separating common from target-specific latent generation factors and reducing the epistemic variability of Contrastive Analysis VAEs. Still, the contributions introduced by Aglinskas et al [6] remain particularly valuable and have provided novel insights and methods when parsing the heterogeneity within psychiatric diseases. They provided a relevant methodology to evaluate Contrastive Analysis VAE in the context of psychiatric disorder heterogeneity parsing and a thoughtful explainability strategy by introducing the concept of a digital healthy avatar, as shown in Fig 4.2 aiming at generating a reconstruction of a diseased brain without the pathological patterns.

[4] In practice β is set to 1 in their code, but when computing the Mean Square Error term, the authors renormalize by the height (64) and width (64) of the image, while we would also expect the depth (64). This is thus equivalent to having $\beta = \frac{1}{64}$

[5] Unfortunately, we were not able to attempt to reproduce their results on their independent TEST set SFARI as we could not access the database.

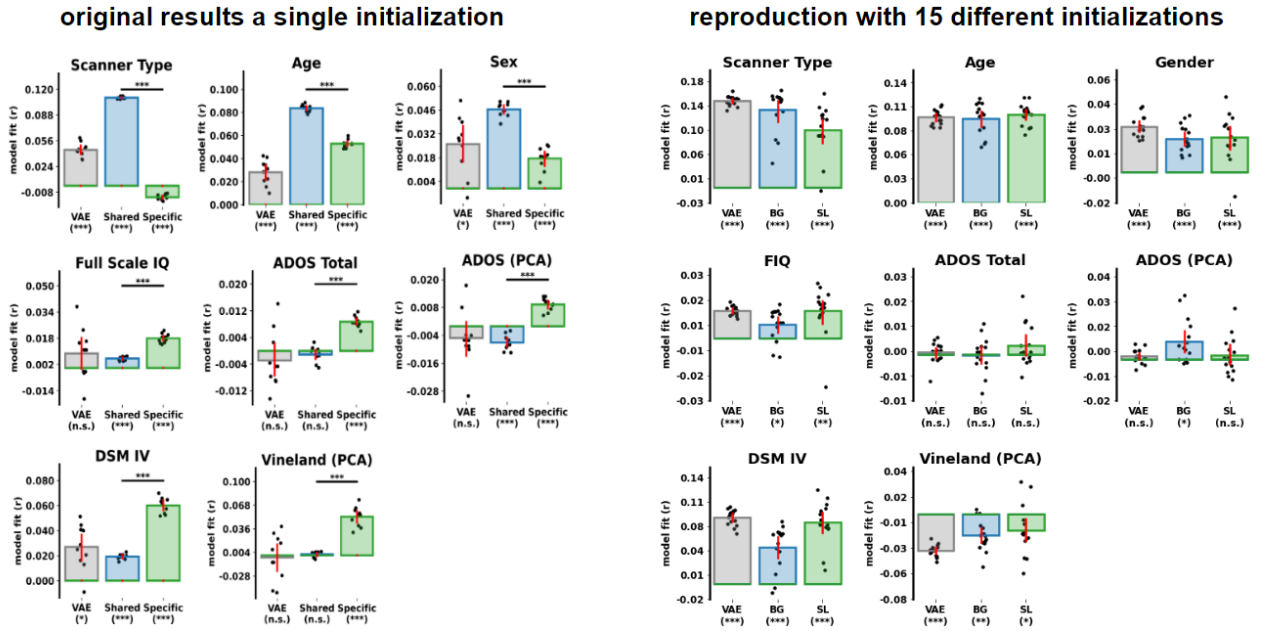


Figure 4.7: Comparison between the original performances reported in Aglinskas et al. [6] and ours. (left) reported RSA measures in the original contribution (standard errors obtained from samples drawn from the latent space distributions). (right) measures from our reproduction attempt (standard errors obtained from 15 different model initializations). The "VAE" bars compute the RSA (Representation Similarity Analysis) between the VAE's latents with a demographic, cognitive, or clinical variable, the higher its absolute value is, the higher the correlation is. The "Shared" and "BG" bars compute the RSA between the CVAE's shared (or common) latents with a variable of interest. The "Specific" and "SL" bars compute the RSA between the CVAE's salient (or pathological) latents with a variable of interest.

4.1.7 Conclusions and Perspectives

In this paper, we developed a novel CA-VAE method entitled SepVAE. Building onto Contrastive Analysis methods, we first criticize previously proposed regularizations about (1) the matching of target and background distributions in the common space and (2) the overlapping of target and background priors in the salient space. These regularizations may fail to prevent information leakage between common and salient spaces, especially when datasets are biased. We thus propose two alternative solutions: salient discrimination between target and background samples, and mutual information minimization between common and salient spaces. We integrate these losses along with the maximization of the ELBO of the joint log-likelihood. We demonstrate superior performances on radiological and two neuro-psychiatric applications, where we successfully separate the pathological information of interest (diagnosis, pathological scores) from the "nuisance" common variations (e.g., age, site). The development of methods

like ours seems very promising and offers a large spectrum of perspectives. For example, it could be further extended to multiple target datasets (*e.g.,* healthy population Vs several pathologies, to obtain a continuum healthy - mild - severe pathology) and to other models, such as GANs, for improved generation quality. Eventually, to be entirely trustworthy, the model must be identifiable, namely, we need to know the conditions that allow us to learn the correct joint distribution over observed and latent variables. We plan to follow [147, 286] to obtain theoretic guarantees of the identifiability of our model.

4.2 SepCLR: Separating common from salient patterns with Contrastive Learning

4.2.1 Abstract

Contrastive Analysis is a sub-field of Representation Learning that aims at separating common factors of variation between two datasets, a background (i.e., healthy subjects) and a target (i.e., diseased subjects), from the salient factors of variation, only present in the target dataset. Despite their relevance, current models based on Variational Auto-Encoders have shown poor performance in learning semantically-expressive representations. On the other hand, Contrastive Representation Learning has shown tremendous performance leaps in various applications (classification, clustering, etc.). In this work, we propose to leverage the ability of Contrastive Learning to learn semantically expressive representations well adapted for Contrastive Analysis. We reformulate it under the lens of the InfoMax Principle and identify two Mutual Information terms to maximize and one to minimize. We decompose the first two terms into an Alignment and a Uniformity term, as commonly done in Contrastive Learning. Then, we motivate a novel Mutual Information minimization strategy to prevent information leakage between common and salient distributions. We validate our method, called SepCLR, on three visual datasets and three medical datasets, specifically conceived to assess the pattern separation capability in Contrastive Analysis. Code available at https://github.com/neurospin-projects/2024_rlouiset_sep_clr

4.2.2 Introduction

In Representation Learning, practitioners estimate parametric models tailored to learn meaningful and compact representations from high-dimensional data. The objective is to capture relevant features to facilitate downstream tasks such as classification, clustering, segmentation, or generation. Contrastive Representation Learning (CL) has made remarkable progress in learning representations that encode high-level semantic information about inputs such as images ([318, 297, 21, 118, 108, 77, 24]) and sequential data ([213, 155, 274, 252, 267]). With a distinct perspective, Contrastive Analysis (CA) approaches aim to discover the underlying generative factors that 1) distinguish a target dataset from a background dataset (*i.e.,* salient factors) and that 2) are shared between them (*i.e.,* common factors). It is usually assumed that target samples comprise additional (or modified) patterns compared to background samples ([4, 327, 328, 254, 240, 100, 169, 329]). The ability to *distinguish* and *separate* common

from salient generative factors is crucial in various domains. For instance, in medical imaging, researchers seek to identify pathological patterns in a population of patients (target) compared to healthy controls (background) [12, 6]. Contrastive Analysis also concerns other domains like drug research (medicated vs. placebo populations), surgery (pre-intervention vs. post-intervention groups), time series (signal vs. signal-free samples), biology and genetics (control vs. characteristic-trait population, [140]).

Current Contrastive Analysis methods are based on VAEs (Variational Auto-Encoders) [150]. This choice is particularly suitable for generation and image-level manipulations. However, as shown in [224], VAE can fail to learn meaningful latent representations, or even learn trivial representations when the decoder is too powerful [47]. Conversely, Contrastive Learning (CL) methods have demonstrated outstanding results in many domains, such as unsupervised learning [50], deep clustering [168], content vs style identification [286], background debiasing [295, 65], and multi-modality [315]. This performance gap might be explained by the following reasons. CL methods produce representations invariant to a set of user-defined image transformations (translation, zoom, color jittering, etc.), whereas VAEs are highly sensitive to these uninteresting variability factors. Furthermore, VAEs maximize the log-likelihood, which is only a function of the marginal distribution of the input data and *not* of the latent representations. Differently, CL methods based on the InfoNCE loss implicitly maximize the Mutual Information (MI) between input data and latent features[6]. From a representation learning point of view, this makes much sense since the MI depends on the joint distribution between input data and representation [330]. Inspired by these works, we propose to reformulate the Contrastive Analysis problem under the lens of the well-known InfoMax principle [31, 123] and leverage the representation power of Contrastive Learning (CL) to estimate the MI terms of our newly proposed Contrastive Analysis setting. We seek to separate the *salient patterns* of the target dataset from the *shared (common) patterns* with the background dataset. Common factors c should be representative of both target and background datasets (respectively y and x). Thus, we propose to maximize the MI between x (resp. y) and c. We compute the Entropy and Alignment to estimate the MI, as in [293] and [234]. Since salient factors s should only describe patterns typical of the target data y, we propose to maximize the MI between s and **only** y. Furthermore, we also add the constraint that background samples' representations should always be equal to an informationless vector s' in the salient space. This objective is close to other recent CA ideas, such as in Contrastive PCA [3] and CA-VAEs, but also to Super-

[6]as shown in Sec.4.2.4, the MI between the latent representation of two views, maximized in many recent methods, is a lower bound of the MI between input data and latent representations.

vised Anomaly Detection intuitions, such as in DeepSAD [239], where the entropy of anomalies (eq. target) is maximized, whereas normal samples (eq. backgrounds) are set to a constant vector. We propose an extension of this salient term when fine-grained target attributes are available and propose disentangling these attributes within the salient space in a supervised manner. Moreover, to avoid information leakage between the common c and salient space s, we constrain the MI to be (exactly) equal to 0. This choice avoids undesirable results since minimizing the MI may bring to a trivial solution where c and/or s would contain no information. Instead, we propose a method to estimate and maximize their joint entropy $H(c, s)$ without requiring any assumptions about the form of its pdf nor a neural network-based approximation. Our contributions are summarized below:

1. We introduce SepCLR, a novel theoretical framework for Contrastive Analysis based on the InfoMax principle. We identify three Mutual Information terms: a common space term, a salient space term, and a common-salient independence term.

2. We leverage Contrastive Learning to estimate the common and salient terms. We show how usual contrastive losses such as InfoNCE and SupCLR can be retrieved from the InfoMax Principle. Likewise, we derive a novel contrastive method to capture target-specific variability while canceling background variability in the salient space.

3. To reduce the information leakage between the common and salient spaces, we suggest a strategy that overcomes the pitfalls of usual Mutual Information Minimization methods.

4.2.3 Related Works

Our work relates to contrastive learning, mutual information, and contrastive analysis.

Contrastive Learning and the InfoMax Principle. Contrastive Learning (CL) hinges on an intuition that dates back to [28]. Given an input sample x (image or sequence) and two different views (*i.e.,* transformations) v and v^+ of x that potentially overlap (spatially or sequentially), CL is based on the assumption that v and v^+ should share a similar information content. A parametric encoder f_θ is then estimated by maximizing their "agreement" in the representation space so that their similarity/dependence is preserved in the embeddings $f_\theta(v)$ and $f_\theta(v^+)$. A commonly used measure of agreement is the Mutual Information between the two views embeddings that is maximized: $\theta^* \leftarrow \text{argmax} I(f_\theta(v); f_\theta(v^+))$, where the choice of f_θ imposes some structural constraints (*i.e.,* inductive bias). As shown in [276], this objective can be seen as a lower bound on the InfoMax principle $\max_\theta I(x; f_\theta(x))$ ([173], [31]). Many

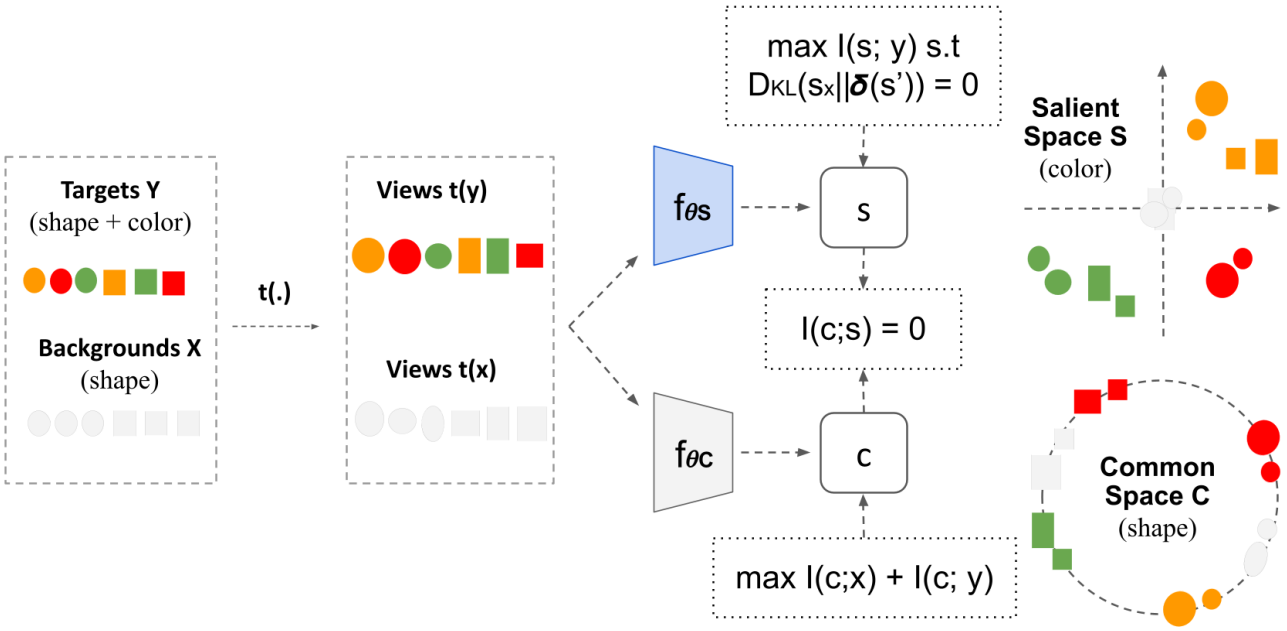


Figure 4.8: SepCLR is trained to identify and separate the salient patterns (color variations) of the target dataset Y from the common patterns (shape) shared between background X and target dataset Y. Views (transformations $t(\cdot)$) of both datasets are fed to two different encoders, one for the salient space (f_{θ_s}) and one for the common space (f_{θ_c}). In the hyperspherical common space, C, embeddings of views of the same image (from both X and Y) are aligned, while embeddings from different images are repelled (max $I(c;x) + I(c;y)$). This enforces C to represent the shared patterns (shape). In the salient space S, which is a Euclidean space, in order not to capture background variability (*i.e:* shape), background embeddings are aligned onto an information-less null vector **s'** ($D_{KL}(s_x||\delta(s')) = 0$). Furthermore, embeddings of views of the same image (only from Y) are aligned while embeddings from different images are pushed away from each other, and they are all repelled from **s'** (max $I(s;y)$). This enforces S to capture only the salient patterns of Y (color). To limit the information leakage between C and S, their MI is constrained to be null, *i.e*: $I(c;s=0)$.

approaches ([155, 274, 21, 213, 275, 24]) propose to maximize $I(f_\theta(v); f_\theta(v^+))$ rather than the original InfoMax objective since the embeddings $f(x)$ have a lower dimension than the original samples x and the choice of the transformation for the views gives more flexibility. [293] simplifies the usual CL loss InfoNCE [50] into an alignment (or reconstruction) and a uniformity (or entropy) term. While the alignment term trains the encoder to assign similar representations to positive views, the uniformity term encourages feature distribution to preserve maximal information *i.e.:* maximal entropy. Recently, [234] demonstrated that these terms could be derived from the maximization of $I(f_\theta(V); f_\theta(V^+))$ and that several clustering methods could be retrieved from this formulation. We build onto these works to introduce the CL framework required to develop the proposed CA losses.

Contrastive Analysis. Contrastive Analysis (CA) methods are designed to separate salient latent variables (*i.e:* patterns that are specific to the target dataset) from common latent variables (*i.e:* patterns that are shared between background and target datasets). Recently, contrastive Variational Auto-Encoders were designed to capture higher-level semantics [3, 328, 299, 180]. These methods usually rely on a latent space split into two parts, common and salient, estimated by two different encoders. To limit information leakage between common and salient spaces, three types of regularization have been proposed. First, a usual solution is to introduce an explicit regularization on the salient encoder to minimize the background information expressivity [3, 299, 180, 328, 329]. This regularization forces the salient vectors of the backgrounds to be close to **s'** (information-less vector, often equal to 0). A second idea, proposed in MM-cVAE [299], is to match the common space distributions of the target and background samples by minimizing their Maximum Mean Discrepancy (MMD) [110]. This regularization reduces the information that would enable discriminating targets from background samples within the common space. In cVAE [3], and SepVAE [180], authors minimize the Mutual Information between the common and salient spaces.

Contrastive Analysis is not disentanglement nor style vs. content separation. CA is not about disentanglement, which aims to isolate independent variation factors in a single dataset [176, 48, 257]. In contrast, CA seeks to separate common from target-specific generative factors without requiring the isolation of independent factors usually defined in a supervised manner using external attributes. Furthermore, CA is not about separating style from content ([144, 286]), where content is usually defined as the invariant part of the latent space, namely the part shared across different views. In contrast, style refers to the varying part that accounts for the differences between views. Content and style depend on the chosen semantic-invariant transformations, and they are defined for a single dataset. In CA, we do not necessarily need transformations or views, and we jointly analyze two different datasets.

Mutual Information Minimization. Mutual Information minimization has gained significant attention in diverse applications such as disentangling [152, 257], domain adaptation [102], style/content identification [144], and Information Bottleneck compression [11]. Typically, it can serve as a regularizer to diminish the dependence between variables. However, computing the value of Mutual Information is hardly possible in cases where closed forms of density functions, joint or marginal, are unknown. In most machine learning setups, access is limited to only samples drawn from the joint distribution. To accommodate, most estimation methods (lower bound, upper bound, and reliable estimators) focus on sample-based estimation. However, most of these works either require strong assumptions about one of the distributions (*e.g.,*

its form) ([11, 226]) or the introduction of an independent neural network to approximate a distribution in a sample-based variational manner. For instance, CLUB [53] derives an upper bound of the Mutual Information $I(X, Y)$ by either assuming the closed-form of $p(y|x)$ or, in its variational form, estimating it with a parameterized neural network $q_\theta(y|x)$. Another example concerns Total Correlation methods [180, 152] that leverage the Density Ratio trick [265, 211] to estimate the density ratio between the joint distribution and the product of the marginals. This technique demands optimizing an independent discriminator to discriminate samples drawn from the joint distribution from those drawn from the product of the marginals.

4.2.4 The InfoMax principle for Contrastive Analysis

Let $X = \{x_i\}_{i=1}^{N_X}$ and $Y = \{y_j\}_{j=1}^{N_Y}$ be the background and target data-sets of images respectively. As it is commonly done in Contrastive Analysis [3, 299, 180], we suppose that both x_i and y_j are drawn i.i.d. from the *same* conditional distribution $p_\theta(\cdot|c, s)$, that is parameterized by unknown parameters θ and that depends on two latent variables: the **common** generative factors $c \in \mathbf{R}^{D_c}$, shared between X and Y, and the **salient** (or **target-specific**) generative factors $s \in \mathbf{R}^{D_s}$, which are only present in Y and not in X. The separation between c and s can be considered a weakly supervised learning problem since the only level of supervision is the population-based label X or Y. The user has no knowledge about the common and salient generative factors at training (or test) time. By grounding our method on the InfoMax principle [31, 123], and since we want the common factors c to be representative of both datasets, we propose to maximize the mutual information I between c and both datasets X and Y. Similarly, we propose maximizing the mutual information between the salient factors s and **only** the target samples Y. Since we want the background samples x to be fully encoded by c, we enforce the salient factors s of x to be always equal to a constant value s' (*i.e.,* no information): $x_i \sim p_\theta(x|c_i, s_i = s')$. Mathematically, we do that by minimizing the Kullback–Leibler divergence D_{KL} between $p(s|x)$ and $\delta(s')$, a Dirac Delta distribution centered at s'. Furthermore, to enforce the separation (*i.e.,* independence) between c and s, we also propose to use $I(c, s) = 0$ as a regularization constraint.

Our objective is to *separate* and *infer* the common c and salient s factors given the input data X and Y. We use two probabilistic encoders, f_{θ_c} and f_{θ_s}, parameterised by θ_c and θ_s, to approximate the conditional distributions $p(c|\cdot)$ and $p(s|\cdot)$ respectively. The two encoders are shared between X and Y. Furthermore, as commonly done in recent representation learning papers, we assume to have multiple views v of each image x (or y) generated via a stochastic augmentation function t: $v = t(\cdot)$. By denoting $c = f_{\theta_c}(v)$, $s = f_{\theta_s}(v)$, $s_x = f_{\theta_s}(t(x))$, our goal

is to find the optimal parameters $\theta^* = \{\theta_c^*, \theta_s^*\}$ that maximize the following cost function:

$$\text{argmax}_\theta \underbrace{\lambda_C(I(x;c) + I(y;c))}_{\text{Common InfoMax}} + \underbrace{\lambda_S I(y;s)}_{\text{Salient InfoMax}} \quad \text{s.t.} \quad \underbrace{D_{KL}(s_x||\delta(s')) = 0}_{\text{Information-less hyp.}} \text{ and } \underbrace{I(c,s) = 0}_{\text{Independence hyp.}} \tag{4.5}$$

In Sec. 4.2.4, we show how to estimate the common terms, $I(x;c)$ and $I(y;c)$, via a formulation similar to the alignment and entropy terms introduced in [293]. In Sec. 4.2.4, we take into account the information-less hypothesis (*i.e.* background embeddings should always be equal to an information-less vector in the salient space) to estimate the salient term $I(y;s)$. Ultimately, in Sec. 4.2.4, we propose a strategy to enforce the independence hypothesis *i.e.* $I(c;s) = 0$, that prevents information leakage between the common and salient space.

Retrieve InfoNCE from InfoMax for common space

In this section, we demonstrate that $I(x;c)$ and $I(y;c)$ can be estimated via the multi-view alignment and uniformity losses inspired by [293]. Full derivation can be found in Appendix Sec. D.1. Let f_{θ_C} be the common encoder and $c \sim f_{\theta_C}(t(.))$ be the common representations. The MI $I(x;c)$ (same reasoning is also valid for $I(y;c)$) can be decomposed into:

$$I(x;c) = \underbrace{-\mathbf{E}_{x\sim p_x} H(c|x)}_{\text{Alignment}} + \underbrace{H(c)}_{\text{Entropy}} \tag{4.6}$$

Entropy (Uniformity). As in [293], the entropy can be computed with a non-parametric estimator described in [8]. To do so, we compute the approximate density function $\hat{p}(c_i)$ with a Kernel Density Estimator as in [217, 237], based on views v_j (random augmentation of an image with index j) uniformly sampled from both the target dataset $f_{\theta_c}(t(y)) \sim p(c|y)$ and the background dataset $f_{\theta_c}(t(x)) \sim p(c|x)$. We choose a Gaussian kernel with constant standard deviation τ, which results in an L2 distance between the views. However, in practice, we constrain the outputs $f_{\theta_C}(.)$ to be unit-normed, which is equivalent to directly choosing a von Mises-Fischer kernel with concentration parameter $\frac{1}{\tau}$. [7] As in [293], we optimize a lower bound

[7] Intuitively, if $||f_{\theta_C}(.)||_2 = 1$, the L2 distance between two representations can be simplified into a negative dot-product: $||f_{\theta_C}(v_i) - f_{\theta_C}(v_j)||_2 = 2 - 2 f_{\theta_C}(v_i)^T . f_{\theta_C}(v_j)$. Full proof in Appendix Sec .D.1.2.

of this estimator in practice, called $-\mathcal{L}_{\text{uniform}}$:

$$\mathcal{L}_{\text{uniform}} = \log \frac{1}{N_X+N_Y} \sum_{i=1}^{N_X+N_Y} \frac{1}{N_X+N_Y} \sum_{j=1}^{N_X+N_Y} \exp \frac{-||f_{\theta_C}(v_i) - f_{\theta_C}(v_j)||_2^2}{2\tau} + \underbrace{\log(\sqrt{2\pi\tau})}_{\text{Constant term}} \tag{4.7}$$

Alignment: Differently from [293], we propose to estimate the conditional entropy $-H(c|x)$ with a re-substitution entropy estimator. We compute the approximate density function $\hat{p}(c_i|x_i)$ with a Kernel Density Estimator based on samples uniformly drawn from the conditional distribution $c_i^k \sim p(c|x_i)$, where $c_i^k = f_\theta(v_i^k)$ and v_i^k are K views obtained via the stochastic process $t(.)$. As for the entropy term, we choose a Gaussian kernel with constant standard deviation τ to derive an L2 distance between the views. Our formulation generalizes [293], as we directly retrieve a multi-view alignment term between K positive views of the same image and not a single-view alignment as in [293]. However, to reduce the computational burden in practice, we also choose a single view K=1, as in [293]. Combining the background alignment $-H(c|x)$ and the target alignment $-H(c|y)$, we obtain:

$$\mathcal{L}_{\text{align}} = -\frac{1}{N_X+N_Y} \sum_{i=1}^{N_X+N_Y} \log \frac{1}{K} \sum_{k=1}^{K} \exp \frac{-||f_{\theta_C}(v_i) - f_{\theta_C}(v_i^k)||_2^2}{2\tau} + \underbrace{\log(\sqrt{2\pi\tau})}_{\text{Constant term}} \tag{4.8}$$

On the relation with $\mathbf{I}(\mathbf{f}_\theta(\mathbf{v}), \mathbf{f}_\theta(\mathbf{v}'))$. Many recent representation learning works ([51, 293]) maximize the MI between two views v and v' of x: $I(f_\theta(v), f_\theta(v')$. Inspired by the InfoMax principle, we propose instead maximizing $I(f_\theta(v), x)$. As shown in [276], by directly applying the *data processing inequality*, one can demonstrate that $I(f_\theta(v), f_\theta(v')$ is a lower bound of $I(f_\theta(v), x)$.

Derive the Background-Contrasting Alignment and Uniformity terms

In this section, we consider the maximization of the salient term $I(s;y)$, which is decomposed into an alignment and uniformity term as before, constrained by the information-less hypothesis:

$$\{textargmax I(s;y) = - \underbrace{\mathbf{E}_{y\sim p_y} H(s|y)}_{\text{Target-only Alignment}} + \underbrace{H(s)}_{s'\text{-Entropy}} \quad \text{s.t.} \quad \underbrace{D_{KL}(s_x||\delta(s')) = 0}_{\text{Information-less hyp.}} \tag{4.9}$$

Target-only alignment. To estimate the target samples' alignment term, we use the same estimation method as in 4.2.4. Namely, we derive an alignment term between two views $v_i = t(y_i)$ and $v_i^k = t(y_i^k)$ of the same target image y_i. As in Sec. 4.2.4, we use a re-substitution

estimator with a Gaussian Kernel Density Estimation with constant standard deviation τ and $K = 1$ in practice.

$$\mathcal{L}_{\text{y-align}} = -\frac{1}{N_Y}\sum_{i=1}^{N_Y}\log\frac{1}{K}\sum_{k=1}^{K}\exp\frac{-||f_{\theta_C}(v_i) - f_{\theta_C}(v_i^k)||_2^2}{2\tau} + \underbrace{\log(\sqrt{2\pi\tau})}_{\text{Constant term}} \tag{4.10}$$

s'-Uniformity. Concerning the Entropy term, as in Sec. 4.2.4, we propose to develop the salient entropy with a re-substitution entropy estimator. Again, we use a lower bound of $\hat{H}(S)$ called $-\mathcal{L}_{s'\text{-unif}}$. Then, we estimate the density $\hat{p}(s)$ with a Gaussian Kernel Density Estimator based on samples uniformly drawn from the target dataset $f_{\theta_S}(t(y)) \sim p(s|y)$ and from the background dataset $f_{\theta_S}(t(x)) \sim p(s|x)$. Importantly, the information-less hypothesis constrains the salient encoder to produce background embeddings always equal to the information-less vector s': $f_{\theta_S}(t(x)) \sim \delta(s')$. Using this hypothesis in the computations (see Sec. D.2.2 of the Supplementary) and ignoring constant terms, we obtain:

$$\mathcal{L}_{s'\text{-unif}} = \log\frac{1}{N_Y}\sum_{i=1}^{N_Y}\left(\exp\frac{-||f_{\theta_s}(t(y_i)) - s'||_2^2}{\tau} + \frac{1}{N_Y}\sum_{j=1}^{N_Y}\exp\frac{-||f_{\theta_s}(t(y_i)) - f_{\theta_s}(t(y_j))||_2^2}{2\tau}\right) \tag{4.11}$$

To respect the Information-less hypothesis, we re-write Eq. 4.9 as a Lagrangian function, with the constraint expressed as a β-weighted ($\beta \geq 0$) KL regularization. Assuming that s_x follows a Gaussian distribution centered on $f_{\theta_s}(x)$ with a standard deviation τ (constant hyper-parameter), we derive the KL divergence as an L2-distance between $f_{\theta_s}(t(x))$ and s', as in [117]:

$$\mathcal{F}(\theta_S, \beta; x, y, s) = -\lambda_S\mathcal{L}_{y\text{-align}} - \lambda_S\mathcal{L}_{s'\text{-unif}} - \beta\frac{1}{N_X}\sum_{i=1}^{N_X}\frac{||f_{\theta_s}(t(x_i)) - s'||_2^2}{2\tau} \tag{4.12}$$

On the null Mutual Information constraint

In Eq. 4.5, to avoid information leakage between common and salient space, we constrain our problem so that the MI between c and s is null. Another common choice would be to simply *minimize* $I(c, s)$ instead than forcing it to be equal to 0. In Tab. 4.5 and 4.6, we show that the latter choice (*i.e.,* $I(c, s) = 0$) clearly outperforms (variational) MI minimization methods, as vCLUB [53], vL1out [226], vUB [11], and TC [180], (see Sec. D.6.7).

Minimizing $H(c)+H(s)$ is detrimental: By def., $I(c; s) = -H(c, s)+H(c)+H(s) \geq 0$, which entails $H(c; s) \leq H(c) + H(s)$. Thus, a trivial way to minimize $I(c; s)$ would be minimizing $H(c) + H(s)$. However, it reduces the quantity of information contained in either c or s,

which could be detrimental. Furthermore, the Common and Salient InfoMax losses of our framework seek to maximize $H(c)$ and $H(s)$ rather than minimizing them.This is why, instead of minimizing $I(c;s)$, we propose to simultaneously maximize $H(c,s)$, $H(c)$ and $H(s)$, until $H(c,s) = H(c) + H(s)$, to respect the constraint $I(c;s) = 0$.[8]

k-JEM: kernel-based Joint Entropy Maximization: Here, we propose a method to estimate and maximize $H(c,s)$ without requiring any assumptions about the form of its pdf nor requiring a neural network-based approximation ([53, 11, 226]). Inspired by [125], we develop $H(c,s)$ with a re-substitution entropy estimator: $-\hat{H}(c,s) = \frac{1}{N_X+N_Y}\sum_{i=1}^{N_X+N_Y} \log \hat{p}(c_i,s_i)$. We estimate the density $\hat{p_\theta}(c_i,s_i)$ with a Gaussian Kernel Density Estimation with a constant standard deviation parameter τ with samples (c,s) uniformly drawn from the target dataset $(f_{\theta_S}(t(y)), f_{\theta_C}(t(y))) \sim p(c,s|y)$ and from the background dataset $(f_{\theta_S}(t(x)), f_{\theta_C}(t(x))) \sim p(c,s|x)$. The indices i and j refer to two different samples in the dataset. Full computations in Appendix, Sec. D.4.

$$-H(c,s) = \frac{1}{N_X+N_Y}\sum_{i=1}^{N_X+N_Y} \log \frac{1}{N_X+N_Y}\sum_{j=1}^{N} \exp\frac{-||c_i-c_j||_2^2}{2\tau} \exp\frac{-||s_i-s_j||_2^2}{2\tau} \tag{4.13}$$

4.2.5 Disentangling attributes in the salient space

Here, we propose to explore an extension of the salient contrastive loss in the case where independent fine-grained attributes about the target dataset $\{a_i \in \mathcal{R}^{D_S}\}_{i=1}^{N_Y}$ are available. We assume the existence of D_S attributes and each attribute a^d is generated by a single factor of generation s_y^d of the target dataset. We also make the hypothesis that the given attributes describe the entire salient variability of the target dataset,[9] and thus construct our salient encoder to output (exactly) D_S latent dimensions. We aim to construct a salient space where each salient latent dimension S^{d_s} only depends on its corresponding attribute a^{d_s}. By leveraging the attributes in a supervised manner, we re-write Eq. 4.5 by replacing $I(y;s)$ with the sum of all attribute Supervised InfoMax terms :

$$\text{argmax } I(x;c)+I(y;c)+\frac{1}{D_S}\sum_{d=1}^{D_s} \underbrace{I(a^d;s^d)}_{\text{d-th SupInfoMax}} \quad \text{s.t. } D_{KL}(s_x||\delta(s')) = 0 \text{ and } I(c,s) = 0 \tag{4.14}$$

[8] In this work, we implicitly assume that the encoders f_{θ_c} and f_{θ_s} can model any distribution.

[9] If it is not true, one can add a Salient InfoMax term (Eq.4.5) and increase the salient space dimension.

Taking inspiration from [74, 75], we then decompose each d-th attribute Supervised InfoMax term in a supervised alignment and uniformity term:

$$I(a^d; s^d) \geq \frac{1}{N_Y} \sum_{i=1}^{N_Y} w_\sigma(a_i^d, a_j^d) \frac{||s_i^d - s_j^d||_2}{\tau} + \hat{H}(s^d) = \mathcal{L}_{\text{d-th SupInfoMax}} \tag{4.15}$$

where the indices i and j refer to two different samples in the target dataset and the scalar weight $w_\sigma(a_i^d, a_j^d) = \frac{K_A(a_i^d, a_j^d)}{\sum_{j=1} K_A(a_i^d, a_j^d)}$ measures the similarity between their attributes. We define K_A as a Gaussian kernel and the entropy $\hat{H}(s^d)$ is also estimated, as before, with a Gaussian kernel.

4.2.6 Experiments

Here, we measure our method's ability to separate common from target-specific variability factors. We train a Logistic (or Linear) Regression on inferred factors to assess whether the information about a characteristic present in both populations or only in the target one, is captured in the common (C) or in the salient (S) latent space. We compute (Balanced) Accuracy scores (=(B-)ACC), or Area-under Curve scores (=AUC) for categorical variables, Mean Average Error (=MAE) for continuous variables, and the sum of the differences (δ_{tot}) between the obtained results and the expected ones.

Hyper-parameters. We empirically choose $\tau = 0.5$ for all experiments and losses. The other hyper-parameters are λ_C and λ_S, which weigh the common terms and salient terms, respectively, and λ, which weighs the independence regularization. The choice of these weights depends on the ratio between common and target-salient information quantity, which might differ among datasets. Architectures and hyper-parameters are chosen as the top-performing ones for each experiment.

SOTA CA methods We have compared the performance of our method with the most recent and best-performing CA-VAE methods whose code was available: cVAE [3] [10], conVAE [6] [11], MM-cVAE [299] and SepVAE [180]. In each experiment, all CA-VAE use the same encoder-decoder architecture, as described in the Supplementary D.6. The architecture used for SepCLR is also described in Sec. D.6.

Digits superimposed on natural backgrounds. In this experiment particularly suited to CA and inspired from [327], we consider CIFAR-10 images as the background dataset ($y = 0$)

[10] Here, for cVAE, we use the fixed version of the TC regularization described in [180].
[11] Here, conVAE corresponds to cVAE method without the TC regularization, as in [6].

and CIFAR-10 images with an overlaid digit as the target dataset ($y = 1$). In Tab .4.5, our model outperforms all other methods in correctly capturing the common factors of variability (*i.e:* objects) in the common space and the target-specific factors of variability (*i.e:* digits) in the salient space.

CelebA accessories dataset. We consider a subset of CelebA [174], already introduced in the previous chapter ([180]). Differently, in this paper, special care was employed in order to balance the individuals of the background and target sets in terms of sex. In Tab .4.6, SepCLR correctly captures the information that enables distinguishing glasses from hats only in the salient space, and it puts the information to distinguish men from females in the common space. Our method globally outperforms all other methods (smallest δ_{tot}). "Best Expected" reports a perfect result (100) when the attribute should be present in that latent space and a random result (50) when it should not.

Table 4.5: Digits on CIFAR-10 (B-ACC). Details in Sec.D.6.7.

	Digits		Objects		$\delta_{\text{tot}} \downarrow$
	S ↑	C ↓	S ↓	C ↑	
cVAE	90.6	23.0	11.2	33.4	90.2
ConVAE	86.2	21.0	10.6	35.6	89.8
MM-cVAE	88.8	19.6	12.2	32.0	93.6
SepVAE	90.6	17.8	10.6	36.6	81.2
SepCLR-vCLUB sym	94.4	18.0	8.0	14.6	97.0
SepCLR-vCLUB C → S	95.2	39.4	9.2	27.2	106.2
SepCLR-vCLUB S → C	95.2	57.0	8.8	31.8	118.8
SepCLR-vL1o sym	95.0	18.4	8.4	15.4	96.4
SepCLR-vL1o C → S	94.0	23.0	10.0	31.8	87.2
SepCLR-vL1o S → C	95.4	41.0	9.2	28.8	106.0
SepCLR-vUB sym	94.6	42.0	8.2	29.0	106.6
SepCLR-vUB C → S	92.8	23.4	7.8	22.6	95.8
SepCLR-vUB S → C	96.6	41.8	8.6	28.6	105.2
SepCLR-TC	95.2	68.6	10.2	24.2	139.4
SepCLR-MMD	94.6	21.2	9.0	62.2	53.4
SepCLR-no k-JEM	95.6	94.4	9.0	42.0	145.8
SepCLR-k-MI	96.2	19.8	8.0	65.8	45.8
SepCLR-k-JEM	96.2	11.0	10.4	73.2	**32.0**
Best expected	100.0	10.0	10.0	100.0	0.0

Table 4.6: CelebA accessories. (B-ACC).

	Hats/Glss		Sex		$\delta_{\text{tot}} \downarrow$
	S ↑	C ↓	S ↓	C ↑	
cVAE	83.89	66.56	60.25	60.60	82.32
ConVAE	81.64	65.94	61.53	58.93	86.90
MM-cVAE	84.60	66.43	60.56	61.57	80.82
SepVAE	84.46	65.19	60.12	59.20	81.65
SepCLR-vCLUB sym	98.98	59.62	65.20	54.23	71.61
SepCLR-vCLUB C → S	98.81	73.71	61.77	53.72	82.95
SepCLR-vCLUB S → C	98.66	95.95	67.65	73.16	91.78
SepCLR-vL1o sym	98.83	56.94	57.97	51.38	64.60
SepCLR-vL1o C → S	99.04	93.17	63.35	59.13	89.91
SepCLR-vL1o S → C	98.46	94.33	65.00	71.77	89.10
SepCLR-vUB sym	98.68	87.33	63.59	56.09	96.15
SepCLR-vUB C → S	98.73	94.40	66.58	71.37	90.88
SepCLR-vUB S → C	98.78	93.92	62.94	61.27	96.81
SepCLR-TC	98.97	98.76	60.39	74.96	85.22
SepCLR-MMD	98.95	67.50	71.83	65.47	74.91
SepCLR-no k-JEM	99.03	66.68	98.48	79.48	86.65
SepCLR-k-MI	98.96	77.10	63.07	71.08	70.13
SepCLR-k-JEM	98.57	55.21	62.52	78.00	**41.16**
Best expected	100.0	50.0	50.0	100.0	0.0

Neuro-imaging: parsing schizophrenia's heterogeneity. Given healthy MRI scans and patients with schizophrenia, we aim to capture pathological patterns only in the salient space that should correlate with clinical scales (SAPS, and SANS) while not being biased by demographic variables (age, sex or acquisition sites), which should be encoded in the common space. As in [180], we gather T1w VBM [17] warped MRIs and evaluate our method in a

cross-validation scheme. In Table 4.7, we can clearly see that our method outperforms all others.

