Classification et modélisation de la croissance du champignon Botrytis cinerea à partir d'imagerie microscopique : vers l'établissement de liens entre phénotypes et molécules antifongiques

Sarah Laroui

UNIVERSITÉ CÔTE D'AZUR

ÉCOLE DOCTORALE
SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION

THÈSE DE DOCTORAT

Classification et modélisation de la croissance du champignon *Botrytis cinerea* à partir d'imagerie microscopique : vers l'établissement de liens entre phénotypes et molécules antifongiques

Sarah LAROUI

Équipe MORPHEME – INRIA Sophia Antipolis/I3S/iBV

Présentée en vue de l'obtention du grade de docteur en Automatique, traitement du signal et des images
d'Université Côte d'Azur
Dirigée par :
Eric Debreuve, Xavier Descombes
Co-encadrée par : Aurélia Vernay
Soutenue le : 23 Novembre 2021

Devant le jury, composé de :
Sylvie Chambon, MCF, ENSEEIHT, IRIT
Patrick Clarysse, DR, CNRS, CREATIS
Eric Debreuve, CR, CNRS, I3S
Xavier Descombes, DR, INRIA
Daniel Racoceanu, Pr, Sorbonne Université, Institut du Cerveau
Florence Tupin, Pr, Télécom Paris, LTCI
Aurélia Vernay, PhD, BAYER

BAYER
Science For A Better Life

ANRT
ASSOCIATION NATIONALE RECHERCHE TECHNOLOGIE

Classification et modélisation de la croissance du champignon *Botrytis cinerea* à partir d'imagerie microscopique : vers l'établissement de liens entre phénotypes et molécules antifongiques.

Jury :

Rapporteurs
Patrick Clarysse, Directeur de Recherche, CREATIS CNRS, Inserm, Lyon
Daniel Racoceanu, Professeur, Sorbonne Université, Institut du Cerveau, Paris

Examinateurs
Sylvie Chambon, Maître de conférence, INP-ENSEEIHT, IRIT, Toulouse
Eric Debreuve, Chargé de Recherche, CNRS, I3S, Sophia-Antipolis
Xavier Descombes, Directeur de Recherche, INRIA, Sophia-Antipolis
Florence Tupin, Professeure, Télécom Paris, LTCI Paris

Invités
Aurélia Vernay, Ingénieur R&D, BAYER, Lyon

Résumé

Les champignons phytopathogènes sont à l'origine d'importantes pertes économiques. Parmi ces champignons, *Botrytis cinerea* est particulièrement destructeur. Afin de protéger les plantes, des molécules anti-fongiques sont développées. Elles sont classées selon leur Mode d'Action dont la compréhension permet d'élucider la façon dont les composés actifs bloquent les fonctions métaboliques ou voie de signalisation intracellulaires du champignon. Certaines molécules peuvent induire chez ce champignon des changements morphologiques dramatiques, ou phénotypes, observables par microscopie et associés au mode d'action (connu ou non) de la molécule étudiée. Chaque molécule est testée à plusieurs concentrations car il existe une dépendance entre la concentration et le phénotype. A ce jour l'analyse des images de microscopie se fait manuellement. Elle représente donc un coût important qui peut être réduit drastiquement par une analyse informatique. Mon projet de thèse s'inscrit dans ce cadre et vise à mettre en évidence les relations « Famille de Molécules <=> Mode d'Action <=> Phénotype » pour de nouvelles molécules testées. Afin de caractériser les différents phénotypes de *Botrytis cinerea*, nous avons mis en place une analyse automatique des images de microscopie. Elle comprend des étapes de traitement d'images et d'extraction de paramètres morphométriques. Puis, une méthode de classification automatique des phénotypes incluant une classe de rejet pour les phénotypes encore inconnus a été développée. Elle propose une stratégie générale dans un contexte supervisé fondée sur trois étapes principales : apprentissage d'un modèle indépendamment pour chaque classe, apprentissage d'un seuil par modèle fondé sur les interactions entre classes, et procédure de prédiction s'appuyant sur les réponses des modèles par rapport à leur seuil. Un "système expert" proposant une hypothèse de Mode d'Action d'une molécule anti-fongique a également été développé pour prendre en compte l'ensemble des décisions aux différentes concentrations de la molécule. Outre la conclusion sur le mécanisme d'action, cette procédure permet d'obtenir une analyse de la molécule testée, notamment en fournissant des indications sur son degré d'efficacité. Nous avons également développé une approche de classification alternative fondée sur le transport optimal. A noter que cette approche offre en outre un moyen original d'estimer la fonction de densité de probabilité sous-jacente à une population. La force du transport optimal réside dans sa capacité à prendre en compte la géométrie de répartition des échantillons. Ainsi, nous avons proposé de transformer les données de sorte qu'elles suivent un modèle simple (en pratique gaussien), la complexité des données étant alors "cachée" dans la transformation de transport. Enfin, nous nous sommes intéressés à la croissance du champignon au cours du temps dans le but de comprendre voire de prédire l'apparition d'un phénotype. Pour chaque phénotype, différents paramètres morphométriques sont estimés d'après des séquences d'images de croissance. Pour cela, nous avons étudié l'évolution de la valeur de ces paramètres en fonction de la molécule testée, de sa concentration, et du temps d'incubation. Ensuite, nous avons conçu des modèles de croissances calibrés à partir de ces données réelles. Les modèles construits sont des processus stochastiques à temps discret utilisant des lois discrètes et continues pour piloter

les différents événements (croissance, création d'une branche...) et leur ampleur. Nous avons alors simulé la croissance de champignons suivant les traitements testés, pour des phénotypes donnés. Ce travail a permis d'acquérir une meilleure compréhension de la croissance de *Botrytis cinerea* en présence d'une molécule antifongique en fonction de son mode d'action.

Mots clés : Classification, traitement et analyse d'images, transport optimal, modèle de croissance d'un champignon, imagerie biologique, mécanisme d'action de molécules antifongiques.

Abstract

Phytopathogenic fungus are the cause of significant economic losses. Among them, *Botrytis cinerea* is particularly destructive. Therefore, anti-fungal molecules are developed for crop protection. They are classified according to their Mode of Action, whose understanding is necessary to infer how the active compounds block the metabolic functions or intracellular signaling pathways of the fungus. Some molecules can induce dramatic morphological changes of a fungus, the so-called phenotypes, that are observable in microscopy and can be associated with the (known or unknown) mode of action of the molecule. Each molecule must be tested at various concentrations since the phenotype is exhibited only above a certain dose. To date, the analysis of microscopy images is done manually. Therefore, it represents a significant cost which can be drastically reduced by computer analysis. Within this framework, this PhD thesis aims at discovering the relationships “Family of Molecules <=> Mode of Action <=> Phenotype” for new molecules. In order to characterize the different phenotypes of *Botrytis cinerea*, we developed an automatic analysis of the microscopy images. The first steps rely on image processing and extraction of morphometric features. Then, a method of automatic classification of the phenotypes including a rejection class for unknown phenotypes was developed. It proposes a general strategy in a supervised context based on three main steps : learning a model independently for each class, learning one threshold per model based on the interactions between the classes, and a prediction procedure based on the responses of models with respect to their threshold. An "expert system" able to take into account all the decisions at the different concentrations of a molecule has been developed to propose a hypothesis of the Mode of Action of the molecule. Besides the conclusion on the mechanism of action, this procedure allows to obtain an analysis of the tested molecule, in particular by providing indications on its degree of effectiveness. We have also developed an alternative classification approach based on optimal transport whose strength lies in its ability to take into account the geometry of the sample distribution. We proposed to transform the data so that they follow a simple model (in practice a Gaussian model), the complexity of the data then being "hidden" in the transport transformation. Note that this approach also offers an original way to estimate the probability density function underlying a population sample. Finally, we studied the growth of the fungus over time in order to understand or even predict the appearance of a phenotype. For each phenotype, different morphometric features are estimated from temporal sequences in microscopy. This is done by analyzing the evolution of these features as a function of the tested molecule, its concentration, and the incubation time. Then, we designed growth models calibrated from these ground-truth data. The models are discrete-time stochastic processes using discrete and continuous probability laws to control the triggering of the different events (growth, creation of a branch, etc.) and their magnitude. We then simulated the growth of fungi in contexts corresponding to different phenotypes. This work has provided a better understanding of the growth of *Botrytis cinerea* in the presence of an antifungal molecule, i.e., for a given mode of action.

Keywords : Classification, image processing and analysis, optimal transport, growth model of a fungus, biological imaging, mechanism of action of antifungal molecules.

Table des matières

Table des matières vii

Liste des figures x

Liste des tableaux xxii

Projet collaboratif entre un laboratoire académique et un laboratoire privé 1
0.1 Équipe Excellence Biochimie : Centre de Recherche de la Dargoire, Bayer . . 2
0.2 Équipe Morpheme : Laboratoire INRIA/I3S/iBV 3

1 Introduction 4
1.1 Modèle biologique étudié : le champignon pythopathogène *Botrytis cinerea* 5
1.2 Traitements fongicides anti-*Botrytis* 7
1.3 Étude du Mode d'Action de nouvelles molécules 9
1.4 Contributions 12
1.5 Organisation du manuscrit 12
1.5.1 Chapitre 1 12
1.5.2 Chapitre 2 13
1.5.3 Chapitre 3 13
1.5.4 Chapitre 4 13
1.5.5 Chapitre 5 13
1.6 Références 14

2 Classification de phénotypes de *Botrytis cinerea* 16
2.1 Quelques outils de traitement d'image 16
2.1.1 Définition d'une image 16
2.1.2 Amélioration de la qualité des images 17
2.1.3 Morphologie mathématique 17
2.1.4 Segmentation 20
2.2 Apprentissage automatique (ou "Machine Learning") 21
2.2.1 Introduction sur la classification 26
2.2.2 Quelques méthodes de classification 27
2.2.3 Évaluation de la précision d'un classifieur 33
2.2.4 Extraction de caractéristiques 34
2.3 Quelques outils d'analyse d'images et de classification utilisés pour le phénotypage 37
2.4 Origine et nature des données à classer 38
2.4.1 Protocole d'expérimentation biologique 38
2.4.2 Images par microscopie à lumière transmise 40
2.4.3 Notre première problématique : la classification des phénotypes . . . 40

