Leveraging Anomaly Detection for Affective Computing Applications

Salam Hamieh

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : EEATS - Electronique, Electrotechnique, Automatique, Traitement du Signal (EEATS)
Spécialité : Signal Image Parole Télécoms
Unité de recherche : Laboratoire d'Electronique et de Technologie de l'Information (LETI)

Utilisation des méthodes de détection d'anomalies pour l'informatique affective

Leveraging Anomaly Detection for Affective Computing Applications

Présentée par :

Salam HAMIEH

Direction de thèse :

Christelle GODIN DIRECTRICE DE RECHERCHE, CEA CENTRE DE GRENOBLE	Directrice de thèse
Vincent HEIRIES INGENIEUR DOCTEUR, Université Grenoble Alpes	Co-encadrant de thèse
Hussein AL OSMAN ASSOCIATE PROFESSOR, University of ottawa	Co-encadrant de thèse

Rapporteurs :

Alice OTHMANI
MAITRESSE DE CONFERENCE HDR, Université Paris-Est Créteil
Tiago FALK
FULL PROFESSOR, Institut national de la recherche scientifique (INRS)

Thèse soutenue publiquement le **19 février 2024**, devant le jury composé de :

Martial MERMILLOD, PROFESSEUR DES UNIVERSITES, Université Grenoble Alpes	Président
Christelle GODIN, DIRECTRICE DE RECHERCHE, CEA CENTRE DE GRENOBLE	Directrice de thèse
Alice OTHMANI, MAITRESSE DE CONFERENCE HDR , Université Paris-Est Créteil	Rapporteure
Tiago FALK, FULL PROFESSOR, Institut national de la recherche scientifique (INRS)	Rapporteur
Fabien RINGEVAL, MAITRE DE CONFERENCES, Université Grenoble Alpes	Examinateur
steven LATRE, FULL PROFESSOR, Universiteit Antwerpen	Examinateur
Jean-Julien AUCOUTURIER, DIRECTEUR DE RECHERCHE, CNRS DELEGATION CENTRE-EST	Examinateur

Invités :

Vincent Heiries
INGENIEUR DE RECHERCHE, CEA
hussein al osman
ASSOCIATE PROFESSOR, Uottawa

Dedication

To *Baba*.

Acknowledgements

I would like to thank everyone who has helped me throughout my doctoral journey. I'm grateful for every bit of assistance, whether small or large, as it has contributed to making this achievement possible.

First and foremost, I would like to express my deepest gratitude to Christelle Godin, Vincent Heiries, and Hussein Al Osman, my supervisors, for their invaluable guidance, support, and encouragement throughout my doctoral research. Their expertise, patience, and unwavering belief in my abilities have been instrumental in shaping this thesis.

I would like to extend my sincere appreciation to my colleagues at CEA, with whom I shared many lunches, coffee breaks, and croissants. I especially want to thank Adrien, Marion, and Raphael for being great friends and for making the PhD journey easier.

I would also like to express my heartfelt gratitude to my friends. I am thankful for the joy and balance you brought into my life during this journey. A special thanks to Sally and Nour.

I am indebted to my family for their unconditional love, encouragement, and understanding throughout this journey. Mama, Baba, Maysaloune, Sirine, Mohammad, Lama, and Tima, celebrating this PhD with you means the world to me.

Finally, I would like to thank my husband Mohammad. You were there for me in every possible way, supporting me unconditionally. You are my partner in this, just like you're my partner in everything else.

Abstract

Recent technological advancements have paved the way for automation in various sectors, from education to autonomous driving, collaborative robots, and customer service. In the near future, the success and acceptance of these automated systems will rely upon their proficiency in assessing and responding to human emotional states, an aspect vital to their effectiveness. This has led to an increasing interest in the development of machine learning models for emotion recognition and interpretation. Nonetheless, the efficient computer-based assessment of affective and mental states faces several significant challenges, which include the difficulty of obtaining sufficient data, the intricacy of labeling, and the complexity of the task. One promising solution to these challenges lies in the field of anomaly detection, which has demonstrated its significance in numerous domains. This thesis is dedicated to addressing the multifaceted challenges in the field of affective computing by leveraging the power of anomaly detection methods. One of the key challenges addressed is data scarcity, a pervasive issue when striving to construct machine learning models capable of accurately identifying rare mental states. We study anomaly detection methods, utilizing unsupervised approaches in two critical applications: Visual Distraction Detection and Psychotic Relapse Prediction. These scenarios represent demanding and sometimes perilous states for data collection in real-world contexts. The study encompasses a comprehensive exploration of traditional and deep learning-based models, such as autoencoders, demonstrating the success of these methods in overcoming the challenges posed by unbalanced datasets. This success suggests the potential for wider applications in the future, which will help us better understand and deal with rare and hard-to-collect mental and affective states across various areas where obtaining sufficient data is not possible.

Furthermore, this research addresses the challenge of inter-variability among individuals in the domain of affective states, particularly in the context of patients with psychotic relapse. The study provides a comparative analysis, exploring the strengths and limitations of both global and personalized models. Personalization is a solution to this challenge, although gathering sufficient personal data, especially for relapse situations, is challenging. However, by employing anomaly detection, it becomes feasible to use an individual's data to model their healthy patterns and detect anomalies when these patterns deviate from the norm. The findings underscore the significance of personalization as an avenue for enhancing the precision of models, especially in scenarios characterized by substantial inter-variability among subjects.

Moreover, the complexity of unbalanced datasets is another focus of this thesis. It explores feature selection methods tailored to address these specific dataset characteristics. By leveraging state-of-the-art techniques, including autoencoders, the research advances novel strategies for addressing feature selection challenges posed by unbalanced datasets in applications such as Visual Distraction Detection and Psychotic Relapse Prediction.

Finally, the study introduces a novel solution for information fusion from multiple sources, enhancing predictive accuracy in affective computing. This novel approach incorporates an innovative difficulty data indicator derived from an autoencoder's reconstruction error. The outcome is the development of multimodal continuous emotion recognition systems that exhibit superior performance. This approach is studied using the ULM TSST dataset for predicting arousal and valence among participants in stress-induced situations.

In this thesis, we investigated various applications of anomaly detection methods in affective computing domain. While these are initial steps showcasing the potential of our proposed approaches, they also lay the groundwork for further exploration in different applications.

Résumé

Cette thèse aborde les défis multifacettes dans le domaine de l'informatique affective en exploitant des méthodes de détection d'anomalies. La prévalence croissante des interactions entre l'homme et l'ordinateur a souligné la nécessité de systèmes capables de comprendre et de réagir aux états émotionnels. Les récents progrès technologiques ont ouvert la voie à l'automatisation dans divers secteurs, de l'éducation à la conduite autonome en passant par le service client. Le succès de ces systèmes automatisés repose sur leur efficacité à évaluer et à répondre aux états émotionnels humains, un aspect essentiel à leur efficacité.

L'un des principaux défis abordés est la rareté des données, un problème répandu lorsqu'il s'agit de construire des modèles d'apprentissage automatique capables d'identifier avec précision des états mentaux rares. Nous étudions les méthodes de détection d'anomalies, en utilisant des approches non supervisées dans deux applications critiques : la détection de distraction visuelle et la prédiction de rechute psychotique. Ces scénarios représentent des états exigeants et parfois dangereux pour la collecte de données dans des contextes réels. L'étude englobe une exploration complète des modèles traditionnels et basés sur l'apprentissage profond, tels que les autoencodeurs, démontrant le succès de ces méthodes pour surmonter les défis posés par des ensembles de données déséquilibrés.
En outre, cette recherche aborde le défi de l'inter-variabilité entre les individus dans le domaine des états affectifs, en particulier dans le contexte des patients en rechute psychotique. L'étude fournit une analyse comparative, explorant les forces et les limites des modèles globaux et personnalisés. Les résultats soulignent l'importance de la personnalisation comme moyen d'améliorer la précision des modèles, notamment dans les scénarios caractérisés par une inter-variabilité substantielle entre les sujets.
De plus, la complexité des ensembles de données déséquilibrés est un autre point focal de cette thèse. Elle explore des méthodes de sélection de caractéristiques adaptées pour aborder ces caractéristiques spécifiques des ensembles de données. En exploitant des techniques de pointe, notamment les autoencodeurs, la recherche propose de nouvelles stratégies pour relever les défis de la sélection de caractéristiques posés par des ensembles de données déséquilibrés dans des applications telles que la détection de distraction visuelle et la prédiction de rechute psychotique.
Enfin, l'étude introduit une nouvelle solution pour la fusion d'informations provenant de sources multiples, améliorant la précision prédictive dans le domaine de l'informatique affective. Cette approche novatrice intègre un indicateur de difficulté des données dérivé de l'erreur de reconstruction de l'autoencodeur. Le résultat est le développement de systèmes de reconnaissance d'émotions continues multimodaux qui présentent des performances supérieures. Cette approche est étudiée à l'aide de l'ensemble de données ULM TSST pour prédire l'excitation et la valence parmi les participants dans des situations induisant du stress.
Dans cette thèse, nous avons étudié diverses applications des méthodes de détection des anomalies dans le domaine de l'informatique affective. Bien qu'il s'agisse d'étapes initiales démontrant le potentiel de nos approches proposées, elles jettent également les bases d'une exploration plus poussée des différentes applications et de leurs variations.

Résumé étendu

Les progrès rapides de l'intelligence artificielle, de la robotique et de l'automatisation remodèlent en profondeur les industries et les sociétés mondiales. Les transformations attendues dans nos routines de travail, nos modes de vie et nos interactions sociales pourraient se dérouler à une vitesse et à une échelle sans précédent, surpassant tous les changements précédents dans l'histoire de l'humanité. Les récentes avancées technologiques ont ouvert la voie à l'automatisation dans divers domaines, notamment l'apprentissage scolaire, la conduite autonome, la médecine, le service à la clientèle, etc. L'informatique affective, qui se situe à l'intersection de l'informatique, de la psychologie et des sciences cognitives, constitue une frontière essentielle pour l'amélioration des interactions homme-machine en intégrant la compréhension des émotions dans la technologie. Ce domaine interdisciplinaire vise à permettre aux machines de reconnaître, d'interpréter et de répondre aux émotions humaines.
Dans le domaine de l'intelligence artificielle, l'intégration de l'informatique affective n'est pas simplement une fonctionnalité avancée ; il s'agit d'un changement fondamental vers des systèmes plus intuitifs, plus réactifs et plus centrés sur l'humain. Cette évolution est particulièrement marquante dans des applications allant des soins de santé, où la prise en charge empathique des patients peut être révolutionnée, au service à la clientèle, où la compréhension et la réponse aux émotions des consommateurs peuvent grandement améliorer la qualité du service. En outre, l'intégration de l'informatique affective dans les systèmes autonomes, tels que les véhicules électriques et les maisons intelligentes, pourrait conduire à des expériences plus adaptées et plus conviviales. Le développement d'algorithmes sophistiqués capables d'analyser et d'interpréter les indices émotionnels - des expressions faciales à la tonalité de la voix - est essentiel. Cela nécessite une approche pluridisciplinaire, mêlant les techniques d'apprentissage automatique (Machine Learning) aux connaissances issues de la recherche psychologique, afin de créer des algorithmes qui soient non seulement techniquement compétents, mais aussi tenant compte des dimensions éthiques et culturelles. Ainsi, la recherche de technologies informatiques affectives avancées n'est pas seulement un effort technologique ; c'est une étape vers des machines plus empathiques, plus compréhensives et, en fin de compte, plus respectueuses de l'être humain. Ce changement de paradigme pourrait redéfinir la dynamique de notre interaction avec la technologie, en la rendant plus transparente, plus intuitive et plus en phase avec nos besoins et nos états émotionnels.
Une grande partie de la recherche en informatique affective est centrée sur la reconnaissance des états émotionnels humains. Il s'agit de tirer parti de diverses modalités telles que les capteurs physiologiques et les caméras pour capturer les réactions humaines telles que les expressions faciales. Cependant, le développement de modèles efficaces et robustes dans le domaine de l'informatique affective se heurte à plusieurs obstacles. Des défis tels que la rareté des données, l'incertitude et le coût de l'étiquetage, le déséquilibre des ensembles de données, la fusion d'informations provenant de sources multiples, l'explicabilité des modèles, les exigences en matière de modèles personnalisés, les préoccupations en matière de respect de la vie privée et les considérations éthiques constituent des obstacles importants. Parmi les domaines prometteurs pour relever ces défis, la détection des anomalies se distingue. Ses méthodes offrent des possibilités d'apprentissage non supervisée rendant la collecte de données annotées moins critique.De plus,

elles génèrent un score d'anomalie, fournissant des informations quantitives sur l'écart à la normale qui peuvent être exploités. Par conséquent, nous avons choisi d'explorer ces méthodes de détection d'anomalies pour relever des défis tels que la rareté des données, le déséquilibre des ensembles de données, la personnalisation des modèles, la sélection des caractéristiques, l'explicabilité des modèles et la fusion des données.
Les contributions de cette thèse peuvent être résumées comme suit :

1. **Détection d'anomalies dans l'informatique affective :** Nous explorons l'application des méthodes de détection d'anomalies pour détecter des états rares dans le domaine de l'informatique affective. Cette exploration est menée par le biais d'une approche non supervisée dans deux applications distinctes : la détection des comportements dangereux au volant, où l'apprentissage supervisé peut être entravé par la collecte de données relatives aux comportements à risque et présente donc des problèmes éthiques, et la prédiction des rechutes pour les patients souffrant de troubles psychotiques, qui implique la détection d'un événement peu fréquent dont la capture peut nécessiter un effort particulier.
2. **Modèles globaux et personnalisés :** L'examen des états en informatique affective révèle une variabilité entre les individus, ce qui représente un défi important lors de l'élaboration de modèles d'évaluation universels. Cette variabilité devient particulièrement importante lors de l'analyse des patients souffrant de rechute psychotique. En réponse au défi susmentionné, cette étude explore le potentiel des méthodes de détection des anomalies pour créer des modèles personnalisés de prédiction des rechutes psychotiques. En tirant parti de techniques de détection d'anomalies non supervisées, nous ouvrons la voie à des systèmes capables de collecter de manière autonome les données de l'utilisateur final dans des scénarios réels, éliminant ainsi la nécessité d'une intervention directe de l'utilisateur ou d'un étiquetage coûteux. Nous menons une analyse comparative avec des modèles globaux et explorons les forces et les limites des deux approches.
3. **Sélection de caractéristiques pour les ensembles de données déséquilibrés :** Lorsque l'on est confronté à des données limitées (par exemple, des modèles personnalisés) ou à des coûts d'étiquetage élevés dans des scénarios réels, la sélection des caractéristiques devient essentielle pour éviter le sur-apprentissage et réduire le coût calculatoire. Nous proposons d'utiliser les erreurs de reconstruction de l'autoencoder comme métrique pour examiner la pertinence des caractéristiques pour le problème de la sélection des caractéristiques. Nous évaluons cette approche dans deux applications critiques : la détection des distractions visuelles et la prédiction des rechutes psychotiques.
4. **Fusion multimodale :** Les états affectifs humains se manifestent par divers canaux tels que la voix, les expressions faciales et les réactions physiologiques, ce qui souligne la nécessité de disposer de sources de données à multiples facettes dans les systèmes d'estimation des états affectifs. L'intégration de modalités multiples, telles que les caméras, les capteurs physiologiques et les microphones, est essentielle pour améliorer l'efficacité de ces systèmes. Cependant, le défi provient de la nature hétérogène de ces modalités, ce qui fait de la fusion multimodale une tâche complexe. Pour y remédier, notre proposition suggère de tirer parti des auto-encoder et de leurs erreurs de reconstruction comme indices pour déterminer la pertinence de chaque modalité. Ces erreurs de reconstruction servent d'indicateurs précieux pour guider les algorithmes de fusion multimodale dans la combinaison efficace de ces diverses sources d'information.

Structure de la thèse et plan

Cette thèse se compose d'un chapitre d'introduction, de quatre chapitres décrivant le travail de la thèse suivis d'un chapitre contenant les conclusions et les perspectives. Les résumés des

quatres chapitres principaux sont présentés ci-dessous.

- *Chapitre 2 : Contexte et défis à relever.*
  Ce chapitre constitue une introduction du domaine de l'informatique affective et jette les bases de nos objectifs de recherche. Nous expliquons les motivations qui sous-tendent notre travail et la nécessité d'avancer dans ce domaine, en mettant l'accent sur les défis qu'il présente. En outre, nous présentons le domaine de la détection des anomalies et son rôle émergeant pour relever ces défis dans le domaine de l'informatique affective. Les chapitres suivants proposeront des solutions basées sur la détection d'anomalies pour une série de défis, les états mentaux rares, la personnalisation des modèles, l'explicabilité, la sélection des caractéristiques et la fusion d'information. Ces solutions seront appliquées à diverses applications telles que la surveillance du comportement des conducteurs, la prédiction des rechutes psychotiques et la reconnaissance des émotions dans des environnements stressants. Notre exploration englobera diverses modalités, notamment les signaux physiologiques, le suivi oculaire, la parole, la vidéo et le texte.
- *Chapitre 3 : Détection des états mentaux rares.* Nous explorons la détection d'états mentaux rares de manière non supervisée.Notre étude se concentre sur deux problèmes du monde réel : La détection des distractions visuelles et la prédiction des rechutes psychotiques. En employant des méthodologies non supervisées, nous visons à mettre en lumière des approches innovantes pour l'identification de ces états peu fréquents mais critiques sans avoir besoin de données étiquetées, ce qui peut être difficile dans ces applications.
  Pour chaque application, nous avons sélectionné une base de données appropriée, choisie stratégiquement pour démontrer l'efficacité de l'approche proposée dans divers contextes. Nos objectifs englobent trois aspects clés : premièrement, la validation de l'efficacité de l'approche proposée dans des applications distinctes ; deuxièmement, la comparaison de diverses méthodes de détection d'anomalies ; et troisièmement, l'évaluation de stratégies multiples impliquant la sélection de caractéristiques, des techniques supervisées et non supervisées pour la conduite, et des approches généralisées ou personnalisées pour la prédiction des rechutes.
  Dans la première étude, nous avons utilisé des approches de détection d'anomalies non supervisées au lieu des méthodes supervisées traditionnelles, qui permettront à terme de surmonter les difficultés liées à la collecte de données sur la distraction au volant, qui peut être dangereuse. En utilisant une base de données obtenue à partir d'un simulateur de conduite, nous avons entraîné nos modèles sur des exemples de conduite sans distraction et évalué leurs performances sur des exemples de conduite avec distraction. Nos résultats ont démontré l'efficacité des modèles non supervisés, la méthode "Isolation Forest" apparaissant comme le meilleur compromis performance/robustesse pour la détection de la distraction. En outre, nous avons comparé les performances des méthodes non supervisées à celles des modèles supervisés traditionnels, en soulignant la supériorité de l'approche que nous proposons pour les ensembles de données déséquilibrés ou même dans les scénarios où aucun échantillon de la classe "anormale" n'est disponible.
  Dans la deuxième étude, nous avons examiné l'efficacité des approches d'apprentissage non supervisé pour la détection des rechutes dans les troubles psychotiques. Nos résultats indiquent que la méthode "Isolation Forest" et les autoencoders ont affiché les meilleures performances dans le schéma global, où un modèle unique a été formé sur les données de tous les patients. Nous avons notamment découvert que les modalités optimales et les combinaisons de caractéristiques variaient d'un patient à l'autre, ce qui souligne l'importance des approches personnalisées dans la détection des rechutes. En outre, nous avons constaté que le choix du modèle de détection des anomalies et des caractéristiques avait un impact significatif sur la précision de la détection des rechutes. En adoptant une approche personnalisée, nous avons obtenu une méthode de détection plus adaptée et individualisée, ce

qui a permis d'améliorer considérablement la détection des rechutes. Cette étude souligne l'importance des modèles personnalisés pour un suivi précis et efficace des rechutes.
Dans l'ensemble, nos contributions à la recherche mettent en évidence l'efficacité des méthodes de détection d'anomalies non supervisées dans les tâches de surveillance du comportement.

- *Chapitre 4 : Explicabilité et sélection des caractéristiques à l'aide de scores d'anomalie.* Dans ce chapitre, nous étudions l'utilisation d'une méthode de sélection de caractéristiques conçue pour des ensembles de données déséquilibrés pour les deux tâches d'informatique affective : la détection de distraction visuelle et les ensembles de données de prédiction de rechute psychotique détaillés dans le chapitre 3.
  Ce chapitre présente l'utilisation de méthodes de détection d'anomalies pour la sélection de caractéristiques dans des ensembles de données déséquilibrés. Notre exploration s'articule autour de l'utilisation des autoencoders entrainés uniquement sur des données de classe normale et du calcul des scores de corrélation entre l'erreur de reconstruction des caractéristiques et les annotations provenant d'échantillons normaux et anormaux. Cette étude comprend une évaluation complète de la méthode de sélection des caractéristiques proposée dans le cadre de deux applications : la détection de la distraction visuelle et la prédiction des rechutes psychotiques.
  Dans l'application de la distraction visuelle, nous avons étudié l'impact de l'utilisation de stratégies de classification binaire, de détection d'anomalies et de modèles de régression. Notre approche, qui s'appuie sur les scores de corrélation, a fourni des perspectives intéressantes sur l'importance des caractéristiques. Elle a notamment mis en évidence les avantages qu'il y a à commencer par les caractéristiques les plus influentes dans les tâches de classification et de régression.
  Cependant, lorsqu'elles ont été étendues à la tâche plus complexe de "prédiction des rechutes psychotiques", qui disposait d'un ensemble de données relativement limité, les stratégies de sélection des caractéristiques ont montré des classements moins cohérents. Cela a mis en évidence le besoin crucial d'un ensemble de données de validation robuste pour ces stratégies. En outre, les résultats ont renforcé la nécessité de modèles personnalisés et souligné la variabilité de l'importance des caractéristiques chez tous les patients.
  Notre étude a présenté des résultats prometteurs en utilisant l'erreur de reconstruction des autoencoders pour obtenir des informations sur l'explicabilité du modèle et l'importance des caractéristiques. Toutefois, pour valider son efficacité et sa généralisation, cette méthode doit être testée plus avant sur différents ensembles de données.

- *Chapitre 5 : Fusion multimodale utilisant les scores d'anomalie.* Ce chapitre explore une nouvelle approche de fusion visant à améliorer la prédiction continue des émotions grâce à une technique de fusion multimodale fondée sur des scores d'anomalie. Ces scores d'anomalie sont dérivés des autoencoders entraînés par modalité. Nous étudions cette technique de fusion en utilisant l'ensemble de données ULM TSST pour la prédiction de l'activation physiologique ("arousal") et de la valence pour les participants dans des situations induites par le stress.
  Cet ensemble de données incorpore diverses modalités d'information, telles que l'audio, la vidéo, les biosignaux et le texte, obtenues auprès d'individus dans des conditions de stress. Notre approche s'est concentrée sur l'utilisation de la fusion tardive avec les modalités audio, vidéo et textuelles. La méthode de fusion proposée, inclue un indicateur de difficulté, dérivé des erreurs de reconstruction des autoencoders, et prend en compte l'aspect temporel des données. Nous avons mené des études pour analyser l'impact de l'indicateur de difficulté des données, sur les niveaux de prédiction unimodale et multimodale. À cette fin, un autoencoder a été entrainé pour chaque ensemble de caractéristiques correspondant à une modalité. Pour l'évaluation, les prédicteurs ont été entrainés avec et sans ces

indicateurs de difficulté au niveau unimodal et multimodal. Les résultats ont démontré une amélioration des modèles incluant les indicateurs de difficulté des données, affirmant la nature informative des erreurs de reconstruction des autoencoders et leur importance dans la fusion unimodale et multimodale.

Conclusions et Perspectives

L'intégration de la détection d'anomalies dans l'informatique affective a permis de relever des défis majeurs dans ce domaine. Nos recherches ont porté sur diverses applications, notamment la détection des comportements de conduite dangereux tels que les distractions visuelles pour la sécurité routière, la prévision des rechutes psychotiques chez les patients souffrant de troubles mentaux et la prévision continue des émotions des individus dans des situations stressantes. En utilisant des méthodes de détection d'anomalies, nous avons dépassé les limites de l'apprentissage supervisé traditionnel, qui nécessite toujours des données étiquetées et un équilibre dans les données, et nous avons réalisé des avancées significatives dans ces domaines. L'une des conclusions commune à ces études est l'importance des erreurs de reconstruction des auto-encoders, qui a été examinée dans trois contextes distincts, mettant en évidence sa valeur substantielle. Il a fonctionné comme un indicateur d'anomalie, aidant à l'identification de modèles rares ou anormaux. En outre, il a servi de métrique pour l'importance des caractéristiques, contribuant ainsi à l'explicabilité, et a agi comme une caractéristique supplémentaire pour améliorer la fusion de sources d'information multiples. L'un de ses avantages par rapport aux scores d'anomalie traditionnels est la possibilité de disséquer les erreurs de reconstruction, ce qui permet de vérifier les caractéristiques qui y contribuent.
Nos recherches ont mis en évidence l'adaptabilité des méthodologies de détection des anomalies dans divers contextes. Au chapitre 3, nous l'avons explorée dans une approche non supervisée, cruciale pour identifier les états rares lorsque les données d'une classe ne sont pas disponibles. Au chapitre 4, nous avons étendu son utilisation dans une approche faiblement supervisée, en utilisant des données étiquetées limitées de la classe minoritaire pour des tâches telles que la sélection des caractéristiques et l'explicabilité. Enfin, le chapitre 5 a exploré son application dans un cadre supervisé, en particulier dans la fusion d'informations.

Perspectives

- L'une des limites que nous avons rencontrées dans l'évaluation de nos approches est la taille restreinte des ensembles de données utilisés. Pour améliorer la robustesse et la généralisation de nos résultats, les études futures devraient se concentrer sur la validation de l'efficacité de nos approches sur des ensembles de données existants plus importants et plus diversifiés.
- Dans notre exploration de la surveillance du comportement des conducteurs, notre application des méthodes de détection des anomalies s'est avérée efficace pour identifier des comportements dangereux spécifiques, tels que les distractions visuelles. Cependant, une piste potentielle de recherche future consiste à valider l'adaptabilité de nos modèles pour reconnaître un spectre plus large de comportements dangereux. Pour ce faire, les modèles pourraient être entraînés sur des données normales enrichies de signaux liés non seulement aux distractions, mais aussi à la fatigue et à la somnolence. Par la suite, en testant ces modèles entraînés sur divers comportements dangereux potentiels, on pourrait vérifier leur capacité à détecter un plus large éventail de comportements dangereux au-delà des distractions visuelles. Une telle extension pourrait considérablement améliorer la polyvalence et l'applicabilité des techniques de détection d'anomalies pour garantir une sécurité

globale des conducteurs. Pour les modèles personnalisés de prédiction des rechutes psychotiques, une stratégie d'amélioration pourrait impliquer une étape initiale de formation d'un modèle général complet utilisant des données agrégées de tous les patients. Par la suite, l'ajustement de ce modèle général basé sur les données individuelles des patients pourrait potentiellement améliorer la précision et la robustesse de la prédiction en adaptant le modèle aux caractéristiques et aux comportements spécifiques des patients.

- En outre, l'ajout de données chronologiques dans l'ensemble de données d'eprevention représenterait une piste d'exploration supplémentaire. L'étude de l'impact de l'ordre chronologique des jours précédant la rechute et des jours de rechute sur les scores d'anomalie pourrait fournir des informations précieuses. Par exemple, l'observation de l'augmentation significative des scores d'anomalies dans la phase précédant la rechute pourrait servir d'indicateur d'alerte précoce. D'autre part, les annotations ne précisaient pas le trouble spécifique du patient. L'étude des performances des modèles non supervisés concernant des troubles spécifiques pourrait nous permettre d'approfondir notre compréhension de leurs capacités prédictives dans différents états de santé mentale.
- Compte tenu des progrès considérables réalisés dans les techniques basées sur les Transformers dans divers domaines de l'intelligence artificielle, leur application potentielle à la détection d'anomalies à l'aide d'ensembles de données plus importants, en particulier dans le traitement des données vocales et vidéo, apparaît comme une orientation convaincante pour les recherches futures. L'intégration des Transformers dans le cadre de la détection d'anomalies pourrait présenter plusieurs avantages. Une approche pourrait consister à tirer parti des architectures de Transformers, telles que BERT ou GPT, pour encoder des informations multimodales à partir de données vocales et vidéo. Ces modèles excellent dans la capture de modèles complexes et de dépendances au sein des séquences, ce qui pourrait s'avérer bénéfique pour la détection des anomalies, en particulier dans les flux de données multimodales.
- Un défi permanent dans la détection d'anomalies dans le comportement humain tourne autour de sa nature intrinsèquement dynamique. Les modèles conçus pour prédire des événements tels que les rechutes psychotiques peuvent déclencher de fausses alertes lorsqu'ils sont confrontés à des changements dans les habitudes quotidiennes des individus. Pour améliorer la robustesse de ces modèles face à de tels changements, l'exploration de méthodologies telles que l'apprentissage actif ou l'apprentissage incrémental pourrait offrir des solutions prometteuses. La mise en œuvre de ces techniques d'apprentissage adaptatif pourrait permettre d'affiner les modèles afin de mieux s'adapter aux changements de comportement, ce qui atténuerait les fausses alertes et améliorerait la précision des prédictions.

Articles et Brevet

Grâce aux recherches menées dans le cadre de cette thèse, nos contributions ont donné lieu à plusieurs publications.

1. **Articles**

   - Hamieh, S., Heiries, V., Al Osman, H., & Godin, C. (2021). Multi-modal fusion for continuous emotion recognition by using auto-encoders. In Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge (pp. 21-27).
   - Hamieh, S., Heiries, V., Al Osman, H., & Godin, C. (2023, June). Relapse Detection in Patients with Psychotic Disorders Using Unsupervised Learning on Smartwatch

Signals. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-2). IEEE.

- Hamieh, S., Heiries, V., Al Osman, H., Godin, C., & Aloui, S. (2023)Driver Visual Distraction Detection Using Unsupervised Learning Techniques. In ITSC 2023 IEEE International Conference on Intelligent Transportation.
- Hamieh, S., Heiries, V., Al Osman, H., & Godin, C. (2023). Psychotic Disorders Relapse Prediction from Passive Signals using Anomaly Detection Methods.(under review)

2. **Brevet**

- Hamieh, S., Heiries, V., Al Osman, H., & Godin, C. Multi-modal prediction system. EP4163830.

Contents

List of Figures

List of Tables

List of Acronyms

AE · Auto-Encoder

AI · Artificial Intelligence

ANOVA · Analysis Of Variance

BERT · Bidirectional Encoder Representations from Transformer

BiGRU · Bidirectional Gated Recurrent Unit

BiLSTM · Bidirectional Long Short-Term Memory

BiRNN · Bidirectional Recurrent Neural Network

BPM · Beats Per Minute

BVP · Blood Volume Pulse

CCC · Concordance Correlation Coefficient

CNN · Convolutional Neural Network

DL · Deep Learning

ECG · Electrocardiogram

EDA · Electrodermal Activity

EEG · Electroencephalogram

eGeMAPS · Extended Geneva Minimalistic Acoustic Parameter Set

EM · Expectation-Maximization

EMG · Electromyography

FN · False Negative

FP · False Positive

FPR · False Positive Rate

GAN · Generative Adversarial Network

GeMAPS · Geneva Minimalistic Acoustic Parameter Set

GMM · Gaussian Mixture Model

GRU · Gated Recurrent Unit

HR · Heart Rate

HRV · Heart Rate Variability

IBI · Interbeat Intervals

KNN · K-Nearest Neighbor

LOF · Local Outlier Factor

LLD · Low-Level Descriptors

LSTM · Long Short-Term Memory

MFCC · Mel Frequency Cepstral Coefficients

ML · Machine Learning

MLP · Multilayer Perceptron

MTCNN · Multi Cascaded Convolutional Neural Networks

MSE · Mean Squared Error

OCSVM · One Class Support Vector Machines

OOD · Out-Of-Distribution

OSR · Open-Set Recognition

PPG · Photoplethysmography

PR AUC · Precision-Recall Area Under Curve

RE · Reconstruction Error

ROC AUC · Receiver Operating Characteristic Area Under Curve

RNN · Recurrent Neural Network

RSP · Respiration

SVM · Support Vector Machines

TP · True Positive

TPR · True Positive Rate

TN · True Negative

TNR · True Negative Rate

VAE · Variational Auto-Encoder

Chapter 1

Introduction

1.1 Goals and contributions of the thesis

The rapid progress of Artificial Intelligence (AI), robotics, and automation is profoundly reshaping global industries and societies. Anticipated transformations in our work routines, lifestyles, and social interactions are expected to unfold at an unprecedented speed and scale, surpassing any previous changes in human history. Recent technological advancements have paved the way for automation across various domains, encompassing educational learning, autonomous driving, medicine, customer service, and more. Affective computing, situated at the intersection of computer science, psychology, and cognitive science, stands as a pivotal frontier in improving human-machine interactions by integrating emotional understanding into technology. This interdisciplinary field aims to enable machines to recognize, interpret, and respond to human emotions.

In the realm of AI, the integration of affective computing is not merely an advanced feature; it's a fundamental shift towards more intuitive, responsive, and human-centric systems. This evolution is particularly salient in applications ranging from healthcare, where empathetic patient care can be revolutionized, to customer service, where understanding and responding to consumer emotions can greatly enhance service quality. Moreover, incorporating affective computing into autonomous systems, such as electric vehicles and smart homes, could lead to more nuanced and user-friendly experiences. The development of sophisticated algorithms capable of analyzing and interpreting emotional cues - from facial expressions to voice tonality - is pivotal. This necessitates a multidisciplinary approach, intertwining ML techniques with insights from psychological research to create algorithms that are not only technically proficient but also culturally and ethically aware. Thus, the pursuit of advanced affective computing technologies is not just a technological endeavor; it's a step towards more empathetic, understanding, and ultimately human-friendly machines. This paradigm shift has the potential to redefine the dynamics of our interaction with technology, making it more seamless, intuitive, and aligned with our emotional needs and states.

Much of the research in affective computing centers around recognizing human emotional states. This involves leveraging various modalities like physiological sensors and cameras to capture human reactions such as facial expressions. However, the development of effective and robust models in this affective computing domain encounters several obstacles. Challenges such as data scarcity, labeling uncertainty and cost, dataset imbalance, information fusion from multiple sources, model explainability, personalized model requirements, privacy concerns, and ethical considerations pose significant obstacles. Among the promising domains to address these challenges, anomaly detection stands out. Its methods often do not rely on labeled data, offering

avenues for exploration in an unsupervised manner. Anomaly detection algorithms generate an anomaly score, providing unsupervised insights into the data. Consequently, we chose to explore anomaly detection methods to tackle challenges such as data scarcity, imbalanced dataset, model personalization, feature selection, model explainability, and data fusion. The contributions of this thesis can be summarized as follows:

1. **Anomaly Detection in Affective Computing:** We explore the application of anomaly detection methods for detecting rare states within the field of affective computing. This exploration is conducted through an unsupervised approach in two distinct applications: Detection of dangerous driving behavior, where supervised learning may be hindered by the collection of data pertaining to risky behavior and hence presents ethical concerns, and Relapse Prediction for patients suffering from psychotic disorders, which involves the detection of an infrequent event that may require extensive recording to capture.
2. **Global and Personalized Models:** The examination of affective computing states reveals an intervariability among individuals, presenting a substantial challenge when developing universal models for assessment. This variability becomes especially prominent when analyzing patients with psychotic relapse. In response to the aforementioned challenge, this study explores the potential of anomaly detection methods to create personalized models for psychotic relapse prediction. By leveraging unsupervised anomaly detection techniques, we pave the way for systems capable of autonomously collecting end-user data in real-life scenarios, eliminating the need for direct user intervention or costly labeling. We conduct a comparative analysis with global models and explore the strengths and limitations of both approaches.
3. **Feature Selection for Imbalanced Datasets:** When confronted with limited data (e.g. personalized models) or high labeling costs in real-life scenarios, feature selection becomes pivotal to prevent overfitting and reduce computational power consumption. We propose using Auto-Encoder (AE) reconstruction errors (RE) as a metric to examine feature relevance for the feature selection problem. We evaluate this approach across two critical applications: Visual Distraction Detection and Psychotic Relapse Prediction.
4. **Multimodal fusion:** Human affective states manifest through diverse channels like voice, facial expressions, and physiological reactions, underscoring the need for multifaceted data sources in affective state estimation systems. Integrating multiple modalities, such as cameras, physiological sensors, and microphones, is crucial for improving the efficiency of these systems. However, the challenge arises from the heterogeneous nature of these modalities, making multimodal fusion a complex task. To address this, our proposal suggests leveraging AE and their RE as cues to determine the relevance of each modality. These RE serve as valuable indicators guiding multimodal fusion algorithms in effectively combining these diverse sources of information.

1.2 Structure of the thesis and outline

This thesis is composed of a brief introduction, four chapters, and a summary of conclusions and perspectives. The main points of each chapter are presented below.

- *Chapter 2: Background and Open Challenges.* This chapter serves as a comprehensive introduction to the field of affective computing, laying the foundation for our research goals. We explain the motivations behind our work and the need for advancements in this field, with a particular focus on the challenges it presents. Additionally, we introduce the domain of anomaly detection and its pivotal role in addressing these challenges within the field of affective computing.

- *Chapter 3: Rare Mental States Detection.* We explore the detection of rare mental states in an unsupervised manner. Our study focuses on addressing two real-world problems: Visual Distraction detection and psychotic relapse prediction. By employing unsupervised methodologies, we aim to shed light on innovative approaches for identifying these infrequent yet critical states without the need for labeled data, which can be difficult in those applications.
- *Chapter 4: Explainability and Feature Selection Using Anomaly Scores.* In this chapter, we study the use of a feature selection method designed for imbalanced datasets for the two affective computing tasks: visual distraction detection and psychotic relapse prediction datasets detailed in Chapter 3.
- *Chapter 5: Multimodal Fusion using Anomaly scores.* This chapter explores a novel fusion approach aimed at enhancing continuous emotion prediction through a multimodal fusion technique founded on anomaly scores. These anomaly scores are derived from AE trained per modality. We study this fusion technique using the ULM TSST dataset for the prediction of arousal and valence for participants in stress-induced situations.

1.3 Publications and Patents

Through the research conducted in this thesis, our contributions have resulted in several publications.

1. **Articles**

    - Hamieh, S., Heiries, V., Al Osman, H., & Godin, C. (2021). Multi-modal fusion for continuous emotion recognition by using auto-encoders. In Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge (pp. 21-27).
    - Hamieh, S., Heiries, V., Al Osman, H., & Godin, C. (2023, June). Relapse Detection in Patients with Psychotic Disorders Using Unsupervised Learning on Smartwatch Signals. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-2). IEEE.
    - Hamieh, S., Heiries, V., Al Osman, H., Godin, C., & Aloui, S. (2023)Driver Visual Distraction Detection Using Unsupervised Learning Techniques. In ITSC 2023 IEEE International Conference on Intelligent Transportation.
    - Hamieh, S., Heiries, V., Al Osman, H., & Godin, C. (2023). Psychotic Disorders Relapse Prediction from Passive Signals using Anomaly Detection Methods.(under review)

2. **PATENT**

    - Hamieh, S., Heiries, V., Al Osman, H., & Godin, C. Multi-modal prediction system. EP4163830.

CHAPTER 2

Background And Open Challenges

In this chapter, we provide an overview of the two fields of affective computing and anomaly detection. We define Affective Computing and explore its wide-ranging applications. We define anomaly detection and describe relevant classical and SOTA methods and evaluation measures. We discuss the challenges of both domains and highlight the ones we will address in the thesis.

Chapter contents

2.1 Introduction

The relevance and significance of understanding human emotions in interaction with machines have surged recently [1] [2] [3]. The primary objective of this thesis is to delve into understanding human mental states by utilizing specific data sources such as speech, physiological signals, and facial expressions. These data are instrumental in estimating these states, requiring models for interpretation. Given the absence of mathematical models, we resort to Machine Learning (ML) methodologies. The conventional approach predominantly relies on supervised ML models, which have been instrumental in interpreting and analyzing emotional cues[4] [5]. However, supervised classical ML approaches come with limitations when dealing with the complexities of human emotions and challenges in data collection. That's why we propose an exploration of anomaly detection methods as a potential avenue for improving mental state recognition. This chapter serves as a comprehensive exploration of Affective Computing, encompassing its diverse applications, modalities, information fusion, and the challenges

encountered in this domain. We devote a section to the most used approaches in the classical approach of supervised learning within affective computing. Subsequently, we delve into anomaly detection, an intriguing field promising to augment affective computing. The discussion spans anomaly detection methods and the associated challenges.

2.2 Affective computing

2.2.1 Affective computing definition

Affective computing stands as an interdisciplinary domain that encompasses computer science, psychology, and cognitive science. It holds significant implications for the domain of human-computer interaction, where the fusion of these disciplines holds the potential for transformative advancements. The concept of affective computing was first introduced by Rosalind Picard in 1995 [6]. It refers to the field of developing devices that can identify, interpret, process, and even mimic human emotions. In Picard's influential work [6], she underscores the essential role that emotions play in human cognition and perception. She goes on to elaborate on how integrating emotional understanding into technology could lead to the creation of more intuitive and efficient intelligent machines.

2.2.2 Emotion recognition

One of the main and initial goals of affective computing is emotion recognition. Psychologists endeavor to describe emotions, and these descriptions are referred to as models. There are generally two approaches to emotional modeling:

- Discrete class model: Categorical Emotions approach.
- Continuous value model: Dimensional Emotion approach.

We will elaborate on each approach in the following subsections.

2.2.2.1 Categorical approach

In the categorical approach, we define emotions as a set of discrete classes. The development of such models enables the scientific community to effectively distinguish and methodically arrange these emotions. Numerous researchers have identified a list of primary or fundamental emotions, considered to be innate and critical for the species' survival. The number of these emotions varies depending on the author. Two well-recognized models frequently cited in literature include Ekman's Basic Emotions model and Plutchik's Wheel of Emotions.

- Ekman [7] identified six emotions, shown in Figure 2.1: anger, disgust, fear, joy, sadness, and surprise that are distinct and universally recognized.
- In 1980, Robert Plutchik [9] developed the Emotion Wheel shown in Figure 2.2, which offers a valuable framework for comprehending emotions and their functions. This wheel is divided into eight sections, each representing one of the eight primary emotions: joy, trust, fear, surprise, sadness, anticipation, anger, and disgust. Each primary emotion has an opposing polar emotion based on the physiological responses it triggers. For example, joy is the opposite of sadness. Additionally, emotions become more intense as they progress from the outer to the inner parts of the wheel, a gradation that is also reflected in their shading; the darker the hue, the more intense the emotion. Furthermore, the Emotion Wheel also illustrates compound emotions that result from the combination of two basic emotions, such as aggressiveness, which arises from the fusion of vigilance and rage emotions.

FIGURE 2.1: *Ekman's six basic emotions. [8]*

2.2.2.2 DIMENSIONAL APPROACH

As for the dimensional approach, its coordinates in the Euclidean space characterize the emotion. Russel's Circumplex Model of Affect [10], represented in figure 2.3, is one of the most widely adopted dimensional representations in affective computing. The circumplex model of emotion proposes that emotions are arranged within a two-dimensional circular space, defined by the dimensions of arousal and valence.

1. **Valence:** Valence is represented along the horizontal axis. The dimensions of valence represent the positivity or the degree of pleasure/displeasure. For instance, both anger and disgust are considered unpleasant emotions and rank high on the displeasure scale.
2. **Arousal:** Arousal is depicted along the vertical axis. The dimension of arousal reflects the activation (low/high) or the stimulation experienced by an individual. For example, fatigue is associated with low activation, indicating an unenergetic state, while emotions like anger are linked to high activation, signifying a heightened and intense state of stimulation.

Another known emotional model is the PAD emotional state model [11]. It proposes a three-dimensional representation of emotions that includes **P**leasure, **A**rousal, and **D**ominance. Dominance represents the degree of control. While dominance is not as commonly used as dimensions like valence and arousal, it can provide additional insights into the impact of emotions on cognitive and behavioral processes.

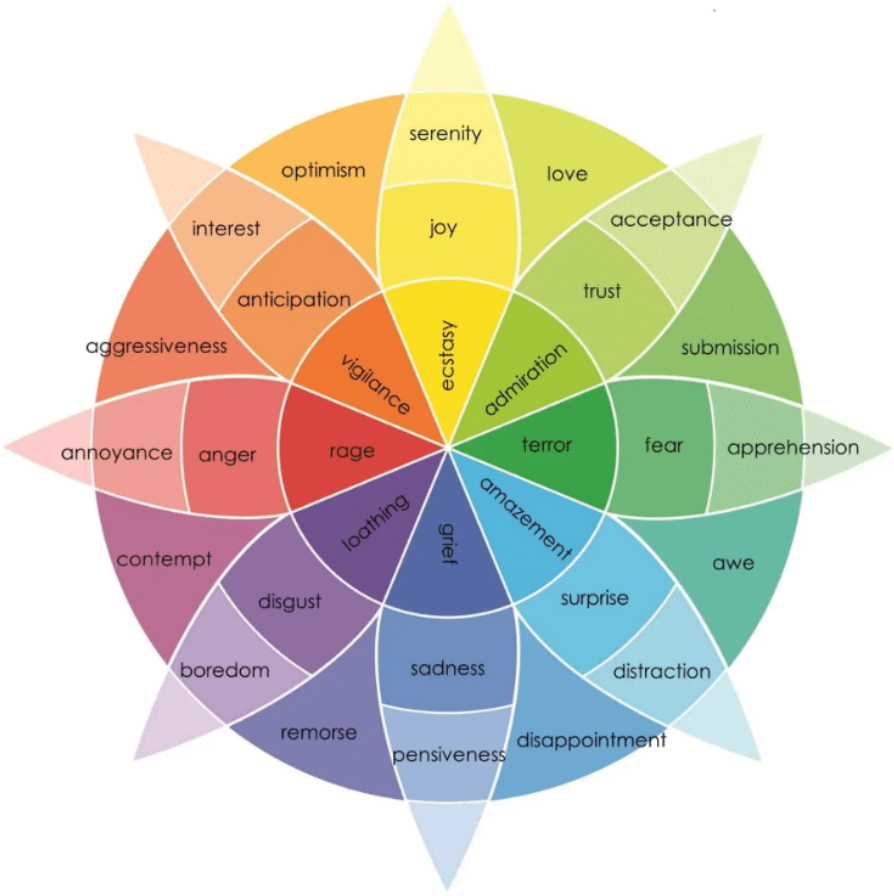

FIGURE 2.2: *Robert Plutchik's Wheel of emotions*

2.2.3 OTHER MENTAL STATE ESTIMATION

Additionally, the scope of research in affective computing has expanded beyond basic emotional states to encompass a wider range of human expressions and mental states. For instance, studies focus on states like laughter [12], attention [13], pain [14], psychological states [15] [16] [17], social behavior analysis to identify people's viewpoints on various subjects [18], and detecting protective behavior [19]. In the following section, we'll explore different domains influenced by affective computing, highlighting diverse human states studied for predictive analysis.

2.2.4 APPLICATIONS OF AFFECTIVE COMPUTING

As technology continues to evolve and automate various sectors, affective computing has the potential to revolutionize numerous industries. Below are several domains showcasing the applications of affective computing.

- **Customer Service and Human-Robot Interaction**: Affective computing can enhance customer service interactions by enabling robots or virtual assistants to recognize and respond appropriately to users' emotions. Examples of such works: developing emotionally smart chatbots [20], personalized robots for better efficiency [21], real-time Speech emotion recognition system [22].
- **Healthcare**: Affective computing can be utilized in mental healthcare to monitor patients' psychological states and provide personalized interventions. Some work focused on proactive identification of mental health concerns in university students [23], depression detec-

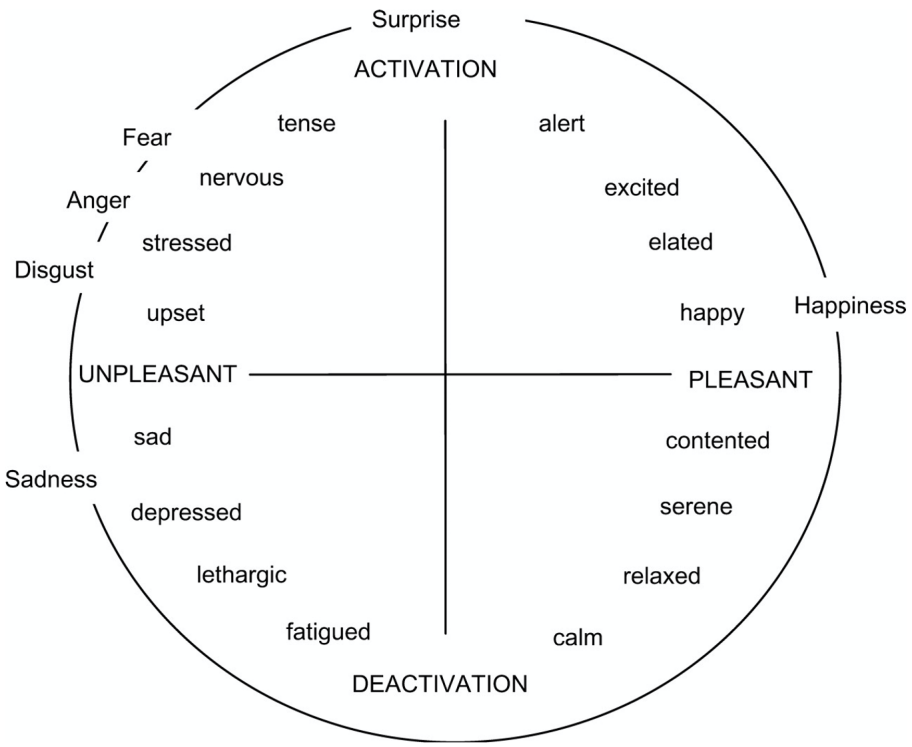

FIGURE 2.3: *Russell's (1980) Circumplex Models*

tion [15], anxiety detection [16] [24], bipolar disorder diagnosis [17] [25] It could help in managing stress, anxiety, and depression by analyzing facial expressions, physiological signals, and speech patterns.

- **Wellbeing**: Affective computing has also been used in monitoring human wellbeing. For example, measuring the effects of social media on its users [26], professional burnout detection [27], stress prediction [28], identification of urban environment effects [29].
- **Automotive sector**: Affective computing can contribute to a safer and more comfortable driving experience. It can help the vehicle adapt its behavior based on the driver's state. Therefore, studies have explored using affective computing for driver drowsiness prediction [30], driving distraction detection [31], and monitoring driver's emotion and behavior in different conditions [32].
- **Education**: In the field of education, affective computing can enhance learning experiences by adapting content and teaching methods based on students' emotional responses. Numerous studies have focused on detecting individuals' emotional states, e.g. engagement and confusion, to enhance the learning process, particularly in online educational settings [33] [34] [35] [36]. For example, [13] proposes a model to detect the level of engagement for special needs students online by analyzing video data.
- **Marketing and Advertising**: Affective computing plays a crucial role in enabling marketers to measure consumers' emotional reactions to advertisements, products, and brands. This information holds the potential to refine marketing strategies, making them more captivating and effective. Consequently, a number of studies have been dedicated to the application of affective computing in the realm of marketing and advertising [37] [38]. For instance, affective computing has been employed to predict purchase likelihood [39], enhance email marketing [40], and forecast emotional (in)congruency [41].

Moreover, affective computing also contributes to enhancing other domains like smart en-

vironment [42], gaming and entertainment [43], therapeutic interventions[44]. These are just a few examples of the diverse applications of affective computing. The field continues to expand as researchers and practitioners explore new ways to integrate emotional understanding into technology to improve various aspects of human life.

As shown above, affective computing, with its ability to perceive, interpret, and respond to human emotions, finds many applications across various domains. More information can be found in [45]. It provides a general overview of incorporating affective computing in software systems.

In our research, we have strategically chosen to focus on three distinct applications within the domains of well-being monitoring, the automotive sector, and mental healthcare. Our selection is motivated by a commitment to proactive and preventative approaches aimed at enhancing the quality of human life through advanced emotionally intelligent technologies.

The first application we introduce in this thesis is **distraction detection** in the Automotive Sector. Distracted driving remains a leading cause of accidents and fatalities on the road. Leveraging affective computing methods, we intend to develop systems capable of detecting driver distraction in real-time. The prevention of accidents and the promotion of road safety are paramount concerns. Our research aims to enhance the driving experience by reducing the risks associated with distracted driving.

The second application we present is **the detection of psychotic relapse** in mental healthcare. Individuals with psychotic disorders often experience relapses that can be challenging to predict and manage. Affective computing offers an opportunity to monitor and predict relapses by analyzing behavioral cues. Early detection and intervention in psychological relapses can be transformative for patients with psychotic disorders. It can lead to faster recoveries, reduce hospitalization rates, and improve the overall quality of life for affected individuals.

The third application is **stress Prediction** in wellbeing monitoring. Stress is a pervasive issue in modern society, with profound implications for individual well-being and overall public health. By harnessing the capabilities of affective computing, we aim to proactively identify and predict stress levels in individuals. The ability to predict stress can be a crucial factor in preventing burnout, promoting mental health, and improving overall quality of life. It enables early interventions and personalized support mechanisms for individuals under stress, leading to healthier and more resilient communities. Moreover, all of these applications could benefit from anomaly detection. This is especially relevant given their requirement for collecting states that are either rare or challenging to record in real-life settings.