Table 4.7: Separate healthy from pathological variability in Schizophrenia disorder. Best in **bold**.

	Age MAE		Sex B-ACC		Site B-ACC	
	C ↓	S ↑	C ↑	S ↓	C ↑	S ↓
cVAE	6.43±0.18	7.27±0.25	75.06±3.48	74.99±2.15	65.12±4.06	59.62±5.42
ConVAE	6.40±0.26	7.46±0.18	74.45±1.80	72.72±1.32	60.42±3.67	54.46±2.46
MM-cVAE	6.55±0.18	7.10±0.34	72.80±3.95	72.15±2.47	63.24±1.41	56.69±9.84
SepVAE	**6.40±0.13**	**7.98±0.25**	74.19±1.81	72.61±2.19	63.89±2.16	44.10±5.78
SepCLR-k-JEM	6.64±0.21	7.72±0.45	**76.5±1.98**	**70.85±1.89**	**66.94±5.06**	**42.40±4.91**
	SANS MAE		SAPS MAE		Diagnosis	
	C ↑	S ↓	C ↑	S ↓	C ↓	S ↑
cVAE	5.89±0.67	4.35±0.26	4.65±0.34	2.98±0.18	60.66±2.63	68.24±5.42
ConVAE	6.17±0.45	3.95±0.28	4.50±0.37	2.76±0.18	61.85±2.60	58.53±4.87
MM-cVAE	6.78±0.54	4.92±0.58	4.52±0.33	3.16±0.05	64.25±2.98	70.94±4.08
SepVAE	7.05±0.67	4.14±0.39	4.79±0.67	2.60±0.27	60.90±1.75	79.15±3.39
SepCLR-k-JEM	**9.17±2.49**	**3.74±0.12**	**5.54±0.70**	**2.52±0.16**	**60.16±1.19**	**79.90±1.57**

Chest and eye pathologies subtyping. We propose two experiments using subsets of CheXpert [136] and ODIR dataset (Ocular Disease Intelligent Recognition dataset) [12] to assess the ability of our method in a controlled environment. About CheXpert, we have healthy X-ray scans (background) and diseased scans (target) divided into 3 distinct subtypes: cardiomegaly, lung edema, and pleural effusion. In the ODIR dataset, there are healthy (background) and diseased fundus images (target) which are divided into 5 subtypes: Diabetes, Glaucoma, Cataract, Age macular degeneration, and pathological Myopia. Sex-related patterns should only be captured in the common encoder.

Table 4.8: CheXpert X-ray scans (B-ACC).

	Subtype		Sex		δ_{tot} ↓
	S ↑	C ↓	S ↓	C ↑	
cVAE	45.77	49.27	54.24	81.26	93.48
ConVAE	42.31	52.53	60.88	79.30	108.8
MM-cVAE	42.50	50.89	57.04	80.19	102.24
SepVAE	42.20	51.10	56.38	79.95	102.34
SepCLR-k-JEM	61.30	52.85	61.57	80.25	**89.87**
Best expected	100.0	33.0	50.0	100.0	0.0

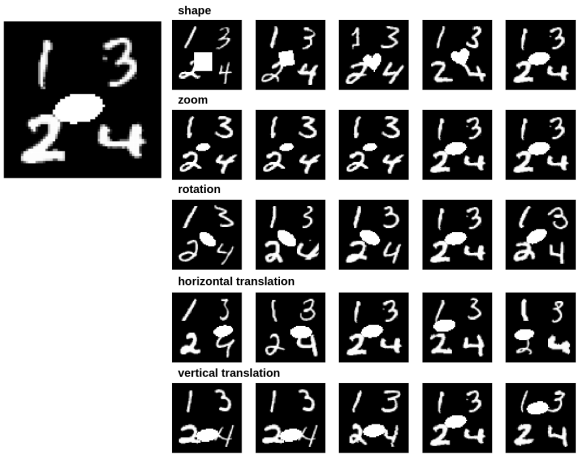
Table 4.9: ODIR images (B-ACC).

	Subtype		Sex		δ_{tot} ↓
	S ↑	C ↓	S ↓	C ↑	
cVAE	46.13	43.91	49.11	51.86	120.03
ConVAE	49.80	52.01	50.82	47.01	131.86
MM-cVAE	42.79	43.66	54.91	53.76	131.02
SepVAE	38.64	41.44	52.91	52.62	124.75
SepCLR-k-JEM	68.54	47.71	52.48	59.62	**97.03**
Best expected	100.0	25.0	50.0	100.0	0.0

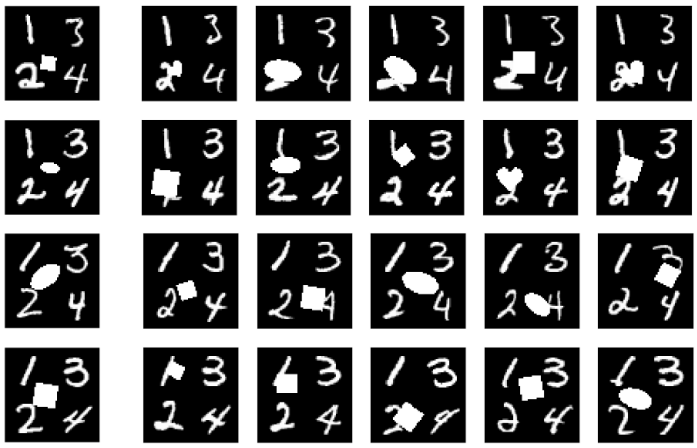
Disentangling dSprites while contrasting with a background. To evaluate CA method

[12] https://www.kaggle.com/datasets/andrewmvd/ocular-disease-recognition-odir5k

enriched with target attributes, we provide a novel toy dataset. Background dataset X consists of 4 MNIST digits (1-4) regularly placed on a grid. Target dataset Y consists of a dSprites item added upon the grid of digits. dSprites only exhibit 5 generation factors (shape, zoom, rotation, X position, Y position). Using Eq. 4.14, we train our salient encoders in a supervised manner to capture and disentangle each attribute in a single salient space dimension (Fig. 4.9a). The common encoder is instead trained to capture the background variability (Fig. 4.9b). Quantitatively, 1st salient dimension distinguishes shapes (B-ACC= 98.23%) while the concatenation of other salient dimensions and common dimensions does not (B-ACC= 36.08%). 2_{nd} predicts zoom attribute ($R^2 = 0.977$) while others don't 0.002. 3_{rd} predicts rotation ($R^2 = 0.947$), others don't $R^2 = 0.0$. 4_{th} predicts horizontal translation ($R^2 = 0.995$), others don't $R^2 = 0.017$. 5_{th} predicts vertical translation ($R^2 = 0.995$), others don't $R^2 = 0.009$. This shows that our method correctly *separate* common from salient information and *disentangle* salient factors (in a supervised manner) *at the same time*.



(a) Given the latent vector of the upper left image, we modify only one salient dimension in each row while freezing the others, then fetch the image in the dataset whose latent vector is the closest.



(b) Given the common latent vector of an image (left column), we fetch the image in the dataset whose inferred common latent vector is the closest in terms of L2 distance.

Figure 4.9: Qualitative results on attribute-supervised SepCLR

4.2.7 Limitations and Perspectives

An important question in Contrastive Analysis, is the identifiability of the models. Namely, under which conditions can the models recover the true latent factors of the underlying data-generating process. Recent works have shown that non-linear models, VAEs included, are generally not identifiable. To obtain identifiability, two different solutions have been proposed:

1) either regularizing [145] the encoder or 2) introducing an auxiliary variable so that the latent factors are conditionally independent given the auxiliary variable [130, 147]. In CA, neither of these solutions may be used [13]. Even though SepCLR effectively *separates* common from salient factors, it does not assure that *all* true generative factors have been identified (like all CA methods). This is a serious limitation of CA methods that we leave for future works. Intriguingly, we also noticed that adding a reconstruction loss during the training degrades performance, see Sec. D.7.2 in Appendix. However, adding a powerful generator, as in [329], on top of the frozen encoders would allow synthesizing new images and increase interpretability.

4.2.8 Conclusion

In this paper, we leverage the power of Contrastive Learning to learn semantically relevant representations for Contrastive Analysis. We reformulate Contrastive Analysis as a constrained InfoMax paradigm. Then, we propose to estimate the Mutual Information terms via alignment and uniformity terms. Importantly, we motivate a novel independence term between common and salient spaces computed via Kernel Density Estimation (KDE). Our method outperforms related works on toy, natural, and medical datasets specifically made to evaluate the common/salient separation ability.

[13] The dataset label could be considered as an auxiliary variable, but it does not make c and s independent

Chapter 5

Perspectives

Chapter summary. This chapter identifies key issues in current approaches to understanding psychiatric diseases and offers insights into unresolved perspectives: the data scarcity and the covariate role. In response to these challenges, this chapter introduces a methodological perspective entitled Normative Contrastive Analysis inspired by Normative Modeling and Transfer Learning research.
Moreover, the chapter addresses the difficulty of current methods to generalize on new independent acquisition sites databases, the use of multiple modalities in computational psychiatry, and the difficulty of interpreting representation learning methods. This chapter advances several strategies to respond to these crucial points.

Contents

The objective of this thesis is to develop methodologies to parse the *neurobiological* disorder heterogeneity within psychiatric disorders with the goal of improving the understanding of mental disorders. Throughout this manuscript, the developed methods operate under the assumption that pathological heterogeneity is distinct from the variability present in the healthy sample. Particularly, the final chapter delves into Contrastive Analysis (CA) methods, which aims at identifying and separating *distinct* and *interpretable* pathological latent generation factors from latent healthy generation factors.

In practical applications, these approaches face several limitations that impede their effectiveness. Firstly, the assumption of independence between pathological variability and the variability observed in the healthy population does not accurately reflect the biological mechanisms underlying psychiatric disorders, as demographic factors strongly correlate with the development of psychiatric diseases [105]. Secondly, the performance of Contrastive Analysis methods is often hindered by the requirement for balanced diseased/control datasets, which frequently suffer from inadequate sample sizes, especially in neuroimaging applications.

To overcome these two major limitations, this chapter proposes to conceive a new class of methods entitled Normative Contrastive Analysis (NCA). Similarly to traditional Normative Models (NM) [195], NCA would enable 1) integrating demographic covariates in the training and inference process and 2) transferring knowledge from a large healthy database to a smaller case/control cohort. Unlike traditional NM, NCA would use CA techniques to separate pathological from healthy latent factors, leading to a better interpretation of the latent space. Compared to traditional NM, integrating patients in the training process would potentially enhance the discriminative power between case and control groups, as further described in the next sections.

Eventually, several challenges and opportunities remain to be addressed to improve the applicability of neuroimaging techniques such as Normative Contrastive Analysis or Subgroup Discovery. These methods often struggle to generalize effectively to new, independent test data samples due to distribution shifts arising from diverse acquisition sites contributing to the databases. Moreover, the frame of studies investigated in this thesis remains restricted to neuroanatomical images. In practice, other modalities, promising for psychiatric disorder biomarkers search (such as diffusion MRIs, functional MRIs, and genomics data) could be further integrated in a multi-modal setting to increase the performance of these methods. While

deep learning methods proposed in neuroimaging generally offer valuable representations, their interpretability remains a significant hurdle, limiting their practical utility.
Below is an outline of the perspectives addressed in this chapter:

1. leveraging large healthy imaging datasets to enhance the model performances via transfer learning techniques,

2. integrating demographic covariates into the learning and inference process, as proposed in normative modeling,

3. remove the data bias of the acquisition sites or acquisition scanner from the learned representations,

4. adding and fusing novel modalities to increase the number of potential biological markers to be discovered,

5. increasing our models' interpretability via enhanced representation disentanglement or enhanced generation performances.

5.1 Toward Normative Contrastive Analysis (NCA)

5.1.1 Transferring the knowledge from a large to a small cohort

Fig. 5.1 describes a methodological perspective that would integrate the ideas pinpointed in the previous paragraph. The upper figure (1) describes leveraging the Transfer Learning strategy, extensively studied in [76], by pre-training a deep convolutional encoder on a large healthy heterogeneous cohort. To successfully model the general physiological variability, it proposes to enrich the training process with demographic variables used as supervision signals. Importantly, this pre-training process is not meant only to classify and regress demographic attributes but also to model the general variability of the healthy population. Therefore, special care should be employed to produce a latent space that captures both these attributes and the general patterns of variability that exist in the input dataset. To this end, training methods such as conditioned VAEs [150, 295], or conditioned contrastive losses [75] could be employed. Also, to reduce the heterogeneity of the representations, debiasing techniques [24] could be employed to encourage the encoder to disregard irrelevant patterns of variability such as acquisition sites or machine setting attributes. This strategy is discussed in Sec. 5.2.

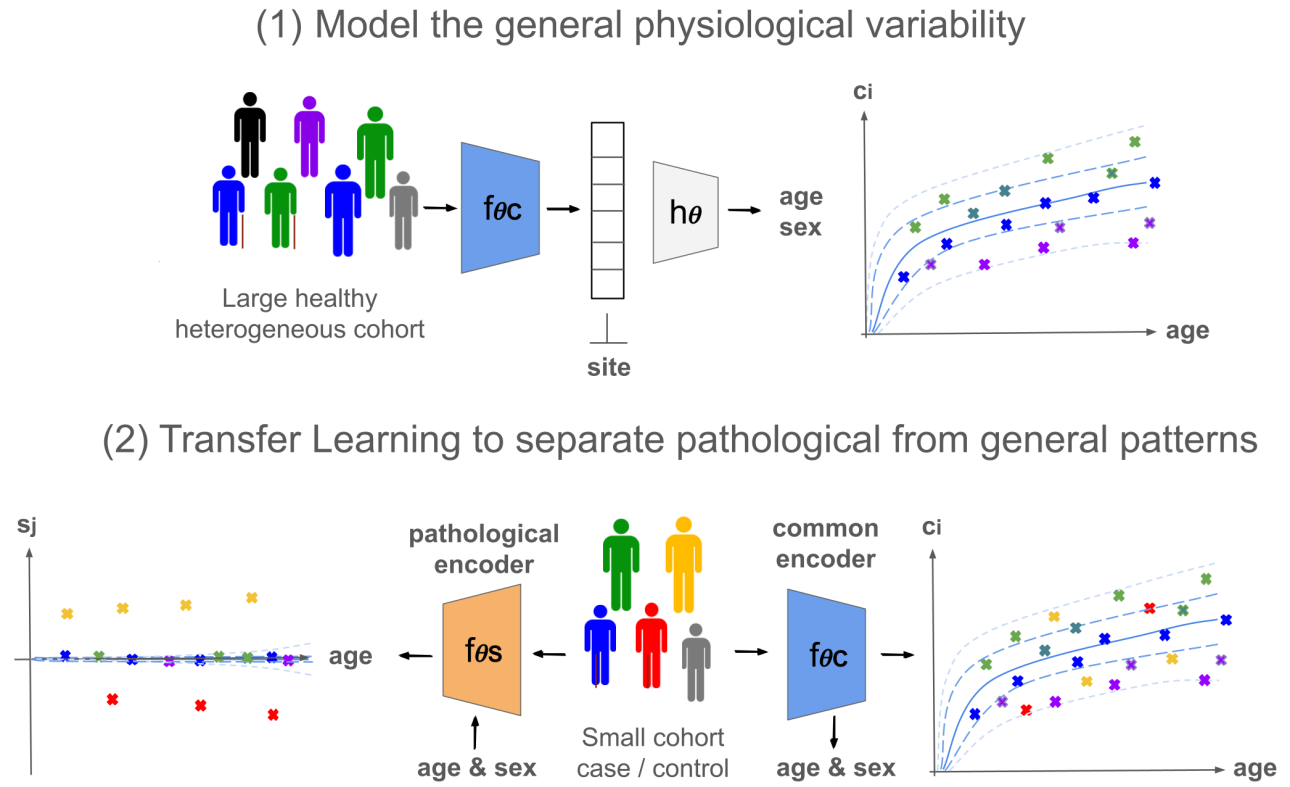


Figure 5.1: Scheme depicting a perspective of debiased, covariate-adjusted Contrastive Analysis strategy. In step (1), a model learns from a large heterogeneous cohort on a self-supervised task. The objective is to produce a latent space that captures demographic attributes (such as age and sex) and the general patterns of variability that exist in the input dataset. To further reduce the heterogeneity of the representations, debiasing techniques should be employed to encourage the encoder to disregard irrelevant patterns of variability, such as acquisition sites. In step (2), the learned representation is transferred as "the common encoder" in order to train a normative contrastive analysis on a disorder-specific cohort. This strategy aims at identifying two separated lower-dimensional representations of the inputs from a small case/control cohort. At the right, the common representations capture the general patterns of variability, for which pathological samples do not necessarily deviate from the normal, healthy deviation. At the left, the pathological (or salient) representations capture the patterns of variability that only exist in the diseased population. Contrary to traditional Contrastive Analysis, covariates such as age and sex are integrated into the training process by using a Conditional Independence constraint (*i.e.* $I(c, s|age, sex) = 0$). This conditioning aims at preventing information leakage between the common and salient space of groups of controls and patients of around the same age and sex. In this setup, step 2) benefits from step 1) thanks to transfer learning and provides better interpretable representations by separating pathological-specific dimensions from those shared between healthy and diseased populations. Ultimately, as in Normative Modeling, a Regression Process could be used to estimate a normative chart conditionally to age and sex, enabling to compute deviation scores on the shared and pathological representations.

The lower part of the figure illustrates a strategy to identify two separated lower dimensional representations of the inputs from a small case/control cohort. At the right, the common representations capture the general patterns of variability, for which pathological samples do not necessarily deviate from the normal, healthy deviation. At the left, the pathological (or salient) representations capture the patterns of variability that only exist in the diseased population. Several choices can be made, both encoders can be initialized from step (1), to allow the low-level features re-use, as described in [210]. The common encoder can be either frozen in step (2) or fine-tuned.
The fine-tuning of the common encoder in step (2) could be beneficial when the small cohort comes from different acquisition sites. In that case, domain adaptation [102] or distribution matching [299] strategies may be considered to align the distribution of diseased and healthy samples from the small cohort of step (2) onto the healthy population of step (1). Besides, covariate-conditioned independence constrained could be considered to ensure that the common space does not capture disorder-specific patterns, as described in Sec.5.1.3. The salient space aims to capture pathological variability patterns that do not exist in the general population (such as patterns that correlate with self-disturbance, auditive or visual hallucinations, anhedonia, or disorganized thinking, for example). The training losses required to perform this task have been extensively studied in SepCLR's paper [181].

5.1.2 Enhancing deep normative models with diagnosis supervision

This section motivates the need for integrating diagnosis labels during the training process of deep normative models. In particular, it encourages separating the latent space of normative models into a healthy and a pathological latent space.

Deep normative models. Normative modeling (NM) is a statistical framework for mapping population trajectories of the relationships between biological readouts (such as height, mass index, or neuroanatomical measures like cortical thickness, and gray matter volumes for ex.) and covariates (such as age, and sex for ex.) while simultaneously preserving individual-level information. This class of methods has been extensively studied in the context of computational psychiatry [194, 195, 196, 148, 149, 241]. As a reason, it overcomes the shortcomings of case-control classification by estimating the deviation scores of the response variables (*e.g.* brain measures) of each individual adjusted on its covariate values (*e.g.* age, sex).
In recent times, numerous methodologies have emerged for the direct computation of normative models from high-dimensional raw inputs utilizing automatic deep learning feature extractors.

Essentially, the primary concept involves estimating normative models within the representation space derived from a deep feature encoder trained using self-supervised techniques such as Variational or Adversarial Auto-Encoder (VAE) [294, 159, 160] or Auto-Encoder (AE) [223]. These approaches closely resemble unsupervised Anomaly Detection (AD) methods, which strive to construct a latent space that retains information about input images while aligning it with a known distribution to facilitate easy computation of density and likelihood.

Toward supervised Anomaly Detection. In a medical context, Anomaly Detection methods [26, 258, 251, 222, 273, 228] generally aim at estimating the general population distribution to compute individual-level deviation scores. Practitioners thus expect pathological samples to deviate from the healthy, general population, which enables the development of a diagnostic tool that does not necessarily fit the case/control paradigm. However, previous work has shown that merely unsupervised anomaly detection methods could underperform when it comes to predicting the diagnosis compared to purely supervised methods. Notably, Gornitz et al. wrote in 2013 that "*the predictive performance of purely unsupervised anomaly detection often fails to match the required detection rates in many tasks and there exists a need for labeled data to guide the model generation*" [107].

In the context of neuropsychiatric applications, Pinaya et al. [223] conducted a comparison between deep normative models and traditional case/control classifiers. They observed that "*Although the traditional classifiers had a better mean performance in most cases, the differences between the two approaches were not statistically significant in most of the cases*". This finding contradicts the conventional assumption often posited by researchers: given the biological heterogeneity inherent in psychiatric disorders, one would anticipate that a method capable of parsing such heterogeneity would outperform traditional case/control classification methods in most cases.

To enhance the classification capabilities of Normative Models, one potential strategy involves training the deep encoder in a semi-supervised manner. This entails integrating the diagnosis within the training and fine-tuning process, as proposed in Gornitz et al. [107] and Ruff et al. [239], or in the salient space of the contributions introduced in this thesis, that is SepVAE [180], and SepCLR [181], whose salient space losses hold close resemblance with the semi-supervised AD method DeepSAD [239]. Still, normative models hold promising performance, and they remain particularly pertinent as they enable integrating covariates in the training and inference process. Therefore, such a covariates-adjustment strategy could be further incorporated into semi-supervised anomaly detection models, or any look-alike methods integrating the diagnosis

as a supervision signal, such as SepCLR [181]. With this objective in mind, the following section elucidates various insights to incorporate covariate conditioning within NCA.

5.1.3 Toward covariate-conditioned Contrastive Analysis

This section describes and analyzes the pitfalls of Contrastive Analysis models in neuroimaging. Then, it provides several strategies to overcome these limits, notably by conditioning and adjusting the representations learned with respect to demographic covariates.

On the limits of Contrastive Analysis. In this manuscript, Contrastive Analysis models are validated by evaluating the performance of linear models on regression and classification tasks that aim at predicting either a) demographic variables or b) pathological variables. A satisfactory Contrastive Analysis model is expected to a) produce shared (or common) vectors that successfully estimate the general demographic attributes, such as age, sex, and acquisition site but not the clinical scale measures, and b) produce salient (or pathological) vectors that correlate with clinical and cognitive scales, but not with demographic variables (age, sex, and acquisition site). While this behavior has been observed to a certain extent, unexpected results were still noticed. In SepVAE and SepCLR's papers [180, 181], the salient vectors produced performances that are better than average/random on age and sex classification and regression tasks. Besides, it was also observed that the sex classification performance in the common space is relatively poor compared to similar unsupervised representation learning strategies.
These observations can be explained by the fact that demographic factors strongly correlate with the development of psychiatric disorders [105]. Therefore, while the salient space captures pathological patterns that vary depending on the individual's sex (via hormonal differences, among other causes) and age (via disorder duration, among other causes), the common space fails to capture the age-related and sex-related natural neuroanatomical variability successfully. Therefore, I believe that the independent hypothesis prevents information leakage between the common and salient space. Therefore, a relevant line of research would be to integrate demographic covariates, which are generally available, in both the evaluation and learning process of common and salient latent spaces.

Integrating covariates in the learning process. Let us provide methodological leads that would help integrate demographic covariates in the Normative Contrastive Analysis training process. As a reminder, in the InfoMax optimization objective, two constraints were introduced:

a) the information-less hypothesis: $D_{KL}(p_{s_X}||\delta(s')) = 0$ that encourages healthy samples distribution to align on a distribution with a null variance centered on a constant information-less vector s', and b) a null Mutual Information constraint: $I(c, s) = 0$, that prevents information leakage between the common and salient space.

The integration of demographic covariates could benefit the produced representation of both Salient and Common space:

- *Salient space:* The salient space disregards healthy neuroanatomical variability patterns (including natural aging and sex-related differences), which is ensured by the information-less hypothesis. Also, the salient space captures pathological neuroanatomical patterns via the salient Infomax term $I(s, y)$. However, as assumed in normative models, diseased representations are said to be deviating from the healthy range when they are discriminable from healthy samples having the same demographic attributes (*i.e.* age and sex). Therefore, constructing a salient representation space that potentially discriminates diseased from healthy samples requires having demographic variables (*i.e.* age and sex) as inputs.

- *Common space:* As they are available at training, demographic attributes such as age and sex could be used to refine the learning of the common space. The common space is expected to capture the neuroanatomical patterns associated with the general healthy variability, encompassing natural aging and sex-related brain morphology differences. These attributes could thus be used as auxiliary supervision signals in the common Infomax terms $I(c, x) + I(s, y)$. Additionally, the common space ignores the pathological variability patterns via the Independence hypothesis $I(c, s) = 0$. This term may potentially produce a common space that conflicts with the salient space, as the salient space captures pathological patterns that likely correlate with age and sex attributes.

To avoid such a pitfall, practitioners should integrate the covariates (*e.g.* a for the age attribute) within a Conditional Independence constraint (*i.e.* $I(c, s|a) = 0$). This conditioning would prevent information leakage between control groups' common and salient space and patients around the same age (and sex). To add this conditional mutual information regularization, one could make use of the insights introduced in [181] by only maximizing the age-conditioned joint entropy term $H(c, s|a)$: This paragraph uses the same notation as Sec. 4.2.4, and describes how to estimate and maximize $H(c, s|a)$. Using a re-substitution entropy estimator and then a Gaussian Kernel Density Estimation, the samples (c, s) should be uniformly drawn from the age-conditioned target distribution $p(c, s|y, a)$ and from the age-conditioned background

distribution $p(c, s|x, a)$. During this sampling, I propose to weight the contribution of each sample j by a similarity measure between a_j and a_i with a similarity function $w(a_i, a_j)$, as in [74]. The indices i and j refer to two different samples in the dataset.

$$-H(c, s|a) = \frac{1}{N_X + N_Y} \sum_{i=1}^{N_X+N_Y} \log \frac{1}{N_X + N_Y} \sum_{j=1}^{N_X+N_Y} \frac{w(a_i, a_j)}{Z(a_i)} \exp \frac{-||c_i - c_j||_2^2}{2\tau} \exp \frac{-||s_i - s_j||_2^2}{2\tau} \tag{5.1}$$

where $Z(a_i) = \sum_{j=1}^{N_X+N_Y} w(a_i, a_j)$, and w is a Kernel, that can be Gaussian or categorical for ex. depending if a is continuous or discrete.

5.2 Debiasing methods to reduce the known irrelevant variability

When addressing diagnosis prediction or clinical scale regression with neuroimaging data, several works have observed poor cross-acquisition site generalization performance and high overfitting on the training acquisition scanner settings for a broad range of prediction tasks. To overcome this issue, several works have proposed covariate adjustment techniques to regress out the effect of acquisition sites in the input data [289, 94, 95, 103]. However, these works rely on linear methods and may fail to effectively generalize to deep representation learning methods that may still overfit on non-linear patterns stemming from the acquisition settings. Moreover, such techniques generally require the acquisition site to be known during inference. In Deep Learning and Medical Imaging, several works have designed debiasing methods for representation learning. In neuroimaging particularly, normative models [149, 27] have proposed to consider acquisition sites as a covariate and propose to fit a Hierarchical Bayesian Repression (HBR) process to build a healthy normative range for each acquisition site in the training data. These methods require the acquisition site to be known during the inference process and may fail when a given inference image comes from a new acquisition site, but remain particularly relevant for research purposes. In natural imaging and medical imaging, several methods [23, 24, 25, 22, 56] propose to learn a representation that captures a target attribute (*e.g.* a diagnosis) while debiasing with a confound attribute (*e.g.* the acquisition site). Debiasing methods have been extensively studied in the literature, and their use should be considered in our paradigm to reduce the distribution shifts of our representations.

Matching biased-aligned and bias-conflicted distributions. Assuming a discrete bias attribute such as acquisition sites or scanner types, several methods emerge as good candidates for reducing the distribution shifts of our representations. Recently, Barbano et al. [24] introduced a debiasing regularization entitled FairKL. The authors denote as "bias-aligned" samples $x^{\cdot,b}$ samples having the same bias attribute (*e.g.* MRIs acquired with the same scanner) and "bias-conflicting" samples $x^{\cdot,b'}$ samples having a different bias attribute (*e.g.* MRIs acquired with a different same scanner). Given an anchor x, if the bias is "strong" and easy to learn, a bias-aligned sample $x^{\cdot,b}$ will probably be closer to the anchor x in the representation space than a bias-conflicting sample $x^{\cdot,b}$, even though they are from the same class (*e.g.* from the healthy class). Additionally, a sample is said to be positive $x^{+,\cdot}$ with respect to the anchor, if it belongs to the same class (healthy or diseased for ex.).
Given these notations, the authors designed a regularization loss term aiming at aligning the distributions of the bias-conflicting samples and the bias-aligned samples given an anchor and its class (healthy or diseased for ex.):

$$\mathcal{R}^{\text{FairKL}} = D_{KL}(\{d_i^{+,b}\}||d_k^{+,b'}\}) \tag{5.2}$$

By minimizing this loss term, one encourages every positive (resp. negative) bias-conflicting representation to have the same distance from the anchor as any other positive (resp. negative) bias-aligned representation. This regularization could be used in practice in a Contrastive Analysis framework, and further works could be led to investigate the impact of the choice of the assumed distributions (Gaussian, von Mises-Fischer, or Dirac for ex.), or divergence measures (the Maximum Mean Discrepancy or Jeffrey Divergence could also be used for ex.).

5.3 Adding modalities to increase the information quantity

The main objective of this thesis is to propose and advance state-of-the-art methods for parsing neuroanatomical heterogeneity in psychiatric disorders. Importantly, the contributions introduced in this thesis are not strictly limited to neuroanatomical imaging only. The methodologies introduced can also be applied to various neuroimaging modalities such as diffusion MRIs [87], functional MRIs [63, 323], as well as EEG recordings [132] or genomics data.

Discovering better biomarkers with other modalities. Adding and integrating the features provided by other neuroimaging modalities could potentially allow the discovery of new

brain-based biomarkers specific to mental illnesses. These ideas are motivated by related works that found functional deviations in schizophrenia [63, 323] (notably in default-mode network, central executive network, and salience network), deviations in diffusion connectivity patterns in various mental disorders [87], and connectivity patterns in electroencephalography [132]. These works motivate the fusion of information from several modalities to potentially enhance the number of biomarkers that would characterize pathological traits.

Adding and interpreting new modalities. When considering an additional modality, a novel signal source is introduced, with potential biomarkers of interest to discover, as well as new sources of variability. Adding a modality also makes hardly interpretable the relative contributions of each modality when estimating a clinical diagnosis or a clinical scale using a deep learning method. As done in several related works [13, 159, 7] (mostly with variational auto-encoders), an interesting perspective consists of designing a representation learning method (*e.g.* with contrastive representation learning, auto-encoders, or diffusion models) that both:

1. reduces the dimensionality of each modality with representation learning methods (such as Contrastive Learning methods, Variational Auto-Encoders, or Latent Diffusion models, for ex.)

2. identifies and separates latent generation factors in the representation space that underpin 1) joint / common patterns of variability between both modalities and 2) modality-specific patterns of variability for each modality.

As an example of application, let me consider joint brain folding images and whole-brain structural observations. In the literature, it is well admitted that the neuroanatomical patterns tangibly result from the interaction between polygenic risk factors and environmental stress factors [231, 111]. It is also well admitted that brain folding patterns mostly stem from genetic factors [39, 281], while most other neuroanatomical variability results from an interaction between genes and environmental factors. Cortical folding could be used to identify the neurodevelopmental variability from another variability that appeared later in life (*e.g.* atrophies related to alcohol, or drug use). Another perspective would be to separate the patterns of variability of structural MRIs that correlate with diffusion-based connectivity measures. Alternatively, we could also explore functional-based connectivity with brain connectome from diffusion MRI. Such applications would have the potential to bridge the gap between different modalities and help understand their relative contributions given a task of interest (*e.g.* clinical scale regression, diagnostic prediction, etc.).

When addressing Contrastive Analysis, two methodological tools have been developed enabling 1) compressing the dimension of an input with Mutual Information maximization or Evidence Lower BOund maximization, or 2) separating common from salient patterns by minimizing the Mutual Information with Total Correlation estimation [180] or kernel-based Joint Entropy estimation [181]. These strategies are now on the shelf to help researchers fuse and integrate several modalities to identify the common and modality-specific information given two modalities to help interpret their relative contributions.

5.4 Enhancing the interpretability of Contrastive Analysis methods

In this thesis, we investigated the use of Contrastive Analysis methods to push further the understanding of psychiatric disorders with neuroimaging data. These methods have shown promising results in producing salient vectors that capture the pathological patterns in patients while disregarding the physiological patterns that are common in the healthy, general population. However, these models remain hardly interpretable and would benefit from research projects in this sense. To interpret the results of Contrastive Analysis, two interesting objectives are 1) the identification of biomarkers in the input data that produce several distinct pathological processes and 2) the artificial generation of a diseased individual's brain without the pathological abnormalities by inferring its healthy digital avatar [6]. With these strategies in mind, let us identify the advantages and drawbacks of each of the classes of methods we have investigated.

Interpretability of Contrastive Analysis representation learning methods. Contrastive representation methods have successfully addressed transfer learning from big to small cohorts [76], and a method like SepCLR successfully separates common from disorder-salient patterns [181]. However, these methods remain hardly interpretable, as they lack a powerful decoder/generator. Thus, a solution could be to add a reconstruction strategy to these methods. However, the preliminary results obtained in this thesis when adding a reconstruction loss term during the training were not satisfactory. One of the reasons could be that the latent space perturbation invariance objective is not consistent with the input image reconstruction objective as it does not allow for a unique reconstruction solution given a data point in the latent space. A perspective would be thus to overcome the need for input perturbations (or

data augmentations) in the positive pair sampling strategy of SepCLR, as suggested in [77].
Another interesting perspective for representation learning Contrastive Analysis methods would be to make use of interpretability methods such as Shapley value (SHAP) [183], GradCam [253, 46], LIME [230], RISE [220], or DRISE [221]. These interpretability methods are model-agnostic, do not require a decoder/generator, and can directly provide saliency maps in the input space that directly highlight the pixels involved in a variable prediction or regression. However, current interpretability methods usually fail to produce unambiguous saliency maps for deep encoders [322].

Performing Contrastive Analysis with a powerful generation method. This thesis has explored, developed, and investigated CA-VAEs. These methods both produce separated common and salient representations and image reconstructions. Theoretically, they allow practitioners to visualize in the original space the effect of 1) modifying latent dimensions in the pathological space and 2) zeroing the pathological space (*i.e.* reconstructing the healthy digital version of the brain). In practice, CA-VAEs produce weaker representations than contrastive representation methods [76], and the image reconstructions they produce are generally blurry and hardly interpretable. Therefore, using other powerful generation methods for Contrastive Analysis could also be considered. Recently, Florence Carton developed a GAN-based Contrastive Analysis method [42] that could potentially overcome CA-VAEs in terms of expressiveness of the representations and quality of the generated image reconstructions. Besides, the use of (Latent) Diffusion Models [236, 124] could also be considered for Contrastive Analysis, as these methods generate images of high quality and fidelity. However, research remains to be led to investigate the quality of such methods' representations in the context of psychiatry-related neuroimaging data.

Encourage disentanglement to enhance interpretability. This manuscript has explored the search for **interpretable** and **distinct** *separated* pathological *and* healthy latent factors of generation. The motivation for learning separated representations of the data (pathological and healthy) was to enhance their interpretability. In this section, an emphasis has been placed on interpretability to improve the applicability of deep learning methods in neuroimaging.
To further increase the interpretability of deep learning methods, this paragraph argues that future works should also focus on improving the disentanglement within each representation (healthy or pathological). Indeed, as suggested by Bengio et al. [30], and Chen et al. [47], disentangled representations offer improved interpretability in unsupervised representation learning

by disentangling underlying factors of variation within the data. This disentanglement allows individual components of the representation to correspond more directly to distinct features or attributes present in the input data. Consequently, disentangled representations empower practitioners to grasp the essential characteristics captured by the model and to discern how each of them contributes to the overall representation, thereby enhancing interpretability in representation learning tasks.

To produce disentanglement, several strategies could be considered. Notably, in the Variational Auto-Encoders (VAEs) literature, unsupervised disentangling regularization has been proposed and discussed in [122, 35, 256, 325, 49, 9, 167] in an unsupervised manner based on statistical independence. With a more general standpoint, in 2021, Peebles et al. [219] proposed a Hessian penalty loss, encouraging unsupervised disentanglement for any method equipped with a deep generator. From a slightly different perspective, weakly-supervised VAEs [199, 151, 192, 142] proposed integrating a supervised signal in their training. These methods would potentially allow conditioning either the common space latent dimensions or the salient space latent dimensions with fine-grained characteristics, if available. For example, the common space could be conditioned with demographic information such as education, urbanicity, body mass index, literacy, or family situation, and the salient space could be conditioned with clinical pathological scale measures such as *e.g.* delusion score, paranoia score, anhedonia score, anxiety score. In the Contrastive Learning paradigm, supervising the salient latent space dimensions with independent attributes has already been proposed previously in [181] and in the manuscript in Sec. 4.2.5 and could be further investigated.

Appendix A

Background

Contents

In this Chapter, we introduce the pivotal ideas that motivate and allow understand this thesis. First, we detail how the neuroimaging data acquisition procedures developed throughout the recent decades have provided a unique window into the brain's structure and function. Likewise, we describe how the data pre-processing and features extraction tools have permitted drawing significant group-level correlations between neuro-anatomical biomarkers and clinical psychiatric measures. Then, we detail a range of useful and relevant Machine Learning algorithms for individual-level prediction, which will served further in the manuscript.

A.1 Acquiring data in neuroimaging

In neuroimage acquisition, multiple non-invasive in-vivo modalities can provide a window to the underlying biological processes in the brain. In terms of spatial imaging, modalities of interest are (CT) scans, computed tomography scanners, dMRI (Diffusion MRIs), and sMRIs (structural MRIs). Regarding functional imaging, which enables visualizing the brain's physiological activity through time, modalities of interest are PET scans (Positron emission tomography), and fMRI (functional MRI). Even though these acquisition techniques are non-invasive, PET and CT scans still involve ionizing radiations, which is not the case for MRIs, making it a safer option for repeated imaging. Let us briefly describe these techniques of interest.
Functional MRI, invented in 1990 by Seiji Ogawa [212] measures changes in blood oxygenation to infer neural activity. During tasks or rest, there is an increased demand for oxygenated blood in active brain regions. fMRI detects this change by utilizing the blood-oxygen-level-dependent (BOLD) contrast. Subjects perform tasks while inside the MRI scanner, and changes in BOLD signal intensity are recorded over time. Functional MRI produces dynamic images reflecting brain activity. These images are often displayed as activation maps, indicating regions with increased neural activity during specific tasks or resting states.
Diffusion MRI [165] measures the diffusion of water molecules in tissues, providing insights into the brain's microstructure. By applying gradients to the magnetic field, dMRI captures the random motion of water molecules. This information is then used to create diffusion tensor imaging (DTI) maps, revealing the orientation of white matter tracts. Diffusion MRI provides valuable information about the connectivity and integrity of white matter pathways in the brain. DTI maps visualize the directionality of water diffusion, offering insights into the organization of neural fibers.
In this thesis, we focus on neuroanatomical MRIs. Noteworthy, the methodologies we developed are general enough to be applied to other modalities.