2.5 Caractérisation et classification des phénotypes . . . 43
2.5.1 Description des phénotypes . . . 43
2.5.2 Deux méthodes de classification . . . 60
2.5.3 Résultats de l'étude comparative des deux méthodes de classification 67
2.5.4 Discussion . . . 70
2.6 Références . . . 72

3 Classification avec classe de rejet 78
3.1 État de l'art . . . 78
3.2 Méthode proposée . . . 83
3.2.1 Notations et principe de la stratégie proposée . . . 83
3.2.2 Sensibilité des méthodes de calcul des seuils aux outliers : . . . 86
3.2.3 Choix du modèle : GMM . . . 86
3.2.4 Paramètres à optimiser . . . 86
3.3 Évaluation et validation de la méthode sur des données synthétiques . . . 88
3.3.1 Description des données . . . 88
3.3.2 Résultats . . . 89
3.3.3 Discussion . . . 99
3.3.4 Conclusion . . . 100
3.4 Test de la méthode sur les images des phénotypes connus et inconnus de *Botrytis cinerea* . . . 101
3.4.1 Description des données . . . 101
3.4.2 Description des caractéristiques . . . 101
3.4.3 Classe de rejet fondée sur un CNN . . . 102
3.4.4 Résultats . . . 105
3.4.5 Discussion . . . 106
3.4.6 Conclusion . . . 107
3.5 De la vignette à la molécule : procédure globale de classification . . . 107
3.5.1 Une classification en deux étapes . . . 107
3.5.2 Règles de prédiction du MoA d'une molécule . . . 108
3.5.3 Résultats (Hypothèse de MoA) . . . 109
3.5.4 Discussions et conclusions . . . 111
3.6 Références . . . 111

4 Une approche de classification par transport optimal 114
4.1 État de l'art . . . 114
4.2 Motivations . . . 121
4.3 Description de la méthode . . . 121
4.3.1 Paramètres à optimiser . . . 123
4.4 Évaluation et validation de la méthode sur des données synthétiques . . . 123
4.4.1 Description des données . . . 123
4.4.2 Méthodes d'évaluation . . . 123
4.4.3 Résultats . . . 125
4.4.4 Discussions et conclusions . . . 131
4.5 Test de la méthode sur les images des phénotypes connus et inconnus de *Botrytis cinerea* . . . 133
4.5.1 Description des données . . . 133
4.5.2 Perspectives . . . 135
4.6 Classification avec rejet . . . 135
4.7 Références . . . 136

5 Modélisation de la croissance du champignon *Botrytis cinerea* au cours du temps 138
5.1 État de l'art . . . 138
5.2 Méthode proposée . . . 141
5.3 Présentation des TimeLapses . . . 142
5.3.1 Nature des données . . . 142
5.3.2 Analyse des images du TimeLapse . . . 143
5.4 Segmentation et squelettisation avec cohérence temporelle . . . 146
5.4.1 Segmentation et division des champignons tangents . . . 146
5.4.2 Squelettisation . . . 147
5.5 Extraction des paramètres morphométriques et topologiques : les paramètres phénotypiques . . . 151
5.6 Étude de l'influence de la concentration en molécule sur le phénotype 1 . . . 152
5.7 Analyses complémentaires . . . 155
5.7.1 Carte de trajectoires : profil de la molécule B . . . 155
5.7.2 Comparaison des profils de différentes molécules . . . 156
5.7.3 Conclusions . . . 158
5.7.4 Perspectives . . . 158
5.8 Développement d'un modèle de croissance . . . 163
5.8.1 Paramètres des modèles . . . 163
5.8.2 Élaboration du modèle . . . 164
5.8.3 Exemples de simulation . . . 166
5.9 Conclusion . . . 173
5.10 Références . . . 173

Conclusion 175

Perspectives 177
6.11 Références . . . 179

Annexes I
A.1 Les grandes étapes de R&D au sein de Bayer CropScience . . . I
A.2 Cycle asexué chez *Botrytis cinerea* . . . II
A.3 Classification des Mode d'Action de molécules antifongiques selon le FRAC (2021). Figure extraite de FRA. . . . III
A.4 Concentrations . . . IV
A.5 Cristaux . . . V
A.6 Approche de classification par transport optimal : Couplage et "mappage" simultanés . . . VI
A.6.1 Principe de la méthode . . . VI
A.6.2 Paramètres à optimiser . . . IX
A.6.3 Résultats . . . IX
A.6.4 Conclusion . . . XI
A.7 Évolution des paramètres longueur et aire des cellules de champignons avant la 30 ème acquisition . . . XII
A.8 Simulation de la croissance d'un champignon . . . XIII
A.9 Références . . . XV

Liste des figures

1.1 La pourriture grise touche toute les parties du végétal parasité : A : Cotylédons de cornichons, B, F : Fleurs de géranium, C, G : Tiges de tomates, D : Baie de raisin, E : Feuille de vigne, H : Tomate. 6

1.2 Observation en microscopie de la croissance polarisée de *Botrytis cinerea* en milieu liquide : Au site de croissance polarisée, la spore du champignon va former une protrusion après 3h. Cette protrusion s'allonge de manière unidirectionnelle pour former un tube germinatif qui continue de croître pour former du mycélium en 24 heures. 7

1.3 Les molécules antifongiques sont classées selon leur Mode d'Action par le FRAC (Fungicide Resistance Action Committee). La figure est extraite de FRA. Le poster est présenté, en plus grand, dans le paragraphe A.3 de l'annexe. Par exemple, comme illustré sur le schéma d'une cellule de champignon, une molécule peut spécifiquement cibler une enzyme de la paroi, un complexe de la chaine respiratoire, le cytosquelette ou une autre cible de la cellule. 8

1.4 Signature phénotypique caractéristique du traitement chimique utilisé : images en microscopie à lumière transmise, Microscope ImageXpress, Objectif x10. 10

1.5 Schéma simplifié de l'étude du Mode d'Action de molécules antifongiques par microscopie : les molécules sont testées en présence du champignon sur des microplaques. Un grand nombre d'images est généré via un microscope automatisé. Un expert procède à la reconnaissance des phénotypes sur les images et il annote ses conclusions à la main dans un tableau Excel. 10

1.6 Partie apprentissage : Les familles de molécules chimiques aux modes d'action connus sont testées et associées aux phénotypes de champignons obtenus. Partie prédiction : De nouvelles molécules sont testées et une hypothèse sur leurs modes d'action est obtenue en comparant les phénotypes observés à ceux de la base d'apprentissage. Un phénotype inconnu indique une molécule avec un nouveau mode d'action. 11

2.1 Possibilités de connexité pour une image en 2D. 17

2.2 La dilatation de l'image A par l'élément structurant e donne l'image B. L'érosion de l'image B par e donne l'image C. 19

2.3 Principe de la dilatation (A) et de l'érosion (B) : élément structurant symétrique (disque) centré en l'origine. 19

2.4 Exemple de résultats de dilatation (A) et d'érosion (B) sur des images synthétiques : élément structurant symétrique centré en l'origine. 19

2.5 Structure d'un neurone : la fonction d'activation prend la somme de ses entrées x, pondérée par les poids w pour calculer sa sortie. 25

2.6 Hyperplan séparateur de marge maximale : deux exemples d'hyperplans possibles 1 et 2, l'hyperplan 1 à la marge c1 la plus forte. L'hyperplan sépare les échantillons. Les vecteurs de supports sont les observations situées sur les droites frontières. 28

2.7 Cas dans lequel la séparation linéaire n'est pas adaptée ; l'aide d'une fonction noyaux est alors nécessaire pour obtenir une fonction de séparation, celle-ci est non-linéaire. Par exemple ici : $\rho = \sqrt{x_1^2 + x_2^2}$ et $\theta = \arctan(x_2/x_1)$. 28

2.8 À gauche : Arbre de décisions permettant la prédiction de la classe de l'observation x (gauche). À droite : la partition associée dans l'espace des paramètres p_n. 29

2.9 Principe des forêts aléatoires : combinaison de plusieurs arbres de classification entraînés sur n sous-ensembles de données différents tirés selon un tirage aléatoire des données d'apprentissage. 30

2.10 Exemple d'un réseau de convolutions avec deux couches de convolution et de sous-échantillonnage représentées et connectées à deux couches entièrement connectées. Partie "Extraction des caractéristiques" : CONVOLUTION : Un filtre passe sur l'image en calculant le produit de convolution entre la caractéristique et chaque portion de l'image balayée. RELU : fonction d'activation qui normalise les cartes de caractéristiques. POOLING : couche permettant la réduction de la taille des images en la découpant en grille régulière et en gardant au sein de chaque cellule qu'une seule valeur (la valeur maximale ou autre). Partie "Classification" : FLATTEN : les valeurs des dernières cartes de caractéristiques sont misent dans un vecteur formant un descripteur. FULLY CONNECTED : couches de neurones connectées entre elles. SOFTMAX : fonction exponentielle normalisée qui modifie les valeurs des sorties en leur attribuant des valeurs positives dont la somme fait 1. Schéma provenant de la page MathWorks. 32

2.11 Exemple de descente de gradient en une dimension (un paramètre à optimiser). La fonction de perte est ici représentée, à partir d'une valeur du paramètre (initialisée de manière aléatoire), on va optimiser la valeur de ce paramètre en cherchant à minimiser l'erreur. 32

2.12 Deux exemples de fonctions d'activation. La variable z est fixée à la somme pondérée des entrées (équation 2.11). À gauche, la variable z est transmise à la fonction sigmoïde (voir l'équation 2.12). À droite, la fonction RELU est définie comme R(z)=max(0,z) c'est-à-dire que les valeurs positives restent inchangées et les valeurs négatives sont mises à 0. 36

2.13 Culture du champignon *Botrytis cinerea* en milieu solide : Développement selon une croissance radiale sur une boite de Pétri, milieu Potato Dextrose Agar. Observation à +1 et +8 jours d'incubation à 21 °C 39

2.14 Protocole d'expérimentation biologique : préparation des solutions de spores (mère et fille) et des différentes concentrations des molécules par dilution en série. 41