2.2.5 Modalities used in affective computing

The focal point of affective computing lies in analyzing human signals to gain insight into human behavior. Within this domain, various types of features have demonstrated effectiveness in recognizing emotions or psychological states, including speech, facial expressions, digital data, and physiological signals.

2.2.5.1 Speech

Speech signals represent the most natural and informative form of human communication. Additionally, it offers the advantage of being easily and cost-effectively collected. Within speech, information is embedded explicitly through linguistic components like words and implicitly through acoustic elements like prosody and specific vocalizations (e.g., laughter or cries).

- **Paralinguistic Acoustic Features** We present an overview of the diverse speech representations applied in the field of affective computing as shown in Figure 2.4.

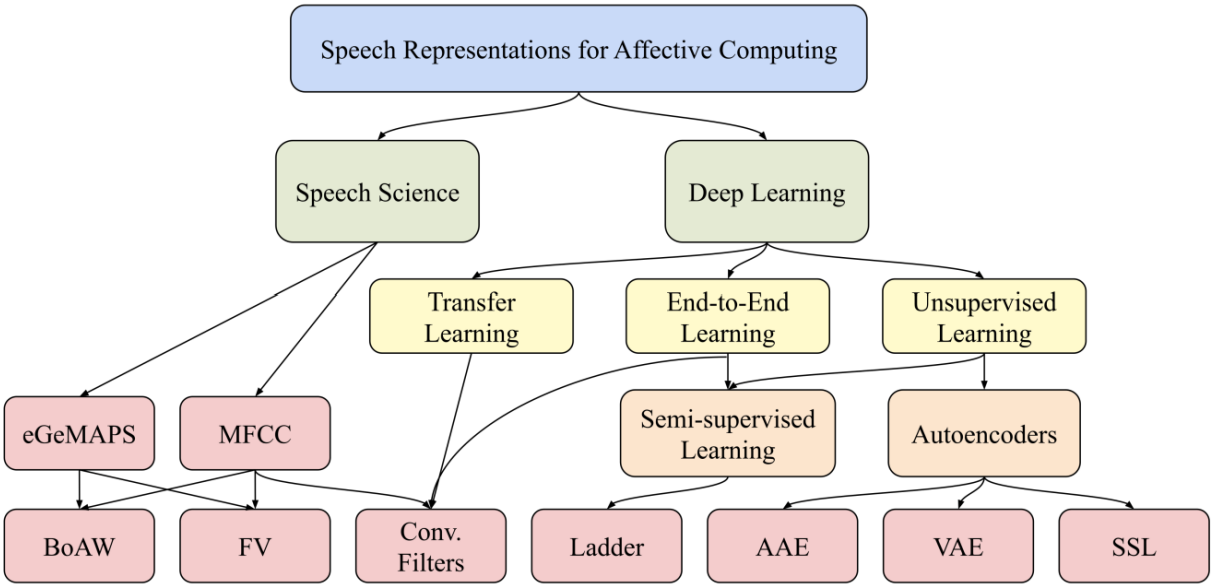

FIGURE 2.4: *Overview of speech representations used in the domain of affective computing from [46]. Abbreviations used in the figure are Geneva Minimalistic Acoustic Parameter Set (GeMAPS): Geneva Minimalistic Acoustic Parameter Set, MFCC: Mel-Frequency Cepstral Coefficients, BoAW: Bag of Audio Word, FV: Fisher Vector, AAE: Advarsarial AE, VAE: Variational AutoEncoder, SSL: SelfSupervised Learning*

Hand-crafted Features: Acoustic speech features can be categorized into three primary groups: prosodic features, qualitative features, and spectral features. Among the prosodic features are energy, zero-crossing rate, and frequency of pitch. Qualitative features encompass jitter and shimmer. The spectral features category includes the Mel-Frequency Cepstral Coefficient Mel Frequency Cepstral Coefficients (MFCC), Linear Prediction Coefficient, Linear Prediction Cepstral Coefficient, Gamma tone Frequency Cepstral Coefficient, Perceptual Linear Prediction, and formants. Numerous techniques have been introduced to extract speech descriptors, often combined with various statistical measures to summarize their temporal patterns. Consequently, the results of early emotion recognition studies lacked comparability and interpretability. Therefore, collaborative efforts have aimed to establish a concise set of acoustic descriptors based on expert knowledge, leading to representations such as the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) [47]. Furthermore, instead of analyzing these acoustic descriptors stochastically, an alternative approach was suggested involving the clustering of these descriptors. The most popular techniques are Bag of Audio Words (BoAWs) [48] and Fisher Vectors (FVs) [49]. However, some studies showed that features extracted from Deep Learning DL models outperform hand-crafted features [50].

DL Features: DL empowers models to extract features at a higher level of abstraction, enabling them to learn information from speech that is unrestricted by human understanding.

1. End-to-end Learning: One approach that has gained prominence is end-to-end learning. Here, raw data is directly input into the model, which independently manages both feature extraction and prediction. This method is typically implemented through Convolutional Neural Network (CNN) and has demonstrated superior performance compared to traditional acoustic features [50]. However, a significant drawback of this approach is its demand for larger datasets due to the heightened complexity and increased number of parameters involved in end-to-end modeling. Additionally, choosing the appropriate network architecture and

learning method can be challenging, leading to various architecture and hyperparameter tuning studies.

2. Transfer Learning: Speech can also be processed as an image by using the spectrogram or Mel-spectrogram. Therefore, some well-established architectures from the field of computer vision for image processing have been adapted for speech e.g., VGGish[51], Deep Spectrum [52].
3. Unsupervised Learning: The scarcity of emotion-labeled data, despite the vast amount of unlabeled recordings, has prompted research in affective computing to explore more label-agnostic approaches. For example, in [53] a Recurrent Neural Network (RNN)-based AE is trained on a large amount of unlabeled data to generate a latent representation of emotional speech, thereby enhancing emotion recognition performance. Other works explored the use of Variational Auto-Encoder (VAE) [54], or Adversarial AE (AAE) [55].
4. Semi-supervised learning: Semi-supervised learning can be achieved through a two-step process. Initially, a generic representation is learned in an unsupervised manner, and subsequently, more task-specific features are extracted using labeled data. In [56], authors employed CNN s to learn local invariant features, and subsequently, these same layers can be fine-tuned to identify emotionally salient features by leveraging labeled data. To eliminate the need for a separate unsupervised pre-training step, researchers extended Ladder networks [57], a variant of denoising AE. This extension allows for the simultaneous minimization of both supervised and unsupervised cost functions [57].
5. Self-supervised learning: In self-supervised learning, the model learns a universal data representation during its training by engaging in a predefined task using only the available data. Among the most commonly used tasks in the literature is Contrastive Predictive Coding (CPC), which involves distinguishing a masked frame from another frame [58]. One of the most known self-supervised models for speech is wav2vec2 which performs better than the best semi-supervised learning methods [59]. WavLM [60], another self-supervised model, leads the SUPERB leaderboard [61], which serves as a benchmark for universal performance in speech processing.

While handcrafted speech representations offer interpretability, they exhibit significant variability across speakers, rendering models vulnerable to generalization errors when dealing with unfamiliar individuals. Conversely, DL based representations can enhance robustness but lack the direct explanatory power of acoustic and linguistic features. Given the diverse real-world applications of affective computing, it becomes imperative to achieve not only accurate assessments but also supplementary insights into decision-making processes. Furthermore, representations based on deep learning often necessitate larger datasets, heightened model complexity, and increased training efforts. Despite numerous studies in this domain, comprehensive comparisons encompassing a wide spectrum of representations for diverse affect-related tasks employing various models across distinct contexts, languages, and cultures, remain scarce. Thus, despite recent advancements, fundamental questions such as the selection of optimal features for emotion recognition persist [46].

- **Linguistic Features** Affect is not only conveyed through the manner of expression but also by the actual words spoken. In the process of text processing, the initial step involves tokenization, where the text is broken down into machine-readable units [62]. Subsequently, an algorithm is applied to generate embeddings, which are vectorized representations of the text's content derived from these tokens. As an example, in the case of word embed-

dings, the text is segmented into words, treating them as fundamental elements of a sentence, with each word corresponding to a distinct feature vector. Consequently, in textual data, each token is associated with and represented by a single, unique, and deterministic embedding. This approach ensures the existence of a finite set of targets for analysis and manipulation. More recently, the use of transformer-based models has enhanced the performance of affect recognition [63] [64] [65] by using models like Bidirectional Encoder Representations from Transformer (BERT) [66], GPT-2 [67], XLNet [68], RoBERTa [69].

For a more comprehensive exploration of extracted features, the following surveys provide in-depth insights: [70], [71].

2.2.5.2 Facial expressions and body movements

One of the most intuitive ways to recognize emotions by humans is by examining facial cues. Mehrabian's research findings suggest that within a message, the verbal component is responsible for only 7 percent of the overall impact of the message. In contrast, the vocal aspect (such as voice intonation) is responsible for a more significant portion, at 38 percent, while the facial expressions of the speaker are responsible for the most substantial influence, accounting for 55 percent of the overall effect of the spoken message [72]. This suggests that facial expressions constitute the primary modality in human communication. Ekman developed the Facial Action Coding System (FACS) [73], a method for describing facial movements. It decomposes the facial expression into a group of particular muscle movements called Action Units (AU), e.g., jaw drop, as shown in Figure 2.5 with other examples. Similar to speech, in earlier research, expert-knowledge features such as Local Binary Patterns (LBP) [74], Histograms of Oriented Gradients (HOG), Multiscale-WLD, Local Directional Patterns (LDP), and Gabor [75] were used as video baseline features for emotion prediction. More recently, deep representations [76] [77] extracted through deep Convolutional Neural Network CNN (e.g. VGG-16 [78]) have been used.

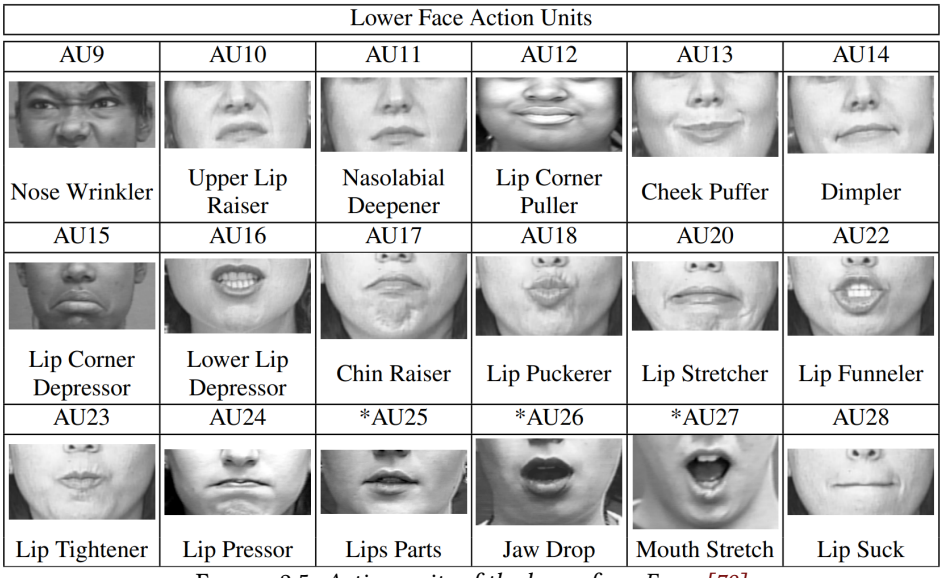

Figure 2.5: *Action units of the lower face. From [79].*

Facial expression, although the most prominent avenue of emotional communication, presents a disadvantage in that it is often more controlled compared to other forms due to its sensitivity to various social contexts.

2.2.5.3 Physiological signals

In addition to observable behavioral changes (e.g. facial expressions and tone) resulting from changes in a person's emotions, there are also noticeable alterations in physiological signals [80]. We present the most relevant features and examples of its use in the domain of affective computing:

- Electrocardiogram (ECG) : ECG records electrical heart activity generated during contraction and relaxation. Common features extracted from ECG signals are Heart Rate (HR)and Heart Rate Variability (HRV). HR represents the number of heartbeats per unit of time, while HRVquantifies the variations over time in the period between successive heartbeats. These features derived from ECG signals play a significant role in the analysis and interpretation of emotional states [81], pain assessment [82], etc.
- Photoplethysmography (PPG) or Blood Volume Pulse (BVP): BVP employs a photodiode to measure backscattered light by a skin voxel, which is proportional to the blood volume. Moreover, PPG is frequently used to measure HR and HRV [83]. In [84], authors transformed monodimensional PPG signals into images and used DL for cognitive load detection.
- Electrodermal Activity (EDA): EDA is assessed through skin resistance by passing current or voltage through the body and recording variations in voltage or current between sensor leads. For quite some time, EDA has been a staple in assessing human emotions. Traditionally, sensors were attached to the fingers for this purpose. However, with the advancements in affective computing and the advent of smartwatches, the sensors have transitioned to the wrist [85]. Driver drowsiness detection [86] and driver stress prediction [87] are some of the applications within the field of affective computing using skin conductance.
- Respiration (RSP) : RSP monitors respiration patterns, including the speed and depth of a person's breathing. Most of the sensors are chest-belt mounted piezoelectric sensors however more technologies are detailed in [88]. Ihmig et al. [89] employed RSP along ECG and EDA signals to detect and assess anxiety levels.
- Electromyography (EMG): EMG records the electrical activity of skeletal muscles, utilizing skin surface electrodes. Findings in [90] indicate that the EMG (Electromyography) signal performed comparably to the well-established ECG signal in the domain of stress detection and demonstrated that the EMG signal from the right trapezius muscle exhibits superior stress recognition capabilities compared to other muscles.
- Electroencephalogram (EEG) : EEG is a technique that records the electrical fields produced when neurons in the cerebral cortex become active during synaptic excitation. EEG is frequently employed as a valuable tool in the supplementary diagnosis of mental health conditions, including disorders like depression [91]. It is used in several affective computing applications such as emotion recognition [92], fatigue estimation [93], etc.

2.2.5.4 Other behavioral indicators

Alongside the previously mentioned features, various additional parameters have been explored for different applications. These include eye-tracker signals [94][95], vehicle information [96], internet data [97], signals retrieved from smartphones—such as call logs, social media activity duration, and overall mobile usage [98]. Moreover, data from smartwatches, encompassing details like sleep patterns and physical activity [99] further enrich the contextual information available for analysis. This diverse array of features augments the depth and breadth of understanding across various scenarios.

2.2.6 Multimodal fusion

Multimodal fusion refers to the integration of diverse data from varying modalities to leverage the complementary nature of these data sources, ultimately enhancing prediction performance. This is particularly relevant in affective computing, where a range of information sources, as shown in the previous section, is available for analysis. Based on the extensive literature on this topic, multimodal fusion is categorized into three categories: feature-level fusion, decision-level fusion, and hybrid fusion [100], shown in Figure 2.6. In this section, we expand upon the fundamental principles of these techniques.

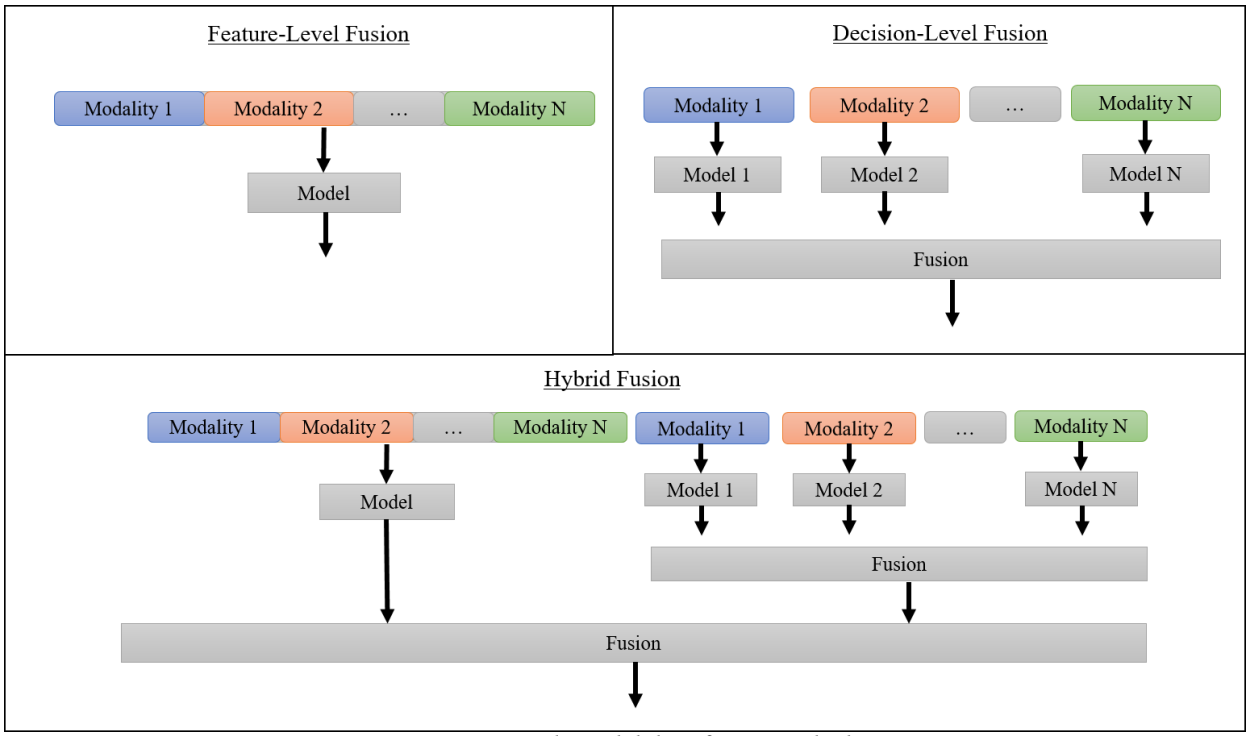

Figure 2.6: *Multimodal data fusion methods.*

2.2.6.1 Feature-level fusion

Early fusion, also known as feature-level fusion, involves the simultaneous incorporation of features from all modalities into a single computational model. In this approach, the raw or preprocessed data from each modality is combined at the very beginning of the analysis pipeline, resulting in a single feature set. Early fusion enables the model to learn from all modalities jointly. It offers the advantage of promptly leveraging inter-modality relationships, facilitating task completion [101]. In [102], a multimodal method was introduced for categorical emotion recognition by incorporating EEG and micro facial expressions. They fused these modalities at the feature-level, and their fusion strategy showed considerable performance enhancement, up to 12% improvement per recognized emotion compared to the unimodal approach.

2.2.6.2 Decision-level fusion

Decision-level Fusion, also referred to as late fusion, in contrast to early fusion, focuses on combining the predictions generated by individual models trained on separate modalities. Each modality is processed independently using dedicated models, and their outputs are later merged or aggregated to make a final decision or prediction. Late fusion provides flexibility by allowing the use of specialized models for each modality, which can be advantageous when dealing with modalities of different complexities. However, it may require careful consideration of how

to combine the predictions and handle potential discrepancies between modalities effectively. Huang et al. [103] showed that late fusion is better at predicting valence and arousal than early fusion. Sun et al. [104] chose late fusion with aLong Short-Term Memory (LSTM) network that captures dynamic information as a fusion model.

2.2.6.3 Hybrid fusion

A hybrid-fusion scheme involves combining feature and decision-level fusion techniques. For instance, Kim et al. [105] integrated audio and physiological data through three stages. The first stage is a feature-level fusion of audio and physiological signals. The second stage is decision-level fusion utilizing unimodal predictions from both audio and physiological sources. Finally, they conducted another round of decision-level fusion on the output produced by each of the preceding fusion steps. Chen et al. [106] proposed the joint use of early and late fusion using bidirectional deep long short-term memory networks. The results showed that early and late information may be complementary [106].

2.2.7 Discussion

Early fusion may face various challenges. Concatenating features from all modalities increases the dimensionality of the input data. This can pose difficulties when the size of the dataset is not sufficiently large to support this expanded feature space. In such cases, the risk is that the model may overfit the data, leading to decreased classification accuracy. Another challenge associated with early fusion involves data alignment and compatibility between modalities. Modalities may capture data at different time scales or have inherent differences in data type and format. Ensuring that these diverse sources of information can be effectively integrated and compared can be a complex task. Additionally, early fusion offers less flexibility in scenarios involving corrupted or missing modalities. When one modality is compromised or unavailable, early fusion may struggle to adapt. In contrast, late fusion offers more flexibility; it can be adapted if a single modality is missing more easily than early fusion; however, it does not take into consideration the inter-modality relationship. There is no consensus in the literature on the best approach, given the limitations and advantages of each technique. The choice between different fusion methods often relies on the dataset specifics and the specific requirements of the task being addressed.
We've highlighted some widely recognized fusion strategies here. However, there are additional surveys that delve further into categorizing various data fusion techniques, including attention-based and DL-based fusion as explored in [101], model-level, rule-based, estimation-based, and classification-based fusion methods as detailed in [107].

2.2.8 Open challenges

Affective computing, although promising, remains in its early stages and requires significant enhancement to align with real-world application constraints. It encounters various challenges, particularly concerning sensors (which should be unobtrusive, consume low power, and ensure data privacy), models for recognizing mental states (which need to be efficient, interpretable, fault-tolerant, and personalized), actions chosen by the models (which should also be personalized and beneficial). As this thesis focuses on mental state estimation, we list some of the main challenges in affective computing and categorize them into three groups: data, models, and ethics.

2.2.8.1 Data challenges

- **Imbalanced dataset:** In affective computing, a significant challenge arises from dealing with unbalanced datasets. This issue occurs when the distribution of affective states in the dataset is uneven, meaning that some classes are more frequently represented than others. For example, the IEMOCAP dataset [108] and FER2013 dataset [109], used for emotion and facial expression recognition, respectively, exhibit imbalanced distributions among emotion categories. Similarly, the eDAIC-Woz dataset [110] is employed for depression detection, also facing imbalanced data concerns. The imbalance can lead ML models to exhibit bias and reduced performance, as they may become more adept at predicting the prevalent emotions while struggling with those less frequently encountered. Therefore, addressing unbalanced datasets is crucial for creating models that generalize well across the entire spectrum of affective states.
- **Labeling:** The main method for acquiring labeled training data in affective computing is by manual annotation of data by experts. However, this approach can be exceedingly time-consuming and costly [111][112]. Furthermore, emotions are inherently ambiguous and are perceived relatively, resulting in uncertainty in their labeling, especially when annotated by a single annotator. Such interpretations can be influenced by various factors, including the annotator's mood and prior exposure to similar examples [113]. error and subjectivity.
- **Large and Diverse Data:** To make progress in the field of affective computing, there is an urgent need to create high-quality and large datasets [114]. This need is particularly crucial when employing advanced methods like DL, known for handling complex tasks and utilizing various sources of information, as seen in many affective computing applications [115] [116].
- **Context:** Human beings instinctively assess emotions in social interactions by considering environmental and social factors. Our comprehension of social interactions is enriched by contextual elements such as the ongoing activity, the individual's identity, their customary emotional expressions, and the presence of other people [117]. In the absence of such context, even humans can misinterpret facial expressions, vocal tones, or body language. However, the majority of released datasets are typically confined to controlled laboratory environments [118] or involve acted performances [119] [120]. Consequently, this limitation has the potential to undermine the model's performance when applied to real-world settings [121].

2.2.8.2 Model challenges

- **Multimodal Fusion** While the advantages of fusion techniques (such as audio-video fusion) for affective computing are anticipated from both engineering and psychological standpoints, our understanding of how humans accomplish this fusion remains quite limited. Neurological studies on the fusion of sensory neurons suggest a preference for early fusion (i.e., feature-level fusion) over late fusion (i.e., decision-level fusion)[122]. Yet, a persistent challenge remains in constructing suitable joint feature vectors that encompass features from diverse modalities characterized by distinct time scales, metric levels, and dynamic structures, all within the constraints of existing methodologies [123] [124]. Given the current knowledge and techniques, several issues related to fusion demand further investigation, including determining the optimal level of information fusion from different data modalities (feature level, decision level, or hybrid), identifying the ideal function for integration and incorporating reliable estimations of each modality's predictions.
- **Explainability:** An ongoing challenge in the field of affective computing is the development of ML models that are both robust and interpretable. Driven by the imperative for

accuracy, there is a discernible trend toward adopting "black-box" solutions, particularly those rooted in neural networks, predominantly Deep Learning [125] [126]. Models like deep neural networks are often considered "black box" models, posing difficulties in comprehending their inner workings and hindering trust and customization [127]. Moreover, the need for explainability is of great importance in fields like education and healthcare, where the stakes are high and incorrect generalizations can have significant human consequences [128].

- **Personalization:** The significance of personalized models in the domain of affective computing is strongly substantiated by empirical findings in social and behavioral sciences. For example, individuals from different cultures may employ nonverbal gestures in distinct ways, and there can be variations in the frequency of their utilization [129]. Moreover, research studies [130] [131] provided evidence of variations in personality traits among diverse user profiles, encompassing distinctions in age groups, genders, and even cultural backgrounds.

2.2.8.3 Ethical challenges

- **Bias:** Bias is a significant concern within affective computing models. Kiritchenko et al. [132] showed gender and race bias in experiments done using different algorithms, including traditional and DL models. During the assessment of affective computing systems, a prominent pattern was observed wherein sentences with African-American names tended to receive higher scores in tasks related to the prediction of anger, fear, and sadness intensity. Conversely, for tasks involving the prediction of joy and valence, most submissions favored sentences featuring European American names. Moreover, Diaz et al. researched age-related bias in sentiment analysis and found that sentences containing "young" adjectives are 66% more likely to receive positive scores compared to identical sentences with "old" adjectives [133]. Many research efforts predominantly focus on identifying and mitigating gender bias despite the existence of more pronounced biases in areas such as race, religion, and intersectionality, which require significant attention [134].
- **Privacy:** In the domain of affective computing, privacy is one of the most important ethical concerns. In numerous affective computing applications, sensitive human data is frequently employed, some of which have the potential to uniquely identify an individual [135]. Studies have shown that data stored in remote cloud servers are vulnerable to attacks, and inferred information can be maliciously used [136] [137]. Moreover, some applications like emotion recognition raise the issue that emotions are private information. Hence, this raises the ethical question: is it acceptable to let computers extract such information? This question should always be taken into consideration to design only valuable affective computing technologies.

2.3 Supervised learning methods in affective computing

2.3.1 Introduction

Supervised learning is used in most affective computing systems. These systems typically utilize various signals as inputs, with the output class labels corresponding to specific affective states. Initial studies employed traditional classifiers, while later advancements led to the widespread use of DL methods. In this section, we highlight some of the prominently employed models in affective computing, spanning from traditional approaches to DL methodologies, many of which will be further used in subsequent chapters.

2.3.2 Classical supervised approaches

We list here some of the most used classification models in supervised learning:

2.3.2.1 Support vector machine

Support Vector Machines (SVM) [138] is a class of ML algorithms known for its efficacy in solving binary classification problems. It has been used heavily in the domain of emotion recognition [139]. SVM aims to identify an optimal hyperplane that maximizes the margin between two classes in the feature space. Let's consider a dataset $\Omega = \{(x_1, y_1), (x_2, y_2), ...(x_N, y_N)\}$ of N data samples, where $x_i \in \mathbb{R}^d$ represents the sample vector and $y_i \in \{1, -1\}$ denotes its class. SVM [138] consists of finding the optimal hyperplane that separates both classes with the maximal margin. The hyperplane is defined by the following equation $w^T x + b = 0$ with $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ such that $y_i(w^T x_i + b) \geq 1$. This last constraint forces each class to be on one side of the hyperplane. In case the data cannot be separated by a hyperplane in their original space, data will be projected to another space of higher dimension by using a non-linear function ϕ where it can be separated. The most popular choices for the kernel function ϕ are polynomial, sigmoid, or Gaussian radial base function. An example is shown in Figure 2.7, where x data cannot be separated by a line. We apply a polynomial kernel where $\phi(x) = x^2$. After projecting the data into its new space, it can be separated by a straight line.

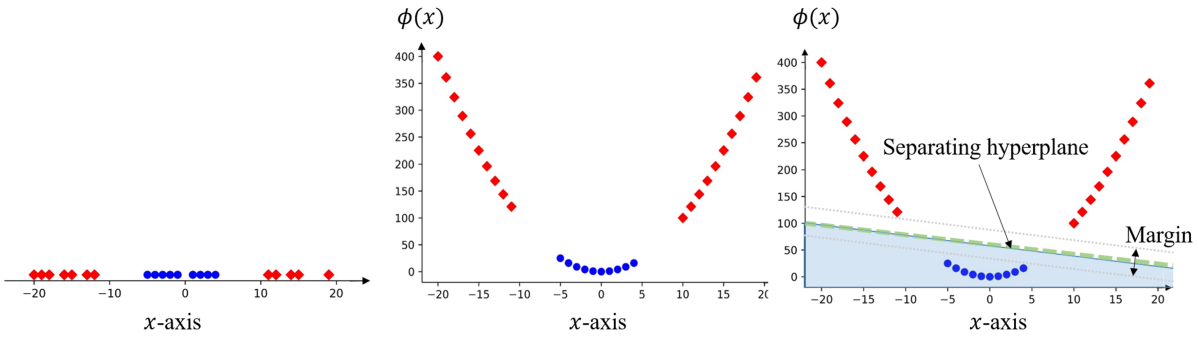

Figure 2.7: *Example of data separated using SVM.*

To make the model more robust to noisy data, slack variables can be introduced ξ_i. These variables allow soft thresholding. Hence the problem of SVM can be reformulated as:

$$\min_{w,b,\xi_i} \frac{1}{2}||w||^2 + C\sum_{i=1}^{N} \xi_i \tag{2.1}$$

subject to:

$$y_i(w^T\phi(x_i) + b) \geq 1 - \xi_i \quad \forall i \tag{2.2}$$

$$\xi_i \geq 0 \quad \forall i \tag{2.3}$$

where $C > 0$ is a constant that determines the compromise between the maximization of the margin and the training error. The aforementioned problem can be solved using Lagrange multipliers.

2.3.2.2 K-Nearest-Neighbors classifier

Another popular classification method in affective computing is K-Nearest Neighbor (KNN) [140]. This classifier [141] is a variant of instance-based learning. The model does not seek a comprehensive internal model but rather preserves instances from the training data. Classification outcomes are determined by a majority vote from the closest neighbors of each data point, associating a query point with the class that appears most frequently among its nearest neighbors. K represents the number of neighbors that will be used to assign the data point's class.

2.3.2.3 Random forest

Some studies also explored Random Forest for emotion recognition [142] [143]. Random Forest Classifier [144] is an ensemble learning technique that combines the predictions of multiple decision trees to improve the accuracy and robustness of the classification task. It builds individual trees by randomly sampling the training data and features, reducing overfitting and introducing diversity among the trees. The final prediction is determined by majority voting.

2.3.2.4 Gaussian naive Bayes

Naive Bayes was used for predicting emotions from tweets [145], speech features [146][147]. The Naïve Bayes simple probabilistic classifier relies on the application of Bayes' theorem. In Naive Bayes, each attribute variable is treated as an independent variable. Therefore for N observations $(x_1, .., x_i, ...x_N)$, the probability of Class y can be written as:

$$P(y|x_1, ..., x_N) = \frac{P(y)\sum_{i=1}^{N} P(x_i|y)}{P(x_1, .., x_N)} \tag{2.4}$$

Therefore the classification rule can be written as:

$$\hat{y} = argmax_y P(y) \sum_{i=1}^{N} P(x_i|y) \tag{2.5}$$

and Maximum A Posteriori (MAP) can be used to estimate $P(x_i|y)$ and $P(y)$.
In the case of Gaussian Naïve Bayes, it is assumed that the continuous values associated with each class follow a Gaussian distribution. Therefore the likelihood of the features is calculated as follows:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}) \tag{2.6}$$

Where σ_y and μ_y are estimated parameters using maximum likelihood.

2.3.3 DL approaches

Following advancements in DL algorithms, they have become extensively employed for affective computing

2.3.3.1 Multi-layer perceptron

Researchers initially evaluated Multilayer Perceptron (MLP), an artificial neural network that typically exhibited superior performance compared to other conventional algorithms [148] [149]. MLP is a feedforward neural network that learns to map a function $f(.) : R^m -> R^o$ by learning

from a dataset of samples, where m is the number of dimensions for input X and o is the number of output dimensions. It is composed of input, output, and hidden layers. Each layer is composed of one or several neurons, where the output of each neuron in a layer is the result of applying an activation function to the weighted sum of its inputs from the previous layer. During training, the network adjusts its weights and biases using the backpropagation algorithm to minimize a loss function.

2.3.3.2 Recurrent neural networks

The Recurrent Neural Network RNN is a specialized form of artificial neural network designed to manage sequential data. Unlike conventional feedforward neural networks, RNNs handle input sequences incrementally. Each step in the sequence computes outputs based not only on the current input but also on previous computations. However, traditional RNNs suffer from issues like vanishing or exploding gradient problems, which constrain their effectiveness in capturing long-term dependencies within sequences.

- **Long short-term memory:**
  Long short-term memory LSTM is a type of recurrent neural network that allows the capture of temporal information. It can process sequential inputs by using its internal state (memory). In contrast to conventional RNN, LSTM has a cell variable c_t and three gates: input gate i_t, output gate o_t, and forget gate f_t as shown in Figure 2.8. These gates help

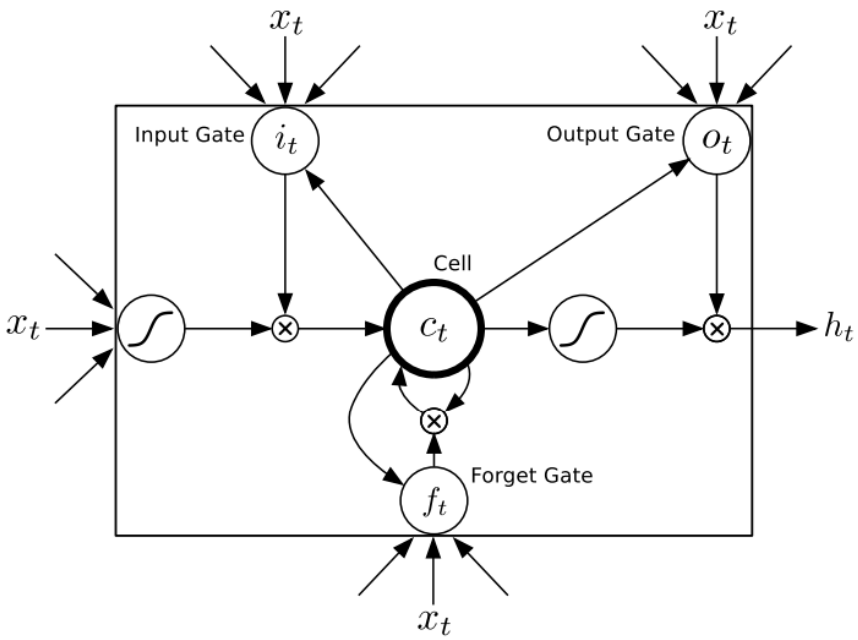

Figure 2.8: *Long Short-term Memory Cell [150].*

the LSTM overcome the vanishing gradient problem that the RNN suffers from. Moreover, it allows it to better handle long input sequences. The equations of the forward pass of an

LSTM are the following:

$$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \tag{2.7}$$
$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \tag{2.8}$$
$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \tag{2.9}$$
$$c_t = f_t \otimes c_{t-1} + i_t \otimes \sigma_h(W_c x_t + U_c h_{t-1} + b_c) \tag{2.10}$$
$$h_t = o_t \otimes \sigma_h(c_t) \tag{2.11}$$

Where x_t is the current passed input, h_t is the current hidden state, σ_g and σ_h are the sigmoid and hyperbolic tangent functions, respectively, and $\otimes$ denotes element-wise multiplication. W, U, and b are the weight matrices and biases.

- **Gated recurrent unit:**
  Gated Recurrent Unit (GRU) is a simplified version of the LSTM. It has only two gates: the update gate z_t and the reset gate r_t [151] as shown in Figure 2.9. It has fewer parameters than the LSTM. It typically has a comparable performance to the LSTM [152] [153]. A forward pass of a sample x_t through the GRU is described in the following equations:

$$z_t = \sigma_g(W_z x_t + U_z h_{t-1} + b_z) \tag{2.12}$$
$$r_t = \sigma_g(W_r x_t + U_r h_{t-1} + b_r) \tag{2.13}$$
$$\hat{h}t = \tanh(W_h x_t + U_h(r_t \otimes ht - 1) + b_h) \tag{2.14}$$
$$h_t = (1 - z_t) \otimes h_{t-1} + z_t \otimes \hat{h}_t \tag{2.15}$$

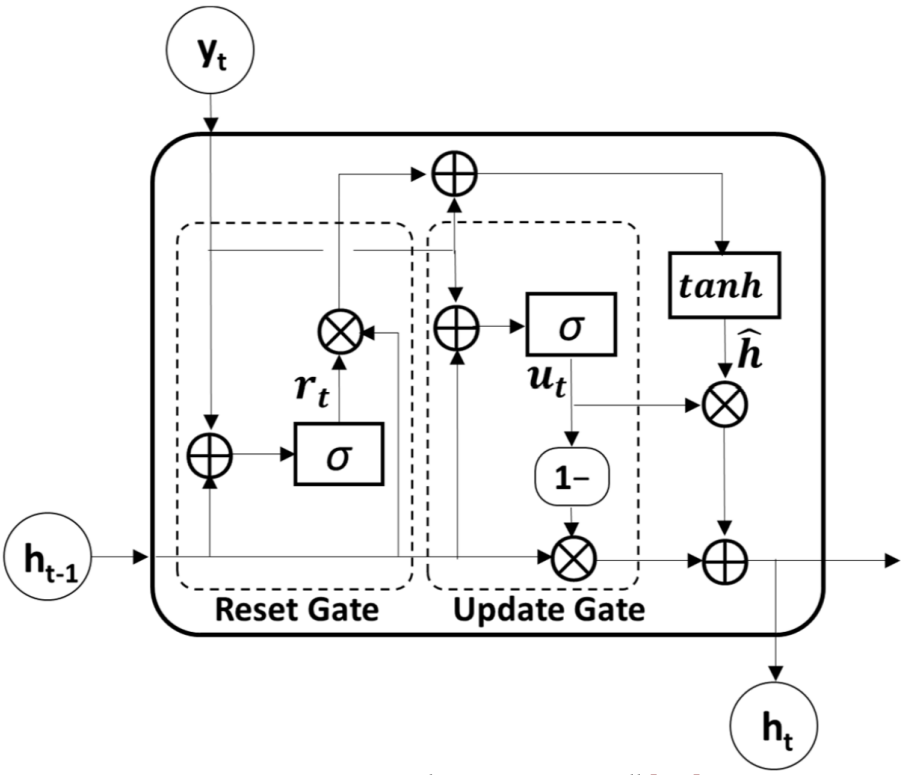

FIGURE 2.9: *Gated Recurrent Unit Cell [154].*

  - x_t: current passed input

- h_t: hidden state at time t
- σ_g and σ_h: sigmoid and hyperbolic tangent functions, respectively
- $\otimes$: element-wise multiplication
- W, U, and b: weight matrices and biases

- **Bidirectional Recurrent Neural Network**
  To incorporate information from both past and future contexts into its predictions, Bidirectional Recurrent Neural Network (BiRNN) analyze input sequences in both forward and backward directions as depicted in Figure 2.10. Numerous studies have investigated the utilization of bidirectional networks for emotion recognition, recognizing the predictive advantage gained from incorporating both future and past inputs for the current prediction [155] [156] [157].

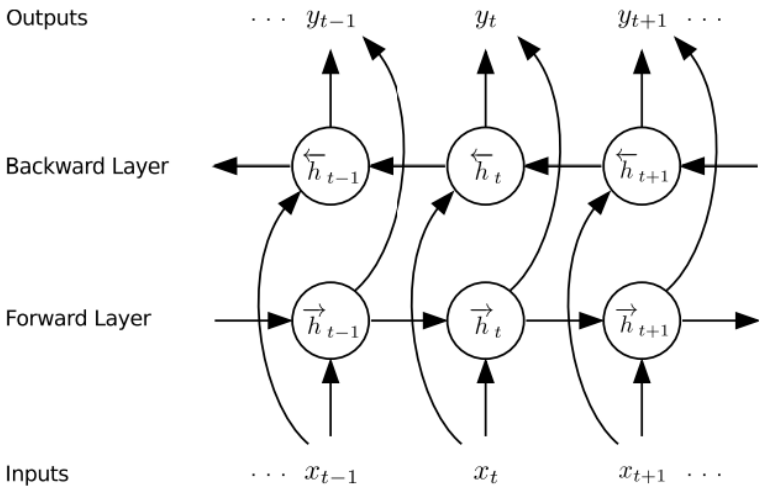

FIGURE 2.10: *Bidirectional RNN [150].*

Numerous studies have employed recurrent models for emotion recognition [4, 158]. These models are often combined with a CNN serving as a feature extractor [50, 159, 160].

2.3.3.3 CONVOLUTIONAL NEURAL NETWORKS

Convolutional Neural Networks CNN have revolutionized the field of computer vision in many domains e.g., image classification, segmentation, object detection, etc. They consist of multiple layers, including convolutional layers, pooling layers, and fully connected layers.

1. Convolutional layers: Contains filters or kernels that have smaller dimensions and width compared to the input volume. The convolutional layer performs dot product calculations between the input image and a kernel by sliding the kernel across the image. This process detects local features at various positions within the 2D input.
2. Pooling layers: are typically positioned between 2 consecutive convolutional layers. The main role of pooling layers is to reduce the number of parameters and computation by downsampling its input. This layer employs a pooling function, often max or average pooling.
3. Fully connected layers: are usually used to integrate the features computed by the previous layers and perform the classification or regression prediction.

CNN have demonstrated success in affective computing across various modalities, including images [161, 162], physiological signals [163, 164], and speech [165, 166].

2.3.3.4 Transformers

Transformer is a ML architecture introduced in 2017 by Vaswani et al. [167]. This technique has revolutionized the field of artificial intelligence in various aspects. It is used in different domains including natural language processing [168] [169], computer vision [170] [171], audio processing [172] [173], and multimodal learning [174][175]. Due to their success, they have been increasingly applied in affective computing, demonstrating superior performance compared to state-of-the-art methods [176] [177] [178]. The transformer has an encoder-decoder structure shown in Figure 2.11 while using the self-attention mechanism.

Attention Mechanism The core innovation of transformers is utilizing self-attention for sequential data. An attention function can be defined as a process that takes a query along with a collection of key-value pairs as input and produces an output. In this context, all these elements - the query, keys, values, and output - are represented as vectors. The output is calculated by performing a weighted summation of the values, where the weight assigned to each value is determined by evaluating a compatibility function that measures the relationship between the query and the corresponding key. This mechanism allows the model to focus on specific pieces of information from the values based on their relevance to the query, enabling it to capture important patterns and relationships in the data. For transformers, typically, a scaled dot-product attention is used where a tuple of inputs, including queries Q and keys K are of dimension d_k and values V of dimension d_v mapped to an output as described in the following equation:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2.16}$$

The inclusion of a scaling factor $\frac{1}{d_k}$ in the dot product in the attention function is essential. This scaling factor is introduced to counteract the issue of dot products becoming excessively large. Without this scaling, the application of the softmax function could yield very small gradients during training, causing a vanishing gradient problem.

Multi-head attention Vaswani et al. [167] employ multi-head attention. Rather than employing a single attention function, the Transformer architecture leverages multi-head attention. In this approach, the initial queries, keys, and values, each with a dimensionality of d_m, are first projected into lower-dimensional spaces: d_k, d_k, and d_v, respectively. These projections are carried out using H distinct sets of learned projections. Subsequently, for each of these newly projected queries, keys, and values, an output is computed through an attention mechanism as specified in the following equation.

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_h)W^O$$
$$where\, head_i = Attention(QW_i{}^Q, KW_i{}^K, VW_i{}^V)$$

Where $W^O, W_i{}^Q, W_i{}^K$, and $W_i{}^V$ are parameter matrices of the projections. The model then consolidates all these individual outputs, concatenates them, and finally projects them.

Positional encoding Since transformers lack the inherent notion of sequential order (unlike recurrent neural networks), positional encodings are added to the input embeddings to provide the model with information about the positions of elements in the input sequence. There exist various options for handling positional encodings, including both learned and fixed approaches.

In [167], they adopt a method involving sine and cosine functions with distinct frequencies. The positional encoding for an element at a given position pos and dimension i is defined as follows:

$$PE_{(pos,2i)} = \sin(pos/10000^{2i/d})$$
$$PE_{(pos,2i+1)} = \cos(pos/10000^{2i/d})$$

Where d is the dimension of the output. In this way, each dimension in the positional encoding is computed using sine and cosine functions with varying frequencies, providing a unique representation for each position in the sequence. The wavelengths create a geometric progression ranging from 2π to 10000.

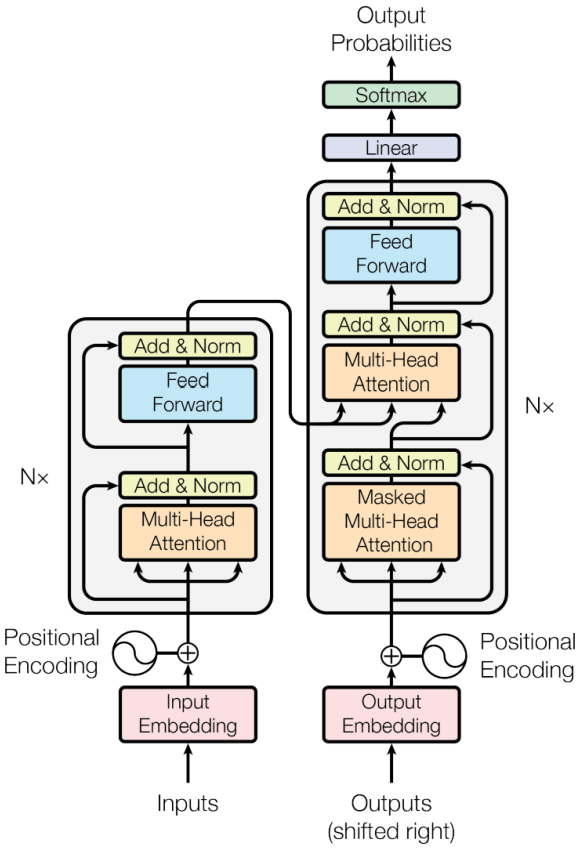

FIGURE 2.11: *The architecture of the Transformer [167]*

2.3.4 CHALLENGES AND LIMITATIONS OF SUPERVISED LEARNING

Similar to other domains, DL methods have enhanced emotion recognition. Nonetheless, DL methods often demand extensive data compared to traditional classifiers. Consequently, when dealing with limited dataset sizes or a restricted number of data points, opting for traditional approaches is recommended [4]. An inherent drawback for both approaches lies in the necessity for labeled data across all classes. This can be challenging if obtaining data for one of the classes is difficult. Consequently, certain works have focused on anomaly detection methods that operate under unsupervised or weakly supervised conditions. We'll delve into the details of this domain in the next section.

2.4 Anomaly detection

2.4.1 Overview of generalized out-of-distribution detection

The Closed-World Assumption in ML supposes that both training and testing data originate from the same distribution [179]. This presupposition suggests that the model has encountered all possible classes and variations during training, and any data encountered during testing belongs to the same set of classes or distributions. Essentially, the model operates within a "closed world," assuming that all necessary information is contained within the training data. However, in real-world scenarios, this assumption may not hold true. In practical terms, this assumption can become problematic. Models may offer misleading confidence values when faced with unseen test samples [180] [181], giving rise to concerns about the reliability of classifiers, particularly in safety-critical applications [182] [183]. Consequently, techniques such as Open-Set Recognition (OSR) and Out-Of-Distribution (OOD) detection have been developed to address scenarios where the closed-world assumption may be invalid. Notably, fields such as anomaly detection, novelty detection, outlier detection, OOD, and OSR have gained prominence. All these domains fall under the Generalized Out of Distribution detection [184]. While these domains tackle similar tasks, the subtle distinctions and connections between them are often overlooked. Our main interest in this thesis is the field of anomaly detection; however, due to the overlap between the terminologies used in each domain, we will offer a concise overview of each domain, outlining their objectives and distinctions.

1. **Anomaly Detection:** Anomaly detection is the identification of rare events, observations, or elements that show significant differences in behavior from normal data. Typically, these points represent areas of interest that require detection.
2. **Novelty Detection:** The term "novel" typically suggests something unknown, new, and intriguing. Similar to anomaly detection, novelty detection seeks to pinpoint test samples that do not fit into any established training category. Consequently, in the community, novelty detection is frequently used interchangeably with anomaly detection [185]. Despite this, the motivation behind each domain differs. Novelty detection regards "novel" test samples as valuable learning resources with a positive learning perspective. [186].
3. **Outlier Detection:** Detecting outliers involves identifying samples within a dataset that notably deviate from others, whether due to changes in covariate or semantic aspects. Unlike previous subdomains of out-of-distribution detection that establish the in-distribution during training, outlier detection defines the "in-distribution" based on the majority of observations. Outliers can arise due to shifts in semantics or covariates within the data. Unlike novelty, outliers are frequently considered as "noise" or "measurement error" to be eliminated in a dataset.
4. **Open Set Recognition:** OSR has been introduced to address the challenge of ML models trained in the closed-world setting, where there is a risk of incorrectly classifying test samples from unknown classes as one of the known categories with high confidence [187]. OSR necessitates that the multi-class classifier simultaneously accomplishes two key objectives:

    - accurately classifying test samples from categories present during training.
    - detecting test samples from categories not belonging to any training category.

5. **Out-of-distribution Detection:** OOD detection is directed at identifying test samples drawn from a distribution dissimilar to the training distribution, with the distribution's definition being context-specific. Usually, the training set encompasses multiple classes, and out-of-distribution detection should not compromise the model's ability to classify

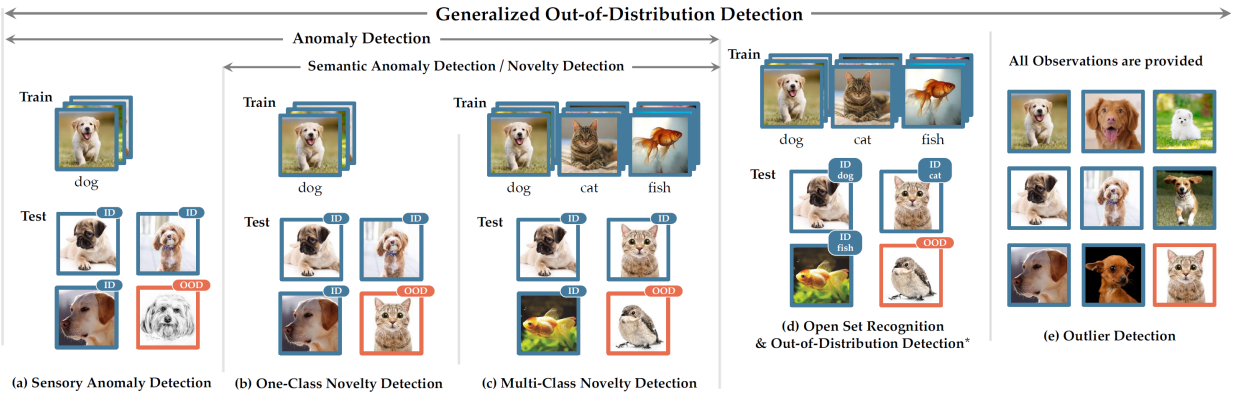

FIGURE 2.12: *Illustration of sub-tasks within a broad out-of-distribution (OOD) detection framework applied to vision-based tasks [184].*

in-distribution samples accurately. Its distinction from anomaly detection and novelty detection arises from operating in a multi-class environment and the essential requirement for in-distribution classification. OSR and OOD detection differ primarily in their benchmark setups, data utilization during training, and the breadth of tasks they cover [184]. OSR splits a multi-class dataset into in-distribution and OOD parts based on classes, while OOD detection uses one dataset as in-distribution and finds distinctOOD datasets. OSR restricts additional data usage for theoretical guarantees [188], contrasting with OOD detection that may leverage extra data for performance. Lastly, OOD detection encompasses a wider range of tasks, including multi-label classification, reflecting a broader solution space compared to OSR's specific focus.

An illustration showcasing various methods for Generalized OOD detection is depicted in Figure 2.12. In (a), a sensory anomaly detection example is presented: in a training set comprising colored dog images, a black and white dog image is deemed an anomaly. In (b) and (c), one-class novelty detection and multi-class novelty detection are demonstrated. In (b), the training set comprises only dog images, while the novelty is represented by a cat image, considered the OOD sample to be detected—an aspect that might be of interest for learning. Contrastingly, in (c), the training set encompasses multiple animal classes, and the novelty emerges as a new class of animals. In (d), an illustration of OSR and OOD detection is similar to multi-class novelty detection, with the model tasked to differentiate between each class in the training set. Last, in (e), outlier detection is applied to all observations without a training/testing scheme, where the majority represents the normal class, and outliers are samples that deviate from the majority.
The following survey offers in-depth definitions, comparisons, and a comprehensive list of methods pertaining to the subtasks of generalized OOD [184].
In our research, our attention is directed toward data samples that deviate from the observed training data but not due to error. Moreover, our primary focus is not on classifying the normal seen classes. Hence, anomaly detection emerges as the primary domain to explore within the field of affective computing.