A.1.1 Structural MRI (sMRI)

Neuroanatomical MRI, invented in the 70s [193], provides detailed images of the brain's structure, revealing diverse components such as gray matter, white matter, and cerebrospinal fluid. The imaging process involves placing the subject within the strong magnetic field of the MRI scanner. By applying an oscillating magnetic field known as radiofrequency current (RF), the scanner stimulates hydrogen protons in water molecules within brain tissues. The pulsing RF current disrupts the alignment of protons, causing them to spin out of equilibrium. When the RF field is turned off, the MRI-receiving coils detect the energy released as protons realign with the magnetic field, allowing estimation of the time taken for protons to return to their equilibrium state.

Different tissues in the imaged brain exhibit independent relaxation processes of T1 (magnetization in the same direction as the static magnetic field) and T2 (transverse to the static magnetic field). For a T1-weighted image, magnetization recovery occurs before measuring the MR signal by adjusting the repetition time (TR). In contrast, for a T2-weighted image, magnetization decay occurs before measuring the MR signal by adjusting the echo time (TE). T1-weighted and T2-weighted scans are the most common neuroanatomical MRI sequences. T1-weighted MRI, utilizing short TE and TR times, enhances fatty tissue signals while suppressing water signals, providing excellent contrast for highlighting anatomical details. This makes it particularly valuable for identifying diseases related to gray matter, including demyelinating diseases, dementia, cerebrovascular disease, infectious diseases, Alzheimer's disease, and epilepsy. On the other hand, T2-weighted MRI, with longer TE and TR times, enhances water signals, aiding in detecting edema, tumors, hemorrhage, inflammation, and revealing white matter lesions.

In this thesis, our objective is to develop methodologies suitable for parsing the heterogeneity of mental disorders given neuroimaging data. To strengthen our contributions, we develop our methods so that they generalize to different types of input modality. Nevertheless, in this thesis, we primarily focus on neuroanatomical T1w MRI images of the brain for stratifying and parsing mood and psychosis mental disorders such as schizophrenia, bipolar disorder, and autism. To justify this choice, let us draw up a non-exhaustive list of group-level evidence of gray matter-based neuroanatomical deviating patterns that correlate well with psychiatric diagnoses documented in the literature. But first, let us describe the pre-processing tools required to derive and obtain input images with an enhanced signal-to-noise ratio (SNR).

A.1.2 The anatomy of the human brain

Foremost, let us provide a brief and coarse overview of the important structures that constitute the brain. The brain floats in the Cerebrospinal Fluid (CSF), which surrounds the brain and fills its ventricles. It is a translucid fluid composed of water, electrolytes, proteins, and other substances. CSF protects the brain and spinal cord from physical shocks and helps maintain a stable environment for the brain by removing waste products and supplying nutrients.
Besides the CSF, the brain can be decomposed into three areas: the brain stem, the cerebellum (10% of the brain weight), and the cerebrum (85% of the brain weight). The cerebellum and the brain stem are responsible for autonomic processes such as keeping a stable heart rate, breathing, and coordinating movements. Conversely, the Cerebrum is responsible for high-level cognitive tasks such as information processing, memory, emotions, learning, and decision-making. The cerebrum is connected by the brainstem to the spinal cord and can be divided into two cerebral hemispheres connected by the corpus callosum. Each hemisphere has an inner core composed of white matter and an outer layer, the cerebral cortex, composed of grey matter. Coarsely, white matter is primarily composed of myelinated axons and acts as a communication network, facilitating the transmission of signals between different regions of the brain and connecting the various gray matter areas. Grey matter consists mainly of neuron cell bodies, dendrites, and unmyelinated axons and is involved in processing and integrating information, including sensory perception, muscle control, and higher cognitive functions
Across the years and decades, the cerebral cortex has been decomposed into four lobes, which are considered to have six lobes each. The lobes are large areas that are anatomically distinguishable and functionally distinct, with numerous ridges, or gyri, and valleys, the sulci:

1. the frontal lobe, in charge of reasoning and decision-making. It notably includes Broca's area, which is associated with language processing;

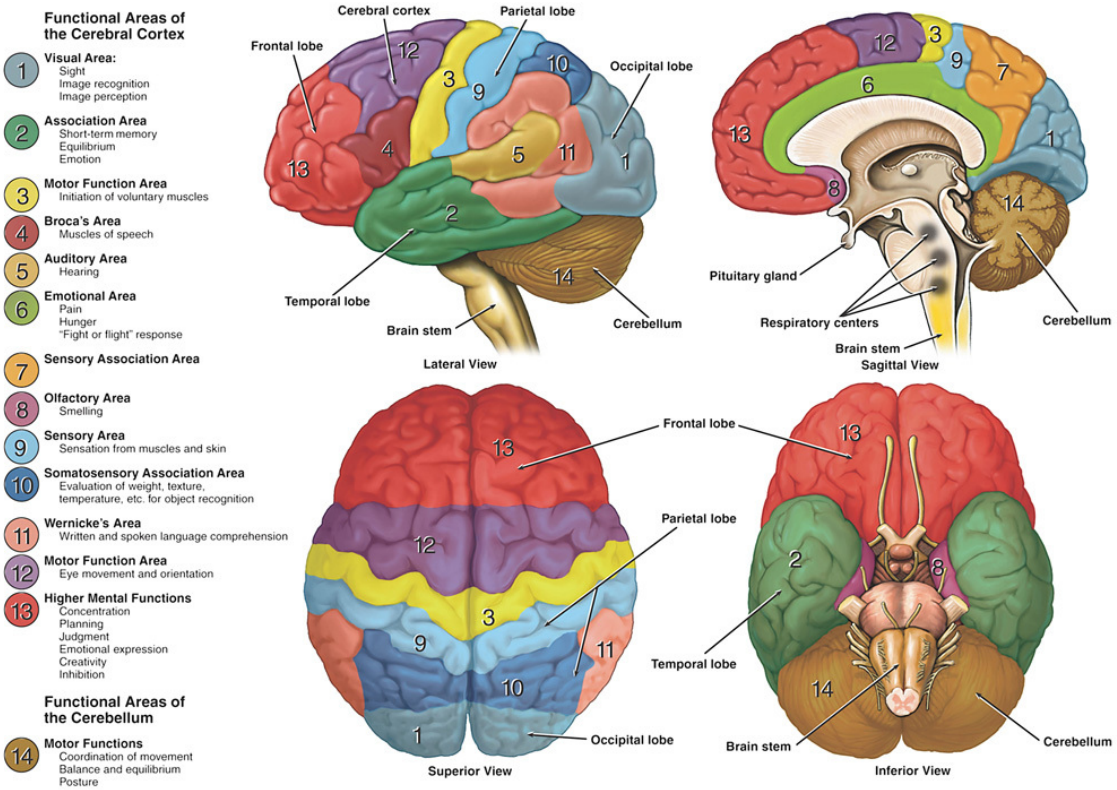
2. the parietal lobe, responsible for sensory integration, visuospatial processing, recognition, and navigation;

3. the occipital lobe, involved with visual processing;

4. the temporal lobe, responsible for short and long-term memory, language processing, and emotion association.

Overall, in recent years, the search and development of finer brain templates have been motivated to quantify the structural variance between individuals. Recently, digital brain templates,

Figure A.1: Scheme of the neuroanatomical atlas that lists well-identified brain regions. Credits to https://dana.org/resources/neuroanatomy-the-basics/.



or atlases, generated from a single-subject or multiple subjects, are replacing conventional printed brain templates (*e.g.*, Talairach and Tournoux atlas [269]). Then, the International Consortium for Brain Mapping (ICBM) adopted MNI-152 as its standard template [200, 201]. This atlas is included in different functional imaging analysis packages, including the statistical parametric mapping package (SPM) and the FMRIB Software Library (FSL) [202]. It has been shown that brain templates using multiple subjects provide a higher signal-to-noise ratio and better contrast between grey matter and white matter. Therefore, the MNI-152 template was designed from 3D brain MRI images of 152 normal subjects [191]. By default, we use this atlas in practice in our pre-processing and feature extraction pipelines.

A.1.3 Data pre-processing tools in neuroanatomical imaging

The pre-processing pipeline for structural MRI (sMRI) involves a well-defined series of steps designed to transform raw images into stripped and registered brain images with an enhanced

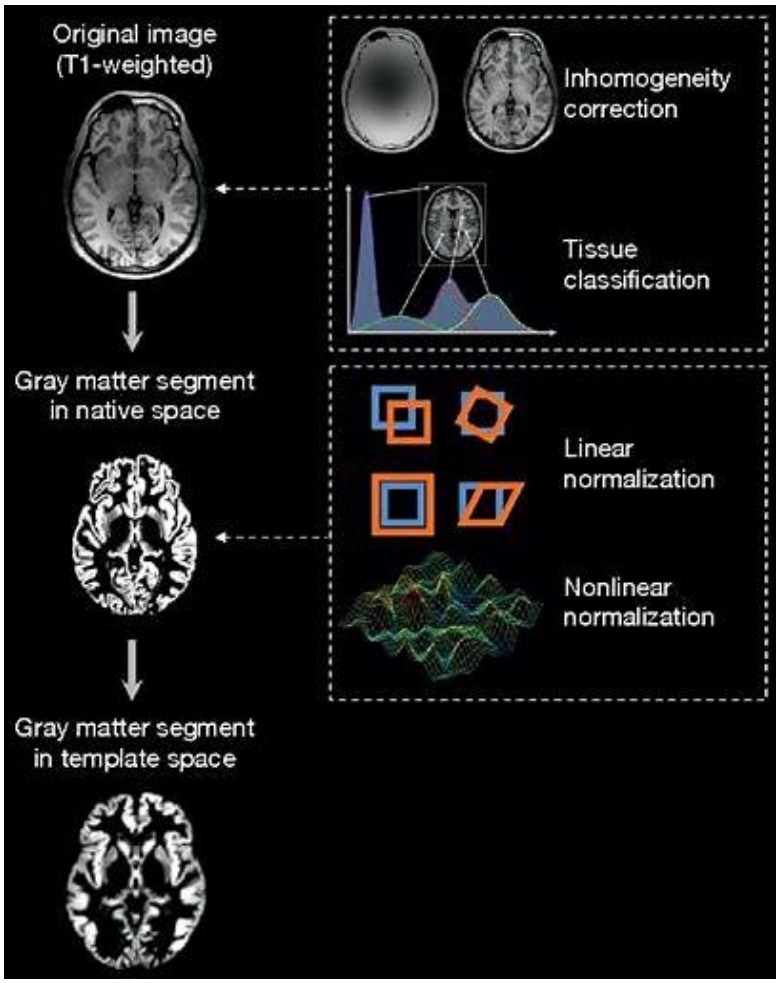
signal-to-noise ratio appropriate for further analysis. As described in detail by, the standard pre-processing procedure generally includes 1) intensity normalization, 2) skull stripping, 3) tissue segmentation, 4) spatial normalization, and 5) intensity modulation.

Step 1) aims at cleaning up certain artifacts from raw scans, such as bias fields. The bias field is a low-frequency spatially varying MRI artifact resulting from spatial inhomogeneity of the magnetic field, variations in the sensitivity of the reception coil, and the interaction between the magnetic field and the human body. The bias field artifact depends on the strength of the magnetic field and causes a smooth signal intensity variation within tissue of the same physical properties. Step 2)'s objective is to separate brain tissues from the non-brain structures such as fat, skin, muscles, eyes, and bones. Then, given brain voxels only, step 3) further categorizes voxels into three distinct categories, the gray matter (GM), the white matter (WM), and the cerebrospinal fluid (CSF), providing a probability density for each tissue type in every voxel. Then, step 4), involves a two-step transformation: a linear component addressing global alignment and a non-linear deformation (e.g., using DARTEL [19]), ensuring local alignment of brain structures. This spatial normalization step consists of locally expanding and contracting the brain regions. As a result, the normalized image needs to be scaled by the amount of contraction to preserve the total amount of Gray Matter. To ensure it, we apply step 5) called intensity modulation. In practice, it consists of multiplying the normalized image by the Jacobian of the transformation. Ultimately, if the global brain size is not of interest, we apply a proportional scaling according to the individual Total Intracranial Volume (TIV), as post-processing, to fully modulated images as suggested in [99]. In this thesis, we consistently apply a VBM pre-processing performed [17] with the Computational Anatomy Toolbox (CAT12) [98, 99] of Statistical Parametric Mapping (SPM) [18]. VBM pre-processing outputs with around 300,000 GM voxels at a resolution of $1.5mm^3$ that quantify the amount of local Gray Matter volume at each voxel.

A.1.4 Evidence of neuro-anatomical patterns in psychiatry

In this thesis, we focus our research on gray matter neuroanatomical deviations captured with T1w MRI acquisition sequence. As a reason, recent years of research on psychiatric conditions have shown that structural alterations in neuroanatomical gray matter have been consistently observed in various mental disorders, such as as schizophrenia, bipolar disorder, and autism spectrum disorder.

Figure A.2: Scheme of an example of a neuroimaging pre-processing pipeline. In step 1), the inhomogeneity aims to correct bias field artifacts. In Steps 2) and 3), non-tissue matter is stripped from tissue matter while different types of tissue matter (GM, WM, and CSF) are segmented. In this particular example, the pipeline conserves only the Gray Matter of the brain. In step 4) the brain is first linearly warped onto a template (*e.g.* MNI-152) and then non-linearly deformed and aligned onto the template. An intensity modulation step is performed to preserve the total amount of Gray Matter. Typically, this step consists of reducing the intensity of the voxels in the brain regions that have been expanded and increasing the intensity of the voxel in the regions that have been shrunk during the template non-linear registration. Credits to [161].



Schizophrenia disorder: For schizophrenia, early studies [141] have reported enlargement of the lateral ventricles based on a comparison between 17 chronic schizophrenic patients and age-matched controls. Empowered by the recently developed pre-processing and template-warping techniques such as Voxel-Based Morphometry (VBM) and Regions-Of-Interest (ROI), several

analyses [255, 244, 163] have revealed that schizophrenia was associated with a global brain volume reduction, particularly in frontal and temporal lobes. A meta-analysis, dating back from 2005 [126], generally outlined gray and white matter deficits in patients with schizophrenia, compared to healthy comparison subjects, in a total of 50 brain regions, with particularly consistent gray matter deficits in specific brain regions, such as the left superior temporal gyrus and parahippocampal, the inferior frontal gyrus, the left superior temporal gyrus and the left medial temporal lobe. Also, sub-cortical structures, notably including the amygdala, hippocampus, thalamus, and accumbens have been shown to exhibit decreased volumes in individuals with schizophrenia in 2016, in a recent ENIGMA large-scale analysis [282], with positive associations found between the volume of certain structures and the duration of illness and age.

Bipolar disorder: Similarly, for bipolar disorder (BD), consistent gray matter alterations have been noted in sub-cortical structures, with studies that date back from 2012 and 2014 revealing differences in hippocampal, thalamic, and amygdala volumes compared to healthy controls [114, 163, 198]. In 2016, a larger study [120], the ENIGMA consortium also reported lower hippocampal, amygdala, and thalamic gray matter volumes in BD individuals and higher bilateral ventricular volumes, as in schizophrenia. In terms of cortical regions, the same study also reported reduced cortical thickness in the anterior cingulate, para-cingulate, superior temporal gyrus, and prefrontal regions. In 2018, gray matter volume thinning was again reported by the ENIGMA consortium in the frontal, temporal, and parietal regions symmetrically in brain hemispheres [121]. Interestingly, [172] also observed gray matter atrophies in frontal and temporal regions of the bipolar group compared to a healthy group and also reported it in schizophrenia which potentially indicates some common underlying pathological processes. Recently, in 2022, an additional large-scale study by the ENIGMA consortium [54] confirmed these findings while expanding the understanding of cortical thinning to include regions such as the inferior parietal, fusiform, and inferior temporal regions associated with disruptions in sensory processing.

Autism disorder: In autism disorder, relatively early studies have identified significant differences in terms of gray matter volume using an approach per regions of interest (ROI). These studies underscored gray matter deviations as well as correlations with clinical severity in the parietal, frontal and limbic regions but also in the basal ganglia and the cerebellum ([204, 235, 290]). Some of these findings were later also observed by Ecker et al. in 2010 [80]

and more recently, in 2022 [81], where deviations were observed in the limbic, frontal-striatal, frontotemporal, frontoparietal and cerebellar systems, but also the cingulate region. However, the authors explain that some results of many studies are in disagreement, and explain that these are due to the heterogeneity of the autism spectrum disorder, which points out the need for more individualized, reliable, and heterogeneity-aware biomarkers discovery methods.

These neuroanatomical findings provide valuable insights into the structural biomarkers associated with mental disorders, contributing to a better understanding of their etiology. Furthermore, even though these analyses have been led at the group level, they provide compelling promises about developing a diagnostic biomarker predictor at an individual level.

A.2 Learn statistical patterns with Machine Learning

Machine learning (ML) is a powerful computational approach that enables systems to learn patterns and predict from individual inputs. At its core, machine learning involves developing and studying statistical algorithms that can effectively generalize and thus perform tasks without explicit instructions. A pivotal concept in machine learning is generalization. Generalization refers to the ability of a learning machine to perform accurately on new, unseen examples/tasks after having fitted a learning data set. The model must achieve robust generalization to provide accurate predictions in real-world scenarios.
Machine learning is particularly suited for individual prediction and knowledge discovery due to its capacity to identify subtle patterns within large and complex datasets. By leveraging diverse features, machine learning models can uncover statistical relationships that may be disregarded by group-level analytical methods. This makes Machine Learning particularly suitable for inferring predictions based on individual characteristics, such as predicting disease outcomes based on unique patient profiles. Moreover, machine learning can also be used in knowledge discovery by drawing insights from data as it can identify relevant associations between input features and variables of interest.

A.2.1 Linear Supervised Learning

Supervised learning is a type of machine learning where the algorithm is trained on a labeled dataset. In this approach, the algorithm learns to map input data to corresponding output labels by observing examples in the training set. Each example in the dataset consists of input-

output pairs, with the algorithm adjusting its internal parameters to minimize the difference between the predicted outputs and the ground truth labeled outputs. Supervised learning aims to enable the algorithm to make accurate predictions on new, unseen data by generalizing patterns learned from the labeled training data. This method is particularly useful for tasks such as classification, where the algorithm is trained to categorize inputs into predefined classes, or regression, where it predicts numerical values based on input features.

Multinomial logistic regression

In classification, logistic regression stands out for its simplicity, interpretability, and efficiency in multiclass classification tasks. Logistic regression is a statistical method widely used in machine learning for predicting the probability of an instance belonging to a specific class. The multinomial logistic regression algorithm models the relationship between independent variables (features, denoted by X, an input matrix of size $[N, D]$) and the probability of an outcome (label, denoted by Y, a one-hot label matrix of size $[N, K]$). The logistic regression model applies the softmax function to a linear combination of input features, where each feature is weighted by a corresponding coefficient (stacked in a weight matrix W, of size $[D, K]$). The weighted sum, augmented by a bias term (denoted by b, a weight vector of size $[K]$), is then transformed through the logistic function to produce a probability score. Given an input x_i of size D, the logits are computed as follows:

$$\text{Logits} = Wx_i + b \tag{A.1}$$

The predicted output for the k-th class is obtained by transforming logits into probabilities using the softmax function:

$$\hat{y}_k = \frac{e^{\text{Logits}_k}}{\sum_{j=1}^{K} e^{\text{Logits}_j}} \tag{A.2}$$

Maximum Likelihood Estimation: To quantify the difference between predictions and ground truth labels, a cost function known as "cross-entropy" is defined over the entire dataset:

$$H(Y, \hat{Y}) = -\sum_{i=1}^{N} \sum_{k=1}^{K} y_{i,k} \cdot \log \hat{y}_{i,k} \tag{A.3}$$

This cost function can be retrieved from a statistical approach by looking for the parameters θ

that maximize $\mathcal{L}(\theta)$ the log-likelihood of input data.

$$\mathcal{L}(\theta) = \log \prod_{i=1}^{N} p(y_{i,k}|x_i;\theta) \tag{A.4}$$

Taking the logarithm of the likelihood simplifies the product into a sum:

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \log p(y_i|x_i;\theta) \tag{A.5}$$

Assuming a generalized Bernoulli distribution for K categorical outcomes. We model the parametric distribution of $p(y_{i,k}|x_i;\theta)$ as:

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \log \left(\prod_{k=1}^{K} p(y_{i,k}|x_i;\theta)^{\delta_{k,y_i}} \right) \tag{A.6}$$

Again, using the logarithm properties simplifies the equation into:

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \sum_{k=1}^{K} \delta_{k,y_i} \log p(y_{i,k}|x_i;\theta) \tag{A.7}$$

Now, assuming that $p(y_{i,k}|x_i;\theta)$ is equal to our parametric estimation $\hat{y_{ik}}$, we obtain the usual "cross-entropy" cost function used in Machine Learning.

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \sum_{k=1}^{K} y_{i,k} \log \hat{y}_{i,k} = -H(Y,\hat{Y}) \tag{A.8}$$

Gradient descent: In the pursuit of optimizing the parameters of our multinomial logistic regression model, we turn to the iterative power of gradient descent. Having established the foundation of our model, characterized by the weight matrix W and bias vector b, the next step is to minimize the cross-entropy cost function $H(Y,\hat{Y})$. This function gauges the disparity between the predicted probabilities $\hat{Y}$ and the actual class labels Y, with Y representing the true class labels in a one-hot encoded form. To navigate this optimization landscape, we compute the gradients of the cost function with respect to the weight matrix W and the bias vector b. The gradients are expressed as follows:

$$\nabla_W H(Y,\hat{Y}) = -\frac{1}{N} X^T \cdot (Y - \hat{Y}) \tag{A.9}$$

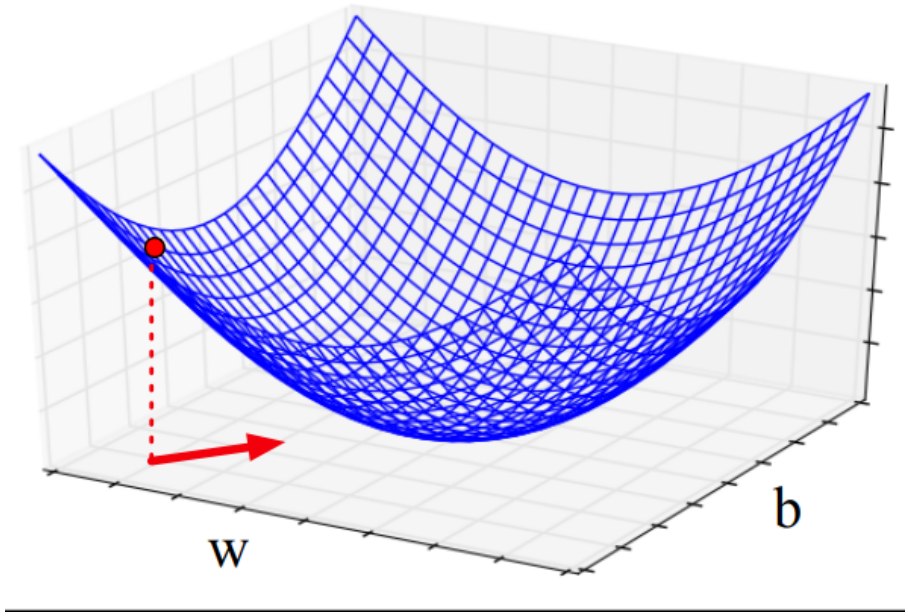
$$\nabla_b H(Y, \hat{Y}) = -\frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i) \tag{A.10}$$

Given the computation of these gradients, we use an iterative optimization process to update the model parameters, guided by the negative gradient's direction and influenced by a learning rate α. The updating equation for updating the parameters $\theta = (W; b)$ based on the gradient is expressed as follows:

$$\theta_{t+1} = \theta_t - \alpha \nabla_{\theta_t} H(Y, \hat{Y}) \tag{A.11}$$

This optimization process converges towards the optimal parameters W^* and b^*, fine-tuning our model parameters to minimize the classification error.

Figure A.3: Scheme of gradient descent optimization process. The blue manifold plots the expression of the loss function $H(Y, \hat{Y})$ for different values of $\theta = \{W, b\}$. At the red point, the red arrow in the x-y plane points in the direction that minimizes the loss function: the opposite direction of the gradient. Image taken from https://web.stanford.edu/~jurafsky/slp3/5.pdf



Linear regression

In regression tasks, linear regression is a key technique. Unlike classification, which aims to predict discrete class labels, linear regression is employed when predicting continuous numerical values. The algorithm models the relationship between input features (X, an input matrix

of size $[N, D]$) and the output variable (Y, a vector of size $[N]$) through a linear equation $Y = XW + b$. Here, W represents the weight matrix of size $[D, 1]$, and b is a bias term.

Maximum Likelihood Estimation: In linear regression, the maximum likelihood estimation seeks parameters (W, b) that maximize the likelihood of the observed data given the model. Assuming that the parametric estimated distribution $p(y|x;\theta)$ follows a Gaussian distribution, the likelihood function is defined as:

$$\mathcal{L}(\theta) = \prod_{i=1}^{N} p(y_i|x_i;\theta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-(Wx_i+b))^2}{2\sigma^2}} \tag{A.12}$$

Applying the logarithm simplifies the product into a sum:

$$\log \mathcal{L}(\theta) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - (Wx_i + b))^2 \tag{A.13}$$

Mean Square Error (MSE) as Cost Function: The negative log-likelihood, which is proportional to the mean square error, serves as the cost function to be minimized during training:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - (Wx_i + b))^2 \tag{A.14}$$

Analytical solution: Given that $y \in \mathbb{R}$ is continuous, the MSE cost function fueled with a linear estimation model constitutes a convex objective. This loss function is known as the Ordinary Least Squares (OLS) regression, and its solution is given by:

$$\theta^* = \text{argmin } \log \mathcal{L}(\theta) = (X^T X)^{-1} X^T Y \tag{A.15}$$

The OLS regression provides a closed-form solution for the optimal parameters θ, making it computationally efficient and widely used for linear regression tasks.

A.2.2 Linear Unsupervised Learning

Unsupervised learning represents a distinctive paradigm in machine learning where the algorithm is entrusted with extracting patterns and structures from unlabeled data. In contrast to supervised learning, there are no explicit output labels to guide the learning process. Instead,

the algorithm explores inherent structures within the input data, aiming to uncover hidden relationships, cluster structures, latent independent dimensions, or latent distinct and interpretable generative factors.

Clustering unlabelled data with K-Means

The k-means algorithm [175, 93], a pivotal method in unsupervised learning, is designed for clustering data into distinct groups based on similarity. The objective of k-means is to identify K centroids, each representing a cluster, and assign data points to the nearest centroid, iterating until convergence. The algorithm's simplicity and efficiency lie in its objective to minimize the within-cluster sum of squares. Interestingly, the k-means algorithm closely relates to Gaussian mixture models (GMMs) [203], specifically in the spherical isotropic case. While k-means simplifies the complexities of GMMs, it offers a computationally efficient means of identifying cluster structures in unlabeled data.

Maximum Likelihood Estimation (MLE): As for supervised methods, the objective function of K-Means can be linked to Maximum Likelihood Estimation. Let X be the dataset, C the latent cluster assignments with k clusters, and θ the parameters (centroids). We look for the parameters θ that maximizes the log-likelihood of observed data X:

$$\mathcal{L}(\theta) = \log \prod_{i=1}^{N} p(x_i|\theta) \tag{A.16}$$

From there, we can introduce the clustering assignment distribution by applying the equality $p(x_i) = \sum_{k=1}^{K} p(x_i|c_i = k, \theta)p(c_i = k|\theta)$. Also, we can use the logarithm to simplify into:

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \log \sum_{k=1}^{K} p(x_i|c_i = k, \theta)p(c_i = k|\theta) \tag{A.17}$$

Given the assumption that $p(c_i = k|\theta)$ follows a spherical isotropic Gaussian distribution with standard deviation σ and mean μ_k, the probability of observing a data point x_i knowing that it belongs to the cluster c_i is modeled as:

$$p(x_i|c_i, \theta) = \frac{1}{(2\pi\sigma^2)^{\frac{D}{2}}} \exp\left(-\frac{\|x_i - \mu_k\|^2}{2\sigma^2}\right) \tag{A.18}$$

We therefore obtain a log-likelihood function equal to:

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \log \sum_{k=1}^{K} p(c_i = k|\theta) \frac{1}{(2\pi\sigma^2)^{\frac{D}{2}}} \exp\left(-\frac{\|x_i - \mu_k\|^2}{2\sigma^2}\right) \tag{A.19}$$

For simplicity, in K-Means (and spherical Gaussian Mixture), it is assumed that σ takes the same value across all clusters.

Expectation-Maximization (EM) Optimization The Expectation-Maximization [209] algorithm optimizes the likelihood function iteratively through Expectation and Maximization steps. In the E-step, the algorithm computes the expected value of the log-likelihood with respect to the current estimates of parameters. For k-means, it is equivalent to re-estimating the cluster assignment $p(c = k|\theta)$ by assigning data points to the nearest centroid.

$$p(c_i = k|\theta) = 1 \text{ if } \exp\left(-\frac{\|x_i - \mu_k\|^2}{2\sigma^2}\right) \geq \exp\left(-\frac{\|x_i - \mu_j\|^2}{2\sigma^2}\right) \ \forall j \neq k, \text{ 0 else.} \tag{A.20}$$

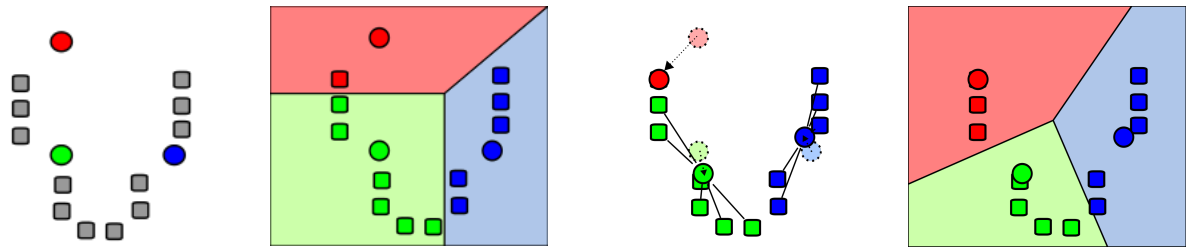
In the M-step, the algorithm maximizes the expected log-likelihood, updating the parameter estimates. For k-means, this step includes recalculating the centroids $\theta = \{\mu_k\}_{k=1}^{K}$ based on the newly assigned data points.

$$\mu_k = \frac{\sum_{i=1}^{N} p(c_i = k|\theta) \cdot x_i}{\sum_{i=1}^{N} p(c_i = k|\theta)} \tag{A.21}$$

These E and M steps iteratively continue until convergence, aligning with the MLE objective of maximizing the likelihood function. In summary, the k-means algorithm, driven by MLE and the EM optimization technique, efficiently discovers cluster structures within unlabeled data, iteratively refining cluster assignments and centroids.

A.3 Learn complex statistical patterns with Deep Learning

In the previous section, we introduced traditional Machine Learning techniques to classify or regress samples in a supervised manner (*e.g.*, logistic regression, linear regression, and support vector methods, etc...) or to discover homogeneous structure (*i.e:* clusters) among a set of data points (*e.g.*, k-means [175, 93], Gaussian mixture models [203], etc...). Even though these methods are well-studied, interpretable, and theoretically grounded, they generally require a limited set of human-defined, well-chosen features as input. Indeed, linear models tend to fail on raw, structured, high-dimensional data inputs. As a reason, the complexity of the decision



(a) K "prototypes" are randomly initialized.

(b) Associate observations with the nearest prototype.

(c) The centroid of each of the k clusters becomes the new prototype.

(d) Steps 2 and 3 are repeated until convergence has been reached.

Figure A.4: Scheme of the K-Means algorithm, taken from https://en.wikipedia.org/wiki/K-means_clustering.

function is limited to a linear function only, which can fail to capture a higher level of pattern complexity.

In contrast, deep learning offers a more parameterized and powerful approach to handling raw, structured, and high-dimensional data. Deep learning models, particularly neural networks, introduce a hierarchy of learned features, allowing them to automatically extract complex patterns and representations from input data. Unlike linear models, which are constrained to linear decision boundaries and may struggle with complex patterns, deep learning architectures employ multiple layers of non-linear transformations. Deep learning excels at automatically learning input representations, eliminating the need for user-based feature extraction engineering. The development of Deep Learning models has enabled improved performance in various domains, including image and speech recognition and natural language processing. As an example, the emergence and development of Deep Learning has enabled the achievement of a classification rate of 95% on the ImageNet benchmark [61], launched in 2009. The three pivotal keys that have fueled this success are the development of better computational resources (notably by letting deep neural networks trained on Graphics Processing Unit (GPU), introduced in 2005, in [263]), data availability (ImageNet dataset has grown from 3 million images in 2009 to 14 million images recently), and the architecture search of the deep neural networks (associated with an increase in the number of learnable parameters).

A.3.1 Deep neural network optimization

In the context of deep neural networks, the optimization process plays a crucial role in the training process. As in traditional machine learning techniques such as logistic regression or linear regression, the training of these networks involves minimizing a chosen loss function,

which, in the supervised case, for example, quantifies the disparity between predicted and actual outputs across the entire dataset. Interestingly, we can derive the same loss functions for classification and regression as with logistic and linear regression by deriving them from the Maximum Likelihood Estimation. The only difference lies in the output estimation, which depends on a (possibly over-)parameterized non-linear, generally non-convex function f_θ rather than a linear matrix weight W and bias vector b. Similar to logistic regression, optimization involves a gradient descent iterative process. A widely used optimization technique for this task is the stochastic gradient descent (SGD) [233]. Its weight updates follow the rule:

$$\theta \leftarrow \theta - \alpha \nabla L(\theta) \tag{A.22}$$

where α denotes the learning rate. An essential advantage of SGD is its scalability, as it can learn from large datasets by dividing the data into batches of samples. The gradient $\nabla L(\theta)$ is approximated using each batch in a stochastic manner, contributing to the efficiency optimization process. Indeed, the non-convex and non-linear nature of neural networks poses challenges in finding global minima, but SGD[233], by iteratively updating model parameters in the direction of the negative gradient, provides a scalable solution that is less likely to stagnate in local minima due to its stochastic property. An important feature contributing to the generalization of Deep Neural Networks is its architecture, as it directly impacts the smoothness of the loss landscape and the expressivity of its decision function.

A.3.2 Deep neural network architectures

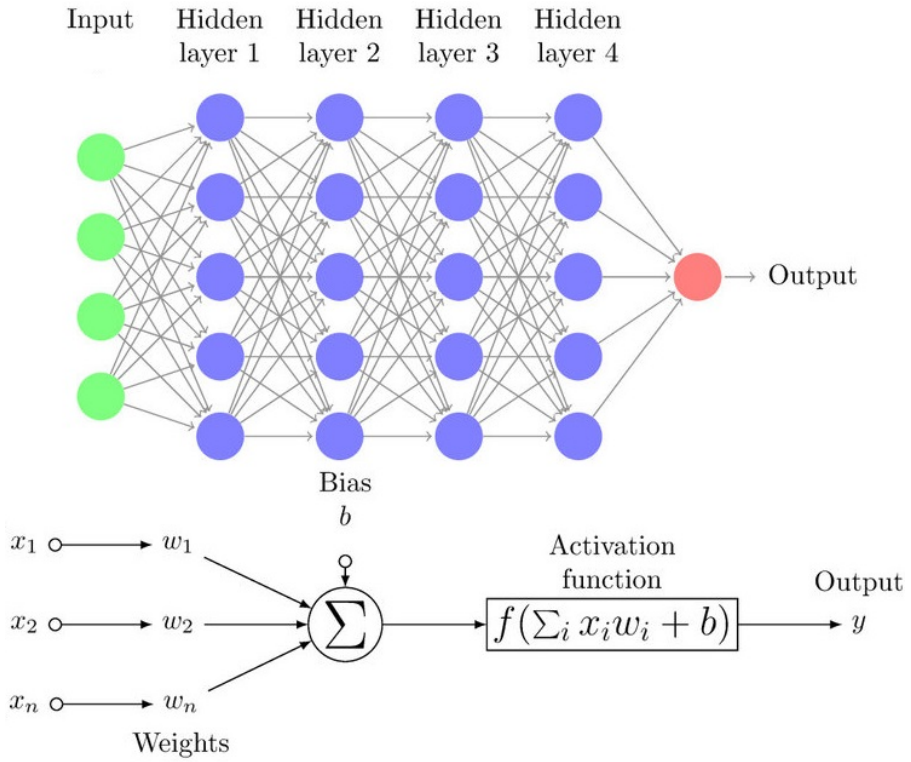
Over the recent years, many endeavors have been invested in neural network architecture design to enhance the expressivity of the decision function while preventing overfitting. In this subsection, we provide a non-exhaustive list of architectures that have enabled performance gains in image recognition and tabular data classification.

Multi-Layer Perceptron

The Multi-Layer Perceptron (MLP) builds upon the concept of the Perceptron, a basic linear model using a non-linear activation function to make binary predictions (0 or 1) from the input. However, MLP takes this idea further by allowing an arbitrary number of layers. We can stack multiple Perceptrons together, forming a composite function $f_\theta = f_{\theta_1} \circ f_{\theta_2} \circ ... \circ f_{\theta_k}$. Each $f_{\theta_i}(x) = \phi(W_i x + b_i)$, where $\phi(x)$ is 1 if $x > 0$ and 0 otherwise (Heaviside activation function). The parameters to be learned are denoted as $\theta_i = \{w_i, b_i\}$, and f_θ represents the decision rule. This

architecture generates multiple intermediate representations after each hidden layer, leading to the final layer where the actual task is performed. The compositional architecture of Multi-Layer Perceptron is particularly interesting to derive compressed representations that capture non-linear feature combinations within the inputs. In practice, other non-linear activation functions, such as the ReLU [104] or the Sigmoid function, were further found to perform better than the Heaviside function.

Figure A.5: Diagram of a Multi-Layer Perceptron (MLP) with four hidden layers which illustrates how to obtain a scalar input given a n dimensional input. One neuron is the result of applying the nonlinear transformations of linear combinations (x_i, w_i, and biases b). Image taken from [82].



Convolutional Neural Networks

LeCun introduced Convolutional Neural Networks (CNNs) in 1989 [166]. These architectures use filter layers designed for processing grid-structured data, such as images. In practice, this type of architecture has demonstrated impressive performance gains, notably in image recognition, which, intuitively, was first attributed to its resemblance with the organization of the animal visual cortex. Unlike Multi-Layer Perceptrons (MLPs), CNNs use convolutional

layers that apply filters to capture patterns of features. These filters enable the network to recognize local patterns like edges and textures. Using them in a hierarchical manner produces a compositional representation of the visual information. Mathematically, this corresponds to rewriting the matrix-vector multiplication $W_k x$ in the k-th layer $f_{\theta_k} = \phi(W_k x + b)$ by a convolution operation $\Omega_k * x$, where Ω_k is the k-th kernel and has a much smaller size than the original matrix $W_k \in \mathbb{R}^{d_{k-1} \times d_k}$. The convolution operation slides the kernel over the grid data, computes element-wise multiplications, and spatially stacks the results. This operation enables the capture of local patterns and reduces the number of learnable parameters of the layer. Let us describe it in detail:

$$f_{\theta_k}(x)_{i,j} = \sum_{a=0}^{d_{k-1}-1} \sum_{b=0}^{d_{k-1}-1} \omega_{ab} y(i+a, j+b) \tag{A.23}$$

where each element is defined as:

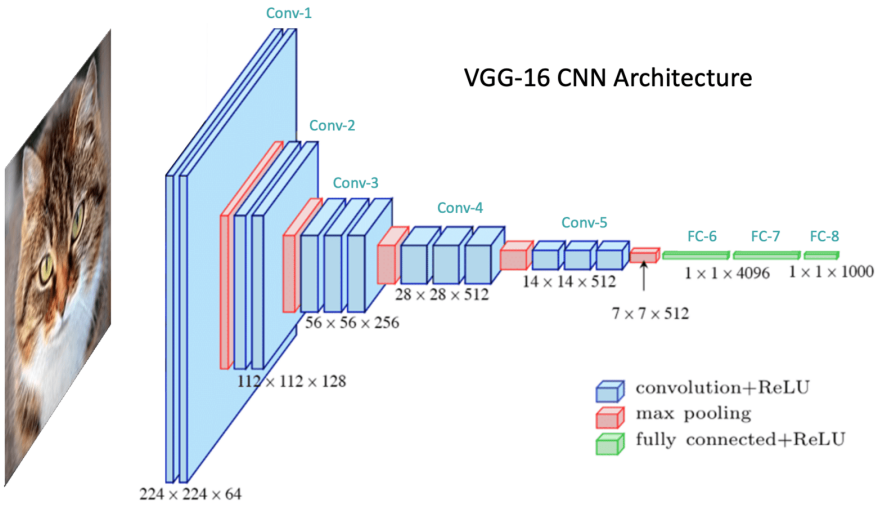
1. $x_{i,j}$ represents the element at position (i, j).
2. ω_{ab} represents the kernel element at position (a, b).
3. $x(i+a, j+b)$ represents the element at position $(i+a, j+b)$.