2.15 Vocabulaire associé à la mise en place d'une plaque 96 puits et de l'acquisition des images par microscopie. 41

2.16 Sept classes : six différentes formes du champignon et la classe crystal qui regroupe les images présentant des artefacts liés à l'agrégation de la molécule antifongique utilisée. Les quatre phénotypes de la deuxième case correspondent à des signatures phénotypiques caractéristiques du traitement chimique utilisé. Ces classes sont scindées en deux sous-classes : "Phénotype" et "Non-phénotype". Images en microscopie à lumière transmise, Microscope ImageXpress, Objectif x10. 43
2.17 Zoom sur les champignons d'une image. (Visualisation avec le logiciel ImageJ) 46
2.18 Histogrammes de deux zones de l'image (zoomée x5) : plage à l'intérieur et à l'extérieur du champignon. Les histogrammes sont générés *via* le logiciel ImageJ. 47
2.19 Zoom sur deux différents phénotypes (x5) : A gauche les champignons présentent une forme moins compacte avec des valeurs globalement plus élevées que celles de l'image de droite (Visualisation avec le logiciel ImageJ). . 47
2.20 Résultats issus des étapes de la méthode "segmentation 2" et du squelette calculé sur le masque binaire ainsi obtenu : les contours du champignon sont identifiés en appliquant le filtre de Canny sur nos images normalisées et dont les contrastes sont rehaussés. L'intérieur de l'objet est rempli et le masque est nettoyé en utilisant des opérations de morphologie mathématique. Les objets au bord de l'image sont également supprimés. Le squelette est calculé sur ce masque final. 48
2.21 Zoom (x4) sur une image de mycelium (Visualisation avec le logiciel ImageJ). 49
2.22 Ligne du haut : résultat de la détection obtenu *via* les méthodes "Segmentation 1", "Segmentation 2" et la combinaison des deux. Ligne du bas : résultat de la squelettisation des objets dans chacun des cas. Les cercles en pointillés orange entourent des exemples d'amélioration du contour des objets. 51
2.23 Résultat de la détection (ligne du haut) et de la squelettisation des champignons (ligne du bas) de phénotypes remarquables (1, 2, 3 et 4). 51
2.24 Organigramme du procédé de segmentation d'une image. La méthode "Segmentation 1" est appliquée à l'image, suivant la valeur de la longueur totale du squelette qui en résulte, l'image est caractérisée comme appartenant à la classe mycelium ou non. La méthode "Segmentation 2" est également effectuée sur l'image afin d'obtenir un masque binaire des champignons qui y sont présents. Dans le cas d'une image caractérisée mycelium, ce masque binaire correspond à la segmentation finale de l'image. Dans le cas contraire, les deux masques résultants des segmentations "Segmentation 1" et "Segmentation 2" sont combinés pour obtenir le résultat final. . . 52
2.25 Résultat des squelettes (première colonne) et graphes de ces squelettes (deuxième colonne) sur quatre champignons de différents phénotypes remarquables (1, 2, 3 et 4). 53
2.26 Liste des paramètres morphométriques calculés sur les trois formes des objets détectés à l'image : le masque binaire, le squelette et le graphe. 54
2.27 Illustration sur un champignon schématique des notions de longueur (L), distance au bord (d), longueur pondérée par la distance au bord (S) d'une branche et des notions de nœuds de jonctions (Nj) et d'extrémités (Ne). . . 54
2.28 Boîtes à moustaches représentant la distribution des longueurs totales des squelettes des objets détectés dans les images des différentes classes. Les cercles correspondent aux valeurs aberrantes. La médiane est représentée par la ligne dans la boîte. 58

2.29 Boîtes à moustaches présentant la dispersion du nombre de nœuds d'extrémités des squelettes des objets détectés dans les images des différentes classes. Les cercles correspondent aux valeurs aberrantes et la médiane est représentée par la ligne dans la boîte. 58

2.30 Boîtes à moustaches illustrant la distribution des aires des objets détectés dans les images des différentes classes. Les cercles correspondent aux valeurs aberrantes. La médiane est représentée par la ligne dans la boîte. . . . 59

2.31 Nombre d'objets détectés en fonction de la longueur totale du squelette des objets. Les échantillons correspondent aux images des classes Germination-Inhibition, Phénotypes 1, 2, 3 et 4 ainsi que celles de la classe Mycelium. . . 59

2.32 Importance des paramètres morphométriques dans la classification des objets détectés des sept différentes classes. Les paramètres sont ordonnés du plus au moins important et leur variabilité inter-arbres (écart-type) est représentée par les barres noires. (*Annotations : Nb = Nombre, Var = Variance, Moy = Moyenne*) 61

2.33 Deux différents stades de développement des champignons des classes 1, 2 et 4. 61

2.34 Figure illustrant la différence entre une convolution standard dans les CNN et une Depthwise Separate Convolution : la depthwise et la pointwise convolution (Figures extraites de Mel [January 10th, 2018]), BN = "Batch Normalization" (Tableaux extraits de TSANG [Oct 14, 2018]). 64

2.35 Architecture du réseau MobileNet (Tableau extrait de HOWARD et collab. [2017]). 65

2.36 Protocole de prédiction du MoA d'une molécule à partir des prédictions images, elles-mêmes obtenues à partir des prédictions objets ou vignettes, suivant la méthode de classification utilisée. 66

2.37 Schéma de la cascade de prédictions. 66

2.38 Résultats préliminaires de shape matching : Étude du gradient de l'image afin de détecter les cercles de rayons prédéfinis correspondants aux spores initiaux de ce phénotype. 72

3.1 L'option de rejet fonctionne sur la base des deux évaluateurs Ψ_a et Ψ_b. Les échantillons rejetés sont ceux dont la valeur de Ψ_a est inférieure à σ_a ou dont la valeur de Ψ_b est inférieure à σ_b. Cette figure correspond à la Figure 5 de 3.1. 82

3.2 Schéma de l'architecture de SelectiveNet. (Figure extraite de GEIFMAN et EL-YANIV [2019]). 83

3.3 Schéma de la procédure d'apprentissage et de prédiction de la méthode de classification avec classe de rejet. Lors de la phase d'apprentissage, les modèles et les seuils sont appris sur les données des différentes classes connues et prennent en compte les relations inter-classes. Dans la phase de prédiction, les nouveaux échantillons sont testés par chaque modèle et les seuils correspondants. Leurs réponses servent à prédire l'appartenance de ces échantillons à l'une de ces classes connues ou à une classe inconnue. 84

3.4 Illustration via trois fonctions Gaussiennes de même variance de la dépendance des seuils FMF à la proximité entre les classes. 84

3.5 Apprentissage du seuil PDF illustré sur des données synthétiques avec les deux options : fondé sur la régression (en haut) et fondé sur le nombre de mauvaises classifications (en bas). Chaque couleur de courbe et de ligne de seuillage en pointillés correspond à un résultat d'apprentissage avec l'ensemble d'apprentissage composé des échantillons de croix et de points noirs plus l'échantillon aberrant "croix" de la même couleur. On peut observer que le décalage de la position de l'échantillon "croix" aberrant déplace également le seuil dans le même sens lors de l'apprentissage avec régression, alors que le seuil appris en utilisant le critère de classifications erronées ne change pas (tant que la valeur aberrante n'entre pas dans le $[T_{min}, T_{max}]$ comme mentionné dans le texte). Notez que les valeurs aberrantes ont reçu une pondération de cinq afin de souligner leur effet. 87

3.6 Ensemble de données des jeux "Moon data", "Ring data" et "S data". 89

3.7 Première colonne : données synthétiques modélisées par une gaussienne. Deuxième colonne : carte de prédictions. Un cercle par population est tracé, de rayon le seuil calculé en fonction des autres modèles et de centre la moyenne des échantillons. Les flèches jaunes indiquent le modèle le plus proche, celui qui permet de définir le seuil sur la probabilité d'appartenance. 90

3.8 Apprentissage des seuils du modèle de la classe 1 (dans le cas d'un ensemble de données à six populations) : Dans le cas de l'**approche "1-vs-1"**, cinq seuils *s* sur les PDFs (log) sont appris : un entre la classe 1 et la classe 0 (A), la classe 1 et la classe 2 (B), la classe 1 et la classe 3 (C), la classe 1 et la classe 4 (D) et enfin entre la classe 1 et la classe 5 (E). Pour chaque calcul, les échantillons de la classe 1 sont étiquetés 1 et ceux de l'autre classe sont étiquetés 0. Les figures de la première colonne illustrent le calcul du seuil par la méthode misclassification. La figure correspond au nombre d'échantillons mal classés en fonction du seuil testé. La seconde colonne présente le calcul du seuil par la méthode de régression. Le seuil calculé est égal à la valeur de l'antécédent de la valeur 0.5 par la fonction de régression estimée sur les valeurs de PDFs du modèle 1 (GMM) pour les échantillons de la classe 1 (label 1) et des échantillons labellisés 0. Le seuil trouvé par misclassification est également indiqué dans le but de comparer les deux possibilités. 92

3.9 Apprentissage des seuils du modèle de la classe 1 (dans le cas d'un ensemble de données à six populations) : Dans le cas de l'**approche "1-vs-all"**, un seul seuil *s* sur les PDFs (log) est appris. Les échantillons de la classe 1 sont étiquetés 1 et les échantillons de toutes les autres classes sont étiquetés 0. Les figures de la première colonne illustrent le calcul du seuil par la méthode misclassification. La figure correspond au nombre d'échantillons mal classés en fonction du seuil testé. La seconde colonne présente le calcul du seuil par la méthode de régression. Le seuil calculé est égal à la valeur de l'antécédent de la valeur 0.5 par la fonction de régression estimée sur les valeurs de PDFs du modèle 1 (GMM) pour les échantillons de la classe 1 (label 1) et des échantillons labellisés 0. 93