2.4.2 Anomaly detection overview

Anomalies can be categorized into three distinct types, as outlined in [189]: Point Anomalies, Contextual Anomalies, and Collective Anomalies. Point anomalies are characterized by data points that deviate from the general pattern of the dataset, such as the presence of an image of a boat within a dataset primarily consisting of car images. Contextual anomalies, on the other hand, are data points that are considered irregular within a specific context. For instance, a heart rate of 120 Beats Per Minute (BPM) during exercise may be normal, but if observed

while at rest, it becomes a cause for concern. Collective anomalies refer to a collection of data points that are regarded as abnormal when analyzed as a group, even if they do not stand out as individual anomalies. For example, a substantial and continuous increase in daily spending could be considered irregular, while a single day of increased spending might not raise an alarm. The definitions of Point and Contextual Anomalies align with the rare affective states of interest in our context. Hence, anomaly detection methods can be effectively applied to identify and detect such states. Anomaly detection methods can be applied in three approaches [189]:

1. **Supervised learning:** In supervised learning, the anomaly detection problem is treated as a classification problem. Labeled data (normal and abnormal) are used to train the model to distinguish between the two classes. Nevertheless, this approach requires a significant number of annotated anomaly instances. This can be problematic since anomalies usually rarely occur. Otherwise, the model will encounter difficulties posed by the unbalanced nature of the dataset.
2. **Semi-supervised learning:** The model will be trained using only normal examples. In this case, the normal characteristics of the data can be captured by the model. Thereby, any data of a different nature from the training data will cause the model to behave differently and can then be reported as an anomaly.
3. **Unsupervised learning:** This approach does not require any labeling. However, it relies on the assumption that anomalies rarely occur in the data. Therefore, if the model is trained on all the data, it will learn its normal patterns. Any example deviating from these patterns is considered an anomaly.

Anomaly detection methods fall into four categories: density-based, reconstruction-based, distance-based, and classification-based methods. We explain techniques from each category.

2.4.3 Density-based methods

Density-based techniques aim to characterize the distribution of normal training data, assuming under the estimated density model, abnormal test data typically holds a lower likelihood, while normal data exhibits a higher likelihood. One of the most popular density-based techniques is the Gaussian Mixture Model (GMM).

2.4.3.1 Gaussian mixture model

In the GMM, it is supposed that the data is formed from several Gaussian distributions. For a multivariate (dimension d) Gaussian distribution, the probability density function of an observation x is given by:

$$G(x|\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{|\Sigma|}} \exp^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)} \tag{2.17}$$

Each distribution k is characterized by its mean μ_k, covariance matrix Σ_k, and its weight/mixing coefficient π_k, such that $\sum_{k=1}^{K}\pi_k = 1$. Supposing there are K distributions, the probability density function is defined as follows:

$$p(x) = \sum_{k=1}^{K}\pi_k G(x|\mu_k,\Sigma_k) \tag{2.18}$$

Using Bayes theorem, for each x, the posterior can be written as:

$$w = p(k|x) = \frac{p(x|k)\pi_k}{\sum_{k=1}^{K}\pi_k p(x|k)} \tag{2.19}$$

For each cluster, μ, Σ, and π can be initialized by the K-means algorithm or randomly. Then, the Expectation-Maximization (EM) algorithm can be applied for the parameter estimation. It can be summarized in two steps.
E-step("Expectation"): For each x_i, we calculate $w_i = p(k|x_i)$ its probability to belong to cluster k.
M-step("Maximization"): To maximize the likelihood function, for each cluster k we update μ_k, Σ_k, and π_k.

$$\mu_k = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} k_i}$$

$$\Sigma_k = \frac{\sum_{i=1}^{N} w_i (x_i - \mu_k)(x_i - \mu_k)^T}{\sum_{i=1}^{N} w_i}$$

$$\pi_k = \frac{1}{N} \sum_{i=1}^{N} w_i$$

N is the number of observations. The steps are repeated until the reach of convergence. The GMM method is trained using only normal examples. Next, having estimated the distribution parameters, we calculate the probability for each testing point. If the probability is lower than a certain threshold, data is considered an anomaly.

2.4.4 Distance-based methods

Distance-based techniques identify anomalies by computing the distance between test samples and several saved training data. One of the well-known distance-based methods is Local Outlier Factor Local Outlier Factor (LOF) .

2.4.4.1 Local Outlier Factor

Local Outlier Factor [190] is one of the unsupervised outlier detection algorithms. It is based on the philosophy that anomalies are typically less densely surrounded by their neighbors in a dataset. The method consists of calculating the deviation of the local density of a certain point from its neighbors. Based on the obtained score of deviation, an example is classified as an outlier or not.
Let $K\text{-}distance(A)$ be the distance between A and its $K\text{-}th$ nearest neighbor and $N_K(A)$ the number of K-nearest neighbors of A. We show an example in Figure 2.13. If we suppose $K = 2$, the K-neighbors of point A will be B, C, and D. Here, the value of $K = 2$ but the $||N_2(A)|| = 3$. $||N_K(point)||$ will always be greater than or equal to K.
Let the reachability distance (RD) from point A to point B be the maximum of the distance between A and B and the K-distance of B.

$$RD_K(A, B) = max\{K\text{-}distance(B), d(A, B)\} \tag{2.20}$$

In Figure 2.14, if $K = 2$ the $RD(A, E) = K\text{-}distance(A)$. However, the $RD_K(A, F)$ is the actual distance between A and F since it is bigger than the $K\text{-}distance(A)$.
Then, the Local Reachability Density (LRD) is defined as follows:

$$LRD_K(A) = \frac{1}{\frac{\sum_{B \in N_K(A)} RD_K(A,B)}{|N_K(A)|}} \tag{2.21}$$

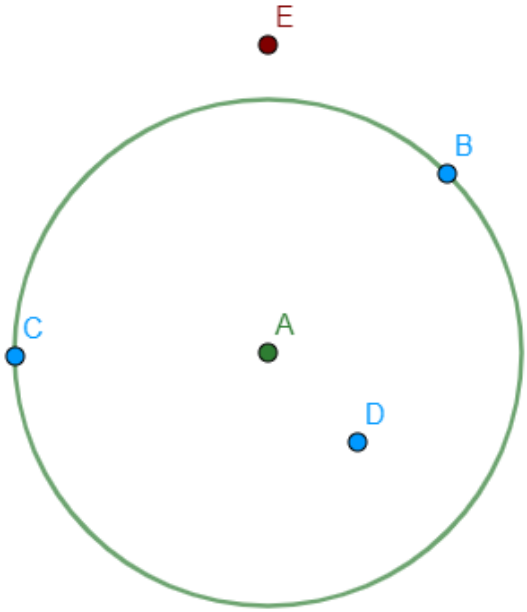

FIGURE 2.13: *Illustration of the K-distance of point A.*

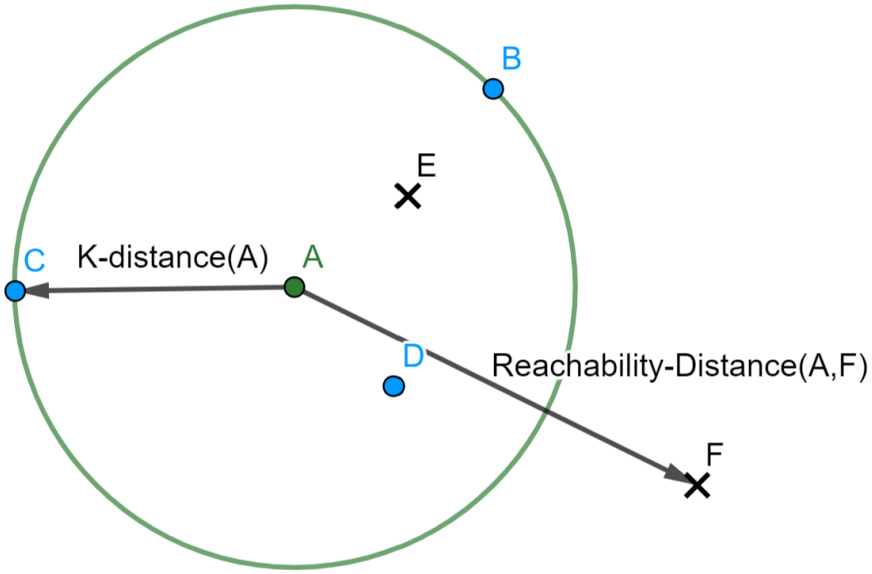

FIGURE 2.14: *Illustration of the RD for point A.*

We calculate the average of the local reachability densities of the neighbors divided by the local reachability density of the point in order to compute the LOF :

$$LOF_k(A) = \frac{\sum_{B \in N_k(A)} \frac{LRD_k(B)}{LRD_k(A)}}{N_k(A)} \tag{2.22}$$

If $LOF(A) > 1$ this indicates that A has a lower density than its neighbors, and it is considered as an outlier. Otherwise, it is considered as normal.

2.4.5 CLASSIFICATION-BASED METHODS

Classification-based methods are based on finding the decision boundary between normal and abnormal samples. Among the well-known techniques in this realm are the One Class Support Vector Machines (OCSVM) , Isolation Forest, and Elliptical Envelope.

2.4.5.1 OCSVM

The fundamental concept of SVM explained in section 2.3.2.1 is versatile and has inspired adaptations to the anomaly detection domain. First, we will explain the mechanism of the traditional SVM , followed by the introduction of the two specialized approaches for anomaly detection: OCSVM by Scholkopf [191] and Class Support Vector Machine by Tax and Duin [192].

- **One Class Support Vector Machine by Scholkopf:** One Class Support Vector Machine OCSVM [191] is a variation of standard SVM ; it was introduced as a novelty detection solution. It groups all the normal data samples as one class and the origin as the second class. The goal is to separate those classes, thereby maximizing the distance between the separating hyperplane and the origin. Therefore, the problem can be reformulated as the following:

$$\min_{w,\rho,\xi_i} \frac{1}{2}||w||^2 + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i - \rho \tag{2.23}$$

subject to:

$$(w.\phi(x_i) + b) \geq \rho - \xi_i \quad \forall i \tag{2.24}$$
$$\xi_i \geq 0 \quad \forall i \tag{2.25}$$

Where ρ is the offset,ν is a value between 0 and 1 that represents the upper bound on the fraction of outliers and the lower bound on the fraction of support vectors. The decision values are obtained using:

$$f(x) = w.\phi(x_i) - \rho \tag{2.26}$$

- **Class Support Vector Machine by Tax and Duin:** Another approach is introduced in [192]; the authors search for an optimal hypersphere that includes all of the data. The volume of the hypersphere of center a and radius R should be minimized, which leads to the following optimization problem:

$$\min_{R,a,\xi_i}(R^2 + C\sum_{i=1}^{n}\xi_i) \tag{2.27}$$

subject to:

$$||x_i - a||^2 \leq R^2 + \xi_i \quad \forall i \tag{2.28}$$
$$\xi_i \geq 0 \quad \forall i \tag{2.29}$$

Similar to the former approaches, slack variables ξ_i are introduced to allow soft thresholding, and C is a constant to penalize the errors.

2.4.5.2 Isolation forest

Isolation forest [193] is an anomaly detection method that does not exploit any density or distance measure. It considers a point as an anomaly if it is easy to separate it from the rest of the data. It operates by modeling normal data to isolate anomalies that are both infrequent and distinctive in the feature space. The algorithm achieves this by constructing a random forest, where decision trees are grown randomly. At each node, features are selected randomly, and a threshold value is chosen to split the dataset. The process continues until all instances are

effectively isolated from each other. The anomaly score of a point x in a dataset of N examples is defined as:

$$AnomalyScore(x) = 2^{-\frac{E\{h(x)\}}{c(N)}} \tag{2.30}$$

where $E\{h(x)\}$ is the expected path length for isolating a point and is calculated as the mean of the path lengths needed to isolate the point across all generated trees. $C(N)$ represents the average value of $h(x)$ for a dataset of size N and can be computed as:

$$c(N) = 2H(N-1) - \frac{2(N-1)}{N} \tag{2.31}$$

where H_N denotes the Nth harmonic number, which can be approximated by $ln(N) + \gamma$ with $\gamma \approx 0.577$ (Euler–Mascheroni's constant). The anomaly threshold is set between 0 and 1. A higher anomaly score, closer to 1, indicates a higher likelihood of the point being an anomaly, while a score closer to 0 suggests a higher likelihood of it being a normal data point.

2.4.5.3 Elliptical envelope

The Elliptical Envelope method adopts a Gaussian distribution model for the data. Its primary goal is to determine a boundary ellipse that encompasses the majority of data points, with any points falling outside this ellipse considered as anomalies or outliers. To achieve this, the routine utilizes the FAST-Minimum Covariance Determinate method to determine the shape and size of the ellipse [194]. This algorithm iteratively calculates the Mahalanobis distance defined by

$$d_{Mahalanobis} = \sqrt{(x_i - \mu)^T C^{-1} (x_i - \mu)} \tag{2.32}$$

which measures how many standard deviations a data point deviates from the mean. The algorithm proceeds by selecting subsamples from the original dataset and calculates the μ and covariance matrix C. Subsequently, the Mahalanobis distance is computed for each data point x_i. The algorithm then selects subsamples corresponding to small values of the Mahalanobis distance. It recalculates the mean, covariance matrix, and Mahalanobis distance values. This iterative process continues until the determinant of the covariance matrix reaches convergence. Among all the subsamples, the covariance matrix with the smallest determinant is identified. This covariance matrix is then utilized to define an ellipse that encompasses a portion of the original data. Data points residing within the surface of this ellipse are categorized as 'inliers' while those situated outside of the ellipse are designated as 'outliers' or anomalous data points. These outliers can subsequently be considered for further analysis or potential removal from the dataset.

2.4.6 Reconstruction-based methods

Reconstruction-based methods rely on the concept that the encoder-decoder framework, trained on normal data, typically generates distinct outputs for normal and abnormal samples. Leveraging the variation in model performance can serve as an indicator for anomaly detection. Various DL methods have been employed through three distinct approaches. The first approach employs a DL model as a feature extractor, which is responsible for transforming high-dimensional data into a lower-dimensional representation. Subsequently, a statistical anomaly detection method is applied to this lower-dimensional data [195]. The second approach adheres to the concept of anomaly detection through failure. Here, a deep model is trained on "normal" data with the goal of accomplishing a specific task. If the model fails to perform the task effectively for one data point[196] [197], then the point is considered to be anomalous. The third

approach combines elements from both of the previously mentioned approaches. In this scenario, two models are trained in tandem. The feature extractor encodes the data into a new latent space, and subsequently, the anomaly detector is applied to the latent features, facilitating a comprehensive approach to anomaly detection. There are various DL architectures popularly used in anomaly detection [198] [199] [200].

2.4.6.1 AEs

AE are a specific type of MLP.
In the case of AE, the output is identical to the input. In the case of a basic AE with one hidden layer, an input example $x \in \mathbb{R}^d$ will pass through the hidden layer $h(x) \in \mathbb{R}^p$ where:

$$h(x) = g(W_1 x + b_1) \tag{2.33}$$

Where $g(z)$ is a non-linear activation function. Then, the model will decode the hidden representation $h(x)$ to produce a reconstruction of the input $\hat{x} \in \mathbb{R}^d$:

$$\hat{x} = g(W_2 h(x) + b_2) \tag{2.34}$$

The training of the AE consists in finding the parameters W_1, W_2, b_1, b_2 that minimize the Reconstruction Error (RE), which is described in the following loss function:

$$RE = \mathbb{L}(W_1, W_2, b_1, b_2) = \sum_{x \in \mathbb{R}^d} ||x - \hat{x}||^2 \tag{2.35}$$

After training the AE successfully on “normal” data, we provide testing data. If, for a particular data point, the RE is large (above a pre-defined threshold), then the AE has failed to reconstruct it correctly. This point is classified as anomalous data. If it’s small, then the data is classified as normal.

2.4.6.2 Variational auto-encoders

In 2014, Kingma et al. [201] proposed the VAE as a solution to the following intractable problem: Given a dataset X of N i.i.d observations that is generated by a hidden variable z where we are interested in a computation of $p(z|x)$, where:

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)} \tag{2.36}$$

As a solution, the Auto-encoding Variational Bayes proposes approximating $p(z|x)$ by $q(z|x)$. The VAE structure, illustrated in Figure 2.15, is composed of two parts Encoder "Recognition model" and the decoder "Generative model". Dissimilar to AEs, VAE encodes the inputs into a mean vector and a standard deviation vector of the latent space distribution instead of a fixed variable vector. Next, a random sampling method is applied to obtain a sampled latent representation z. Finally, the sampled z is fed into the decoder part of the VAE that aims to reconstruct the initial input.
In VAE, the prior of the latent space $p_\theta(z)$ is assumed to be a normal distribution, and $p_\theta(x|z)$ is the marginal likelihood following the distribution $\sim N(\mu(z); \sigma(z)I)$. Therefore, $q_\phi(z|x)$ an approximate posterior, can be assumed following the distribution $\sim N(\mu(x); \sigma(x)I)$. The loss function of the VAE can be reformulated as the following:

$$L(\theta, \phi; x, z) = E_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\phi(z|x)||p_\theta(z)) \tag{2.37}$$

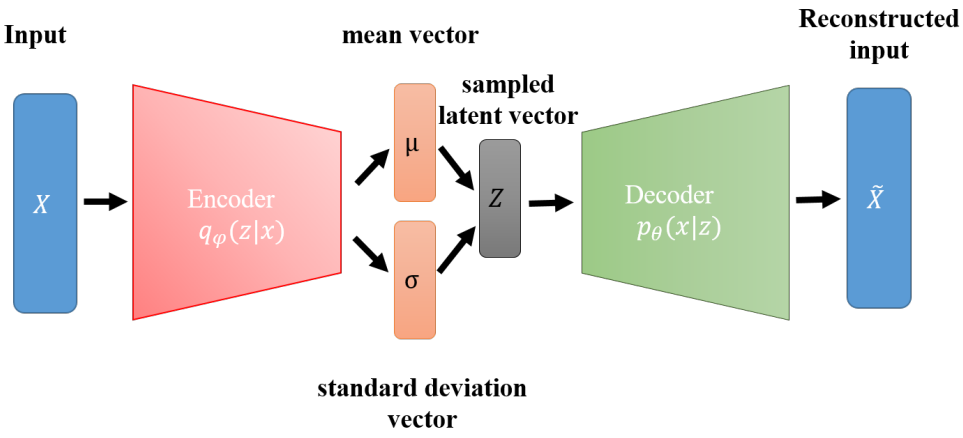

FIGURE 2.15: *Illustration of a Variational Auto-Encoder.*

where the first term $E_{q_\phi(z|x)}[\log p_\theta(x|z)]$ represents the RE between the original input and the decoded output from the sampled latent attributes, whereas the second term represents the Kullback-Leibler divergence between the variational posterior approximate $q_\phi(z|x)$ and the prior of the latent variables $p_\theta(z)$. In essence, this term forces the model to approximate a latent distribution that is as close as possible to a normal distribution. However, to optimize the loss function, we must be able to run back-propagation through the model, which is problematic due to the presence of the sampling node. To resolve this issue, a "Reparameterization trick" is introduced. Since z is a random variable from a Gaussian distribution, it can be expressed as $z = \mu + \sigma.\epsilon$ where $\epsilon \sim N(0,1)$. This approach facilitates the integration of the necessary random component for sampling from the latent distribution while maintaining a series of differentiable operations crucial for backpropagation. VAE can be used for data augmentation. The estimate of the θ parameters can be used to generate data that is similar to the original dataset. It can also be used for data representation by using $q_\phi(z|x)$. Furthermore, it can be exploited in any case, whereas prior over x is required for inference tasks, e.g., denoising. Although Variational Auto-Encoders and traditional AE have a similar architecture, they have a fundamental difference in their mechanisms. As mentioned before, AE encodes each input sample into a fixed vector, whereas VAE encodes it into a latent distribution. Transforming the inputs into a distribution of variables instead of fixed values of variables allows the model to learn a smooth representation of the data. It will not only be able to discriminate between classes but also distribute the data evenly in the latent space. This provides a wider coverage of the possible values of data.
Many studies explored the use of VAE for anomaly detection [202] [203] [204]. In [205], the authors developed a VAE model to capture resilient local features across short windows followed by an LSTM module for anomaly detection in time series data. An alternative approach is presented in [206], employing a β-VAE for anomaly detection. Notably, the calculation of the anomaly score in this approach involves a combination of the input's reconstruction and gradient loss. Following the computation of anomaly scores, a decision rule is applied, comparing against a predefined threshold to ascertain whether a given sample qualifies as an anomaly.

2.4.6.3 GENERATIVE ADVERSARIAL NETWORK

Generative Adversarial Network (GAN) is a type of generative artificial NN introduced by Goodfellow et al. in 2014 [207]. GAN has also been utilized for anomaly detection in several domains, including medical diagnosis, network security, finance fraud, infrastructure inspection, and industrial defect detection [208]. Given the success of GAN , many variations have been developed for anomaly detection in image data [209][210], time series data [211] [212] [213], and

other fields[214][215]. However, most of those variations are based on GANomaly, AnoGAN, and EGBAD [216].

Vanilla GAN The concept of GAN lies in training two artificial NN adversarially: the discriminator D and the generator G. The goal of the generator's training is to model the underlying distribution of the input data and generate data from random noise that is as similar as possible to the training or 'real' data. Whereas the discriminator plays the role of a binary classifier that's able to distinguish between real training data and fake generated from the Generator data. When the discriminator evaluates real data, it assigns a value of "1" while it assigns "0" for generated data. The loss function guiding the discriminator's training can be expressed as shown in the following formula:

$$maxV(D,G) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))] \tag{2.38}$$

where x is the data point from the training dataset, and z is a random noise vector sampled from a noise distribution $p_z(z)$. In contrast, the generator's objective is to produce outputs that the discriminator classifies as real and assigns a value of "1." Consequently, the corresponding loss function for the generator takes the form shown in the formula:

$$minV(D,G) = \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))] \tag{2.39}$$

The combination of both formulas yields:

$$min_G max_D V(D,G) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))] \tag{2.40}$$

The generator continuously adjusts its parameters based on the feedback from the discriminator. This adversarial loop continues until the GAN reaches a state of equilibrium where the generator generates data that closely matches the distribution of real data. An example is shown in Figure 2.16, where a GAN is trained on a dataset of digits.

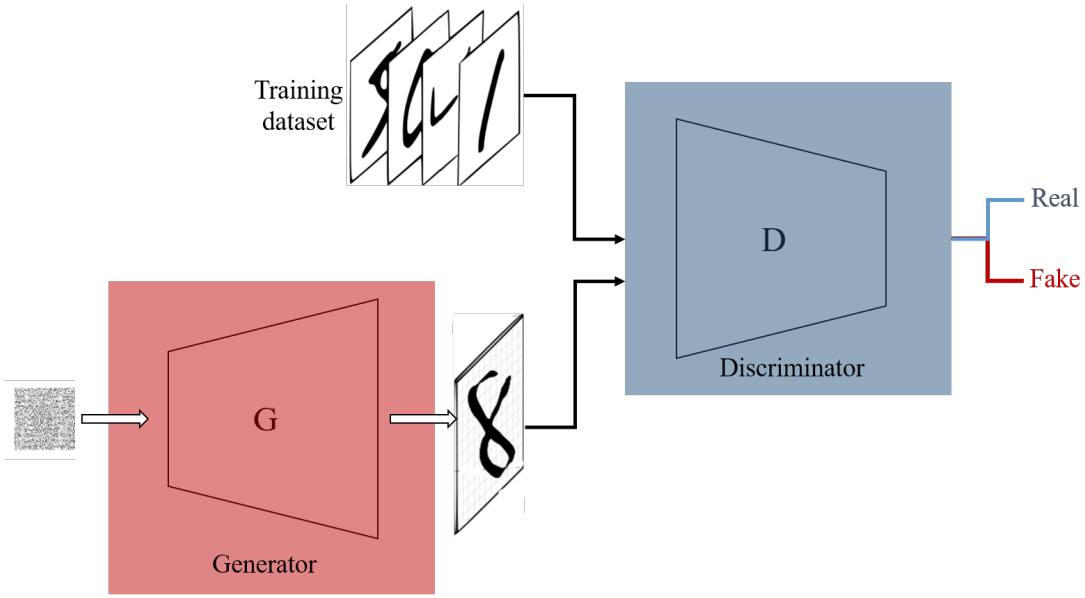

FIGURE 2.16: *Example of a Generative Adversarial Networks(GAN)*

AnoGAN AnoGAN [197] has a generator and a discriminator like a traditional GAN . AnoGAN learns to create a mapping from the latent space representation (z) to a sample similar to training data. Subsequently, it leverages this acquired representation to map novel, unseen samples

back to the latent space using backpropagation. For testing, the difference or residual between the original test data point and the reconstructed data point is used as an anomaly score. This architecture was first proposed to detect abnormal samples of optical coherence tomography images of the retina. However, the approach suffers from several disadvantages, including the difficulty of anomaly score interpretation and bad testing time performance [217]. Efficient GAN-Based Anomaly Detection (EGBAD)[218] leverages the Bidirectional Generative Adversarial Network (BiGAN) architecture to enhance anomaly detection and overcome limitations seen in the AnoGAN approach. BiGAN [219] extends the traditional GAN framework by incorporating an encoder E alongside the generator G. The encoder learns the inverse mapping of examples in data space to latent variable space. Unlike standard GAN that operate solely in data space (comparing data x to generated data G(z)), BiGAN's discriminator D evaluates both data and latent space, specifically comparing tuples of (data $x, E(x)$) to (generated data $G(z), z$). The data space is flattened and concatenated with the latent space vector before being fed into the discriminator D. In this context, the latent representation z can be seen as a "label" for the corresponding data x, obtained without the need for explicit supervision. BiGAN's training objective is defined as a minimax objective, optimizing the interplay between the generator and discriminator. One of the main advantages of EGBAD is that it doesn't require backpropagation to calculate the anomaly score, making it more computationally efficient. This approach outperformed the competing architectures on MNIST [220] and KDD19 [221].

GANomaly In 2018, Ackay et al. [222] introduced GANomaly, an approach inspired by AnoGAN [197] and EGBAD[218]. In Figure 2.17, we present the architecture of GANomaly. The difference between GANomaly and other GAN architectures is the generator consists of an AE G composed of encoder G_E, decoder G_D, and an encoder E. First, the AE learns to map the input x to a compressed latent representation z, where $z = G_E(x)$. Then, the decoder part of the AE reconstructs the initial input using z. The generated input $\hat{x} = G_D(z)$ is then passed to E where it is downscaled to $\hat{z} = E(\hat{x})$. For testing data x, the anomaly score is defined as:

$$A(x) = ||G_E(x) - E(G(x))||_1 \tag{2.41}$$

The authors conducted experiments in different scenarios and found that the GANomaly [222] exhibited superior performance over AnoGAN [197] and EGBAD [218].

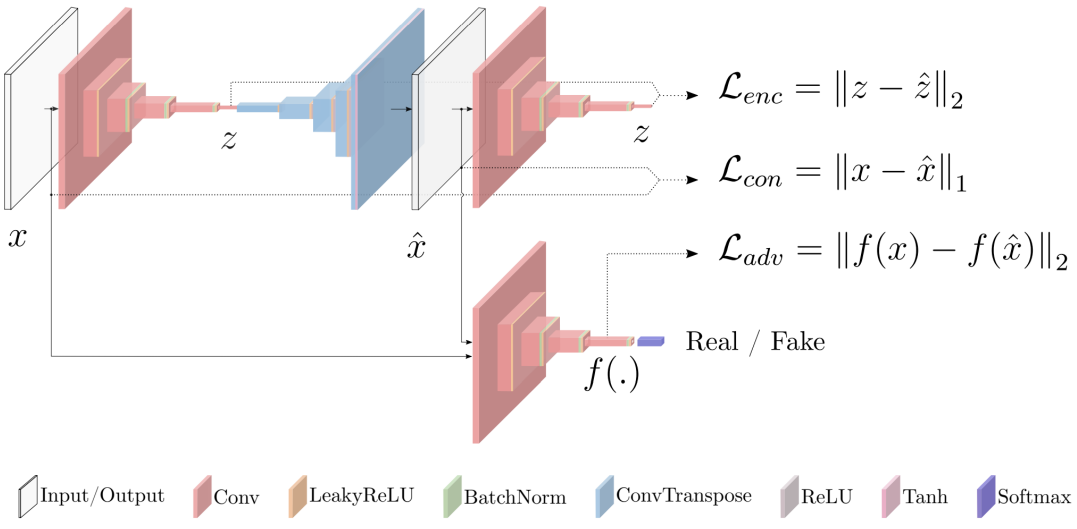

FIGURE 2.17: *Architecture of GANomaly [222]*

2.4.6.4 Transformers-based anomaly detection

Given their ability to process large sequences of data, transformers have also been utilized for anomaly detection. In [223], the authors proposed a time-series anomaly detection approach utilizing a Transformer architecture to effectively capture dynamic patterns in sequential data through self-attention mechanisms. The model comprises multiple Transformer encoder layers in the encoder and a 1D convolutional layer in the decoder. The anomaly detection is based on the discrepancy between predicted and actual values at each timestamp, with data exceeding a predefined threshold classified as anomalies. Their results show that the transformer approach outperforms traditional Long LSTM and CNN approaches across several time-series datasets benchmark [223]. Huang et al. [224] introduced a novel log-based model that leverages a hierarchical transformer structure for anomaly detection in system log data. Comparative experiments demonstrate the superior performance of the transformer model, particularly in modeling log sequences, as the proposed approach is evaluated on three log datasets, surpassing existing anomaly detection methods. It was also implemented for anomaly detection in images [225], acoustics [226], video[227], and wearable signals [228][229][230].

2.4.7 Comparison

Anomaly detection is a critical task in various domains, and both ML and DL methods have been employed to tackle it. Each approach offers distinct advantages and has its own set of limitations. Traditional methods prove efficient with small datasets and are less computationally intensive. However, classical approaches may struggle with high-dimensional, intricate data and often require manual feature engineering. On the other hand, DL methods excel in handling complex, high-dimensional data and automatically learn relevant features, reducing the need for feature extraction. Nevertheless, these models are data-hungry and demand substantial computational resources, potentially rendering them impractical in situations where large datasets are unavailable, or resource constraints apply [231]. Moreover, due to their complex architectures, they suffer from training problems. For example, GAN suffer from mode collapse, gradient vanishing, stopping problem, and instability [232]. Skavara et al. [233] highlight that the choice of the optimal anomaly detection model depends on various contextual factors, including data type and hyperparameter tuning strategies, emphasizing the need for a thoughtful selection process tailored to specific use cases.

2.4.8 Hyperparameter tuning for anomaly detection models

Tuning hyperparameters in anomaly detection presents a notable challenge, particularly due to data scarcity and the preference to minimize reliance on labeled data in model training. Despite this, models inherently possess hyperparameters that require optimization. Broadly, two approaches exist in this domain:

1. Opting for default parameters recommended in the literature [234] [235][236]. [237] [238] [239]
2. Selecting parameters that yield the best performance on the testing dataset [240] [241] [242]. The first approach tends to provide a pessimistic estimate of the model's performance, while the second approach often leads to unrealistic performance maximization.

In our work, we choose the first approach due to its feasibility of being employed in the real world for data-scarce applications.

2.4.9 Performance evaluation metrics

Irrespective of the anomaly detection method chosen, assessing its performance is crucial to determine its suitability for a particular application. Therefore, this section introduces widely recognized evaluation metrics for anomaly detection, which we will use throughout our experiments:

2.4.9.1 Receiver operating characteristic area under curve

In the context of anomaly detection, algorithms typically produce real-valued anomaly scores as output. The determination of True Positive (TP) number of anomalies correctly predicted as positives, True Negative (TN) number of normal examples classified as negatives, False Positive (FP) number of normal samples misclassified as positives, and False Negative (FN) number of anomalies misclassified as negatives relies on the selection of a threshold. Usually, there exists a trade-off between the quantity of TP and FN generated by an algorithm. The Receiver Operating Characteristic Area Under Curve (ROC AUC) is a graphical representation that illustrates this trade-off. The ROC AUC displays, for various threshold selections, the True Positive Rate (TPR) and the False Positive Rate (FPR). TPR is calculated as:

$$TPR = \frac{TP}{TP + FN} \tag{2.42}$$

representing the ratio of TP to the sum of TP and FN (i.e. number of positive examples). On the other hand, FPR is computed as:

$$FPR = \frac{FP}{FP + TN} \tag{2.43}$$

A ROC AUC of 1 indicates a perfect classifier that perfectly separates positive and negative instances. A ROC AUC of 0.5 suggests a classifier with no discriminatory power, essentially performing as well as random guessing. A ROC AUC between 0.5 and 1 indicates varying degrees of classifier performance, with a higher value indicating a better performance.

2.4.9.2 Precision-Recall Area Under Curve

The Precision-Recall Area Under Curve (PR AUC) curve shows, for all possible threshold choices, the Precision vs. Recall. The precision is defined as follows:

$$Precision = \frac{TP}{TP + FP} \tag{2.44}$$

The recall is defined as follows:

$$Recall = \frac{TP}{TP + FN} \tag{2.45}$$

PRAUC is particularly useful when dealing with class imbalance [243].

2.4.9.3 F-score

The F-score or F1 score is a metric used in classification tasks, especially when dealing with imbalanced classes. It's the harmonic mean of precision and recall and is calculated using the following formula:

$$F_1 = 2\frac{Precision * Recall}{Precision + Recall} \tag{2.46}$$

2.4.9.4 Balanced accuracy

Balanced accuracy finds application in both binary and multi-class classification scenarios. It represents the average between Recall and specificity, offering value, particularly in situations involving imbalanced data, where one target class significantly outweighs the other in representation.

$$Balancedaccuracy = \frac{Recall + Specify}{2} \tag{2.47}$$

where Specifity is the True Negative Rate (TNR) and can be calculated as follows:

$$Specify = \frac{TN}{TN + FP} \tag{2.48}$$

2.4.10 Challenges in anomaly detection

Anomaly detection exhibits versatility across numerous domains and applications; nonetheless, it encounters both universal challenges and application-specific challenges. In the subsequent discussion, we highlight some of the most critical challenges in this domain.

1. **What is Normal?** One of the main difficulties lies in precisely defining the parameters of normal behavior. Setting the boundaries for a normal region that includes every possible normal instance is challenging.
2. **Dynamic Normalcy** Normal behavior exhibits a dynamic nature, constantly evolving and adapting. Consequently, maintaining an accurate representation of normality over time becomes a challenging endeavor.
3. **Contextual Anomaly:** Anomaly, as a concept, is highly context-dependent. What may constitute an anomaly in one domain, such as the medical field, might be deemed entirely normal in another domain, like financial markets. This contextual relativity presents obstacles when transferring anomaly detection techniques across different domains or subjects.
4. **Lack of labeled data:** Finding labeled data as normal/abnormal is rare. Moreover, anomalies, by their infrequent nature, restrict the possibility of rigorously evaluating the robustness of anomaly detection systems.
5. **Explainability:** Most anomaly detection methods do not provide an explanation of why data is classified as abnormal. This could be problematic in critical domains, e.g., medicine.
6. **Distinguishing Noise from Anomalies** Discriminating between random noise and genuine anomalies can be a challenging task. This differentiation demands the application of advanced techniques and robust algorithms.

In summary, the domain of anomaly detection underscores the need for a comprehensive understanding of these challenges. Effectively addressing these complexities is paramount for the development of accurate and dependable anomaly detection systems across diverse domains. Many of these challenges find potential solutions through extensive data collection. For instance, gathering extensive and diverse data enables the establishment of normal behavior or patterns, contingent on the scale and diversity of the dataset. Continual data collection addresses the challenge of evolving normal behavior by facilitating model updates. Additionally, a large amount of normal and anomaly examples increases the robustness of the model in distinguishing between noise and genuine anomalies. In our research, we concentrate on addressing the following challenges: "lack of labeled data", "contextual anomaly", and "explainability." Our emphasis on unsupervised models in Chapter 3 addresses the challenge of unlabeled data, focusing solely on normal behavior. Relapse behavior varies among patients, illustrating a contextual

anomaly. Exploring the advantages of unsupervised methods allows us to construct personalized models with less data. Moreover, we delve into explainability using AE, aiming to determine the most important features influencing the model's decisions in two applications: visual distraction detection and psychotic relapse prediction.

2.5 Anomaly detection in affective computing

Given the persistent challenge of acquiring sufficient data in the domain of affective computing, particularly for relatively rare and ethically sensitive behaviors, e.g., aggression, researchers have recently begun to explore the use of anomaly detection techniques. These methods allow researchers to identify and analyze rare behavior within limited datasets by treating them as anomalies or deviations from normal behavior patterns. In [244], anomaly detection is used to detect inattentive pupils from surveillance video. Baliniskite et al. [245] employed an anomaly detector (Isolation Forest) along with an emotion predictor to discern anomalies in behavior, thereby reducing the likelihood of flagging non-threatening abnormal behaviors. Ding et al. employed GMM to detect abnormal emotional states in drivers [246]. In [247], the authors introduced the EMO&LY dataset, designed specifically to provoke abnormal emotional reactions in participants by altering the established protocol. They employed OCSVM and GMM to detect these abnormal emotions on audio and video-extracted features. Their study highlighted the efficacy of unsupervised classifiers in effectively discerning anomalous samples from normal ones. Hu et al. used an AE with a recurrent graph attention network to detect social behavior anomalies in highway drivers [248]. Pillai et al. have showcased the potential of mobile sensing for the detection of rare life events in real-world scenarios [249]. In their research, the authors present a multitask learning architecture composed of two core elements: first, an encoder-decoder framework based on LSTM for the computation of an anomaly score, and second, a sequence predictor aimed at providing context to the anomaly score by deducing transitions in workplace performance [249]. Zhu et al. [250] developed an unsupervised method for abnormal emotion detection by merging Gaussian Mixture VAE with CNN . Their approach was compared against state-of-the-art methods, demonstrating superior performance. Ye et al. proposed an anomaly detection model for autonomous behavior public transport, addressing the scarcity of relevant datasets by leveraging similar datasets. Their model, based on a VAE trained on normal data, demonstrated efficacy in detecting anomalies akin to scenarios like someone falling, attacking, and fighting [251]. Across the studies, the definition of abnormal emotions or behavior tends to be specific to the dataset. Since there is no clear standard for abnormal state or emotion, the authors establish distinct criteria for abnormal state based on the characteristics of each dataset.

At the start of this PhD, there was limited research in applying anomaly detection methods to affective computing. Previous efforts were constrained to employing anomaly detection solely as classifiers. This involved training on a predominant class to establish it as the norm and identify deviations such as inattention, negative emotions, rare events, etc., as illustrated in earlier examples. Our work further explored this application in visual distraction detection and psychotic relapse prediction. Moreover, we utilized anomaly detection methods to tackle crucial challenges in affective computing, including combining information, personalizing models, selecting key features, and improving comprehensibility.

2.6 Conclusion

This chapter introduced affective computing and elucidated the challenges it encounters, particularly concerning the complexities in acquiring relevant data. Conventional supervised learning approaches often face limitations in these scenarios. Hence, we delved into the do-

main of anomaly detection, citing pertinent works that underscore its significance in addressing the challenges encountered in affective computing. Subsequent chapters will offer anomaly detection-based solutions for a range of challenges, including rare mental states, model personalization, explainability, feature selection, and information fusion. These solutions will be applied across diverse applications, such as driver behavior monitoring, psychotic relapse prediction, and emotion recognition in stress-inducing environments. Our exploration will encompass various modalities, including physiological signals, eye-tracking, speech, video, and text.

CHAPTER 3

RARE MENTAL STATES DETECTION

This chapter highlights our contributions to utilizing anomaly detection methods for the detection of rare mental states. Our focus is on two specific applications: visual distraction and psychotic relapse detection. Each application is presented in its own dedicated section.

CHAPTER CONTENTS

3.1 Introduction

In Chapter 2, we highlighted numerous challenges within the realm of affective computing, notably emphasizing the issue of data scarcity. This challenge becomes more pronounced when attempting to capture data in authentic real-life settings, especially when targeting rare occurrences. One strategy to tackle this challenge involves simulating these infrequent events within a controlled laboratory environment. For instance, many emotion datasets are constructed through acted or induced scenarios. For the IEMOCAP dataset, ten trained actors illustrate emotions through scripted and non-scripted interactions [108]. However, the efficacy of models trained on simulated data often falters when deployed in genuine real-world settings [252] [253]. Hence, the acquisition of authentic, real-world data becomes crucial. Nevertheless, collecting such data is impeded by obstacles such as data labeling, and data diversity which are discussed in Chapter 2. An illustrative case is driver behavior monitoring, where the sheer volume of data and a multitude of signals make manual labeling an impractical task [96].

In affective computing, numerous applications aim to detect rare yet critical situations (like aggression) or undesirable mental states (such as depression), which are fortunately inherently infrequent. Conversely, acquiring normal data is comparably easier. For instance, a train company might readily have videos depicting passengers' regular behaviors but scarcely capture instances of passenger aggression. This imbalance in data distribution frequently causes conventional and widely used supervised learning models to underperform. One proposed strategy to address the detection of these rare mental states or undesired events involves reframing the classification problem as an anomaly detection problem. This reframing aims to develop robust models better equipped to handle real-world settings. As illustrated in Section 2.5, while a few works have begun exploring this concept, it remains relatively novel and is at its inception stage. The initial results are promising, urging further investigation and validation of this approach before implementing it in real-life applications. In this chapter, we investigate this concept within two preventive applications ´Abnormal driving behavior detection´ and ´Psychotic relapse prediction´. Each application underscores the key specifications of our approach: First, the imperative need to detect rare and critical events that must be averted. Second, the rarity of these events implies that continuous data collection in real-life scenarios predominantly yields normal data instances.

One of the critical situations to be detected is abnormal driving behavior that might endanger the driver´s life and other people on the road. Data collected by the World Health Organization shows that the principal cause of death for children and adults between ages 5 and 29 was road traffic injuries. Furthermore, traffic crashes result in approximately 1.3 million deaths [254]. Monitoring driving behavior and detecting anomalies can significantly improve road safety and coordination, empowering drivers to make informed decisions. However, gathering data on dangerous or anomalous behavior exposes drivers to risks. Hence, employing an anomaly detection approach becomes pertinent for this application. To assess its relevance, we initially employ data acquired in a simulated environment for the specific task of visual distraction detection. If proven effective, this method could potentially be adopted for this task and for a broader spectrum of abnormal behaviors in real-world settings.

Another significant application involves predicting relapses in patients with psychotic disorders. Given the infrequent occurrence of relapses and the challenges and expenses associated with labeling such data, employing anomaly detection becomes promising for predicting these rare mental states. The initial findings outlined in [99] have encouraged further exploration of this task. This exploration was further stimulated by the e-Prevention challenge [255] introduced in ICASSP´23, which focused on unsupervised learning and specifically emphasized anomaly detection.

The contributions presented in this chapter can be summarized as follows:

- Anomaly detection remains a promising yet underutilized category of ML methods within affective computing. Our research stands out as one of the few efforts dedicated to leveraging anomaly detection methods for the identification of rare mental states.
- Our exploration spans two domains: visual distraction detection and psychotic relapse prediction. Notably, our research is the first to use anomaly detection techniques for visual distraction, specifically with signals from non-invasive sensors. As for the application for psychotic relapse prediction, we further explored the use of anomaly detection for psychotic relapse prediction, building upon prior prevention works. We emphasized classical methods due to their efficacy with limited data, particularly relevant in a data-scarce domain.
- We also investigated personalized models tailored for each patient for psychotic relapse prediction, considering the implications and advantages brought about by the utilization of anomaly detection methods. Also, we conducted a comprehensive study encompassing diverse features and models for better understanding and applicability.

The remainder of this chapter is structured as follows: We begin this chapter by providing a SOTA of the current advancements in both visual distraction detection and psychotic relapse prediction in Section 3.2. Then, The chapter is organized into two parts dedicated each to one application. The first part focuses on our driver monitoring application as a proof of concept for visual distraction detection. In section 3.3.1, we elaborate on the dataset we used for visual distraction detection. It's followed by a detailed description of our experimental setup in section 3.3.2. The subsequent section, Section 4.4.3, outlines our findings. This encompasses comparisons between various anomaly detection methods and classical supervised techniques across different data balance scenarios, along with an assessment of whether anomaly scores effectively estimate the degree of abnormality. Finally, Section 3.3.4 encapsulates the conclusion drawn from this section. In 3.4, we go into the details of our study on psychotic relapse prediction. In Section 3.4, we delve into our study on psychotic relapse prediction, introducing the dataset utilized, data pre-processing steps, and feature extraction detailed in Section 3.4.1. Subsequently, we expand on our proposed methodologies in Section 3.4.2, where we introduce and explore two distinct schemes: the global scheme and the personalized scheme. The outcomes and findings derived from our experiments are presented in Section 3.4.3. Finally, we conclude this chapter in Section 3.5.

3.2 State of the art for rare mental state detection

In Chapter 2 section 2.5, we discussed the application of anomaly detection methods in various affective computing scenarios to identify rare states that are challenging to collect. In this section, we delve deeper into the specific research approaches we are interested in for detecting rare states related to abnormal driving behavior and predicting psychotic relapse.

3.2.1 Abnormal driving behavior

- Supervised learning approach: Initially, traditional supervised ML algorithms were used for abnormal driving behavior, e.g., Singular Value Decomposition (SVD) [256], SVM [95], and Random Forest [257]. However, with the recent advances in DL, researchers are increasingly deploying DL algorithms to predict inattention. Wollmer et al. used a LSTM network that outperformed the traditional SVM approach [258]. Moreover, some studies have used modern computer vision techniques for image and video processing to extract features automatically, which outperformed the traditional ML methods [259], [260]. Chen et al. developed an automatically constructed deep CNN that extracted high dimensional mappings of various sources of information and detected different types of driver distraction [261]. Recently, [262] demonstrated the effectiveness of an ensemble model for driver anomaly detection in a supervised learning approach. However, supervised classical approaches, which depend on labeled data, may face limitations. The expense involved in data collection could prevent the coverage of all potential abnormal driving behaviors. In contrast, unsupervised approaches offer the advantage of gathering extensive real-life data at a significantly lower cost because it does not require labeling. As a result, researchers have explored weakly-supervised and unsupervised anomaly detection methods as an alternative for identifying dangerous driving behaviors.

- weakly supervised: Some works tried to alleviate the need for dangerous driving behavior data collection by developing weakly supervised models [263] [264] [265]. In [266], a novel contrastive learning method is introduced to distinguish between normal and anomalous driving behaviors using the Driver Anomaly Detection (DAD) dataset. This dataset comprises video clips illustrating normal driving and instances of abnormal behavior depicted as distracting acts. The approach involves learning embeddings of these clips using CNN layers and employing contrastive learning techniques. Specifically, the method maximizes the similarity among embeddings of normal driving clips while emphasizing dissimilarity between representations of normal and abnormal driving. Their proposed similarity measure successfully identifies abnormal driving in new samples, showcasing promising results, including the detection of unseen anomalous actions. However, a notable limitation remains: the model still requires abnormal data during the training process.
- Unsupervised approach: Zhang et al. [96] introduced SafeDrive, an unsupervised approach tailored for the detection of abnormal driving behaviors through the analysis of extensive vehicle data. Utilizing a Statistical Graph derived from normal behaviors in a sizable dataset, SafeDrive effectively identifies anomalies within real-time driving data streams. However, their methodology predominantly relies on vehicle data (e.g. RPM, swerve angle, and gear position), which presents a challenge as certain behaviors flagged as unsafe might be contextually safe given specific environmental conditions or influenced by the behaviors of other drivers. In [267], the authors introduce the utilization of conditional GAN for detecting driving anomalies, employing physiological and CAN-Bus data [1]. Their findings support the efficacy of unsupervised methods; however, the evaluation lacks comparison with other anomaly detection techniques. Moreover, Dairi et al. [268] proposed detecting drunk driving behavior by combining t-distributed stochastic neighbor embedding (t-SNE) as a feature extractor with the Isolation Forest algorithm. Qiu et al. [269] introduced an unsupervised technique to identify irregular driving behaviors employing conditional GANs and contrastive learning. Their method involves training a conditional GAN for each modality, predicting forthcoming signals. They subsequently fuse the in-

[1] A sequential broadcast bus was developed to facilitate communication among the electronic control units installed in the vehicle.

formation from these modalities by utilizing the layer embedding from the discriminator. Their contrastive loss implementation employs a triplet loss function, aiming to minimize the distance between predicted and observed data while maximizing the distance between predicted data and a randomly selected segment within the dataset.

The literature has witnessed a surge in research leveraging normal data for unsupervised driver behavior anomaly detection, yet few studies offer comparative analyses among these methods. Our work aligns with these studies but focuses on eliminating the need for labeled abnormal data, enhancing real-world applicability. We conduct a comprehensive evaluation spanning classical to DL-based anomaly detection methods. Furthermore, we concentrate on leveraging non-intrusive sensors, particularly eye-tracking features.

3.2.2 Psychotic relapse detection

Similar to visual distraction detection, Classical supervised learning methods have been applied to detect psychotic disorders and relapses, such as Support Vector Machine [270], deep CNN [271], LSTM neural network [272], etc. Othmani et al. [273] developed two models, one for depression (anomaly model) and one for non-depression (anomaly-free model), and identified relapse by determining if a sample is more strongly correlated with the depression model.

Recently, due to the rare occurring nature of relapse, unsupervised anomaly detection methods have been proposed for relapse detection [274] [275]. In a two-country longitudinal study, authors investigated the application of anomaly detection to predict relapses in patients exhibiting symptoms of psychosis [276]. This recent study showed that the frequency of anomalies increased 2.12 times in the month before and 2.78 times in the month after a relapse compared to other times. In [277], authors studied relapse prediction using anomaly detection-based convolutional variational autoencoder (CVAE) on speech signals. They compared the performances of CVAE and a deterministic convolutional autoencoder CAE baseline for the global and personalized schemes. Their results showed that the CVAEs and CAE baseline achieved a similar performance for the personalized scheme. However, the CVAE performed significantly better than the CAE baseline for the global scheme. Furthermore, [99] compared the use of four different AE architecture models for detecting relapses in patients with different psychotic disorders using physiological signals collected by smartwatches. Calgagno et al. employed transformer models personalized per patient to detect psychotic relapse using wearable signals.

The prediction of psychotic relapse remains a complex and underexplored area, demanding further investigation to refine outcomes and deepen our understanding, especially concerning unsupervised methodologies. Few studies have tackled this area with limited datasets, underscoring the need for more research to validate existing findings. Moreover, there's an ongoing requirement to identify optimal features for improved prediction accuracy. Our research focuses on advancing anomaly detection-based models for predicting psychotic relapses. We employ methodologies such as AE and compare their efficacy against classical methods. This exploration is relevant, especially for detecting rare states within smaller databases. Additionally, our study explores the personalization of models across various levels, including anomaly detection methods and feature combinations.

3.3 Anomaly detection methods for driver monitoring: a proof of concept

The first application under consideration is monitoring driving behavior. Given the complexity of the driving task, drivers must maintain physical, mental, and visual engagement. Any

compromise in these aspects causes anomalous behavior and can lead to severe or fatal accidents. Anomalous driving behaviors encompass a range, from fatigue and drowsiness to distraction and alcohol consumption. However, most research tends to concentrate on detecting these behaviors individually [278] [279]. Numerous studies have pursued this task by employing the conventional supervised learning approach, necessitating instances of both normal and abnormal behavior in a balanced distribution. However, as previously mentioned, this requirement either endangers drivers to induce such events or demands data acquired within a simulated environment, compromising performance. Hence, we propose to construct a model that characterizes normal driving behavior through anomaly detection methods. By leveraging anomaly scores, we aim to establish an indicator of a driver's capability to operate a vehicle safely. Our initial step, as a contribution, involves validating the relevance and feasibility of this idea before employing it in the real world. As a proof of concept, we focus first on one type of dangerous driving: visual distraction detection. Although visual distraction isn't directly a mental state; it occurs when a driver's attention shifts away from the road, causing their brain to engage with stimuli unrelated to driving. Moreover, driver monitoring falls within the realm of affective computing, where technologies gauge and respond to human emotions and behaviors, aiding in the assessment of a driver's cognitive state and potential distractions. We choose the distraction detection task as it has been found to be one of the primary causes of car accidents [280] [281], [282]. Several experiments have established that engaging in a secondary task other than driving caused a delay in response time [283] [284], delay in the detection of visual stimuli [285], and a weakened driving performance [286]. Moreover, several accidents have been reported in autonomous driving vehicles that resulted in fatalities. With respect to autonomous driving, driver distraction during emergencies prevented the drivers in these scenarios from taking the appropriate corrective actions [261]. Our primary contribution is to ascertain the relevance and feasibility of this idea. To achieve this, we require data that simulate dangerous driving situations, specifically focusing on distracted driving. Distracted driving serves as an ideal starting point due to the ease of accessing related datasets and its significant impact on safety. The objectives of this following section are to:

1. Select a comprehensive dataset for this proof of concept.
2. Compare various anomaly detection methods.
3. Compare anomaly detection methods with supervised learning, particularly in scenarios where the dataset exhibits an imbalance, containing a higher frequency of normal instances compared to abnormal ones. To simulate this data imbalance, we will deliberately create a dataset that mirrors such real-world conditions. This comparative analysis will shed light on the efficacy and performance of these two approaches when handling imbalanced datasets.
4. Evaluate our proposal to use anomaly scores as indicators of abnormal driving (distraction levels during driving).