This formula describes the convolution operation, where a single kernel Ω is applied to the input data at different positions, and the results are summed to produce the output of the convolutional layer. In practice, in a traditional convolutional neural network layer, multiple kernels are used, and resulting filtered features are stacked along the channel dimension, allowing a greater expressivity of the features extractor.

Dense and Residual Neural Networks

Importantly, the stacking of convolutional layers has been a key component of the increase of the performances of convolutional neural networks. However, it has been observed that training very deep convolutional neural networks hardly leads to an optimal solution. As a reason, the *vanishing gradient* problem. This issue has been first observed in 1994 [29] with recurrent neural networks. It was shown that the gradient norms are smaller when the gradient chain rule requires to be applied too many times, typically when the task requires long-range dependencies in the sequential input when using a recurrent neural network. In deep convolutional neural networks, the same optimization problem was observed as the depth

Figure A.6: Scheme of VGG-16 [259], a Deep Convolutional Neural Architecture. Image taken from https://learnopencv.com/tag/convolutional-neural-networks/.



of the convolutional model increases, which leads to frozen first layers weights that are barely modified during the optimization process.

As a solution, Kaiming He et al. [116] introduced residual neural networks in 2016. Residual layers in residual neural networks (ResNets) operate by explicitly fitting a residual mapping. In these networks, the desired underlying mapping is denoted as $h(x)$, and the stacked non-linear layers are encouraged to fit another mapping, $f(x) = h(x) - x$. This technique consists of recasting the original mapping into $f(x) + x$. The key hypothesis behind using residual layers is that estimating the residual mapping $f(x)$ is supposedly more tractable than estimating the original, mapping $h(x)$. Essentially, this approach assumes that it is easier to drive the residual to zero than to train a stack of nonlinear layers to fit an identity mapping, particularly if an identity mapping is considered optimal.

In the same vein, Gao Huang et al. [129] introduced dense neural networks in 2018. They further established a dense connectivity pattern. In this architecture, all layers with matching feature-map sizes are directly connected to each other. To maintain a feed-forward nature, each layer receives additional inputs from all preceding layers and passes its own feature maps to all subsequent layers. Consequently, dense connected layered networks thus require fewer parameters than traditional convolutional networks, as they do not need to relearn redundant feature maps. Besides better parameter efficiency, DenseNets improves the flow of information and the scalability of the gradients throughout the network, as each layer has direct access to the gradients from the loss function.

Universal Approximation Theorem

As previously explained, many endeavors have searched relevant architectures allowing an expressive decision function while not being too prone to overfitting, undesirable behavior that occurs when the model gives accurate predictions for training data but not for evaluation data that have not been fed during the training process. Concerning the expressivity of the decision function, theoretical results have been given as evidence to justify using MLPs and CNNs in practice. In the mathematical theories of Deep Neural Networks, universal approximation theorems are results that enable delineating what an architecture can possibly learn. Notably, Stinchcombe, White, and Hornik [128, 264], showed in 1989 that n-layers (providing that n is greater than 2) MLPs with Sigmoid or ReLU activation functions had the potential to theoretically learn any continuous function between two Euclidian spaces as soon as sufficient number of hidden units are available at each layer. This theorem is extremely relevant as it states that, theoretically, the choice of the MLP architecture can potentially solve any task, as long as it does not overfit, or end up in a local minima during optimization. Indeed, a major weakness of the universal approximation theorem is that while any continuous function can be arbitrarily well approximated in a bounded input region, it does not hold outside of the input region (*i.e.* approximated functions do not extrapolate outside of the input training region), which makes models prone to overfitting. Given these results, a relevant question to be raised would be: if MLPs are universal approximators, why would we need Convolutional and Residual Neural Networks? The answer needs two arguments to be complete. Primarily, it has been recently shown that deep convolutional ([331]) and residual ([171]) neural networks were also able to approximate any continuous function to an arbitrary accuracy provided that their depth is large enough. Secondary, it has been shown that convolutional neural networks are actually more sample efficient than traditional fully connected networks, notably because they are not orthogonal-equivariant or permutation equivariant, as shown in [168]. Despite these results, deep neural architecture used nowadays could theoretically over-fit the training dataset. However, it was observed that these models generalize well on new unseen data. Understanding this behavior remains an open problem as explained in [321].

Deep representation learning An interesting property of deep learning models is their ability to automatically extract compressed representations of the inputs, *i.e.* features. This field of interest so-called "Representation Learning", where practitioners aim at estimating parametric models tailored to learn meaningful and compact representations from high-dimensional data. The objective is reduce the input dimensionality and capture relevant features to facil-

itate downstream tasks such as classification [158, 259], clustering [40, 41, 168, 275, 278] or generation [150, 106].

A.3.3 Variational-Auto-Encoders

Variational Autoencoders (VAEs) [150] are a type of deep generative model that can be used to learn a compact, continuous latent representation of a dataset. They are based on the idea of using an encoder network to map input data points x (*e.g:* an image) to a latent space z and a decoder network to map points in the latent space back to the original data space. Given a dataset $X = \{x_i\}_{i=1}^N$ and a VAE model with encoder $q_\phi(z|x)$ and decoder $p_\theta(x|z)$. The VAE's objective seeks the parameters ϕ, θ to maximize a lower bound of the input distribution likelihood, which is entitled the ELBO (Evidence Lower BOund):

$$\log p_\theta(x) \leq \mathbf{E}_{z\sim q_\phi(z|x)} \log p_\theta(x|z) - KL(q_\phi(z|x)||p_\theta(z)) \tag{A.24}$$

where $p_\theta(x|z)$ is the likelihood of the input space, and $\mathrm{KL}(q_\phi(z|x)||p(z))$ is the Kullback-Leibler divergence between $q_\phi(z|x)$, the approximation of the posterior distribution, and $p(z)$ the prior over the latent space (often chosen to be a standard normal distribution). The first term in the objective function, $\mathbf{E}_{z\sim q_\phi(z|x)} \log p_\theta(x|z)$, is the negative reconstruction error, which measures how well the decoder can reconstruct the input data from the latent representation. The second term, $\mathrm{KL}(q_\phi(z|x)||p(z))$, encourages the encoder distribution to be similar to the prior distribution, which helps to prevent over-fitting and encourages the learned latent representation to be continuous and smooth.

Evidence Lower BOund: Interestingly, let us note that the ELBO objective can be retrieved from the maximization of the log-likelihood of the input distribution in a simple way. Starting with the log-likelihood of the data, $\log p_\theta(x)$, and using the properties of the logarithm, we can express it as follows:

$$\log p_\theta(x) = \mathbf{E}_{z\sim q_\phi(z|x)} \log p_\theta(x) = \mathbf{E}_{z\sim q_\phi(z|x)} \log \frac{p_\theta(x,z)}{p_\theta(z|x)} \tag{A.25}$$

Applying the log product rule, introducing the auxiliary distribution $q_\phi(z|x)$ and rearranging terms, we get:

$$\log p_\theta(x) = \mathbf{E}_{z\sim q_\phi(z|x)} \log p_\theta(x|z) - \mathrm{KL}(q_\phi(z|x)||p_\theta(z)) + \mathrm{KL}(q_\phi(z|x)||p_\theta(z|x)) \tag{A.26}$$

Considering that the latest term is positive (KL divergence is always positive or null), we have demonstrated the ELBO inequality introduced in the upper equation. Thus, by maximizing the ELBO, we are effectively maximizing a lower bound on the log-likelihood of the input distribution. This generation method is particularly suited to learning a compressed representation z that can be further investigated to uncover the latent structure of the input data.

VAE loss: Given the previous equation, we can derive the Variational Auto-Encoder loss that is optimized in practice by making assumptions about the form of the distributions during the ELBO derivation. To formulate the Variational Autoencoder (VAE) loss, we combine two key components: the reconstruction error and the Kullback-Leibler (KL) divergence between the prior and posterior distributions.

The reconstruction error quantifies the dissimilarity between the input data x and its reconstructed counterpart $\hat{x}$. This term assesses how well the decoder captures the original data during the generative process. In practice, one usually assumes that $p_\theta(x|z)$ follows an isotropic Gaussian distribution with a fixed diagonal unit covariance matrix and a mean equals to $f_\theta(z)$, where $d_\theta(.)$ is the decoder function. Using a Monte-Carlo estimation to estimate the Expectations, these choices enable to derive the reconstruction term into a Mean Squared Error function between the input data x and its reconstructed counterpart $f_\theta(z)$:

$$\mathbf{E}_{z\sim q_\phi(z|x)} \log p_\theta(x|z) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{l=1}^{L} |||x_i - d_\theta(z_i^l)|||_2^2 \tag{A.27}$$

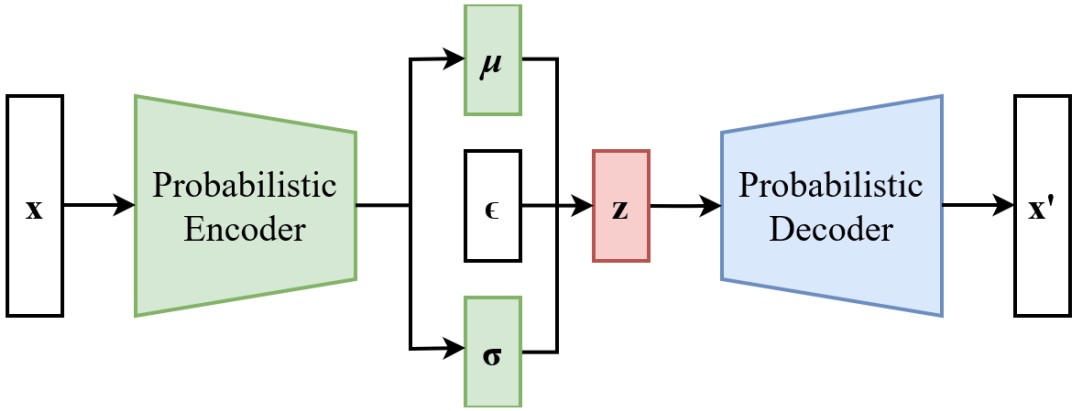
where z_i^l is the l-th latent vector sampled from the posterior distribution $q_\phi(z|x_i)$ via the reparametrization trick. $q_\phi(z|x_i)$ is assumed to follow a Guaussian distribution, whose parameters $(\mu(x_i), \sigma(x_i))$ are inferred by the encoder $f_\phi(.)$.

Simultaneously, the KL divergence between the prior distribution $p_\theta(x)$ and the posterior distribution $q_\phi(z|x)$ measures the discrepancy between these two distributions. For a standard normal prior and a parameterized Gaussian posterior (as described in the previous paragraph), the KL divergence analytical form can be computed as:

$$\mathrm{KL}(q_\phi(z|x)||p_\theta(z)) = \sum_{i=1}^{N} \frac{1}{2}\left[-\log(\sigma_i^2) - 1 + \sigma_i^2 + \mu_i^2\right] \tag{A.28}$$

The derived Variational Autoencoder loss, combining reconstruction error and KL divergence, presents a powerful framework for learning compact and continuous latent representations.

Figure A.7: Scheme of VAE's training process [150]. The probabilistic encoder estimates a mean μ and a standard deviation σ from the input x. These two parameters are regularized to map a normal distribution. Then it samples a latent vector z given μ and σ. Importantly, VAEs employ a reparameterization trick, which consists of computing $z = \mu + \sigma\epsilon$ where ϵ is stochastically drawn from a normal distribution. Given z, the decoder processes the latent vector to produce a reconstruction x', trained to be as close as x possible via a L2-loss.



These two loss terms enable reconstructing input data while respecting a specified prior distribution, making VAEs a versatile and effective tool for latent feature discovery in complex datasets. Moreover, the simple and versatile manner that enables obtaining the Evidence Lower BOund has inspired many works in many different representation learning and generative methods sub-domains such as disentanglement [256, 122, 35, 49], interpretability [9, 325], conditioned generation [167, 151, 142].

A.3.4 Contrastive Representation Learning

In the realm of Representation Learning, a recent class of methods entitled Contrastive Representation Learning (CL) has made remarkable progress in learning representations that encode high-level semantic information about inputs such as images ([318, 297, 21, 118, 52, 108, 77, 275]) and sequential data ([213, 155, 274, 252, 267]). Contrastive Learning methods are particularly tailored for representation learning problems as their encoders capture compact and semantically meaningful representations that generalize well to downstream tasks.

Contrastive Learning (CL) hinges on an intuition that dates back to Becket and Hinton, in 1992 [28], and Bell and Sejnowski in 1995 [31]. Given an input sample x (image or sequence) and two different views (*i.e.,* transformations) v and v^+ of x that potentially overlap (spatially or

sequentially), CL is based on the assumption that v and v^+ should share a similar information content. A parametric encoder f_θ is then estimated by maximizing their "agreement" in the representation space so that their similarity/dependence is preserved in the embeddings $f_\theta(v)$ and $f_\theta(v^+)$. A commonly used measure of agreement is the Mutual Information between the two views embeddings that is maximized: $\theta^* \leftarrow argmax\ _\theta I(f_\theta(v); f_\theta(v^+))$, where the choice of f_θ imposes some structural constraints (*i.e.,* inductive bias). While earlier contrastive learning approaches were confined to specific transformations like rotation, spatial re-organization, or context patching, contemporary methods, such as SimCLR [50], MoCo [118], [52], and CPC [213], have emerged to accommodate diverse data transformations. These methods rely on loss equal (or almost equal) to InfoNCE, which has inspired the representation learning community to pursue efforts in contrastive learning approaches. As an illustration, SimCLR's training process is illustrated in Fig. Importantly, the InfoNCE loss is written as:

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{2N}\sum_{i=1}^{N}\log\left(\frac{\exp(\text{sim}(f_\theta(v_i), f_\theta(v_i^+)))}{\sum_{j=1}^{N}\exp(\text{sim}(f_\theta(v_i), f_\theta(v_j^+)))}\right) + \log\left(\frac{\exp(\text{sim}(f_\theta(v_i^+), f_\theta(v_i)))}{\sum_{j=1}^{N}\exp(\text{sim}(f_\theta(v_i^+), f_\theta(v_j)))}\right) \tag{A.29}$$

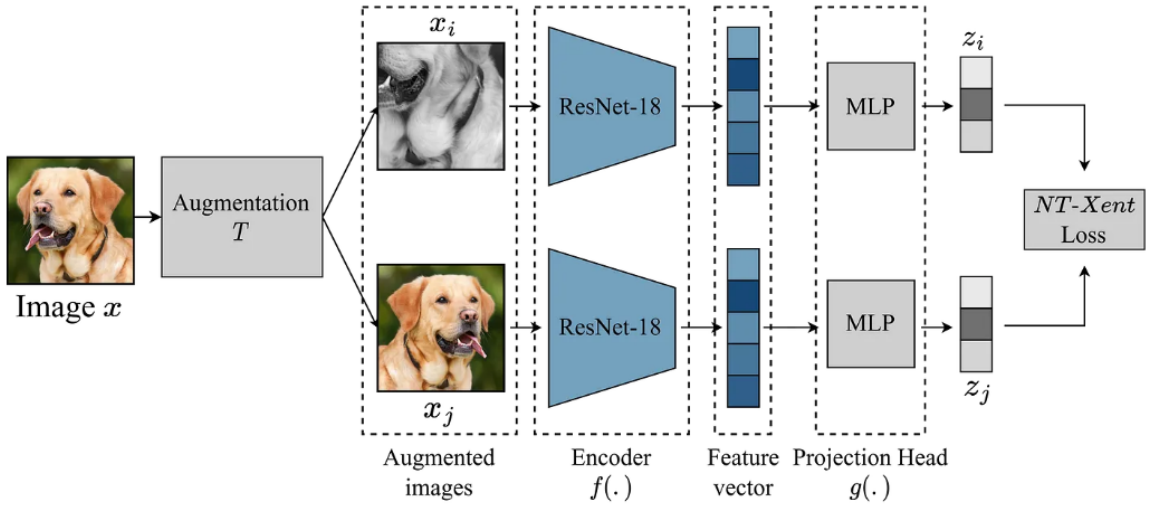
where each term is defined as:

- N is the number of positive pairs (corresponding views of the same sample).
- v_i and v_i^+ are a pair of views for the i-th sample.
- f_θ is the parametric encoder, where $f_\theta(a)$ is generally unit-normalized.
- $\text{sim}(a, b)$ denotes the dot product between the embeddings a and b.

Intuitively, the InfoNCE loss [50], [213] encourages the model to maximize the similarity between positive pairs (i.e., $f_\theta(v_i)$ and $f_\theta(v_i^+)$) while minimizing the similarity with other negative samples in the dataset. This encourages the model to learn representations where positive pairs are more similar than negative pairs, effectively capturing meaningful information from the input data.

Understanding InfoNCE through Alignment and Uniformity: Interestingly, [293] further simplified the InfoNCE loss into an alignment (or reconstruction) and a uniformity (or entropy) term. While the alignment term trains the encoder to assign similar representations

Figure A.8: Scheme of SimCLR's training process [50]. Two distinct data augmentation operators t and t' are sampled from the same family of augmentations T and applied to each image in the batch to obtain two augmented views per example. A base encoder network $f(.)$ and a projection head $g(.)$ are trained to maximize agreement using a contrastive loss. After completing training, we throw away the projection head $g(.)$ and representations produced by the encoder $f(.)$ as inputs for downstream tasks. Credits to https://medium.com/mlearning-ai/self-supervised-pre-training-with-simclr-79830997be34.



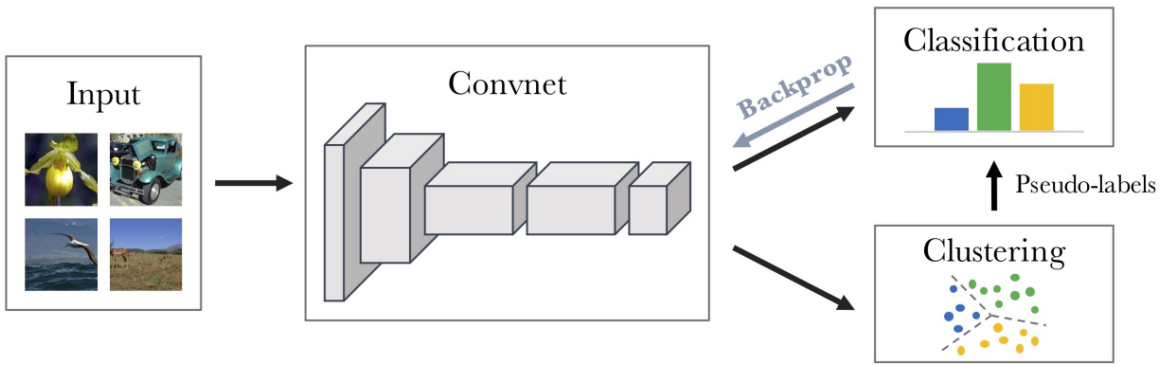
to positive views, the uniformity term encourages feature distribution to preserve maximal information *i.e.:* maximal entropy, that is, it encourages feature distribution toward a uniform distribution in a hyperspherical space. This analysis is interesting as it enables us to intuitively understand the behavior of such a method. Additionally, it intuitively provides insights into the commonly done choice of the L1 normalization of the latent features as the authors explain that connected sets of samples with soft boundaries (*i.e.,* well-clustered data) are almost linearly separable in the hyperspherical space.

A.3.5 Deep Clustering

In the past decades, unsupervised and self-supervised learning techniques have proven to be particularly effective at identifying relevant patterns and factors of variation within a dataset. Combined with powerful Neural Networks (NNs), these methods can produce semantically rich representations [50, 52, 118, 326]. Notably, unsupervised Deep Clustering (DC) methods [16, 40, 41, 169, 278] seek to produce a suitable representation space for identifying homogeneous latent clusters based on the patterns present in the dataset. Previous methods [288]

have already been led to develop methods that simultaneously learn feature representations and cluster assignments using deep neural networks. Recent works such as [16, 40, 41, 184, 319], proposed to uncover semantically groups of samples without using other auxiliary tasks (such as a reconstruction task for example). Notably, a simple method is Deep Cluster [40], which alternates between a pseudo-label estimation phase (*i.e.*: the clustering estimation step) and the update of encoder parameters (*i.e.*: the descent gradient step). In detail, at each epoch, previous clustering assignments produced by a K-Means method are used as pseudo-labels for minimizing a Cross-Entropy loss with the network output logits. Interestingly, as in classification methods such as Logistic Regression, this alternating double optimization setup can be justified via a statistical framework, by assuming that pseudo-labels are the estimates of $\hat{y}_{i,k}$ between each gradient descent (see Sec 3.2.2).

Figure A.9: Scheme of DeepCluster's training process. Image taken from the original paper [40]. Clusters are alternatively estimated from deep features then, cluster assignments are used as as pseudo-labels to learn and update the parameters of the deep convolutional neural networks.



Later, several methods such as SwAv [41] and PCL [168] have introduced an additional constraint on the features encoder: the augmentation invariance with respect to user-designed transformation functions. This choice enabled the practitioners to further guide the deep neural network representation learning by distillating domain-specific a prioris about the input-augmentation invariance required to perform clustering. Such Deep Clustering approaches perform feature extraction with a Deep Neural Network (can also be called "dimension reduction") jointly with clustering identification in the representation space. This kind of method is directly relevant to our thesis objective as it lays the basis of subgroup structure discovery performed with an automatic deep features encoder.

Appendix B

Deep UCSL's appendix

B.1 Sinkhorn-Knopp Soft K-Means

In this section, we describe the pseudo-code algorithm (See Alg. 0.) for the Soft K-Means algorithm regularized with Sinkhorn-Knopp [59]. We implement this algorithm on GPU. The Sinkhorn-Knopp algorithm directly comes from [41] and uses the same hyperparameter choice.

Algorithm 5 SK regularized Soft K-Means pseudo-code

Input:
    Disease representations: $Z \in \mathbf{R}^{N_{y=1} \times D}$,
    K: subgroups number, λ: SK temperature
    N: iterations
Output:
    Centroids: $\mu = \{\mu^k\}_{k \in |[1,K]|}$.
Initialization step:
    Initialize centroids μ with K-Means ++ algorithm.
for i in N iterations **do**
        Compute soft clustering probabilities $Q(c_i)$ given a representation Z_i: $Q(c_i) = \frac{1/||Z_i - \mu_i||_2^2}{\sum_{j=1}^{K}(1/||Z_i - \mu_j||_2^2)}$.
        Apply SK regularization: $Q = SK(Q, \lambda)$.
        Compute one-hot clustering matrix Q_{hot}:
            $Q_{hot} = OneHot(Q.argmax(dim = 1))$.
    for k in K subgroups **do**
            Update centroid k: $\mu^k = \frac{(Z * Q_{hot}[:,k]).sum()}{Q_{hot}[:,k].sum()}$.
    end for
end for
Return centroids $\mu = \{\mu^k\}_{k \in |[1,K]|}$.

The OneHot(.) function consists of transforming a smooth probability clustering vector (e.g.: $[0.2, 0.1, 0.7]$) into the hard version (i.e.: $[0, 0, 1]$).

B.2 Clustering re-identification

In the clustering re-identification paragraph, we aim to identify each updated cluster (epoch $t+1$) with its most similar previous cluster (epoch t). Let us clarify the notation. At epoch t, we have estimated K subtypes, we can compute their respective centroids with the following formula:

$$\mu_k^t = \sum_{i=1}^{N} 1_{c_i^t=k} f_\theta(x_i) \tag{B.1}$$

where f_θ is the encoder, x_i is an input image, associated with an inferred C^t at epoch t.
At epoch $t+1$, we update our subtype estimation, and we estimate K updated subtypes, once again, we can compute their centroids:

$$\mu_k^{t+1} = \sum_{i=1}^{N} 1_{c_i^{t+1}=k} f_\theta(x_i) \tag{B.2}$$

We wish to permute the labels of the clusters (and their centroids) estimated at epoch $t+1$ so that there is a continuity between clusters estimated at epoch t and those estimated at epoch $t+1$. In practice, we aim to compute a permutation function σ that maps an updated cluster (epoch $t+1$) onto its most similar former cluster (epoch t). Given a similarity function $s(\mu, \mu')$ between two centroids μ and μ'. We are seeking the optimal permutation σ^*, which maximizes the average similarity:

$$\sigma^* = \max_{\sigma} \sum_{k=1}^{K} s(\mu_k^t, \mu_{\sigma^{-1}(k)}^{t+1}) \tag{B.3}$$

Importantly, we wish to construct a function σ that is bijective. Indeed, as explained in the main text, a non-bijective mapping could potentially allow for more than one previous cluster to be merged into a single updated cluster, which may produce one or more empty clusters. For example, assuming that $K = 2$ and that the estimated mapping gives $\sigma(0) = 1$, $\sigma(1) = 1$, then after having permuted the indices of the updated clusters, we would get $C_0^{t+1} = \varnothing$ because $\sigma^{-1}(0) = \varnothing$). Thus, to ensure the bijectivity of σ, we propose casting our problem into a conceptually different one. Let us explain it in detail.
Let assume that we are given K data-points: $\{c_j^{t+1}, j \in |[1, K]|\}$ (in our experiment, it corre-

sponds to the K centroids of clusters estimated at epoch $t+1$). Now, let's say that we are given K categories (which, in our case, correspond to the K clusters estimated at epoch t). Given a similarity measure, the probability of a sample j to belong to a given category i can be computed with the following formula:

$$p(c_i^t|\mu_j^{t+1}) = \frac{s(\mu_j^{t+1}, \mu_i^t)}{\sum_{k=1}^{K} s(\mu_j^{t+1}, \mu_k^t)} \tag{B.4}$$

Algorithm 6 Subgroups re-identification pseudo-code

Inputs:
    K: subgroups number
    Previous Subgroups Centroids: $\mu^t = \{\mu_k^t\}_{k\in|[1,K]|}$
    Subgroups Centroids: $\mu^{t+1} = \{\mu_k^{t+1}\}_{k\in|[1,K]|}$
Output:
    Permuted Subgroups Centroids: $\mu^{t+1} = \{\mu^{t+1}_{\sigma^{-1}(k)}\}_k$
Initialization step: Compute the similarity matrix S:
    $S = (\frac{\mu^t}{||\mu^t||_2})^T \cdot \frac{\mu^{t+1}}{||\mu^{t+1}||_2})$
while len(np.unique(σ)) $\leq K$
    Apply SK regularization: $S_{SK} = SK(S, \lambda)$
    Compute permutation: σ = np.argmax(S_{SK}, axis=1)
    Increase SK strength: $\lambda = 1.1 \times \lambda$
endwhile
Return permuted centroids $\mu^{t+1} = \mu^{t+1}[\sigma, :]$

We wish to find the closest solution where the samples get assigned to a category, and each category has the same number of attributed samples (equipartition property). This problem has a simple solution that can be easily estimated via an optimal transport algorithm: the Sinkhorn-Knopp algorithm. See Alg. 0.
Importantly, note that in our case, as we have K samples for K classes, the equipartition property is respected if and only if each sample gets mapped to a single category, which is equivalent to having a bijective mapping between samples and categories.

B.3 Convergence guarantee

Here, we provide proof that the proposed Expectation-Maximization optimization process yields a monotonic increase of the log of the joint conditional likelihood. The proof is very similar to the one proposed in [209]. Calling $F(\theta, \phi, \psi)$ the joint conditional likelihood, namely our cost

function), we have:

$$\begin{aligned} F(\theta, \phi, \psi) &= \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} Q(c_i = k) \frac{p_{\theta,\phi,\psi}(y_i, c_i = k | x_i)}{Q(c_i = k)} \right) \\ &\geq \sum_{i=1}^{n} \sum_{k=1}^{K} Q(c_i = k) \log p_{\theta,\phi}(y_i | c_i = k, x_i) - D_{KL}(Q(c) || p_{\theta,\psi}(c|x)) \end{aligned} \tag{8}$$

Given a guess of the parameters $\theta^{(t)}$ at the t-th step, the E-step consists in choosing $Q^{(t)} = p_{\theta^{(t)}}(c_i|x_i, y_i)$ which makes the previous bound (Eq. 8) tight (*i.e.,* the inequality holds with equality). This means that, with this choice of $Q^{(t)}$, we have:

$$F(\theta^{(t)}, \phi^{(t)}, \psi^{(t)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} Q^{(t)}(c_i = k) \log p_{\theta^{(t)},\phi^{(t)}}(y_i | c_i = k, x_i) - D_{KL}(Q(c) || p_{\theta^{(t)},\psi^{(t)}}(c|x)) \tag{9}$$

At the t-th M-step, we freeze $Q^{(t)}$ and we obtain the parameters $\theta^{(t+1)}$, $\psi^{(t+1)}$ and $\phi^{(t+1)}$ by maximizing the right-hand side of the equation above. Thus:

$$\begin{aligned} F(\theta^{(t+1)}, \phi^{(t+1)}, \psi^{(t+1)}) &\geq \sum_{i=1}^{n} \sum_{k=1}^{K} Q^{(t)}(c_i = k) \log p_{\theta^{(t+1)},\phi^{(t+1)}}(y_i | c_i = k, x_i) - D_{KL}(Q^{(t)} || p_{\theta^{(t+1)},\psi^{(t+1)}}(c|x)) \\ &\geq \sum_{i=1}^{n} \sum_{k=1}^{K} Q^{(t)}(c_i = k) \log p_{\theta^{(t)},\phi^{(t)}}(y_i | c_i, x_i) - D_{KL}(Q^{(t)} || p_{\theta^{(t)},\psi^{(t)}}(c|x)) = F(\theta^{(t)}, \phi^{(t)}, \psi^{(t)}) \end{aligned} \tag{10}$$

where the first inequality comes from Eq. 8 and the second one is true since we look for the parameters $\theta^{(t+1)}, \phi^{(t+1)}, \psi^{(t+1)}$ that maximizes $F(\theta^{(t)}, \phi^{(t)}, \psi^{(t)})$. The above result suggests that $F(\theta, \phi, \psi)$ monotonically increases.

B.4 Robustness of the results to initialization

Please note that Deep UCSL is robust to different initializations. Indeed, variability in the results (*i.e.,* $\pm$) of the paper was obtained based on 5 different initializations.
Also, for the neuro-psychiatric case, the variability in the results (*i.e.,* $\pm$) of the paper was obtained based on 5 different initializations and 5 different TRAIN/VAL splits (0.9, 0.1) to also account for training data uncertainty.

B.5 Implementation details

B.5.1 MNIST

The employed MNIST dataset is balanced, with 6265 digit 7 samples and 6265 samples for the other digits. Image sizes are reduced to 32×32 with only one color channel. When used, data augmentation is the same for all methods and it comprises a RandomRotation with +/- 25 degrees, a RandomAffine with translate parameters equal to (0.1, 0.1), and a shear set to 0.1. Morphological variants (*i.e.,* morpho in the main article) had an additional erosion and dilation (kernel randomly chosen between 2 and 4 at each image) applied with a 0.4 probability. Concerning the model of the encoder in the MNIST experiment, we chose the same as in [272] (*i.e:* four convolutional layers with 7×7 kernels, padding of 3, batch norms between each layer, numbers of channels: 16, 32, 64, 128). We obtain a representation space of size 128 after an average pooling layer. For Deep UCSL, the optimizer chosen was Adam with a learning rate of 1e-4, trained during 75 epochs. The batch size was chosen as 256. Before each SK + Soft K-Means fitting, the positive sample features were scaled with a Pytorch Robiust Scaler. UCSL experiment was led with the same hyperparameters as in the original UCSL paper [179]. The only difference with the UCSL paper is that we balanced the MNIST dataset in this contribution.

B.5.2 Neuro-psychiatric experiment

We gathered multiple multi-site datasets, notably SCHIZCONNECT-VIP [292], BIOBD [245], and Bipolar and Schizophrenia Network for Intermediate Phenotype (BSNIP), all comprising of T1-weighted 3D MRI scans. SCHIZCONNECT-VIP encompasses 4 publicly available cohorts of controls and patients with schizophrenia. These cohorts present heterogeneous acquisition scanners and geographical sites. As for BSNIP, the MRI images were acquired at 5 different centers with 3T scanners across the USA. BIOBD has images of control and bipolar disorder patients. BSNIP [271] is only used as a test set throughout this study, while SCHIZCONNECT-VIP and BIOBD are mixed together (686 HC and 275 SZ, 307 BP patients) and then split during training and validation splitting. VBM pre-processing was performed with CAT12 [98, 99] from the SPM toolbox. It consists of a correction of the bias field and the noise in MRI images, the segmentation of Gray Matter (GM), White Matter (WM), and the cerebrospinal fluid (CSF). Images are normalized into a common standard MNI space using a linear transformation that accounts for global alignment (rotation, translation, and global brain size), with a non-linear

deformation [17] that locally aligns brain structures. Finally, normalized images are modulated by the Jacobian of their transformation to account for the quantity of tissue. Images were uniformized isotropic 1.5mm3 spatial resolution, output dimension is 121×145×121. From there, images are cropped to 121×121×121 and padded to reach a dimension of 128x128x128. Voxels are centered on a unit-gaussian distribution per image (*i.e.* mean of voxels of 0, std of voxels of 1). All methods were trained with a batch size of 8. This is explained by the fact that images have a huge dimension and take up a lot of memory space on the GPU. For all neuro-psychiatric experiments, we chose the same data augmentation transformations as in [73], that is: horizontal flip with probability 0.5; blur with probability 0.5, sigma=(0.1, 0.1); noise with probability 0.5, sigma=(0.1, 0.1); CutOut with probability 0.5, patch size equal to 32x32x32, RandomCrop of size (96x96x96) with probability 0.5. PCL [169] implementation parameters were m=0.9, and temperature=0.1. SimCLR [50], SupCon [153] and DeepCluster-v2 [41] temperature was chosen as 0.1. For all methods, we chose a non-linear head with a linear layer going to dimension 512, a ReLU layer, and another linear layer going from dimension 512 to 128 as in [73]. DeepUCSL was trained with a learning rate of 2e-5 without a scheduler. Clusters are inferred with our SK + Soft KMeans clustering method after applying a Standard Scaler on the latent space. For SimCLR, SupCon, and DeepCluster, clustering probabilities are estimated by fitting a K-Means on the patients after scaling the features with a Standard Scaler. Classification and contrastive methods were trained with a learning rate of 1e-5 during 100 epochs. Variations in the results (*i.e.,* ±) were obtained with 5 different train-validation splits, but were each time evaluated on the same independent test set. This enables us to account for initialization and data uncertainty as it is common in neuro-psychiatric experiments.

B.5.3 Pneumonia experiment

In the pneumonia dataset (Fig. B.1), Chest X-ray images come from retrospective cohorts of pediatric patients from one to five years old. All chest radiographs were filtered for quality control to remove unreadable scans. The diagnosis labels were determined by two experts practicing and the test set was also checked by a third expert. We chose 1340 healthy samples, 1340 viral pneumonia samples, and 1340 bacterial pneumonia samples from it[1]. Radiographies were selected from a cohort of pediatric patients aged between one and five years old from Guangzhou Women and Children's Medical Center, Guangzhou. TRAIN set images were graded by 2 radiologist experts and a third expert graded the independent TEST set to account

[1] https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia

Figure B.1: Pneumonia dataset image description of [146], see dataset URL and original contribution. The left panel of the chest X-ray shows lungs that are free of any abnormal opacification. On the other hand, bacterial pneumonia often displays a concentrated lobar consolidation, as shown by the white arrows in the middle panel, which in this case is located in the upper lobe of the right lung. Meanwhile, viral pneumonia, as seen in the right panel, typically presents with a more widespread "interstitial" pattern affecting both lungs.

for label uncertainty. Image sizes are reduced to (224, 224). For all experiments, we trained during 50 epochs. For Deep UCSL and the binary cross entropy method ("CE" in the main article), we weighted each sample by dividing it by the proportion of its class (HEALTHY / PNEUMONIA). The chosen encoder was an ImageNet pre-trained ResNet-18. For the Deep UCSL experiment, the encoder was followed by a single linear layer of size 128. SK + Soft K-Means was applied on a representation space of size 128 after applying a StandardScaler; batch size was chosen as 256. For SupCon we found it unnecessary to re-weight due to class imbalance. For the binary cross-entropy method experiment, the encoder was directly followed by the classification prediction linear layer, K-Means is applied on a representation space of size 512; batch size was chosen as 256. For SupCon and SimCLR, we chose a temperature of 0.1, a non-linear head with a linear layer going to dimension 128, a ReLU layer, and another linear layer going from dimension 128 to 128. For SupCon [153], K-Means is applied on a representation space of size 512 after a Standard Scaler, as detailed in the original paper, and classification performance is obtained with a linear probe on the representations. Variability in the results (*i.e.,* $\pm$) was obtained by relaunching the experiment with 3 different initializations.

B.5.4 Retinal OCT experiment

In the retinal OCT dataset (Fig. B.2), the Retinal optical coherence tomography (OCT) imaging technique was used to capture cross-sectional images of the retinas of patients. [2].

[2] https://www.kaggle.com/datasets/paultimothymooney/kermany2018

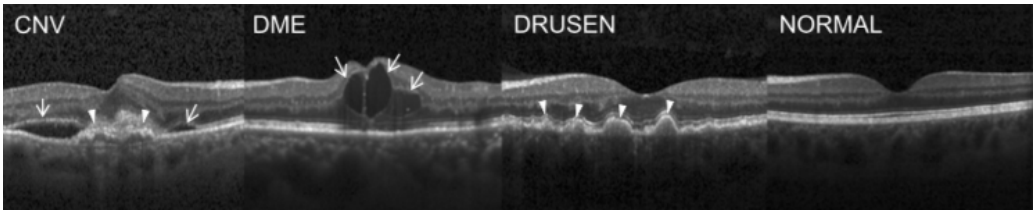


Figure B.2: Retinal OCT dataset image description of [146], see dataset URL and original contribution. The far left image displays choroidal neovascularization (CNV) featuring a neovascular membrane (indicated by white arrowheads) and accompanying subretinal fluid (indicated by arrows). Moving to the middle left, we see diabetic macular edema (DME) characterized by intraretinal fluid associated with retinal thickening (arrows). The middle right image shows the presence of multiple drusen (arrowheads), which is a hallmark of early age-related macular degeneration (AMD). Finally, the far right image depicts a normal retina with a well-preserved foveal contour and no signs of retinal fluid or edema.

Image sizes are reduced to (224, 224). For all experiments, we trained during 50 epochs. The dataset was divided into equal background class (HEALTHY) and target class (CNV: choroidal neovascularization, DME: Diabetic macular edema, DRUSEN: aging-related deposit on the retina). The chosen encoder was ResNet-18 pre-trained on Image-Net. For the Deep UCSL experiment, the encoder was followed by a single linear layer of size 128, SK + Soft K-Means is applied on a representation space of size 128 after having applied a StandardScaler; batch size was chosen as 256. For SimCLR, SCAN, DeepCluster, and SupCon, we chose a temperature of 0.1, a non-linear head with a linear layer going to dimension 128, a ReLU layer, and another linear layer going from dimension 128 to 128. For SimCLR, BYOL, DeepCluster and SupCon [153], K-Means is applied on a representation space of size 512 after a Standard Scaler, as detailed in the original paper, and classification performance is obtained with a linear probe on the representations. Variability in the results (*i.e.,* $\pm$) was obtained by relaunching the experiment with 3 different initializations.