3.10 Apprentissage des seuils du modèle 1 : Les figures de la première colonne illustrent le calcul du seuil *s* sur les PDFs (log) par la méthode misclassification. Le nombre d'échantillons mal classés en fonction du seuil testé est calculé pour 1000 valeurs de seuils (dans l'intervalle de la valeur minimale à la valeur maximale des PDFs de l'ensemble d'apprentissage). Les figures de la deuxième colonne présentent le calcul du seuil par la méthode de régression, le seuil est égal à la valeur de l'antécédent de la valeur 0.5 par la fonction de régression estimée sur les valeurs de PDFs du modèle 1. Les échantillons des classes 1 et 2 sont respectivement étiquetés 1 et 0. 95

3.11 Illustration des résultats de classification sur les bases de données "Moon data" et "Ring data". Les croix noires correspondent aux échantillons de l'ensemble d'apprentissage. (A) Ensembles d'apprentissage et de test. (B) Résultats de classification sans classe de rejet des échantillons de test. La couleur correspond aux valeurs des PDFs. (C) Résultats de classification sans classe de rejet des échantillons d'une grille régulière définie. La couleur d'un échantillon correspond à la classe prédite. (D) Résultats de classification avec classe de rejet des échantillons d'une grille régulière définie. La couleur correspond aux valeurs des PDFs et la couleur saumon indique la zone de rejet dans l'espace des caractéristiques. 96

3.12 Illustration des résultats de classification sur la base de données "S data". Les croix noires correspondent aux échantillons de l'ensemble d'apprentissage. (A) Ensembles d'apprentissage et de test. (B) Résultats de classification sans classe de rejet des échantillons de test. La couleur correspond aux valeurs des PDFs. (C) Résultats de classification sans classe de rejet des échantillons d'une grille régulière définie. La couleur d'un échantillon correspond à la classe prédite. (D) Résultats de classification avec classe de rejet des échantillons d'une grille régulière définie. La couleur correspond aux valeurs des PDFs et la couleur saumon indique la zone de rejet dans l'espace des caractéristiques. 97

3.13 Phénotypes connus (1-4) correspondant à des signatures phénotypiques caractéristiques du traitement chimique utilisé et quelques exemples de nouveaux phénotypes. Images en microscopie à lumière transmise, Microscope ImageXpress, Objectif x10. 101

3.14 Exemple de deux phénotypes différents considérés comme nouveaux phénotypes car ne faisant pas partie des sept classes sur lesquelles le CNN a été entraîné. 103

3.15 Illustration du découpage en vignettes d'une image dont l'information relative aux champignons est répartie de manière hétérogène entraînant des sous-images quasiment vides (vignettes encadrées en rouge). 103

3.16 Exemple de symétrisation de données Output à quatre dimensions (données illustrées ici seulement de manière indépendante par dimension). Estimation de gaussiennes sur les histogrammes des quatre caractéristiques des échantillons d'une classe donnée : la première et deuxième ligne représentent respectivement les données originales et les données symétrisées. Les Gaussiennes se rapprochent mieux des données lorsqu'elles sont estimées sur les données symétrisées. 104

3.17 Classification en deux étapes. Étape 1 : classification des images dans les classes Germination-Inhibition, Crystal, mycelium et Phénotypes en fonction des scores obtenus par le CNN. Étape 2 : classification des images dans les classes Phénotypes 1-4 et Nouveau en fonction des réponses des modèles GMM et des seuils appris. 108
3.18 Schéma général de classification. 110

4.1 Cas d'un transport d'une distribution discrète μ (points x_i) vers une distribution ν (points y_j). Les masses des points sont représentées via leurs aires et le transport est tel que T($x1$) = T($x2$) = $y1$ et T($x3$) = T($x4$) = T($x5$) = $y2$. . 114
4.2 A gauche : Cas de deux droites perpendiculaires. À droite : Cas d'un nombre de masses dans l'espace source inférieur à celui dans l'espace cible. 115
4.3 Transfert de la palette de couleurs de Picasso sur un tableau de Cézanne : Le transport T est calculé en utilisant un coût $C_{x,y}$ dont la valeur dépend de la similarité entre les couleurs p_x et p_y des deux pixels. Moins les couleurs sont proches plus $C_{x,y}$ est élevé. La valeur du pixel p_x de l'image résultante est remplacée par celle de p_y. – source PEYRÉ [19 janvier 2017] 117
4.4 Égalisation d'histogramme dans le but de créer une image contrastée : les niveaux de gris sont redistribués par TO en fonction de t qui paramètre l'interpolation du déplacement entre les histogrammes. Ligne du haut : Évolution des images, Ligne du bas : évolution de l'histogramme des niveaux de gris des images correspondantes – source COT [2019]. 117
4.5 Le transport optimal pour l'adaptation du domaine. À gauche : les données d'apprentissage dans le domaine source, et les données tests dans le domaine cible. Le classifieur estimé ne permet pas la classification des données cibles. Au milieu : transport des observations d'apprentissage dans le domaine cible via une carte de transport T ∗ γ0 . La transformation n'est généralement pas linéaire. À droite : calcul d'un classifieur efficace dans le domaine cible. – source COURTY et collab. [2016]. 118
4.6 Les distances définies par le TO (wasserstein) entre deux distributions de probabilité 1D bleue et rouge. Les moyennes des distributions sont initialisées à 80 puis la distribution rouge se décale. 119
4.7 Influence du paramètre de régularisation entropique sur la dispersion de la masse des distributions α à β sur le plan de TO régularisé γ ∗ (λ). Le plan de TO indique la proportion de matière au voisinage de x dans α que l'on va transporter au voisinage de y dans β. À mesure que ϵ > 0 croit la solution γ ∗ (λ) "s'étale" de plus en plus. – source FLAMARY [novembre 2017] où λ le terme de régularisation est noté ϵ dans le cas de la régularisation par entropie. 121
4.8 Lignes de niveau des PDFs théoriques des distributions des deux classes de données synthétiques. 124
4.9 Échantillons des deux classes correspondant à notre ensemble de données synthétiques. 124
4.10 PDFs théoriques des échantillons des deux classes 1 et 2. 124
4.11 A gauche : les échantillons de la classe 1 (décrits dans le paragraphe 4.4.1) dans l'espace source. À droite : les échantillons générés suivant une loi normale, de mêmes moyenne et variance correspondants aux données cibles. . 126
4.12 A gauche : la matrice de coût des échantillons source de la classe 1. À droite : la matrice de coût des échantillons cible générés. 126

4.13 Plan de TO des échantillons source et cible de la classe 1 obtenue avec l'approche par transport optimal sans régularisation ("EMD"). Cette matrice indique la proportion de matière de x dans la distribution source que l'on va transporter sur un point y de la distribution cible. La matière est envoyée vers un unique point cible, le poids d'un point est fixé à 0.00025. 127

4.14 Plan de TO des échantillons source et cible de la classe 1 obtenue avec l'approche par transport optimal avec régularisation ("Sinkhorn"). Cette matrice indique la proportion de matière de x dans la distribution source que l'on va transporter sur plusieurs points de la distribution cible. La couleur du point indique la valeur du poids de l'échantillon source transporté vers le point cible. 127

4.15 A gauche : les échantillons de la classe 1 (décrits dans le paragraphe 4.4.1) dans l'espace source. Au milieu et à droite : les échantillons dits transportés issus du résultat du couplage par "EMD" et "Sinkhorn". 127

4.16 A gauche : les PDFs des échantillons de la classe 1 dans l'espace source. À droite : les PDFs des échantillons transportés correspondants. 128

4.17 A gauche : masque binaire délimitant la zone de validité des populations de la classe 1. À droite : résultat de l'interpolation des valeurs de PDFs des échantillons de l'ensemble d'apprentissage de la classe 1 pour les points présents à l'intérieur de la zone de validité. 128

4.18 Cartes de prédiction obtenues avec : Figure A : les PDFs théoriques et Figure B : les PDFs estimées par l'ajustement d'une gaussienne directement sur les données. La figure C correspond aux points de la grille dont la prédiction est différente de celle de la carte figure A. 130

4.19 Cartes de prédiction obtenues avec : Figure A : les PDFs théoriques et Figure B : les PDFs estimées par la méthode utilisant le transport optimal ("EMD"). La figure C correspond aux points de la grille dont la prédiction est différente de celle de la carte figure A. La carte dans le cas de l'approche "Sinkhorn" n'est pas affichée car très proche de celle-ci. 130

4.20 Exemple de distributions étiquetées en 4 couches. **A :** Résultat obtenu avec la méthode "Alpha-shape". **B :** Résultat obtenu avec la méthode "GMM pdf". Les couplages sont contraints de se faire entre couches source et cible de même label et le nombre de couches est un paramètre à fixer. 132

4.21 **(A) :** Distance de Wasserstein. **(B) :** Distance de Gromov. 132

4.22 Données Output et Bottlenecks dont la dimension est réduite par ACP à 2 dimensions. 133

4.23 Données Output et Bottlenecks dont la dimension est réduite par ACP à 2 dimensions. Les échantillons appartiennent aux classes connues (1-4) ou correspondent à des nouveaux phénotypes. 136

5.1 Courbe de Koch générée via l'approche du L-système, après un nombre de 1 (**A**) et 4 (**B**) itérations. 139

5.2 A : Illustration de l'architecture végétale sous la forme d'un graphe arborescent multi-échelles (MTG). La figure est extraite de l'article GODIN et CARAGLIO [1998]. B : Graphe arborescent correspondant. 140

5.3 Illustration du procédé du logiciel AMAPmod. La figure est extraite de l'article GODIN et collab. [1999] : (a) Mesures réelles issues d'observations sur le terrain, (b) Code informatique décrivant la topologie de la plante à différentes échelles (observations faites en (a)), (c) construction d'une représentation interne de l'architecture de la plante (graphes arborescents multi-échelles) par le logiciel à partir du code. (d) extraction d'informations à partir de la représentation interne, (e) analyses statistiques de ces données *via* des modèles probabilistes ou stochastiques. **AML :** langage de modélisation AMAP 141

5.4 **A :** Expérimentation biologique, **B :** Acquisition des images par microscopie, **C :** Traitement des images, **D :** Mesures des caractéristiques, **E :** Analyse des données, **F :** Mise en place des modèles et simulation de la croissance d'un phénotype remarquable suivant une molécule et une concentration. 142