3.3.1 Dataset

The dataset utilized in our research originates from the European project HADRIAN ("Holistic Approach for Driver Role Integration and Automation Allocation for European Mobility Needs" [287]), developed in collaboration with our laboratory at CEA. Contrary to traditional automotive approaches that primarily focus on integrating automated driving as a new vehicle function for commercial purposes, HADRIAN places a premium on crafting comprehensive mobility services. This broader approach encompasses road infrastructure elements and accounts for human drivers, shaping the dataset with a specific emphasis on the drivers' states and requirements. Given the critical importance of this task, numerous research endeavors have concentrated on constructing datasets for monitoring driver behavior [266] [288] [289]. Some of

these datasets concentrate solely on one modality [266] [288], while others focus on specific types of abnormal driving [290] [291] [292]. However, the distinctive feature of the HADRIAN project lies in its inclusion of simulations and its broad scope. It does not confine itself to a single form of abnormal behavior but encompasses various risky behaviors such as 'Fatigue', 'Visual distraction', and 'Stress'. This characteristic opens up possibilities to further explore our anomaly detection method for a range of abnormal behaviors. Furthermore, the dataset employs diverse sensor types. The dataset includes video recordings, physiological measurements, eye-tracking data, and signals linked to driving simulators. Another advantage of this dataset is the provision of two driving modes, manual and autonomous, offering a comprehensive representation of varied driving scenarios. It also allows us to test our method on two different datasets that share the same abnormal behavior ¨distraction¨ but do differ in normal driving behavior.

3.3.1.1 Participant selection

The dataset consists of 43 participants aged between 26 and 54 years old (40 of them are male). All participants hold a valid B driving license and have jobs that require daily driving. All the realized experiments were conducted in agreement with the code of Ethics of the World Medical Association[293].

3.3.1.2 Driving simulation environment

The experiments were conducted using a Nervtech driving simulator system [294] that has a real car seat and steering wheel as shown in Figure 3.1. In addition, a surround sound system, a four-degree freedom motion platform, and three screens are used.

Figure 3.1: *A driving session setup*

3.3.1.3 Scenarios

As depicted in Figure 3.3, each participant engaged in two driving sessions: a "control session" and a "distraction session". Within each session, two modes of driving, "manual "and "autonomous driving" modes were sequentially explored, resulting in four distinct sessions. The numbers of examples in each partition are presented in Figure 3.2. The distribution of the control (non-distracted) versus distraction scenarios in the dataset does not mirror their occurrence rates in real-life settings, where distracted driving is notably less frequent than normal driv-

ing. However, our primary focus was on acquiring as many distracted examples as possible to thoroughly test the models' capabilities in detecting distraction.

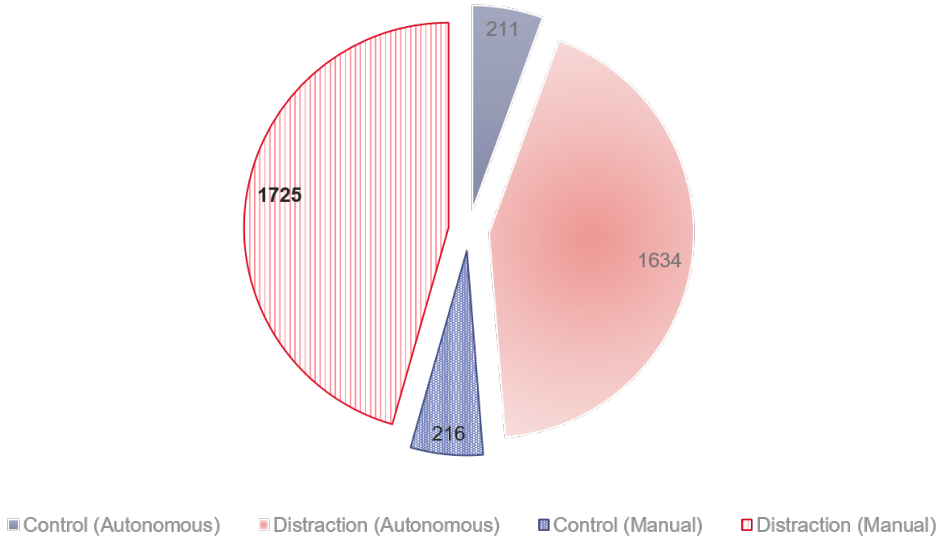

FIGURE 3.2: *Number of examples in the driving scenarios*

In the autonomous driving mode, the participant did not participate in driving. During the "control scenario" the participant drove in an undisturbed environment for approximately five minutes, representing our definition of normal behavior setting. Conversely, in the "distraction scenario", the participant drove while simultaneously using a tablet for about 20 minutes ($\sim$10 minutes in manual driving mode and $\sim$10 minutes in autonomous driving mode). Therefore, the distraction scenario encompasses both normal and abnormal behaviors. The subject was also asked to use the tablet in three ways: while it is mounted to the dashboard, while holding it with one hand, and while holding it with both hands. Such interactions pose potential risks in real-world scenarios, which is precisely why we opted for simulated data and why we exclusively utilize the control scenario to train our anomaly detection model.

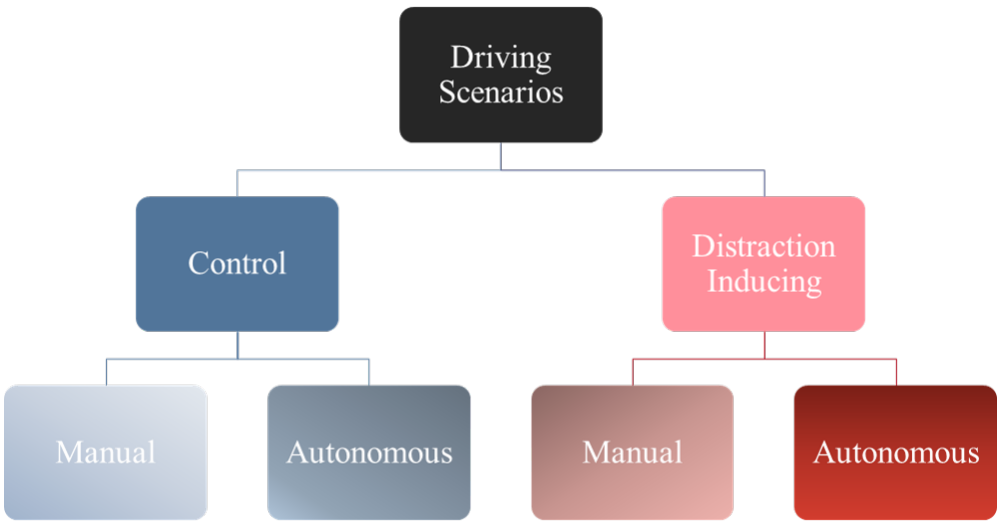

FIGURE 3.3: *Driving scenario*

3.3.1.4 Sensors and features

During the driving session, the drivers' eye movements are tracked using a Smart Eye Pro system [295]. Also, the participants' physiological signals ECG , EDA, respiration movements, and blood volume pressure were recorded using the Plux system [296] and video using Intel Realsense D435i [297]. Moreover, the face temperature was measured using FLIR A325sc [298]. The SCANeR studio [299] driving simulator software tracks dynamic vehicle information e.g., speed, lane crossing rate, etc.
Given our focus on visual distraction detection and the consistent findings in prior research highlighting the significance of eye-tracking features in this context [94][95][300], we specifically concentrate on eye-tracker features calculated by the smart-eye system within a 30-second time window. These features include 'Saccade magnitude', 'Saccade rate', 'Saccade peak velocity', 'Eye position entropy', 'Gaze heading mean', 'Gaze heading standard deviation', 'Gaze pitch mean', 'Gaze pitch standard deviation', 'Head heading mean', 'Head heading standard deviation', 'Head pitch mean', and 'Head pitch standard deviation'.

3.3.1.5 Annotation

In the dataset, two judges continuously annotated the data by observing videos of the driver and the driving scene. Throughout the distraction session, they identified instances of distraction-related activities such as "Looking out the left window" or "Writing on a device," noting both the start and end times of these acts. In order to evaluate our ideas, we developed a distraction level indicator from those objective annotations. It will be considered as our "gold standard" for our evaluations. We segmented the recordings into smaller windows of 30 seconds each, comprising 25 frames per second. For each frame, we binarize the annotation: if one of the annotators indicated at least one distraction-related action, the frame will be labeled 1, otherwise 0. As shown in Figure 3.5, for each clip, the level of distraction is equal to the number of frames labeled 1 divided by the total number of frames. We illustrate an example of the evolution of distraction levels for a single driver during the distraction scenario in Figure 3.4. In Table 3.1, we show the

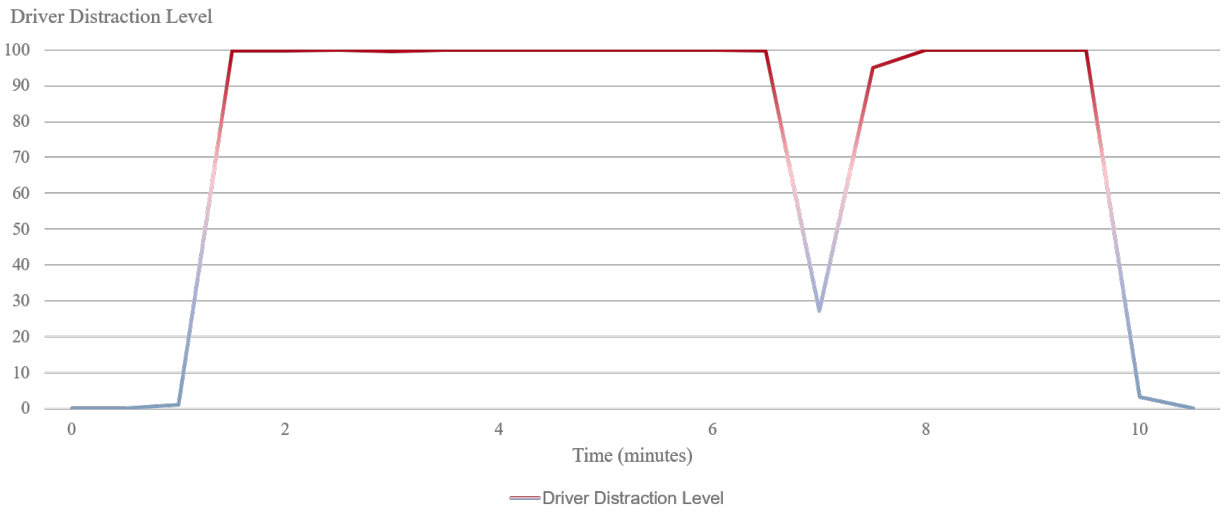

Figure 3.4: *Temporal evolution of driver distraction levels during distraction scenario in autonomous driving mode*

distribution of distraction levels across the two driving modes data in all the distraction sessions.

TABLE 3.1: *Distraction level distribution in autonomous and manual driving mode data*

Scenario	Driving mode	Distraction level l	Number of examples
Distraction	Autonomous	$l = 0$	181
		$0 < l \leq 25$	87
		$25 < l \leq 75$	132
		$l > 75$	1234
Distraction	Manual	$l = 0$	291
		$0 < l \leq 25$	89
		$25 < l \leq 75$	148
		$l > 75$	1197

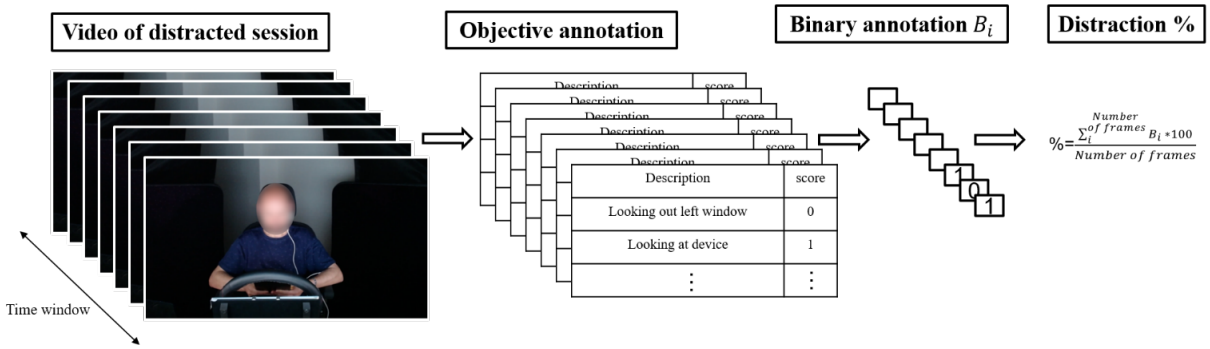

FIGURE 3.5: *Distraction level "gold standard" for our experiments*

3.3.2 ML EXPERIMENTAL SETUP AND DATA PARTITIONING

3.3.2.1 OUR IDEA: LEARNING USING ONLY "NORMAL" DATA

Our approach can be summarized in two steps: Model fitting and Anomaly Detection. During the training phase, the models are exclusively exposed to normal data, representing the majority class. In this stage, the model learns the patterns and distributions within the normal data (non-distracted driving). During the testing phase, we introduce a mix of normal and rare data (distracted data), designated as abnormal data, falling under the minority class. The model then produces an anomaly score, quantifying the dissimilarity of the new data compared to the normal data it encountered during training. Subsequently, these scores can be thresholded to generate binary predictions classifying the data as normal or abnormal.
While we explore a variety of techniques, the training process remains consistent across all models, as illustrated in Figure 3.6. We implement OCSVM , LOF , Elliptic Envelope, and Isolation Forest using the Scikit-learn library [301]. As mentioned in Chapter 2 Section 2.4.8, we choose the pessimistic approach and adopt the default parameters of the anomaly detection models. For OCSVM , we use rbf kernel, degree = 3, and nu= 0.1. For the isolation forest, we set the number of estimators to 100, and contamination to 0.1. The number of neighbors is set to 20 for the LOF , and the distance used for computation is 'Minkowski', with a contamination rate of 0.1 for the elliptical envelope. We use Tensorflow [302] to implement the auto-encoder. We choose an encoder composed of one dense layer comprising 4 neurons that projects data into a 2-dimensional space. The decoder is also composed of one dense layer of 4 neurons. The choice of the hyperparameters was based on having compressing layers. We set the threshold by taking the right value of the 99% confidence interval on the RE of the normal data.

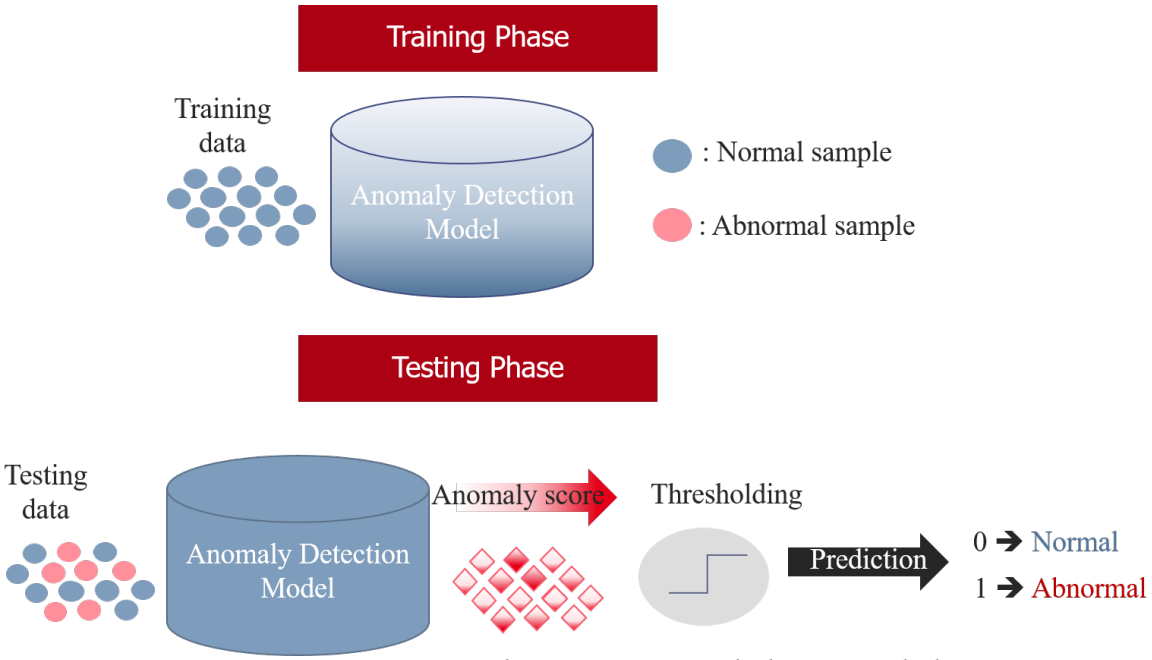

FIGURE 3.6: *Rare states detection using anomaly detection methods.*

3.3.2.2 CLASSICAL APPROACH: SUPERVISED LEARNING

Furthermore, we train classical supervised models to compare with our proposed method. We employ five classical supervised models: MLP, SVM , K-neighbors classifier, Random Forest (RF), and Gaussian Naïve Bayes classifier detailed in Chapter 2. Our MLP is configured with two dense layers, one containing four neurons and the other containing two neurons. The loss function is the classical binary cross-entropy loss.

3.3.2.3 COMPARISON BETWEEN ANOMALY DETECTION AND CLASSICAL APPROACH

To strengthen the validation of our anomaly detection-based approach, we conduct several classical methods evaluation across various simulated data distribution scenarios. These scenarios encompass different degrees of data imbalance, closely mirroring realistic data distributions found in real-world settings. Imbalanced datasets often pose challenges for classical learning methods, leading to less emphasis on the underrepresented class. Approaches to address this imbalance, such as duplicating samples from the underrepresented class or adjusting the loss function, have been explored. Thus, we have included scenarios that incorporate augmented data solutions.

3.3.2.4 LEVEL OF DISTRACTION ESTIMATION USING ANOMALY SCORE

In addition to evaluating the utility of anomaly scores for binary classification (identification of distraction), we leverage the continuous nature of available annotations to test if the anomaly score can provide an estimation of the distraction level. To assess the viability of this approach, we calculate the correlation between the anomaly scores obtained from the models in Experiment 1 and the distraction level labels within our dataset. For the OCSVM, the anomaly score is the signed distance of each point to the separating hyperplane. As for LOF, we take the negative of the outlier factor. For Isolation Forest, we use the number of splits required to isolate the data as the anomaly score. For the AE, we use the reconstruction error. Last, the elliptic envelope algorithm employs the squared Mahalanobis distance between the observation and training data distribution to derive its anomaly score.

3.3.2.5 Data partitions

In our study, we conduct a series of experiments considering the two distinct driving modes: manual and autonomous. Given the variance in normal driving behavior between these modes, we train separate models and present the resulting outcomes for each driving mode. Each experiment utilizes the dataset corresponding to the respective driving mode:

- Supervised Classification: For training, we use the N (number of examples in the control scenario) control samples as the normal examples (Negative Class) and randomly choose N samples with extreme distraction level >75% as the abnormal examples (Positive Class) to create a balanced training dataset where N is the number of examples in the control session. To test the models' performance, we test using data from the distraction session. Similar to the training data, we take the data annotated with 0 as the level of distraction for the negative class and data annotated higher than 75% as the level of distraction for the positive class.
- Anomaly Detection: The training data includes the N examples from the control scenario only. We use the same testing dataset used in supervised classification.
- Comparison between supervised and unsupervised: We present four scenarios for this experiment.

  1. Scenario 1: We train the model using N normal examples from the control and 0.1×N anomalous examples randomly chosen from the distraction session (with distraction level > 75 %).
  2. Scenario 2: We train the model using N normal examples from the control and 0.3×N anomalous examples randomly chosen from the distraction session (with distraction level > 75 %).
  3. Scenario 3: We train the model using the data from scenario 1 with augmentation, where we duplicate the examples of distracted driving until we reach an equal number of normal and abnormal instances.
  4. Scenario 4: We train the model using the data from scenario 2 with augmentation, where we duplicate the examples of distracted driving until we reach an equal number of normal and abnormal instances.
     All scenarios are tested using the same testing dataset used in supervised classification.

- Level of distraction estimation using anomaly score: after training the anomaly detection models, we calculate the Pearson correlation coefficient between the annotated distraction level and the anomaly scores of each method. For this experiment, we use all the examples in the distraction scenario to have data with different levels of distraction ranging from 0% to 100%.

3.3.3 Anomaly detection-based methods evaluation

In this section, our objective is to evaluate anomaly detection methods. To accomplish this, we first assess supervised classification methods, aiming to later compare their performance with that of anomaly detection methods.

3.3.3.1 Supervised classification results

Table 3.2 illustrates the outcomes of binary classification through various supervised classical models. The results indicate comparable performances among the models. SVM stands out as the best-performing model for manual driving data, while KNN exhibit superior performance for autonomous driving data. The performance on the autonomous dataset is superior to the ones achieved on the manual dataset for all models. It is worth noting that further improvements can be achieved by applying grid search to optimize the model's hyperparameters. However, the primary objective here is not solely focused on achieving the highest performance but rather gaining insights into the performance range of classical supervised methods.

Table 3.2: *Comparison of supervised methods using F1 and balanced accuracy on the manual driving data.*

Driving Scenario	Model	MLP	SVM	K-Neighbors	RF	Naïve Bayes
Manual	$F1$	0.81	0.84	0.82	0.83	0.83
	Balanced accuracy	81.97	84.96	81.97	83.92	83.33
Autonomous	$F1$	0.89	0.89	0.89	0.89	0.87
	Balanced accuracy	88.72	88.4	89.31	88.77	87.41

3.3.3.2 Anomaly detection results

In Table 3.3, we present the $F1$ and balanced accuracy of each model on the manual and autonomous driving modes data separately. For the manual driving scenario, results show that the elliptic envelope outperforms the rest of the methods with 0.83% $F1$ and 82.28% balanced accuracy. However, AE shows the worst performance $F1$ at 0.57% and balanced accuracy at 68.22%. For the autonomous driving data, results show that the best-performing model is the LOF . Similarly to supervised models, we also observe that the models generally perform better on autonomous driving mode data. An explanation for this result may be that in the autonomous driving mode, the driver can possibly be completely engaged in the distracting task, whereas in the manual driving mode, they may still have to steer the wheel or check the road. This can make the separation between distracted and not distracted data easier in the autonomous mode. For both driving modes, LOF and Isolation Forest exhibit robustness, which could potentially be attributed to lower sensitivity to hyperparameter tuning compared to other anomaly detection methods. To comprehensively evaluate the performance of anomaly detection methods,

Table 3.3: *Comparison of unsupervised methods by F1 and balanced accuracy.*

Driving Scenario	Model	OCSVM	LOF	Isolation Forest	Elliptic Envelope	AE
Manual	$F1$	0.58	0.81	0.82	0.83	0.57
	Balanced accuracy	68.9	78.22	82.46	82.28	68.22
Autonomous	$F1$	0.5	0.9	0.87	0.84	0.79
	Balanced accuracy	65.63	89.13	87.35	84.85	82.64

we compute the ROC AUC and PR AUCscores. As elaborated in Chapter 2, these metrics serve as prominent benchmarks for assessment. In Figure 3.7, we present the ROC AUC values for various models concerning the manual and autonomous driving datasets. For the manual data (depicted on the left side of Figure 3.7), the models exhibit fairly similar performance when the hard thresholding step of the anomaly scores is removed. Most models achieve ROC AUC scores around 0.88, demonstrating consistent performance, except for the AE model, which registers a

slightly lower score of 0.83 compared to other methods.
On the right side, the ROC AUC scores for the autonomous dataset mirror the trend observed in the F1 score and balanced accuracy metrics, with all models displaying enhanced performance. Each model attains a ROC AUC of 0.94, except for the AE model, which achieves a slightly lower score of 0.91. This consistent improvement underscores the efficacy of these models in anomaly detection for autonomous driving scenarios. Additionally, the results underscore a lower performance by the AE , suggesting a potential need for more fine-tuning to reach parity with other methods.
We also provide the PR AUC scores in Table 3.4 which further confirm our findings in Table 3.3

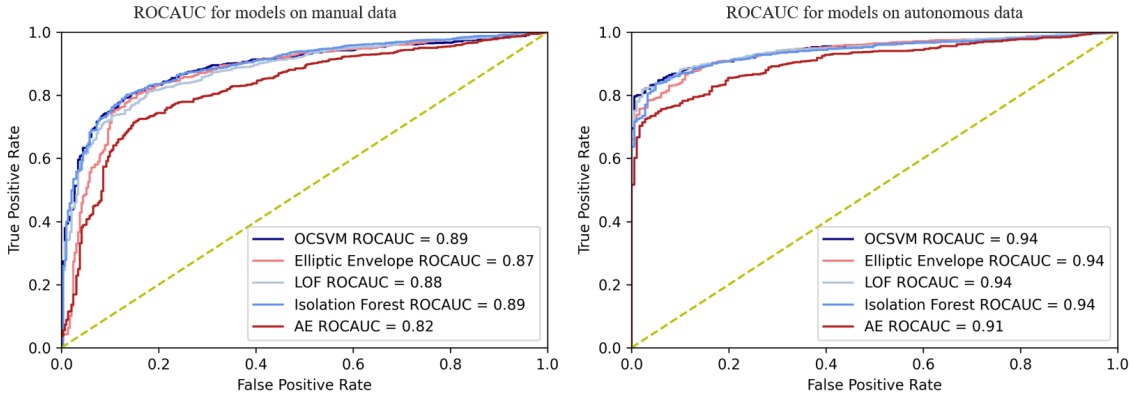

FIGURE 3.7: *ROC AUC performance for the anomaly detection on autonomous and manual driving modes.*

and Figure 3.7.

TABLE 3.4: *PR AUC performance for the anomaly detection on autonomous and manual driving modes.*

Model	**OCSVM**	**Isolation Forest**	**LOF**	**Elliptic Envelope**	**AE**
Manual Driving	0.97	0.97	0.97	0.96	0.94
Autonomous Driving	0.99	0.99	0.99	0.99	0.98

Based on the outcomes obtained from anomaly detection methods, we observed that detecting distractions in autonomous mode is more straightforward compared to manual mode. Furthermore, most anomaly detection methods exhibited comparable performances, except for AE, which might require fine-tuning to yield better results.

3.3.3.3 Comparison with supervised learning approach performance

To assess the sensitivity of supervised learning models to class imbalance, we evaluate the best supervised models (SVM for the manual dataset and KNN for the autonomous dataset) in various data distribution scenarios. Table 3.5 demonstrates that supervised learning, even with data augmentation, is penalized by the imbalance in the dataset, even when using 0.3xN distracted samples, which does not correspond to a significant imbalance in class distribution. In the case of a balanced dataset, anomaly detection methods yield a performance that is similar to that of supervised methods. Consequently, being agnostic of anomalous examples, our unsupervised methods perform better than the supervised models when the dataset is imbalanced. This outcome showcases the effectiveness of our approach in simulating real-world scenarios with limited or imbalanced labeled data, achieving comparable results to a supervised learning model trained on a balanced dataset.

TABLE 3.5: *Comparison of best-supervised models performance using varying number of anomalous examples*

	manual data		autonomous data	
	SVM		KNN	
Number of anomalous examples	F1	Balanced Accuracy	F1	Balanced Accuracy
Scenario 1: 0.1*N	0.39	61.95	0.68	75.85
Scenario 2: 0.3*N	0.7	76.76	0.89	90.12
Scenario 3: 0.1*N+ Data augmentation	0.39	62.24	0.74	79.51
Scenario 4: 0.3*N + Data augmentation	0.71	76.87	0.9	90.64
Balanced dataset N	0.84	84.96	0.89	89.31
	Elliptic Envelope		LOF	
0	0.83	82.28	0.9	89.13

3.3.3.4 LEVEL OF DISTRACTION ESTIMATION USING ANOMALY SCORE

TABLE 3.6: *Correlation between anomaly scores of models and the annotated level of distraction*

Model	**OCSVM**	**Isolation Forest**	**LOF**	**Elliptic Envelope**	**AE**
Manual Driving	0.59	0.57	0.49	0.4	0.38
Autonomous Driving	0.65	0.6	0.56	0.43	0.42

In the preceding subsections, we demonstrated the utility of anomaly scores derived from anomaly detection methods for identifying distracted instances. We showcased the significance of this approach in learning from data without or with limited anomalous examples. Furthermore, we want to explore if the anomaly detection methods can yield information about the level of distraction. The hypothesis is that the anomaly score could serve as an estimate of the distraction level. To assess this, we compute the Pearson correlation coefficient between the annotated distraction level and the anomaly scores obtained from each method. We use all the data from the distraction scenario to have data with different levels of distraction ranging from 0% to 100%. Table 3.6 shows that all obtained anomaly scores are correlated with the level of distraction. Results also show that OCSVM shows the strongest correlation, with 0.59 on manual driving data and 0.65 on autonomous driving data. Furthermore, our analysis reveals that the Isolation Forest consistently demonstrates robust performance across all evaluations, emerging as one of the most resilient models. In contrast, the AE exhibits lower performance compared to other classical methods. The dataset's dimensionality and size may contribute to classical methods outperforming the AE. This initial attempt to estimate the level of abnormality using anomaly scores in an affective computing application showcased a significant correlation, providing encouragement for further exploration of this approach.

3.3.4 CONCLUSION

In this study, our objective was to assess anomaly detection methods for predicting abnormal driving behavior, negating the need for risky data collection. Utilizing a database derived from a driving simulator ensured safe data procurement, facilitating the evaluation, with visual distraction detection chosen as the targeted abnormal behavior. Our comparisons spanned various anomaly detection techniques, encompassing classical and DL methods, using multiple metrics (F1 score, balanced accuracy, ROC AUC, and PR AUC). Our analysis revealed LOF and Isolation

Forest as the most robust models across all metrics for both manual and autonomous datasets. Additionally, we conducted experiments in supervised classification, demonstrating that our anomaly detection methods outperformed supervised approaches in imbalanced datasets and yielded similar performance in balanced datasets, underscoring the effectiveness of our approach. Moreover, we introduced the concept of estimating the level of abnormal behavior from anomaly scores in affective computing applications. The correlation results encourage further investigation of this approach. This holds promise in tailoring feedback to users based on this estimated level. The evaluation of our work encompassed both manual and autonomous driving modes. In autonomous vehicles, our model's scores can serve the crucial purpose of determining when it's safe to transition control back to the driver. Conversely, in manual mode, it can offer insights into detecting potentially hazardous driving behaviors. Nonetheless, for real-world implementation, it's essential to test smaller time window—a facet not extensively explored in our study, as our primary aim was to establish the viability of anomaly detection methods for identifying dangerous driving behavior as a proof of concept.

3.4 Learning behavioral patterns to detect psychotic relapses

In the preceding section, we evaluated our anomaly detection approach in the domain of driver monitoring, demonstrating its potential usefulness for detecting dangerous driving behavior. Mental health represents another promising avenue for this approach, as highlighted in previous studies in section 3.3.2.1. In mental health, relapse detection of patients with psychotic disorders holds paramount importance. Based on the results of previous studies [99], we further investigate tackling this problem with an anomaly detection approach.

Psychotic Relapse is defined as the reappearance of psychotic symptoms following a period of remission. Detecting relapse in the early stages of mental illness can facilitate quicker and easier recovery [303]. The increased prevalence of smart devices, especially smartwatches [304], [305] presents an opportunity for continuous behavior monitoring, potentially aiding in the early detection of relapses. However, obtaining relapse data to train and test models is a challenge. Relapse is a rare occurrence that affects a relatively small percentage of the population, necessitating extended periods of data acquisition to record its incidence. As a result, researchers are inclined to develop unsupervised models that only require non-relapse (i.e., normal) data. This approach helps address the challenge posed by the absence of relapse data or imbalanced datasets. Therefore, employing anomaly detection methods where relapse is considered as an anomaly presents a viable solution to tackle these challenges effectively. Expanding on previous research [99], our study delves deeper into this concept. The subject was also highlighted in a challenge held at ICASSP23 [255], utilizing the same dataset. Hence, we've chosen to utilize the e-prevention dataset for our exploration. Following data pre-processing and feature extraction, akin to the driver monitoring study, we conducted a comparative analysis of various anomaly detection methods, one of which ranked second in the e-prevention challenge [306]. Moreover, recognizing the significance of personalization in mental health applications due to the diverse and individualized nature of mental disorders, we delved deeper into this aspect. To explore the implications of personalization, we compared models trained using all patient data against models trained solely on patient-specific datasets. This nuanced analysis aimed to discern the impact of personalization on the accuracy and efficacy of our anomaly detection approach in the context of mental health and relapse detection.
The contributions outlined in this section can be summarized as follows:

- Assessment and comparison of multiple anomaly detection methods for relapse detection, employing signals from wearable devices.

- Comparing general models trained on collective patient data against models customized for individual patients.
- A comprehensive exploration of diverse feature combinations and models across the entire patient cohort.

3.4.1 Dataset

Relapse prediction, an increasingly important area of study, faces challenges in obtaining datasets due to privacy concerns and regulatory restrictions. Several datasets pertinent to this field are either unavailable [307] [308] to the public or limited to specific data sources, such as mobile signals [309] [310]. In our study, we utilized the dataset from the ICASSP'23 Grand Challenge e-Prevention [255], offering an opportunity to examine patient data and detect relapses among individuals within the psychotic spectrum without relying on labeled relapse data. This dataset comprises daily signal recordings from a smartwatch from ten patients diagnosed with various psychotic disorders, including Schizoaffective Disorder, Bipolar I Disorder, Brief Psychotic Episode, Schizophreniform Disorder, and Schizophrenia. We provide a summary of the mental disorders encompassed within the dataset:

- Bipolar disorder, formerly recognized as manic depression, is a psychological disorder marked by alternating episodes of depression and extended periods of abnormally heightened mood, each persisting for days to weeks [311]. When the heightened mood is intense and linked to psychosis, it is referred to as mania; if it's less severe, it's referred to as hypomania. In the state of mania, individuals exhibit abnormal levels of energy and can feel excessively joyful or irritable, frequently leading to impulsive decisions [312].
- Schizophrenia is a mental disorder marked by persistent or recurring episodes of psychosis[313]. Psychosis is a state of the mind that leads to challenges in distinguishing reality from non-reality. Symptoms can encompass delusions, hallucinations, and various other characteristics. Furthermore, individuals with schizophrenia may exhibit incoherent speech and engage in behavior that is not contextually appropriate [314].
- Schizoaffective disorder is a mental condition distinguished by irregular thought processes and mood instability. Individuals diagnosed with Schizoaffective disorder exhibit symptoms of both schizophrenia, typically involving psychosis, and a mood disorder, which could be either bipolar disorder or depression [315].
- Schizophreniform disorder is a mental health condition that is diagnosed when symptoms resembling those of schizophrenia are present for a substantial duration of time, usually at least a month. However, the individual does not display the necessary signs of disruption persisting for the entire six-month period required for a schizophrenia diagnosis [316].

We selected this dataset for its public availability and previous use in anomaly detection studies. It provides insights into the feasibility of our work, facilitates comparisons, and includes diverse psychotic disorders. Also, it includes long-term recordings of patients, enabling monitoring over extended periods. Furthermore, these signals are collected via wearable devices. Studies demonstrate the growing acceptance of smartwatches for monitoring daily activities or signals [304], [305]. Additionally, research findings [317] indicate that most patients found consistent data tracking motivating and expressed a desire to continue such monitoring. This highlights the practicality of implementing a monitoring device in real-life scenarios.

The daily signals were recorded using a Samsung Gear S3 Frontier smartwatch equipped with an accelerometer, gyroscope, and non-invasive heart rate monitor. The recorded data includes heart rate, RR interval, accelerometer and gyroscope coordinates, sleep state at five-second intervals throughout the day, periods of physical activity, and the total number of steps. The annotated data distinguishes between Relapse and Normal days, identified by clinicians based on

monthly mental health assessments, questionnaires filled out regarding general psychopathology and relapse state, communication between patients and their physician and family, and the necessity for hospitalization. However, the nature of the relapse was not disclosed, and they did not indicate the chronological order of relapse days.

The dataset is partitioned into three subsets: Training (containing only normal data), Validation (containing normal and relapse data), and Testing data (containing normal and relapse data). The testing dataset and validation dataset have approximately the same data distribution. An inherent imbalance between relapse and non-relapse data is observed, common in real-world settings. The recorded days vary among patients, with patient 1 having the highest number of training days (248) and patient 9 with the lowest (105), detailed in Table 3.7. It's important to note that certain patients display extremely low relapse days, such as 3 days for patients 5 and 9, and 4 days for patient 7. These instances are highlighted in italic and bold in the results, indicating their significance as findings to be carefully considered due to the limited number of relapse days. This scarcity of relapse data poses challenges in evaluation, highlighting the need for models that can be trained without relying heavily on relapse days.

TABLE 3.7: *Number of days for each patient per each data partition.*

Patient	Training Non Relapse	Validation Non Relapse	Validation Relapse	Testing Non Relapse	Testing Relapse
1	248	31	9	31	10
2	179	22	57	23	57
3	204	25	13	26	13
4	168	21	17	21	17
5	176	22	3	23	4
6	217	27	22	28	22
7	210	26	4	27	5
8	230	29	93	29	94
9	105	13	3	14	4
10	169	21	73	74	22

3.4.2 Proposed methodology

Figure 3.8 illustrates the key steps of our method, which include data pre-processing, model fitting using non-relapse data, and evaluation. In the following sections, we will provide a detailed explanation of each step, clarifying the techniques and processes employed.

3.4.2.1 Data pre-processing and features extraction

- **Data cleaning:**
  Before feature extraction, we cleaned the collected data files by removing duplicate time intervals in the recordings, discarding heart rate values outside the acceptable range of 30 to 200 BPM, and removing heart rate values deviating more than 20% from the heart rate calculated using RR intervals. We also discarded accelerometer and gyroscope norm data outside the intervals [-19.6, 19.6], [-573, 573] following the challenge guidelines and eliminated instances of negative step counts in physical activity data.

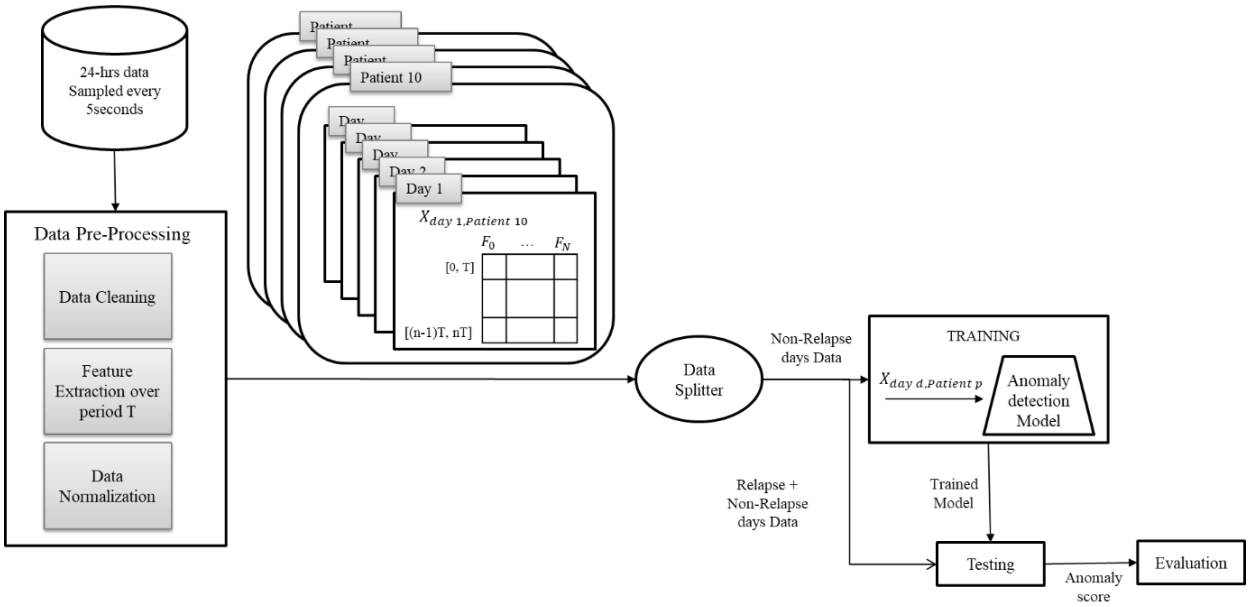

FIGURE 3.8: *Proposed Method schema for psychotic relapse prediction.*

Missing data resulting from data cleaning or the patient not wearing the watch were replaced by the median of the following feature over all days in the training dataset for the specific patient.

- **Feature Extraction:**
  Physical activity disruption has been shown to be a key feature for psychological state classification [318], [319]. Chapman et al. [320] conducted a study on 99 adult patients with different psychological disorders. Their results show that physical activity patterns extracted using accelerometers differ with the nature of the disorder. Moreover, Spulber et al. [321] showed that physical activity is also linked to the severity of the symptoms. In a study conducted on participants with Bipolar Affective Disorder, a decrease in physical activity was indicative of an increase in clinical symptoms of depression [322]. Sleep disturbance is also linked to mental health disorders such as schizophrenia [323]. It is also used as an indicator for relapse detection [274], [324]. Lambrichts et al. found that patients with sleep disturbance are more likely to relapse [325]. Physiological signals such as heart rate variability have been shown to strongly correlate with several psychotic disorders [326], [327]. For example, Esaki et al. observed an association between circadian activity rhythm, mood, and depressive episode relapses in patients with bipolar disorder [328].
  Therefore, we segmented the signals and extracted the following features from each segment, including mean heart rate, the standard deviation of heart rate, the norm of accelerometer coordinates, the norm of gyroscope coordinates, the percentage of sleeping time, and the total number of steps. Supposing that the act of wearing the watch and maintaining the setup might be affected during relapse periods, we computed the watch-wearing duration by considering the initial missing data as the time during which the patient removed the watch. In total, we obtained 7 features. We refer to the features % of sleeping time by sleep, mean of the heart rate and its standard deviation by HR, accelerometer norm by Acc, Gyroscope norm by Gyr, and Percentage of time wearing the watch by %WW. Finally, we standardize data per patient similarly to [99]. We illustrate an example of the sleeping pattern for patient 6 on both a relapse and a non-relapse day in Figure 3.9. The figure demonstrates a notable variation between the two.
  Additionally, we sought to investigate whether changes in daily patterns during relapse

were more pronounced within specific time windows. To explore this, we conducted experiments using four distinct time window lengths: 5 minutes, 1 hour, 4 hours, and 24 hours.

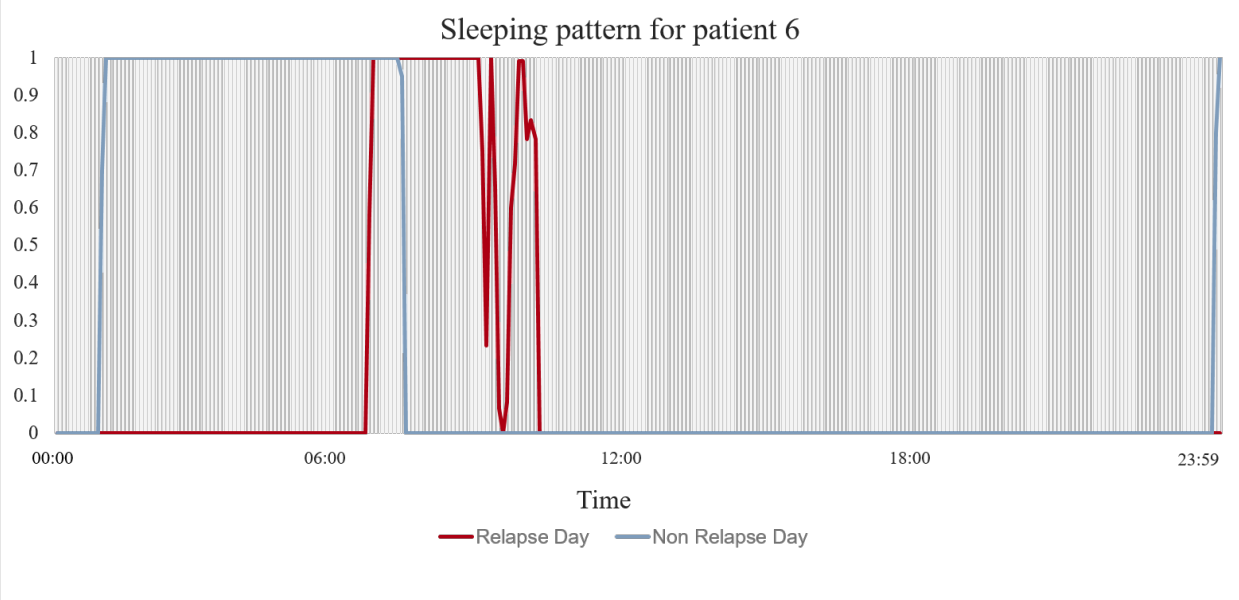

FIGURE 3.9: *Difference in sleeping behavior between relapse day and normal day for patient 6.*

3.4.2.2 MODELS

Our study focuses on learning patterns of physical, sleep, and physiological activity during normal non-relapse days to detect any deviations from relapse. To accomplish this objective, we adopt a methodology similar to the driving behavior monitoring study. We utilize classical and DL anomaly detection methods and test: AE, LOF , Elliptical Envelope, and One-Class SVM. Additionally, we introduce an LSTM AE, a pertinent addition in this context as temporal patterns related to relapse might be present, unlike in the case of visual distraction detection. The input for classical models involves concatenating features calculated over the time segment length for 24 hours. For instance, with a time window length of 1 hour, the input will be a vector of size 7 * 24 (7 features, 24 hours). Conversely, for a time window length of 24 hours, it reduces to a vector of size 7. In the case of LSTM, the data is inputted sequentially, where each input dimension represents the time window length by 7.
For OCSVM , isolation forest, LOF , and elliptical envelope we use the same parameters as for the study for visual distraction detection detailed in section 3.3.2.1. As for the AE, to account for varying input sizes, which depend on the number of features and time window size for calculating the features, we opt to have one hidden layer with N/2 number of hidden neurons for both AE and LSTM AE, where N is the number of features. We train the model using the Adam optimizer, the Mean Squared Error (MSE) as the loss function, a batch size of 16, and a maximum of 1000 epochs. We use 20% of the training dataset for early stopping with patience of 10 epochs. In the case of LOF , IF, OCSVM , and elliptical envelope, we use the distance function as an anomaly score, while the averaged RE of the input is used in the case of the AE.

3.4.2.3 TRAINING SETTINGS AND EVALUATION

After the phase of feature extraction, several pivotal questions have emerged during the construction of our relapse detection system that we aim to address. These questions primarily revolve around determining:

- The pertinent selection of features for the relapse detection system.
- We have calculated features over different time lengths, and what time window is the most efficient.
- The most effective anomaly detection method to employ.
- The suitability and performance comparison between general and personalized models for this application.

These questions will be key in shaping the subsequent steps of our approach. Our study aims to offer a thorough exploration of how the selection of features interacts with the choice of models and time window for feature extraction. This investigation will encompass both a global perspective and a personalized approach, providing a deeper understanding of the complex dynamics at play in relapse prediction .

- **Global scheme:** It involves training a single model using data from all patients. This approach enables us to evaluate the generalization potential of unsupervised learning techniques for detecting relapses in psychotic disorders. Prior to training, we apply the min-max normalization technique on every feature across all patients' data after normalizing the data per patient. This normalization process ensures consistency and comparability of the features used in the training process.
- **Personalized scheme:** We train, validate, and test the model exclusively on each patient's data. This approach may result in models that are tailored to the unique characteristics of each individual's data, potentially increasing the accuracy of predictions for that specific individual.

We assess the predictive performance of each method by comparing them to the relapse labels, using two commonly used metrics introduced in section 2.4.9: ROC AUC and PR AUC. We also compare using the average of both those metrics, which was used in the e-prevention challenge to rank the best models [255].

3.4.3 Results and discussion

In this section, we present our results in the global and personalized scheme. All the results presented were evaluated on the validation dataset, except for the comparison between personalized and global models at the end of the study was done on the testing dataset.

3.4.3.1 Global scheme

In this section, we showcase the results of our global scheme performance assessment. Initially, we conduct an exhaustive search over all potential combinations of features, models, and time window durations (TMW) to identify the best-performing configurations. Furthermore, we scrutinize the influence of each variable individually—namely features, models, and TMW—by maintaining one variable fixed while altering the others, calculating the average performance to assess the robustness of each model. Additionally, we employ Analysis Of Variance (ANOVA) tests to determine if the selection of the model or feature combination significantly impacts model performance.

1. **Exhaustive search best models**
   Table 3.8 presents the top-performing models in the global scheme, irrespective of the feature set and TMW. In contrast to the results found in 4.4.3 where AE was least performant for visual distraction detection, we found that the top three performing models are AE. It's probable that the hyperparameters are better suited for this particular problem. However, further analysis of the models' performance on average should be provided to verify

the difference in performance. The feature combinations that produce the best results are sleep ratio, heart rate mean and standard deviation, and accelerometer and gyroscope norm, yielding a mean performance score of 0.671. Moreover, the TMW of the best 2 models is 24 hours, and for the third is 4 hours.

TABLE 3.8: *Top 3 performing models in global scheme using exhaustive search*

Featureset	Model	ROC AUC	PR AUC	Average	TMW
Sleep, HR, acc,gyr	AE	0.658	0.684	0.671	24hrs
Sleep, HR, steps	AE	0.639	0.678	0.659	24hrs
Sleep, %WW, HR	AE	0.655	0.661	0.658	4hrs

2. **Sensor selection**
Table 3.9 presents the performance metrics for all possible combinations of modalities within the global scheme. In the table, the entries under the ROC AUC and PR AUC columns represent the average ROC AUC and PR AUC values, respectively, calculated across all models and all time window lengths (TMW): 5 minutes, 1 hour, 4 hours, and 24 hours. The column labeled 'Mean Performance Score' gives the mean value of the corresponding ROC AUC and PR AUC entries in the same row, providing an overall performance measure for each feature set. Based on the Table results, the best features for global relapse detection on average are sleep and heart rate features (including mean and standard deviation) with a mean performance score of 0.6. These results agree with the results of the exhaustive search in Table 3.8. Moreover, HRis the best-performing single feature, with a mean performance score of 0.59. Whereas steps is the worst performing feature with a mean performance score of 0.51.

TABLE 3.9: *Global scheme performance average across all feature sets on the validation set*

Featureset	ROC AUC	PR AUC	Average
sleep, HR	0.596	0.618	0.607
sleep, % WW, HR	0.581	0.609	0.595
HR	0.578	0.609	0.593
Sleep, HR, acc,gyr	0.579	0.607	0.593
Sleep, HR, steps	0.573	0.599	0.586
% WW,HR	0.568	0.602	0.585
sleep, % WW, HR, acc,gyr	0.566	0.598	0.582
sleep	0.564	0.598	0.581
HR, acc,gyr	0.562	0.591	0.576
sleep, % WW, HR, steps	0.56	0.588	0.574
HR, steps	0.555	0.59	0.572
sleep, HR, acc,gyr, steps	0.557	0.588	0.572
% WW, HR, acc,gyr	0.552	0.588	0.57
sleep, % WW, HR, acc,gyr, steps	0.548	0.584	0.566
% WW, HR, steps	0.545	0.581	0.563
HR, acc,gyr, steps	0.545	0.578	0.561
sleep, steps	0.536	0.573	0.555
sleep, % WW	0.537	0.573	0.555
Sleep ,acc,gyr	0.528	0.573	0.551
% WW, HR, acc,gyr, steps	0.53	0.57	0.55
sleep, % WW, acc,gyr	0.525	0.568	0.546
sleep,acc,gyr,steps	0.514	0.559	0.537
% WW	0.512	0.561	0.536
Sleep, % WW, steps	0.514	0.556	0.535
% WW, acc,gyr	0.505	0.555	0.53
Acc,gyr	0.502	0.553	0.527
sleep, % WW, acc,gyr, steps	0.502	0.551	0.526
Steps	0.485	0.546	0.515
% WW, acc,gyr, steps	0.483	0.54	0.512
acc,gyr, steps	0.481	0.539	0.51
% WW, steps	0.476	0.538	0.507

3. **Model selection**
Furthermore, in Table 3.10, we calculate the average ROC AUC, PR AUC, and mean performance score (mean of ROC AUC and PR AUC) for each model across all possible feature combinations and all TMW. These feature combinations were presented in 3.9. As shown in 3.10, the isolation forest model achieves the highest mean performance score (0.572) whereas LSTM AE achieves the worst mean performance score (0.526). These results further show the robustness of isolation forest which was found in 4.4.3. Additionally, the AE model emerged as the second-best performer on average, with approximately same performance as the Isolation Forest model. Interestingly, it secured the top position in the exhaustive search.

TABLE 3.10: *Global scheme performance average across all models on the validation set*

Model	ROC AUC	PR AUC	Average
Isolation Forest	0.552	0.592	0.572
AE	0.555	0.587	0.571
Elliptical Envelope	0.552	0.586	0.569
OCSVM	0.524	0.566	0.545
LOF	0.523	0.564	0.544
LSTMAE	0.501	0.551	0.526

4. **TMW selection**
To assess the effect of varying the TMW used to extract the features, in Table 3.11, we calculated the average ROC AUC and PR AUC across all feature sets and models. Moreover, as in the previous tables, we computed the mean performance score (mean of ROC AUC and PR AUC). We considered a TMW of '5 minutes', '1 hour', '4 hours', and '24 hours'. Table 3.11 illustrates that there is a slight variation in the performance depending on the TMW used for feature extraction, indicating that daily patterns may differ significantly across patients. Consequently, their use might not be globally relevant. These features may be more pertinent for personalized models, as we'll explore further in the personalized scheme results.