B.5.5 ODIR experiment

The Ocular Disease Intelligent Recognition (ODIR) dataset is a structured ophthalmic database. A color fundus imaging technique was used to capture left and right eye images of patients. [3]. Image sizes are reduced to (224, 224). For all experiments, we trained during 50 epochs. The dataset was divided into equal background class (HEALTHY) and target class (Normal (N); Diabetes (D); Glaucoma (G); Cataract (C), Age-related Macular Degeneration (A); Patholog-

[3] https://www.kaggle.com/datasets/tanjemahamed/odir5k-classification

ical Myopia (M)). The chosen encoder was ResNet-18 pre-trained on ImageNet. For the Deep UCSL experiment, the encoder was followed by a single linear layer of size 128. SK + Soft K-Means was applied on a representation space of size 128 after applying a RobustScaler; batch size was chosen as 256. For SimCLR, DeepCluster and SupCon [153], K-Means is applied on a representation space of size 512 after a Standard Scaler, as detailed in the original paper, and classification performance is obtained with a linear probe on the representations. Variability in the results (*i.e.,* $\pm$) was obtained by relaunching the experiment with 3 different initializations.

B.5.6 On the input augmentation choice in Contrastive Learning

We explore the use of contrastive methods [50, 326, 213, 52, 118, 109] to enforce transformation invariance. To summarize, a contrastive method such as SimCLR [50] uses input-distortion invariance as a pretext task for representation learning. Specifically, it encourages two different augmentations of an image to be closer in the representation space with respect to all other dataset images. In this way, it can disregard irrelevant transformations, not suited for the upstream task. This idea develops an efficient manner to regress out non-semantically relevant image distortion such as rotation, flip, crop, blur, noise, or color distortion for example. However, invariance is restricted to feasible *a priori* transformations, which is a non-negligible shortcoming. Furthermore, a poor augmentation search could let irrelevant variability persist in the representation space.

To illustrate this behavior, we analyzed the subtype discovery performance on the digit 7 of MNIST dataset under various augmentation strategies. We proposed to enforce boldness invariance by simulating it through morphological deformations. In Fig. B.3, we show the differences between the subtypes obtained with and without this strategy. Therefore, assuming that they are *a priori* known and feasible, we demonstrate that wiping uninteresting general variability through the design of input augmentations remains possible. Yet, we show that an insufficient invariance search leads to bad performance. Notably, in our method, our regularization naturally erases such variability with a simple intuition. The intuition that given a class, negative samples should not belong to any of its subtypes (*i.e.:* random probability). Thus, there is no need for a difficult, human-based input augmentations design step.

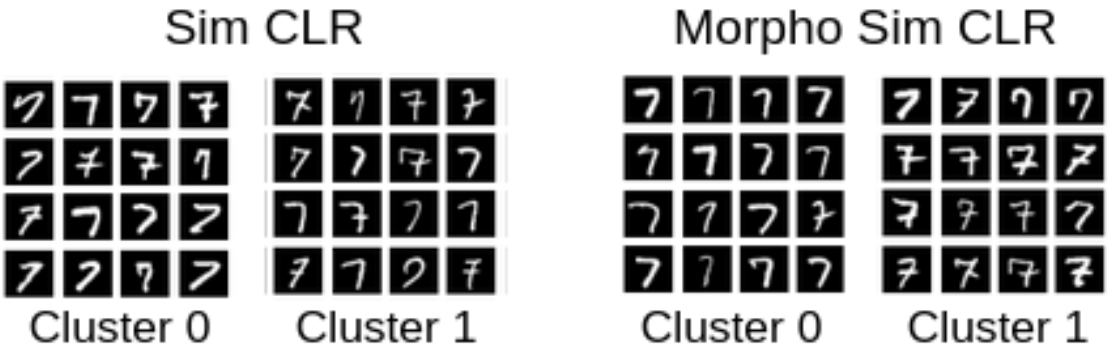


Figure B.3: Comparison between k-means clusters estimated using SimCLR and Morpho SimCLR. In Morpho SimCLR we added the morphological transformation that aims at simulating the boldness of the digits. SimCLR clusters seem to depict a general boldness variability, rather than true semantic differences. On the other hand, Morpho SimCLR focuses on more interesting patterns such as the middle bar.

Appendix C

SepVAE's appendix

C.1 Salient posterior sampling for background samples

In the original contribution, we motivated the choice of a peaked Gaussian prior for salient background distribution with a user-defined σ_p. This way, the derivation of the Kullback-Leiber divergence is directly analytically tractable as in standard VAEs.
To simplify the optimization scheme, we could also set and freeze the standard deviations $\sigma_q^{y=0}$ of the salient space of the background samples. This way, it reduces the Kullback-Leiber divergence between $q_\phi(s|x, y = 0)$ and $p_\theta(s|x, y = 0)$ to a $\frac{1}{\sigma_p}$-weighted Mean Squared Error between $\mu_s(x|y = 0)$ and s' : $\frac{||\mu_s^{x_i|y=0} - s'||_2^2}{\sigma_p}$. This choice in our code simplifies the training scheme ($\sigma_q^{y=0}$ does not need to be estimated). If a continuum exists between healthy and diseased populations, $\sigma_q^{y=0}$ should be estimated.
Also, the choice of a frozen $\sigma_q^{y=0}$ allows controlling the radius of the classification boundary between background and target samples in the salient space. Indeed, the classifier is fed with samples from the target distributions ($q_{\phi_s(s|x,y=1)} \sim N(\mu_s(x), \sigma_s(x))$), and background distributions ($q_{\phi_s(s|x,y=0)} \sim N(\mu_s(x|y = 0), \sigma_q)$. This implicitly avoids the overlap of both distributions with a margin proportional to σ_q. See Fig. C.1 for a visual explanation.

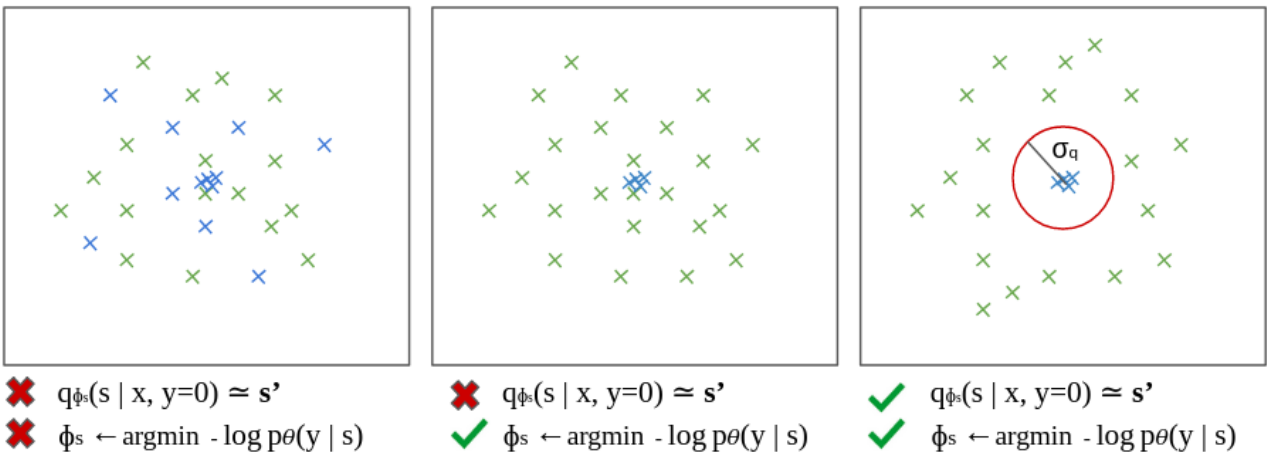


Figure C.1: Illustration of the regularization loss within the salient space. As in MM-cVAE, the prior $q_{\phi_s(s|x,y=0)} \sim$ **s'** on the background samples (blue) forces their variance to be as small as possible. However, as the prior on-target samples (green) follow a normal distribution, they may overlap with the background distribution. To avoid this case, our method trains a non-linear classifier to avoid the overlap of both distributions with a margin proportional to σ_q.

C.2 Implementation Details

C.2.1 CelebA glasses and hat versus no accessories

We used a train set of 20000 images (10000 no accessories, 5000 glasses, 5000 hats) and an independent test set of 4000 images (2000 no accessories, 1000 glasses, 1000 hats) and ran the experiment 5 times to account for initialization uncertainty. Images are of size 64×64, pixel were normalized between 0 and 1. A dataset illustration is shown in Fig.C.2 For this

Figure C.2: Celeba accessories dataset. The upper row consists of background images. The lower row shows target images.

experiment, we use a standard encoder architecture composed of 5 convolutions (channels 3,

32, 32, 64, 128, 256), kernel size 4, stride 2, and padding (1, 1, 1, 1, 1). Then, for each mean and standard deviations predicted (common and salient) we used two linear layers going from 256 to hidden size 32 to (common and salient) latent space size 16. The decoder was set symmetrically. We used the same architecture across all the concurrent works we evaluated. We used a common and latent space dimension of 16 each. The learning rate was set to 0.001 with an Adam optimizer. Oddly we found that re-instantiating it at each epoch led to better results (for concurrent works also), we think that it is because it forgets momentum internal states between the epochs. The models were trained during 250 epochs. To note, MM-cVAE used latent spaces of 16 (salient space) and 6 common space and a different architecture but we noticed that it led to artifacts in the reconstruction (see original contribution). Also, we did not succeed in reproducing their performances with their code, their model, and their latent spaces, even with the same experimental setup. We, therefore, used our model setting which led to better performances across each method with batch size equal to 512. We used $\beta_c = 0.5$ and $\beta_s = 0.5$, $\kappa = 2$, $\gamma = 1e - 10$, $\sigma_p = 0.025$. For MM-cVAE we used the same learning rate, $\beta_c = 0.5$ and $\beta_s = 0.5$, the background salient regularization weight 100, common regularization weight of 1000.

C.2.2 Pneumonia

Train set images were graded by 2 radiologist experts and a third expert graded the independent test set. The experiment was run 5 times to account for initialization uncertainty. Images are of size 64×64, pixel were normalized between 0 and 1. For this experiment, we use a standard encoder architecture composed of 4 convolutions (channels 3, 32, 32, 32, 256), kernel size 4, and padding (1, 1, 1, 0). Then, for each mean and standard deviations predicted (common and salient) we used two linear layers going from 256 to hidden size 256 to (common and salient) latent space size 128. The decoder was set in a symmetrical manner. We used the same architecture across all the concurrent works we evaluated. We used a common and latent space dimension of 128 each. The learning rate was set to 0.001 with an Adam optimizer. Oddly we found that re-instantiating it at each epoch led to better results (for concurrent works also). We think that it is because it forgets momentum internal states between the epochs. The models were trained during 100 epochs with batch size equal to 512. We used $\beta_c = 0.5$ and $\beta_s = 0.1$, $\kappa = 2$, $\gamma = 5e - 10$, $\sigma_p = 0.05$. For MM-cVAE, we used the same learning rate, $\beta_c = 0.5$ and $\beta_s = 0.1$, the background salient regularization weight 100, common regularization weight of 1000.

C.2.3 Neuro-psychiatric experiments

The task of identifying consistent correlations between neuro-anatomical biomarkers and observed symptoms in psychiatric diseases is important for developing more precise treatment options. Separating the different latent mechanisms that drive neuro-anatomical variability in psychiatric disorders is a challenging task. Contrastive Analysis (CA) methods such as ours have the potential to identify and separate healthy from pathological neuro-anatomical patterns in structural MRIs. This ability could be a key component to push forward the understanding of the mechanisms that underlie the development of psychiatric diseases. As explained in the main text, given a background population of Healthy Controls (HC) and a target population suffering from a Mental Disorder (MD), the objective is to capture the pathological factors of variability in the salient space, such as psychiatric and cognitive clinical scores, while isolating the patterns related to demographic variables, such as age and sex, or acquisition sites to the common space. For each experiment, we gather T1w anatomical VBM [17] pre-processed images of HC and MD subjects of size $128 \times 128 \times 128$. We divide them into 5 TRAIN, VAL splits (0.75, 0.25) and evaluate the performance of SOTA CA-VAEs in a cross-validation scheme. Let us note that this is a challenging problem, especially due to the high dimensionality of the input and the scarcity of the data. Notably, the measures of psychiatric and cognitive clinical scores are only available for some patients, making it scarce and precious information.

Images are of size $128 \times 128 \times 128$ with voxels normalized on a Gaussian distribution per image. Experiments were run 3 times with a different train/val/test split to account for initialization and data uncertainty. For this experiment, we use a standard encoder architecture composed of 5 3D-convolutions (channels 1, 32, 64, 128), kernel size 3, stride 2, and padding 1 followed by batch normalization layers. Then, for each mean and standard deviations predicted (common and salient), we used two linear layers going from 32768 to hidden size 2048 to (common and salient) latent space size 128. The decoder was set symmetrically, except that it has four transposed convolutions (channels 128, 64, 32, 16, 1), kernel size 3, stride 2, and padding 1 followed by batch normalization layers. We used the same architecture across all the concurrent works we evaluated. We used a common and latent space dimension of 128 each. The models were trained during 51 epochs with a batch size equal to 32 with an Adam optimizer. For the Schizophrenia experiment, for Sep VAE, we used a learning rate of 0.00005, $\beta_c = 1$ and $\beta_s = 0.1$, $\kappa = 10$, $\gamma = 1e-8$, $\alpha = \frac{1}{0.01}$. For MM-cVAE we used the same learning rate, $\beta_c = 1$ and $\beta_s = 0.1$, the background salient regularization weight 100, common regularization weight of 1000. For the Autism disorder experiment, we used a learning rate of 0.00002, $\beta_c = 1$ and $\beta_s = 0.1$, $\kappa = 10$, $\gamma = 1e-8$, $\sigma_p = 0.01$. For MM-cVAE we used the same learning rate, $\beta_c = 1$

and $\beta_s = 0.1$, the background salient regularization weight 100, common regularization weight of 1000.

Appendix D

SepCLR's appendix

D.1 Retrieve the InfoNCE loss

Let $X = \{(x_i)\}_{i=1}^N$ be the data-set of background images x_i, and $Y = \{(x_i)\}_{i=1}^N$ be the data-set of target images y_i. Input samples are assumed to be independently generated from latent unobserved variables $C = \{c_i \in \mathbf{R}^D\}_{i=1}^N$. We aim to estimate an encoder f_{θ_C} that infers latent factors of generation c from the inputs (and its views v).
To do so, we entitle c the latent codes produced by the common encoder $f_{\theta_C}(t(.))$, where $t(.) = v$ are the views generated from either x or y via a stochastic augmentation function $t(.)$. The objective is to construct an encoder $f_{\theta_C}(t(.))$ that is invariant to data augmentation. From the InfoMax perspective, we seek the optimal parameters θ^* that maximize the MI between x and $c \sim f_{\theta_C}(t(x))$. Foremost, we decompose the MI $I(x; c)$ into:

$$I(x; c) = \underbrace{-\mathbf{E}_{x \sim p_x} H(c|x)}_{\textbf{Alignment}} + \underbrace{H(c)}_{\textbf{Entropy}} \tag{D.1}$$

but the same reasoning is valid for the target dataset: $I(c; y) = -\mathbf{E}_{y \sim p_y} H(c|y) + H(c)$.

D.1.1 Derive the Uniformity term from the Entropy term

In this section, we propose to make the correspondence between the concept of Entropy, well-known in Mutual Information literature, and the concept of Uniformity introduced in [293]. The entropy can be derived with a non-parametric estimator described in [8] with samples

uniformly drawn from both the target and background datasets.

$$\hat{H}(c) = -\frac{1}{N_X + N_Y} \sum_{i=1}^{N_X+N_Y} \log \hat{p}(c_i) \tag{D.2}$$

Then, we compute the approximate density function $\hat{p}(c_i)$ with a Kernel Density Estimator, based on samples uniformly drawn from both the target dataset $f_{\theta_c}(t(y)) \sim p(c|y)$ and the background dataset $f_{\theta_c}(t(x)) \sim p(c|x)$:

$$\hat{H}(c) = -\frac{1}{N_X + N_Y} \sum_{i=1}^{N_X+N_Y} \log \frac{1}{N_X + N_Y} \sum_{j=1}^{N_X+N_Y} K_C(c_i, c_j) \tag{D.3}$$

For simplicity, we choose a Gaussian kernel with constant standard deviation τ to derive an L2 distance between the views. This enables us to obtain:

$$\hat{H}(c) = -\frac{1}{N_X + N_Y} \sum_{i=1}^{N_X+N_Y} \log \frac{1}{N_X + N_Y} \sum_{j=1}^{N} \exp \frac{-||f_\theta(v_i) - f_\theta(v_j)||_2^2}{2\tau} + \underbrace{\log(\sqrt{2\pi\tau})}_{\text{Constant term}} \tag{D.4}$$

where $c_j = f_\theta(v_j)$ and $c_i = f_\theta(v_i)$. And where v_i and v_j are the views obtained by feeding the input with index i (can be a target or a background sample) through the stochastic data augmentation function $t(.)$. In practice, [293] minimize the asymptotic lower bound of this term entitled Uniformity term. Using Jensen's inequality, we obtain:

$$\begin{aligned} &\underbrace{-\frac{1}{N_X + N_Y} \sum_{i=1}^{N_X+N_Y} \log \frac{1}{N_X + N_Y} \sum_{j=1}^{N_X+N_Y} \exp \frac{-||f_\theta(v_i) - f_\theta(v_j)||_2^2}{2\tau}}_{=\hat{H}(c)-\log(\sqrt{2\pi\tau})} \geq \\ &\underbrace{-\log \frac{1}{N_X + N_Y} \sum_{i=1}^{N_X+N_Y} \frac{1}{N_X + N_Y} \sum_{j=1}^{N_X+N_Y} \exp \frac{-||f_{\theta_C}(v_i) - f_{\theta_C}(v_j)||_2^2}{2\tau}}_{=-\mathcal{L}_{\text{uniform}}} \end{aligned} \tag{D.5}$$

Given a bounded support, minimizing $\mathcal{L}_{\text{uniform}}$ encourages the latent vectors to match a uniform distribution (*e.g:* spherical uniform distribution on unit-norm support in [293]).

D.1.2 Derive the Multi-View Alignment term

Differently from [293], we propose to estimate the conditional entropy on background samples $-H(c|x)$ with a re-substitution entropy estimator.

$$-H(c|x) = \frac{1}{N_X}\sum_{i=1}^{N_X} \log \hat{p}(c_i|x_i) \tag{D.6}$$

We compute the approximate density function $\hat{p}(c_i|x_i)$ with a Kernel Density Estimator based on samples uniformly drawn from the conditional distribution $c_i^k \sim p(c|x_i)$, where $c_i^k = f_\theta(v_i^k)$ and v_i^k are K views obtained via the stochastic process $t(.)$.

$$-H(c|x) = \frac{1}{N_X}\sum_{i=1}^{N_X} \log \frac{1}{K}\sum_{k=1}^{K} K_C(f_{\theta_C}(v_i)), f_{\theta_C}(v_i^k)) \tag{D.7}$$

K_C is chosen as a von Mises-Fisher kernel with a constant concentration parameter $\kappa = \frac{1}{\tau}$. These choices enable us to retrieve a Multi-View Alignment term with K positive views rather than only 1 as in [293]:

$$-H(c|x) + \log(C(\kappa)) = \frac{1}{N_X}\sum_{i=1}^{N_X} \log \frac{1}{K}\sum_{k=1}^{K} \exp \frac{-||f_{\theta_C}(v_i) - f_{\theta_C}(v_i^k)||_2^2}{2\tau} \tag{D.8}$$

By estimating the conditional entropy on target samples $-H(c|y)$ in the same fashion and summing both, we can retrieve the Alignment term written in Eq. 4.8. For computational reasons, we restrict to only one view in this paper: $K = 1$.

On the connection between the Gaussian kernel and the von Mises-Fisher kernel

Let us note the kernel similarity between two representations: $f_{\theta_C}(x_i)$ and $f_{\theta_C}(x_j)$ as $K(f_{\theta_C}(x_i), f_{\theta_C}(x_j))$. Assuming that we are given a Gaussian kernel with a constant standard deviation σ, this term can be estimated as:

$$K_{\text{Gaussian}}(f_{\theta_C}(x_i), f_{\theta_C}(x_j)) = \frac{1}{\sqrt{2\pi\tau}} \exp \frac{-||f_{\theta_C}(x_i) - f_{\theta_C}(x_j)||_2^2}{2\tau} \tag{D.9}$$

Now, we can divide the square norm into three terms:

$$K_{\text{Gaussian}}(f_{\theta_C}(x_i), f_{\theta_C}(x_j)) = \frac{1}{\sqrt{2\pi\tau}} \exp \frac{-||f_{\theta_C}(x_i)||_2^2 - 2f_{\theta_C}(x_i)^T.f_{\theta_C}(x_j) + ||f_{\theta_C}(x_j)||_2^2}{2\tau} \tag{D.10}$$

Let assume that $f_{\theta_C}(x_i)$ and $f_{\theta_C}(x_j)$ are unit-normed, then this estimation get simplified into:

$$K_{\text{Gaussian}}(f_{\theta_C}(x_i), f_{\theta_C}(x_j)) = \frac{1}{\sqrt{2\pi\tau}} \exp \frac{-1 + f_{\theta_C}(x_i)^T . f_{\theta_C}(x_j)}{\tau} \tag{D.11}$$

which can be further simplified:

$$K_{\text{Gaussian}}(f_{\theta_C}(x_i), f_{\theta_C}(x_j)) = \frac{1}{e^1\sqrt{2\pi\tau}} \exp \frac{f_{\theta_C}(x_i)^T . f_{\theta_C}(x_j)}{\tau} \tag{D.12}$$

Ignoring the normalization terms, we recognize the von Mises-Fisher kernel with concentration hyper-parameter $\kappa = \frac{1}{\tau}$:

$$K_{\text{vMF}} = \frac{1}{C(\kappa)} \exp \frac{f_{\theta_C}(x_i)^T . f_{\theta_C}(x_j)}{\tau}$$

D.2 Derive the Background-Contrasting InfoNCE loss in the salient space

In this section, we propose deriving the salient term $I(s; y)$ into a novel loss entitled BC-InfoNCE. Foremost, let us decompose the constrained Mutual Information maximization:

$$\text{argmax} \; - \underbrace{\mathbf{E}_{y\sim p_y} H(s|y)}_{\text{Target Alignment}} + \underbrace{H(s)}_{s'\text{-Entropy}} \qquad \text{s.t.} \; \underbrace{D_{KL}(s_x || \delta(s')) = 0}_{\text{Information-less hyp.}} \tag{D.13}$$

D.2.1 Alignment of target samples:

In order to estimate the target samples' alignment term, we use the same estimation method as in 4.2.4. First, we derive an alignment term between two views $f_{\theta_S}(v_i)$ and $f_{\theta_S}(v_i^+)$ of the same target image y_i using re-substitution estimation:

$$-\mathbf{E}_{y\sim p_y} \hat{H}(s|y) = \frac{1}{N_Y} \sum_{i=1}^{N_Y} \hat{p}(s_i|y_i) \tag{D.14}$$

Then, the density $\hat{p}(s_i|y_i)$ is estimated with a Kernel Density Estimator based on samples uniformly drawn from $p_{s|y_i}$, *i.e.:* $\{f_\theta(t(y_i)^k)|y_i\}_{k=1}^K$, where $t(y_i)^k = v_i^k$ are K views uniformly

drawn from the stochastic input-transformation process $t(y_i)$:

$$-\mathbf{E}_{y\sim p_y}\hat{H}(s|y) = \frac{1}{N}\sum_{i=1}^{N}\log\frac{1}{N}\sum_{k=1}^{K}K_Z(f_\theta(v_i), f_\theta(v_i^k)) \tag{D.15}$$

K_Z is chosen as a von Mises-Fisher kernel with a constant concentration parameter $\kappa = \frac{1}{\tau}$ and only $K = 1$ positive view is chosen. These choices enable us to derive the target alignment term:

$$-\mathbf{E}_{y\sim p_y}\hat{H}(s|y) = \frac{1}{N_Y}\sum_{i=1}^{N_Y}\frac{-||f_\theta(v_i) - f_\theta(v_i^+)||_2^2}{2\tau} \tag{D.16}$$

D.2.2 s'-Uniformity:

Now, concerning the Entropy term, we propose to develop the salient entropy with a resubstitution entropy estimator from samples drawn from $X \cup Y$.

$$\hat{H}(S) = -\frac{1}{(N_Y + N_X)}\sum_{v\in t(X\cup Y)}\log\hat{p}(f_{\theta_S}(v)) \tag{D.17}$$

Then we estimate the density $\hat{p}(f_{\theta_S}(v))$ with a Gaussian Kernel Density Estimator based on latent vectors drawn from the target view $f_{\theta_S}(t(y))$ and from the background views $f_{\theta_S}(t(x))$.

$$\hat{H}(s) = -\frac{1}{N_Y + N_X}\sum_{v\in t(X\cup Y)}\log\frac{1}{N_Y + N_X}\sum_{v^+\in t(X\cup Y)}\exp\frac{-||f_{\theta_s}(v) - f_{\theta_s}(v^+)||_2^2}{\tau} \tag{D.18}$$

We consider the asymptotic form of the Entropy. Therefore, we pull the log out of the exterior sum. In practice, it is equivalent to considering a lower bound of the Entropy. Now, separating the background and the target datasets inside the log yields:

$$
\begin{aligned}
\exp \mathcal{L}_{s'\text{uniform}} = &- \frac{1}{N_Y + N_X} \sum_{i=1}^{N_Y} \frac{1}{N_Y + N_X} \sum_{j=1}^{N_X} \exp \frac{-||f_{\theta_s}(y_i) - f_{\theta_s}(x_j)||_2^2}{\tau} \\
&- \frac{1}{N_Y + N_X} \sum_{i=1}^{N_Y} \frac{1}{N_Y + N_X} \sum_{j=1}^{N_Y} \exp \frac{-||f_{\theta_s}(y_i) - f_{\theta_s}(y_j)||_2^2}{\tau} \\
&- \frac{1}{N_Y + N_X} \sum_{i=1}^{N_X} \frac{1}{N_Y + N_X} \sum_{j=1}^{N_X} \exp \frac{-||f_{\theta_s}(x_i) - f_{\theta_s}(x_j)||_2^2}{\tau} \\
&- \frac{1}{N_Y + N_X} \sum_{i=1}^{N_X} \frac{1}{N_Y + N_X} \sum_{j=1}^{N_Y} \exp \frac{-||f_{\theta_s}(x_i) - f_{\theta_s}(y_j)||_2^2}{\tau}
\end{aligned}
\tag{D.19}
$$

Importantly, the information-less hypothesis constrains the specific encoder to produce background embeddings aligned on the information-less vector s'. This property implies that background samples should not have any variability expressed in the latent space. Assuming that the salient encoder respects this property yields $f_{\theta_S}(t(x)) = s'$, it enables to express $\hat{H}(S)$ as:

$$
\begin{aligned}
- \exp \mathcal{L}_{s'\text{uniform}} = &2 \frac{1}{N_Y + N_X} \sum_{i=1}^{N_Y} \frac{N_X}{N_Y + N_X} \exp \frac{-||f_{\theta_s}(y_i) - s'||_2^2}{\tau} + \frac{N_X}{N_Y + N_X} \frac{N_X}{N_Y + N_X} \\
&\frac{1}{N_Y + N_X} \sum_{i=1}^{N_Y} \frac{1}{N_Y + N_X} \sum_{j=1}^{N_Y} \exp \frac{-||f_{\theta_s}(y_i) - f_{\theta_s}(y_j)||_2^2}{\tau}
\end{aligned}
\tag{D.20}
$$

Assuming that the target and background datasets are balanced: $N_X = N_Y = N$ and ignoring the constant terms, we obtain:

$$
\mathcal{L}_{s'\text{uniform}} = - \log \frac{1}{N_Y} \sum_{i=1}^{N_Y} \left(\exp \frac{-||f_{\theta_s}(y_i) - s'||_2^2}{\tau} + \frac{1}{2N_Y} \sum_{j=1}^{N_Y} \exp \frac{-||f_{\theta_s}(y_i) - f_{\theta_s}(y_j)||_2^2}{\tau} \right)
\tag{D.21}
$$

D.2.3 On the Information-less hypothesis:

To respect the Information-less hypothesis, we re-write Eq. 4.9 as a Lagrangian function, with the constraint expressed as a β-weighted ($\beta \geq 0$) KL regularization. Assuming that s_x follows a Gaussian distribution centered on $f_{\theta_s}(x)$ with a constant standard deviation τ permits deriving

the KL divergence into an L2-distance between $f_{\theta_s}(x)$ and s'. Let us re-write Eq. 4.9 under the KKT conditions:

$$-\mathcal{F}(\theta_S, \beta; x, y, s) = \mathcal{L}_{y\text{-alignment}} + \mathcal{L}_{s'\text{-uniformity}} + \beta \frac{1}{N_X} \sum_{i=1}^{N_X} ||f_{\theta_s}(x_i) - s'||_2^2 \tag{D.22}$$

D.3 Retrieve the Supervised InfoNCE loss

The Supervised counterpart of the InfoNCE loss has been introduced in [153]. Compared to SimCLR, it consists of choosing positive pairs from the same class, while the negative pairs term remains unchanged. Let $X = \{(x_i)\}_{i=1}^N$ be a data-set of images x_i, $Y = \{(y_i)\}_{i=1}^N$ be their associated discrete or continuous labels y_i, and $Z = \{(z_i)\}_{i=1}^N$ the associated latent codes z_i. Let us introduce the maximization of Mutual Information between the labels Y and latent vectors Z. The Mutual Information can be decomposed as follows:

$$I(z; y) = \underbrace{-\mathbf{E}_{y\sim p_y} H(z|y)}_{\textbf{Supervised Alignment term}} + \underbrace{H(z)}_{\textbf{Uniformity term}} \tag{D.23}$$

The Supervised counterpart of the InfoNCE loss has been introduced in [153]. In this section, we show that it can be derived from the MI between Y and $f_\theta(t(X))$. Compared to InfoNCE, it consists in aligning positive views $(t(x_i), t(x_j))$ from the same class $y_i = y_j$ via a supervised alignment term $-\mathbf{E}_{y\sim p_y} H(z|y)$, while the entropy term estimation $H(z)$ remains the same. Using the re-substitution estimator and the KDE, we derive the supervised alignment term into the alignment term of $\mathcal{L}_{\text{sup}}^{\text{in}}$ in [153]:

$$-\mathbf{E}_{y\sim p_y} H(z|y) + \log(\sqrt{2\pi\tau}) = \frac{1}{N} \sum_{i=1}^{N} \log \frac{1}{|P(i)|} \sum_{j\in P(i)} \exp \frac{-||f_\theta(v_i)^T - f_\theta(v_j)||_2^2}{2\tau} \tag{D.24}$$

where $P(i)$ is the set of indices of samples belonging to class y_i and $|P(i)|$ is its cardinality. In [74], the authors proposed a generalized version of SupCon, which accounts for continuous label y.

D.3.1 On the distinction between $\mathcal{L}_{\text{sup}}^{\text{in}}$ and $\mathcal{L}_{\text{sup}}^{\text{out}}$:

In [153], the authors show that it is preferable to optimize $\mathcal{L}_{\text{sup}}^{\text{out}}$, a variant of $\mathcal{L}_{\text{sup}}^{\text{in}}$ where positive samples are summed outside of the logarithm. We propose to derive $\mathcal{L}_{\text{sup}}^{\text{out}}$ rather than $\mathcal{L}_{\text{sup}}^{\text{in}}$ by

simply computing a lower bound of the Alignment term via Jensen's inequality:

$$-\mathbf{E}_{y\sim p_y} H(z|y) + \log(\sqrt{2\pi\tau}) \geq -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|P(i)|}\sum_{j\in P(i)}\frac{||f_\theta(v_i)^T - f_\theta(v_j)||_2^2}{2\tau} \tag{D.25}$$

D.3.2 Quantify the Jensen Gap for SupInfoNCE:

In Eq.D.25, we derived a lower bound of the Conditional Entropy of the Supervised InfoMax formulation via Jensen's inequality. In this paragraph, we propose to a) quantify Jensen's Gap between both formulations and b) describe under which condition these formulations are equal (tight bound). The Jensen's Gap can be computed as:

$$J_{\text{GAP}} = \frac{1}{N}\sum_{i=1}^{N}\underbrace{\log\frac{1}{|P(i)|}\sum_{j\in P(i)}\exp\frac{-||f_\theta(v_i)^T - f_\theta(v_j)||_2^2}{2\tau} - \frac{1}{|P(i)|}\sum_{j\in P(i)}\frac{-||f_\theta(v_i)^T - f_\theta(v_j)||_2^2}{2\tau}}_{\text{D}_{\text{GAP}}^{\text{sup}}} \tag{D.26}$$

where $J_{\text{GAP}} \geq 0$. Let us note $J_{\text{GAP}} = 0$ if and only if $\text{D}_{\text{GAP}}^{\text{sup}} = 0$. We simplified $\text{D}_{\text{GAP}}^{\text{sup}}$ into the difference between a LogSumExp and a SumLogExp of $f_\theta(v_i)^T.f_\theta(v'_j)$. Using the fact that LogSumExp consists of a smooth approximation of the max function, $\text{D}_{\text{GAP}} = 0$ if and only if:

$$\max_j ||f_\theta(v_i)^T - f_\theta(v_j)||_2^2 + \log\frac{1}{N} = \frac{1}{N}\sum_{j=1}^{N}||f_\theta(v_i)^T - f_\theta(v_j)||_2^2 \quad , \forall i \text{ in } |[1,N]| \tag{D.27}$$

where $y_i = y_j$, *i.e:* v_i, v_j and v'_j are views from images from the same class y.

D.3.3 The case of a continuous y:

In [74], the authors proposed a generalized version of SupCon, which accounts for continuous label y. It adds a weight $w_\sigma(y_i, y_j)$ before the similarity term. Let us explain how to retrieve this formulation. From Eq. D.23, we use the resubstitution estimator:

$$-\mathbf{E}_{y\sim p_y} H(z|y) + \log(\sqrt{2\pi\tau}) = \frac{1}{N}\sum_{i=1}^{N}\log\hat{p}(z_i|y_i) \tag{D.28}$$

From there, we can use a Kernel Density estimation for the conditional distributions in the case where we only have access to samples from the joint distribution:

$$-\mathbf{E}_{y\sim p_y} H(z|y) + \log(\sqrt{2\pi\tau}) = \frac{1}{N}\sum_{i=1}^{N} \log \frac{\frac{1}{N}\sum_{j=1} K_Y(y_i, y_j) K_Z(f_\theta(x_i), f_\theta(x_j))}{\frac{1}{N}\sum_{j=1} K_Y(y_i, y_j)} \tag{D.29}$$

By choosing K_Y as a gaussian kernel: $K_y(y_i, y_j) = \frac{1}{\sigma\sqrt{2\pi}} \exp -\frac{(y_i - y_j)^2}{2\sigma^2}$ and K_Z as a von Mises-Fisher kernel as usually done in Contrastive Learning literature, we retrieve [74]'s $L_{\text{sup}}^{\text{in}}$ formulation.

$$-\mathbf{E}_{y\sim p_y} H(z|y) + \log(\sqrt{2\pi\tau}) = \frac{1}{N}\sum_{i=1}^{N} \log \frac{1}{N}\sum_{j=1} w_\sigma(y_i, y_j) \exp \frac{-||f_\theta(x_i) - f_\theta(x_j)||_2^2}{2\tau} \tag{D.30}$$

where $w_\sigma(y_i, y_j) = \frac{K_Y(y_i, y_j)}{\frac{1}{N}\sum_{j=1} K_Y(y_i, y_j)}$.
Now, the Jensen's inequality can be utilized to retrieve [74]'s exact formulation.

D.4 Maximize the Joint Entropy via Kernel Density-based Estimation

In Sec. 4.2.4, we proposed a method to estimate and minimize $-H(c, s)$ without requiring any assumptions about the form of its pdf nor requiring a neural network-based approximation ([53, 11, 226]). Inspired by [125], we develop $H(c, s)$ with a re-substitution entropy estimator:

$$-\hat{H}(c, s) = \frac{1}{N_X + N_Y} \sum_{i=1}^{N_X+N_Y} \log \hat{p}(c_i, s_i) \tag{D.31}$$

To do so, we estimate the density $\hat{p_\theta}(c_i, s_i)$ with a Kernel Density Estimation:

$$-\hat{H}(c, s) = \frac{1}{N_X + N_Y} \sum_{i=1}^{N_X+N_Y} \log \frac{1}{N_X + N_Y} \sum_{k=1}^{N_X+N_Y} K_C(c_i, c_j) K_S(s_i, s_j) \tag{D.32}$$

where c_j and s_j are drawn from the joint distribution $p(c, s)$. In practice, we will draw pairs (c, s) from $(f_{\theta_C}(x), f_{\theta_S}(x))$ and $(f_{\theta_C}(y), f_{\theta_S}(y))$ where x and y are respectively uniformly drawn from X and Y. Importantly, as in Sec. 4.2.4, the information-less constraint still holds: $f_{\theta_S}(x) = s', \forall x$.
For simplicity, we choose Gaussian kernels for K_C and K_S with a constant standard deviation

parameter τ, which simplifies the estimation of the joint entropy into:

$$-\hat{H}(c,s) = \frac{1}{N_X+N_Y}\sum_{i=1}^{N_X+N_Y}\log\frac{1}{N_X+N_Y}\sum_{j=1}^{N_X+N_Y}\exp\frac{-||c_i-c_j||_2^2}{2\tau}\exp\frac{-||s_i-s_j||_2^2}{2\tau} \quad \text{(D.33)}$$

D.5 Capturing independent attributes and Disentangle with Contrastive Learning

D.5.1 Supervised disentanglement

We can also use our framework to derive a supervised disentangling loss with known variability factors. In this section, we propose to explore an extension of BC-InfoNCE in the case where independent fine-grained attributes about the target dataset: $\{a_i \in \mathcal{R}^{D_S}\}_{i=1}^{N_Y}$ are available. Given this set of independent observed characteristics, we can leverage these observations in a supervised manner to identify the independent factors of generation of the target dataset.

We assume the existence of D_S attributes and construct our salient encoder to output D_S latent dimensions. We aim to construct a salient space where each salient latent dimension S^{d_s} only depends on its corresponding attribute a^{d_s}. Let us re-write Eq. 4.5 by replacing the salient InfoMax term by each d-th attribute Supervised InfoMax term:

$$\text{argmax } I(x;c) + I(y;c) + \frac{1}{D_S}\sum_{d=1}^{D_s}\underbrace{I(a^d;s^d)}_{\text{d-th SupInfoMax}} \qquad \text{s.t. } D_{KL}(s_x||\delta(s')) = 0 \text{ and } I(c,s) = 0 \quad \text{(D.34)}$$

From there, we take inspiration from [74] to decompose each d-th attribute Supervised InfoMax term in a supervised alignment and a uniformity term:

$$I(a^d;s^d) \geq \frac{1}{N_Y}\sum_{i=1}^{N_Y} w_\sigma(a_i^d,a_j^d)\frac{||s_i^d-s_j^d||_2}{2\tau} + \hat{H}(s^d) = \mathcal{L}_{\text{d-th SupInfoMax}} \quad \text{(D.35)}$$

We propose to develop the Entropy term for each d-th salient dimension as in Sec. D.3.3.

D.6 Datasets and Implementation Details

D.6.1 dSprites watermarked on a grid of digits experiment

We provide a novel toy dataset to evaluate the Contrastive Analysis method enriched with target attributes. The background dataset X consists of 4 MNIST digits (1-4) regularly placed on a grid. The target dataset Y consists of a dSprites item added upon the foreground of this grid of digits. dSprites is a dataset introduced to evaluate disentanglement. Its images are of size 64x64 pixels. Its elements only exhibit 5 generation factors, see Fig. D.1, making it easy to evaluate the disentanglement. Possible variations are 1) shape (heart, ellipse, and square), 2) size, 3) position in X, 4) position in Y, and 5) orientation (i.e. rotation). To construct the Contrastive Analysis dataset we use in this paper, we randomly sample MNIST images of digits 1, 2, 3, and 4 and regularly place them on a grid. We create 25,000 background images with this method. Then, we superimpose a random dSprite element on 25,000 distinct digit grids to create 25 000 target images. We use the same method to derive 5,000 test images equally balanced between the target and background classes. Importantly, we constrain the dSprites elements to have a rotation attribute between -45 and $+45$ degrees. Downstream task performances are computed on the projection head.