5.5 Images à 12 heures après inoculation des spores. 143

5.6 Image 2 du champ pris dans le puits 7 aux temps 0, 3.75 et 7.5 heures, illustrant le changement de place des spores dans le milieu. Entourés d'un cercle jaune : les spores. À 7.5 heures, la flèche rouge montre le déplacement de la spore entre les temps 3.75 et 7.5 heures. Le cercle en pointillés rouges entoure une spore absente du champ de l'image à 3.75 heures. 144

5.7 Image 2 du champ pris dans le puits 7 aux temps 32.75 et 36.25 heures, illustrant l'explosion des cellules de champignon. 145

5.8 Image 3 du champ pris dans le puits 7 aux temps 24, 32.75 et 37.75 heures, illustrant le développement des cellules de champignon de phénotype 2. . . 145

5.9 Somme des masques binaires au cours du temps de certains objets (exemples de phénotypes 1,2,3, mycelium et spore) du temps 0 au temps 100. Lorsque la couleur du pixel tend vers le rouge, le pixel est présent dans un nombre croissant de masques, et inversement lorsque sa couleur tend vers le bleu. . 146

5.10 Résultats de segmentation de deux images d'un même TimeLapse du phénotype 1, images prises dans le puits 7 de la plaque (concentration en molécule de 1.23µm) : à T_0 (à gauche) et T_{100} (à droite). Les masques binaires sont multipliés par les images originales permettant une meilleure vérification de la segmentation des objets. 148

5.11 Résultats de labellisation des objets de deux images d'un même TimeLapse du phénotype 1, images prises dans le puits 7 (concentration en molécule de 1.23µm) : à T_0 (à gauche) et T_{100} (à droite). Les objets sont mis en correspondance d'un temps à l'autre. 148

5.12 Résultat de l'utilisation de l'algorithme de segmentation Watershed pour la séparation des champignons d'un même objet au temps T_{89}. 149

5.13 Résultats de la correction de labellisation des pixels (objets aux temps T_{88}, T_{89}, T_{90} et T_{91}) par l'algorithme de Watershed dans le cas de la segmentation des champignons initialement numérotés 8 et 9. 149

5.14 Résultats de labellisation des objets des images du TimeLapse : à T_0 et T_{100} . Les objets sont mis en correspondance d'un temps à l'autre. La labellisation est corrigée grâce à l'algorithme de Watershed. 149

5.15 Résultats sur onze temps, de la superposition du graphe obtenu, sur le masque binaire du champignon détecté (de phénotype 1). 150

5.16 Résultats sur onze temps, de la superposition du graphe obtenu, sur le masque binaire du champignon détecté (de phénotype 2). 150

5.17 Deux clusters de phénotype de champignons. Les champignons sont regroupés suivant leurs valeurs de paramètres phénotypiques, par la méthode de clustering k-Means. À gauche : les champignons et leurs graphes correspondants. À droite : masque binaire correspondant. 151
5.18 Deux images de champignons traités avec la molécule A qui présentent pour certains le phénotype 2 (cercles bleu) et pour d'autres, la forme spore (cercles jaune). 152
5.19 Boîtes à moustaches avant (à gauche) et après (à droite) suppression des valeurs aberrantes dans le cas des champignons traités avec la molécule B aux concentrations 08 et 09. La barre noire correspond à la médiane des observations et le rectangle de la boîte va du premier quartile au troisième quartile. La longueur des moustaches vaut 1,5 fois l'intervalle inter-quartile. 152
5.20 Évolution des paramètres phénotypiques (moyenne sur les champignons d'une même concentration) du phénotype 1 au cours du temps. **B :** molécule dont le MoA entraîne le phénotype 1. **06 à 09 :** concentrations décroissantes en molécule. 153
5.21 Distribution des valeurs (à T_{100}) des quatre paramètres phénotypiques pour les concentrations où le phénotype 1 est observable. **A :** aire, **B :** longueur, **C :** nombre de branches et **D :** nombre de branches partants de la spore initiale. Les cercles correspondent aux outliers. La médiane est représentée par la ligne noire dans la boîte. 155
5.22 Exemples de champignons présents sur les images, issues du test de la molécule B, aux concentrations 02, 03, 04, 05, 06, 07, 08, 09, 10, 11 et 12. 159
5.23 Carte de trajectoires des valeurs moyennes des paramètres aire et longueur des objets détectés dans les images issues du test de la molécule B. Les courbes bleues correspondent au phénotype spore (concentrations 03, 04 et 05), les rouges au phénotype 1 (concentrations 06, 07, 08 et 09) et les vertes à la forme mycelium (concentrations 02, 10, 11 et 12). 159
5.24 Carte de trajectoires des valeurs moyennes des paramètres aire et longueur des objets détectés dans les images issues du test des molécules A (courbes rouges), B (courbes bleues) et C (courbes vertes) aux concentrations 06 et 07. 160
5.25 Carte de trajectoires des valeurs moyennes des paramètres aire et longueur des objets détectés dans les images issues du test des molécules A (courbes rouges) et D (courbes bleues) aux concentrations où le phénotype 2 apparaît. 160
5.26 Carte de trajectoires des valeurs moyennes des paramètres aire et longueur des objets détectés dans les images issues du test des molécules et de leurs réplicats. Dans l'ordre de haut en bas : cas des molécules G et F, puis des molécules H et I et enfin des molécules K et J. 161
5.27 Carte de trajectoires des valeurs moyennes des paramètres aire et longueur des objets détectés dans les images issues du test des molécules B (courbes roses), E (courbes rouges) F (courbes bleues), H (courbes moutardes), et J (courbes vertes) aux concentrations où le phénotype 1 apparaît. 162
5.28 Fonctions exponentielles finales (courbes noires) dont les coefficients sont estimés à partir des valeurs des paramètres phénotypiques (points en couleur). 167
5.29 Distributions des coefficients a et b des fonctions exponentielles ajustées sur l'évolution des paramètres aire et longueur pour les différentes concentrations. 168

5.30 Nombre de branches créées à chaque intervalle de temps en considérant l'ensemble des objets à chaque concentration. 169
5.31 Illustration de la recherche des coordonnées d'un nouveau nœud et d'une nouvelle branche en fonction de la longueur calculée, et de l'angle tiré aléatoirement dans un intervalle donné. 170
5.32 Exemple de résultat de la simulation, au temps T_{final}, d'un champignon du phénotype 1 (molécule B à la concentration 07) *via* notre modèle de croissance : Le graphe, le squelette et l'objet correspondant. 170
5.33 Exemple de résultats, au temps 100, de la simulation *via* notre modèle de croissance de neuf champignons du phénotype 1 (molécule B à la concentration 07). **A :** les squelettes et **B :** les objets correspondants. 171
5.34 Évolution des paramètres longueur, nombre de branches terminales et nombre de branches primaires (moyenne sur les champignons d'une même concentration) du phénotype 1 au cours du temps. B : molécule dont le MoA entraîne le phénotype 1. 06 à 09 : concentrations décroissantes en molécule. **A :** Données réelles et **B :** Données issues de la simulation (20 simulations par concentration). 172
35 Cycle asexué de Botrytis : Les sclérotes germent et pénètrent dans le tissu végétal de l'hôte quand les conditions de température et d'humidité sont favorables. Le développement aboutit à la formation d'un mycélium portant des conidiophores; hyphe aérienne portant une ou plusieurs conidies (Ascospore). Après la sporulation, la spore va adhérer à la surface de la plante et va germer. Cela mène à la formation d'un appressorium, qui en exerçant une pression mécanique (pression de turgescence) va passer la paroi pectocellulosique et s'accoler à la membrane plasmique de la cellule hôte. II
36 Six exemples de cristaux obtenus en conséquence des problèmes de solubilité de certaines molécules dans le solvant utilisé. V
37 Illustration de l'apprentissage des transports des échantillons de chaque classe de l'espace source vers les échantillons générés dans l'espace cible selon la loi correspondante. VI
38 Illustration du transport de l'ensemble des échantillons de l'ensemble d'apprentissage via le transport optimal appris T1 (Seul le modèle de la classe 1 est ici considéré). VII
39 Illustration de l'apprentissage du seuil sur des données de synthèse (Seul le modèle de la classe 1 est ici concidéré). Les échantillons de la classe 1 reçoivent l'étiquette 1. Les autres (soit les échantillons de toutes les autres classes en "1-vs-all", soit ceux des autres classes à tour de rôle en "1-vs-1") reçoivent l'étiquette 0. VII
40 Illustration du protocole de prédiction dans le cas de deux classes connues (1 et 2) pour lesquels le TO et le seuil sont optimisés lors de la phase d'apprentissage. VIII
41 À gauche : Les échantillons de l'ensemble d'apprentissage des deux classes. À droite : Les échantillons de l'ensemble de test des deux classes colorés en fonction de la valeur de la PDF cible aux échantillons transportés par le transport optimal appris pour la classe 2. X
42 Boîtes à moustaches des distributions des valeurs des paramètres aire et longueur des squelettes des objets détectés sur les images aux temps T_1, T_5, T_{10}, T_{15}, T_{20}, T_{25} et T_{29}. Les cercles correspondent aux valeurs aberrantes et la ligne noire dans chaque boîte représente la médiane. XII

43 Évolution du graphe au cours du temps. XIV

Liste des tableaux

2.1 Exemple de matrice de confusion illustrant les résultats de classification normalisés d'échantillons de trois classes (A, B et C). Chaque ligne de cette matrice correspond à une classe réelle et chaque colonne correspond à une classe prédite. Cette matrice indique pour chaque classe une valeur entre 0 et 1, indiquant le rapport des observations prédites comme appartenant à chacune des classes (la somme de ces rapports fait 1). 33

2.2 Caractéristiques morphologiques identifiées décrivant les Phénotypes 1, 2, 3 et 4. Informations relatives à la spore initiale du champignon, aux branches (forme, composition et degrés des branches) et à l'homogénéité de l'effet de la molécule sur les champignons. 44

2.3 Valeurs des paramètres pour les classifications en une et deux étapes (décrites dans le paragraphe 2.5.2) : nombre d'arbres de la forêt et nombre de caractéristiques sélectionnées aléatoirement. 63

2.4 Résultats globaux de classification (en pourcentages) des images issus des classifications par "objets" (scores ramenés à un résultat de classification sur les images par vote majoritaire) et par "images" (médianes des valeurs des paramètres des objets). 68