TABLE 3.11: *Global scheme performance average across all time windows on the validation set*

Period	ROC AUC	PR AUC	Total
5 minutes	0.544	0.582	0.563
1 hour	0.537	0.572	0.554
4 hours	0.541	0.577	0.559
24 hours	0.529	0.575	0.552

5. **Sensor and model selection effect significance**
ANOVA is a statistical method based on the law of total variance to determine if there is a difference between the means of several groups. It is commonly used to determine if a quantitative variable is dependent on a categorical variable. Therefore, we conduct a two-way ANOVA test to evaluate: 1. the effect of the choice of the anomaly detection model, and 2. the choice of the feature set on the relapse detection performance. The results are shown in Table 3.12. The obtained p-value, denoted as p_1, for the effect of the choice of feature set on the model's performance, is approximately 0, indicating that the selection of features

significantly affects the model's performance. This result suggests that different feature combinations have a notable impact on the accuracy of relapse detection. Similarly, the p-value p_2 obtained for the effect of the choice of the model on relapse detection accuracy is also approximately 0. This finding indicates that the choice of the model significantly influences the overall accuracy of relapse detection. Different models exhibit varying levels of performance in detecting relapses among patients with psychotic disorders. Lastly, the p-value p_3 tests the interaction between the choice of model and the choice of feature set. The value of p_3 reveals that there is a significant combined effect of the model choice and the choice of the selected feature set.

TABLE 3.12: *2way-Anova Results for global models results*

p_1	p_2	p_3
0.000	0.0009	0.0235

3.4.3.2 PERSONALIZED SCHEME

In this subsection, we present the results of the personalized scheme. Similar to the global scheme, we test different combinations of features, different anomaly detection methods, and different TMW for feature extraction.

1. **Exhaustive search**
   In Table 3.13, we present the best-performing model for each patient irrespective of the feature set and TMW for feature extraction. Notably, the results demonstrate that the optimal modalities and feature combinations and TMW vary across different patients, highlighting the importance of personalized approaches in relapse detection. The best-performing model is patient 1's model with a mean performance score (mean of ROC AUC and PR AUC) of 0.938. Conversely, the worst performing model is patient 3's model with a score of 0.727. The obtained results demonstrate notably superior performance compared to those derived from the exhaustive search within the global scheme for all patients in Table 3.8. This outcome signifies a positive stride toward the necessity of personalized models for relapse prediction. However, to draw more definitive conclusions, additional assessments on the testing set are required which will be provided at the end of the study. Furthermore, LOF and Isolation Forest, which demonstrated prominence in the visual distraction detection results outlined in Section 4.4.3, also emerge as prominent models in this context. We also observed that features related to sleep and heart rate are consistently prevalent in the selected best-performing configurations.

TABLE 3.13: *Best personalized model performance for each patient*

Patient	Period	Model	Featureset	ROC AUC	PR AUC	Average
1	1hr	Isolation Forest	sleep, %WW, acc, gyr, steps	0.953	0.924	0.938
2	24hrs	LOF	sleep, acc, gyr	0.698	0.853	0.776
3	24hrs	LOF	sleep	0.752	0.702	0.727
4	1hr	Isolation Forest	%WW, HR	0.866	0.849	0.857
5	5min	AE	sleep	0.833	0.738	0.786
6	4hrs	LOF	%WW, HR, acc, gyr, steps	0.785	0.749	0.767
7	5min	LOF	HR, steps	0.74	0.721	0.73
8	4hrs	AE	Sleep, %WW, HR, acc,gyr	0.769	0.911	0.84
9	1hr	OCSVM	%WW	0.949	0.867	0.908
10	4hrs	Elliptical Envelope	sleep, %WW	0.743	0.896	0.82

2. **Sensor selection**
   To further analyze the performance of unsupervised learning approaches in personalized relapse detection, we evaluated the average performance of different feature combinations for each patient across all models and all TMW. In Table 3.14, we present the top five performing feature sets over 31 combinations in terms of the average of ROC AUC and PR AUC for each patient. As in Table 3.13, the results of Table 3.14 show that the best feature set and best model differ from one patient to another. However, it is worth noting that some features appear in the best-performing feature sets for most of the patients. For instance, '%WW' is part of the best-performing feature sets for all patients. 'sleep' appears in the best-performing feature sets for all patients, except patient 6. 'HR' features in the best sets of all patients, except patient 3. 'Sleep', 'HR', and '%*WW*' were among the top features in the global model, showcasing their robust generalization. 'acc_gyr' appears in all sets except for those of patients ***5*** and ***9***. In contrast, 'steps' only features for patients 3 and ***5***.

TABLE 3.14: *Top five performing features for each patient*

Patient	featureset	ROC AUC	PR AUC	Average
1	sleep, HR, acc,gyr	0.706	0.417	0.561
	sleep,%WW, HR, acc,gyr	0.703	0.419	0.561
	sleep, acc,gyr	0.69	0.425	0.558
	sleep, HR	0.71	0.401	0.556
	sleep, %WW, acc,gyr	0.683	0.424	0.553
2	sleep, %WW, acc,gyr	0.55	0.675	0.613
	%WW, acc,gyr	0.543	0.668	0.605
	sleep, %WW, HR, acc,gyr	0.536	0.662	0.599

continues on the next page

Table 3.14 – Continuation

Patient	featureset	ROC AUC	PR AUC	Average
	acc,gyr	0.532	0.662	0.597
	%WW	0.525	0.667	0.596
3	sleep, %WW, acc,gyr	0.608	0.497	0.553
	%WW, acc,gyr	0.594	0.486	0.54
	acc,gyr	0.584	0.486	0.535
	sleep, acc,gyr	0.588	0.472	0.53
	acc,gyr, steps	0.581	0.469	0.525
4	sleep, %WW, HR	0.704	0.667	0.685
	%WW, HR, acc,gyr	0.681	0.659	0.67
	sleep, %WW, HR, acc,gyr	0.68	0.661	0.67
	%WW, HR	0.683	0.653	0.668
	sleep, HR, acc,gyr	0.674	0.658	0.666
5	%WW	0.419	0.388	0.403
	sleep, %WW	0.699	0.31	0.504
	sleep	0.656	0.275	0.466
	%WW	0.64	0.281	0.46
	sleep, %WW, HR	0.641	0.262	0.452
6	acc,gyr, steps	0.574	0.495	0.535
	acc,gyr	0.576	0.481	0.528
	steps	0.551	0.502	0.527
	%WW, HR, acc,gyr, steps	0.562	0.481	0.522
	%WW, acc,gyr, steps	0.558	0.485	0.522
7	Sleep	0.556	0.231	0.394
	sleep, %WW	0.513	0.217	0.365
	sleep, %WW, acc,gyr	0.431	0.194	0.313
	%WW	0.416	0.179	0.297
	sleep, %WW,HR	0.411	0.166	0.288
8	sleep, %WW,HR	0.654	0.759	0.707
	sleep, %WW, HR, acc,gyr	0.633	0.755	0.694
	sleep, HR	0.634	0.748	0.691
	sleep, %WW	0.622	0.754	0.688

continues on the next page

Table 3.14 – Continuation

Patient	featureset	ROC AUC	PR AUC	Average
	sleep, HR, acc,gyr	0.625	0.747	0.686
9	%WW	0.633	0.45	0.542
	sleep, %WW	0.556	0.325	0.441
	sleep, %WW,HR	0.532	0.305	0.418
	%WW,HR	0.516	0.309	0.413
	sleep	0.491	0.313	0.402
10	sleep	0.631	0.753	0.692
	sleep, %WW	0.594	0.748	0.671
	Sleep, acc,gyr	0.567	0.725	0.646
	sleep, %WW, acc,gyr	0.56	0.731	0.646
	sleep, HR	0.562	0.724	0.643

3. **Model selection**
   In figure 3.10, we show the average of ROC AUC and PR AUC of each model across all feature sets and possible TMW. Unlike the global scheme where models perform similarly on all patients' data, in the personalized scheme, models show significant differences in their performance on each patient's data. The results further highlight the importance of individualized models across different patients for relapse prediction.

4. **TMW selection**
   In Figure 3.11, the average of ROC AUC and PR AUC performance of each time window (TMW) for feature extraction across all feature sets and models are presented for each patient. Unlike the global scheme results in Table 3.11, where different TMWs showed similar averaged performance, the personalized scheme exhibits significant differences. For instance, for patient 1, a 5-minute TMW yields an average of 0.46, while a 4-hour TMW yields 0.56, a noticeable variation across different time windows for each patient. This discrepancy suggests that the detection of temporal patterns may vary for different time periods among patients. Furthermore, the absence of a universal TMW for all patients might explain why, in the global scheme results, all TMWs exhibited similar performance levels. Additionally, we notice a distinct trend in the scores—they progressively increase until reaching the optimal TMW and subsequently decrease. This consistent trend strongly suggests a tangible influence of the time window on the personalized models.

5. **Sensor and model selection effect significance**

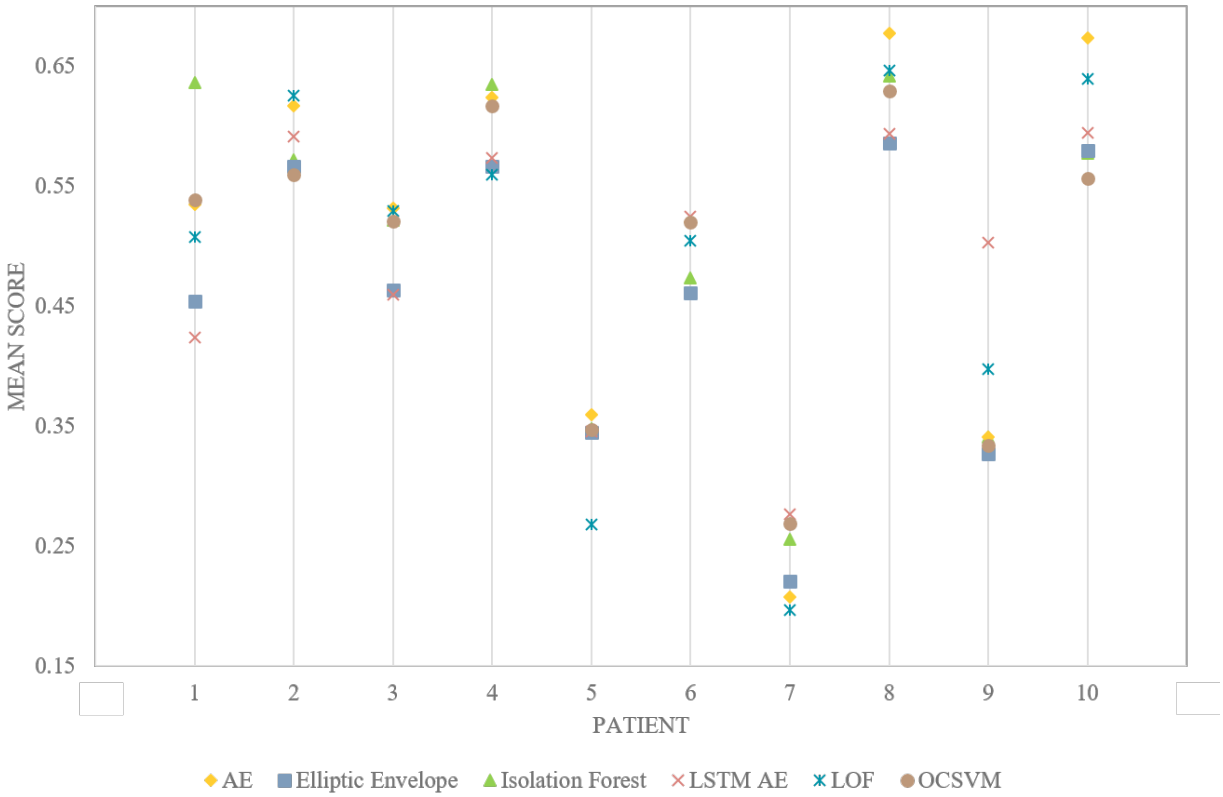

FIGURE 3.10: *Personalized scheme performance average across all models on the validation set*

TABLE 3.15: *2-way ANOVA on personalized models results*

Patient	p_1	p_2	p_3
1	0.000	0.000	0.8446
2	0.0000	0.0000	0.0118
3	0.0000	0.0001	0.1320
4	0.0000	0.0000	0.1578
5	0.0000	0.0000	0.0035
6	0.0000	0.0000	0.3394
7	0.0000	0.0000	0.0038
8	0.0000	0.0000	0.4908
9	0.0000	0.0024	0.4159
10	0.0000	0.0000	0.2721

We also conducted a two-way ANOVA test to evaluate the effect of the choice of the anomaly detection model and the choice of the features on the relapse detection performance on personalized models. The results are shown in Table 3.15. The value p_1 presents the p-value for the null hypothesis that the group means of each feature set are not significantly different. For all patients, $p_1 \sim 0$ which suggests that the choice of the feature set has a significant effect on the performance of the model. Similarly, the value p_2 presents the p-value for the null hypothesis that the group means of the models are not significantly different. For all patients, $p_2 \sim 0$, which implies that the choice

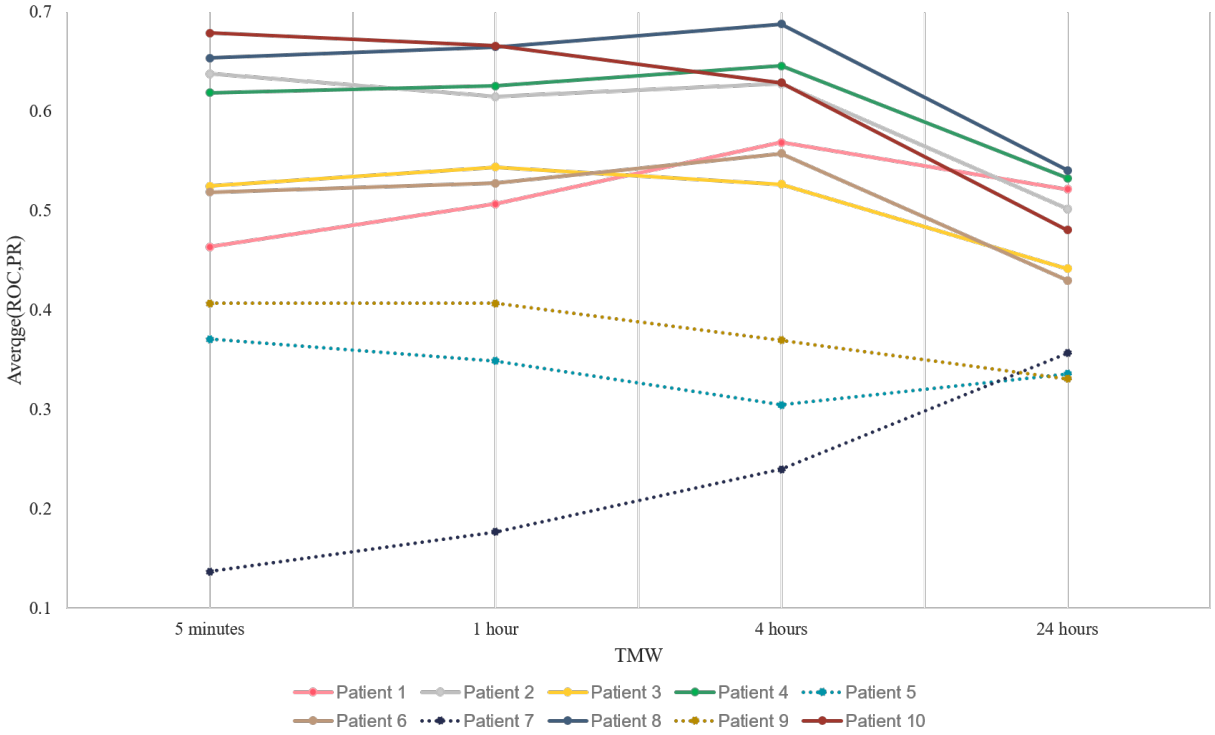

FIGURE 3.11: *Personalized scheme performance for each period per patient*

of the model has a significant effect on the relapse detection accuracy. Finally, p_3 value assesses whether there is an interaction between the choice of the model and the choice of the feature set. According to the ANOVA test, such interaction is only evident for patients 2, ***5***, and ***7***.

3.4.3.3 Final evaluation on testing dataset

Table 3.16: *Model performance for each patient on testing dataset*

Patient	ROC AUC			PR AUC		
	All features	**Best features**	**Global model**	**All features**	**Best features**	**Global model**
1	0.706	0.81	0.703	0.504	0.698	0.483
2	0.518	0.439	0.561	0.744	0.677	0.761
3	0.512	0.559	0.654	0.347	0.413	0.467
4	0.683	0.773	0.538	0.661	0.754	0.59
5	0.804	0.783	0.37	0.499	0.521	0.145
6	0.49	0.524	0.659	0.466	0.527	0.569
7	0.4	0.496	0.526	0.157	0.275	0.193
8	0.548	0.506	0.427	0.81	0.79	0.734
9	0.357	0.536	0.446	0.209	0.303	0.24
10	0.479	0.539	0.476	0.767	0.797	0.756
AVERAGE	0.549	0.596	0.536	0.516	0.575	0.493
AVERAGE -Patients5,7,9	0.558	0.592	0.574	0.613	0.658	0.617

Furthermore, we assess the performance of both the optimal global model and the optimal personalized model for each patient. These models are determined through an exhaustive search on the validation dataset. Additionally, we evaluate the performance of the best model achieved via an exhaustive search, considering all available features. We calculated the averaged performance also without the patients with very low number of patients as well. As shown in Table 3.16, on average, employing the global model yields the least favorable performance, while utilizing the personalized model leads to the best results. This observation underscores the superiority of personalized models in comparison to the global approach in terms of relapse detection accuracy. Moreover, the notable performance difference between personalized models employing all features (scoring 0.549 in ROC AUC and 0.516 PR AUC) compared to those utilizing only the best features (scoring 0.596 in ROC AUC and 0.575 in PR AUC) distinctly highlights the substantial influence of feature selection in improving the model´s robustness and generalization capacity which will be further explored in the next chapter.

3.5 Conclusion

In this chapter, We considered real-life applications where our aim is to identify undesirable events, inherently rarer than positive patterns. For this purpose, we proposed leveraging anomaly detection as a solution and conducted an evaluation to assess its applicability. To achieve this, we selected two pivotal applications within the affective computing domain: driver monitoring and relapse detection for patients with mental disorders. Both applications target predicting rare states whose data collection is not feasible and can benefit from anomaly detec-

tion. For each application we selected a suitable database, strategically chosen to demonstrate the effectiveness of our proposed approach across diverse contexts. Our objectives encompassed three key aspects: first, the validation of the proposed approach's efficiency in distinct applications; second, the comparison of various anomaly detection methods; and third, the assessment of multiple strategies involving feature selection, supervised and unsupervised techniques for driving, and generalized or personalized approaches for relapse prediction.

In the first study, we focused on distraction detection in the context of road safety. By employing unsupervised anomaly detection approaches instead of traditional supervised ML methods, we overcame the challenges associated with collecting distracted driving data, which can be dangerous. Using a database obtained from a driving simulator, we trained our models on non-distracted driving examples and evaluated their performance on distracted driving examples. Our findings demonstrated the efficiency of unsupervised models, with Isolation Forest emerging as the best model for distraction detection. Additionally, we compared the performance of unsupervised methods to traditional supervised models, highlighting the superiority of our proposed approach for imbalanced datasets or even in scenarios where no samples of the "anomalous" class were available

In the second study, we investigated the effectiveness of unsupervised learning approaches for relapse detection in psychotic disorders. Our results indicated that the Isolation Forest and the AE anomaly detection method exhibited the highest performance in the global scheme, where a single model was trained on data from all patients. Notably, we discovered that the optimal modalities and feature combinations varied across different patients, emphasizing the significance of personalized approaches in relapse detection. Moreover, we found that the choice of anomaly detection model and features significantly impacted the accuracy of relapse detection. By adopting a personalized approach, we achieved a more tailored and individualized detection method, leading to substantial improvements in relapse detection. This study underscores the importance of personalized models for accurate and effective relapse monitoring.
Overall, our research contributions highlight the efficiency and efficacy of unsupervised anomaly detection methods in behavior monitoring tasks. By circumventing the limitations of traditional supervised approaches and leveraging the power of anomaly detection, we have demonstrated advancements in both distraction detection and relapse detection. An intriguing finding from the personalized scheme results underscores the significance of feature selection in determining performance. In the upcoming chapter, we will further explore feature selection using the same datasets for detecting rare mental states such as visual distraction detection and psychotic relapse prediction.

CHAPTER 4

Explainability and Feature Selection Using Anomaly Scores

In this chapter, we propose the use of AE-based reconstruction errors for feature selection and explainability in the context of detecting rare mental states. We explore the use of this idea in two distinct scenarios: "visual distraction detection" and "psychotic relapse prediction".

Chapter contents

4.1 Introduction

In the previous chapter, we tackled the challenge of detecting rare states through an unsupervised approach, bypassing the issue of imbalanced data. We examined two real-life applications in affective computing: visual distraction detection and psychotic relapse prediction. When constructing models for real-world applications, identifying the crucial features and sensors for data collection stands as a critical step. For instance, collecting signals related to heart rate variability solely for detecting distracted driving might incur unnecessary costs and computational demands. Hence, feature selection plays a pivotal role in efficiently gathering data. Feature selection not only enhances performance and reduces computational costs but also provides valuable insights into the decision-making processes of ML models [329]. Unfortunately, many existing feature selection methods often overlook the critical issue of class imbalance, resulting in suboptimal performance for the detection of the minority class [330][331]. Paradoxically, the minority class often holds greater significance in various applications [332] [333]. This underscores the need for its careful recognition and accurate classification. For instance, consider the context of fraud detection, where it is of utmost importance to detect a rare malicious transaction within a larger population of normal transactions. Moreover, in certain domains like medicine, the ability to identify discriminative traits or features distinguishing the majority class (typically representing normal or healthy instances) from the minority class (typically denoting disease or anomaly) is of utmost importance. For instance, it can be valuable in comprehending their origin or enabling early intervention by tracking these features across specific targeted or at-risk populations.

In this chapter, we extend the work from Chapter 3 by introducing a feature selection method that assesses the importance of features based on anomaly detection methods. Our proposed approach involves training an AE solely on normal data—undistracted driving for visual distraction detection and non-relapse data for psychotic relapse prediction. Subsequently, we evaluate the correlation scores between annotations and the reconstruction error of each feature within a dataset containing both normal and abnormal instances (i.e., distracted and relapse data). These correlation scores serve as indicators of feature importance, enabling us to rank the features. Finally, we conduct an evaluation of the feature selection by assessing model performance using the top-ranked and lowest-ranked features.

This chapter also tackles the issue of explainability. Humans have continuously endeavored to seek explanations to grasp and interpret their environment [334]. This pursuit extends to understanding black box decisions made by ML systems. It's crucial to comprehend the rationale behind a model's predictions to guarantee its reliability and safety when used in real-world applications. This understanding aids in detecting biases and enhancing trust [335]. Hence, we discuss how our proposed method contributes to enhancing explainability for affective computing models. We apply this method to two distinct applications introduced in the previous chapter, both facing imbalanced data settings. These applications are well-suited for this approach due to the challenges associated with data scarcity—either due to the risks involved in data collection or the infrequent occurrence of the event in focus. Additionally, both applications stand to gain from an explainable approach. Lastly, we engage in two tasks with differing levels of complexity: one yielding easily interpretable results and the other showcasing the method's performance in handling more intricate tasks. The key contributions of this chapter can be outlined as follows:

- Introduction of a novel feature selection technique designed specifically for unbalanced datasets, leveraging the power of AE.
- In-depth analysis of feature influence for visual distraction detection and psychotic relapse prediction personalized for each patient.

The remainder of the chapter follows this structure: In Section 4.2, we introduce various known types of feature selection methods. Section 4.3 details our proposed method specifically tailored

for imbalanced datasets. This method is examined within the context of two distinct applications outlined in Chapter 3: Visual Distraction Detection and Psychotic Relapse Prediction. Section 4.4 shows into the impact of the feature selection method applied for driver distraction detection across three tasks: anomaly detection, classification, and regression, with results presented in Section 4.4.3. Following that, we present the results obtained regarding psychotic relapse prediction in Section 4.5. Additionally, in Section 4.6, we compare our proposed method with a similar approach from the literature. Furthermore, we discuss how our approach contributes to explainability in Section 4.7. Finally, we conclude our chapter in Section 4.8.

4.2 Feature selection methods

Feature selection is a key process within ML and data analysis that involves choosing a subset of relevant features (also known as variables or attributes) from a larger set of available features in a dataset. The goal of feature selection is to improve the performance and efficiency of a ML model by focusing only on the most important and informative features while discarding or ignoring less pertinent or redundant ones. This holds particular importance with smaller databases, aiming to mitigate model complexity and prevent overfitting issues.

4.2.1 Feature selection methods categories

Feature selection methods can be broadly categorized into three types, which we present in Figure 4.1.

- **Filter methods:** Filter methods [336] evaluate the relevance of features using statistical measures or scores, often independently of the ML algorithm to be used as shown in Figure 4.1. Various assessment criteria have been introduced for filter methods. Some key criteria include the feature's discriminative capability in separating samples [337], mutual information [338], and feature correlation [339].
- **Wrapper methods** Wrapper methods [340] assess the performance of a ML model using different subsets of features. These methods involve training and evaluating the model multiple times with different feature combinations. However, these methods are generally computationally expensive, and their convergence to a global optimum is not guaranteed. Recursive Feature Elimination (RFE) [341] and forward/backward selection [342] are examples of wrapper methods. Another example involves exhaustive feature selection, where they perform a brute-force assessment of feature subsets. The ideal subset is chosen by optimizing a designated performance metric while employing any given regressor or classifier.
- **Embedded methods** Embedded methods [343] integrate feature selection into the process of constructing the model. Techniques like $L_{2,1}$-norm regularization [344] and tree-based feature importance [345] are embedded methods, as they naturally select relevant features during the training process.

Each method carries its own set of advantages and drawbacks. Wrapper methods, although effective, suffer from high computational cost, restricting their practical application [347]. On the other hand, filter methods are computationally more efficient, assign a score for each feature, and are model-independent but tend to overlook feature interdependencies. Embedded methods strike a balance between filter and wrapper approaches by integrating feature selection into model learning. However, they still maintain a dependence on the model employed and often involve hyperparameters that require tuning to achieve optimal performance.

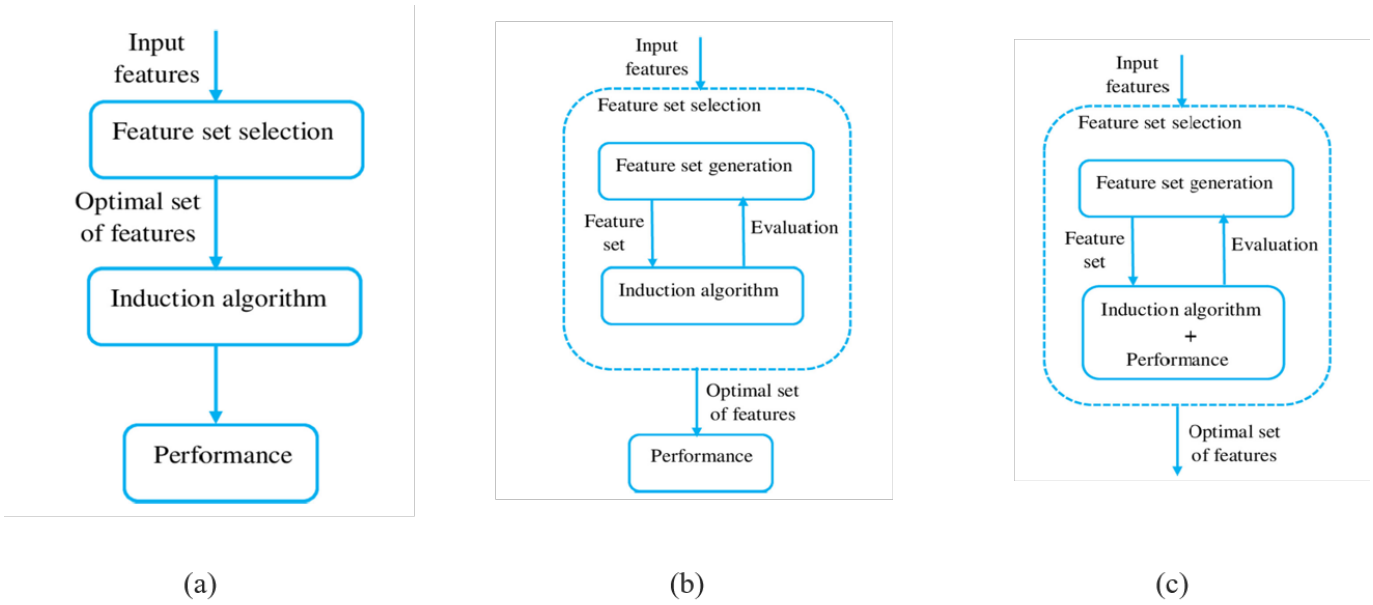

FIGURE 4.1: *General Framework of the process of (a): filter method, (b): wrapper method, and (c): embedded method. Adapted from [346].*

4.2.2 FEATURE SELECTION FOR IMBALANCED DATASETS

Data imbalance remains a persistent concern in feature selection research [346]. Certain research findings indicate that employing conventional feature selection methods without accounting for class imbalance can result in a performance decline. Yin et al. [330] conducted a study demonstrating that feature selection has the potential to amplify the overlap between distributions of different classes. This amplification is traced back to the significant bias towards the majority class, which subsequently hampers the efficiency of classification tasks. Despite the widespread domain of feature selection, only a limited amount of literature focuses on addressing data imbalance [348]. Some approaches have been developed to tackle feature selection challenges within unbalanced data scenarios. For wrapper methods, researchers have explored metrics less sensitive to class imbalance like the ROC AUC [349], F-measure[350] [351], and balanced loss function[352]. Embedded methods, on the other hand, handle this issue by incorporating a regularization term [344].

Our contribution lies in providing a solution to the feature selection within imbalanced datasets by leveraging anomaly detection techniques. Specifically, we employ the reconstruction error generated by an AE trained exclusively on normal data (majority class), thus reducing reliance on minority class data. This method, similar to filter methods, assigns scores to each feature. However, it is crucial to note that all features are interrelated within the autoencoder architecture, which can be used as the classifier as well.

4.3 PROPOSED METHOD BASED ON ANOMALY DETECTION

In Chapter 2, we explained the utilization of AE for anomaly detection, primarily by training them solely on normal data. During the test phase, when a new sample is presented, any instance significantly divergent from the learned normal data pattern results in the AE's inability to reconstruct the input effectively. This inability manifests as a higher reconstruction error across all features, subsequently serving as an anomaly score. In our study, we operate under the assumption that the AE struggles to accurately reconstruct features crucial for distinguishing between the normal and abnormal classes. Our aim is to leverage this discrepancy in reconstruction errors to discern the relative importance of these features.

Our method focuses on an imbalanced setup where the minority class is notably smaller than the majority class. The training dataset consists of (input, target) pairs $\{(x_1, y_1), \ldots, (x_N, y_N)\}$ where y_i is the target (discrete values of 1 and 0 or a continuous variable) and $X \in \mathbb{R}^{P \times N}$ is the input matrix and we have N_0 examples from class 0 and N_1 from class 1 and $N_1 \ll N_0$, with the objective of distinguishing between two conditional multivariate probability distributions, denoted as P_0 and P_1. A critical aspect of this method is the identification of a feature set F that enables the differentiation between these distributions, with $|F| < P$.
To achieve this, an AE is used in two phases presented in 4.2.

- **Phase 1:** Similar to the AE-based anomaly detection method, the data points are divided into training D_1 and feature selection D_2 datasets, where the training set includes N' samples from the majority class only (class 0), and D_2 has N'' examples from majority and minority class. This enables, in the training phase, unsupervised training on the overrepresented majority class.
- **Phase 2:** After training the AE, we compute the RE of the samples of set D_2. Following this, we compute the correlation score between the reconstruction error of each feature and the corresponding annotation. These correlation scores serve as an indicator of feature importance, where higher scores indicate greater importance. The selection of discriminating features is achieved by applying a threshold.

This feature selection method is designed specifically for imbalanced datasets. During the training phase, the autoencoder leverages the available data from the majority class to train. In contrast, the feature selection step using simple correlation doesn't demand an extensive number of examples.

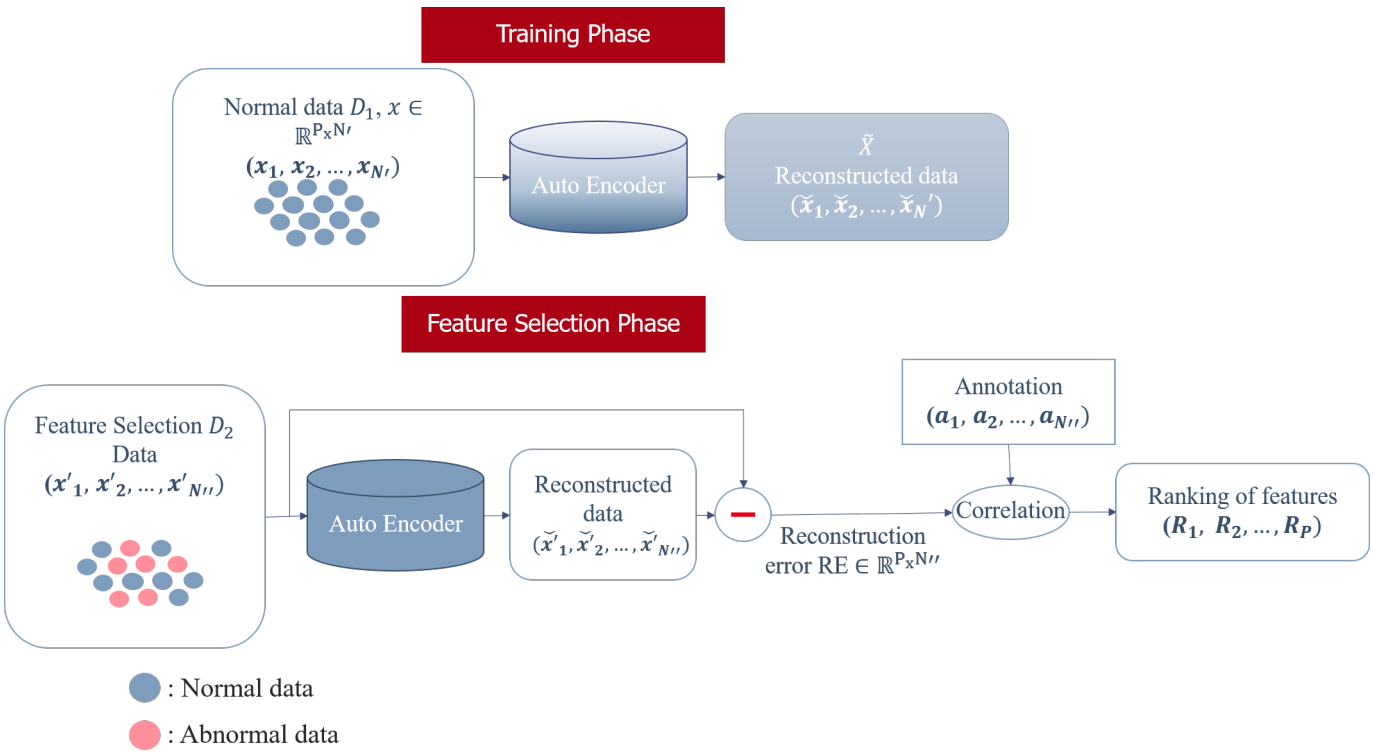

FIGURE 4.2: *Our proposed feature selection approach using AE.*

4.4 Feature selection for driver distraction detection

As previously mentioned, gathering instances of distracted driving in real-world scenarios poses significant risks. Hence, we require a specialized feature selection method suitable for imbalanced data settings. We evaluate our approach specifically on the HADRIAN subset dedicated to visual distraction detection, aiming to discern the significance of features employed in this context. The specifics of the dataset are outlined in Chapter 3. Our assessment of feature selection encompasses three distinct tasks: classification, regression, and anomaly detection, allowing us to explore its impact across varied domains.

4.4.1 Data

We use the dataset introduced in Chapter 3 for visual distraction detection of autonomous driving mode.
In this chapter, the data partitioning differs from that in Chapter 3 due to several reasons, including the execution of multiple tasks beyond anomaly detection. Additionally, this data requires labeled samples from the anomalous class for feature rankings, necessitating a separate dataset that was not previously seen during the training of the autoencoder or used in calculating the correlation scores for evaluating the estimation of the level of distraction. To thoroughly evaluate the performance of our proposed feature selection method, we divided the available data into three distinct subsets: D_1, D_2, and D_3, each serving different purposes within our study.

- D_1 comprises data exclusively from non-distracted driving scenarios, totaling 211 examples from the control scenario. This subset is primarily utilized for fitting the AE model, which will subsequently be employed for feature selection.
- D_2 encompasses a more diverse dataset, including both non-distracted and distracted driving data from the distraction scenario, comprising a total of 235 examples. It serves multiple purposes within our analysis. It is used for calculating correlation scores and conducting feature selection. Within this set, 84 examples fall within the category of distractions with a severity level of less than or equal to 25, 66 examples pertain to distractions with severity levels between 25 and 75, and the remaining 85 examples correspond to distractions with a severity level exceeding 75. Additionally, D_2 is employed for training supervised models in our study. Furthermore, it is divided into the same distraction severity categories as previously mentioned.

- D_3 dataset mirrors the composition of D_2, containing a total of 235 examples from both non-distracted and distracted driving scenarios. This subset serves as the evaluation dataset for testing the performance of our models.

In our feature selection process, we have incorporated an expanded set of features compared to the ones utilized in Chapter 3. The selected features for this analysis encompass 'Looking at road', 'Saccade magnitude', 'Saccade rate', 'Saccade peak velocity', 'Eye position entropy', 'Gaze heading mean', 'Gaze heading standard deviation', 'Gaze pitch mean', 'Gaze pitch standard deviation', 'Head heading mean', 'Head heading standard deviation', 'Head pitch mean', 'Head pitch standard deviation', 'ECG Interbeat Intervals (IBI)', and 'BVP IBI'. We have expanded the feature set to facilitate a more comprehensive evaluation of our feature selection method. This augmented feature set has been carefully curated to include features that are known to be crucial for distraction detection including 'Looking at road', as well as less influential features for visual distraction detection, including 'ECG IBI' and 'BVP IBI'. In order to compute the "Looking at road" feature, we define a plane that covers the windshield surface. Then, we use the gaze

estimation to determine if the gaze vector intersects with the plane. If true, we consider that the driver is looking at the road. Else, the driver is not looking at the road.

4.4.2 Experiments

In this section, we present the series of experiments conducted. Initially, we perform feature selection for visual distraction detection, followed by an evaluation of our outcomes across three tasks: classification, anomaly detection, and regression.

- **Feature Selection:** To obtain feature rankings, we employ an AE model consisting of three dense layers with 4, 2, and 4 neurons, respectively. This is the same architecture that was used in Chapter 3. The AE is trained exclusively using the D_1 dataset, which contains non-distracted driving data. To obtain feature scores, we feed the model with data from D_2 and subsequently calculate the correlation scores.
- **Classification:** To assess the impact of the feature selection method on the classification task, which aims to distinguish between two classes (Class 0 representing non-distracted driving and Class 1 representing distracted driving), we employ C-Support Vector Classification implemented using the sklearn library. Our model is trained using examples from the D_2 subset. Subsequently, the model's performance is evaluated using the D_3 subset. Examples with a distraction level of <=50% represent "Class 0" and examples with a distraction level greater than 50% represent "Class 1". This approach allows us to determine how effectively the selected features discriminate between the most significant distraction levels in our classification task.
- **Anomaly Detection:**

  Regarding the anomaly detection task, we employ a One-Class SVM approach. During the training phase, the model is exclusively trained using normal data, which corresponds to scenarios with a distraction level of <=50%, derived from the D_2 subset. Subsequently, for testing, the model is evaluated using D_3 subset, where examples with a distraction level <=50% are considered normal samples, and examples with a distraction level greater than 50% are treated as abnormal samples.
- **Regression:**

  For the regression task, we opt for Epsilon-Support Vector Regression as our modeling approach. The model is trained using all of the examples in the D_2, which includes both non-distracted and distracted driving scenarios. Subsequently, we evaluate the model's performance using the D_3 subset. This regression task aims to predict and assess the accuracy of distraction level estimates based on the selected features, encompassing a wide range of distraction levels.

4.4.3 Results

In this section, we present the results of our evaluation of the proposed feature selection method across three model types: binary classification, anomaly detection, and regression models. We investigate two distinct approaches: "Best to worst" and "worst to best," where features were added to the models incrementally based on their importance rankings. Our evaluation considered a range of feature subset sizes, from 1 to 15 features. In the "best to worst" approach, we systematically add features to the model, beginning with the feature having the highest importance ranking. Conversely, in the "worst to best" approach, we initiate the model with the feature that had the lowest importance ranking or correlation score.

- **Feature Ranking:**
  In Table 4.1, we present the obtained correlation scores between each feature and our tar-

TABLE 4.1: *Feature ranking using our proposed feature selection method.*

Feature Name	Correlation Score
Looking at road	0.64
Gaze pitch mean	0.54
Head pitch standard deviation	0.47
Gaze pitch standard deviation	0.40
Head pitch mean	0.38
BVP IBI	0.33
ECG IBI	0.33
Saccade rate	0.31
Eye position entropy	0.28
Head heading mean	0.28
Head heading standard deviation	0.23
Saccade magnitude	0.18
Saccade peak velocity	0.17
Gaze heading mean	0.16
Gaze heading standard deviation	0.14

get variable 'distraction level'. The feature 'Looking at road', which we added due to its relevance to distraction, emerges as the most crucial feature with a high correlation score of 0.64, suggesting its strong positive association with distraction levels. Features such as 'Gaze pitch mean', 'Head pitch standard deviation', and 'Gaze pitch standard deviation' also exhibit notable correlations, highlighting their relevance. Whereas the features 'ECG IBI' and 'BVP IBI', which are not relevant to visual distraction obtain 0.33, significantly less than the feature "Looking at road". We use these scores in the best-to-worst and worst-to-best approaches, as described before in the following experiments, to evaluate the effectiveness of our feature selection method.

- **Classification Task:** In Figure 4.3, we present the outcomes of our feature selection approach evaluation using a binary classifier. In "best to worst", we initially start with the most influential feature, "Looking at road". Consequently, the classifier has a high classification accuracy of 83.4% using just this single feature. As we continue to add features in descending order of importance, the accuracy of the classifier slightly improves and reaches its peak at 12 features with an accuracy of 86.38% . The little improvement in performance (3%) indicates that the remaining features bring little relevant information in comparison to solely using the best feature. This strategy allows us to leverage the most critical features right from the beginning, potentially leading to a fast and effective classification model. In contrast, we start with 'Gaze heading standard deviation' in "worst to best". As a result, the initial classification accuracy was low. This observation is significant as it highlights that the initial feature lacked discriminatory power for binary classification, with an accuracy of 51.91%, closely resembling random guessing's 50% accuracy rate, which confirms that a low correlation between reconstruction error of the feature and the annotations indicates a

less relevant feature. However, as we gradually incorporate features in ascending order of importance, the classification accuracy improves progressively. This strategy is informative as it demonstrates the challenges posed by less influential features initially and highlights the gradual improvement that can be achieved by adding more important features. Both strategies offer valuable insights into how the order of feature inclusion impacts the performance of a binary classifier. "Best to Worst" highlights the potential benefits of leveraging highly influential features early on, while "Worst to Best" underscores the initial difficulty of using less important features and their limited impact on classification accuracy. Moreover, the results reveal that the rankings generated by the feature selection method indeed reflect the discriminatory capability of each feature. The highest-ranked feature achieved an accuracy of 83.4%, whereas the lowest-ranked feature yielded an accuracy of 51.91%.

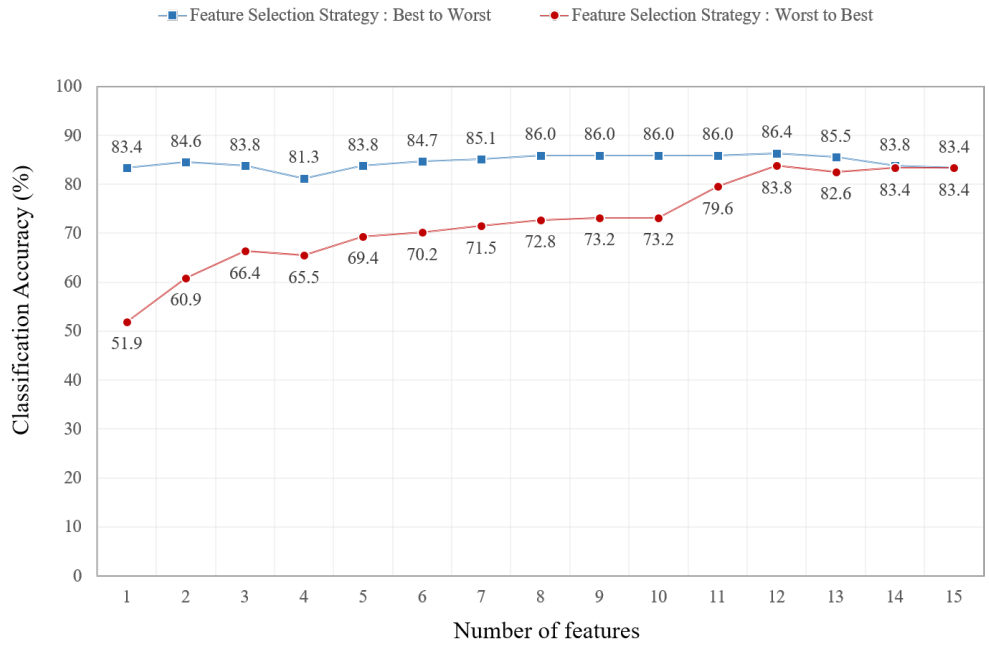

FIGURE 4.3: *Distraction detection classification performance using best to worst and worst to best strategies.*

- **Anomaly Detection Task:** Moreover, we evaluate the performance of an anomaly detection model using ROC AUC and PR AUCmetrics for both the "best to worst" and "worst to best" feature selection strategies. Table 4.2 summarizes the results of this evaluation. For the "best to worst" strategy, we observed consistently high ROC AUC scores ranging from 0.918 to 0.931, indicating the model's ability to distinguish between normal (non-distracted) and anomalous (distracted) data effectively. PR AUCscores also remained consistently high, demonstrating the model's precision in identifying anomalies. The highest performance achieved on average was 0.936 with 10 features close to the performance achieved by the top feature. Conversely, the "worst to best" strategy showed contrasting results. Initially, the ROC AUC and PR AUC scores were notably lower 0.513 and 0.511 respectively, reflecting the challenge of beginning with less important features. However, as more crucial features were incorporated, we observed a gradual improvement in both ROC AUC and PR AUC scores. The highest average performance of 0.924 was achieved when utilizing all features, which was approximately achieved by the model's performance when using only the top-ranked feature. These results emphasize the critical role of feature selection in anomaly detection which leads to models with better generalization capabilities.
- **Regression Task:**

  In our regression task, we assessed the performance of our model using MSE and the coeffi-

TABLE 4.2: *Anomaly detection performance*

Number of Features	Best to Worst			Worst to Best		
	ROC AUC	PR AUC	Average	ROC AUC	PR AUC	Average
1	0.923	0.936	0.93	0.513	0.511	0.512
2	0.92	0.935	0.928	0.588	0.613	0.6
3	0.918	0.931	0.925	0.73	0.727	0.728
4	0.928	0.938	0.933	0.732	0.72	0.726
5	0.924	0.937	0.931	0.731	0.721	0.726
6	0.927	0.939	0.933	0.743	0.738	0.74
7	0.929	0.94	0.934	0.747	0.743	0.745
8	0.93	0.941	0.936	0.757	0.753	0.755
9	0.927	0.938	0.932	0.765	0.762	0.764
10	0.931	0.942	0.936	0.764	0.762	0.763
11	0.927	0.938	0.932	0.782	0.778	0.78
12	0.928	0.937	0.933	0.849	0.844	0.846
13	0.928	0.936	0.932	0.855	0.85	0.852
14	0.919	0.928	0.924	0.886	0.891	0.889
15	0.919	0.928	0.924	0.919	0.928	0.924

cient of determination (R^2) for both the "best to worst" and "worst to best" feature selection strategies, as summarized in Table 4.3. For the "best to worst" strategy, MSE and R^2 reached 0.073 and 0.602, respectively, starting from the first feature, and achieved their best performance using only two features, with an MSE of 0.067 and an R^2 of 0.633. Conversely, the "worst to best" strategy initially exhibited higher MSE and a lower R^2 value, indicating poorer predictive performance when less important features were introduced. The peak performance for the "worst to best" strategy was achieved when all features were used, resulting in an MSE of 0.082 and an R^2 of 0.552. These results reinforce the critical role of feature selection in regression tasks.

Throughout the analysis of all tasks, the performance of the highest ranked feature ¨Looking at road¨ achieved high results. The inclusion of additional features contributes minimally to performance improvement. It strongly indicates that the feature "looking at the road" encapsulates the majority of critical information.

4.5 Feature selection for personalized psychotic relapse prediction

In the second application, we apply our method to predict psychotic relapses. Here, we experiment with employing the feature selection method to develop personalized models. Therefore, for each patient, we train an AE for feature selection and an anomaly detection model for relapse

TABLE 4.3: *Regression performance on the visual distraction dataset.*

Number of Features	Best to Worst		Worst to Best	
	MSE	R^2	MSE	R^2
1	0.073	0.602	0.214	-0.173
2	0.067	0.633	0.189	-0.035
3	0.079	0.57	0.15	0.18
4	0.07	0.614	0.146	0.202
5	0.068	0.628	0.144	0.211
6	0.071	0.609	0.139	0.24
7	0.072	0.605	0.133	0.271
8	0.073	0.6	0.124	0.32
9	0.074	0.594	0.122	0.333
10	0.074	0.594	0.124	0.324
11	0.073	0.599	0.115	0.372
12	0.073	0.599	0.09	0.506
13	0.072	0.604	0.095	0.477
14	0.084	0.543	0.091	0.502
15	0.082	0.552	0.082	0.552

prediction

4.5.1 DATA

In our study, we employ the dataset discussed in Chapter 3. However, we utilize subsets of this dataset differently for various purposes in our analysis. Therefore, we changed their names to avoid confusion about the purpose of their use. We follow the same data division into three subsets:

- D_1 is the training data used in Chapter 3. This set exclusively comprises non-relapse data. We use it for training the AE used for feature selection and relapse predictor model.
- D_2 is the validation dataset. It consists of both relapse and non-relapse days data. We use it primarily for the ranking features by computing the correlation scores.
- D_3 is the testing dataset. It is employed to evaluate the performance of the feature selection methods.

The table 3.7 in Chapter 3 shows the distribution of relapse and non-relapse days in the partitions.

4.5.2 EXPERIMENTS

We aim to evaluate feature selection methods for each patient. Consequently, for each patient, we execute the following steps:

- **AE for Feature Selection:**

For each patient, we employ a simple AE architecture featuring a single hidden layer comprising 4 neurons (similar to Chapter 3). During the model training phase, we utilize the Adam optimizer and implement the MSE as the loss function. The batch size is set to 16, with a maximum training duration of 1000 epochs. To enhance training efficiency and avoid overfitting, we implement an early stopping mechanism, using 20% of the training dataset for this purpose and setting a patience threshold of 10 epochs.

- **Anomaly Detection Models:**

  To assess the effectiveness of our feature selection methods, we conducted tests using anomaly detection models tailored to each patient's data. We use the anomaly detection models that were identified as the best-performing models through a grid search process, as detailed in Chapter 3. We present the selected models for each patient in Table 4.4.

TABLE 4.4: *Anomaly detection models for patients*

Patient	Anomaly Detection Model
1	Isolation Forest
2	LOF
3	LOF
4	Isolation Forest
5	AE
6	LOF
7	LOF
8	AE
9	OCSVM
10	Elliptical Envelope

- **Features:**

  As for the features, we use all the features used in Chapter 3 calculated over a 24-hour time window, which includes: the mean heart rate, the standard deviation of the heart rate, the norm of the accelerometer coordinates, the norm of the gyroscope coordinates, the percentage of sleeping time, and the total number of steps. To validate the efficacy of the proposed approach, we introduce a noise feature created by random sampling from a uniform distribution within the range [0,1]. This noise feature serves as a reference for features expected to have low scores and minimal relevance.

4.5.3 RESULTS

Table 4.5 presents feature data rankings for patients, in the context of detecting relapse in patients with psychotic disorders, from patient 1 (P1) through patient 10 (P10). Patient 1 emerges as distinctive, exhibiting the highest correlation score (0.75) for the "HR_mean" feature, implying its strong predictive power for this patient. Moreover, patient 4 has 3 features approximately around 0.4 while the rest of the patients have low correlation scores. In contrast, Patient 5 demonstrates predominantly zero correlations across all features, however, this patient had only 3 relapse days in the validation set. The results may indicate that for the available features, the distinction of relapse is easier for patients 1 and 4 and more difficult for the remaining patients

which is consistent with the results obtained in 3.16 except for patient 5 whose relapse days are very low. Furthermore, the results show that there is no feature that consistently demonstrates high efficiency in predicting relapse for all patients. All features exhibit negative correlations with relapse for at least one patient, indicating that none of the features are universally reliable for relapse prediction across the entire patient cohort. This observation underscores the individualized nature of psychotic disorder relapse prediction. Furthermore, when compared to the correlation scores observed in the visual distraction detection task, the scores are notably lower, indicating the heightened complexity of this particular task in contrast to distraction detection and the need for more relevant features for relapse prediction. Moreover, these scores highlight the varying degrees of difficulty across different patients within this task and the need for more relevant features that possess a higher discriminative power.