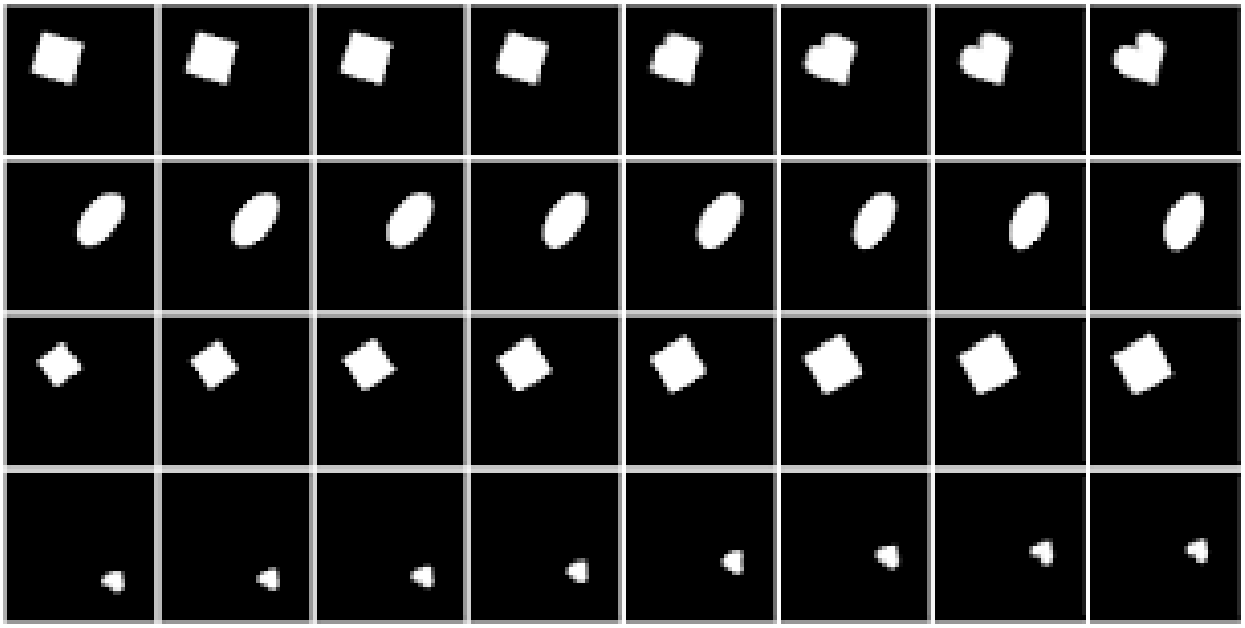


Figure D.1: Illustration of the dSprites dataset and its different independent variability factors: shape, zoom, rotation, Y position, and X position.

D.6.2 MNIST digit superimposed on CIFAR-10 background

MNIST digit superimposed on CIFAR-10 background is a simple intuitive dataset inspired from [327]. We consider as the background dataset ($y = 0$) CIFAR-10 images, and as the target dataset ($y = 1$) CIFAR-10 images (background) with an overlaid digit (target pattern), see Fig. D.2. This experiment is particularly suited to CA, we expect our model to successfully capture the background variability (*i.e:* natural objects semantic) in the common space and to capture the digits variability in the salient space. In practice, we used a train set of 50000 images (25000 Cifar-10 images, 25000 Cifar-10 images with random MNIST digits overlaid) and an independent test set of 1000 images (500, 500). Images are of size 32×32. Pixels were normalized between 0 and 1.
In terms of Data Augmentation for the stochastic transformation process $t(.)$, we remained close to SimCLR [50], as we used a RandomCrop(size=(24, 24), scale=(0.2, 1.0)) augmentation, then a RandomHorizontalFlip(p=0.5) augmentation, a RandomColorJitter(0.4, 0.4, 0.4, 0.1) applied with a probability 0.8 followed from a RandomGrayScale(p=0.2) augmentation.
Concerning the Neural Network architecture, both common and salient encoders were chosen as ResNet-18 with a representation linear layer as follows: linear(512, 32) and a non-linear projector layer as follows: (linear(32, 128), batch norm(128), relu(), linear(128, 32)). We used an Adam optimizer with learning rate of 5e-4, batch size of 512, and trained it during 250 epochs.
As for the SepCLR's hyper-parameters, we chose $\lambda_C = 1$, $\lambda_S = \beta = 1000$, and $\lambda = 10$. Downstream task performances are computed before the projection head, as in [50].

Concerning Contrastive Analysis VAE methods, we took inspiration from experimental setups in [180]. Namely, we used a standard encoder architecture composed of 4 convolutions (channels 3, 32, 32, 32, 256), kernel size 4, and padding (1, 1, 1, 0). Then, for each mean and standard deviations predicted (common and salient), we used two linear layers going from 256 to hidden size 256 to (common and salient) latent space size 32. The decoder was set symmetrically. We used the same architecture across all the CA-VAEs concurrent works we evaluated. Interestingly, we also tried with ResNet-18 encoders but the results actually remained similar. The learning rate was set to 0.001 with an Adam optimizer. The models were trained during 250 epochs with batch size equal to 512. We used $\beta_c = 0.5$ and $\beta_s = 0.5$, $\kappa = 2$, $\gamma = 1e-9$, $\alpha = \frac{1}{0.025}$. For cVAE, we used $\beta_c = 0.5$ and $\beta_s = 0.5$, $\kappa = 2$ and $\kappa = 0$ for conVAE. For MM-cVAE we used the same learning rate, $\beta_c = 0.5$ and $\beta_s = 0.5$, the background salient regularization weight 100, common regularization weight of 1000.

Concerning Mutual Information minimization methods, we used the same hyper-parameters

as for k-JEM, except for λ. λ was set to 0.1 for CLUB, as in the original paper Domain Adaptation section [53]. Please note that we also tried values of 1 and 10, but it did not give better results. We also chose 0.1 for vUB and vL1out. For TC, we used $\lambda = 10$. For MMD, we used $\lambda = 50$; we motivate this choice in the Sec. D.7.1.

Figure D.2: the Superimposed MNIST digits on CIFAR background dataset. Target images are CIFAR-10 images overlaid with an MNIST digit. Background images are CIFAR-10 images.

D.6.3 CelebA accessories

In CelebA with accessories [299], we consider a subset of CelebA [174]. It contains two sets, target and background, from a subset of CelebA [174], one with images of celebrities wearing glasses or hats (target) and the other with images of celebrities not wearing any of these accessories (background). Importantly, and contrarily to MM-cVAE [299] and SepVAE [180], we take care to balance the distribution of males and females in the background and the target dataset to avoid gender bias with respect to the accessories. We used a train set of 20000 images, (10000 no accessories, 5000 glasses, 5000 hats) and an independent test set of 4000 images (2000 no accessories, 1000 glasses, 1000 hats). Images are of size 128×128, normalized between 0 and 1.

In terms of Data Augmentation for the stochastic transformation process $t(.)$, we remained close to SimCLR [50], as we used a RandomCrop(size=(128, 128), scale=(0.2, 1.0)) augmentation, then a RandomHorizontalFlip(p=0.5) augmentation, a RandomColorJitter(0.4, 0.4, 0.4, 0.1) applied wit a probability 0.8 followed from a RandomGrayScale(p=0.2) augmentation.

Concerning the Neural Network architecture, both common and salient encoders were chosen as ResNet-18 with a representation linear layer as follows: linear(512, 16) and a non-linear projector layer as follows: (linear(16, 128), batch norm(128), relu(), linear(128, 16)). We used an Adam optimizer with learning rate 5e-4, a batch size of 256, and trained it during 250 epochs.

As for the SepCLR's hyper-parameters, we chose, as in MNIST superimposed on CIFAR-10 experiment, $\lambda_C = 1$, $\lambda_S = \beta = 1000$, and $\lambda = 10$. Downstream task performances are computed before the projection head, as in [50].
Concerning Contrastive Analysis VAE methods, we took inspiration from experimental setups in [180]. Notably, we used images of size $64x64$ pixels. Namely, we use a standard encoder architecture composed of 5 convolutions (channels 3, 32, 32, 64, 128, 256), kernel size 4, stride 2, and padding (1, 1, 1, 1, 1). Then, concerning the mean and standard deviations predicted (common and salient), we used two linear layers going from 256 to hidden size 32 to (common and salient) latent space size 16. The decoder was set symmetrically. We used the same architecture across all the CA-VAEs concurrent works we evaluated. The learning rate was set to 0.001 with an Adam optimizer. The models were trained during 250 epochs with batch size equal to 512. We used $\beta_c = 0.5$ and $\beta_s = 0.5$, $\kappa = 2$, $\gamma = 1e - 10$, $\sigma_p = 0.025$. For cVAE, we used $\beta_c = 0.5$ and $\beta_s = 0.5$, $\kappa = 2$ and $\kappa = 0$ for conVAE. For MM-cVAE, we used the same learning rate, $\beta_c = 0.5$ and $\beta_s = 0.5$, the background salient regularization weight 100, common regularization weight of 1000.
Concerning Mutual Information minimization methods, we used the same hyper-parameters as for k-JEM, except for λ. λ was set to 0.1 for CLUB, as in the original paper Domain Adaptation section [53]. Please note that we also tried values of 1 and 10, but it did not give better results. We also chose 0.1 for vUB and vL1out. For TC, we used $\lambda = 10$. For MMD, we used $\lambda = 50$; we motivate this choice in the Sec. D.7.1.

D.6.4 CheXpert

In the CheXpert subtyping experiment, we select a subset of CheXpert separated in the background dataset: 10,000 healthy X-rays and the target dataset: 3,000 with edema, 3,000 with pleural effusion, and around 2,000 images with cardiomegaly. Images are resized to 224x224 pixels. Pixels are normalized between 0 and 1.
For SepCLR, in terms of Data Augmentation for the stochastic transformation process $t(.)$, we remained close to SimCLR [50], as we used a RandomCrop(size=(224, 224), scale=(0.2, 1.0)) augmentation, a RandomColorJitter(0.4, 0.4, 0.4, 0.1) applied with a probability 0.8 followed from a RandomGrayScale(p=0.2) augmentation, a RandomRotation(degrees=45), and then a RandomHorizontalFlip(p=0.5) augmentation.
Concerning the Neural Network architecture, both common and salient sizes are 32. Both common and salient encoders were chosen as a pre-trained ResNet-18 with a representation linear layer as follows: linear(512, 32) and a non-linear projector layer as follows: (linear(32, 128),

batch norm(128), relu(), linear(128, 32)). We used an Adam optimizer with a learning rate of 5e-4, a batch size of 256, and trained it during 200 epochs. As for the SepCLR's hyper-parameters, we chose $\lambda_C = 1$, $\lambda_S = 1$, $\beta = 10$, and $\lambda = 5$. Downstream task performances are computed before the projection head, as in [50].

Concerning the Contrastive VAEs, we use the same common and salient encoders. For the decoders, we chose an architecture composed of a linear layer, taking into input the concatenation of common and salient space, mapping it to a size of 256. Then 7 deconvolution layers were used with a kernel size of 4, stride of 2, and padding of 1 with filters (256 to 512, 256, 128, 64, 32, 16, 3). Output images are of size 256×256 and are cropped to 224×224. The final activation layer is chosen as a sigmoid layer.

We used the same architecture across all the CA-VAEs concurrent works we evaluated. The learning rate was set to 0.001 with an Adam optimizer. The models were trained during 200 epochs with batch size equal to 256. We used $\beta_c = 0.5$ and $\beta_s = 0.5$, $\kappa = 2$, $\gamma = 1e-9$, $\sigma_p = 0.05$. For cVAE, we used $\beta_c = 0.5$ and $\beta_s = 0.5$, $\kappa = 2$ and $\kappa = 0$ for conVAE. For MM-cVAE, we used the same learning rate, $\beta_c = 0.5$ and $\beta_s = 0.5$, the background salient regularization weight 100, common regularization weight of 1000.

D.6.5 ODIR (Ocular Disease Image Recognition)

In the ODIR subtyping experiment, we select a subset of the ODIR dataset separated into a background and a target dataset. Train dataset contains 1890 healthy images, 363 diabetes images, 278 glaucoma images, 281 cataract images, 242 age-related macular degeneration images, and 227 pathological myopia images. On the other hand, TEST dataset contains respectively 210 healthy, 37 diabetes, 26 glaucoma, 39 cataract, 23 macular degeneration, 30 myopia images. Pixels are normalized between 0 and 1.

For SepCLR, in terms of Data Augmentation for the stochastic transformation process $t(.)$, we remained close to SimCLR [50], as we used a RandomCrop(size=(224, 224), scale=(0.75, 1.0)) augmentation, a RandomColorJitter(0.4, 0.4, 0.4, 0.1) applied wit a probability 0.8 followed from a RandomGrayScale(p=0.2) augmentation, a RandomRotation(degrees=45), and then a RandomVerticalFlip(p=0.5) augmentation.

Concerning the Neural Network architecture, both common and salient encoders were chosen as a pre-trained ResNet-18 with a representation linear layer as follows: linear(512, 32) and a non-linear projector layer as follows: (linear(32, 128), batch norm(128), relu(), linear(128, 32)). We used an Adam optimizer with a learning rate of 5e-4, a batch size of 256, and trained it during 200 epochs. As for the SepCLR's hyper-parameters, we chose $\lambda_C = 1$, $\lambda_S = 1$, $\beta = 100$,

and $\lambda = 10$. Downstream task performances are computed before the projection head, as in [50].
Concerning the Contrastive VAEs, we use the same common and salient encoders. For the decoders, we chose an architecture composed of a linear layer, taking into input the concatenation of common and salient space, mapping it to a size of 256. Then 7 deconvolution layers were used with a kernel size of 4, stride of 2, and padding of 1 with filters (256 to 512, 256, 128, 64, 32, 16, 3). Output images are of size 256×256 and are cropped to 224×224. The final activation layer is chosen as a sigmoid layer.

We used the same architecture across all the CA-VAEs concurrent works we evaluated. The learning rate was set to 0.001 with an Adam optimizer. The models were trained during 200 epochs with batch size equal to 256. We used $\beta_c = 0.5$ and $\beta_s = 0.5$, $\kappa = 2$, $\gamma = 1e-9$, $\sigma_p = 0.05$. For cVAE, we used $\beta_c = 0.5$ and $\beta_s = 0.5$, $\kappa = 2$ and $\kappa = 0$ for conVAE. For MM-cVAE, we used the same learning rate, $\beta_c = 0.5$ and $\beta_s = 0.5$, the background salient regularization weight 100, common regularization weight of 1000.

D.6.6 Schizophrenia experiment

In this study, we analyzed neuroimaging data from several sources, including the SCHIZCONNECT database (which includes 368 healthy controls and 275 patients with schizophrenia) and the BSNIP database (which includes 199 healthy controls and 190 patients with schizophrenia). The data used in this study was collected from various scanners and locations and included brain scans from individuals in the United States. Images are of size $128 \times 128 \times 128$ with voxels normalized on a Gaussian distribution per image. Experiments were run 5 times with a different train/val split (respectively 75% and 25% of the dataset) to account for initialization and data uncertainty. Inspired by [180], common and salient convolutional encoders were chosen as 5 3D-convolutions (channels 1, 32, 64, 128, 256, 512), kernel size 4, stride 2, and padding 1 followed by batch normalization layers. Then, we used a linear layer from 32768 to representations (sizes 128 for common and 32 for salient). Then, the projection heads were set as non-linear with hidden sizes 128 for common and 32 for salient, with batch normalization(128) and relu() activation functions.
For SepCLR, the data augmentations were inspired from [74], that is: horizontal flip with probability 0.5; blur with probability 0.5, sigma=(0.1, 0.1); noise with probability 0.5, sigma=(0.1, 0.1); CutOut with probability 0.5, patch size equal to 32x32x32, RandomCrop of size (96x96x96) with probability 0.5. The models were trained during 50 epochs with a batch size equal to 32 with an Adam optimizer of learning rate of 0.0005. As for the SepCLR's hyper-parameters, we

chose $\lambda_C = 1$, $\lambda_S = 1$, $\beta = 1$, and $\lambda = 5$. Downstream task performances are computed before the projection head, as in [50]. Importantly, the classification task is computed with a 2 layers MLPs to be comparable with SepVAE [180]

Concerning the Contrastive Analysis VAEs methods we compared with, we use the same experimental setup in terms of hyper-parameters and architecture as in [180]. Concerning the architecture, in detail, the common and salient convolutional encoders were chosen as 5 3D-convolutions (channels 1, 32, 64, 128, 256, 512), kernel size 4, stride 2, and padding 1 followed by batch normalization layers. Then, we used a non-linear layer from 32768 to directly predict mean and standard deviations (sizes 256 for common and 256 for salient) with 2048 as hidden size with batch normalization and relu as activation functions. The decoder was set symmetrically, except it has 6 transposed convolutions (channels 512, 256, 128, 64, 32, 16, 1), kernel size 3, stride 2, and padding 1, followed by batch normalization layers.

D.6.7 Mutual Information Minimization methods

To compare with k-JEM (kernel-Joint Entropy Maximization), we used the implementation of several Mutual Information variational upper bound, namely vCLUB [53], vUB [11] and vL1out [226] available at `https://github.com/Linear95/CLUB/tree/master`. Interestingly, these methods can be implemented with a variational approximation of S from C, vice-versa (C from S), or symmetrically (mean of both). We tried all three possibilities with different weights and chose the best results each time to set in Tab .4.5 and Tab .4.6.
We also compared with the exact Mutual Information estimator TC of [180] and [3] inspired by the Total Correlation introduced in [152].
In Sec. 4.2.4, we motivated the idea of minimizing the negative joint entropy ($-H(C, S)$) rather than the Mutual Information ($H(C) + H(S) - H(C, S)$). To prove our point, we implemented k-MI, a version of k-JEM where we also minimize the entropies $H(C) + H(S)$. To do so, we estimate $H(C)$ as in the common entropy estimation in Eq. D.4 and $H(S)$ as in the salent entropy estimation in Eq. D.18. Interestingly, we can see that k-MI indeed underperforms compared to k-JEM.

D.7 Supplementary results

D.7.1 On Mutual Information minimization versus target and background distributions matching

In Contrastive Analysis, practitioners use various regularizations to respect properties established a priori. Recent works agree that background input should be mapped to a single information-less vector in the salient space. However, two regularizations have been proposed to reduce the information leakage between the common and the salient space: 1- match the distributions of targets and backgrounds in the common space, 2- minimize the mutual information between the common and salient distributions. In our framework, the latter was naturally derived from the InfoMax principle. In Tab .D.1 and Tab .D.2, we propose to compare both strategies on a) CelebA accessories and b) Digits superimposed on CIFAR-10 to assess their effect on the common space. In both experiments, we observe that the stronger the regularization is, the less common information (objects and sex) is captured. Also, we observe that k-JEM 's ability to diminish target-specific information (digits and accessories) remains relatively consistent across the regularization strength. Concerning MMD (Maximum Mean Discrepancy), a high regularization strength is needed to reduce target-specific information despite its detrimental effect on capturing common patterns. We conclude that a low-strength k-JEM regularization (we choose $\lambda = 10$ in practice) is the right trade-off for capturing common patterns while canceling salient patterns.

Table D.1: Digits watermarked on CIFAR-10 (B-ACC). Comparison of k-JEM with MMD given different strengths.

	Digits		Objects		$\delta_{\text{TOT}} \downarrow$
	S $\uparrow$	C $\downarrow$	S $\downarrow$	C $\uparrow$	
SepCLR-no k-JEM	95.6	94.4	9.0	42.0	145.8
SepCLR-10 MMD	95.4	86.8	10.8	48.2	56.8
SepCLR-50 MMD	94.6	21.2	9.0	62.2	134.0
SepCLR-100 MMD	95.2	13.8	11.0	56.4	53.2
SepCLR-10 k-JEM	96.2	11.0	10.4	73.2	**32.0**
SepCLR-50 k-JEM	95.2	13.2	8.6	59.2	47.4
SepCLR-100 k-JEM	95.0	12.0	9.2	52.4	53.8
Best expected	100.0	10.0	10.0	100.0	0.0

Table D.2: CelebA accessories (B-ACC). Comparison of k-JEM with MMD given different strengths.

	Hats/Glss		Sex		$\delta_{\text{TOT}} \downarrow$
	S ↑	C ↓	S ↓	C ↑	
SepCLR-no k-JEM	99.03	66.68	98.48	79.48	86.65
SepCLR-10 MMD	98.99	81.53	60.87	76.19	87.14
SepCLR-50 MMD	98.95	67.50	65.47	71.83	62.19
SepCLR-100 MMD	99.03	53.25	67.12	52.51	68.83
SepCLR-10 k-JEM	98.57	55.21	62.52	78.00	**41.16**
SepCLR-50 k-JEM	98.83	58.27	62.45	68.38	53.51
SepCLR-100 k-JEM	98.73	62.00	68.92	57.10	75.09
Best Expected	100.0	50.0	50.0	100.0	0.0

D.7.2 On the add of a reconstruction term

Contrastive Analysis, jointly performed with a generative process, enables performing salient or common characteristics swapping, salient attribute generation or deletion, and novel sample generation. Therefore, we investigated the addition of a decoder jointly trained with the encoder parameters to reconstruct the input images (with a Mean Square Error Cost Function) during the optimization process. We added a reconstruction term from the concatenation of the common and salient space (as in CA-VAEs but without the need for a re-parameterization trick) with the same decoder as in CA-VAEs, and it degrades the results. Intriguingly, we found that it tends to degrade the results (see Tab. D.4 and Tab. D.3), which could be explained by the fact that the reconstruction task tries to conserve unnecessary noisy information in the latent space. However, an interesting perspective could be to include and train a generator or a decoder for generation and interpretability purposes, given frozen representations learned with SepCLR.

D.7.3 On the comparison with Contrastive methods

In this section, we propose to compare SepCLR with self-supervised methods that are not based on the encoder-decoder architecture. As no Contrastive Learning methods are tailored for Contrastive Analysis, we propose to design a naive and simple strategy to compare with

Table D.3: Digits watermarked on CIFAR-10 (B-ACC). On the impact of a reconstruction term in addition to SepCLR.

	Digits		Objects		$\delta_{\text{tot}} \downarrow$
	S ↑	C ↓	S ↓	C ↑	
SepCLR-k-JEM	96.2	11.0	10.4	73.2	**32.0**
SepCLR-k-JEM with 0.1 rec	98.8	10.8	40.6	47.2	85.4
SepCLR-k-JEM with 1 rec	94.4	22.2	51.8	27.4	132.2
Best expected	100.0	10.0	10.0	100.0	0.0

Table D.4: CelebA accessories (B-ACC). On the impact of a reconstruction term in addition to SepCLR.

	Hats/Glss		Sex		$\delta_{\text{tot}} \downarrow$
	S ↑	C ↓	S ↓	C ↑	
SepCLR-k-JEM	98.57	55.21	62.52	78.00	**41.16**
SepCLR-k-JEM with 0.1 rec	97.27	67.81	67.53	62.38	75.69
SepCLR-k-JEM with 1 rec	91.51	68.87	62.77	64.39	78.98
Best expected	100.0	50.0	50.0	100.0	0.0

SepCLR. First, we infer the common features with the features of SimCLR trained on the background dataset only (as it should have common features only). Then, we propose to infer the salient space with a SupCon method trained to discriminate the background samples from the target samples. This way, such a method should capture target-specific patterns while discarding common features. Additionally, we compare with SimCLR trained on both datasets to get a reference point (even though, in that case, the common space and the salient space are the same unique space, which cannot perform the separation of common and salient patterns). See Tab. D.5, Tab. D.6 and Tab. D.7 for the results. SepCLR always performs better in terms of δ_{tot}.

D.7.4 Ablation study

In the main text, we investigated the effect of a null Mutual Information constraint by removing the proposed loss (No k-JEM) or by minimizing the Mutual Information estimate (k-MI) rather than the Joint Entropy estimate (see Tab. 4.5 and Tab. 4.6). Here, we propose a further ablation study in Tab. D.8 and Tab. D.9. We report the results of our method when removing all proposed losses one by one. We can observe that each loss is important since, when removing

Table D.5: Comparison of SepCLR-k-JEM with Contrastive methods on Digits watermarked on CIFAR-1 (B-ACC)

	Digits		Objects		$\delta_{\text{TOT}} \downarrow$
	S ↑	C ↓	S ↓	C ↑	
SimCLR on TG and BG	44.0	44.0	94.6	94.6	180.0
SimCLR + SupCon	41.4	51.4	19.0	50.0	159.0
SepCLR-k-JEM	96.2	11.0	10.4	73.2	**32.0**
Best expected	100.0	10.0	10.0	100.0	0.0

Table D.6: Comparison of SepCLR-k-JEM with Contrastive methods on CelebA accessories (B-ACC).

	Hats/Glss		Sex		$\delta_{\text{TOT}} \downarrow$
	S ↑	C ↓	S ↓	C ↑	
SimCLR on BG and TG	98.92	98.92	84.16	84.16	100.0
SimCLR + SupCon	97.93	82.15	59.98	80.76	63.44
SepCLR-k-JEM	98.57	55.21	62.52	78.00	**41.16**
Best Expected	100.0	50.0	50.0	100.0	0.0

Table D.7: Comparison of SepCLR-k-JEM with Contrastive methods on ODIR dataset (B-ACC).

	Subtype		Sex		$\delta_{\text{TOT}} \downarrow$
	S ↑	C ↓	S ↓	C ↑	
SimCLR on BG and TG	66.10	66.10	57.20	57.20	125.0
SimCLR + SupCon	68.70	57.17	51.94	58.41	107.0
SepCLR-k-JEM	68.54	47.71	52.48	59.62	**97.03**
Best Expected	100.0	25.0	50.0	100.0	0.0

it, we either degrade the capture of salient patterns or we fail to disregard the common features in the salient space.

D.7.5 Performances on the background datasets

In the main text, we evaluated our method on the ability to linearly predict common attributes in the common space only and salient attributes in the salient space only on target samples only (as they are generated from both common and target-specific factors of variability). In this section, we evaluate the ability to linearly predict common attributes only in the common space. In Tab. D.10 and Tab D.11, we can see that the common performances remain good

Table D.8: Ablation Study on Digits watermarked on CIFAR-10 (B-ACC).

	Digits		Objects		$\delta_{\text{tot}} \downarrow$
	S ↑	C ↓	S ↓	C ↑	
SepCLR-$\lambda = 0$ (no k-JEM)	95.6	94.4	9.0	42.0	145.8
SepCLR-$\beta = 0$ (no Infoless reg)	96.2	11.6	10.4	71.8	34.0
SepCLR-$\lambda_S = 0$ (no Salient term)	93.4	42.0	18.6	40.0	90.25
SepCLR-$\lambda_C = 0$ (no Common term)	94.4	10.4	18.8	20.4	94.4
SepCLR-k-JEM	96.2	11.0	10.4	73.2	**32.0**
Best expected	100.0	10.0	10.0	100.0	0.0

Table D.9: Ablation Study on CelebA accessories (B-ACC).

	Hats/Glss		Sex		$\delta_{\text{tot}} \downarrow$
	S ↑	C ↓	S ↓	C ↑	
SepCLR - $\lambda = 0$ (no k-JEM)	99.03	66.68	98.48	79.48	86.65
SepCLR - $\beta = 0$ (no Infoless reg)	99.12	53.88	68.82	77.29	46.29
SepCLR - $\lambda_S = 0$ (no Salient term)	77.50	87.73	53.30	77.55	85.98
SepCLR - $\lambda_C = 0$ (no Common term)	98.38	56.32	66.44	53.09	71.29
SepCLR - k-JEM	98.57	55.21	62.52	78.00	**41.16**
Best Expected	100.0	50.0	50.0	100.0	0.0

on background samples while the salient space is non-informative as it is supposed to be. We can also notice that, compared to concurrent CA-VAE methods, our method is still the best-performing one in terms of δ, as it predicts common attributes way better.

Table D.10: Balanced Accuracy results on Digits watermarked on CIFAR-10 on background samples.

	Digits		Objects		$\delta \downarrow$
	S ↑	C ↓	S ↓	C ↑	
MM-cVAE	×	×	18.2	32.8	75.4
SepVAE	×	×	20.0	34.4	75.6
SepCLR-k-JEM	×	×	28.0	74.0	**44.0**
Best expected	×	×	10.0	100.0	0.0

Table D.11: Balanced Accuracy results on CelebA accessories on background samples.

	Hats/Glss		Sex		$\delta \downarrow$
	S $\uparrow$	C $\downarrow$	S $\downarrow$	C $\uparrow$	
MM-cVAE	$\times$	$\times$	64.27	70.48	43.79
SepVAE	$\times$	$\times$	56.42	70.19	36.23
SepCLR - k-JEM	$\times$	$\times$	64.10	86.63	**27.47**
Best Expected	$\times$	$\times$	50.0	100.0	0.0

D.7.6 On the impact of $\mathcal{L}_{\text{unif}}$ or $\mathcal{L}_{\text{log-sum-exp}}$

As shown in Sec. D.1, we estimate the entropy using a resubstitution entropy estimator. This results in one of the terms of the standard Contrastive loss (i.e., InfoNCE) that accounts for the negative samples. As shown in Wang et Isola, this term has the same minimizer as the L_{unif} loss when the number of negatives tends to be infinite. We decided to use the L_{unif} loss instead of the Contrastive loss because it is computationally less expensive, and it has been shown by Wang et Isola to lead to good representations and good downstream task performance. Furthermore, we have also compared the two losses in the CIFAR10-MNIST dataset the CelebA accessories (see Tab. D.12 and Tab. D.13) and found that the results are slightly better or similar using L_{unif}.

Table D.12: Balanced Accuracy results on Digits watermarked on CIFAR-10. Comparison between $\mathcal{L}_{\text{unif}}$ and $\mathcal{L}_{\text{log-sum-exp}}$ to estimate and minimize $H(C)$ and $H(S)$.

	Digits		Objects		$\delta \downarrow$
	S $\uparrow$	C $\downarrow$	S $\downarrow$	C $\uparrow$	
SepCLR-k-JEM ($\mathcal{L}_{\text{unif}}$)	96.2	11.0	10.4	73.2	32.0
SepCLR-k-JEM (log-sum-exp)	96.6	11.6	11.0	71.6	34.4
Best expected	100.0	10.0	10.0	100.0	0.0

D.7.7 On the impact of the encoders

In this section, we justify our choices in terms of architecture. In Tab. D.14 and Tab. D.15, we show the performance of SepVAE and SepCLR on Digits watermarked on CIFAR-10 and CelebA with accessories with different architectures. SepVAE with ResNet-18 performs less better or similarly than the one described in our original paper. Conversely, SepCLR with ResNet-18

Table D.13: Balanced Accuracy results on CelebA accessories.
Comparison between $\mathcal{L}_{\text{unif}}$ and $\mathcal{L}_{\text{log-sum-exp}}$ to estimate and minimize $H(C)$ and $H(S)$.

	Digits		Objects		$\delta \downarrow$
	S $\uparrow$	C $\downarrow$	S $\downarrow$	C $\uparrow$	
SepCLR-k-JEM ($\mathcal{L}_{\text{unif}}$)	98.57	55.21	62.52	78.00	41.16
SepCLR-k-JEM (log-sum-exp)	98.73	55.06	61.36	76.94	**40.75**
Best expected	100.0	10.0	10.0	100.0	0.0

performs better. Overall, SepCLR remains largely better than SepVAE, a consistent method across Contrastive Analysis VAEs.

Table D.14: Results of several different encoder architectures on Digits watermarked on CIFAR (B-ACC).

	Digits		Objects		$\delta_{\text{tot}} \downarrow$
	S $\uparrow$	C $\downarrow$	S $\downarrow$	C $\uparrow$	
SepVAE	90.6	17.8	10.6	36.6	81.2
SepVAE - ResNet 18 encoder	90.8	23.2	10.2	34.0	88.24
SepCLR with SepVAE's encoder	75.6	28.8	16.2	52.6	96.8
SepCLR-k-JEM	96.2	11.0	10.4	73.2	**32.0**
Best expected	100.0	10.0	10.0	100.0	0.0

Table D.15: Results of several different encoder architectures on CelebA accessories (B-ACC).

	Hats/Glss		Sex		$\delta_{\text{tot}} \downarrow$
	S $\uparrow$	C $\downarrow$	S $\downarrow$	C $\uparrow$	
SepVAE	84.46	65.19	60.12	59.20	81.65
SepVAE with ResNet-18 encoder	86.13	67.47	60.04	61.93	81.45
SepCLR - k-JEM with SepVAE's encoder	97.89	60.01	51.07	70.51	42.68
SepCLR - k-JEM	98.57	55.21	62.52	78.00	**41.16**
Best Expected	100.0	50.0	50.0	100.0	0.0

D.7.8 dSprites element superimposed on a digit grid

We show supplementary qualitative results on the salient space disentanglement in Fig. D.3. We qualitatively show that the common space captures background variability rather than foreground variability in Fig. D.4. We qualitatively show that the salient space captures foreground variability rather than background variability in Fig. D.5.

Figure D.3: Attribute Supervised SepCLR on dSprites superimposed on a digits grid. Given an image, sampling of the nearest neighbor images in the latent space given small displacement given an axis of the salient space. Each row represents the variation of only one element of the salient factor s while keeping c fixed. We can see a certain disentanglement: shape (line 1), zoom (line 2), orientation (line 3), X and Y position (lines 4 and 5).

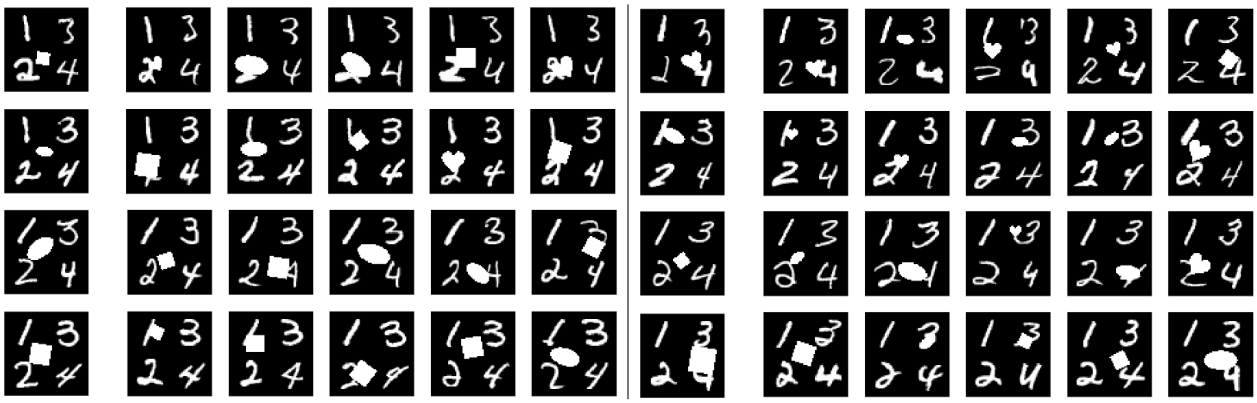


Figure D.4: Attribute Supervised SepCLR on dSprites superimposed on a digits grid. Given random target images on the left, we sample the nearest neighbors in the dataset with respect to their L2 distance in the common space only. We can see that the dSprite object remains the same while the MNIST digit grid in the background changes across the neighbors.

We also show quantitative results in Tab. D.16 by computing the Mutual Information Gap (MIG) score to measure the disentanglement in the DSprite-MNIST experiment. Results are reported below, and it can be noticed that the proposed method obtains good results (MIG is bounded by 0 and 1, where 1 indicates a perfect result).

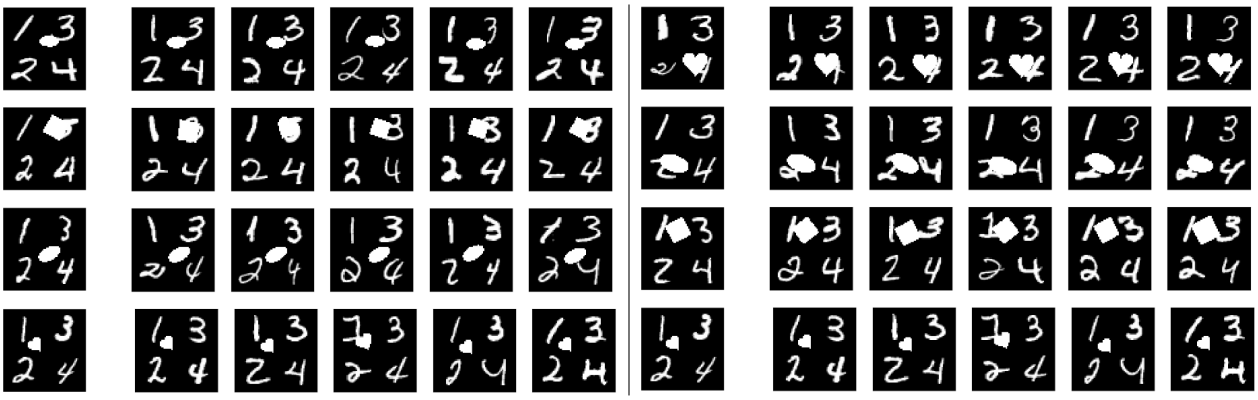


Figure D.5: Attribute Supervised SepCLR on dSprites superimposed on a digits grid. Given random target images on the left, we sample the nearest neighbors in the dataset with respect to their L2 distance in the salient space only, we can see that the dSprite object remains the same while the MNIST digits in the background change across the neighbors.

Table D.16: Computation of MIG on the salient space in dSprites on MNIST digit grid experiment.

	Z1 (Shape)	Z2 (Zoom)	Z3 (Rotation)	Z4 (Trans X)	Z5 (Trans
Best Expected	1	1	1	1	1
Attr Sup SepCLR - k-JEM	0.915	0.909	0.674	0.823	0.835
Random Vector	0.002	0.003	0.007	0.007	0.0008

D.7.9 Qualitative results on CelebA with accessories

In this section, we propose to display qualitative results on the CelebA accessories dataset. In Fig. D.6 and Fig. D.7, we propose to respectively display a 2D t-SNE plot for the salient and common latent space of SepCLR-k-JEM on the target dataset (portraits of celebrities with accessories). Yellow points represent people with hats, Purple points represent people with glasses. We can clearly see that our method has correctly encoded the patterns related to the accessories in the salient space and not in the common space.

Figure D.6: 2D t-SNE plot of the salient space of SepCLR-k-JEM on CelebA with accessories. We highlight in yellow and purple the actual labeled subgroups (people with hats or with glasses), respectively. We can see that the two subgroups are clearly separated in the salient space. Furthermore, we train a K-Means (K=2), which successfully identifies the two subgroups, and we propose to display the 6 nearest images from both centroids. Interestingly, we observe various backgrounds, poses, and people of different genders but with the same accessories (hats in cluster 0 and glasses in cluster 1). This clearly shows that our method has correctly encoded the patterns related to the accessories in the salient space and not the general ones (e.g., background, pose, gender, etc.).