2.5 Résultats de classification en deux étapes (en pourcentages) des images. G-I correspond à Germination-Inhibition. 69

2.6 Matrice de confusion des résultats de classification (en pourcentages) des images obtenus avec la méthode des forêts aléatoire. G-I correspond à Germination-Inhibition. 69

2.7 Résultats de classification (en pourcentages) des mécanismes d'action des molécules testées calculées suivant les règles de prédiction citées dans le paragraphe 2.5.2 *via* les résultats obtenues sur les images. 69

2.8 Résultats de classification en pourcentages par classes obtenus par le réseau de neurones MobileNet entrainé sur nos sept classes d'images. La première ligne correspond aux scores de prédiction sur la classe des vignettes. La seconde ligne fait état des résultats de classification des images qui sont obtenus par vote majoritaire des prédictions de la classe des vignettes correspondantes. (G-I correspond à Germination-Inhibition) 70

2.9 Matrice de confusion des résultats de classification (en pourcentages) des images. (G-I correspond à Germination-Inhibition) 70

2.10 Résultats de classification (en pourcentages) des mécanismes d'action des molécules testées calculées suivant les règles de prédiction citées dans le paragraphe 2.5.2 *via* les résultats obtenues sur les images. 70

2.11 Récapitulatif des résultats de classification des images en pourcentages par classes obtenus par les deux méthodes de classification (RF et CNN). 70

3.1 Nombre d'échantillons par classe dans les trois ensembles de données. . . . 88
3.2 Matrices de confusion des résultats de classification (en pourcentages) des échantillons des bases de données (dans l'ordre) "Moon data", "Ring data" et "S data" obtenus avec le modèle GMM sans la classe de rejet. 98
3.3 Matrices de confusion des résultats de classification (en pourcentages) obtenus avec le modèle GMM avec la classe de rejet. Les échantillons considérés sont ceux des bases de données "Moon data" (en haut), "Ring data" (au milieu) et "S data" (en bas) ainsi que dans chaque cas, des échantillons dits "nouveaux". Les seuils sur les PDF sont estimés à l'aide d'une régression logistique, dans un contexte "1-vs-1". 98
3.4 Scores moyens de classification (en pourcentages) des échantillons des bases de données "Moon", "Ring" et "S", obtenus avec différents classifieurs ; RF : Random Forest, KNeighbors : méthode des k-plus proches voisins, SVM : les machines à vecteurs de support, AdaBoost : algorithme de boosting et la méthode GMM sans la classe de rejet. Les moyennes des scores obtenus avec la méthode de classification avec classe de rejet sont indiquées de manière indépendante sur les classes connues et inconnue, puis une moyenne globale est également calculée. 99
3.5 Précisions de classification en pourcentages pour les différents paramètres expérimentaux. Les valeurs réelles ont été arrondies à l'entier le plus proche. 106
3.6 Matrice de confusion des résultats de classification (en pourcentages) des images. (G-I correspond à Germination-Inhibition) 108
3.7 Résultats de classification (en pourcentages) des modes d'action des molécules testées obtenus suivant les règles de prédiction citées dans le paragraphe 3.5.2. Les six classes de MoAs sont les quatre MoA connus (1-4), la classe nouveau MoA et la classe NA (« No Answer ») qui contient les molécules dont les images ne permettent pas d'émettre une hypothèse de MoA. . 109

4.1 Scores moyens de classification issus de la validation croisée (en pourcentages) des échantillons de l'ensemble de données de synthèses, obtenus avec une Gaussienne et une approche par transport optimal en utilisant deux algorithmes ("Sinkhorn" et "EMD"). 129
4.2 Scores de classification (en pourcentages) des échantillons de la grille régulière, obtenus en utilisant les valeurs de PDFs théoriques et celles estimées par une gaussienne ajustée aux données ou par les approches par transport optimal (appris respectivement avec "EMD" et "Sinkhorn"). 130
4.3 Nombre de composantes des mélanges de gaussiennes (estimé par l'indice BIC), par classe et par type de données considéré (Bottlenecks et Output). . 134
4.4 Scores moyens de classification (en pourcentages) des échantillons Output et Bottlenecks, obtenus avec les GMM puis les approches par transport optimal, "EMD" et "Sinkhorn" permettant le calcul du transport optimal avec et sans régularisation. 134

5.1 Concentrations en spores testées dans le cadre de notre étude. 143
5.2 Paramètres phénotypiques calculés sur les trois formes des objets détectés à l'image : le masque binaire, le squelette et le graphe. 151
5.3 Tableau indiquant les molécules testées dans le cadre de cette étude. Les couleurs regroupent les molécules testées en réplicats. 156
4 Concentrations par puits (en µM). IV

5 Précisions de classification en pourcentage pour les deux méthodes de classification avec classe de rejet (GMM et TO) et les deux types de données (Bottlenecks et Output). Les valeurs réelles ont été arrondies à l'entier le plus proche. X

Projet collaboratif entre un laboratoire académique et un laboratoire privé

« *Even if I knew that tomorrow the world would go to pieces, I would still plant my apple tree.* »

Martin Luther King

Le présent mémoire rend compte des travaux de recherche effectués de concert entre le laboratoire du Centre de Recherche de la Dargoire de Bayer CropScience, un centre de recherche sur les fongicides basé à Lyon ainsi que l'Institut National de Recherche en Automatique et en Informatique (INRIA) et le laboratoire d'Informatique Signaux et Systèmes (I3S), tous deux situés à Sophia Antipolis. Dans la mesure où nous nous plaçons dans un cadre de recherches scientifiques d'une part et vis-à-vis de l'entreprise, dans un contexte industriel d'autre part, cette thèse présente aussi bien un aspect applicatif que théorique. Le fondement théorique de chaque fonctionnalité présentée est étudié et expliqué. Nous tenons également à avoir des résultats exploitables dans un contexte industriel, c'est-à-dire à obtenir des performances élevées et un temps d'exécution acceptable. Dans ce projet mis en place et financé par Bayer CropScience dans le cadre d'une convention CIFRE, chaque laboratoire a une expertise bien définie qui est décrite ci-après.

0.1 Équipe Excellence Biochimie : Centre de Recherche de la Dargoire, Bayer

Ce paragraphe est issu du rapport de stage de Master 2 LAROUI (2017).

Bayer CropScience est l'une des filiales du groupe allemand Bayer, une entreprise combinant la recherche de solutions innovantes dédiées aux maladies des plantes et aux maladies humaines (Bayer HealthCare et Bayer ConsumerHealth).

Situé à Lyon, le Centre de Recherche de la Dargoire (CRLD), créé dans les années 60, est aujourd'hui dédié à l'identification et à l'optimisation de molécules fongicides, qui visent à éliminer ou limiter le développement des champignons parasites des végétaux. Le but est de protéger les plantes contre des maladies fongiques affectant la qualité et la quantité des récoltes et finalement leur rendement. Les solutions proposées peuvent aujourd'hui combiner des approches conventionnelles et/ou des solutions complémentaires comme l'utilisation d'agents de bio-contrôle en tant que fongicides. Ainsi, il faut compter une dizaine d'années depuis la découverte d'une nouvelle molécule, les études sur champignons et sur plantes, les études en champs, les études toxicologiques avant l'obtention d'une homologation et autorisation de mise sur le marché. Le challenge est de trouver une molécule qui empêchera le développement du champignon parasite sans pour autant affecter la plante hôte et sans conséquences pour l'homme et l'environnement.

L'une des missions de l'équipe Excellence Biochimie est d'élucider le mode d'action d'une molécule, c'est-à-dire de comprendre son mécanisme d'action sur le champignon modèle d'intérêt. Afin d'élucider et comprendre le mode d'action d'une molécule, différentes techniques sont utilisées : méthodes biochimiques ou biophysiques, biologie moléculaire, méthodes d'omiques, et microcopie. Certains traitements chimiques peuvent induire chez les cellules des champignons des changements morphologiques dramatiques, ou « phénotypes », observables par microscopie et associés au mode d'action connu ou inconnu des molécules testées. Il s'agit donc de générer des images par microscopie en lumière transmise sur des cellules de champignons non-modifiés puis de visualiser et quantifier l'action de molécules chimiques sur les champignons. L'emmergence de ce projet vient de la volonté d'automatiser ce procédé de reconnaissance des phénotypes obtenus après traitement, en développant des solutions d'analyse d'images et d'intelligence artifi-

cielle pour caractériser l'effet de la molécule sur le modèle biologique étudié ainsi que de développer un modèle de prédiction du mode d'action connu ou inconnu des molécules testées. Ainsi, informatique, mathématiques appliquées et biologie sont combinés dans cette optique.

0.2 Équipe Morpheme : Laboratoire INRIA/I3S/iBV

Ce paragraphe est inspiré de la description de l'équipe sur le site de l'Inria MORPHEME.

Le travail accompli dans ce projet est en total adéquation avec les deux grandes thématiques de l'équipe Morpheme, à savoir : la Biologie Numérique et la BioInformatique. Un des objectifs scientifiques de Morpheme est la caractérisation et la modélisation du développement et des propriétés morphologiques de structures biologiques, ces structures pouvant aller de l'échelle cellulaire à supra-cellulaire (notions de tissu). Le but est de comprendre les changements morphologiques qui apparaissent lors du développement en conjuguant l'imagerie *in vivo* avec l'analyse d'images et la modélisation numérique. En effet, la morphologie et la topologie des structures mésoscopiques (échelle intermédiaire entre le microscopique et le macroscopique) ont une influence majeure sur les multiples fonctionnalités des organes. La forme des structures (cellulaires ou supra-cellulaires) est analysée dans différentes populations et/ou des conditions de développement différentes. Si le développement de solutions d'analyses d'images est le coeur de métier de Morpheme, le développement de modèles de prédiction est aussi un aspect clé de leur expertise. Aujourd'hui, différentes techniques de microscopie telle que la microscopie confocale, le bi-photon, la vidéo-microscopie et la micro-tomographie permettent de générer des images 2D, 2D+t, 3D ou 3D+t à partir desquelles il est possible d'extraire des informations quantitatives. Ces informations caractérisent la morphométrie des échantillons et leur évolution temporelle. Les formes et les structures complexes sont ensuite soumises à des analyses statistiques afin d'identifier des marqueurs significatifs et développer des outils de classification. Le but final est de proposer des modèles pour expliquer l'évolution au cours du temps des échantillons observés et cela afin d'affiner la compréhension du développement des tissus. Par exemple, ces analyses sont réalisées dans des tissus sains mais aussi pour l'étude de différentes pathologies (à un niveau supra-cellulaire) tels que différents cancers.