TABLE 4.5: *Feature scores for patients using our proposed approach.*

Features	Correlation Scores									
	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
sleep	0.70	0.10	-0.06	0.38	-0.01	**0.27**	0.07	**0.31**	-0.27	**0.15**
sleep_index	0.15	0.02	-0.07	0.20	-0.13	-0.02	0.06	0.17	-0.32	-0.07
HR_mean	**0.75**	-0.03	0.14	0.37	-0.00	0.14	0.01	0.28	-0.01	-0.03
HR_std	0.51	-0.05	0.11	**0.40**	-0.02	0.08	**0.21**	0.18	-0.27	0.01
acc	0.13	0.08	0.25	0.29	-0.20	-0.05	0.04	0.00	-0.11	0.00
gyr	0.52	**0.18**	-0.15	0.31	-0.16	-0.12	-0.13	0.24	-0.26	-0.06
steps	0.21	-0.06	**0.27**	0.33	0.00	-0.02	0.15	0.02	-0.17	-0.08
noise	0.12	0.03	0.03	-0.10	-0.05	0.00	-0.13	-0.14	0.06	-0.12

Table 4.6 presents the performance of anomaly detection models using two different strategies for feature selection across multiple patients. The table is divided into two sections: "Best to Worst" and "Worst to Best". The performance metrics evaluated are ROC AUC, PR AUC, and the total score. the analysis of the anomaly detection models' performance shows that for all patients the best-ranked feature consistently outperforms the worst-ranked feature. However, there is an exception for Patient 7, where the worst feature performs better in terms of ROC AUC and PR AUCscores. This can be attributed to the limitation of the small validation dataset for calculating correlation scores. Moreover, the best testing performances (highest total scores) are often achieved with fewer features than the total number of available features. However, for certain patients, the "Worst to Best" scenario leads to the optimal feature combination. This suggests that the strategy's effectiveness might be compromised due to the relatively small dataset and high complexity of the task.

TABLE 4.6: *Anomaly detection models performance using feature selection.*

Patient	Number of features	Best to Worst			Worst to Best		
		ROC AUC	PR AUC	Total	ROC AUC	PR AUC	Total
	1	0.7	0.571	**0.636**	0.513	0.252	0.382
continues on the next page							

Table 4.6 – Continuation

Patient	Number of features	Best to Worst			Worst to Best		
		ROC AUC	PR AUC	Total	ROC AUC	PR AUC	Total
	2	0.606	0.453	0.53	0.519	0.261	0.39
	3	0.629	0.464	0.546	0.426	0.263	0.344
	4	0.594	0.407	0.5	0.461	0.363	0.412
	5	0.677	0.504	0.59	0.577	0.384	0.48
	6	0.671	0.522	0.596	0.723	0.564	**0.644**
	7	0.661	0.44	0.55	0.561	0.339	0.45
	8	0.652	0.551	0.602	0.652	0.551	0.602
2	1	0.486	0.747	0.616	0.445	0.684	0.564
	2	0.423	0.661	0.542	0.619	0.816	0.718
	3	0.439	0.677	0.558	0.651	0.817	**0.734**
	4	0.394	0.644	0.519	0.617	0.784	0.7
	5	0.511	0.708	0.609	0.578	0.768	0.673
	6	0.431	0.682	0.556	0.518	0.735	0.626
	7	0.482	0.713	0.597	0.542	0.742	0.642
	8	0.55	0.752	**0.651**	0.55	0.752	0.651
3	1	0.584	0.437	**0.51**	0.405	0.296	0.351
	2	0.389	0.282	0.336	0.524	0.439	0.482
	3	0.462	0.315	0.388	0.562	0.465	**0.514**
	4	0.497	0.405	0.451	0.586	0.441	0.514
	5	0.429	0.41	0.42	0.473	0.328	0.4
	6	0.482	0.368	0.425	0.536	0.352	0.444
	7	0.462	0.338	0.4	0.393	0.289	0.341
	8	0.473	0.383	0.428	0.473	0.383	0.428
4	1	0.689	0.663	**0.676**	0.532	0.538	**0.535**
	2	0.683	0.623	0.653	0.485	0.49	0.488
	3	0.588	0.589	0.588	0.448	0.461	0.454
	4	0.594	0.573	0.583	0.485	0.516	0.5
	5	0.625	0.564	0.594	0.471	0.506	0.488
	6	0.583	0.53	0.556	0.501	0.48	0.49

continues on the next page

Table 4.6 – Continuation

Patient	Number of features	Best to Worst			Worst to Best		
		ROC AUC	PR AUC	Total	ROC AUC	PR AUC	Total
	7	0.594	0.526	0.56	0.476	0.46	0.468
	8	0.552	0.516	0.534	0.552	0.516	0.534
5	1	0.38	0.151	0.266	0.228	0.12	0.174
	2	0.554	0.195	**0.375**	0.185	0.116	0.15
	3	0.424	0.159	0.292	0.185	0.112	0.148
	4	0.467	0.172	0.32	0.457	0.158	**0.308**
	5	0.446	0.165	0.306	0.261	0.122	0.192
	6	0.37	0.14	0.255	0.217	0.116	0.166
	7	0.304	0.132	0.218	0.174	0.11	0.142
	8	0.315	0.134	0.224	0.359	0.14	0.25
6	1	0.666	0.634	**0.65**	0.477	0.451	0.464
	2	0.461	0.44	0.45	0.558	0.504	**0.531**
	3	0.445	0.402	0.424	0.448	0.41	0.429
	4	0.424	0.404	0.414	0.385	0.448	0.416
	5	0.533	0.462	0.498	0.473	0.474	0.474
	6	0.459	0.428	0.444	0.403	0.441	0.422
	7	0.481	0.442	0.462	0.495	0.472	0.484
	8	0.502	0.459	0.481	0.502	0.459	0.481
7	1	0.367	0.155	0.261	0.504	0.183	**0.344**
	2	0.53	0.179	0.355	0.43	0.153	0.292
	3	0.526	0.19	0.358	0.496	0.19	0.343
	4	0.57	0.504	**0.537**	0.356	0.162	0.259
	5	0.578	0.239	0.408	0.274	0.126	0.2
	6	0.467	0.176	0.322	0.348	0.16	0.254
	7	0.333	0.136	0.234	0.348	0.209	0.278
	8	0.267	0.127	0.197	0.267	0.127	0.197
8	1	0.473	0.752	0.612	0.441	0.75	0.596
	2	0.393	0.722	0.558	0.447	0.753	0.6
	3	0.4	0.726	0.563	0.534	0.804	0.669

continues on the next page

Table 4.6 – Continuation

Patient	Number of features	Best to Worst			Worst to Best		
		ROC AUC	PR AUC	Total	ROC AUC	PR AUC	Total
	4	0.401	0.726	0.564	0.539	0.81	**0.675**
	5	0.42	0.737	0.578	0.506	0.771	0.639
	6	0.385	0.724	0.554	0.486	0.771	0.629
	7	0.42	0.738	0.579	0.454	0.751	0.602
	8	0.475	0.751	**0.613**	0.396	0.731	0.564
9	1	0.571	0.297	0.434	0.304	0.195	0.25
	2	0.714	0.381	**0.548**	0.321	0.204	**0.262**
	3	0.589	0.287	0.438	0.179	0.169	0.174
	4	0.482	0.246	0.364	0.161	0.167	0.164
	5	0.411	0.224	0.318	0.125	0.162	0.144
	6	0.464	0.244	0.354	0.125	0.162	0.144
	7	0.321	0.199	0.26	0.161	0.168	0.164
	8	0.161	0.168	0.164	0.161	0.168	0.164
10	1	0.441	0.757	0.599	0.387	0.737	0.562
	2	0.523	0.772	0.648	0.456	0.746	0.601
	3	0.517	0.769	0.643	0.555	0.783	0.669
	4	0.507	0.775	0.641	0.518	0.77	0.644
	5	0.602	0.818	0.71	0.536	0.78	0.658
	6	0.619	0.837	**0.728**	0.57	0.801	0.686
	7	0.577	0.809	0.693	0.585	0.811	**0.698**
	8	0.557	0.805	0.681	0.557	0.805	0.681

4.6 Comparison with related work

Following the development of our method, we came across a similar technique designed for feature selection in imbalanced data settings. Massi et al. introduced a feature selection method specifically tailored to identifying significant features capable of effectively distinguishing between the minority and majority classes in highly imbalanced binary classification scenarios [353]. They use an ensemble of deep sparse AE to obtain the ranking of the features. They train an ensemble of B learners (autoencoders). Each learner is trained using a tailored sampling procedure, utilizing subsets of the data. The dataset for each learner is split into training and testing sets. The training sets comprise only the majority class, while the testing set consists of an equal distribution of majority and minority class data points. In contrast to our approach, in their method, they compute feature rankings(Δ) by generating RE matrices from test sets and calcu-

lating average REs per feature per class (minority and majority classes). This vector identifies the features that display noteworthy differences in REs between the majority and minority classes. Discriminative feature selection is accomplished by implementing a threshold $\delta \in \{0,1\}$. We implement this method in both applications (visual distraction detection and psychotic relapse prediction) and proceed to compare the outcomes between this approach and ours. Due to the limited size of our datasets, we made an adjustment in our evaluations by employing a single AE instead of multiple deep stacked AE as shown in Figure 4.4.

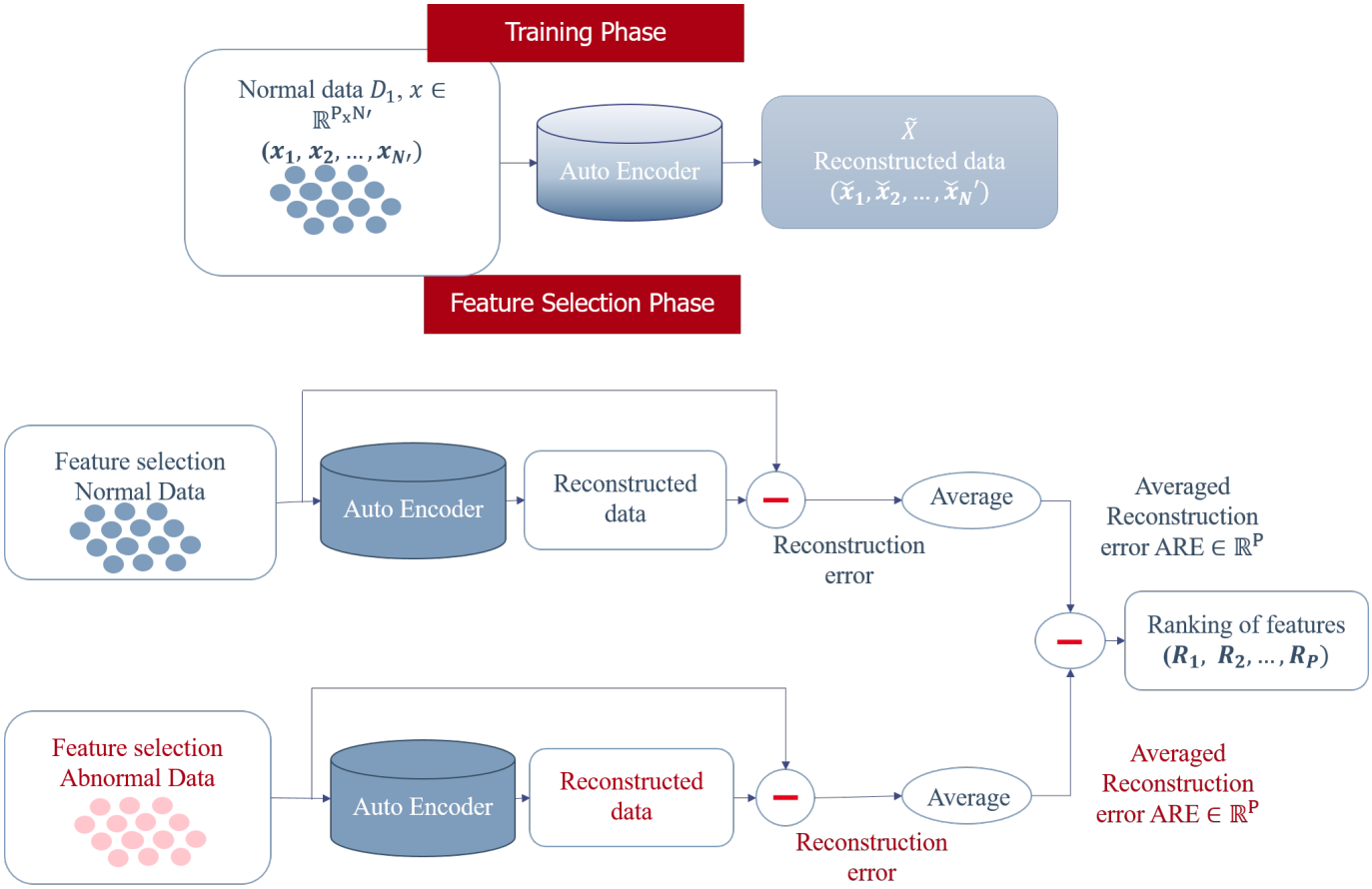

FIGURE 4.4: *Our adapted implementation of the proposed method by Massi et al. [353].*

4.6.1 VISUAL DISTRACTION DETECTION

In Table 4.7, we provide the Delta Scores obtained for each feature using the method proposed by Massi et al [353]. Notably, the feature 'Looking at road' emerges as the most important feature with a high score of 2.77, while 'Gaze heading standard deviation' is ranked as the least significant feature with a comparatively lower Delta Score of 0.18. It is worth noting that the overall ranking of features is similar to that obtained with our approach; however, there are differences in the rankings for features occupying positions from the 9th to the 14th. In Figure 4.5, we compare the performance improvements achieved by adding the most important features one by one for the three tasks: anomaly detection, classification, and regression. The graphs reveal that both strategies yield similar performance improvements. In the regression task, our feature selection slightly outperforms the other strategy.

TABLE 4.7: *Delta scores*

Feature Name	Delta Score
Looking at road	2.77
Gaze pitch mean	2.14
Head pitch standard deviation	0.82
Gaze pitch standard deviation	0.77
Head pitch mean	0.65
BVP IBI	0.57
ECG IBI	0.57
Saccade rate	0.51
Head heading mean	0.49
Eye position entropy	0.41
Gaze heading mean	0.41
Head heading standard deviation	0.38
Saccade magnitude	0.33
Saccade peak velocity	0.26
Gaze heading standard deviation	0.18

4.6.2 PSYCHOTIC RELAPSE PREDICTION

We also evaluated the approach on psychotic relapse prediction. Table 4.8 presents feature data ranking for patients in the context of detecting relapse in patients with psychotic disorders. Each column represents a different patient, from patient 1 (P1) through patient 10 (P10). The table includes various features, and for each feature, the table displays the corresponding delta values for each patient. The delta values highlight the importance of different features in detecting relapse. Notably, the features of heart rate, gyroscope, and steps seem to have relatively higher values for patient 1 than sleep. Similar to our method, patients 1 and 4 display the highest scores. Moreover, patient 8 also exhibits notable scores. However, the feature rankings for patient 4 appear to offer more discriminatory information compared to patient 1, which is inconsistent with the outcomes from our proposed approach in Table 3.6 and the testing results detailed in the previous chapter in Table 3.16. In the table, it is intriguing to observe that patient 9 has a relatively high positive value for this noise feature. This finding raises several possibilities. It could suggest that, for this specific patient, none of the features are inherently strong indicators for relapse prediction. Alternatively, it might indicate that the feature selection method applied in this context may not be capturing the critical features effectively. However, due to the number of abnormal examples in the validation and testing sets (only three), it is difficult to conclude. Moreover, it is essential to note that the importance of these features varies among patients. For instance, while "HR_std" is particularly important for patient 4 with a value of 0.575, it may not be as relevant for other patients, (e.g. negative values for patients 2,5,7, and 9). This diversity underscores the personalized nature of healthcare for individuals with psychotic disorders, highlighting the need to tailor feature selection and modeling to each patient's unique characteristics.

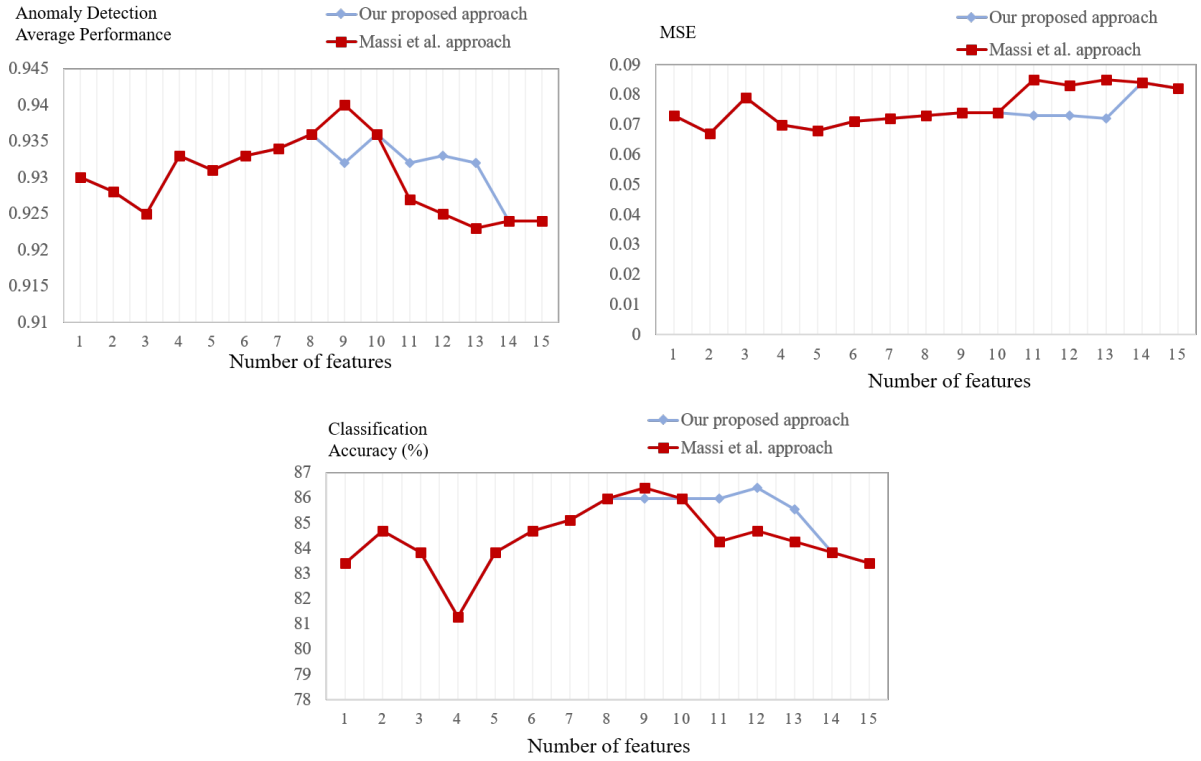

FIGURE 4.5: *Performances obtained with feature selection using our proposed approach and the approach proposed by Massi et al.*

TABLE 4.8: *Feature Data for Patients*

Features	Delta Scores									
	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
sleep	0.112	0.038	-0.05	0.128	-0.181	-0.095	0.008	0.215	-0.073	0.049
sleep_index	0.099	-0.056	-0.083	0.094	-0.07	**0.109**	0.017	0.114	0.002	-0.035
HR_mean	**0.311**	0.072	0.102	0.229	**0.196**	0.059	0.029	0.189	0.048	**0.062**
HR_std	0.196	-0.103	0.092	**0.575**	-0.25	0.083	-0.11	**0.31**	-0.148	0.031
acc	0.036	**0.091**	0.063	0.123	-0.107	-0.025	0.071	0.035	-0.178	0.011
gyr	0.26	-0.004	-0.047	0.21	-0.054	-0.076	-0.209	0.095	0.074	-0.009
steps	0.185	-0.064	**0.108**	0.21	-0.052	-0.004	**0.135**	0.007	-0.141	-0.117
noise	-0.061	-0.122	-0.065	0.049	-0.013	-0.044	-0.134	0.022	**0.094**	-0.045

The feature rankings show variations between both approaches, we compare across all patients the performance of their best and worst ranked features, as depicted in Figure 4.6. In 4.6.a, we calculate the performance of the best ranked feature for each patient and in 4.6.b we calculate the performance of the worst ranked feature for each patient. The performances of the best-ranked features from both approaches are generally very close, indicating that both strategies tend to identify strong features similarly across most patients except for patient 6 where our strategy's best feature significantly outperforms the one obtained through their approach. However, in the case of the least ranked feature, the worst-ranked features from the approach

by Massi et al. [353] outperformed those of our approach for most patients, thereby rendering the rankings from our approach more consistent. Moreover, considering the results from both strategies, it becomes apparent that there is a need for a larger dataset to establish more robust feature rankings.

Both methods exhibited comparable performance in the visual distraction task, showcasing

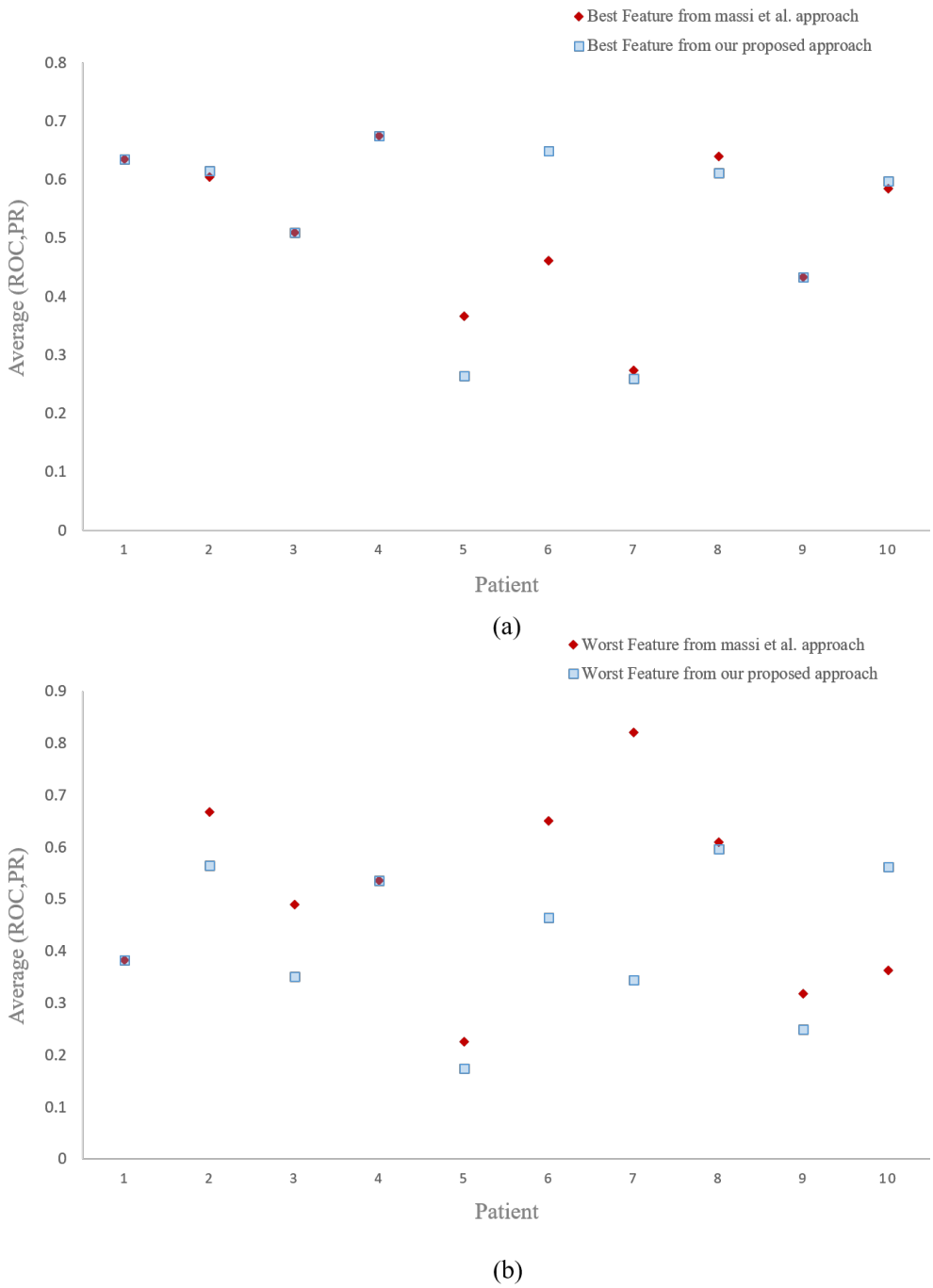

FIGURE 4.6: *(a): Anomaly detection averaged performance for each patient using the best ranked feature our proposed strategy and Massi et al. approach. (b): Anomaly detection averaged performance for each patient using the worst ranked feature our proposed strategy and Massi et al. approach.*

good efficiency. However, when applied to the psychotic relapse prediction, discrepancies arose between the outcomes of the two approaches. Due to the relatively low results of the feature scores in the second task, the identification of crucial features became challenging, especially

given the small validation set. An advantageous aspect of our approach lies in its consideration of the continuous value of the annotation. Nonetheless, reaching conclusive results demands further testing and analysis.

4.7 Potential exploitation of the proposed method

Our method can be exploited in several ways to provide more explainability and insights into the model:

- **Feature Importance:** The feature rankings offer valuable insights into the significance of each feature within the anomaly detection process. This grants explanations regarding which features the model deems relevant. For example, for distraction detection, we found that the feature 'Looking at road' is very important for the detection of distraction, whereas 'ECG IBI' was not relevant. Furthermore, it sheds light on variations in the manifestation of certain phenomena, such as personalized relapse prediction. Without conducting the exhaustive search as we had done in the previous chapter, the results obtained through the feature selection method illustrate the varying importance of the features for each individual patient.
- **Multiple types of anomaly classification:** In scenarios involving anomaly detection, such as identifying abnormal behaviors in driving like fatigue, distraction, or stress, conventional models often detect anomalies without specifying the specific behavior triggering them. However, our approach can enhance this process by targeting and identifying various types of dangerous driving behaviors. To achieve this, we can train a unique AE for different behavior-related signals within the normal state. To compute the correlation scores, we can utilize a small dataset encompassing both normal and abnormal instances specifically tailored to each subgroup, such as fatigue, drowsiness, and distraction. This approach will enable us to isolate and identify the key features critical for detecting each specific behavior within the driving context. During testing, when an instance is flagged as an anomaly, our approach can verify the features contributing most to the RE. Then, by leveraging our feature rankings obtained during training, we can categorize abnormal behavior into known driving anomaly categories. This approach represents a deeper dive into anomaly detection, aiming not only to detect anomalies but also to attribute them to specific known driving behaviors, potentially offering more nuanced insights into various dangerous driving patterns.
- **Sensors Selection:** The application of this method extends to scenarios where a multitude of signals are collected without a clear understanding of their significance. For example, in the domain of driving visual distraction detection, our method's findings showcase that signals like ECG and BVP contribute minimally to distinguishing between distracted and non-distracted behavior. This insight can aid in prioritizing and focusing resources on more influential sensors, potentially streamlining data collection efforts and computational resources in scenarios where signal acquisition is resource-intensive or costly.

4.8 Conclusion

This chapter explored the use of anomaly detection methods for feature selection within imbalanced datasets. Our exploration revolves around leveraging AE trained solely on normal class data and calculating correlation scores between the reconstruction error of features and annotations from both normal and abnormal samples. This study encompasses a comprehensive evaluation of our proposed feature selection method within two applications: visual distraction detection and psychotic relapse prediction.

In the visual distraction application, we investigated the impact of employing strategies for binary classification, anomaly detection, and regression models. Our approach, relying on correlation scores, provided insightful perspectives on feature importance. Particularly, it highlighted the benefits of beginning with highly influential features in classification tasks and regression.

However, when extended to the more intricate "psychotic relapse prediction" task, which had a relatively constrained dataset, the feature selection strategies showcased less consistent feature rankings. This highlighted the crucial need for a robust validation dataset for these strategies. Additionally, the results reinforced the necessity for personalized models and underscored the variability in feature importance across all patients.

Our study presented promising results by utilizing AE reconstruction error for gaining insights into model explainability and feature importance. However, to validate its efficiency and generalizability, this method warrants further testing across various datasets.

CHAPTER 5

Multimodal Fusion Using Anomaly scores

In this chapter, we present and evaluate a new multimodal fusion approach. We introduce difficulty indicators for each modality's data as additional inputs to enhance signal fusion. Our proposed method is tested on the ULM-TSST dataset from the Muse-Stress challenge at ACM Multimedia21, where it secured the second position.

Chapter contents

5.1 Introduction

As discussed in Chapter 2, affective expressions inherently involve multiple modalities. An emotion encompasses three distinct components: a subjective experience, a physiological response, and an expressive response [354]. The latter two are most prominently manifested through facial expressions, speech, and physiological indicators like heart rate or electrodermal activity. Both humans and computers recognize emotions by analyzing several types of signals. Extensive research has been dedicated to identifying the most effective modalities and features for affective computing. Some studies focus solely on predicting emotions through specific modalities such as speech [355], video [356], or physiological signals [357]. Conversely, other works explore the multimodal nature of emotion expression by simultaneously fusing multiple modalities to predict emotions [358, 359].

In Chapter 2 Section 2.2.6, we explain in detail the most adopted fusion techniques in modalities fusion in affective computing. It can be summarized into three categories: early, late, and hybrid fusion. However, there remains no consensus on the optimal approach for fusing multiple modalities in emotion recognition [124].

Instead of focusing on the techniques and models used for fusion, some works focused on enhancing the data prediction and fusion process by including additional information by calculating an additional learning task to the model. Zhang et al.[360] used data difficulty indicators in dynamic difficulty awareness training (DDAT). DDAT relies on the assumption that a model will perform better if it is provided with the learning difficulty of the data. They train a model that reconstructs the input and predicts emotions in a multi-task learning framework. They calculate the RE of the inputs and use it as a difficulty indicator to update the model. The RE is re-injected into the model to update its weights accordingly. For fusion, they used a linear regression model which input are the original prediction and the corresponding difficulty indicator, without considering temporal influences on the fusion of predictions.

In our work, we have drawn inspiration from their approach but extended it further. Our architecture is designed to leverage data difficulty indicators for multimodal fusion, utilizing recurrent models to fuse predictions. Moreover, we specifically evaluate the impact of adding the RE in the fusion process alone, which was not covered in their research. Our evaluation centers on the ULM TSST emotion database, encompassing multiple modalities with the goal of predicting arousal and valence.

The rest of this chapter is structured as follows: Section 5.2 provides a detailed overview of our proposed approach. Following this, we introduce the dataset and the computed features used to evaluate our fusion methods in Section 5.3.1.1. We present the results of our ablation studies and comparisons with other works in Section 5.4. Finally, we conclude our findings in Section 5.5.

5.2 Proposed multimodal fusion scheme

In multimodal prediction, the significance of each modality can fluctuate based on various factors. An example in emotion prediction: if an unseen facial expression accompanies a familiar speech pattern indicating happiness, the model should emphasize the speech modality more for the prediction. Another instance could involve a malfunctioning sensor providing erratic signals different from what it learned during training; here, the model should rely more on the consistent, normal signals it's accustomed to. These scenarios exemplify one factor influencing the change in modalities' importance, namely encountering signals dissimilar to the model's training data. However, there are other contributing factors to consider e.g. context. In our work, we focus on the difficulty indicator reflecting the dissimilarity from the training dataset. In this section, we explain our proposed approach for multimodal fusion prediction that takes into account data

difficulty indicators. We adopt the late fusion strategy where the fusion is done on the decision level. The proposed approach is divided into three steps: Data difficulty indicator estimation, Unimodal predictions, and Multimodal fusion.

5.2.1 Step 1: Data difficulty indicator estimation

As mentioned in Chapter 2, AE are used to solve the high dimensionality problem since an increase in dimensions raises the required complexity of the model, the demand in data, and the computation capacities. More recently, AE is being used to detect anomalies as we explored in Chapter 3. In this context, the average RE of the input determines whether it's considered an anomaly. It serves as a measure of dissimilarity from the training data. Similar to the approach in Zhang et al. [360], we are exploring this metric as a measure of difficulty. We train a specific AE for each feature set. Post-training the AE, we calculate a difficulty indicator for each feature set of every input. We define the difficulty indicator as the averaged MSE between the input and its reconstruction.

5.2.2 Step 2: Unimodal predictions

Since we are adopting late fusion, we perform predictions of each feature set separately. For each feature set, we train a regressor where the input is the feature and the data difficulty indicator is obtained from the AE specific to the feature set.

5.2.3 Step 3: Multimodal predictions

The last step is multimodal fusion. We feed our fusion model with the predictions from each feature set and the difficulty indicator of the respective input, as shown in Figure 5.1.

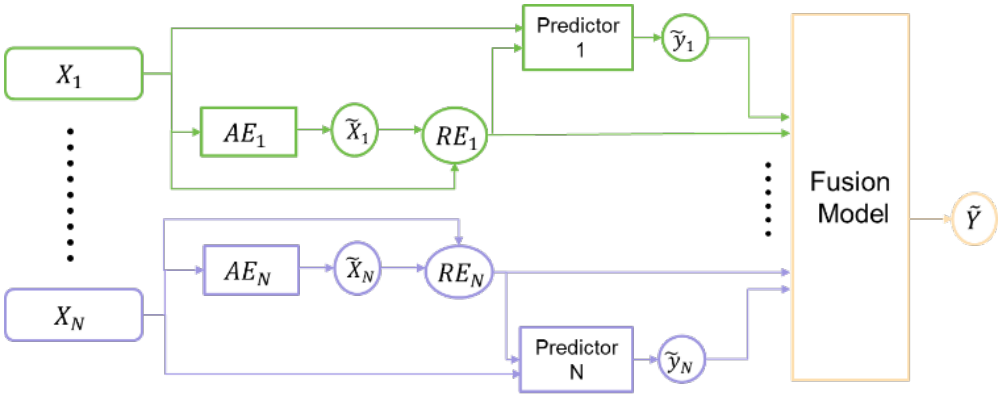

Figure 5.1: *Diagram of the proposed solution. X_i refers to the ith unimodal features set. $\tilde{X}_i$ refers to their reconstruction using the AE. RE refers to the averaged reconstructed error. $\tilde{Y}_i$ refers to the unimodal prediction using the ith features set and $\tilde{Y}_i$ refers to the multi-modal prediction.*

5.3 Model architecture and training settings

For each feature set, we train a separate AE for calculating the data difficulty indicators and a regressor for arousal and valence continuous prediction. We implement our solution using the Pytorch toolkit [361]. For each feature set, our unimodal predictors, AE, and late fusion model share the same architecture: four-layered bi-directional RNN with 64 hidden neurons followed by a feedforward layer. We evaluate two types of RNNs: a LSTM and a GRU networks. The choice between LSTM , GRU , Bidirectional Long Short-Term Memory (BiLSTM), and Bidirectional Gated Recurrent Unit (BiGRU) as recurrent layer is determined based on the

results on the validation set. We utilize the Adam optimizer and varied learning rates (0.001, 0.005, 0.0005). As a form of regularization, we apply dropout and evaluate with several rates (0.1, 0.2, or 0.5). We train the model for 100 epochs at most and apply early stopping if the validation performance does not improve after 15 epochs. For the loss function proposed in the challenge [362], we use the Concordance Correlation Coefficient (CCC) loss [363] which is defined as:

$$L = 1 - CCC, \tag{5.1}$$

where

$$CCC = \frac{2\rho\hat{\sigma}_y\sigma_y}{\hat{\sigma}_y^2 + \sigma_y^2 + (\hat{\mu}_y - \mu_y)^2} \tag{5.2}$$

and $\hat{\mu}_y$ and μ_y are the mean of the prediction $\hat{Y}$ and the label Y, and $\hat{\sigma}_y$ and σ_y are the corresponding standard deviations. ρ is the Pearson Correlation Coefficient (PCC) between $\hat{Y}$ and Y. Thereby the predictions that exhibit a strong correlation with the gold standard but deviate in value are penalized according to the extent of their deviation [364].

5.3.1 Dataset

5.3.1.1 Datasets for emotion recognition

The first step in constructing an effective ML model is identifying an appropriate dataset. In emotion recognition, datasets typically fall into three categories: acted, induced, and natural [365]. Acted datasets involve individuals being directed to display specific emotions. Induced datasets create controlled settings specifically designed to elicit particular emotions. Meanwhile, natural datasets are gathered from people spontaneously expressing or reacting to emotions. It's important to note that collecting natural datasets poses the greatest challenge due to ethical and privacy concerns.

Certain datasets are limited to a single modality. For instance, RAF-DB [366] exclusively comprises facial images, while Berlin DB [367] solely contains utterances from German-speaking individuals. Exploring these databases is intriguing as they concentrate on a single modality, making them suitable for scenarios where only one type of signal can be recorded. However, this approach doesn't leverage multiple sources of information, potentially limiting robustness. Hence, our focus lies on databases encompassing multiple modalities exclusively. Among the notable multimodal datasets, RAVDESS [368] offers audiovisual recordings featuring 24 professional actors annotated across various emotion categories. IEMOCAP [369], another renowned database in affective computing, includes audio, speech, motion capture, and textual recordings of both scripted and unscripted conversations. It encompasses annotations for categorical emotions as well as dimensional labels like valence, activation, and dominance. However, findings suggested that real-world expressions are often subtler than acted ones [370], potentially leading to reduced model performance in practical applications. Therefore, we aimed towards induced or natural datasets.

Moreover, our focus on emotion prediction in the dimensional approach, unconstrained by specific emotion categories, led us to consider datasets with continuous annotations. The following datasets are also multimodal and have continuous annotations of emotions. SEWA DB [365] is a rich database collected in the wild in diverse settings and includes several ethnicities. RECOLA [371] encompasses audio, video, and physiological recordings of participants in a spontaneous collaborative environment. ULM-TSST [362], includes audio, video, and physiological signals captured within a stress-inducing environment. Our choice to work with ULM TSST stems from its emphasis on stress-inducing settings, aligning with our interest in proactive well-being applications. Furthermore, ULM TSST was introduced in the MUSE challenge [362], providing us with an opportunity to position our work within the broader research community for comparison and evaluation.

5.3.1.2 ULM-TSST DATASET DESCRIPTION

It consists of 69 German-speaking participants, aged between 18 and 39 years, in a stress-inducing setup following the Trier Social Stress Test procedure [372]. Following a short preparation period, participants are instructed to deliver an oral presentation. This takes place in a simulated job interview environment where two interviewers are present, though they remain silent for a duration of five minutes. Video and physiological recordings are taken during the presentation. The total duration of the database is 5h: 47 min: 27s. Arousal and valence are annotated by three raters and the fusion of the annotations is done using the Rater Aligned Annotation Weighting method. The given modalities are audio, video, and transcripts in addition to the physiological signals EDA,ECG, respiration, and heart rate (BPM).

5.3.1.3 DATASET PARTITIONNING

For a thorough evaluation, the dataset is divided into three subsets: train, devel (validation), and test. The partition was provided by the organizers of the challenge. We present the number of unique sessions provided in each partition in Table 5.1. Each session corresponds to a participant. Additionally, Figure 5.2 illustrates the distributions of arousal and valence values across the dataset partitions. It reveals a consistent distribution pattern of arousal and valence across all partitions.

TABLE 5.1: *Number of unique videos and total duration of data in each partition of the dataset ULM-TSST*

Partition	Number of sessions	Total duration
Train	41	3:25:56
Devel	14	1 :10 :50
Test	14	1 :10 :41
$\sum$	69	5 :47 :27

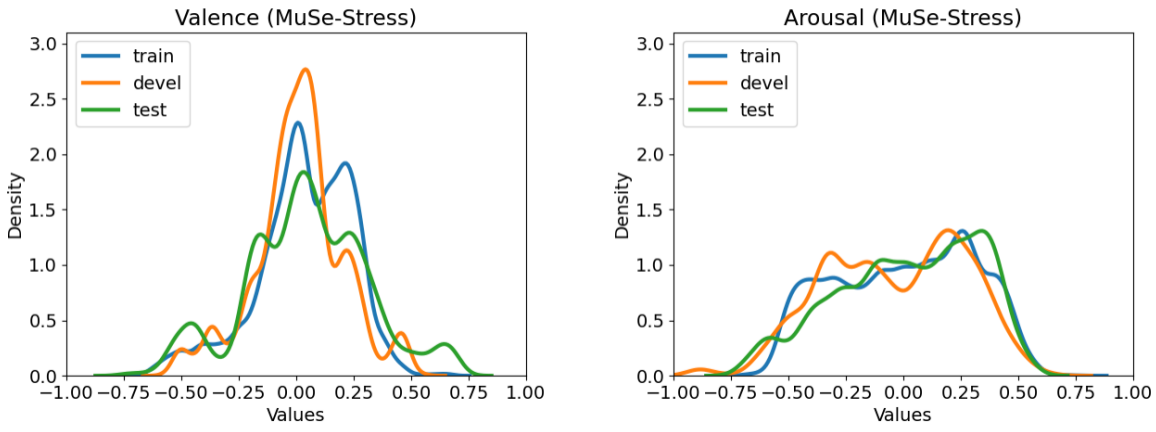

FIGURE 5.2: *Frequency distribution in the partitions train, development, and test for the continuous values of arousal and valence [362].*

5.3.2 Features

The MuSe2021 provides a range of relevant extracted acoustic, visual, and textual features for the participants to use. However, we discarded the physiological signals due to their low performance in this dataset [362]. We explored the performance of high-level and low-level features for continuous emotion prediction. In our approach, we used the following parameter sets provided in the dataset due to their reported good performance for emotion recognition tasks [362][373][374] [375].

5.3.2.1 Acoustic features

Extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) features We explore using the extended Geneva Minimalistic Acoustic Parameter Set eGeMAPS [47] which can be extracted using the free openSMILE toolkit [376]. It consists of Low-Level Descriptors (LLD) including:

- **Frequency Parameters:** Pitch, jitter, the bandwidth of first formant(formant 1), and formant 1,2, and 3 frequency, and Formant 2-3 bandwidth.
- **Energy Parameters:** Shimmer, loudness, harmonics-to-noise ratio
- **Spectral Parameters:** Alpha Ratio, Hammarberg Index, Spectral Slope 0-500 Hz and 500-1500 Hz, Formant 1, 2, and 3 relative energy, Harmonic difference H1-H2, Harmonic difference H1-A3, Mel-Frequency Cepstral Coefficients 1-4, spectral flux.
  The functionals arithmetic mean and coefficient of variation are applied to all these LLD over voiced regions only except for MFCC 1-4 and spectral flux. For MFCC 1-4 and spectral flux, the functionals are calculated over voiced regions only and voiced and unvoiced regions together. For both loudness and pitch, an additional set of 8 functionals is employed. These functionals encompass the 20th, 50th, and 80th percentiles, the range spanning from the 20th to the 80th percentile, and the mean and standard deviation of the slopes in the rising and falling segments of the signal. In addition, the set includes the mean value of the Alpha Ratio, the Hammarberg Index, and the spectral slopes within the 0-500 Hz and 500-1500 Hz frequency ranges across all unvoiced segments. Furthermore, the parameter set incorporates the equivalent sound level.
- **Temporal Parameters:** rate of loudness peaks, mean length and the standard deviation of continuously voiced regions, mean length and the standard deviation of unvoiced regions, and number of continuous voiced regions per second.

The total number of parameters in eGeMAPS is 88. We normalize the eGeMAPS features.

DeepSpectrum DeepSpectrum features [52] were also tested in our experiments. DeepSpectrum had been trained on spectrograms of audio snores, utilizing a deep CNN (VGG-19) that had been pre-trained for image recognition. The default extraction settings were maintained to yield a feature set with dimensions of 4096.

5.3.2.2 Visual features

Face action unit Using the Multi Cascaded Convolutional Neural Networks (MTCNN)[377] shown in Figure 5.3, 17 facial action unit intensities are obtained from the center and left sides of the face.

VGGFace The VGGface [378] architecture was initially intended for supervised facial recognition tasks. The network was trained on a substantial dataset consisting of 2.6 million faces,

representing more than 2,500 unique identities. Moreover, by detaching the top layer of a pre-trained version, it can provide a 512-feature vector output referred to as VGGface. The input to the network was obtained using the MTCNN.

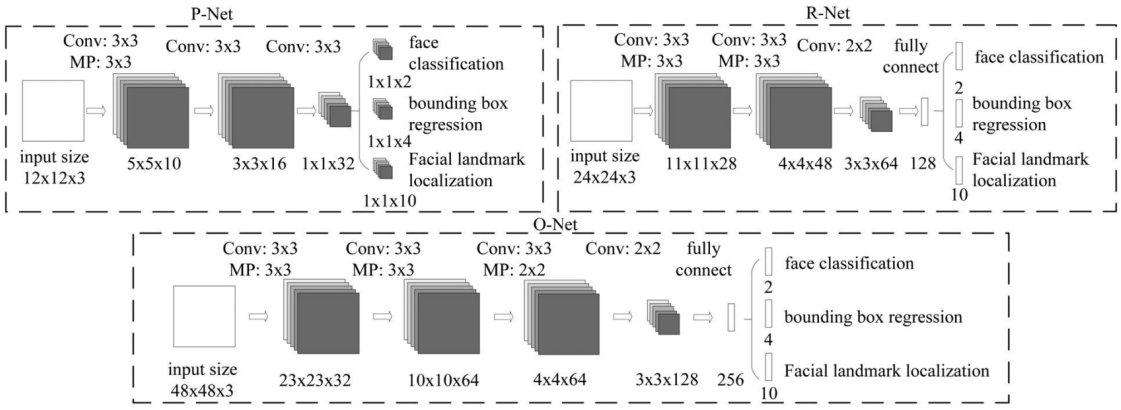

FIGURE 5.3: *MTCNN architecture [377].*

5.3.2.3 TEXTUAL FEATURES

BERT : For the textual modality, BERT [379] features are provided. The high-level contextual embedding proved to deliver state-of-the-art results for several Natural Language Processing (NLP) tasks [380] [381]. BERT is specifically engineered to pre-train deep bidirectional representations from unannotated text data. It achieves this by considering both left and right contexts simultaneously in all layers. Consequently, the pre-trained BERT model can be fine-tuned with the addition of just one extra output layer. During the inference process, the context-driven representations are retained, with a specific vector associated with each word. This differs from static word embeddings, which provide a single vector for each word regardless of context. The features are computed by summing the outputs from the last four BERT layers, resulting in a 768-dimensional similar to [104]. Since the Ulm-TSST is a German database, BERT (base) that has been pre-trained on German texts is used.

5.4 RESULTS

In this section, we will outline the outcomes of our experiments aimed at identifying the optimal recurrent layer and loss function for our model. Additionally, we will show the results of unimodal predictions across different modalities and features, examining the impact of integrating the data difficulty indicator as supplementary input for these predictions. Furthermore, we present our exploration of combining unimodal predictions to identify the most effective feature combinations and assess the effect of incorporating the difficulty indicator into the fusion process. Finally, we will compare our findings with other methods applied to the ULM TSST dataset.

5.4.1 ABLATION STUDY

5.4.1.1 NETWORK ARCHITECTURE STUDY

We conduct an ablation study to determine the best type of recurrent layer for our model. We explore four types of models: LSTM , GRU , BiGRU, and BiLSTM using the standard approach where only unimodal features are inputted to the model. All of the four models are 4-layered networks with 64 hidden neurons. We present the results of the performance of each model in

Tables 5.2 and 5.3 for arousal and valence respectively. Generally, the BiGRU and BiLSTM models outperform both LSTM and GRU models. These results show that both past and future information is relevant for emotion prediction. Since BiGRU and BiLSTM have close performances and BiGRU have fewer parameters, we choose to continue with BiGRU model to target better generalization on the testing set.

TABLE 5.2: *CCC performance comparison between recurrent models for unimodal predictions on arousal dimension on the validation set.*

Features	LSTM	BiLSTM	GRU	BiGRU
eGeMAPS	0.4714	**0.5466**	0.4739	0.5322
VGGface	0.1809	**0.3561**	0.1283	0.2293
FAU	0.3260	0.3637	0.3641	**0.3688**
BERT	0.2250	**0.3166**	0.2349	0.2681
DeepSpectrum	**0.3339**	0.2617	0.2538	0.2185

TABLE 5.3: *CCC performance comparison between recurrent models for unimodal predictions on valence dimension on the validation set.*

Features	LSTM	BiLSTM	GRU	BiGRU
eGeMAPS	**0.5926**	0.5597	0.5671	0.5646
VGGface	0.5650	0.5671	0.5414	**0.6481**
FAU	0.5480	**0.5952**	0.5531	0.5143
BERT	0.3025	0.2538	0.2828	**0.4473**
DeepSpectrum	0.5548	**0.5678**	0.5532	0.5630

5.4.1.2 LOSS FUNCTION STUDY

We also experiment with 3 loss functions on the BiGRU model: "CCC", "MSE", and "L1" loss. Generally, CCC gives better performance as shown in Table 5.4 and 5.5. The results agree with the findings in [382] [50]. This coherence is expected as our primary focus was enhancing the CCC between the predictions and annotations.

TABLE 5.4: *CCC performance on the arousal obtained by using different loss functions on the validation set.*

Features	MSE	L1	CCC
eGeMAPS	0.3038	0.3306	**0.5322**
VGGFace	0.0772	0.0159	**0.2293**
FAU	0.3083	**0.4354**	0.3688
BERT	**0.3277**	0.3160	0.2681
DeepSpectrum	0.0666	0.1201	**0.2185**

Table 5.5: *CCC performance on the valence obtained by using different loss functions on the validation set.*

Features	MSE	L1	CCC
eGeMAPS	0.4465	0.4414	**0.5646**
VGGFace	0.5354	0.3980	**0.6481**
FAU	0.4011	0.2885	**0.5143**
BERT	0.3658	0.3095	**0.4473**
DeepSpectrum	0.5262	0.4984	**0.5630**

5.4.2 Influence of difficulty data indicator

5.4.2.1 Unimodal level

This experiment aimed to assess the impact of the data difficulty indicator on the model's prediction. Tables 5.6 and 5.7 showcase the performance on the validation set for arousal and valence prediction under two scenarios. In the first, the predictor receives input from a single feature set. In the second, the model receives the feature set along with data difficulty indicators, acquired via the AE trained on this feature set's data. Across most modalities for both tasks, the results demonstrate improvement, indicating that incorporating difficulty information enhances the model's performance. These findings align with [360], highlighting that providing insights into prediction difficulty notably enhances model performance. These results strongly advocate further exploration of this indicator's influence at the multimodal fusion level.

Table 5.6: *CCC performance comparison for unimodal predictions on arousal dimension on the validation set*

Modalities	Features	Model Inputs	
		Unimodal Features	Unimodal Features + RE
Audio	**eGeMAPS**	0.5322	**0.5829**
	DeepSpectrum	0.2185	**0.2498**
video	**VGGFace**	0.2293	**0.3926**
	FAU	**0.3688**	0.3668
Text	**BERT**	0.2681	**0.3457**

Table 5.7: *CCC performance comparison for unimodal predictions on valence dimension on the validation set*

Modalities	Features	Model Inputs	
		Unimodal Features	Unimodal Features + RE
Audio	**eGeMAPS**	0.5646	**0.6353**
	DeepSpectrum	0.5630	**0.5676**
Videoo	**VGGFace**	0.6481	**0.6798**
	FAU	0.5143	**0.5307**
Text	**BERT**	0.4473	**0.4655**

5.4.2.2 Multimodal level

In this section, we aim to identify which feature combination gives the best results. For this purpose, we take the trained models predictions which we trained separately using the feature set and its data difficulty indicator. The results here present the fusion of predictions using the outputs of the models only. We try several combinations for fusion, as shown in Table 5.8. We observe that adding the textual modality causes a drop in performance; this can be explained by the fact that textual information in interviews may not reflect true emotions. We also observe that adding DeepSpectrum features does not improve the performance. This could be the result of the high dimensionality of this feature set (4096 features), which would necessitate a more complex model to leverage the information of these features. When fusing several modalities, the performance is improved significantly, further affirming the advantage of utilizing several modalities for emotion prediction. The optimal combination involves audio and video features, precisely the feature sets comprising eGeMAPS, VGGFace, and FAU. It's noteworthy that the best-performing feature combination encompasses both low-level features (FAU) and high-level features (VGGFace). We aimed to assess the impact of integrating data difficulty levels into the

Table 5.8: *CCC performance of multi-modal features on the arousal and valence dimension on the validation set.*

Modalities	Features	Arousal	Valence
A+V	**eGeMAPS + VGGFace**	0.6205	0.7024
A+V+T	**eGeMAPS + VGGFace + BERT**	0.6031	0.6811
A+V	**eGeMAPS + VGGFace + DeepSpectrum**	0.6199	0.7320
A+V	**eGeMAPS + VGGFace + FAU**	**0.6469**	**0.7653**

multimodal fusion process. We conducted the multimodal fusion in two scenarios: first, using predictions from each feature set solely, and second, fusing both predictions and the data difficulty indicator of each feature set. Leveraging the optimal combination of unimodal predictions (eGeMAPS + VGGFace + FAU) outlined in Table 5.8, our results, as depicted in Table 5.9, demonstrate improved performance when incorporating the data difficulty indicator as additional features during fusion for both arousal and valence. This substantiates the utility and effectiveness of this feature in the fusion process.