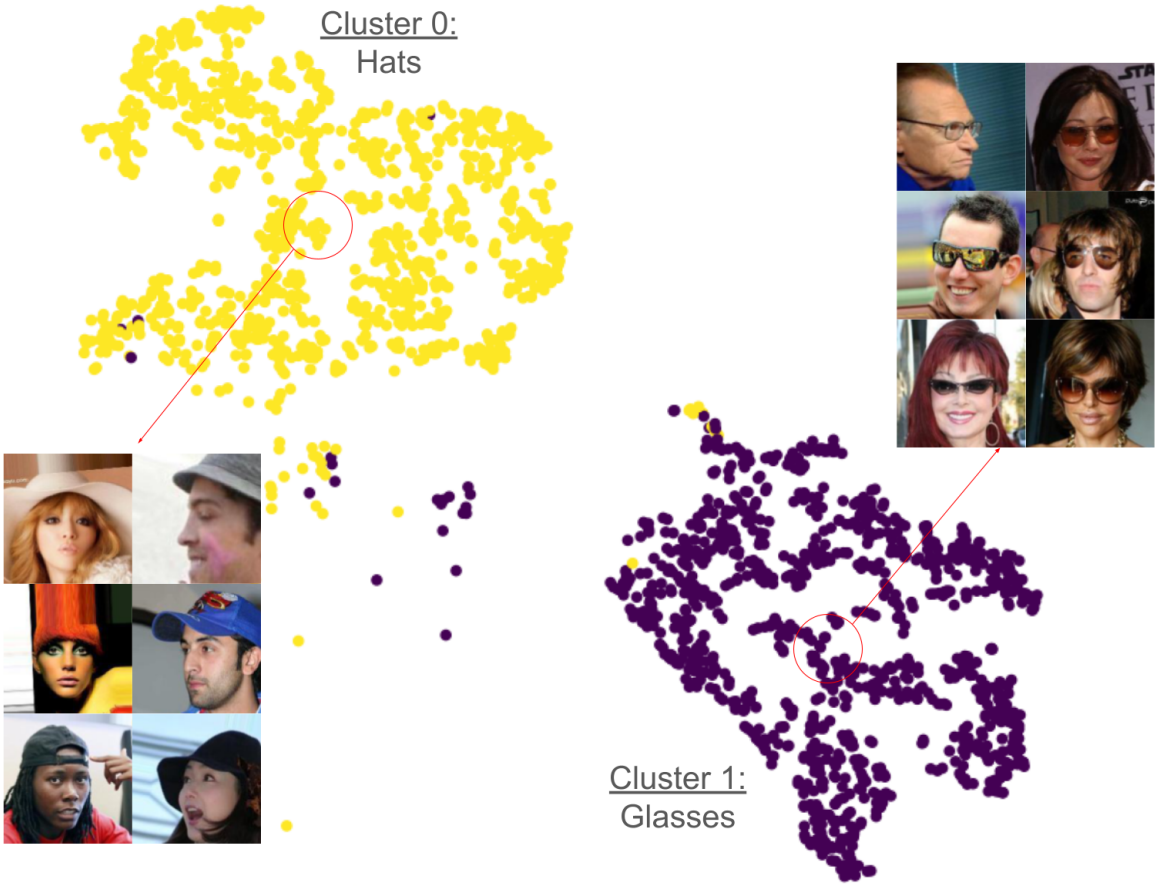
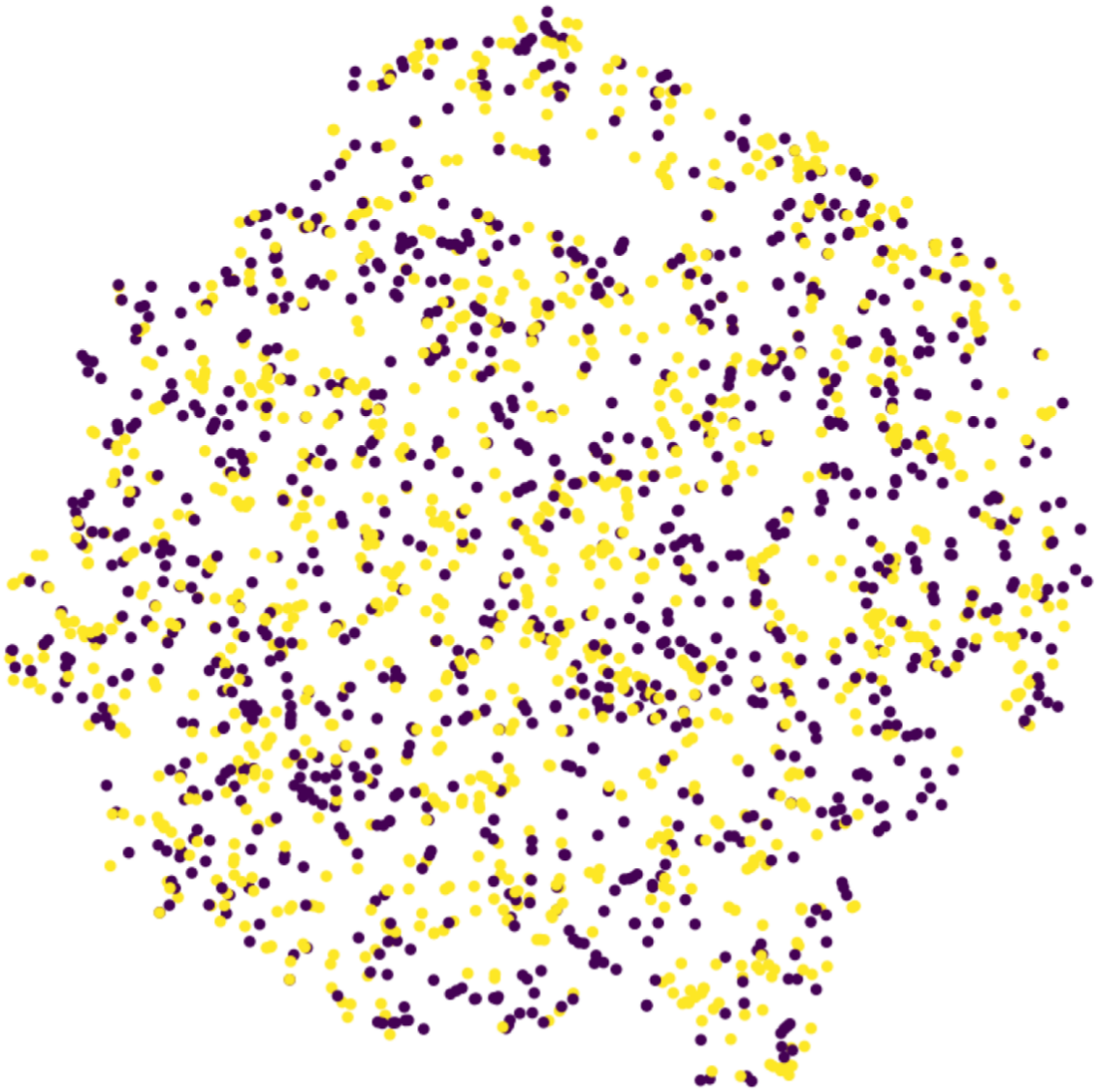


Figure D.7: 2D t-SNE plot of the common space of SepCLR-k-JEM on CelebA with accessories (target dataset). We highlight in yellow and purple the actual labeled subgroups (people with hats or with glasses), respectively. We can see that the two subgroups overlap in the common space. This clearly confirms that our method does not encode the patterns related to the accessories in the common space.



Bibliography

[1] Abi-Dargham, A., Horga, G. "The search for imaging biomarkers in psychiatric disorders". in *Nat Med* n. 22, pp. 1248-–1255, 2016.

[2] Abi-Dargham, A., Moeller, S.J., Ali, F., DeLorenzo, C., Domschke, K., Horga, G., Jutla, A., Kotov, R., Paulus, M.P., Rubio, J.M., Sanacora, G., Veenstra-VanderWeele, J. and Krystal, J.H. "Candidate biomarkers in psychiatric disorders: state of the field." in *World Psychiatry*, n.22, pp:236–262, 2023.

[3] Abid, A. and Zou, J. Contrastive Variational "Autoencoder Enhances Salient Features", in *arXiv:1902.04601 [cs, stat]*. Feb. 2019.

[4] Abid, A., Zhang, M. J., Bagaria, V. K., and Zou, J. "Exploring patterns enriched in a dataset with contrastive principal component analysis". in *Nature Communications*, vol. 9(1), pp. 2134, May 2018.

[5] A. Abrol, Z. Fu, M. Salman, R. Silva, Y. Du, S. Plis, and V. Calhoun. "Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning". in *Nature communications*, vol. 12(1). pp. 1–17, 2021.

[6] Aidas Aglinskas, Joshua K. Hartshorne, and Stefano Anzellotti. "Contrastive machine learning reveals the structure of neuroanatomical variation within autism". in *Science*, vol. 376(6597), pp. 1070–1074, 2022.

[7] A.L. Aguila, J. Chapman and A. Altmann "Multi-modal Variational Autoencoders for Normative Modelling Across Multiple Imaging Modalities" in *Medical Image Computing and Computer Assisted Intervention,* 2023.

[8] Ahmad and Pi-Erh Lin. "A nonparametric estimation of the entropy for absolutely continuous distributions". in *IEEE Transactions on Information Theory*, vol; 22(3), pp. 372–375, May 1976.

[9] Ainsworth, S. K., Foti, N. J., Lee, A. K. C., and Fox, E. B. oi-VAE: Output Interpretable VAEs for Nonlinear Group Factor Analysis. in *Proceedings of the 35th International Conference on Machine Learning*, pp. 119–128. Jul. 2018.

[10] Akshoomoff, N., Corsello, C., and Schmidt, H. "The Role of the Autism Diagnostic Observation Schedule in the Assessment of Autism Spectrum Disorders in School and Community Settings". in *The California school psychologist.* vol. 11 pp. 7–19, 2006.

[11] Alexander A. Alemi. "Variational Predictive Information Bottleneck". in *PMLR*, pp. 1–6, Feb. 2020.

[12] Antelmi, L., Ayache, N., Robert, P., and Lorenzi, M. "Sparse multi-channel variational autoencoder for the joint analysis of heterogeneous data". in *Proceedings of the 36th International Conference on Machine Learning*, vol. 97, pp. 302–311, Jun. 2019.

[13] C. Ambroise, A. Grigis, E. Duchesnay, and V. Frouin, "Multi-View Variational Autoencoders Allow for Interpretability Leveraging Digital Avatars: Application to the HBN Cohort", in *IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, 2023.

[14] Arbabshirani, Mohammad R., Sergey Plis, Jing Sui, and Vince D. Calhoun. "Single Subject Prediction of Brain Disorders in Neuroimaging: Promises and Pitfalls." in *NeuroImage.*

[15] Arnedo, J. et al. "Uncovering the hidden risk architecture of the schizophrenias: confirmation in three independent genome-wide association studies". in *Am. J. Psychiatry* vol. 172, pp. 139–153, 2015.

[16] Asano, Y.M., Rupprecht, C., Vedaldi, A., "Self-labelling via simultaneous clustering and representation learning." in *Proc. Int. Conf. Learn. Represent. ICLR,* May. 2020.

[17] Ashburner, J., Friston, K.J., "Voxel-based morphometry–the methods." in *NeuroImage,* vol. 11, no. 4 Pt 1., Jun. 2000, pp. 805-–21

[18] Ashburner J, Friston KJ. "Unified segmentation." in *Neuroimage.* n. 26(3), pp. 839–51, Jul. 2005.

[19] Ashburner J. "A fast diffeomorphic image registration algorithm." in *Neuroimage.* n. 38(1), pp. 95–113, Oct. 2007.

[20] Atzmueller, M. "Subgroup discovery." in *WIREs Data Mining and Knowledge Discovery* vol. 5. no. 1., pp. 35–49, 2015

[21] Philip Bachman, R. Devon Hjelm, and William Buchwalter. Learning Representations by Maximizing Mutual Information Across Views. in *Advances in Neural Information Processing Systems*. Jun 2019.

[22] H. Bahng, S. Chun, S. Yun, J. Choo, and S. J. Oh. "Learning de-biased representations with biased representations". in *International Conference on Machine Learning*, pp. 528–539, 2020.

[23] C.A. Barbano et al., "EnD: Entangling and Disentangling deep representations for bias correction" in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[24] Carlo Alberto Barbano, Benoit Dufumier, Enzo Tartaglione, Marco Grangetto, and Pietro Gori. "Unbiased Supervised Contrastive Learning". in *International Conference on Learning Representations (ICLR)*, 2023.

[25] C.A. Barbano, B. Dufumier, E. Duchesnay, M. Grangetto, P. Gori, "Contrastive learning for regression in multi-site brain age prediction" in *International Symposium on Biomedical Imaging (ISBI)*, 2023.

[26] Baur, C., Denner, S., Wiestler, B., Albarqouni, S., and Navab, N. "Autoencoders for Unsupervised Anomaly Segmentation in Brain MR Images: A Comparative Study". in *arXiv:2004.03271 [cs, eess]*, 2020.

[27] J.M.M Bayer, "Accommodating site variation in neuroimaging data using normative and hierarchical Bayesian models", in *Neuroimaging*, vol.264, 2022.

[28] Suzanna Becker and Geoffrey E. Hinton. Self-organizing neural network that discovers surfaces in random-dot stereograms. in *Nature*, vol. 355(6356). pp. 161-–163, 1992.

[29] Y. Bengio, P. Simard and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," in *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157-166, March 1994.

[30] Yoshua Bengio, Aaron Courville, and Pascal Vincent. "Representation learning: A review and new perspectives". in *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35(8), pp. 1798–1828, August 2013.

[31] A. J. Bell and T. J. Sejnowski. "An information-maximization approach to blind separation and blind deconvolution". in *Neural Computation*, vol. 7(6), pp. 1129–1159, Nov. 1995.

[32] Bell M.D., Corbera S., Johannesen J.K., Fiszdon J.M., Wexler B.E. "Social cognitive impairments and negative symptoms in schizophrenia: Are there subtypes with distinct functional correlates?" in *Schizophr Bull.* vol. 39, pp. 186–196, 2013.

[33] Y. Bengio. "Deep learning of representations for unsupervised and transfer learning". in *Proceedings of ICML workshop on unsupervised and transfer learning*, pp. 17–36, 2012.

[34] Buch, A.M, Liston C. Dissecting diagnostic heterogeneity in depression by integrating neuroimaging and genetics. in *Neuropsychopharmacology.* n. 46(1), pp. 156-175, Jan; 2021.

[35] Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., and Lerchner, A. "Understanding disentangling in beta-VAE". in *arXiv:1804.03599 [cs, stat]*, Apr. 2018.

[36] Brodersen K.H., Deserno L., Schlagenhauf F., Lin Z., Penny W.D., Buhmann J.M., et al. "Dissecting psychiatric spectrum disorders by generative embedding." in *Neuroimage Clin.* vol. 4, pp. 98–111, 2013.

[37] Bzdok D., Meyer-Lindenberg A. "Machine Learning for Precision Psychiatry: Opportunities and Challenges.", in *Biol Psychiatry Cogn Neurosci Neuroimaging.* n. 3(3) pp. 223–230, Mar. 2018.

[38] Califf, R.M. "Biomarker definitions and their applications". in *Exp Biol Med.* 243(3), pp. 213–221, Feb. 2018.

[39] Camino de Juan Romero, Víctor Borrell, "Genetic maps and patterns of cerebral cortex folding", in *Current Opinion in Cell Biology*, vol. 49, pp 31–37, 2017.

[40] Caron, M., Bojanowski, P., Joulin, A., Douze, M. "Deep Clustering for Unsupervised Learning of Visual Features." in *Europ. Conc. in Comp. Vis. ECCV.*, 2020

[41] Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A., "Unsupervised Learning of Visual Features by Contrasting Cluster Assignments." in *Proc. Adv. Neural Inf. Process. Syst. NeurIPS.,* Dec. 2020.

[42] Carton, Florence., Louiset, Robin., and Gori, Pietro. "Double InfoGAN for Contrastive Analysis" in *Artificial Intelligence and Statistic AISTATS*, 2024.

[43] Carey, L.A. et. al. "Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study." in *JAMA*, vol. 295 no. 21, pp. 2492–2502, 2006

[44] R. Caruana. "Multitask learning". in *Machine learning*, vol. 28(1), pp. 41–75, 1997.

[45] Chand, G.B., et al., "Two distinct neuroanatomical subtypes of schizophrenia revealed using machine learning." in *Brain,* vol. 143, no. 3, 2020, pp. 1027–1038.

[46] Chattopadhay, Aditya and Sarkar, Anirban and Howlader, Prantik and Balasubramanian, Vineeth N. "Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks." in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018.

[47] Xi Chen, Diederik P. Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. "Variational Lossy Autoencoder". in *International Conference on Learning Representations*, 2017.

[48] Ricky T. Q. Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. "Isolating Sources of Disentanglement in Variational Autoencoders". in *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[49] Chen, R. T. Q., Li, X., Grosse, R., and Duvenaud, D. Isolating Sources of Disentanglement in Variational Autoencoders. in *Advances in Neural Information Processing Signal*, Apr. 2019.

[50] Chen, T., Kornblith, S., Norouzi, M., Hinton, G., "A Simple Framework for Contrastive Learning of Visual Representations." in *Proc. Int. Conf. Mach. Learn. ICML,* Jul. 2020.

[51] Xinlei Chen and Kaiming He. "Exploring Simple Siamese Representation Learning", in *Computer Vision and Pattern Recognition (CVPR 2020)*, November 2020.

[52] Chen, X., Fan, H., Girshick, R., He, K., "Improved Baselines with Momentum Contrastive Learning." in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. CVPR.* Jun. 2020.

[53] Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and L. Carin. "CLUB: A Contrastive Log-ratio Upper Bound of Mutual Information". in *International Conference on Learning Representations.* June 2020.

[54] C. R. Ching, D. P. Hibar, T. P. Gurholt, A. Nunes, S. I. Thomopoulos, C. Abe, I. Agartz, R. M. Brouwer, D. M. Cannon, S. M. de Zwarte, et al. "What we learn about bipolar disorder from large-scale neuroimaging: Findings and future directions from the enigma bipolar disorder working group". in *Human brain mapping*, vol. 43(1), pp. 56-–82, 2022.

[55] Choudhuri, A., Makkuva, A. V., Rana, R., Oh, S., Chowdhary, G., and Schwing, A. "Towards Principled Objectives for Contrastive Disentanglement". Dec. 2019.

[56] C. Clark, M. Yatskar, and L. Zettlemoyer. "Don't take the easy way out: Ensemble based methods for avoiding known dataset biases". in *preprint arXiv:1909.03683*, 2019.

[57] Clementz, B.A., Sweeney, J.A., Hamm, J.P., Ivleva, E.I., Ethridge, L.E., Pearlson, G.D., Keshavan, M.S., Tamminga, C.A. "Identification of Distinct Psychosis Biotypes Using Brain-Based Biomarkers". in *Am J Psychiatry.* Apr. 2016.

[58] Costa Dias T.G., Iyer S.P., Carpenter S.D., Cary R.P., Wilson V.B., Mitchell S.H. et al. "Characterizing heterogeneity in children with and without ADHD based on reward system connectivity." in *Dev Cogn Neurosci.* vol. 11, pp. 155–174, 2015.

[59] Cuturi, M.: "Sinkhorn Distances: Lightspeed Computation of Optimal Transport." in *Proc. Adv. Neural Inf. Process. Syst. NeurIPS.,* Dec. 2013.

[60] Dang, Z., Deng, C., Yang, X., Wei, K., Huang, H., "Nearest Neighbor Matching for Deep Clustering." in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. CVPR.*. Jun. 2021. pp. 13688— 13697.

[61] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. "Imagenet: A large-scale hierarchical image database". in *IEEE conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

[62] Dong, A., Honnorat, N., Gaonkar, B. Davatzikos, C. "CHIMERA: Clustering of Heterogeneous Disease Effects via Distribution Matching of Imaging Patterns". in *IEEE Trans. Med. Imaging.* vol. 35, pp. 612–621 2016.

[63] De Pierrefeu, Amicie, Thomas Fovet, Fouad Hadj-Selem, Tommy Löfstedt, Philippe Ciuciu, Stephanie Lefebvre, Pierre Thomas, Renaud Lopes, Renaud Jardri, and Edouard Duchesnay. "Prediction of Activation Patterns Preceding Hallucinations in Patients with Schizophrenia Using Machine Learning with Structured Sparsity." in *Human Brain Mapping* vol. 39(4), pp. 1777–1788, 2018.

[64] Derks, E. M. et al. "Kraepelin was right: a latent class analysis of symptom dimensions in patients and controls". in *Schizophr. Bull.* vol. 38, pp. 495–505, 2012.

[65] Shuangrui Ding, Maomao Li, Tianyu Yang, Rui Qian, Haohang Xu, Qingyi Chen, Jue Wang, and Hongkai Xiong. "Motion-aware Contrastive Video Representation Learning via Foreground-background Merging", in *Computer Vision and Pattern Recognition (CVPR 2022)* March 2022.

[66] R. Dinga, et al. "Evaluating the evidence for biotypes of depression: Methodological replication and extension of Drysdale et al. (2017)" in *NeuroImage: Clinical*, vol. 22, 2019.

[67] A. Di Martino, C.-G. Yan, Q. Li, E. Denio, F. X. Castellanos, K. Alaerts, J. S. Anderson, M. Assaf, S. Y. Bookheimer, M. Dapretto, et al. "The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism". in *Molecular psychiatry*, vol. 19(6), pp. 659–667, 2014.

[68] A. Di Martino, D. O'connor, B. Chen, K. Alaerts, J. S. Anderson, M. Assaf, J. H. Balsters, L. Baxter, A. Beggiato, S. Bernaerts, et al. "Enhancing studies of the connectome in autism using the autism brain imaging data exchange". in *Scientific data*, vol. 4(1), pp. 1–15, 2017.

[69] American Psychiatric Association. "Diagnostic and statistical manual of mental disorders." in *American Psychiatric Association.* vol. 5, 2013.

[70] American Psychiatric Association. "Diagnostic and statistical manual of mental disorders, fifth edition, text revision". in *American Psychiatric Association*, 2022.

[71] Andrew T. Drysdale, Logan Grosenick et al. "Resting-state connectivity biomarkers define neurophysiological subtypes of depression." in *Nat Med.* 23(1). pp. 28–38. 2017.

[72] Duchesnay Edouard, "Neuroimaging Signatures of Brain Disorders: Fighting Overfitting in Predictive Models." in *Habilitation à Diriger des Recherches*, 2019.

[73] Dufumier B., Gori P., Battaglia I., Victor J., Grigis A., and Duchesnay E., "Benchmarking CNN on 3D Anatomical Brain MRI: Architectures, Data Augmentation and Deep Ensemble Learning" in *arXiv.2106.01132 [cs.CV]*, 2023.

[74] B. Dufumier, P. Gori, A. Grigis, and E. Duchesnay. "Contrastive Learning with Continuous Proxy Meta-data for 3D MRI Classification". in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021* pp. 58–68, 2021a.

[75] B. Dufumier, Pietro Gori, Julie Victor, Antoine Grigis, and Edouard Duchesnay. "Conditional Alignment and Uniformity for Contrastive Learning with Continuous Proxy Labels". in *MedNeurIPS, Workshop NeurIPS*, 2021b.

[76] Dufumier, B. "Representation learning in neuroimaging : transferring from big healthy data to small clinical cohorts." Thesis Manuscript. 2022.

[77] B. Dufumier., C.A. Barbano, R. Louiset, E. Duchesnay, and P. Gori "Integrating Prior Knowledge in Contrastive Learning with Kernel". in *International Conference on Machine Learning*, 2023.

[78] Dwyer, D. B. et al. "Brain Subtyping Enhances The Neuroanatomical Discrimination of Schizophrenia". in *Schizophr. Bull.* vol. 44, pp. 1060–1069, 2018.

[79] H. Eavani, M.K. Hsieh, Y. An, G. Erus, L. Beason-Held, S. Resnick, et al. "Capturing heterogeneous group differences using mixture-of-experts: Application to a study of aging Neuroimage", vol. 125, pp. 498–514, 2016.

[80] Ecker, C., et al. "Investigating the predictive value of whole-brain structural MR scans in autism: A pattern classification approach.", in *NeuroImage*, vol. 49(1), pp. 44–56, 2010.

[81] Ecker, C., et al. "Interindividual Differences in Cortical Thickness and Their Genomic Underpinnings in Autism Spectrum Disorder." in *Am J Psychiatry.* vol. 179(3), pp. 242–254, Mar. 2022.

[82] Pérez-Enciso, Zingaretti, Laura. "A Guide for Using Deep Learning for Complex Trait Genomic Prediction". in *Genes.* vol. 10. pp. 553. 2019.

[83] European Medicines Agency (EMA). "Definitions for genomic biomarkers, pharmacogenomics, pharmacogenetics, genomic data and sample coding categories". 2007.

[84] Erro, R., Vitale, C., Amboni, M., Picillo, M., et al. "The heterogeneity of early Parkinson's disease: a cluster analysis on newly diagnosed untreated patients". in *PLoS One* vol. 8(8), 2013.

[85] Fair D.A., Bathula D., Nikolas M.A., Nigg J.T., "Distinct neuropsychological subgroups in typically developing youth inform heterogeneity in children with ADHD." in *Proc. Natl. Acad. Sci.* vol. 109, pp. 6769–6774, 2012.

[86] Farmer, A.E., McGuffin, P., Spitznagel, E.L. "Heterogeneity in schizophrenia: a cluster-analytic approach." in *Psychiatry Res.* n.8(1), pp. 1–12, Jan. 1983.

[87] Favre, P., Pauling, M., Stout, J. et al. "Widespread white matter microstructural abnormalities in bipolar disorder: evidence from mega- and meta-analyses across 3033 individuals". Neuropsychopharmacol. vol. 44, pp. 2285–2293, 2019.

[88] FDA-NIH Biomarker Working Group. "BEST (Biomarkers, EndpointS, and other Tools) resource". *Silver Spring: US Food and Drug Administration*, 2016.

[89] Eric, Feczko., Oscar, M.D., Mollie, M., Alice, M. G. Joel, T. N., Damien, A. F., "The Heterogeneity Problem: Approaches to Identify Psychiatric Subtypes", in *Trends in Cognitive Sciences*, vol. 23(7), pp. 584–601, 2019.

[90] Ferreira, D., Verhagen, C., Hernandez-Cabrera, J.A., Cavallin, L., et al. "Distinct subtypes of Alzheimer's disease based on patterns of brain atrophy: longitudinal trajectories and clinical applications." in *Scientific Report* vol. 7, 2017.

[91] R. Filipovych, S.M. Resnick, C. Davatzikos. "JointMMCC: Joint maximum-margin classification and clustering of imaging data." in *IEEE Transactions on Medical Imaging*, vol. 31, pp. 1124–1140, 2012.

[92] Fitzgerald, G.A. "Measure for Measure: Biomarker standards and transparency". in *Sci Trans Med* n.8, pp.343, 2016.

[93] Forgy, Edward W. "Cluster analysis of multivariate data: efficiency versus interpretability of classifications". in *Biometrics.* vol. 21(3), pp. 768–769, 1962.

[94] Fortin, J.-P. et al. "Harmonization of cortical thickness measurements across scanners and sites". in *NeuroImage* vol. 167, pp. 104–120, 2018.

[95] Fortin, J.-P. et al. "Harmonization of multi-site diffusion tensor imaging data". in *NeuroImage.* vol. 161, pp. 149–170, 2017.

[96] García-Gutiérrez, M.S., Navarrete, F., Sala, F., Gasparyan, A., Austrich-Olivares, A, Manzanares. J. "Biomarkers in Psychiatry: Concept, Definition, Types and Relevance to the Clinical Reality". in *Frontiers in Psychiatry.* n.11, pp. 432, May. 2020.

[97] C. Gaser, K. Franke, S. Kloppel, N. Koutsouleris, H. Sauer, and A. D. N. Initiative. "Brain age in mild cognitive impaired patients: predicting the conversion to alzheimer's disease". in *PloS one*, vol. 8(6), pp. 67346, 2013.

[98] Kurth, F., Luders, E. and Gaser, C. "Voxel-Based Morphometry", in *Brain Mapping*, pp. 345-–349. 2015.

[99] Christian, G., Robert, D., Paul, M. T., Florian, K., Eileen, L., Alzheimer's Disease Neuroimaging Initiative, "CAT – A Computational Anatomy Toolbox for the Analysis of Structural MRI Data - new results" in *bioRxiv.*, Jun. 2022.

[100] Ge, R. and Zou, J. Rich component analysis. in *Proceedings of the 33rd International Conference on International Conference on Machine Learning*. vol. 48, pp. 1502–1510, Jun. 2016.

[101] Genevay, A., Dulac-Arnold, G., Vert, J.P. "Differentiable Deep Clustering with Cluster Size Constraints." in *arXiv:1910.09036 [cs, stat]*, Oct. 2019.

[102] Behnam Gholami, Pritish Sahu, Ognjen Rudovic, Konstantinos Bousmalis, and Vladimir Pavlovic. "Unsupervised Multi-Target Domain Adaptation: An Information Theoretic Approach", in *IEE TIP* October 2018.

[103] Glocker, B., Robinson, R., Castro, D. C., Dou, Q., Konukoglu, E. "Machine Learning with Multi-Site Imaging Data: An Empirical Study on the Impact of Scanner Effects". in *ArXiv191004597 cs Eess Q-Bio*, 2019.

[104] X. Glorot, A. Bordes, and Y. Bengio. "Deep sparse rectifier neural networks". in *Proceedings of the fourteenth international conference on artificial intelligence and statistics, JMLR*, pp. 315-–323, 2011.

[105] Gobinath, A.R., Choleris, E. and Galea, L.A. "Sex, hormones, and genotype interact to influence psychiatric disease, treatment, and behavioral research," in *Journal of neuroscience research*, 95(1-2), pp.50-64, 2017.

[106] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S. Courville, A., Bengio, Y. "Generative Adversarial Nets". in *Proceedings of the International Conference on Neural Information Processing Systems*. pp. 2672–2680, 2014.

[107] N. Gornitz, M. Kloft, K. Rieck, and U. Brefeld, "Toward Supervised Anomaly Detection", in *Journal of Artificial Intelligence Research*, vol. 46, pp. 235-262, 2013.

[108] Goyal, P., Mathilde C., Benjamin L., Min X., Pengchao W., Vivek P., Mannat S, Vitaliy L., Ishan M., Armand J. and Piotr B.. "Self-supervised Pretraining of Visual Features in the Wild." in *ArXiv abs/2103.01988*, 2021.

[109] Grill, J.B. et al. "Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning." in *Advances in Neural Information Processing Systems.* vol. 33, pp. 21271–21284, 2020

[110] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Scholkopf, and Alexander Smola. "A kernel two-sample test". *The Journal of Machine Learning Research*, vol; 13, pp. 723–773, March 2012.

[111] Green, I. W. Glausier, J. R. "Different Paths to Core Pathology: The Equifinal Model of the Schizophrenia Syndrome". in *Schizophr. Bull.* vol. 42, pp. 542–549, 2016.

[112] Haber, A. et al., "A single-cell survey of the small intestinal epithelium". in *Nature*, vol. 551(7680), pp. 333–339, Nov. 2017.

[113] Hajek, T., Gunde, E., Slaney, C., Propper, L., MacQueen, G., Duffy, A., Alda, M. "Amygdala and hippocampal volumes in relatives of patients with bipolar disorder: a high-risk study". in *Can J Psychiatry.* vol. 54(11), pp. 726–733, Nov. 2009.

[114] Hajek, T., Kopecek, M., Höschl, C., Alda, M. "Smaller hippocampal volumes in patients with bipolar disorder are masked by exposure to lithium: a meta-analysis". in *J Psychiatry Neurosci.* vol. 37(5), pp. 333–343, Sep. 2012.

[115] L. K. Han, R. Dinga, T. Hahn, C. R. Ching, L. T. Eyler, L. Aftanas, M. Aghajani, A. Aleman, B. T. Baune, K. Berger, et al. "Brain aging in major depressive disorder: results from the enigma major depressive disorder working group". in *Molecular psychiatry*, pp. 1–16, 2020.

[116] He, K., Zhang, X., Ren, S., Sun, J. "Deep Residual Learning for Image Recognition." in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. CVPR.* Jun. 2016. pp. 770—778.

[117] Yihui He, Chenchen Zhu, Jianren Wang, Marios Savvides, and Xiangyu Zhang. "Bounding Box Regression With Uncertainty for Accurate Object Detection". in *Computer Vision and Pattern Recognition* pp. 2883–2892, June 2019.

[118] He, K., Fan, H., Wu, Y., Xie, S., Girshick, R. "Momentum Contrast for Unsupervised Visual Representation Learning." in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. CVPR.* pp. 9726–9735. Jun. 2020.

[119] Heinsfeld, A. S., Franco, A. R., Craddock, R. C., Buchweitz, A., and Meneguzzi, F. Identification of autism spectrum disorder using deep learning and the ABIDE dataset. in *NeuroImage : Clinical*, vol. 17. pp. 16–23, Aug. 2017.

[120] D. Hibar, L. T. Westlye, T. G. van Erp, J. Rasmussen, C. D. Leonardo, J. Faskowitz, U. K. Haukvik, C. B. Hartberg, N. T. Doan, I. Agartz, et al. "Subcortical volumetric abnormalities in bipolar disorder". in *Molecular psychiatry*, vol. 21(12), pp. 1710–1716, 2016.

[121] D. Hibar, L. T. Westlye, N. T. Doan, N. Jahanshad, J. Cheung, C. R. Ching, A. Versace, A. Bilderbeck, A. Uhlmann, B. Mwangi, et al. "Cortical abnormalities in bipolar disorder: an mri analysis of 6503 individuals from the enigma bipolar disorder working group". in *Molecular psychiatry*, vol. 23(4), pp. 932–942, 2018.

[122] Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. in *International Conference for Learning Representations*, 2017.

[123] R. Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. "Learning deep representations by mutual information estimation and maximization", in *arXiv:1808.06670 [cs, stat].* 2019.

[124] Jonathan Ho, Ajay Jain, Pieter Abbeel. "Denoising Diffusion Probabilistic Models." in *arXiv:2006.11239*, 2020.

[125] Michael P. Holmes, Alexander G. Gray, and Charles Lee Isbell. "Fast nonparametric conditional density estimation". in *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, pp. 175–182, July 2007.

[126] Honea, R., Crow, T.J., Passingham, D., Mackay, C.E. "Regional deficits in brain volume in schizophrenia: a meta-analysis of voxel-based morphometry studies". in *Am J Psychiatry.* vol. 162(12), pp. 2233–2245, Dec. 2005.

[127] Honnorat, N., Dong, A., Meisenzahl-Lechner, E., Koutsouleris, N., Davatzikos, C. "Neuroanatomical heterogeneity of schizophrenia revealed by semi-supervised machine learning methods." in *Schizophr. Res.* vol.214, pp. 43-50, 2019

[128] Hornik, K., Stinchcombe, M., White, H. "Multilayer feedforward networks are universal approximators", in *Neural Networks*, vol. 2(5), pp. 359–366, 1989.

[129] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. "Densely connected convolutional networks". in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

[130] Aapo Hyvarinen, Hiroaki Sasaki, and Richard E Turner. "Nonlinear ICA Using Auxiliary Variables and Generalized Contrastive Learning". in *AISTATS*, 2019.

[131] Harrison, J.E., Weber, S., Jakob, R., Chute, C.G. "ICD-11: an international classification of diseases for the twenty-first century". in *BMC Med Inform Decis Mak.* n. 21, pp. 206, Nov. 2021.

[132] Iftimovici A, Marchi A, Férat V, Pruvost-Robieux E, Guinard E, Morin V, Elandaloussi Y, D'Halluin A, Krebs MO, Chaumette B, Gavaret M. "Electroencephalography microstates imbalance across the spectrum of early psychosis, autism, and mood disorders". in *Eur Psychiatry*, vol. 66(1), May. 2023.

[133] Insel TR, Quirion R. "Psychiatry as a clinical neuroscience discipline". in *JAMA* vol. 294(17), pp. 2221-–2224, 2005.

[134] T. Insel, B. Cuthbert, M. Garvey, R. Heinssen, D. S. Pine, K. Quinn, C. Sanislow, and P. Wang. "Research domain criteria (rdoc): toward a new classification framework for research on mental disorders", in *Am J Psychiatry.* vol. 167, pp. 748–751, 2010.

[135] Insel, T. R. Cuthbert, B. N. "Brain disorders? Precisely". in *Science* vol. 348, pp. 499–500, 2015.

[136] Jeremy Irvin et al. "CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison". in *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, pp.590–597, Jan. 2019.

[137] Galatzer-Levy, I.R., Bryant, R.A. "636,120 Ways to Have Posttraumatic Stress Disorder.", in *Perspect Psychol Sci.* n. 8(6), pp.651–62, Nov. 2013.

[138] Jack, C.R. Jr., Petersen, R.C., Xu, Y.C., Waring, S.C., O'Brien, P.C., Tangalos, E.G., Smith, G.E., Ivnik, R.J., Kokmen, E. "Medial temporal atrophy on MRI in normal aging and very mild Alzheimer's disease". in *Neurology* n. 49(3), pp. 786–94, Sep. 1997.

[139] Jack, C. R., Bennett, D. A., Blennow, K., Carrillo, M. C., et al. "Toward a biological definition of Alzheimer's disease. Alzheimer's Dementia." in *The Journal of the Alzheimer's Association*, vol. 14(4), pp. 535–562, April 2018.

[140] Jones, A., Townes, F. W., Li, D., and Engelhardt, B. E. "Contrastive latent variable modeling with application to case-control sequencing experiments", in *Annals of Applied Statistics*, February 2021.

[141] Johnstone, E.C., Crow, T.J., Frith, C.D., Husband, J., Kreel, L. "Cerebral ventricular size and cognitive impairment in chronic schizophrenia". in *Lancet.* vol. 2, pp. 924–926, Oct. 1976.

[142] Joy, T., Schmon, S., Torr, P., Siddharth, N., Rainforth, T., "Capturing Label Characteristics in VAEs." in *Proc. Int. Conf. Learn. Represent. ICLR,* May. 2021.

[143] Karalunas S.L., Fair D., Musser E.D., Aykes K. Iyer S.P., Nigg J.T., "Subtyping attention-deficit/hyperactivity disorder using temperament dimensions toward biologically based nosologic criteria." in *JAMA Psychiatry.* vol. 71, pp. 1015–1024, 2014.

[144] Hadi Kazemi, Seyed Mehdi Iranmanesh, and Nasser Nasrabadi. "Style and Content Disentanglement in Generative Adversarial Networks". in *Winter Applications Computer Vision (WACV 2019)*, pp. 848–856, January 2019.

[145] Bohdan Kivva, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. "Identifiability of deep generative models without auxiliary information". in *Advances in Neural Information Processing Systems*, 2022.

[146] Kermany, D.S., Goldbaum, et al. "Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning." in *Cell* vol. 172, no.5, pp 1122–1131, Feb. 2018.

[147] Khemakhem, I., Kingma, D., Monti, R., and Hyvarinen, A. "Variational Autoencoders and Nonlinear ICA: A Unifying Framework". in *AISTATS*, 2020.

[148] S.M. Kia, and A.F. Marquand, "Neural Processes Mixed-Effect Models for Deep Normative Modeling of Clinical Neuroimaging Data", in *Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning, PMLR*, vol. 102, pp. 297-314, 2019.

[149] S.M. Kia et al. "Hierarchical Bayesian Regression for Multi-Site Normative Modeling of Neuroimaging Data" in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2020.

[150] Kingma, Diederik P. and Max Welling."Auto-Encoding Variational Bayes." in *CoRR abs/1312.6114*, Dec. 2013.

[151] Kingma, D. P., Mohamed, S., Rezende, D. J., and Welling, M. "Semi-supervised Learning with Deep Generative Models", in *in Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

[152] Kim, H. and Mnih, A. Disentangling by Factorising, in *Advances in Neural Information Processing Systems.* July 2019.

[153] Khosla, P., Tian, Y., Teterwak, P., Wang, C., Isola, P., Maschinot, A., Krishnan, D., Sarna, A. "Supervised Contrastive Learning." in *Proc. Adv. Neural Inf. Process. Syst. NeurIPS.,* Dec. 2020.

[154] Klosgen, W. "Explora: A multipattern and multistrategy discovery assistant." in *Advances in Knowledge Discovery and Data Mining*, pp. 249–271, Feb. 1996

[155] Lingpeng Kong, Cyprien de Masson d'Autume, Lei Yu, Wang Ling, Zihang Dai, and Dani Yogatama. "A Mutual Information Maximization Perspective of Language Representation Learning". in International Contrastive Learning of Representation, Sep. 2019.

[156] Koutsouleris, N. et al. "Structural correlates of psychopathological symptom dimensions in schizophrenia: a voxel-based morphometric study". in *NeuroImage* vol. 39, pp. 1600–1612, 2008.