Chapitre 1

Introduction

La science des données est un domaine interdisciplinaire émergent à l'ère numérique qui unifie les statistiques, l'analyse de données et l'apprentissage automatique pour extraire des connaissances à partir de données. L'apprentissage automatique ou "Machine Learning" en anglais, est un sous-domaine de l'intelligence artificielle San, définissant globalement l'idée selon laquelle : l'algorithme, sur la base d'approches mathématiques et statistiques, apprend à effectuer une tâche à partir de données. Le but est d'automatiser le processus de conversion des données en connaissances en construisant des modèles analytiques. Ce type d'algorithmes peut être utilisé à des fins diverses telles que, par exemple, l'exploration de données, le traitement d'images et la prédiction analytique. Toutes ces notions seront utilisées dans le cadre de mes travaux.
Mon projet de thèse voit le jour lorsqu'un laboratoire de recherche chargé de la mission d'élucider le mode d'action de molécules chimiques antifongiques a vu l'intérêt, compte-tenu du coût important, financier d'une part mais aussi temporel et humain (temps de l'expertise elle-même et temps d'attente de disponibilité de l'expert), d'automatiser ce processus. Ce processus entre dans le cadre d'études menées dans le but de contrôler les maladies provoquées par les champignons phytopathogènes, un contrôle indispensable au maintien d'un bon niveau de production agricole. En effet, plus de 10 000 espèces de champignons phytopathogènes ravagent chaque année un grand nombre de cultures d'intérêt agronomique (cultures céréalières, maraichères et horticoles) conduisant à d'importantes pertes de production Koeck et collab. [2011]. Les maladies causées par ces pathogènes (oïdium, septoriose, rouille, etc.) représentent près de 85% des maladies observées sur les plantes. Un exemple marquant l'importance économique de ces organismes serait celui de *Magnaporthe oryzae*, l'agent pathogène de la pyriculariose du riz duquel dépend la nutrition de la moitié de la population mondiale. Les diverses maladies et les dégâts causés par ces champignons exigent donc en retour une variété de solutions de lutte, stimulant ainsi l'intérêt pour le développement de nouveaux produits fongicides. Le développement de nouvelles solutions implique des équipes pluridisciplinaires de chimistes, biologistes, biochimistes, bio-informaticiens et toxicologistes afin de synthétiser, tester et étudier ces nouveaux produits (voir le paragraphe en annexe A.1). L'une des étapes clefs de cette lutte est l'identification des mécanismes d'action des nouvelles molécules antifongiques produites. Parmi les techniques d'identification de ces modes d'action, on retrouve la notion de phénotypage, à savoir : la détermination (par un expert) du phénotype [1] d'un organisme via des images en microscopie, l'hypothèse étant que mode d'action de molécules antifongiques et phénotypes de champignons sont liés. Or l'extraction d'informations et de variables pertinentes à l'échelle microscopique pose

1. Ensemble des caractères observables d'un organisme.

de nombreux problèmes. Les trois principaux sont la subjectivité de l'expert, la mauvaise reproductibilité ainsi que le coût certain en ressources (financier, temporel et humain). Dans l'optique d'éviter ce lourd travail manuel d'analyse d'images et les difficultés qui en découlent, les algorithmes de traitement d'images constituent un outil puissant. Un des principaux objectifs de ma thèse est donc de mettre au point un système capable de reconnaître de façon automatique les différents phénotypes de champignon, à partir d'images en microscopie (images 2D et 2D+t[2]). Les travaux rapportés dans ce manuscrit sont ancrés dans une discipline scientifique à la frontière entre l'informatique, les mathématiques, les statistiques et la biologie : la bio-informatique.

Ce premier chapitre est consacré à la description du contexte biologique de ce projet et des enjeux de nos travaux. Un organisme modèle est une espèce étudiée pour comprendre un phénomène biologique particulier. Du fait de principes biologiques fondamentaux partagés avec d'autres organismes (voies métaboliques, régulatoires, ..), les résultats d'expériences approfondies effectuées sur le modèle sont généralisables et permettent d'élargir nos connaissances. Au centre de Recherche de la Dragoire, de nombreux champignons parasites des végétaux sont étudiés comme la rouille du soja et le mildiou de la pomme de terre. Notre étude est concentrée sur un seul modèle de champignon mais le but est d'étendre par la suite les conclusions obtenues à d'autres espèces. L'espèce ici considérée est le champignon *Botrytis cinerea*, récemment classé par la communauté des phytopathologistes parmi les pathogènes les plus importants scientifiquement et économiquement DEAN et collab. [2012].

1.1 Modèle biologique étudié : le champignon pythopathogène *Botrytis cinerea*

Botrytis cinerea (Botrytis c.) est une espèce de champignons phytopathogène de la classe des Ascomycètes GROVES et LOVELAND [1953]. L'étymologie de son nom fait référence à sa morphologie : « *Botrytis* » signifie « en forme de grappe ». Cette morphologie correspond à celle des conidiophores, organes contenants les spores ou conidies. Le nom « cinerea » renvoie à la couleur gris-cendrée de la sporulation. *Botrytis cinerea* est responsable de la pourriture noble du raisin mais également de la « pourriture grise » de nombreuses plantes LEROUX [2007]. Du fait des énormes dégâts qu'il engendre en agriculture, *Botrytis cinerea* fait figure de référence parmi les champignons pathogènes des plantes. En s'attaquant aux cultures d'intérêt agronomique, il entraîne environ 20% des pertes de cultures dans le monde. Ce champignon polyphage a la capacité d'attaquer et d'infecter toutes les parties aériennes de plus de 200 hôtes différents GOVRIN et LEVINE [2000]; WILLIAMSON et collab. [2007]. Cette maladie sévit dans toutes les zones de production du monde, sur diverses cultures d'intérêt agronomique majeur, sous serre ou en plein champ, telles que la vigne, le tournesol, la tomate, la fraise, la courgette en production légumière ou des plantes ornementales comme la rose. Un exemple de cette large gamme d'hôtes est présenté dans la figure 1.1. Au cours de son cycle biologique, *Botrytis cinerea* peut produire du mycélium, des spores asexuées (ou conidies), des spores sexuées ainsi que des sclérotes. Son cycle asexué est décrit dans le paragraphe A.2. Parmi les conditions favorables au développement de ce champigon sont retrouvées une humidité relativement importante (entre 80 et 90 % HR) et des températures comprises entre

2. Séquences temporelles d'images 2D

17 et 23 °C [WANG et collab., 1986; WILCOX et SEEM, 1994]. *Botrytis cinerea* est un champignon présentant une forte diversité phénotypique et génétique. Cette diversité est liée à ses exigences nutritives ou encore à son niveau d'agressivité. Sa morphologie est également influencée par son milieu de culture et par sa sensibilité aux fongicides (voir paragraphe 1.2). *Botrytis cinerea* cause une macération, accompagnée d'une sporulation gris-cendrée sur tiges, feuilles, fleurs et/ou fruits, pouvant aller jusqu'à la perte totale de la récolte. C'est un champignon nécrotrophe qui se nourrit de tissus morts digérés par un arsenal d'enzymes préalablement produites et sécrétées VAN KAN [2006]. En effet, cette batterie d'enzymes GONZÁLEZ-FERNÁNDEZ et collab. [2015] lui permet de dégrader la matière, de l'absorber et de l'utiliser comme source de nutriments pour son développement fongique. Son large spectre d'hôtes fait de ce phytopathogène une menace économique importante avec des pertes de 10 à 100 milliards d'euros par an à l'échelle mondiale. Rien que sur les vignes, la perte mondiale est estimée à 2 milliards de dollars par an ELMER et MICHAILIDES [2007]. Il est considéré comme un problème phytosanitaire majeur en viticulture dans le monde MARTINEZ et collab. [2005].
Botrytis cinerea est facilement cultivable en laboratoire et son cycle infectieux (Annexe A.2) est reproductible sur différentes plantes. Le champignon peut être recueilli en milieu liquide sous forme de cellules rondes dite « spores ». Ces spores placées dans un milieu liquide contenant les nutriments nécessaires à leur croissance, vont pouvoir former une protrusion. Cette protrusion s'allonge de manière unidirectionnelle pour former un tube germinatif qui va grandir au cours du temps jusqu'à la formation d'un réseau matriciel appelé « mycélium » (voir la figure 1.2).

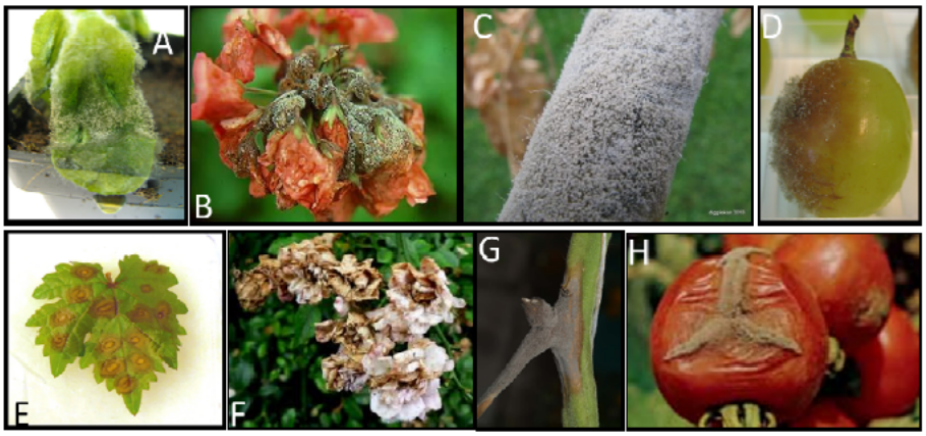


FIGURE 1.1 – La pourriture grise touche toute les parties du végétal parasité : A : Cotylédons de cornichons, B, F : Fleurs de géranium, C, G : Tiges de tomates, D : Baie de raisin, E : Feuille de vigne, H : Tomate.