Table 5.9: *Multimodal fusion CCC performance on arousal and valence using as fusion model inputs, unimodal predictions only (first row) and unimodal prediction along with RE (second row) on the validation set.*

Inputs	Arousal	Valence
Fused Unimodal predictions	0.6469	0.7653
Fused Unimodal predictions + RE	**0.6554**	**0.8036**

5.4.3 Results comparison of methods proposed in the literature

Our work on the ULM-TSST dataset marked one of the initial studies conducted with this specific dataset. In this section, we present concurrent studies that were done around the same period as our research and the results obtained from each research on the testing dataset. The research listed here represents the highest-ranking solutions alongside our work for the MUSE stress sub-challenge [362]. In their study [383], authors employed a combination of acoustic features like eGeMAPS, DeepSpectrum, MFCC, INTERSPEECH 2009 (IS09), and INTERSPEECH 2010 (IS10), along with visual features such as FAU, Emonet, and SENET. Their exploration involved both early and late fusion techniques by a bidirectional LSTM model. Their results showed

that late fusion has a superior performance. Their findings revealed that the optimal combination for arousal was FAUs+eGeMAPS, whereas, for valence, FAUs+eGeMAPS+Emonet+IS10 showed the best performance. Duong et al. [384] examined all available features within the MUSE dataset, including eGeMAPS, DeepSpectrum, VGGish, FAU, VGGface and BERT. They introduced a positional encoding created by encoding timestamps for each frame within the input sequence. These encoded features were then fed into a self-attention temporal CNN and LSTM. Their multimodal predictions were fused using two fully connected layers. For valence and arousal predictions, they selected the best audio and video features for arousal and valence (FAU + eGeMAPS) determined through an ablation study conducted on the development set. Ma et al. [385] used a model consisting of a self-attention layer with a LSTM followed by a fully connected layer. They predicted from the following features eGeMAPS, VGGface, ECG, RESP, and BPM fed [385] and performed late fusion with BiLSTM. A common finding across all research was the performance drop between the development set and the testing set. This suggests a potential difference in data distribution between these two datasets. Also, similar to our findings, the text modality didn't appear to be efficient for predicting arousal and valence in this dataset.

The results in Table 5.10 showcase various approaches, with the top-performing model incorporating biosignals, signifying their efficacy in predictions. Additionally, the integration of a self-attention mechanism, as observed in [385] and [384], led to improvements. Our work secured the second rank among these studies. An intriguing avenue for future research could involve testing whether our proposed solution can be further enhanced by incorporating a self-attention mechanism.
In the following year, the ULM TSST dataset featured in a MUSE subchallenge [386], prompting several research efforts to predict arousal and valence [387] [388] [389] [390] [391] [392]. However, the challenge organizers altered the annotation for arousal, preventing direct comparison without retraining our models.

TABLE 5.10: *Proposed solutions for continuous emotion prediction on the ULM TSST dataset in the literature.*

	Arousal	Valence	Total
[385]	0.615	0.460	0.538
Our approach [393]	0.595	0.427	0.511
[384]	0.613	0.405	0.509
[383]	0.3054	0.664	.485

5.5 CONCLUSION

In this chapter, we investigated the integration of anomaly detection methods to enhance multimodal fusion in predicting continuous emotions, specifically arousal and valence, using the ULM TSST dataset. This dataset incorporates various information modalities, such as audio, video, biosignals, and text, obtained from individuals in stress-induced conditions. Our approach focused on employing late fusion with audio, video, and text modalities.

We conducted studies to analyze the impact of a data difficulty indicator, derived from the RE of AE, on both unimodal and multimodal prediction levels. To this purpose AE were trained for each feature set specifically to obtain the data difficulty indicators. For evaluation, predictors were trained with and without these difficulty indicators both on unimodal and multimodal levels. The outcomes demonstrated an improvement in models trained with the data difficulty indicators, affirming the informative nature of the AE RE and its significance in unimodal and multimodal fusion.

Additionally, our findings revealed that the most effective modalities for emotion prediction were audio and video, specifically eGeMAPS, FAU, and VGGFace as the optimal combination. Surprisingly, the text modality exhibited a drop in performance, contributing less effectively to the predictive model.

However, a limitation of our work was the potential overfitting of the model to the validation set, a common issue observed in other proposed solutions in the literature. The disparity in CCC performance between the validation and test sets suggests a need for additional data to ensure generalization. Implementing data augmentation methods could potentially address this limitation.

CHAPTER 6

Conclusion

6.1 Conclusions

Exploring anomaly detection within affective computing has provided solutions to major challenges in the field. Our research spanned various applications, including detecting dangerous driving behavior, such as visual distractions for road safety, predicting psychotic relapses in mental health patients, and continuously predicting emotions for individuals under stressful situations. By utilizing anomaly detection methods, we overcame the limitations of traditional supervised learning, which requires labeled data and balanced datasets, and achieved significant advancements in these domains. Moreover, our work explored several types of modalities: speech, video, text, physiological signals, and eye-tracking features.

First, we explored detecting rare mental states using anomaly detection methods. The fundamental principle underlying our approach involved training a model to comprehend the distribution patterns within the normal data. When presented with new samples, the model calculates their distance or dissimilarity in comparison to the data it has been trained on. Any significant deviation from this learned pattern indicates that the sample belongs to the rare class that we aim to detect. We explored applying this approach to two affective computing applications: Dangerous driving behavior and psychotic relapse prediction.

The dangerous driving behavior detection was tested on a dataset containing normal driving behavior and distracted driving behavior. For the relapse prediction, the data contained the logged daily activity of patients during relapse and non-relapse days. In both applications, the models were trained on normal data, which was represented by non-distracted driving for the first application and non-relapse days for the second application. Our study encompassed various anomaly detection-based methods, including OCSVM , Isolation Forest, Elliptic Envelope, LOF , and AE. Our comparison with supervised models in various data imbalances reinforced the validity of applying these approaches to real-world scenarios to overcome the obstacle of collecting dangerous or costly data. Specifically, the Isolation Forest demonstrated robustness in imbalanced datasets in most of the experiments, delivering superior performance. Moreover, for the psychotic relapse prediction, we leveraged the use of anomaly detection methods to develop personalized models for each patient and conducted a thorough study on the influence of the choice of models, features, and time window for calculating the features on the performance of relapse prediction. Our results underscored the necessity for tailored approaches in mental health monitoring, establishing a foundation for more accurate and personalized detection methods.

Moreover, we aimed to provide deeper insights and a better understanding of key features crucial in identifying rare states. Therefore, we explored the use of AE's RE, evaluating its effec-

tiveness across tasks of varying complexity levels: visual distraction detection (low complexity) and psychotic relapse prediction (high complexity). Our methodology involved calculating correlation scores between data annotations and the RE of samples, both normal and abnormal. Higher correlations indicated greater feature relevance. This approach efficiently identified the most and least influential features in visual distraction detection. However, with the more complex dataset for relapse prediction, our findings unveiled varying and low feature importance across individual patients. This emphasized the essentiality of personalized models in mental health monitoring and highlighted the crucial need for more pertinent features, especially in the context of psychotic relapse prediction.

In the final chapter, our primary focus was integrating data difficulty indicators into the fusion model to augment its performance. These indicators are extracted from AE trained for each modality by calculating the averaged RE for each sample. These indicators offer additional insights into the dissimilarity of the sample from the observed training data and, hence, their difficulty. We evaluated the approach and the impact of these indicators in the application of predicting emotions, specifically in stressful scenarios. Our findings highlighted how these indicators substantially improved the fusion process involving multiple modalities. This suggests their potential to enhance robustness, especially in environments where a single modality might be compromised.

A consistent finding throughout these studies was the significance of the AE's RE, which was examined across three distinct contexts, showcasing its substantial value. It functioned as an anomaly indicator, aiding in the identification of rare or abnormal patterns. Additionally, it served as a metric for feature importance, contributing to explainability, and acted as an augmenting feature for enhancing the fusion of multiple information sources. One of its distinct advantages over traditional anomaly scores is the ability to dissect the RE, allowing verification of the contributing features.

Our research has highlighted the adaptability of anomaly detection methodologies across diverse contexts. In Chapter 3, we explored it in an unsupervised approach, which is crucial for identifying rare states when data from one class is unavailable. In Chapter 4, we extended its use in a weakly supervised approach, utilizing limited labeled data from the minority class for tasks like feature selection and explainability. Finally, Chapter 5 explored its application in a supervised setting, particularly in information fusion.

6.2 Perspectives

While working on this thesis, we have identified several challenges along with ideas for future work. In this section, we delve into perspectives pertaining to datasets, model evaluation, and model improvement.

6.2.1 Datasets related perspectives

- The absence of chronological data in the eprevention dataset represents an intriguing avenue for exploration. Investigating the impact of the chronological order of pre-relapse and relapse days on anomaly scores could yield valuable insights. For instance, observing whether anomaly scores increase significantly in the pre-relapse phase might serve as an early warning indicator. The annotation do not include the specific disorder of the patient. Exploring the performance of unsupervised models concerning specific disorders could deepen our understanding of their predictive capabilities across different mental health conditions. This exploration could significantly deepen our understanding of their predictive capabilities concerning distinct disorders. Such insights could be particularly valuable in feature selection and explainability efforts, allowing for the identification of which fea-

tures correspond to specific relapses, whether globally or for individual patients or groups, facilitating tailored treatment strategies.

- A fundamental challenge encountered in this thesis revolves around the scarcity of extensive databases, a prevalent issue in various domains within affective computing. This scarcity is particularly problematic when addressing the detection of rare behaviors, such as anxiety attacks or relapses. The limited availability of larger datasets poses a significant obstacle in exploring anomaly detection techniques, especially concerning mental health applications, where datasets are both small and exceedingly complex. Therefore, future work can rely on datasets that depict normal behavior without the need for manual labeling.
- Anomaly detection methods in affective computing, while showing promise, still linger in a relatively early developmental phase. A common challenge observed across various works, including our own, is the reliance on individually defined norms or abnormalities. While this approach aids method exploration, it restricts comparability across studies. The subjectivity in defining what's normal or abnormal, often context-dependent and relative, necessitates standardization in datasets for anomaly detection in affective computing. Initiatives like the E-prevention challenge provide an opportunity for researchers to benchmark various agnostic methods using the same dataset and specific metrics. Standardized datasets and shared evaluation frameworks could substantially enhance comparability and collective advancements within the research community. Moreover, fostering collaboration and extending datasets to include multiple modalities would further enrich research exploration and enhance the efficacy of anomaly detection methods in affective computing.

6.2.2 Model and approach evaluation perspectives

- One limitation we encountered in evaluating our approaches was the restricted size of the datasets used. To improve the robustness and generalizability of our findings, future studies should focus on validating the efficiency of our approaches on larger and more diverse existing datasets.
- In our exploration of driver behavior monitoring, our application of anomaly detection methods proved effective in identifying specific dangerous behaviors, such as visual distractions. However, a potential avenue for future research lies in validating the adaptability of our models to recognize a broader spectrum of hazardous behaviors. Models can be trained on normal data only using signals related not just to distractions but also to fatigue and drowsiness. Subsequently, testing these trained models across various behaviors could ascertain their capability to detect a wider range of hazardous behaviors beyond visual distractions. Such an extension could significantly enhance the versatility and applicability of anomaly detection techniques in ensuring comprehensive driver safety.
- Exploring the wider applicability of the feature selection method in diverse regression tasks could provide a more comprehensive assessment of its effectiveness and robustness. Specifically, delving deeper into its explainability potential within tasks could yield valuable insights into how and why certain features are deemed important by the model. Additionally, conducting a comparative evaluation with other commonly used feature selection techniques, particularly in terms of computational efficiency, would be beneficial. Understanding how this method stacks up against existing techniques could provide a clearer picture of its advantages and areas for improvement.
- The proposed multimodal fusion scheme using reconstruction errors and temporal dependencies have shown promise in our study. To further validate its efficiency, an interesting approach would involve intentionally adding artificial noise to certain modalities. This

would allow us to observe how the fusion model responds, verifying if it assigns less importance to the noisy modalities and prioritizes the non-noisy ones. This would verify our hypothesis that RE helps the fusion to use most reliable modalities. This kind of experimentation could potentially enhance our understanding of the model's robustness and adaptability. Moreover, expanding the application of these fusion techniques to other domains could be valuable. For instance, investigating their efficacy in different contexts or industries, such as healthcare or manufacturing, might uncover their broader applicability and potential advantages in various real-world scenarios.

6.2.3 Model improvement perspectives

- Given the significant strides witnessed in transformer-based techniques across diverse AI domains, their potential application in anomaly detection using larger datasets, especially in processing speech and video data, emerges as a compelling direction for future investigation. Integrating transformers into anomaly detection frameworks could introduce several advantages. One approach could involve leveraging transformer architectures, like BERT or GPT, to encode multimodal information from speech and video data. These models excel in capturing complex patterns and dependencies within sequences, which could prove beneficial in detecting anomalies, particularly in multimodal data streams.
- For personalized models in predicting psychotic relapse, an enhancement strategy could involve an initial step of training a comprehensive, general model using data aggregated from all patients. Subsequently, fine-tuning this general model based on individual patient data could potentially improve predictive accuracy and robustness by tailoring the model to specific patient characteristics and patterns.
- An ongoing challenge within anomaly detection in human behavior revolves around its inherently dynamic nature. Models designed for predicting events like psychotic relapse might trigger false alarms when confronted with alterations in individuals' daily routines. To enhance the robustness of these models in the face of such changes, exploring methodologies like active learning or incremental learning could offer promising solutions. Implementing these adaptive learning techniques might assist in fine-tuning the models to better adapt to shifts in behavior, thereby mitigating false alarms and improving the accuracy of predictions.

Bibliography

[1] R. Gervasi, F. Barravecchia, L. Mastrogiacomo, and F. Franceschini. "Applications of affective computing in human-robot interaction: State-of-art and challenges for manufacturing". In: *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture* 237.6-7 (2023), pp. 815–832.

[2] S. Afzal et al. "A Comprehensive Survey on Affective Computing; Challenges, Trends, Applications, and Future Directions". In: *arXiv preprint arXiv:2305.07665* (2023).

[3] T. Olugbade et al. "Touch Technology in Affective Human–, Robot–, and Virtual–Human Interactions: A Survey". In: *Proceedings of the IEEE* (2023).

[4] Y. S. Can, B. Mahesh, and E. André. "Approaches, applications, and challenges in physiological emotion recognition—a tutorial overview". In: *Proceedings of the IEEE* (2023).

[5] P. Wu et al. "Automatic depression recognition by intelligent speech signal processing: A systematic survey". In: *CAAI Transactions on Intelligence Technology* 8.3 (2023), pp. 701–711.

[6] R. W. Picard. "MIT Media Laboratory; Perceptual Computing; 20 Ames St., Cambridge, MA 02139 picard@media.mit.edu, http://www.media.mit.edu/~picard/". en. In: ().

[7] P. Ekman. "Basic emotions". In: *Handbook of cognition and emotion*. New York, NY, US: John Wiley & Sons Ltd, 1999, pp. 45–60.

[8] P. Ekman. "Pictures of facial affect". In: *(No Title)* (1976).

[9] R. Plutchik. "A general psychoevolutionary theory of emotion". In: *Theories of emotion*. Elsevier, 1980, pp. 3–33. URL: https://www.sciencedirect.com/science/article/pii/B9780125587013500077 (visited on 10/30/2023).

[10] J. Russell. "A Circumplex Model of Affect". In: *Journal of Personality and Social Psychology* 39 (Dec. 1980), pp. 1161–1178.

[11] A. Mehrabian. "Basic dimensions for a general psychological theory: Implications for personality, social, environmental, and developmental studies". In: (1980). URL: https://philpapers.org/rec/MEHBDF (visited on 11/02/2023).

[12] H. Bohy, K. El Haddad, and T. Dutoit. "A new perspective on smiling and laughter detection: Intensity levels matter". In: *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE. 2022, pp. 1–8.

[13] K. A. Islam et al. "Online Detection of Attentiveness of Students with Special Needs". In: *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*. ISSN: 2156-8111. Oct. 2022, pp. 1–8.

[14] T. Olugbade et al. "Emopain (at) home: Dataset and automatic assessment within functional activity for chronic pain rehabilitation". In: *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE. 2022, pp. 1–8.

[15] P. Mann, E. H. Matsushima, and A. Paes. "Detecting Depression from Social Media Data as a Multiple-Instance Learning Task". In: *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*. ISSN: 2156-8111. Oct. 2022, pp. 1–8.

[16] A. Al-Ezzi et al. “Machine learning for the detection of social anxiety disorder using effective connectivity and graph theory measures”. In: *Frontiers in Psychiatry* 14 (2023). URL: https://www.frontiersin.org/articles/10.3389/fpsyt.2023.1155812 (visited on 08/31/2023).

[17] P. Mikolas et al. “Prediction of estimated risk for bipolar disorder using machine learning and structural MRI features”. en. In: *Psychological Medicine* (May 2023). Publisher: Cambridge University Press, pp. 1–11. URL: https://www.cambridge.org/core/journals/psychological-medicine/article/prediction-of-estimated-risk-for-bipolar-disorder-using-machine-learning-and-structural-mri-features/36C2E6ABABE679A885AB8A6AA83E452E (visited on 08/31/2023).

[18] R. Deng, L. Panl, and C. Clavel. “Domain Adaptation for Stance Detection towards Unseen Target on Social Media”. In: *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE. 2022, pp. 1–8.

[19] G. Cen et al. “Exploring multimodal fusion for continuous protective behavior detection”. In: *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE. 2022, pp. 1–8.

[20] B. Luo, R. Y. K. Lau, and C. Li. “Emotion-regulatory chatbots for enhancing consumer servicing: An interpersonal emotion management approach”. In: *Information & Management* 60.5 (July 2023), p. 103794. URL: https://www.sciencedirect.com/science/article/pii/S0378720623000423 (visited on 08/31/2023).

[21] O. Rudovic et al. “Personalized machine learning for robot perception of affect and engagement in autism therapy”. en. In: *Science Robotics* 3.19 (June 2018), eaao6760. URL: https://robotics.sciencemag.org/lookup/doi/10.1126/scirobotics.aao6760 (visited on 07/23/2021).

[22] W. Alsabhan. “Human–Computer Interaction with a Real-Time Speech Emotion Recognition with Ensembling Techniques 1D Convolution Neural Network and Attention”. en. In: *Sensors* 23.3 (Jan. 2023). Number: 3 Publisher: Multidisciplinary Digital Publishing Institute, p. 1386. URL: https://www.mdpi.com/1424-8220/23/3/1386 (visited on 08/31/2023).

[23] R. Majethia, V. P. Sharma, and R. Dwaraghanath. “Mental Health Indices as Biomarkers for Assistive Mental Healthcare in University Students”. In: *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*. ISSN: 2156-8111. Oct. 2022, pp. 1–8.

[24] S. Bhatnagar, J. Agarwal, and O. R. Sharma. “Detection and classification of anxiety in university students through the application of machine learning”. In: *Procedia Computer Science*. International Conference on Machine Learning and Data Engineering 218 (Jan. 2023), pp. 1542–1550. URL: https://www.sciencedirect.com/science/article/pii/S1877050923001321 (visited on 08/31/2023).

[25] W. A. Campos-Ugaz et al. “An Overview of Bipolar Disorder Diagnosis Using Machine Learning Approaches: Clinical Opportunities and Challenges”. en. In: *Iranian Journal of Psychiatry* (Apr. 2023). URL: https://publish.kne-publishing.com/index.php/IJPS/article/view/12372 (visited on 08/31/2023).

[26] N. Cosmann et al. “The value of mood measurement for regulating negative influences of social media usage: A case study of TikTok”. In: *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*. ISSN: 2156-8111. Oct. 2022, pp. 1–7.

[27] P. Zhernova, Y. Bodyanskiy, B. Yatsenko, and I. Zavgorodnii. “Detection and Prevention of Professional Burnout Using Machine Learning Methods”. In: *2020 IEEE 15th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET)*. Feb. 2020, pp. 218–221.

[28] C. Ding, Y. Zhang, and T. Ding. “A systematic hybrid machine learning approach for stress prediction”. en. In: *PeerJ Computer Science* 9 (Feb. 2023). Publisher: PeerJ Inc., e1154. URL: https://peerj.com/articles/cs-1154 (visited on 08/31/2023).

[29] D. Dritsa and N. Biloria. "Context- and Movement-Aware Analysis of Physiological Responses in The Urban Environment Using Wearable Sensors". In: *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*. ISSN: 2156-8111. Oct. 2022, pp. 1–8.

[30] M. Ebrahim Shaik. "A systematic review on detection and prediction of driver drowsiness". In: *Transportation Research Interdisciplinary Perspectives* 21 (Sept. 2023), p. 100864. URL: https://www.sciencedirect.com/science/article/pii/S2590198223001112 (visited on 08/31/2023).

[31] G. K. Sahoo, S. K. Das, and P. Singh. "A deep learning-based distracted driving detection solution implemented on embedded system". en. In: *Multimedia Tools and Applications* 82.8 (Mar. 2023), pp. 11697–11720. URL: https://doi.org/10.1007/s11042-022-13450-6 (visited on 08/31/2023).

[32] A. Tavakoli, V. Balali, and A. Heydarian. "A Multimodal Approach for Monitoring Driving Behavior and Emotions". In: *Mineta Transportation Institute Publications* (July 2020). URL: https://scholarworks.sjsu.edu/mti_publications/312.

[33] L. Shen, M. Wang, and R. Shen. "Affective e-Learning: Using "motional" data to improve learning in pervasive learning environment". English. In: *Educational Technology and Society* 12.2 (2009), pp. 176–189.

[34] S. R. Rathi and Y. D. Deshpande. "Embedding Affect Awareness in e-Learning: A Systematic Outline of the Literature". en. In: *AI, IoT, Big Data and Cloud Computing for Industry 4.0*. Ed. by A. Neustein, P. N. Mahalle, P. Joshi, and G. R. Shinde. Signals and Communication Technology. Cham: Springer International Publishing, 2023, pp. 39–63. URL: https://doi.org/10.1007/978-3-031-29713-7_3 (visited on 08/30/2023).

[35] Y. Wang. "Affective State Analysis During Online Learning Based on Learning Behavior Data". English. In: *Technology, Knowledge and Learning* 28.3 (2023), pp. 1063–1078.

[36] Y. Xu et al. "Spontaneous visual database for detecting learning-centered emotions during online learning". English. In: *Image and Vision Computing* 136 (2023).

[37] P. Bruno, V. Melnyk, and F. Völckner. "Temperature and emotions: Effects of physical temperature on responses to emotional advertising". In: *International Journal of Research in Marketing* 34.1 (Mar. 2017), pp. 302–320. URL: https://www.sciencedirect.com/science/article/pii/S0167811616301148 (visited on 08/30/2023).

[38] S. W. Naidoo, N. Naicker, S. S. Patel, and P. Govender. "Computer Vision: The Effectiveness of Deep Learning for Emotion Detection in Marketing Campaigns". en. In: *International Journal of Advanced Computer Science and Applications* 13.5 (2022). URL: http://thesai.org/Publications/ViewPaper?Volume=13&Issue=5&Code=IJACSA&SerialNo=100 (visited on 08/31/2023).

[39] P. Ładyżyński, K. Żbikowski, and P. Gawrysiak. "Direct marketing campaigns in retail banking with the use of deep learning and random forests". In: *Expert Systems with Applications* 134 (Nov. 2019), pp. 28–35. URL: https://www.sciencedirect.com/science/article/pii/S0957417419303471 (visited on 08/31/2023).

[40] M. Paulo, V. L. Miguéis, and I. Pereira. "Leveraging email marketing: Using the subject line to anticipate the open rate". In: *Expert Systems with Applications* 207 (Nov. 2022), p. 117974. URL: https://www.sciencedirect.com/science/article/pii/S0957417422012040 (visited on 08/31/2023).

[41] T. J. Wen, C.-H. Chuan, W. S. Tsai, and J. Yang. "Decoding Emotional (In)Congruency: A Computational Approach Toward Ad Placement on YouTube". en. In: *Journal of Interactive Marketing* 57.3 (Aug. 2022). Publisher: SAGE Publications, pp. 421–441. URL: https://doi.org/10.1177/10949968221095546 (visited on 08/31/2023).

[42] A. Fernández-Caballero et al. "Smart environment architecture for emotion detection and regulation". In: *Journal of Biomedical Informatics* 64 (Dec. 2016), pp. 55–73. URL: https://www.sciencedirect.com/science/article/pii/S1532046416301289 (visited on 08/31/2023).

[43] M. S. Benlamine, A. Dufresne, M. H. Beauchamp, and C. Frasson. "BARGAIN: behavioral affective rule-based games adaptation interface–towards emotionally intelligent games: application on a virtual reality environment for socio-moral development". en. In: *User Modeling and User-Adapted Interaction* 31.2 (Apr. 2021), pp. 287–321. URL: https://doi.org/10.1007/s11257-020-09286-0 (visited on 08/31/2023).

[44] M. Cheng et al. "Computer-Aided Autism Spectrum Disorder Diagnosis With Behavior Signal Processing". In: *IEEE Transactions on Affective Computing* (2023). Conference Name: IEEE Transactions on Affective Computing, pp. 1–18.

[45] M. Alharbi and S. Huang. "A survey of incorporating affective computing for human-system co-adaptation". In: *Proceedings of the 2nd World Symposium on Software Engineering*. 2020, pp. 72–79.

[46] S. Alisamir and F. Ringeval. "On the Evolution of Speech Representations for Affective Computing: A brief history and critical overview". In: *IEEE Signal Processing Magazine* 38.6 (Nov. 2021). Conference Name: IEEE Signal Processing Magazine, pp. 12–21.

[47] F. Eyben et al. "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing". In: *IEEE Transactions on Affective Computing* 7.2 (Apr. 2016). Number: 2, pp. 190–202.

[48] S. Pancoast and M. Akbacak. "Bag-of-Audio-Words Approach for Multimedia Event Classification". en. In: (2012).

[49] T. Jaakkola and D. Haussler. "Exploiting Generative Models in Discriminative Classifiers". In: *Advances in Neural Information Processing Systems*. Vol. 11. MIT Press, 1998. URL: https://proceedings.neurips.cc/paper_files/paper/1998/hash/db1915052d15f7815c8b88e879465a1e-Abstract.html (visited on 09/13/2023).

[50] G. Trigeorgis et al. "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network". en. In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Shanghai: IEEE, Mar. 2016, pp. 5200–5204. URL: http://ieeexplore.ieee.org/document/7472669/ (visited on 07/17/2021).

[51] S. Hershey et al. "CNN architectures for large-scale audio classification". en. In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. New Orleans, LA: IEEE, Mar. 2017, pp. 131–135. URL: http://ieeexplore.ieee.org/document/7952132/ (visited on 07/17/2021).

[52] S. Amiriparian et al. "Snore Sound Classification Using Image-Based Deep Spectrum Features". en. In: *Interspeech 2017*. ISCA, Aug. 2017, pp. 3512–3516. URL: http://www.isca-speech.org/archive/Interspeech_2017/abstracts/0434.html (visited on 07/21/2021).

[53] M. Neumann and N. T. Vu. "Improving Speech Emotion Recognition with Unsupervised Representation Learning on Unlabeled Speech". In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. ISSN: 2379-190X. May 2019, pp. 7390–7394.

[54] S. Latif, R. Rana, J. Qadir, and J. Epps. *Variational Autoencoders for Learning Latent Representations of Speech Emotion: A Preliminary Study*. arXiv:1712.08708 [cs, eess, stat]. July 2020. URL: http://arxiv.org/abs/1712.08708 (visited on 09/13/2023).

[55] S. Sahu et al. *Adversarial Auto-encoders for Speech Based Emotion Recognition*. arXiv:1806.02146 [cs, stat]. June 2018. URL: http://arxiv.org/abs/1806.02146 (visited on 09/13/2023).

[56] Q. Mao, M. Dong, Z. Huang, and Y. Zhan. "Learning Salient Features for Speech Emotion Recognition Using Convolutional Neural Networks". In: *IEEE Transactions on Multimedia* 16.8 (Dec. 2014). Conference Name: IEEE Transactions on Multimedia, pp. 2203–2213.

[57] A. Rasmus et al. “Semi-supervised Learning with Ladder Networks”. In: *Advances in Neural Information Processing Systems*. Vol. 28. Curran Associates, Inc., 2015. URL: https://proceedings.neurips.cc/paper_files/paper/2015/hash/378a063b8fdb1db941e34f4bde584c7d-Abstract.html (visited on 09/13/2023).

[58] A. v. d. Oord, Y. Li, and O. Vinyals. *Representation Learning with Contrastive Predictive Coding*. arXiv:1807.03748 [cs, stat]. Jan. 2019. URL: http://arxiv.org/abs/1807.03748 (visited on 09/13/2023).

[59] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli. “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations”. In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 12449–12460. URL: https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html (visited on 10/04/2022).

[60] S. Chen et al. “Wavlm: Large-scale self-supervised pre-training for full stack speech processing”. In: *IEEE Journal of Selected Topics in Signal Processing* 16.6 (2022), pp. 1505–1518.

[61] S.-w. Yang et al. “Superb: Speech processing universal performance benchmark”. In: *arXiv preprint arXiv:2105.01051* (2021).

[62] M. Polignano, P. Basile, M. De Gemmis, and G. Semeraro. “A Comparison of Word-Embeddings in Emotion Detection from Text using BiLSTM, CNN and Self-Attention”. en. In: *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*. Larnaca Cyprus: ACM, June 2019, pp. 63–68. URL: https://dl.acm.org/doi/10.1145/3314183.3324983 (visited on 09/14/2023).

[63] L. Peng et al. “Customising General Large Language Models for Specialised Emotion Recognition Tasks”. In: *arXiv preprint arXiv:2310.14225* (2023).

[64] W. Shen, J. Chen, X. Quan, and Z. Xie. “Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. Issue: 15. 2021, pp. 13789–13797.

[65] X. Wang et al. “A novel end-to-end speech emotion recognition network with stacked transformer layers”. In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6289–6293.

[66] J. Devlin, M.-W. Chang, K. Lee, and K. N. Toutanova. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: 2018. URL: https://arxiv.org/abs/1810.04805 (visited on 09/14/2023).

[67] A. Radford et al. “Language Models are Unsupervised Multitask Learners”. en. In: ().

[68] Z. Yang et al. “XLNet: Generalized Autoregressive Pretraining for Language Understanding”. In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html (visited on 09/14/2023).

[69] Y. Liu et al. “RoBERTa: A Robustly Optimized BERT Pretraining Approach”. en. In: (Sept. 2019). URL: https://openreview.net/forum?id=SyxS0T4tvS (visited on 09/14/2023).

[70] K Ezzameli and H Mahersia. “Emotion recognition from unimodal to multimodal analysis: A review”. In: *Information Fusion* (2023), p. 101847.

[71] S. K. Khare, V. Blanes-Vidal, E. S. Nadimi, and U. R. Acharya. “Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations”. In: *Information Fusion* (2023), p. 102019.

[72] A. Mehrabian. “Communication Without Words”. In: *Communication Theory*. 2nd ed. Num Pages: 8. Routledge, 2008.

[73] P. Ekman, W. V. Friesen, J. C. Hager, and A. H. F. F. *Facial action coding system*. English. ISBN: 9780931835018 Place: Salt Lake City, UT OCLC: 58460796. 2002.

[74] T. Ojala, M. Pietikainen, and T. Maenpaa. "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns". en. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24.7 (July 2002), pp. 971–987. URL: http://ieeexplore.ieee.org/document/1017623/ (visited on 07/18/2021).

[75] D. Gabor. "Theory of communication. Part 1: The analysis of information". en. In: *Journal of the Institution of Electrical Engineers - Part III: Radio and Communication Engineering* 93.26 (Nov. 1946). Publisher: IET Digital Library, pp. 429–441. URL: https://digital-library.theiet.org/content/journals/10.1049/ji-3-2.1946.0074 (visited on 07/18/2021).

[76] F. Ringeval et al. "AVEC 2019 Workshop and Challenge: State-of-Mind, Detecting Depression with AI, and Cross-Cultural Affect Recognition". en. In: *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop - AVEC '19*. Nice, France: ACM Press, 2019, pp. 3–12. URL: http://dl.acm.org/citation.cfm?doid=3347320.3357688 (visited on 07/18/2021).

[77] L. Stappen et al. "MuSe 2020 Challenge and Workshop: Multimodal Sentiment Analysis, Emotion-target Engagement and Trustworthiness Detection in Real-life Media: Emotional Car Reviews in-the-wild". en. In: *Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop*. Seattle WA USA: ACM, Oct. 2020, pp. 35–44. URL: https://dl.acm.org/doi/10.1145/3423327.3423673 (visited on 07/18/2021).

[78] K. Simonyan and A. Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". en. In: *arXiv:1409.1556 [cs]* (Apr. 2015). arXiv: 1409.1556. URL: http://arxiv.org/abs/1409.1556 (visited on 07/18/2021).

[79] J. A. Coan and J. J. B. Allen. *Handbook of Emotion Elicitation and Assessment*. en. Google-Books-ID: 9xhnDAAAQBAJ. Oxford University Press, USA, Apr. 2007.

[80] W. Lin et al. "Looking At The Body: Automatic Analysis of Body Gestures and Self-Adaptors in Psychological Distress". en. In: *arXiv:2007.15815 [cs]* (July 2020). arXiv: 2007.15815. URL: http://arxiv.org/abs/2007.15815 (visited on 03/15/2022).

[81] Y.-L. Hsu, J.-S. Wang, W.-C. Chiang, and C.-H. Hung. "Automatic ECG-Based Emotion Recognition in Music Listening". In: *IEEE Transactions on Affective Computing* 11.1 (Jan. 2020). Conference Name: IEEE Transactions on Affective Computing, pp. 85–99.

[82] S. Shilaskar, D. Bobby, A. Dusane, and S. Bhatlawande. "Fusion of EEG, EMG, and ECG Signals for Accurate Recognition of Pain, Happiness, and Disgust". In: *2023 International Conference in Advances in Power, Signal, and Information Technology (APSIT)*. June 2023, pp. 142–149.

[83] A. Alqaraawi, A. Alwosheel, and A. Alasaad. "Heart rate variability estimation in photoplethysmography signals using Bayesian learning approach". In: *Healthcare technology letters* 3.2 (2016), pp. 136–142.

[84] F. Gasparini, A. Grossi, and S. Bandini. "A Deep Learning Approach to Recognize Cognitive Load using PPG Signals". en. In: *Proceedings of the 14th PErvasive Technologies Related to Assistive Environments Conference*. Corfu Greece: ACM, June 2021, pp. 489–495. URL: https://dl.acm.org/doi/10.1145/3453892.3461625 (visited on 09/15/2023).

[85] S. Ollander, C. Godin, A. Campagne, and S. Charbonnier. "A comparison of wearable and stationary sensors for stress detection". In: *2016 IEEE International Conference on systems, man, and Cybernetics (SMC)*. IEEE. 2016, pp. 004362–004366.

[86] A. Amidei et al. "Driver Drowsiness Detection: A Machine Learning Approach on Skin Conductance". en. In: *Sensors* 23.8 (Jan. 2023). Number: 8 Publisher: Multidisciplinary Digital Publishing Institute, p. 4004. URL: https://www.mdpi.com/1424-8220/23/8/4004 (visited on 09/15/2023).

[87] P. Zontone et al. "Stress detection through electrodermal activity (EDA) and electrocardiogram (ECG) analysis in car drivers". In: *2019 27th European Signal Processing Conference (EUSIPCO)*. IEEE. 2019, pp. 1–5.

[88] T. Hussain, S. Ullah, R. Fernández-García, and I. Gil. "Wearable sensors for respiration monitoring: A review". In: *Sensors* 23.17 (2023), p. 7518.

[89] F. R. Ihmig et al. "On-line anxiety level detection from biosignals: Machine learning based on a randomized controlled trial with spider-fearful individuals". en. In: *PLOS ONE* 15.6 (June 2020). Publisher: Public Library of Science, e0231517. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0231517 (visited on 09/15/2023).

[90] S. Pourmohammadi and A. Maleki. "Stress detection using ECG and EMG signals: A comprehensive study". In: *Computer Methods and Programs in Biomedicine* 193 (Sept. 2020), p. 105482. URL: https://www.sciencedirect.com/science/article/pii/S0169260719320528 (visited on 09/14/2023).

[91] P. Giannakopoulos, P. Missonnier, G. Gold, and A. Michon. "Electrophysiological markers of rapid cognitive decline in mild cognitive impairment". In: *Dementia in Clinical Practice* 24 (2009). Publisher: Karger Publishers, pp. 39–46.

[92] *A comprehensive survey on emotion recognition based on electroencephalograph (EEG) signals | SpringerLink*. URL: https://link.springer.com/article/10.1007/s11042-023-14489-9 (visited on 09/14/2023).

[93] B. Venkata Phanikrishna, A. Jaya Prakash, and C. Suchismitha. "Deep Review of Machine Learning Techniques on Detection of Drowsiness Using EEG Signal". In: *IETE Journal of Research* 69.6 (Aug. 2023). Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/03772063.2021.1913070, pp. 3104–3119. URL: https://doi.org/10.1080/03772063.2021.1913070 (visited on 09/14/2023).

[94] B. Brousseau, J. Rose, and M. Eizenman. "Hybrid Eye-Tracking on a Smartphone with CNN Feature Extraction and an Infrared 3D Model". en. In: *Sensors* 20.2 (Jan. 2020), p. 543. URL: https://www.mdpi.com/1424-8220/20/2/543 (visited on 05/06/2022).

[95] Y. Liang, M. Reyes, and J. Lee. "Real-Time Detection of Driver Cognitive Distraction Using Support Vector Machines". In: *Intelligent Transportation Systems, IEEE Transactions on* 8 (July 2007), pp. 340–350.

[96] M. Zhang et al. "SafeDrive: Online driving anomaly detection from large-scale vehicle data". In: *IEEE Transactions on Industrial Informatics* 13.4 (2017), pp. 2087–2096.

[97] M. R. Islam et al. "Depression detection from social network data using machine learning techniques". In: *Health information science and systems* 6 (2018), pp. 1–12.

[98] R. Razavi, A. Gharipour, and M. Gharipour. "Depression screening using mobile phone usage metadata: a machine learning approach". In: *Journal of the American Medical Informatics Association* 27.4 (2020), pp. 522–530.

[99] M. Panagiotou et al. "A comparative study of autoencoder architectures for mental health analysis using wearable sensors data". en. In: *2022 30th European Signal Processing Conference (EUSIPCO)*. Belgrade, Serbia: IEEE, Aug. 2022, pp. 1258–1262. URL: https://ieeexplore.ieee.org/document/9909697/ (visited on 02/13/2023).

[100] H. Al Osman and T. H. Falk. "Multimodal affect recognition: Current approaches and challenges". In: *Emotion and attention recognition based on biological signals and images* (2017), pp. 59–86.

[101] N. Ahmed, Z. Al Aghbari, and S. Girija. "A systematic survey on multimodal emotion recognition using learning algorithms". In: *Intelligent Systems with Applications* 17 (2023), p. 200171.

[102] V. Chaparro et al. "Emotion recognition from EEG and facial expressions: a multimodal approach". In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2018, pp. 530–533.

[103] J. Huang et al. "Multimodal Continuous Emotion Recognition with Data Augmentation Using Recurrent Neural Networks". en. In: *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*. Seoul Republic of Korea: ACM, Oct. 2018, pp. 57–64. URL: https://dl.acm.org/doi/10.1145/3266302.3266304 (visited on 07/19/2021).

[104] L. Sun et al. "Multi-modal Continuous Dimensional Emotion Recognition Using Recurrent Neural Network and Self-Attention Mechanism". en. In: *Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop*. Seattle WA USA: ACM, Oct. 2020, pp. 27–34. URL: https://dl.acm.org/doi/10.1145/3423327.3423672 (visited on 07/16/2021).

[105] J. Kim et al. "Integrating information from speech and physiological signals to achieve emotional sensitivity". en. In: *Interspeech 2005*. ISCA, Sept. 2005, pp. 809–812. URL: https://www.isca-speech.org/archive/interspeech_2005/kim05c_interspeech.html (visited on 09/08/2023).

[106] H. Chen et al. "Efficient Spatial Temporal Convolutional Features for Audiovisual Continuous Affect Recognition". en. In: *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop - AVEC '19*. Nice, France: ACM Press, 2019, pp. 19–26. URL: http://dl.acm.org/citation.cfm?doid=3347320.3357690 (visited on 07/20/2021).

[107] S. Poria, E. Cambria, R. Bajpai, and A. Hussain. "A review of affective computing: From unimodal analysis to multimodal fusion". In: *Information fusion* 37 (2017), pp. 98–125.

[108] C. Busso et al. "IEMOCAP: Interactive emotional dyadic motion capture database". In: *Language resources and evaluation* 42 (2008), pp. 335–359.

[109] I. J. Goodfellow et al. "Challenges in representation learning: A report on three machine learning contests". In: *Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20*. Springer. 2013, pp. 117–124.

[110] D. DeVault et al. "SimSensei Kiosk: A virtual human interviewer for healthcare decision support". In: *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. 2014, pp. 1061–1068.

[111] P.-Y. Hsueh, P. Melville, and V. Sindhwani. "Data Quality from Crowdsourcing: A Study of Annotation Selection Criteria". In: *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*. Boulder, Colorado: Association for Computational Linguistics, June 2009, pp. 27–35. URL: https://aclanthology.org/W09-1904 (visited on 09/11/2023).

[112] A. Kittur, E. H. Chi, and B. Suh. "Crowdsourcing for Usability: Using Micro-Task Markets for Rapid, Remote, and Low-Cost User Measurements". en. In: ().

[113] J. H. Shen, A. Lapedriza, and R. W. Picard. "Unintentional affective priming during labeling may bias labels". In: *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE. 2019, pp. 587–593.

[114] S. Zhao et al. "Affective Computing for Large-scale Heterogeneous Multimedia Data: A Survey". en. In: *ACM Transactions on Multimedia Computing, Communications, and Applications* 15.3s (Nov. 2019), pp. 1–32. URL: https://dl.acm.org/doi/10.1145/3363560 (visited on 09/12/2023).

[115] M. Naegelin et al. "An interpretable machine learning approach to multimodal stress detection in a simulated office environment". In: *Journal of Biomedical Informatics* 139 (2023), p. 104299.

[116] M. Fang et al. "A multimodal fusion model with multi-level attention mechanism for depression detection". In: *Biomedical Signal Processing and Control* 82 (2023), p. 104561.

[117] A. Vinciarelli, M. Pantic, and H. Bourlard. "Social signal processing: Survey of an emerging domain". In: *Image and Vision Computing*. Visual and multimodal analysis of human spontaneous behaviour: 27.12 (Nov. 2009), pp. 1743–1759. URL: https://www.sciencedirect.com/science/article/pii/S0262885608002485 (visited on 09/12/2023).

[118] Lin Shu et al. "A Review of Emotion Recognition Using Physiological Signals". In: *Sensors (14248220)* 18.7 (2018), p. 2074.

[119] M. El Ayadi, M. S. Kamel, and F. Karray. "Survey on speech emotion recognition: Features, classification schemes, and databases". In: *Pattern Recognition* 44.3 (Mar. 2011), pp. 572–587. URL: https://www.sciencedirect.com/science/article/pii/S0031320310004619 (visited on 09/12/2023).

[120] P. Naga, S. D. Marri, and R. Borreo. "Facial emotion recognition methods, datasets and technologies: A literature survey". In: *Materials Today: Proceedings*. SI:5 NANO 2021 80 (Jan. 2023), pp. 2824–2828. URL: https://www.sciencedirect.com/science/article/pii/S2214785321048987 (visited on 09/12/2023).

[121] S. K. D'mello and J. Kory. "A Review and Meta-Analysis of Multimodal Affect Detection Systems". In: *ACM Computing Surveys* 47.3 (Feb. 2015), 43:1–43:36. URL: https://dl.acm.org/doi/10.1145/2682899 (visited on 09/12/2023).

[122] B. E. Stein and M. A. Meredith. *The Merging of the Senses*. en. Google-Books-ID: _r5NEAAAQBAJ. MIT Press, Jan. 1993.

[123] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. "A survey of affect recognition methods: audio, visual and spontaneous expressions". en. In: *Proceedings of the 9th international conference on Multimodal interfaces*. Nagoya Aichi Japan: ACM, Nov. 2007, pp. 126–133. URL: https://dl.acm.org/doi/10.1145/1322192.1322216 (visited on 09/11/2023).

[124] A. Gandhi et al. "Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions". In: *Information Fusion* 91 (2023), pp. 424–444.

[125] Y. LeCun, Y. Bengio, and G. Hinton. "Deep learning". en. In: *Nature* 521.7553 (May 2015). Number: 7553 Publisher: Nature Publishing Group, pp. 436–444. URL: https://www.nature.com/articles/nature14539 (visited on 09/11/2023).

[126] Z. Cui and W. Zheng. *Deep Learning Techniques Applied to Affective Computing*. Frontiers Media SA, 2023.

[127] E. Cambria, S. Poria, A. Gelbukh, and M. Thelwall. "Sentiment Analysis Is a Big Suitcase". In: *IEEE Intelligent Systems* 32.6 (Nov. 2017). Conference Name: IEEE Intelligent Systems, pp. 74–80.

[128] W. Samek et al. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. en. Google-Books-ID: j5yuDwAAQBAJ. Springer Nature, Sept. 2019.

[129] D. Archer. "Unspoken Diversity: Cultural Differences in Gestures". en. In: *Qualitative Sociology* 20.1 (Mar. 1997), pp. 79–105. URL: https://doi.org/10.1023/A:1024716331692 (visited on 09/08/2023).

[130] M. Akyunus, T. Gençöz, and B. T. Aka. "Age and sex differences in basic personality traits and interpersonal problems across young adulthood". en. In: *Current Psychology* 40.5 (May 2021), pp. 2518–2527. URL: https://doi.org/10.1007/s12144-019-0165-z (visited on 09/08/2023).

[131] Y. Weisberg, C. DeYoung, and J. Hirsh. "Gender Differences in Personality across the Ten Aspects of the Big Five". In: *Frontiers in Psychology* 2 (2011). URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2011.00178 (visited on 09/08/2023).

[132] S. Kiritchenko and S. Mohammad. "Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems". In: *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 43–53. URL: https://aclanthology.org/S18-2005 (visited on 09/08/2023).

[133] M. Diaz et al. "Addressing Age-Related Bias in Sentiment Analysis". en. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Montreal QC Canada: ACM, Apr. 2018, pp. 1–14. URL: https://dl.acm.org/doi/10.1145/3173574.3173986 (visited on 09/08/2023).

[134] A. K., M. P. Gangan, D. P., and L. V. L. "Towards an Enhanced Understanding of Bias in Pre-trained Neural Language Models: A Survey with Special Emphasis on Affective Bias". In: vol. 940. arXiv:2204.10365 [cs]. 2022, pp. 13–45. URL: http://arxiv.org/abs/2204.10365 (visited on 09/08/2023).

[135] A. Agarwal, P. Chattopadhyay, and L. Wang. "Privacy preservation through facial de-identification with simultaneous emotion preservation". en. In: *Signal, Image and Video Processing* 15.5 (July 2021), pp. 951–958. URL: https://link.springer.com/10.1007/s11760-020-01819-9 (visited on 09/12/2023).

[136] J. Domingo-Ferrer, O. Farràs, J. Ribes-González, and D. Sánchez. "Privacy-preserving cloud computing on sensitive data: A survey of methods, products and challenges". In: *Computer Communications* 140-141 (May 2019), pp. 38–60. URL: https://www.sciencedirect.com/science/article/pii/S0140366418310740 (visited on 09/12/2023).

[137] P. R. Kumar, P. H. Raj, and P. Jelciana. "Exploring Data Security Issues and Solutions in Cloud Computing". In: *Procedia Computer Science*. The 6th International Conference on Smart Computing and Communications 125 (Jan. 2018), pp. 691–697. URL: https://www.sciencedirect.com/science/article/pii/S1877050917328570 (visited on 09/12/2023).

[138] C. Cortes and V. Vapnik. "Support-vector networks". In: *Machine learning* 20 (1995), pp. 273–297.

[139] L. Shu et al. "A review of emotion recognition using physiological signals". In: *Sensors* 18.7 (2018), p. 2074.

[140] P. J. Bota, C. Wang, A. L. Fred, and H. P. Da Silva. "A review, current challenges, and future possibilities on emotion recognition using machine learning and physiological signals". In: *IEEE Access* 7 (2019), pp. 140990–141020.

[141] T. Cover and P. Hart. "Nearest neighbor pattern classification". In: *IEEE transactions on information theory* 13.1 (1967), pp. 21–27.

[142] W. Wen et al. "Emotion recognition based on multi-variant correlation of physiological signals". In: *IEEE Transactions on Affective Computing* 5.2 (2014), pp. 126–140.

[143] T. Iliou and C.-N. Anagnostopoulos. "Comparison of different classifiers for emotion recognition". In: *2009 13th Panhellenic Conference on Informatics*. IEEE. 2009, pp. 102–106.

[144] L. Breiman. "Random forests". In: *Machine learning* 45 (2001), pp. 5–32.

[145] H. Krishnan, M. S. Elayidom, and T Santhanakrishnan. "Emotion detection of tweets using naïve bayes classifier". In: *Emotion* 4.11 (2017), pp. 457–62.

[146] S. K. Bhakre and A. Bang. "Emotion recognition on the basis of audio signal using Naive Bayes classifier". In: *2016 International conference on advances in computing, communications and informatics (ICACCI)*. IEEE. 2016, pp. 2363–2367.

[147] A. Khan and U. K. Roy. "Emotion recognition using prosodie and spectral features of speech and Naïve Bayes Classifier". In: *2017 international conference on wireless communications, signal processing and networking (WiSPNET)*. IEEE. 2017, pp. 1017–1021.

[148] V. Kolodyazhniy et al. "An affective computing approach to physiological emotion specificity: Toward subject-independent and stimulus-independent classification of film-induced emotions". In: *Psychophysiology* 48.7 (2011), pp. 908–922.

[149] Y. S. Can, N. Chalabianloo, D. Ekiz, and C. Ersoy. "Continuous stress detection using wearable sensors in real life: Algorithmic programming contest case study". In: *Sensors* 19.8 (2019), p. 1849.

[150] A. Graves, A.-r. Mohamed, and G. Hinton. "Speech recognition with deep recurrent neural networks". en. In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. Vancouver, BC, Canada: IEEE, May 2013, pp. 6645–6649. URL: http://ieeexplore.ieee.org/document/6638947/ (visited on 10/25/2023).

[151] K. Cho et al. "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation". In: *arXiv:1406.1078 [cs, stat]* (Sept. 2014). arXiv: 1406.1078. URL: http://arxiv.org/abs/1406.1078 (visited on 08/06/2021).

[152] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling". In: *arXiv:1412.3555 [cs]* (Dec. 2014). arXiv: 1412.3555. URL: http://arxiv.org/abs/1412.3555 (visited on 08/06/2021).

[153] M. Ravanelli, P. Brakel, M. Omologo, and Y. Bengio. "Light Gated Recurrent Units for Speech Recognition". In: *IEEE Transactions on Emerging Topics in Computational Intelligence* 2.2 (Apr. 2018). Conference Name: IEEE Transactions on Emerging Topics in Computational Intelligence, pp. 92–102.

[154] W. Choi, M.-J. Kim, M.-S. Yum, and D.-H. Jeong. "Deep Convolutional Gated Recurrent Unit Combined with Attention Mechanism to Classify Pre-Ictal from Interictal EEG with Minimized Number of Channels". In: *Journal of Personalized Medicine* 12 (May 2022), p. 763.

[155] Z.-T. Liu, M.-T. Han, B.-H. Wu, and A. Rehman. "Speech emotion recognition based on convolutional neural network with attention-based bidirectional long short-term memory network and multi-task learning". In: *Applied Acoustics* 202 (2023), p. 109178.

[156] R. Febrian et al. "Facial expression recognition using bidirectional LSTM-CNN". In: *Procedia Computer Science* 216 (2023), pp. 39–47.

[157] S. Liu et al. "Multi-modal fusion network with complementarity and importance for emotion recognition". In: *Information Sciences* 619 (2023), pp. 679–694.

[158] N Priyadarshini and J Aravinth. "Emotion Recognition based on fusion of multimodal physiological signals using LSTM and GRU". In: *2023 Third International Conference on Secure Cyber Computing and Communication (ICSCCC)*. IEEE. 2023, pp. 1–6.

[159] P. Tzirakis, J. Zhang, and B. W. Schuller. "End-to-end speech emotion recognition using deep neural networks". In: *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2018, pp. 5089–5093.

[160] M.-y. Zhong et al. "EEG emotion recognition based on TQWT-features and hybrid convolutional recurrent neural network". In: *Biomedical Signal Processing and Control* 79 (2023), p. 104211.

[161] K. Sarvakar et al. "Facial emotion recognition using convolutional neural networks". In: *Materials Today: Proceedings* 80 (2023), pp. 3560–3564.

[162] H. Shahzad et al. "Hybrid Facial Emotion Recognition Using CNN-Based Features". In: *Applied Sciences* 13.9 (2023), p. 5572.

[163] H. P. Martinez, Y. Bengio, and G. N. Yannakakis. "Learning deep physiological models of affect". In: *IEEE Computational intelligence magazine* 8.2 (2013), pp. 20–33.

[164] S. Oh, J.-Y. Lee, and D. K. Kim. "The design of CNN architectures for optimal six basic emotion classification using multiple physiological signals". In: *Sensors* 20.3 (2020), p. 866.

[165] J. Zhao, X. Mao, and L. Chen. "Speech emotion recognition using deep 1D & 2D CNN LSTM networks". In: *Biomedical signal processing and control* 47 (2019), pp. 312–323.