[157] N. Koutsouleris, C. Davatzikos, S. Borgwardt, C. Gaser, R. Bottlender, T. Frodl, P. Falkai, A. Riecher-Rossler, H.-J. Moller, M. Reiser, et al. "Accelerated brain aging in schizophrenia and beyond: a neuroanatomical marker of psychiatric disorders". in *Schizophrenia bulletin*, vol. 40(5), pp. 1140–1153, 2014.

[158] Krizhevsky, Alex, Ilya Sutskever and Geoffrey E. Hinton. "ImageNet classification with deep convolutional neural networks." in *Communications of the ACM*, vol. 60, pp. 84–90, 2012.

[159] S. Kumar, and A. Sotiras, "NormVAE: Normative Modeling on Neuroimaging Data using Variational Autoencoders", preprint, under review., 2021.

[160] "Normative modeling using multimodal variational autoencoders to identify abnormal brain volume deviations in Alzheimer's disease" in *Workshop DGM4H in NeurIPS*, 2023.

[161] Kurth, F., Luders, E. "Voxel-Based Morphometry" in *Brain Mapping: An Encyclopedic Reference*, vol. 1, pp. 345–349, 2015.

[162] Lamers F., de Jonge P., Nolen W.A., Smit J.H., Zitman F.G., ATF Beekman, et al., "Identifying depressive subtypes in a large cohort study: Results from the Netherlands Study of Depression and Anxiety (NESDA)." in *J Clin Psychiatry.* vol. 71, pp. 1582-1589, 2010.

[163] L. M. Rimol, R. Nesvag, D. J. Hagler Jr, Ø. Bergmann, C. Fennema-Notestine, C. B. Hartberg, U. K. Haukvik, E. Lange, C. J. Pung, A. Server, et al. "Cortical volume, surface area, and thickness in schizophrenia and bipolar disorder". in *Biological psychiatry*, vol. 71(6): pp. 552—560, 2012.

[164] Lauterbur, P.C. "Image Formation by Induced Local Interactions: Examples Employing Nuclear Magnetic Resonance". in *Nature.* vol. 242(5394), pp. 190–191, 1973

[165] Le Bihan D. "Diffusion MRI: what water tells us about the brain". in *EMBO Mol Med.* vol. 6, pp. 569–573, 2014.

[166] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel. "Handwritten digit recognition with a back-propagation network". in *Advances in neural information processing systems*, vol. 2, 1989.

[167] Li, Y., Pan, Q., Wang, S., Peng, H., Yang, T., and Cambria, E. Disentangled Variational Auto-Encoder for Semi-supervised Learning. in *arXiv:1709.05047 [cs]*, Dec. 2018.

[168] Li, Zhiyuan et al. "Why Are Convolutional Nets More Sample-Efficient than Fully-Connected Nets?" in *Proc. Int. Conf. Learn. Represent. ICLR,* 2020.

[169] Li, J., Zhou, P., Xiong, C., Hoi, S.C.H.: "Prototypical Contrastive Learning of Unsupervised Representations." in *Proc. Int. Conf. Learn. Represent. ICLR,* May. 2021.

[170] Lieberman, DZ., Peele, R., Razavi, M.: "Combinations of DSM-IV-TR criteria sets for bipolar disorders." in *Psychopathology. 2008;41(1):35-8.*, Oct. 2007.

[171] Lin, H., and Jegelka, S. "ResNet with one-neuron hidden layers is a universal approximator". in *Proceedings of the 32nd International Conference on Neural Information Processing Systems.* pp. 6172–6181, 2018.

[172] L. C. Hanford, A. Nazarov, G. B. Hall, and R. B. Sassi. "Cortical thickness in bipolar disorder: a systematic review". in*Bipolar disorders*, vol. 18(1), pp. 4-–18, 2016.

[173] R. Linsker. "Self-organization in a perceptual network". in *Computer.* vol. 21(3), pp; 105–117, March 1988.

[174] Liu, Z., Luo, P., Wang, X., and Tang, X. "Deep Learning Face Attributes in the Wild", in *arXiv:1411.7766 [cs].* Sep 2015.

[175] Lloyd, Stuart P. "Least square quantization in PCM". in *Bell Telephone Laboratories Paper IEEE Transactions on Information Theory.* vol. 28(2), pp. 129–137, 1957.

[176] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Ratsch, Sylvain Gelly, Bernhard Scholkopf, and Olivier Bachem. "A Commentary on the Unsupervised Learning of Disentangled Representations". in *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(09), 2020.

[177] Lorenzetti, V., Allen, N.B., Fornito, A., Yucel, M. "Structural brain abnormalities in major depressive disorder: a selective review of recent MRI studies". in *Journal of Affect Disord.* n.117, pp. 1–17, Sep. 2009.

[178] M. Lorenzi, X. Pennec, G. B. Frisoni, N. Ayache, and Alzheimer's Disease Neuroimaging Initiative, "Disentangling normal aging from Alzheimer's disease in structural magnetic resonance images", in *Neurobiology of aging*, vol.36, pp. 42-52, 2015.

[179] Louiset, R., Gori, P., Dufumier, B., Houenou, J., Grigis, A., Duchesnay, E. "UCSL: A Machine Learning Expectation-Maximization Framework for Unsupervised Clustering Driven by Supervised Learning." in *Proc. Europ. Conf. on Machine Learn. and Princ. and Pract. of Know. Discov. in Data.*, vol. 12975, Jun. 2021.

[180] Robin Louiset, Edouard Duchesnay, Antoine Grigis, Benoit Dufumier, and Pietro Gori. "SepVAE: a contrastive VAE to separate pathological patterns from healthy ones", in *Workshop on Interpretable ML in Healthcare at International Conference on Machine Learning (ICML).* July 2023.

[181] Robin Louiset, Edouard Duchesnay, Antoine Grigis, and Pietro Gori. "Separating common from salient patterns with Contrastive Representation Learning", in *International Conference on Learning Representations ICLR* 2024.

[182] Lundberg, S.M., Erion, G.G., Lee, S.I. "Consistent Individualized Feature Attribution for Tree Ensembles". in *Workshop International Conference on Machine Learning*, 2017.

[183] Lundberg, S.M., Lee, S.I. "A unified approach to interpreting model predictions". in *Advances in Neural Information Processing Systems.* pp. 4768–4777, 2017.

[184] Lv, J., Kang, Z., Lu, X., Xu, Z. "Pseudo-Supervised Deep Subspace Clustering." in *IEEE Transactions on Image Processing*, 2021.

[185] Lydiard, J., Nemeroff, C.B., "Biomarker-Guided Tailored Therapy". in *Adv Exp Med Biol.* n. 1192, pp. 199–224, 2019.

[186] Nenadic, I., Sauer, H. Gaser, C. "Distinct pattern of brain structural deficits in subsyndromes of schizophrenia delineated by psychopathology". in *NeuroImage* vol. 49, pp. 1153–1160, 2010.

[187] Newson, J.J., Hunter, D., Thiagarajan, T.C. "The Heterogeneity of Mental Health Assessment". in *Front Psychiatry.* n. 11, pp. 76, Feb. 2020.

[188] Niculescu, A.B., Le-Niculescu, H. "Precision medicine in psychiatry: biomarkers to the forefront". in *Neuropsychopharmacology.* n. 47, pp. 422-–423, 2022.

[189] A. Nunes, H. G. Schnack, C. R. Ching, I. Agartz, T. N. Akudjedu, M. Alda, D. Alnæs, S. Alonso-Lana, J. Bauer, B. T. Baune, et al. "Using structural MRI to identify bipolar disorders–13 site machine learning study in 3020 individuals from the enigma bipolar disorders working group". in *Molecular psychiatry*, vol. 25(9), pp. 2130–2143, 2020.

[190] Maaten, L.v.d., Hinton, G. "Visualizing Data using t-SNE. Journal of Machine Learning Research." vol. 9(86), pp. 2579–2605, 2008.

[191] Maintz, J.B., Viergever, M.A. "A survey of medical image registration". in *Med Image Analysis.* n. 2, pp. 1–36, 1998.

[192] Maaløe, L., Sønderby, C. K., Sønderby, S. K., and Winther, O. "Auxiliary Deep Generative Models". in *International Conference on Machine Learning*, pp. 1445–1453, June 2016.

[193] Mansfield P, Grannell PK. "Diffraction and microscopy in solids and liquids by NMR". in *Physical Review B*. vol. 12(9), pp. 3618–34, 1975.

[194] Marquand, A.F., Wolfers, T., Mennes, M., Buitelaar, J., Beckmann, C.F.: "Beyond Lumping and Splitting. A Review of Computational Approaches for Stratifying Psychiatric Disorders." in *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*. vol. 1 no. 5, pp. 433–447, 2016.

[195] Marquand, A.F., Iead, R., Jan, B., J., Beckmann, C.F.: "Understanding Heterogeneity in Clinical Cohorts Using Normative Models: Beyond Case-Control Studies." in *Biological Psychiatry*. vol. 80(7), pp. 552–561, Oct. 2016.

[196] A. F. Marquand, S. M. Kia, M. Zabihi, T. Wolfers, J. K. Buitelaar, and C. F. Beckmann. Conceptualizing mental disorders as deviations from normative functioning. Molecular psychiatry, 24(10):1415–1424, 2019.

[197] Marusyk, A., Polyak, K. "Tumor heterogeneity: causes and consequences." in *Biochim Biophys Acta* vol. 1805, no. 1, pp. 105–117, 2010.

[198] M. L. Phillips and H. A. Swartz. "A critical appraisal of neuroimaging studies of bipolar disorder: toward a new conceptualization of underlying neural circuitry and a road map for future research". in *American Journal of Psychiatry*, vol. 171(8), pp. 829–843, 2014.

[199] Mathieu, E., Rainforth, T., Siddharth, N., and Teh, Y. W. "Disentangling Disentanglement in Variational Autoencoders". in *International Conference on Machine Learning, ICLR 2019*, pp. 4402–4412, 2019.

[200] Mazziotta, J.C., Toga, A.W., Evans, A.C., Fox, P.T., Lancaster, J.L. "Digital brain atlases." in *Trends Neurosci*. n. 18, pp. 210-–211, 1995.

[201] Mazziotta, J.C., Toga, A.W., Evans, A., Fox, P., Lancaster, J.: The International Consortium for Brain Mapping (ICBM). "A probabilistic atlas of the human brain: Theory and rationale for its development." in *Neuroimage*. n. 2, pp. 89-–101, 1995.

[202] Mazziotta J., et al. International Consortium for Brain Mapping (ICBM). "A probabilistic atlas and reference system for the human brain." in *Philos Trans R Soc Lond B Biol Sci.* n. 356, pp. 1293–1322, 2001.

[203] McLachlan, G.J.; Basford, K.E., "Mixture Models: inference and applications to clustering", in *Statistics: Textbooks and Monographs*, 1988.

[204] McAlonan, G.M., Cheung, V., Cheung, C., Suckling, J., Lam, G.Y., Tai, K.S., Yip, L., Murphy, D.G., Chua, S.E. "Mapping the brain in autism. A voxel-based MRI study of volumetric differences and inter-correlations in autism." in *Brain*, n. 128, pp. 268-–276, 2005.

[205] McGrath, J.J., et al. "Comorbidity within mental disorders: a comprehensive analysis based on 145 990 survey respondents from 27 countries.", in *Epidemiol Psychiatr Sci.* n. 29, pp. 153, Aug. 2020.

[206] McInnes, L., Healy, J., Melville, J. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction". in *arXiv:1802.03426 [cs, stat]*, 2020.

[207] Menyhart, O., Gyorffy, B.: "Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis". in *Computational and Structural Biotechnology Journal.* vol. 19, pp. 949–960, 2021.

[208] Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., and et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE Transactions on Medical Imaging, vol. 34(10), pp. 1993–2024, Oct. 2015.

[209] T. K. Moon, "The expectation-maximization algorithm," in *IEEE Signal Processing Magazine*, vol. 13, no. 6, pp. 47-60, Nov. 1996

[210] B. Neyshabur, H. Sedghi, and C. Zhang. "What is being transferred in transfer learning?", in *arXiv:2008.11687*, 2020.

[211] Nguyen, X., Wainwright, M. J., and Jordan, M. I. "Estimating divergence functionals and the likelihood ratio by convex risk minimization". *IEEE Transactions on Information Theory*, vol. 56(11), pp. 5847–5861, Nov. 2010.

[212] Ogawa, S., Lee, T. M., Nayak, A. S., Glynn, P. "Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields", in *Magnetic Resonance in Medicine*, vol. 14(1), pp. 68–78, 1990.

[213] Oord, Aaron van den, Yazhe Li and Oriol Vinyals. "Representation Learning with Contrastive Predictive Coding." in *ArXiv abs/1807.03748*, 2018.

[214] Y. Ostby, C. K. Tamnes, A. M. Fjell, L. T. Westlye, P. Due-Tønnessen, and K. B. Walhovd. "Heterogeneity in subcortical brain development: a structural magnetic resonance imaging study of brain maturation from 8 to 30 years". in *Journal of Neuroscience*, 29(38): pp. 11772–11782, 2009.

[215] Oyelade, J., Isewon, I., Oladipupo, F., Aromolaran, O., Uwoghiren, E., Ameh, F., Achas, M., Adebiyi, E. "Clustering Algorithms: Their Application to Gene Expression Data". in *Bioinform Biol Insights.* vol. 10, pp. 237–253. 2016.

[216] Park, S.C., Kim, J.M., Jun, T.Y., Lee, M.S., Kim, J.B., Yim, H.W, Park, Y.C. "How many different symptom combinations fulfill the diagnostic criteria for major depressive disorder? Results from the CRESCEND study." in *Nord J Psychiatry.*, n. 71(3), pp. 217–222, Apr. 2017.

[217] Emanuel Parzen. "On Estimation of a Probability Density Function and Mode". in *The Annals of Mathematical Statistics*, vol; 33(3), pp. 1065–1076, 1962.

[218] T. Pattyn, F. Van Den Eede, F. Lamers, D. Veltman, B.G. Sabbe, B.W. Penninx "Identifying panic disorder subtypes using factor mixture modeling". in *Depression Anxiety*, vol. 32, pp. 509–517, 2015.

[219] W. Peebles, J. Peebles, J-Y. Zhu, A. Efros, and A. Torralba, "The Hessian Penalty: A Weak Prior for Unsupervised Disentanglement", in *European Conference on Computer Vision (ECCV)*, 2021.

[220] Petsiuk, V., Das, A., and Saenko, K. "Rise: Randomized input sampling for explanation of black-box models". in *arXiv preprint arXiv:1806.07421*, 2018.

[221] Petsiuk, V., Jain, R., Manjunatha, V., Morariu, V. I., Mehra, A., Ordonez, V., and Saenko, K. "Black-box explanation of object detectors via saliency maps". in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11443–11452, 2021.

[222] Pinaya, W. H. L., Mechelli, A., and Sato, J. R. "Using deep autoencoders to identify abnormal brain structural patterns in neuropsychiatric disorders: A large-scale multi-sample study". in *Human Brain Mapping*, vol. 40(3), pp. 944–954, 2019.

[223] Walter H. L. Pinaya, "Normative modeling using deep autoencoders: a multi-cohort study on mild cognitive impairment and Alzheimer's disease", in *bioRxiv preprint,* 2020.

[224] Mary Phuong, Max Welling, Nate Kushman, Ryota Tomioka, and Sebastian Nowozin. "The Mutual Autoencoder: Controlling Information in Latent Code Representations". in *Semantic Scholer, Computer Science*, 2018.

[225] Planey, C.R., Gevaert, O. "CoINcIDE: A framework for discovery of patient subtypes across multiple datasets." in *Genome Medicine.* vol. 8, no.27, 2016.

[226] Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. "On Variational Bounds of Mutual Information". in *Proceedings of the 36th International Conference on Machine Learning*, pp. 5171–5180. May 2019.

[227] Radford, A., Metz, L., Chintala, S. "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks". in *arXiv:1511.06434 [cs].* Jan. 2016.

[228] Muñoz-Ramírez, V., Kmetzsch, V., Forbes, F., Meoni, S., Moro, E., and Dojat, M. Subtle anomaly detection: Application to brain MRI analysis of de novo Parkinsonian patients. *Artificial Intelligence in Medicine*, 2022.

[229] Rawat, K.S., Malhan, I.V. "A Hybrid Classification Method Based on Machine Learning Classifiers to Predict Performance in Educational Data Mining". in *ICCCN*. pp. 677–684, 2019.

[230] Marco Tulio Ribeiro and Sameer Singh and Carlos Guestrin. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier." in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016.

[231] Richetto, J. Meyer, U. "Epigenetic Modifications in Schizophrenia and Related Disorders: Molecular Scars of Environmental Exposures and Source of Phenotypic Variability". in *Biol. Psychiatry* vol. 89, pp. 215–226, 2021.

[232] Robb, M.A., McInnes, P.M., Califf, R.M. "Biomarkers and Surrogate Endpoints: Developing Common Terminology and Definitions." in *JAMA Network*. n. 315(11), pp. 1107-–1108, 2016.

[233] H. Robbins and S. Monro. "A stochastic approximation method". in *The annals of mathematical statistics*, pp. 400–407, 1951.

[234] Borja Rodriguez-Galvez, Arno Blaas, Pau Rodriguez, Adam Golinski, Xavier Suau, Jason Ramapuram, Dan Busbridge, and Luca Zappella. "The Role of Entropy and Reconstruction

in Multi-View Self-Supervised Learning", in *International Conference on Machine Learning (ICML 2023)* July 2023.

[235] Rojas, D.C., Peterson, E., Winterrowd, E., Reite, M.L., Rogers, S.J., Tregellas, J.R. "Regional gray matter volumetric changes in autism associated with social and repetitive behavior symptoms". in *BMC Psychiatry.* vol. 6, pp. 56, Dec. 2006.

[236] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Björn Ommer. "High-Resolution Image Synthesis with Latent Diffusion Models" in *Computer Vision and Pattern Recognition.* 2022.

[237] Murray Rosenblatt. "Remarks on Some Nonparametric Estimates of a Density Function". in *The Annals of Mathematical Statistics*, vol. 27(3), pp. 832–837, Sep. 1956.

[238] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S.A. Siddiqui, A. Binder, E. Müller, and M. Kloft, "Deep One-Class Classification" in *Proceedings of the 35th International Conference on Machine Learning (PMLR)*, vol. 80, pp. 4393-4402, 2018.

[239] Lukas Ruff, Robert A. Vandermeulen, Nico Gornitz, Alexander Binder, Emmanuel Muller, Klaus, Robert Muller, and Marius Kloft. "Deep Semi-Supervised Anomaly Detection". in *International Conference on Learning Representations (ICLR 2019)* September 2019.

[240] Ruiz, A., Martinez, O., Binefa, X., and Verbeek, J. "Learning Disentangled Representations with Reference-Based Variational Autoencoders", in *arXiv:1901.08534 [cs].* Jan. 2019.

[241] Saige Rutherford, Pieter Barkema, Ivy F Tso, Chandra Sripada, Christian F Beckmann, Henricus G Ruhe, Andre F Marquand, "Evidence for embracing normative modeling" in *eLife*, vol. 12, 2023.

[242] R. Sacco, C. Lenti, M. Saccani, P. Curatolo, B. Manzi, C. Bravaccio, et al. "Cluster analysis of autistic patients based on principal pathogenetic components". in *Autism Res*, vol. 5, pp. 137–140, 2012.

[243] Saito, S., Tan, R.T. "Neural clustering: concatenating layers for better projections". in *Workshop in International Conference on Learning Representations.* 2017.

[244] Haijma, S.V., et al. "Brain volumes in schizophrenia: a meta-analysis in over 18 000 subjects." in *Schizophr Bull.* vol. 39(5), pp. 1129–38, 2013.

[245] Sarrazin, S., et. al., "Neurodevelopmental subtypes of bipolar disorder are related to cortical folding patterns: An international multicenter study." in *Bipolar Disorders.* vol.20, no. 8, pp 721–732, Dec. 2018.

[246] H. G. Schnack, N. E. Van Haren, M. Nieuwenhuis, H. E. Hulshoff Pol, W. Cahn, and R. S. Kahn. "Accelerated brain aging in schizophrenia: a longitudinal pattern recognition study". in *American Journal of Psychiatry*, vol. 173(6), pp. 607–616, 2016.

[247] M.-A. Schulz, B. T. Yeo, J. T. Vogelstein, J. Mourao-Miranada, J. N. Kather, K. Kording, B. Richards, and D. Bzdok. "Different scaling of linear models and deep learning in UK biobank brain images versus machine-learning datasets". in *Nature communications*, vol. 11(1). pp. 1–15, 2020.

[248] M.-A. Schulz, D. Bzdok, S. Haufe, J.-D. Haynes, and K. Ritter. "Performance reserves in brain-imaging-based phenotype prediction". in *bioRxiv*, 2022.

[249] Schulz, M.A., Chapman-Rounds, M., Verma, M., Bzdok, D., Georgatzis, K. "Inferring disease subtypes from clusters in explanation space". in *Scientific Reports* vol. 10(1), 2020.

[250] Schumann G., Binder E.B., Holte A., de Kloet E.R., Oedegaard K.J., Robbins T.W. et al., "Stratified medicine for mental disorders. Eur Neuropsychopharmacol". vol. 24, pp. 5–50, 2014.

[251] Schlegl, T., Seeböck, P., Waldstein, S. M., Langs, G., and Schmidt-Erfurth, U. "f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks". in *Medical Image Analysis*, vol. 54, pp. 30–44, 2019.

[252] Steffen Schneider, Jin Hwa Lee, and Mackenzie Weygandt Mathis. "Learnable latent embeddings for joint behavioural and neural analysis". in *Nature*, vol. 617(7960), pp. 360-–368, May 2023.

[253] Selvaraju, Ramprasaath R. and Cogswell, Michael and Das, Abhishek and Vedantam, Ramakrishna and Parikh, Devi and Batra, Dhruv. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization." in *International Conference in Computer Vision 2016*, vol. 128(2), 2016.

[254] Severson, K. A., Ghosh, S., and Ng, K. "Unsupervised Learning with Contrastive Latent Variable Models". in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33(01), pp. 4862–4869, July 2019.

[255] Shenton, M.E., Dickey, C.C., Frumin, M., McCarley, R.W. "A review of MRI findings in schizophrenia". in *Schizophr Res.* vol. 49(1-2), pp. 1-52, Apr. 2001.

[256] Shu, R., Zhao, S., and Kochenderfer, M. J. Rethinking style and content disentanglement in variational auto encoders. in *International Conference of Leaning Representations Workshop.*. 2018.

[257] Siddharth N, Brooks Paige, Jan-Willem van de Meent, Alban Desmaison, Noah Goodman, Pushmeet Kohli, Frank Wood, and Philip Torr. "Learning Disentangled Representations with Semi-Supervised Deep Generative Models". in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[258] Simarro Viana, J., de la Rosa, E., Vande Vyvere, T., Robben, D., Sima, D. M., and Investigators, C.-T. P. a. "Unsupervised 3D Brain Anomaly Detection". in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pp. 133–142, 2021.

[259] Simonyan, K. and Zisserman, A. "Very deep convolutional networks for large-scale image recognition". in *arXiv preprint arXiv:1409.1556*, 2014.

[260] Smeland, O. B., Frei, O., Dale, A. M. Andreassen, O. A. "The polygenic architecture of schizophrenia — rethinking pathogenesis and nosology". in *Nat. Rev. Neurol.* vol. 16, pp. 366–379, 2020.

[261] Sonpatki, P., Shah, N. "Recursive Consensus Clustering for novel subtype discovery from transcriptome data". in *Scientific Reports* vol. 10(1). 2020.

[262] Strawbridge, R., Young, A.H., Cleare, A.J. "Biomarkers for depression: recent insights, current challenges and future prospects". in *Neuropsychiatr Dis Treat.* n. 13, pp. 1245–1262, May 2017.

[263] D. Steinkraus, I. Buck, and P. Simard. "Using gpus for machine learning algorithms." in *IEEE ICDAR 2005*, pp. 1115–1120, 2005.

[264] Stinchcombe, White, "Universal approximation using feedforward networks with non-sigmoid hidden layer activation functions," in *International 1989 Joint Conference on Neural Networks, Washington, DC, USA*, vol. 1, pp. 613–617, 1989.

[265] Sugiyama, M., Suzuki, T., and Kanamori, T. "Density-ratio matching under the Bregman divergence: a unified framework of density-ratio estimation". in *Annals of the Institute of Statistical Mathematics*, vol. 64, pp. 1009–1044, 2012.

[266] Sun, Jinghan, et. al. "Unsupervised Representation Learning Meets Pseudo-Label Supervised Self-Distillation: A New Approach to Rare Disease Classification." in *Proc. Med. Imag. Comput. and Comput. Assist. Interv. MICCAI.* pp. 519–529, 2021.

[267] Sun, C., Baradel, F., Murphy, K.P., Schmid, C. "Learning Video Representations using Contrastive Bidirectional Transformer". in *arXiv:1906.05743* , 2019.

[268] Tager-Flusberg, H., Joseph, R.M. "Identifying neurocognitive phenotypes in autism." in *Philos. Trans. R. Soc. Biol. Sci.* vol. 358. no. 1430, pp. 303–314, 2003

[269] J. Talairach and P. Tournoux, "Co-planar stereotaxic atlas of the human brain." in *Thieme Medical Publishers*, 1988

[270] Talpalaru, A., Bhagwat, N., Devenyi, G. A., Lepage, M. Chakravarty, M. M. "Identifying schizophrenia subgroups using clustering and supervised learning". in *Schizophr. Res.* vol. 214, pp. 51–59, 2019.

[271] Tamminga, C.A. et. al. "Bipolar and Schizophrenia Network for Intermediate Phenotypes: Outcomes Across the Psychosis Continuum." in *Schizophrenia Bulletin* vol. 40 no. 2. pp. 131—137, Mar. 2014.

[272] Tartaglione, E., Barbano, C.A., Grangetto, M. "EnD: Entangling and Disentangling deep representations for bias correction." in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. CVPR.*. Jun. 2021. pp. 13503–13512.

[273] Tian, Y., Pang, G., Liu, F., chen, Y., Shin, S. H., Verjans, J. W., Singh, R., and Carneiro, G. "Constrained Contrastive Distribution Learning for Unsupervised Anomaly Detection and Localisation in Medical Images". in *arXiv:2103.03423[cs]*, 2021.

[274] Yonglong Tian, Dilip Krishnan, and Phillip Isola. "Contrastive Multiview Coding". in *European Conference of Computer Vision*, pp. 776–794, 2020

[275] Tsai, T.W., Li, C., Zhu, J. "MiCE: Mixture of Contrastive Experts for Unsupervised Image Clustering." in *Proc. Int. Conf. Learn. Represent. ICLR,* May. 2021.

[276] Michael Tschannen, Josip Djolonga, Paul K. Rubenstein, Sylvain Gelly, and Mario Lucic. "On Mutual Information Maximization for Representation Learning". in *International Conference on Learning Representation 2020* September 2019

[277] Tsuang MT. "Heterogeneity of schizophrenia." in *Biol Psychiatry.* n. 10(4) pp. 465–74, Aug. 1975.

[278] Van Gansbeke, W., Vandenhende, S., Georgoulis, S., Proesmans, M., Van Gool, L. "SCAN: Learning to Classify Images Without Labels." in *Europ. Conc. in Comp. Vis.* vol. 12355. pp. 268–285, 2020

[279] E. Varol, A. Sotiras, C. Davatzikos, "Disentangling disease heterogeneity with max-margin multiple hyperplane classifier". in *Medical Image Computing and Computer-Assisted Intervention, MICCAI*, pp. 702-709, 2015.

[280] Varol, E., Sotiras, A., Davatzikos, C. "HYDRA: Revealing heterogeneity of imaging and genetic patterns through a multiple max-margin discriminative analysis framework." in *NeuroImage* vol. 145, pp 346–364, 2017.

[281] van der Meer D, Kaufmann T, Shadrin AA, Makowski C, Frei O, Roelfs D, Monereo-Sánchez J, Linden DEJ, Rokicki J, Alnæs D, de Leeuw C, Thompson WK, Loughnan R, Fan CC, Westlye LT, Andreassen OA, Dale AM. "The genetic architecture of human cortical folding". in *Sci Adv.* vol.7(51), 2017

[282] van Erp TG, et al. "Subcortical brain volume abnormalities in 2028 individuals with schizophrenia and 2540 healthy controls via the ENIGMA consortium". in *Mol Psychiatry.* vol. 21(4), pp. 547–553, 2016.

[283] van Hulst B.M., de Zeeuw P. Durston S., "Distinct neuropsychological profiles within ADHD: A latent class analysis of cognitive control, reward sensitivity and timing." in *Psychol Med.* vol. 45, pp. 735–745, 2015.

[284] van Loo H.M., de Jonge P., Romeijn J.-W., Kessler R.C., Schoevers R.A. "Data-driven subtypes of major depressive disorder: A systematic review." in *BMC Med.* vol. 10, pp. 156, 2010.

[285] Vapnik, V. N., Chervonenkis, A. Ya. "On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities". in *Theory of Probability Its Applications.* vol. 16(2), pp. 264. 1971.

[286] von Kugelgen, J., Sharma, Y., Gresele, L., Brendel, W., Scholkopf, B., Besserve, M., and Locatello, F. "Self-Supervised Learning with Data Augmentations Provably Isolates

Content from Style". in *Advances in Neural Information Processing Systems, NeurIPS*, 2021.

[287] Voineskos, A. N. et al. "Neuroimaging evidence for the deficit subtype of schizophrenia". in *JAMA Psychiatry* vol. 70, pp. 472–480, 2013.

[288] Junyuan Xie, Ross Girshick, and Ali Farhadi. "Unsupervised deep embedding for clustering analysis". in *International Conference on International Conference on Machine Learning* vol. 48, pp. 478-–487, 2016.

[289] Wachinger, C., Rieckmann, A. Pölsterl, S. "Detect and correct bias in multi-site neuroimaging datasets". in *Med. Image Anal.* vol. 67, pp. 101879, 2021.

[290] Waiter, G.D., et al. "A voxel-based investigation of brain structure in male adolescents with autistic spectrum disorder". in *Neuroimage.* vol. 22, pp. 619—625, 2004.

[291] Wahlstedt, C., Thorell, L.B., Bohlin, G. "Heterogeneity in ADHD: Neuropsychological Pathways, Comorbidity and Symptom Domains". in *J Abnorm Child Psychol* vol. 37(4), pp. 551–564, 2009.

[292] Wang, L.e.a. "SchizConnect: Mediating Neuroimaging Databases on Schizophrenia and Related Disorders for Large-Scale Integration." in *NeuroImage* vol. 124, pp. 1155–1167, Jan. 2016.

[293] Wang, T., Isola, P. "Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere." in *Proc. Int. Conf. Mach. Learn. ICML.* Jul. 2020.

[294] X. Wang et al., "Normative modeling via conditional variational autoencoder adversarial learning to identify brain dysfunction in Alzheimer's Disease." in *IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, 2023.

[295] Ke Wang, Harshitha Machiraju, Oh-Hyeon Choung, Michael Herzog, and Pascal Frossard. "CLAD: A Contrastive Learning based Approach for Background Debiasing". in *British Machine Vision Conference (BMVC) 2022*, 2022.

[296] Wang, Y., Zhao, Y., Therneau, T.M., Atkinson, E.J., Tafti, A.P., Zhang, N., Amin, S., Limper, A.H., Khosla, S., Liu, H.: "Unsupervised machine learning for the discovery of latent disease clusters and patient subgroups using electronic health records". in *Journal of Biomedical Informatics.* vol. 102, 2020.

[297] Chen Wei, Huiyu Wang, Wei Shen, and Alan Yuille. "CO2: Consistent Contrast for Unsupervised Visual Representation Learning". in *International Contrastive Learning of Representations. ICLR 2020.* Sep. 2020.

[298] Weinberg, D. et al. "Cognitive Subtypes of Schizophrenia Characterized by Differential Brain Volumetric Reductions and Cognitive Decline". in *JAMA Psychiatry* vol. 73, pp. 1251–1259, 2016.

[299] Weinberger, E., Beebe-Wang, N., and Lee, S.-I. "Moment Matching Deep Contrastive Latent Variable Models". in *AISTATS*, 2022.

[300] Wen, J., Varol, E., Chand, G., Sotiras, A., Davatzikos, C. "MAGIC: Multi-scale Heterogeneity Analysis and Clustering for Brain Diseases." in *Proc. Med. Imag. Comput. and Comput. Assist. Interv. MICCAI.,* 2020.

[301] Wenzel, J. et al. "Cognitive subtypes in recent onset psychosis: distinct neurobiological fingerprints?" in *Neuropsychopharmacology* vol. 46, pp. 1475–1483, 2021.

[302] Wisco BE, Miller MW, Wolf EJ, Kilpatrick D, Resnick HS, Badour CL, et al. "The impact of proposed changes to ICD-11 on estimates of PTSD prevalence and comorbidity". in *Psychiatry Res.* vol. 240, pp. 226–233, 2016.

[303] Wolfers T, Doan NT, Kaufmann T, Alnæs D, Moberget T, Agartz I, Buitelaar JK, Ueland T, Melle I, Franke B, Andreassen OA, Beckmann CF, Westlye LT, Marquand AF. "Mapping the Heterogeneous Phenotype of Schizophrenia and Bipolar Disorder Using Normative Models". in *JAMA Psychiatry.* vol. 75(11). pp. 1146–1155, Nov 2018.

[304] Woo, C.-W., Chang, L. J., Lindquist, M. A. Wager, T. D. "Building better biomarkers: brain models in translational neuroimaging". in *Nat. Neurosci.* vol. 20, pp. 365–377, 2017.

[305] Woodward, N. D. Heckers, S. "Brain Structure in Neuropsychologically Defined Subgroups of Schizophrenia and Psychotic Bipolar Disorder". in *Schizophr. Bull.* vol. 41, pp. 1349–1359, 2015.

[306] Auerbach, R.P., WHO WMH-ICS Collaborators. "WHO World Mental Health Surveys International College Student Project: Prevalence and distribution of mental disorders". in *J Abnorm Psychol.*, n. 127(7), pp. 623–638, Oct. 2018.

[307] Wu, M.Y., Dai, D.Q., Zhang, X.F., Zhu, Y. "Cancer Subtype Discovery and Biomarker Identification via a New Robust Network Clustering Algorithm." in *PLOS ONE* vol. 8. no. 6, 2013

[308] Yang, B., Fu, X., Sidiropoulos, N.D., Hong, M. "Towards K-means-friendly Spaces: Simultaneous Deep Learning and Clustering". in *International Conference on Machine Learning, ICML 2017*. pp. 3861–3870. 2017.

[309] Yang, Z., Wen, J., Davatzikos, C. "Smile-GANs: Semi-supervised clustering via GANs for dissecting brain disease heterogeneity from medical images". in *arXiv:2006.15255.* 2020.

[310] Yang, T., Frangou, S., Lam, R.W., Huang, J., Su, Y., Zhao, G., Mao, R., Zhu, N., Zhou, R., Lin, X., Xia, W., Wang, X., Wang, Y., Peng, D., Wang, Z., Yatham, L.N., Chen, J., Fang, Y. "Probing the clinical and brain structural boundaries of bipolar and major depressive disorder". in *Translational Psychiatry.* vol. 11(1), pp. 1–8, 2021.

[311] Yang, J., et. al. "Novel subtypes of pulmonary emphysema based on spatially-informed lung texture learning." in *IEEE Transactions on Medical Imaging.* vol. 40, no. 12, pp. 3652–3662, 2021

[312] Yang, T., et. al. "Probing the clinical and brain structural boundaries of bipolar and major depressive disorder." in *Translational Psychiatry* vol. 11. no. 1, pp. 1-8, 2021.

[313] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. "How transferable are features in deep neural networks ?" in *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[314] Young, G., Lareau, C., Pierre, B., "One Quintillion Ways to Have PTSD Comorbidity: Recommendations for the Disordered DSM-5." in *Psychol. Inj. and Law*, no. 7, pp.61–74, 2014.

[315] Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faieta. "Multimodal Contrastive Training for Visual Representation Learning". in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6991–7000, June 2021.

[316] Zabihi, M., Oldehinkel, M., Wolfers, T., Frouin, V., Goyard, D., et al. "Dissecting the Heterogeneous Cortical Anatomy of Autism Spectrum Disorder Using Normative Models". in *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging.* vol. 4(6), pp. 567–578, 2019.

[317] M. Zabihi, D. L. Floris, S. M. Kia, T. Wolfers, J. Tillmann, A. L. Arenas, C. Moessnang, T. Banaschewski, R. Holt, S. Baron-Cohen, et al. "Fractionating autism based on neuroanatomical normative modeling". in *Translational psychiatry*, vol. 10(1), pp. 1–10, 2020.

[318] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, Stéphane Deny: "Barlow Twins: Self-Supervised Learning via Redundancy Reduction". in *International Conference of Machine Learning.* pp. 12310-12320, 2021.

[319] Zhan, X., Xie, J., Liu, Z., Ong, Y.S., Loy, C.C. "Online Deep Clustering for Unsupervised Representation Learning." in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. CVPR.* pp. 6687–6696, Jun. 2020.

[320] Zhang, W. et al. "Brain Structural Abnormalities in a Group of Never-Medicated Patients With Long-Term Schizophrenia". in *Am. J. Psychiatry* vol. 172, pp. 995–1003, 2015.

[321] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. "Understanding deep learning (still) requires rethinking generalization". in *Communications of the ACM*, vol. 64(3), pp. 107-–115, 2021.

[322] Jiajin Zhang, Hanqing Chao, Giridhar Dasegowda, Ge Wang, Mannudeep K. Kalra, and Pingkun Yan. "Overlooked Trustworthiness of Saliency Maps" in *MICCAI 2022*, pp. 451–461, 2022.

[323] Chao Zhao, Jiajia Zhu, Xiaoyi Liu, Chengcheng Pu, Yunyao Lai, Lei Chen, Xin Yu, Nan Hong, "Structural and functional brain abnormalities in schizophrenia: A cross-sectional study at different stages of the disease", in *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, vol. 83, pp. 27–32, 2018.

[324] Zheng, G. X. Y., Terry, et al. "Massively parallel digital transcriptional profiling of single cells". in *Nature Communications*, vol. 8(1) pp. 14049, Jan. 2017.

[325] Zheng, Z. and Sun, L. "Disentangling Latent Space for VAE by Label Relevant/Irrelevant Dimensions". in *arXiv:1812.09502 [cs]*, Mar. 2019.

[326] Zheng, M., Wang, F., You, S., Qian, C., Zhang, C., Wang, X., Xu, C. "Weakly Supervised Contrastive Learning." in *IEEE/CVF Int. Conf. on Comput. Vis. ICCV.* pp. 10022–10031. IEEE, Montreal, QC, Canada, Oct. 2021

[327] Zou, J. Y., Hsu, D. J., Parkes, D. C., and Adams, R. P. "Contrastive Learning Using Spectral Methods". in *Advances in Neural Information Processing Systems*, vol. 26. 2013.

[328] Zou, K., Faisan, S., Heitz, F., and Valette, S. "Joint Disentanglement of Labels and Their Features with VAE". in *IEEE International Conference on Image Processing (ICIP)*, pp. 1341–1345, 2022.

[329] Kaifeng Zou, Sylvain Faisan, Fabrice Heitz, and Sebastien Valette. "Disentangling high-level factors and their features with conditional vector quantized VAEs". in *Pattern Recognition Letters*, vol. 172, pp. 172–180, August 2023.

[330] Shengjia Zhao, Jiaming Song, and Stefano Ermon. InfoVAE: Balancing Learning and Inference in Variational Autoencoders. in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33(01), pp. 5885–5892, July 2019.

[331] Zhou, D-X, "Universality of deep convolutional neural networks", in *Applied and Computational Harmonic Analysis*, vol. 48(2), pp. 787–794, 2020.

[332] Zixiang, C. and Yihe, D. and Yue, W. and Quanquan, G. and Yuanzhi, Li. "Towards Understanding the Mixture-of-Experts Layer in Deep Learning" in *Proc. Adv. Neural Inf. Process. Syst. NeurIPS.,* 2022