[166] A. Othmani et al. "Towards robust deep neural networks for affect and depression recognition from speech". In: *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part II*. Springer. 2021, pp. 5–19.

[167] A. Vaswani et al. "Attention Is All You Need". In: *arXiv:1706.03762 [cs]* (Dec. 2017). arXiv: 1706.03762. URL: http://arxiv.org/abs/1706.03762 (visited on 11/08/2021).

[168] J. W. Rae, A. Potapenko, S. M. Jayakumar, and T. P. Lillicrap. *Compressive Transformers for Long-Range Sequence Modelling*. arXiv:1911.05507 [cs, stat]. Nov. 2019. URL: http://arxiv.org/abs/1911.05507 (visited on 09/06/2023).

[169] Z. Fan et al. “Mask Attention Networks: Rethinking and Strengthen Transformer”. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, June 2021, pp. 1692–1701. URL: https://aclanthology.org/2021.naacl-main.135 (visited on 09/06/2023).

[170] Y. Jiang, S. Chang, and Z. Wang. “TransGAN: Two Pure Transformers Can Make One Strong GAN, and That Can Scale Up”. In: *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc., 2021, pp. 14745–14758. URL: https://proceedings.neurips.cc/paper_files/paper/2021/hash/7c220a2091c26a7f5e9f1cfb099511e3-Abstract.html (visited on 09/06/2023).

[171] Z. Liu et al. “Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows”. en. In: 2021, pp. 10012–10022. URL: https://openaccess.thecvf.com/content/ICCV2021/html/Liu_Swin_Transformer_Hierarchical_Vision_Transformer_Using_Shifted_Windows_ICCV_2021_paper (visited on 09/06/2023).

[172] X. Chen et al. “Developing Real-Time Streaming Transformer Transducer for Speech Recognition on Large-Scale Dataset”. In: *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. ISSN: 2379-190X. June 2021, pp. 5904–5908.

[173] W. Yu, J. Zhou, H. Wang, and L. Tao. “SETransformer: Speech Enhancement Transformer”. en. In: *Cognitive Computation* 14.3 (May 2022), pp. 1152–1158. URL: https://doi.org/10.1007/s12559-020-09817-2 (visited on 09/06/2023).

[174] Y. Zhang et al. *Meta-Transformer: A Unified Framework for Multimodal Learning*. arXiv:2307.10802 [cs]. July 2023. URL: http://arxiv.org/abs/2307.10802 (visited on 09/06/2023).

[175] J. Liu et al. “A Transformer-based multimodal-learning framework using sky images for ultra-short-term solar irradiance forecasting”. In: *Applied Energy* 342 (July 2023), p. 121160. URL: https://www.sciencedirect.com/science/article/pii/S030626192300524X (visited on 09/06/2023).

[176] U. Naseem, I. Razzak, K. Musial, and M. Imran. “Transformer based deep intelligent contextual embedding for twitter sentiment analysis”. In: *Future Generation Computer Systems* 113 (2020), pp. 58–69.

[177] M. Jiang, J. Wu, X. Shi, and M. Zhang. “Transformer based memory network for sentiment analysis of web comments”. In: *IEEE Access* 7 (2019), pp. 179942–179953.

[178] D. Wang et al. “TETFN: A text enhanced transformer fusion network for multimodal sentiment analysis”. In: *Pattern Recognition* 136 (2023), p. 109259.

[179] M. Salehi et al. “A unified survey on anomaly, novelty, open-set, and out-of-distribution detection: Solutions and future challenges”. In: *arXiv preprint arXiv:2110.14051* (2021).

[180] X. Sun et al. “Conditional gaussian distribution learning for open set recognition”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 13480–13489.

[181] D. Miller, N. Sunderhauf, M. Milford, and F. Dayoub. “Class anchor clustering: A loss for distance-based open set recognition”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2021, pp. 3570–3578.

[182] D. Hendrycks, N. Carlini, J. Schulman, and J. Steinhardt. “Unsolved problems in ml safety”. In: *arXiv preprint arXiv:2109.13916* (2021).

[183] D. Hendrycks and M. Mazeika. “X-risk analysis for ai research”. In: *arXiv preprint arXiv:2206.05862* (2022).

[184] J. Yang, K. Zhou, Y. Li, and Z. Liu. “Generalized out-of-distribution detection: A survey”. In: *arXiv preprint arXiv:2110.11334* (2021).

[185] G. Pang, C. Shen, L. Cao, and A. V. D. Hengel. “Deep learning for anomaly detection: A review”. In: *ACM computing surveys (CSUR)* 54.2 (2021), pp. 1–38.

[186] L. Ruff et al. “A unifying review of deep and shallow anomaly detection”. In: *Proceedings of the IEEE* 109.5 (2021), pp. 756–795.

[187] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult. “Toward open set recognition”. In: *IEEE transactions on pattern analysis and machine intelligence* 35.7 (2012), pp. 1757–1772.

[188] T. E. Boult et al. “Learning and the unknown: Surveying steps toward open world recognition”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01. 2019, pp. 9801–9807.

[189] V. Chandola, A. Banerjee, and V. Kumar. “Anomaly detection: A survey”. en. In: *ACM Computing Surveys* 41.3 (), p. 58.

[190] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. “LOF: identifying density-based local outliers”. In: *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*. 2000, pp. 93–104.

[191] B. Schölkopf et al. “Support vector method for novelty detection.” In: *NIPS*. Vol. 12. Citeseer. 1999, pp. 582–588.

[192] D. M. Tax and R. P. Duin. “Support vector data description”. In: *Machine learning* 54.1 (2004), pp. 45–66.

[193] F. T. Liu, K. M. Ting, and Z.-H. Zhou. “Isolation forest”. In: *2008 eighth ieee international conference on data mining*. IEEE. 2008, pp. 413–422.

[194] *A Fast Algorithm for the Minimum Covariance Determinant Estimator: Technometrics: Vol 41, No 3*. URL: https://www.tandfonline.com/doi/abs/10.1080/00401706.1999.10485670 (visited on 09/04/2023).

[195] M. Sabokrou et al. “Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes”. en. In: *Computer Vision and Image Understanding* 172 (July 2018), pp. 88–97. URL: https://www.sciencedirect.com/science/article/pii/S1077314218300249 (visited on 09/08/2022).

[196] Z. Chen, C. K. Yeo, B. S. Lee, and C. T. Lau. “Autoencoder-based network anomaly detection”. In: *2018 Wireless Telecommunications Symposium (WTS)*. Apr. 2018, pp. 1–5.

[197] T. Schlegl et al. “Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery”. en. In: *Information Processing in Medical Imaging*. Ed. by M. Niethammer et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2017, pp. 146–157.

[198] Z. Zhu et al. “Using a VAE-SOM architecture for anomaly detection of flexible sensors in limb prosthesis”. In: *Journal of Industrial Information Integration* 35 (2023), p. 100490.

[199] S. Yan et al. “Hybrid robust convolutional autoencoder for unsupervised anomaly detection of machine tools under noises”. In: *Robotics and Computer-Integrated Manufacturing* 79 (2023), p. 102441.

[200] R. Liu et al. “Anomaly-GAN: A data augmentation method for train surface anomaly detection”. In: *Expert Systems with Applications* 228 (2023), p. 120284.

[201] D. P. Kingma and M. Welling. “Auto-encoding variational bayes”. In: *arXiv preprint arXiv:1312.6114* (2013).

[202] A. A. Pol et al. “Anomaly detection with conditional variational autoencoders”. In: *2019 18th IEEE international conference on machine learning and applications (ICMLA)*. IEEE. 2019, pp. 1651–1657.

[203] J. An and S. Cho. “Variational autoencoder based anomaly detection using reconstruction probability”. In: *Special lecture on IE* 2.1 (2015), pp. 1–18.

[204] Z. Xie et al. “Unsupervised Anomaly Detection on Microservice Traces through Graph VAE”. In: *Proceedings of the ACM Web Conference 2023*. 2023, pp. 2874–2884.

[205] S. Lin et al. "Anomaly detection for time series using vae-lstm hybrid model". In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Ieee. 2020, pp. 4322–4326.

[206] F. Ulger, S. E. Yuksel, and A. Yilmaz. "Anomaly detection for solder joints using β-VAE". In: *IEEE Transactions on Components, Packaging and Manufacturing Technology* 11.12 (2021), pp. 2214–2221.

[207] I. Goodfellow et al. "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems*. Vol. 27. Curran Associates, Inc., 2014. URL: https://proceedings.neurips.cc/paper_files/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html (visited on 09/05/2023).

[208] X. Xia et al. "GAN-based anomaly detection: A review". In: *Neurocomputing* 493 (July 2022), pp. 497–535. URL: https://www.sciencedirect.com/science/article/pii/S0925231221019482 (visited on 09/05/2023).

[209] L. Deecke et al. "Image Anomaly Detection with Generative Adversarial Networks". en. In: *Machine Learning and Knowledge Discovery in Databases*. Ed. by M. Berlingerio et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019, pp. 3–17.

[210] T. Schlegl et al. "f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks". In: *Medical Image Analysis* 54 (May 2019), pp. 30–44. URL: https://www.sciencedirect.com/science/article/pii/S1361841518302640 (visited on 09/06/2023).

[211] Y. Choi, H. Lim, H. Choi, and I.-J. Kim. "GAN-Based Anomaly Detection and Localization of Multivariate Time Series Data for Power Plant". In: *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)*. ISSN: 2375-9356. Feb. 2020, pp. 71–74.

[212] X. Xu, H. Zhao, H. Liu, and H. Sun. "LSTM-GAN-XGBOOST Based Anomaly Detection Algorithm for Time Series Data". In: *2020 11th International Conference on Prognostics and System Health Management (PHM-2020 Jinan)*. ISSN: 2166-5656. Oct. 2020, pp. 334–339.

[213] M. A. Bashar and R. Nayak. "TAnoGAN: Time Series Anomaly Detection with Generative Adversarial Networks". In: *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*. Dec. 2020, pp. 1778–1785.

[214] R. Xu and W. Yan. "A Comparison of GANs-Based Approaches for Combustor System Fault Detection". In: *2020 International Joint Conference on Neural Networks (IJCNN)*. ISSN: 2161-4407. July 2020, pp. 1–8.

[215] W. Huo, W. Wang, and W. Li. "AnomalyDetect: An Online Distance-Based Anomaly Detection Algorithm". en. In: *Web Services – ICWS 2019*. Ed. by J. Miller, E. Stroulia, K. Lee, and L.-J. Zhang. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019, pp. 63–79.

[216] H. Li and Y. Li. "Anomaly detection methods based on GAN: a survey". en. In: *Applied Intelligence* 53.7 (Apr. 2023), pp. 8209–8231. URL: https://doi.org/10.1007/s10489-022-03905-6 (visited on 09/06/2023).

[217] F. Di Mattia, P. Galeone, M. De Simoni, and E. Ghelfi. *A Survey on GANs for Anomaly Detection*. arXiv:1906.11632 [cs, stat]. Sept. 2021. URL: http://arxiv.org/abs/1906.11632 (visited on 09/05/2023).

[218] H. Zenati et al. "Efficient GAN-Based Anomaly Detection". en. In: (June 2018). URL: https://openreview.net/forum?id=BkXADmJDM (visited on 09/05/2023).

[219] J. Donahue, P. Krähenbühl, and T. Darrell. *Adversarial Feature Learning*. arXiv:1605.09782 [cs, stat]. Apr. 2017. URL: http://arxiv.org/abs/1605.09782 (visited on 09/05/2023).

[220] L. Y. "THE MNIST DATABASE of handwritten digits". In: *http://yann.lecun.com/exdb/mnist/* (). URL: https://cir.nii.ac.jp/crid/1571417126193283840 (visited on 09/06/2023).

[221] *UCI Machine Learning Repository*. URL: http://archive.ics.uci.edu/ (visited on 09/06/2023).

[222] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon. "GANomaly: Semi-supervised Anomaly Detection via Adversarial Training". en. In: *Computer Vision – ACCV 2018*. Ed. by C. V. Jawahar, H. Li, G. Mori, and K. Schindler. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019, pp. 622–637.

[223] J. Kim, H. Kang, and P. Kang. "Time-series anomaly detection with stacked Transformer representations and 1D convolutional network". In: *Engineering Applications of Artificial Intelligence* 120 (Apr. 2023), p. 105964. URL: https://www.sciencedirect.com/science/article/pii/S0952197623001483 (visited on 09/07/2023).

[224] S. Huang et al. "HitAnomaly: Hierarchical Transformers for Anomaly Detection in System Log". In: *IEEE Transactions on Network and Service Management* 17.4 (Dec. 2020). Conference Name: IEEE Transactions on Network and Service Management, pp. 2064–2076.

[225] X. Cai et al. "ITran: A novel transformer-based approach for industrial anomaly detection and localization". In: *Engineering Applications of Artificial Intelligence* 125 (Oct. 2023), p. 106677. URL: https://www.sciencedirect.com/science/article/pii/S0952197623008618 (visited on 09/07/2023).

[226] O. H. Anidjar et al. "A Stethoscope for Drones: Transformers-Based Methods for UAVs Acoustic Anomaly Detection". In: *IEEE Access* 11 (2023). Conference Name: IEEE Access, pp. 33336–33353.

[227] W. Pang, Q. He, and Y. Li. "Predicting skeleton trajectories using a Skeleton-Transformer for video anomaly detection". en. In: *Multimedia Systems* 28.4 (Aug. 2022), pp. 1481–1494. URL: https://doi.org/10.1007/s00530-022-00915-9 (visited on 09/07/2023).

[228] S. Panchavati et al. "Pretrained Transformers for Seizure Detection". In: *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. ISSN: 2379-190X. June 2023, pp. 1–2.

[229] A. Raza, K. P. Tran, L. Koehl, and S. Li. "AnoFed: Adaptive anomaly detection for digital health using transformer-based federated learning and support vector data description". In: *Engineering Applications of Artificial Intelligence* 121 (May 2023), p. 106051. URL: https://www.sciencedirect.com/science/article/pii/S095219762300235X (visited on 09/07/2023).

[230] M. Gu et al. "A lightweight convolutional neural network hardware implementation for wearable heart rate anomaly detection". In: *Computers in Biology and Medicine* 155 (Mar. 2023), p. 106623. URL: https://www.sciencedirect.com/science/article/pii/S0010482523000884 (visited on 09/07/2023).

[231] R. Chalapathy and S. Chawla. *Deep Learning for Anomaly Detection: A Survey*. arXiv:1901.03407 [cs, stat]. Jan. 2019. URL: http://arxiv.org/abs/1901.03407 (visited on 09/14/2023).

[232] G. Iglesias, E. Talavera, and A. Díaz-Álvarez. "A survey on GANs for computer vision: Recent research, analysis and taxonomy". In: *Computer Science Review* 48 (May 2023), p. 100553. URL: https://www.sciencedirect.com/science/article/pii/S1574013723000205 (visited on 09/14/2023).

[233] V. Škvára et al. "Comparison of Anomaly Detectors: Context Matters". In: *IEEE Transactions on Neural Networks and Learning Systems* 33.6 (June 2022). Conference Name: IEEE Transactions on Neural Networks and Learning Systems, pp. 2494–2507.

[234] A. Emmott et al. *A Meta-Analysis of the Anomaly Detection Problem*. arXiv:1503.01158 [cs, stat]. Aug. 2016. URL: http://arxiv.org/abs/1503.01158 (visited on 05/24/2023).

[235] X.-R. Sheng, D.-C. Zhan, S. Lu, and Y. Jiang. "Multi-View Anomaly Detection: Neighborhood in Locality Matters". en. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33.01 (July 2019). Number: 01, pp. 4894–4901. URL: https://ojs.aaai.org/index.php/AAAI/article/view/4418 (visited on 05/24/2023).

[236] G. Steinbuss and K. Böhm. "Benchmarking Unsupervised Outlier Detection with Realistic Synthetic Data". en. In: *ACM Transactions on Knowledge Discovery from Data* 15.4 (Aug. 2021), pp. 1–20. URL: https://dl.acm.org/doi/10.1145/3441453 (visited on 05/24/2023).

[237] X. Xu, H. Liu, and M. Yao. "Recent Progress of Anomaly Detection". en. In: *Complexity* 2019 (Jan. 2019), pp. 1–11. URL: https://www.hindawi.com/journals/complexity/2019/2686378/ (visited on 05/24/2023).

[238] H. Trittenbach, A. Englhardt, and K. Böhm. "An overview and a benchmark of active learning for outlier detection with one-class classifiers". en. In: *Expert Systems with Applications* 168 (Apr. 2021), p. 114372. URL: https://linkinghub.elsevier.com/retrieve/pii/S0957417420310496 (visited on 05/24/2023).

[239] V. Vincent, M. Wannes, and D. Jesse. "Transfer Learning for Anomaly Detection through Localized and Unsupervised Instance Selection". en. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.04 (Apr. 2020). Number: 04, pp. 6054–6061. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6068 (visited on 05/24/2023).

[240] E. Marchi, F. Vesperini, S. Squartini, and B. Schuller. "Deep Recurrent Neural Network-Based Autoencoders for Acoustic Novelty Detection". en. In: *Computational Intelligence and Neuroscience* 2017 (2017), pp. 1–14. URL: https://www.hindawi.com/journals/cin/2017/4694860/ (visited on 07/21/2021).

[241] G. O. Campos et al. "On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study". en. In: *Data Mining and Knowledge Discovery* 30.4 (July 2016), pp. 891–927. URL: https://doi.org/10.1007/s10618-015-0444-8 (visited on 05/24/2023).

[242] I. Golan and R. El-Yaniv. "Deep Anomaly Detection Using Geometric Transformations". In: *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc., 2018. URL: https://papers.nips.cc/paper_files/paper/2018/hash/5e62d03aec0d17facfc5355dd90d441c-Abstract.html (visited on 05/26/2023).

[243] T. Saito and M. Rehmsmeier. "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets". In: *PloS one* 10.3 (2015), e0118432.

[244] V. Rothoft, J. Si, F. Jiang, and R. Shen. "Monitor Pupils' Attention by Image Super-Resolution and Anomaly Detection". In: *2017 International Conference on Computer Systems, Electronics and Control (ICCSEC)*. Dec. 2017, pp. 843–847.

[245] G. Baliniskite, E. Lavendelis, and M. Pudane. "Affective state based anomaly detection in crowd". In: *Applied Computer Systems* 24.2 (2019), pp. 134–140.

[246] N. Ding et al. "Driver's emotional state-based data anomaly detection for vehicular ad hoc networks". In: *2019 IEEE International Conference on Smart Internet of Things (SmartIoT)*. IEEE. 2019, pp. 121–126.

[247] C. Fayet, A. Delhay, D. Lolive, and P.-F. Marteau. "EMO&LY (EMOtion and AnomaLY) : A New Corpus for Anomaly Detection in an Audiovisual Stream with Emotional Context". en. In: (), p. 7.

[248] Y. Hu, Y. Zhang, Y. Wang, and D. Work. "Detecting Socially Abnormal Highway Driving Behaviors via Recurrent Graph Attention Networks". en. In: *Proceedings of the ACM Web Conference 2023*. Austin TX USA: ACM, Apr. 2023, pp. 3086–3097. URL: https://dl.acm.org/doi/10.1145/3543507.3583452 (visited on 09/07/2023).

[249] A. Pillai, S. Nepal, and A. Campbell. "Rare Life Event Detection via Mobile Sensing Using Multi-Task Learning". en. In: *Proceedings of the Conference on Health, Inference, and Learning*. ISSN: 2640-3498. PMLR, June 2023, pp. 279–293. URL: https://proceedings.mlr.press/v209/pillai23a.html (visited on 09/08/2023).

[250] J. Zhu et al. "UAED: Unsupervised Abnormal Emotion Detection Network Based on Wearable Mobile Device". In: *IEEE Transactions on Network Science and Engineering* (2023). Conference Name: IEEE Transactions on Network Science and Engineering, pp. 1–14.

[251] Y. Ye and P. Li. "Anomaly Detection For Autonomous Driving Public Transports". In: *2023 IEEE 6th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*. Vol. 6. IEEE. 2023, pp. 1430–1434.

[252] G. Debard et al. "Camera-based fall detection using real-world versus simulated data: How far are we from the solution?" In: *Journal of Ambient Intelligence and Smart Environments* 8.2 (2016), pp. 149–168.

[253] B. Osiński et al. "Simulation-based reinforcement learning for real-world autonomous driving". In: *2020 IEEE international conference on robotics and automation (ICRA)*. IEEE. 2020, pp. 6411–6418.

[254] *Road traffic injuries*. en. URL: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries (visited on 05/02/2022).

[255] filby. *SP Grand Challenge e-Prevention*. en-US. URL: https://robotics.ntua.gr/eprevention-sp-challenge/ (visited on 08/09/2023).

[256] H. Almahasneh et al. "EEG based driver cognitive distraction assessment". In: *2014 5th International Conference on Intelligent and Advanced Systems (ICIAS)* (2014).

[257] A. Ragab, C. Craye, M. Kamel, and F. Karray. "A visual–based driver distraction recognition and detection using random forest". English. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 8814 (2014). ISBN: 9783319117577, pp. 256–265.

[258] M. Wollmer et al. "Online Driver Distraction Detection Using Long Short-Term Memory". In: *IEEE Transactions on Intelligent Transportation Systems* 12.2 (June 2011). Conference Name: IEEE Transactions on Intelligent Transportation Systems, pp. 574–582.

[259] B. Baheti, S. Gajre, and S. Talbar. "Detection of Distracted Driver Using Convolutional Neural Network". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2018).

[260] M. Leekha et al. "Are You Paying Attention? Detecting Distracted Driving in Real-Time". In: *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)* (2019).

[261] J. Chen et al. "Fine-Grained Detection of Driver Distraction Based on Neural Architecture Search". en. In: *IEEE Transactions on Intelligent Transportation Systems* 22.9 (Sept. 2021), pp. 5783–5801. URL: https://ieeexplore.ieee.org/document/9352235/ (visited on 05/05/2022).

[262] T. J. Chengula, J. Mwakalonge, G. Comert, and S. Siuhi. "Improving Road Safety with Ensemble Learning: Detecting Driver Anomalies Using Vehicle Inbuilt Cameras". In: *Machine Learning with Applications* (2023), p. 100510.

[263] V. Sadhu, T. Misu, and D. Pompili. "Deep multi-task learning for anomalous driving detection using CAN bus scalar sensor data". In: *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2019, pp. 2038–2043.

[264] H. Oikawa et al. "Fast semi-supervised anomaly detection of drivers' behavior using online sequential extreme learning machine". In: *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*. IEEE. 2020, pp. 1–8.

[265] H. Yang et al. "Quantitative Identification of Driver Distraction: A Weakly Supervised Contrastive Learning Approach". In: *IEEE Transactions on Intelligent Transportation Systems* (2023).

[266] O. Kopuklu, J. Zheng, H. Xu, and G. Rigoll. "Driver anomaly detection: A dataset and contrastive learning approach". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2021, pp. 91–100.

[267] Y. Qiu, T. Misu, and C. Busso. "Use of triplet-loss function to improve driving anomaly detection using conditional generative adversarial network". In: *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*. IEEE. 2020, pp. 1–7.

[268] A. Dairi, F. Harrou, and Y. Sun. "Efficient Driver Drunk Detection by Sensors: A Manifold Learning-Based Anomaly Detector". In: *IEEE Access* 10 (2022), pp. 119001–119012.

[269] Y. Qiu, T. Misu, and C. Busso. "Unsupervised scalable multimodal driving anomaly detection". In: *IEEE Transactions on Intelligent Vehicles* (2022).

[270] Z. Pan et al. "Detecting Manic State of Bipolar Disorder Based on Support Vector Machine and Gaussian Mixture Model Using Spontaneous Speech". English. In: *Psychiatry Investigation* 15.7 (July 2018). Publisher: Korean Neuropsychiatric Association, pp. 695–700. URL: http://www.psychiatryinvestigation.org/journal/view.php?doi=10.30773/pi.2017.12.15 (visited on 04/14/2023).

[271] X. Zhou, K. Jin, Y. Shang, and G. Guo. "Visually Interpretable Representation Learning for Depression Recognition from Facial Images". In: *IEEE Transactions on Affective Computing* 11.3 (July 2020). Conference Name: IEEE Transactions on Affective Computing, pp. 542–552.

[272] B. Lamichhane, J. Zhou, and A. Sano. *Psychotic Relapse Prediction in Schizophrenia Patients using A Mobile Sensing-based Supervised Deep Learning Model*. arXiv:2205.12225 [cs, eess]. May 2022. URL: http://arxiv.org/abs/2205.12225 (visited on 04/14/2023).

[273] A. Othmani, A.-O. Zeghina, and M. Muzammel. "A Model of Normality Inspired Deep Learning Framework for Depression Relapse Prediction Using Audiovisual Data". en. In: *Computer Methods and Programs in Biomedicine* (Sept. 2022), p. 107132. URL: https://linkinghub.elsevier.com/retrieve/pii/S0169260722005132 (visited on 09/26/2022).

[274] D. A. Adler et al. "Predicting Early Warning Signs of Psychotic Relapse From Passive Sensing Data: An Approach Using Encoder-Decoder Neural Networks". In: *JMIR mHealth and uHealth* 8.8 (Aug. 2020), e19962. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7490673/ (visited on 02/14/2023).

[275] K. Wang et al. "Research on Healthy Anomaly Detection Model Based on Deep Learning from Multiple Time-Series Physiological Signals". en. In: *Scientific Programming* 2016 (Sept. 2016). Publisher: Hindawi, e5642856. URL: https://www.hindawi.com/journals/sp/2016/5642856/ (visited on 02/14/2023).

[276] A. Cohen et al. "Relapse prediction in schizophrenia with smartphone digital phenotyping during COVID-19: a prospective, three-site, two-country, longitudinal study". English. In: *Schizophrenia* 9.1 (2023).

[277] C. Garoufis et al. "An Unsupervised Learning Approach for Detecting Relapses from Spontaneous Speech in Patients with Psychosis". In: *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*. ISSN: 2641-3604. July 2021, pp. 1–5.

[278] C. R. Bingham, J. T. Shope, and J. Zhu. "Substance-involved driving: Predicting driving after using alcohol, marijuana, and other drugs". In: *Traffic injury prevention* 9.6 (2008), pp. 515–526.

[279] A.-C. Phan, T.-N. Trieu, and T.-C. Phan. "Driver drowsiness detection and smart alerting using deep learning and IoT". In: *Internet of Things* 22 (2023), p. 100705.

[280] J. Wang, R. R. Knipling, and M. J. Goodman. "The role of driver inattention in crashes: new statistics from the 1995 Crashworthiness Data System". en. In: *Annual proceedings of the Association for the Advancement of Automotive Medicine* 40 (1996), pp. 377–392. URL: https://www.safetylit.org/citations/index.php?fuseaction=citations.viewdetails&citationIds[]=citjournalarticle_84079_17 (visited on 05/03/2022).

[281] T. A. Ranney et al. "NHTSA driver distraction research: past, present and future". In: *17th International Technical Conference on the Enhanced Safety of Vehicles*. 2001.

[282] H. B. Sundfør, F. Sagberg, and A. Høye. "Inattention and distraction in fatal road crashes – Results from in-depth crash investigations in Norway". en. In: *Accident Analysis & Prevention* 125 (Apr. 2019), pp. 152–157. URL: https://www.sciencedirect.com/science/article/pii/S0001457519301988 (visited on 05/03/2022).

[283] J. Laberge, C. Scialfa, C. White, and J. Caird. “Effects of passenger and cellular phone conversations on driver distraction”. English. In: *Transportation Research Record* 1899 (2004), pp. 109–116.

[284] J. Lee, B. Caven, S. Haake, and T. Brown. “Speech-based interaction with in-vehicle computers: The effect of speech-based e-mail on drivers’ attention to the roadway”. English. In: *Human Factors* 43.4 (2001), pp. 631–640.

[285] J. Harbluk and S. Lalande. *Performing E-mail Tasks While Driving: The Impact of Speech-Based Tasks on Visual Detection*. Pages: 317. June 2005.

[286] J. Li et al. “Distracted driving caused by voice message apps: A series of experimental studies”. en. In: *Transportation Research Part F: Traffic Psychology and Behaviour* 76 (Jan. 2021), pp. 1–13. URL: https://www.sciencedirect.com/science/article/pii/S1369847820305568 (visited on 05/05/2022).

[287] *Hadrian Project | Holistic Approach for Driver Role Integration and Automation Allocation for European Mobility Needs*. en-US. URL: https://hadrianproject.eu/ (visited on 12/14/2022).

[288] J. D. Ortega et al. “Dmd: A large-scale multi-modal driver monitoring dataset for attention and alertness analysis”. In: *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*. Springer. 2020, pp. 387–405.

[289] W. Othman, A. Kashevnik, A. Ali, and N. Shilov. “DriverMVT: In-cabin dataset for driver monitoring including video and vehicle telemetry information”. In: *Data* 7.5 (2022), p. 62.

[290] S. Jha et al. “The multimodal driver monitoring database: A naturalistic corpus to study driver attention”. In: *IEEE Transactions on Intelligent Transportation Systems* 23.8 (2021), pp. 10736–10752.

[291] Y. Abouelnaga, H. M. Eraqi, and M. N. Moustafa. “Real-time distracted driver posture classification”. In: *arXiv preprint arXiv:1706.09498* (2017).

[292] S. Abtahi, M. Omidyeganeh, S. Shirmohammadi, and B. Hariri. “YawDD: A yawning detection dataset”. In: *Proceedings of the 5th ACM multimedia systems conference*. 2014, pp. 24–28.

[293] *WMA - The World Medical Association-Declaration of Helsinki 2008*. en-US. URL: https://www.wma.net/what-we-do/medical-ethics/declaration-of-helsinki/doh-oct2008/ (visited on 05/05/2022).

[294] M. Vengust, B. Kaluža, K. Stojmenova, and J. Sodnik. *NERVteh Compact Motion Based Driving Simulator*. Pages: 243. Sept. 2017.

[295] *SE PRO*. en-US. URL: https://smarteye.se/research-instruments/se-pro/ (visited on 06/10/2022).

[296] *Professional Kit*. en. URL: https://www.pluxbiosignals.com/en-fr/products/professional-kit (visited on 06/09/2022).

[297] *Depth Camera D435*. en-US. URL: https://www.intelrealsense.com/depth-camera-d435/ (visited on 06/09/2022).

[298] *Support for FLIR A325sc | Teledyne FLIR*. URL: https://www.flir.fr/support/products/a325sc/#Overview (visited on 06/09/2022).

[299] *SCANeR studio*. fr-FR. URL: https://www.avsimulation.com/scaner-studio/?lang=fr (visited on 06/09/2022).

[300] T. Liu et al. “Driver Distraction Detection Using Semi-Supervised Machine Learning”. In: *IEEE Transactions on Intelligent Transportation Systems* 17.4 (Apr. 2016). Conference Name: IEEE Transactions on Intelligent Transportation Systems, pp. 1108–1120.

[301] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12.85 (2011), pp. 2825–2830. URL: http://jmlr.org/papers/v12/pedregosa11a.html (visited on 06/09/2022).

[302] *TensorFlow*. en. URL: https://www.tensorflow.org/ (visited on 06/09/2022).

[303] F. Španiel et al. "ITAREPS: Information Technology Aided Relapse Prevention Programme in Schizophrenia". en. In: *Schizophrenia Research* 98.1 (Jan. 2008), pp. 312–317. URL: https://www.sciencedirect.com/science/article/pii/S0920996407003994 (visited on 04/04/2023).

[304] M. Pobiruchin, J. Suleder, R. Zowalla, and M. Wiesner. "Accuracy and Adoption of Wearable Technology Used by Active Citizens: A Marathon Event Field Study". EN. In: *JMIR mHealth and uHealth* 5.2 (Feb. 2017). Company: JMIR mHealth and uHealth Distributor: JMIR mHealth and uHealth Institution: JMIR mHealth and uHealth Label: JMIR mHealth and uHealth Publisher: JMIR Publications Inc., Toronto, Canada, e6395. URL: https://mhealth.jmir.org/2017/2/e24 (visited on 04/03/2023).

[305] M. Dehghani, K. J. Kim, and R. M. Dangelico. "Will smartwatches last? factors contributing to intention to keep using smart wearable technology". en. In: *Telematics and Informatics* 35.2 (May 2018), pp. 480–490. URL: https://www.sciencedirect.com/science/article/pii/S0736585317307141 (visited on 04/03/2023).

[306] S. Hamieh, V. Heiries, H. Al Osman, and C. Godin. "Relapse Detection in Patients with Psychotic Disorders Using Unsupervised Learning on Smartwatch Signals". In: *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. June 2023, pp. 1–2.

[307] M. L. Birnbaum et al. "Detecting relapse in youth with psychotic disorders utilizing patient-generated and patient-contributed digital data from facebook". In: *NPJ schizophrenia* 5.1 (2019), p. 17.

[308] A. Sato et al. "Psychotic relapse in people with schizophrenia within 12 months of discharge from acute inpatient care: protocol for development and validation of a prediction model based on a retrospective cohort study in three psychiatric hospitals in Japan". In: *Diagnostic and Prognostic Research* 6.1 (2022), pp. 1–9.

[309] D. Adler et al. *Predicting early warning signs of psychotic relapse from passive sensing data: an approach using encoder-decoder neural networks. JMIR Mhealth Uhealth 8 (8), e19962 (2020).*

[310] D. Ben-Zeev et al. "CrossCheck: Integrating self-report, behavioral sensing, and smartphone use to identify digital indicators of psychotic relapse." In: *Psychiatric rehabilitation journal* 40.3 (2017), p. 266.

[311] I. M. Anderson, P. M. Haddad, and J. Scott. "Bipolar disorder". eng. In: *BMJ (Clinical research ed.)* 345 (Dec. 2012), e8508.

[312] A. P. American Psychiatric Association and A. P. Association. *Diagnostic and statistical manual of mental disorders: DSM-IV*. Vol. 4. American psychiatric association Washington, DC, 1994. URL: https://www.gammaconstruction.mu/sites/default/files/webform/cvs/pdf-diagnostic-and-statistical-manual-of-mental-disorders-dsm-iv-american-psychiatric-association-pdf-download-free-book-9223cc7.pdf (visited on 10/25/2023).

[313] M. J. Owen, A. Sawa, and P. B. Mortensen. "Schizophrenia". In: *Lancet (London, England)* 388.10039 (July 2016), pp. 86–97. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4940219/ (visited on 10/25/2023).

[314] D. B. Arciniegas. "Psychosis". en-US. In: *CONTINUUM: Lifelong Learning in Neurology* 21.3 (June 2015), p. 715. URL: https://journals.lww.com/continuum/abstract/2015/06000/psychosis.15.aspx (visited on 10/25/2023).

[315] D. Malaspina et al. "Schizoaffective Disorder in the DSM-5". In: *Schizophrenia Research*. DSM-5 150.1 (Oct. 2013), pp. 21–25. URL: https://www.sciencedirect.com/science/article/pii/S0920996413002260 (visited on 10/25/2023).

[316] "Schizophreniform Disorder: Practice Essentials, Background, Pathophysiology". In: (June 2023). Publication: Medscape - eMedicine. URL: https://emedicine.medscape.com/article/2008351-overview?form=fpf (visited on 10/25/2023).

[317] S. Sigurðardóttir, A. Islind, and M. Óskarsdóttir. "Collecting Data from a Mobile App and a Smartwatch Supports Treatment of Schizophrenia and Bipolar Disorder". In: *Studies in health technology and informatics*. Vol. 294. Journal Abbreviation: Studies in health technology and informatics. May 2022.

[318] A. Maxhuni et al. "Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients". In: *Pervasive and Mobile Computing* 31 (2016), pp. 50–66.

[319] M. Faurholt-Jepsen et al. "Differences in psychomotor activity in patients suffering from unipolar and bipolar affective disorder in the remitted or mild/moderate depressive state". en. In: *Journal of Affective Disorders* 141.2 (Dec. 2012), pp. 457–463. URL: https://www.sciencedirect.com/science/article/pii/S0165032712001164 (visited on 04/14/2023).

[320] J. J. Chapman, J. A. Roberts, V. T. Nguyen, and M. Breakspear. "Quantification of free-living activity patterns using accelerometry in adults with mental illness". eng. In: *Scientific reports* 7 (Mar. 2017), p. 43174. URL: https://europepmc.org/articles/PMC5339808 (visited on 04/14/2023).

[321] S. Spulber et al. "Patterns of activity correlate with symptom severity in major depressive disorder patients". en. In: *Translational Psychiatry* 12.1 (June 2022). Number: 1 Publisher: Nature Publishing Group, pp. 1–8. URL: https://www.nature.com/articles/s41398-022-01989-9 (visited on 04/14/2023).

[322] M. Martinato et al. "Usability and Accuracy of a Smartwatch for the Assessment of Physical Activity in the Elderly Population: Observational Study". EN. In: *JMIR mHealth and uHealth* 9.5 (May 2021). Company: JMIR mHealth and uHealth Distributor: JMIR mHealth and uHealth Institution: JMIR mHealth and uHealth Label: JMIR mHealth and uHealth Publisher: JMIR Publications Inc., Toronto, Canada, e20966. URL: https://mhealth.jmir.org/2021/5/e20966 (visited on 04/03/2023).

[323] K. L. Benson. "Sleep in Schizophrenia: Impairments, Correlates, and Treatment". English. In: *Psychiatric Clinics* 29.4 (Dec. 2006). Publisher: Elsevier, pp. 1033–1045. URL: https://www.psych.theclinics.com/article/S0193-953X(06)00073-6/fulltext (visited on 02/14/2023).

[324] I. Barnett et al. "Relapse prediction in schizophrenia through digital phenotyping: a pilot study". en. In: *Neuropsychopharmacology* 43.8 (July 2018). Number: 8 Publisher: Nature Publishing Group, pp. 1660–1666. URL: https://www.nature.com/articles/s41386-018-0030-z (visited on 04/14/2023).

[325] S. Lambrichts et al. "Which residual symptoms predict relapse after successful electroconvulsive therapy for late-life depression?" English. In: *Journal of Psychiatric Research* 154 (2022), pp. 111–116.

[326] R. Hartmann, F. M. Schmidt, C. Sander, and U. Hegerl. "Heart Rate Variability as Indicator of Clinical State in Depression". In: *Frontiers in Psychiatry* 9 (2019). URL: https://www.frontiersin.org/articles/10.3389/fpsyt.2018.00735 (visited on 04/14/2023).

[327] B. L. Henry et al. "Heart rate variability in bipolar mania and schizophrenia". en. In: *Journal of Psychiatric Research* 44.3 (Feb. 2010), pp. 168–176. URL: https://www.sciencedirect.com/science/article/pii/S0022395609001745 (visited on 04/14/2023).

[328] Y. Esaki et al. "Association between circadian activity rhythms and mood episode relapse in bipolar disorder: a 12-month prospective cohort study". In: *Translational Psychiatry* 11 (Oct. 2021), p. 525. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8514471/ (visited on 02/14/2023).

[329] I. Guyon and A. Elisseeff. "An introduction to variable and feature selection". In: *Journal of machine learning research* 3.Mar (2003), pp. 1157–1182. URL: https://www.jmlr.org/papers/volume3/guyon03a/guyon03a.pdf?ref=driverlayer.com/web (visited on 10/10/2023).

[330] L. Yin et al. "Feature selection for high-dimensional imbalanced data". en. In: *Neurocomputing* 105 (Apr. 2013), pp. 3–11. URL: https://linkinghub.elsevier.com/retrieve/pii/S0925231212007126 (visited on 08/08/2023).

[331] N. Anwar, G. Jones, and S. Ganesh. "Measurement of data complexity for classification problems with unbalanced data". en. In: *Statistical Analysis and Data Mining: The ASA Data Science Journal* 7.3 (2014). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sam.11228, pp. 194–211. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sam.11228 (visited on 10/10/2023).

[332] L. Annemans, K. Redekop, and K. Payne. "Current methodological issues in the economic assessment of personalized medicine". In: *Value in Health* 16.6 (2013). Publisher: Elsevier, S20–S26. URL: https://www.sciencedirect.com/science/article/pii/S1098301513018640 (visited on 10/10/2023).

[333] F. Thabtah, S. Hammoud, F. Kamalov, and A. Gonsalves. "Data imbalance in classification: Experimental evaluation". In: *Information Sciences* 513 (2020). Publisher: Elsevier, pp. 429–441. URL: https://www.sciencedirect.com/science/article/pii/S0020025519310497 (visited on 10/10/2023).

[334] J. Lear. *Aristotle: the desire to understand*. Cambridge University Press, 1988.

[335] C. Molnar. *Interpretable machine learning*. Lulu. com, 2020.

[336] N. Sánchez-Maroño, A. Alonso-Betanzos, and M. Tombilla-Sanromán. "Filter Methods for Feature Selection – A Comparative Study". en. In: *Intelligent Data Engineering and Automated Learning - IDEAL 2007*. Ed. by H. Yin et al. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2007, pp. 178–187.

[337] J. Tang, S. Alelyani, and H. Liu. "Feature selection for classification: A review". In: *Data classification: Algorithms and applications* (2014), p. 37.

[338] S. Gao, G. Ver Steeg, and A. Galstyan. "Variational information maximization for feature selection". In: *Advances in neural information processing systems* 29 (2016).

[339] I. Guyon and A. Elisseeff. "An introduction to variable and feature selection". In: *Journal of machine learning research* 3.Mar (2003), pp. 1157–1182.

[340] H. Liu and H. Motoda. *Computational methods of feature selection*. CRC press, 2007.

[341] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. "Gene Selection for Cancer Classification using Support Vector Machines". en. In: *Machine Learning* 46.1 (Jan. 2002), pp. 389–422. URL: https://doi.org/10.1023/A:1012487302797 (visited on 08/08/2023).

[342] T. Rückstieß, C. Osendorfer, and P. van der Smagt. "Sequential Feature Selection for Classification". en. In: *AI 2011: Advances in Artificial Intelligence*. Ed. by D. Wang and M. Reynolds. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2011, pp. 132–141.

[343] T. N. Lal, O. Chapelle, J. Weston, and A. Elisseeff. "Embedded Methods". en. In: *Feature Extraction: Foundations and Applications*. Ed. by I. Guyon, M. Nikravesh, S. Gunn, and L. A. Zadeh. Studies in Fuzziness and Soft Computing. Berlin, Heidelberg: Springer, 2006, pp. 137–165. URL: https://doi.org/10.1007/978-3-540-35488-8_6 (visited on 08/08/2023).

[344] F. Nie, H. Huang, X. Cai, and C. Ding. "Efficient and Robust Feature Selection via Joint \mathscrl2,1-Norms Minimization". In: *Advances in Neural Information Processing Systems*. Vol. 23. Curran Associates, Inc., 2010. URL: https://proceedings.neurips.cc/paper_files/paper/2010/hash/09c6c3783b4a70054da74f2538e Abstract.html (visited on 08/08/2023).

[345] B. H. Menze et al. "A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data". en. In: *BMC Bioinformatics* 10.1 (July 2009), p. 213. URL: https://doi.org/10.1186/1471-2105-10-213 (visited on 08/08/2023).

[346] P. Dhal and C. Azad. "A comprehensive survey on feature selection in the various fields of machine learning". In: *Applied Intelligence* (2022), pp. 1–39.

[347] J. Li et al. "Feature selection: A data perspective". In: *ACM computing surveys (CSUR)* 50.6 (2017), pp. 1–45.

[348] F. Kamalov, F. Thabtah, and H. H. Leung. "Feature selection in imbalanced data". In: *Annals of Data Science* 10.6 (2023), pp. 1527–1541.

[349] X.-w. Chen and M. Wasikowski. "FAST: a roc-based feature selection metric for small samples and imbalanced data classification problems". en. In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. Las Vegas Nevada USA: ACM, Aug. 2008, pp. 124–132. URL: https://dl.acm.org/doi/10.1145/1401890.1401910 (visited on 10/11/2023).

[350] A. Ali, S. M. Shamsuddin, and A. L. Ralescu. "Classification with class imbalance problem". In: *Int. J. Advance Soft Compu. Appl* 5.3 (2013), pp. 176–204. URL: https://www.researchgate.net/profile/Aida-Ali-4/publication/288228469_Classification_with_class_imbalance_problem_A_review/links/57b556d008ae19a365faff16/Classification-with-class-imbalance-problem-A-review.pdf (visited on 10/10/2023).

[351] Y. Liu et al. "A classification method based on feature selection for imbalanced data". In: *IEEE Access* 7 (2019). Publisher: IEEE, pp. 81794–81807. URL: https://ieeexplore.ieee.org/abstract/document/8740942/ (visited on 10/11/2023).

[352] S. Maldonado, R. Weber, and F. Famili. "Feature selection for high-dimensional class-imbalanced data sets using support vector machines". In: *Information sciences* 286 (2014). Publisher: Elsevier, pp. 228–246. URL: https://www.sciencedirect.com/science/article/pii/S0020025514007154?casa_token=DV19sK27H-cAAAAA:WyNU1P-wwCQq3gecZ3pOPc1bgB_nd4S7tHVFXVrM4wP2HNP4XerbwyG89J1f (visited on 10/11/2023).

[353] M. C. Massi, F. Gasperoni, F. Ieva, and A. M. Paganoni. "Feature selection for imbalanced data with deep sparse autoencoders ensemble". en. In: *Statistical Analysis and Data Mining: The ASA Data Science Journal* 15.3 (June 2022), pp. 376–395. URL: https://onlinelibrary.wiley.com/doi/10.1002/sam.11567 (visited on 10/10/2023).

[354] D. H. Hockenbury and S. E. Hockenbury. *Discovering psychology, 4th ed.* Discovering psychology, 4th ed. Pages: xlii, 587. New York, NY, US: Worth Publishers, 2007.

[355] L.-W. Chen and A. Rudnicky. "Exploring Wav2vec 2.0 Fine Tuning for Improved Speech Emotion Recognition". In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2023, pp. 1–5.

[356] S. Zhou et al. "Emotion Recognition from Large-Scale Video Clips with Cross-Attention and Hybrid Feature Weighting Neural Networks". In: *International Journal of Environmental Research and Public Health* 20.2 (2023), p. 1400.

[357] S. Liu et al. "EEG emotion recognition based on the attention mechanism and pre-trained convolution capsule network". In: *Knowledge-Based Systems* 265 (2023), p. 110372.

[358] B. Mocanu, R. Tapu, and T. Zaharia. "Multimodal emotion recognition using cross modal audio-video fusion with attention and deep metric learning". In: *Image and Vision Computing* 133 (2023), p. 104676.

[359] J. Wen et al. "Dynamic interactive multiview memory network for emotion recognition in conversation". In: *Information Fusion* 91 (2023), pp. 123–133.

[360] Z. Zhang, J. Han, E. Coutinho, and B. Schuller. "Dynamic Difficulty Awareness Training for Continuous Emotion Prediction". In: *IEEE Transactions on Multimedia* 21.5 (May 2019). Conference Name: IEEE Transactions on Multimedia, pp. 1289–1301.

[361] *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. URL: https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html (visited on 08/07/2021).

[362] L. Stappen et al. "The MuSe 2021 Multimodal Sentiment Analysis Challenge: Sentiment, Emotion, Physiological-Emotion, and Stress". en. In: (2021), p. 10.

[363] L. I.-K. Lin. "A Concordance Correlation Coefficient to Evaluate Reproducibility". In: *Biometrics* 45.1 (1989). Publisher: [Wiley, International Biometric Society], pp. 255–268. URL: https://www.jstor.org/stable/2532051 (visited on 07/17/2021).

[364] F. Ringeval et al. "AV+EC 2015: The First Affect Recognition Challenge Bridging Across Audio, Video, and Physiological Data". en. In: *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. Brisbane Australia: ACM, Oct. 2015, pp. 3–8. URL: https://dl.acm.org/doi/10.1145/2808196.2811642 (visited on 07/17/2021).

[365] J. Kossaifi et al. "Sewa db: A rich database for audio-visual emotion and sentiment research in the wild". In: *IEEE transactions on pattern analysis and machine intelligence* 43.3 (2019), pp. 1022–1040.

[366] S. Li, W. Deng, and J. Du. "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2852–2861.

[367] F. Burkhardt et al. "A database of German emotional speech." In: *Interspeech*. Vol. 5. 2005, pp. 1517–1520.

[368] S. R. Livingstone and F. A. Russo. "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English". In: *PloS one* 13.5 (2018), e0196391.

[369] C. Busso et al. "IEMOCAP: interactive emotional dyadic motion capture database". en. In: *Language Resources and Evaluation* 42.4 (Nov. 2008). Number: 4, p. 335. URL: https://doi.org/10.1007/s10579-008-9076-6 (visited on 04/09/2021).

[370] B. Schuller, A. Batliner, S. Steidl, and D. Seppi. "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge". In: *Speech communication* 53.9-10 (2011), pp. 1062–1087.

[371] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne. "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions". In: *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. Apr. 2013, pp. 1–8.

[372] C. Kirschbaum, K.-M. Pirke, and D. H. Hellhammer. "The 'Trier Social Stress Test' – A Tool for Investigating Psychobiological Stress Responses in a Laboratory Setting". In: *Neuropsychobiology* 28.1-2 (1993). Publisher: Karger Publishers, pp. 76–81. URL: https://www.karger.com/Article/FullText/119004 (visited on 07/20/2021).

[373] A. Pentari, G. Kafentzis, and M. Tsiknakis. "Graph-based representations of speech signals: A novel approach for emotion recognition". In: *Available at SSRN 4402871* ().

[374] A. Greco, N. Strisciuglio, M. Vento, and V. Vigilante. "Benchmarking deep networks for facial emotion recognition in the wild". In: *Multimedia tools and applications* 82.8 (2023), pp. 11189–11220.

[375] X. Qin et al. "BERT-ERC: Fine-tuning BERT is enough for emotion recognition in conversation". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 11. 2023, pp. 13492–13500.

[376] F. Eyben, M. Wöllmer, and B. Schuller. "Opensmile: the munich versatile and fast open-source audio feature extractor". en. In: *Proceedings of the international conference on Multimedia - MM '10*. Firenze, Italy: ACM Press, 2010, p. 1459. URL: http://dl.acm.org/citation.cfm?doid=1873951.1874246 (visited on 07/17/2021).

[377] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. "Joint face detection and alignment using multitask cascaded convolutional networks". In: *IEEE signal processing letters* 23.10 (2016), pp. 1499–1503.

[378] O. M. Parkhi, A. Vedaldi, and A. Zisserman. "Deep face recognition". en. In: (2015). Publisher: British Machine Vision Association. URL: https://ora.ox.ac.uk/objects/uuid:a5f2e93f-2768-45bb-8508-74747f85cad1 (visited on 07/21/2021).

[379] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *arXiv:1810.04805 [cs]* (May 2019). arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805 (visited on 07/21/2021).

[380] B. W. Schuller et al. *The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 Cough, COVID-19 Speech, Escalation & Primates.* arXiv:2102.13468 [cs, eess]. Feb. 2021. URL: http://arxiv.org/abs/2102.13468 (visited on 11/06/2023).

[381] B. W. Schuller et al. "The interspeech 2020 computational paralinguistics challenge: Elderly emotion, breathing & masks". In: (2020). URL: https://opus.bibliothek.uni-augsburg.de/opus4/files/90751/schuller20_interspeech.pdf (visited on 11/06/2023).

[382] H. Sun et al. "Multi-Modal Adaptive Fusion Transformer Network for the Estimation of Depression Level". en. In: *Sensors* 21.14 (July 2021), p. 4764. URL: https://www.mdpi.com/1424-8220/21/14/4764 (visited on 05/02/2022).

[383] C. Cai et al. "Multimodal sentiment analysis based on recurrent neural network and multimodal attention". In: *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*. 2021, pp. 61–67.

[384] A.-Q. Duong et al. "Multi-modal Stress Recognition Using Temporal Convolution and Recurrent Network with Positional Embedding". In: *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*. 2021, pp. 37–42.

[385] Z. Ma, F. Ma, B. Sun, and S. Li. "Hybrid mutimodal fusion for dimensional emotion recognition". In: *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*. 2021, pp. 29–36.

[386] S. Amiriparian et al. "Muse 2022 challenge: Multimodal humour, emotional reactions, and stress". In: *Proceedings of the 30th ACM International Conference on Multimedia*. 2022, pp. 7389–7391.

[387] S. Yadav et al. "Comparing biosignal and acoustic feature representation for continuous emotion recognition". In: *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*. 2022, pp. 37–45.

[388] H.-m. Park et al. "Towards Multimodal Prediction of Time-continuous Emotion using Pose Feature Engineering and a Transformer Encoder". In: *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*. 2022, pp. 47–54.

[389] Y. Liu, W. Sun, X. Zhang, and Y. Qin. "Improving Dimensional Emotion Recognition via Feature-wise Fusion". In: *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*. 2022, pp. 55–60.

[390] J. Li et al. "Hybrid multimodal feature extraction, mining and fusion for sentiment analysis". In: *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*. 2022, pp. 81–88.

[391] C. Li, L. Xie, and H. Pan. "Branch-fusion-net for multi-modal continuous dimensional emotion recognition". In: *IEEE Signal Processing Letters* 29 (2022), pp. 942–946.

[392] Y. He et al. "Multimodal Temporal Attention in Sentiment Analysis". In: *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*. 2022, pp. 61–66.

[393] S. Hamieh, V. Heiries, H. Al Osman, and C. Godin. "Multi-modal Fusion for Continuous Emotion Recognition by Using Auto-Encoders". In: *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*. MuSe '21. New York, NY, USA: Association for Computing Machinery, Oct. 2021, pp. 21–27. URL: https://doi.org/10.1145/3475957.3484455 (visited on 11/17/2021).