La maladie fongique causée par *Botrytis c.* affecte donc la qualité et la quantité des récoltes. Afin de protéger les plantes, des molécules antifongiques appelées fongicides sont développées. Dans le paragraphe suivant 1.2, nous définissons ce qu'est un fongicide.

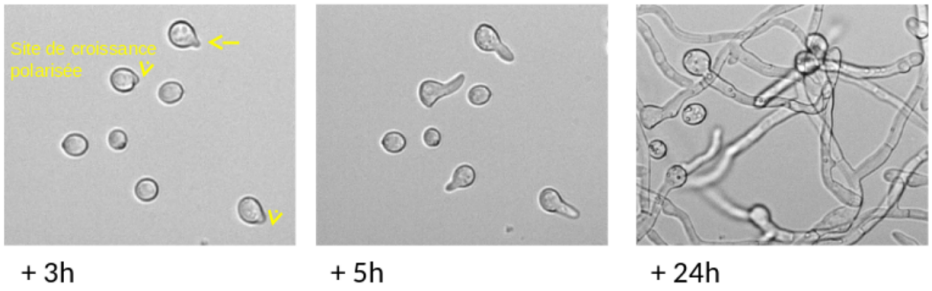


FIGURE 1.2 – Observation en microscopie de la croissance polarisée de *Botrytis cinerea* en milieu liquide : Au site de croissance polarisée, la spore du champignon va former une protrusion après 3h. Cette protrusion s'allonge de manière unidirectionnelle pour former un tube germinatif qui continue de croître pour former du mycélium en 24 heures.

1.2 Traitements fongicides anti-*Botrytis*

Dans le but de maintenir un bon niveau de production agricole, il est essentiel de mettre en place des moyens de contrôler les maladies causées par des pathogènes tels que *Botrytis cinerea*. Un fongicide est une substance capable d'éliminer ou limiter de façon exclusive le développement des champignons parasites des végétaux. Pour combattre la maladie, l'utilisation de fongicides entraîne des coûts financiers importants. Le marché mondial des produits de contrôle de *Botrytis cinerea* est estimé de 15 à 25 millions de dollars par an ELAD et STEWART [2007]. Les fongicides sont constitués de plusieurs matières actives ayant chacune des propriétés et des Modes d'Action (MoA) différents. Sur le marché actuel les fongicides sont classés selon leur Mode d'Action et sont répertoriés par le « Fungicide Resistance Action Committee » (FRAC, voir la figure 1.3). Ces Modes d'Action sont regroupés à ce jour en neuf classes ciblant : la synthèse des acides nucléiques, la mitose et la division cellulaire, la respiration, la synthèse des acides aminés et des protéines, la transduction du signal, la synthèse des lipides et des membranes, la synthèse des stérols, la synthèse des glucanes et de la chitine, et enfin la synthèse de la mélanine.

Néanmoins, malgré l'efficacité de cette stratégie, l'utilisation accrue de fongicides génère des problèmes de pollution et a fait apparaître de nouvelles souches pathogènes résistantes LEROUX [2007]. La résistance acquise aux fongicides est définie, par l'Organisation Européenne de Protection des Plantes (OEPP), comme une réduction stable de la sensibilité d'un champignon à un produit fongi-toxique WALKER [2013], résultant d'un changement génétique. Autrement dit, des souches résistantes sont apparues. Leur croissance et leur développement ne sont plus inhibés à une concentration donnée en fongicide, une concentration qui affecte en revanche fortement les souches restées sensibles. En effet, diverses études ont mis en évidence des niveaux de résistance aux fongicides dans les populations de *Botrytis cinerea* LIU et collab. [2019]. L'apparition de résistances aux fongicides conduit entre autres à des pertes d'efficacité de ces molécules et donc à la nécessité d'augmenter les doses appliquées. À terme, ces substances peuvent entraîner dans certains cas, des problèmes phytosanitaires et sont donc abandonnées. Les problèmes environnementaux liés à la protection des cultures sont de plus en plus préoccupants. Le contexte politique et économique actuel visant une agriculture durable a entraîné la mise en place de mesures telle que le plan Ecophyto 2025. Ce plan matérialise les engagements pris par les gouvernements. Ainsi, les législations européenne et française visent à réduire, d'ici 2025, de 50% l'usage des produits phytosanitaires jugés trop dangereux pour l'équilibre et la survie des écosystèmes (Grenelle de l'environne-

ment, France). D'autre part, les nouvelles souches résistantes sont actuellement contrôlées par l'alternance et l'utilisation raisonnée de fongicides. En effet, à l'heure actuelle, les fongicides restent des outils indispensables pour lutter efficacement contre *Botrytis cinerea* et assurer une production suffisante. L'ensemble de ces éléments conduit d'une part à rechercher de nouvelles solutions de protection mais également à l'identification et l'optimisation de nouvelles molécules antifongiques avec des nouveaux Modes d'Action. Les recherches effectuées au Centre de Recherche de la Dargoire sont focalisées sur ce deuxième point.

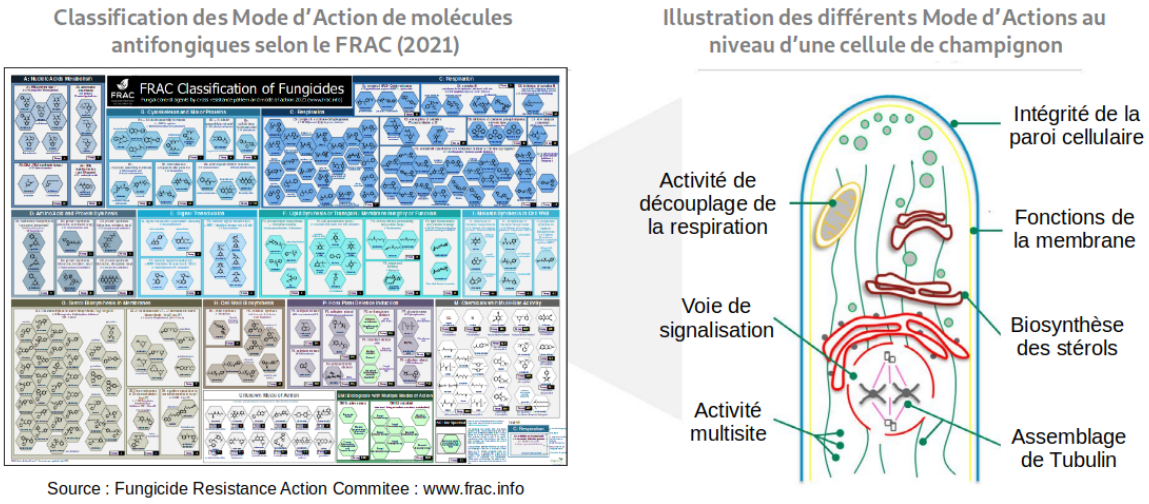


FIGURE 1.3 – Les molécules antifongiques sont classées selon leur Mode d'Action par le FRAC (Fungicide Resistance Action Committee). La figure est extraite de FRA. Le poster est présenté, en plus grand, dans le paragraphe A.3 de l'annexe. Par exemple, comme illustré sur le schéma d'une cellule de champignon, une molécule peut spécifiquement cibler une enzyme de la paroi, un complexe de la chaine respiratoire, le cytosquelette ou une autre cible de la cellule.

Il faut compter une dizaine d'années depuis la découverte d'une nouvelle molécule, les études sur champignons et sur plantes, les études en champs et les études toxicologiques avant l'obtention d'une homologation et d'une autorisation de mise sur le marché. Nos travaux sont axés sur la compréhension des mécanismes d'action des fongicides sur le champignon modèle. Dans le paragraphe suivant 1.3 nous décrivons les aspects de cette étude sur lesquels nous nous sommes focalisés ainsi que les techniques employées à cet effet.

1.3 Étude du Mode d'Action de nouvelles molécules

Pour rappel, les matières actives constituant les fongicides présentent des propriétés et des Modes d'Action (MoA) différents. Comprendre ces Modes d'Action permet d'élucider la façon dont ces composés actifs bloquent les activités cellulaires du champignon. De nombreuses méthodes existent pour cela. Comme décrit ci-avant (dans le paragraphe 0.1), au sein de l'équipe Excellence Biochimie, la microscopie est l'une des techniques utilisée pour identifier le mode d'action d'une molécule antifongique. Il s'agit de visualiser directement l'effet de celle-ci sur le développement du champignon, à des stades précoces (observations entre 0 et 48 heures de traitement généralement). Les études sont soit menées par microscopie à lumière transmise soit par microscopie à fluorescence sur des souches reportrices de voies métaboliques, de voies de signalisation ou d'organites via l'utilisation de protéines fluorescentes verte (GFP) activées sous l'action d'une enzyme (cas non traité au cours de ma thèse).

Choix de la microscopie à lumière transmise En microscopie à lumière transmise, les images sont le résultat de la lumière émanant d'une lampe halogène qui passe à travers l'échantillon. Dans le cas où les observations et le fond altèrent de manière différente la phase de la lumière, les « détails » des observations sont alors visibles. En effet, cela crée un contraste entre échantillons et fond de l'image. Si les images ne sont pas assez contrastées, une manière d'obtenir l'image d'un échantillon qui est peu visible est d'augmenter le contraste par la fixation et la coloration histologique des échantillons. Il faut cependant noter que le protocole de génération des images est complexifié par l'ajout des étapes de fixation et de coloration des échantillons. Une autre méthode est d'utiliser les optiques du microscope afin d'accentuer les petites différences dans la phase de la lumière causées par l'échantillon (exemple avec le contraste de phase). Toutefois, les cellules du champignon *B. cinerea* modifient suffisamment le passage de la lumière pour renseigner une bonne partie de l'image. Ainsi, les images utilisées dans le cadre de mes travaux de thèse sont donc des images en microscopies à lumière transmise.

Motivations du projet Dans certain cas, un traitement chimique donné (famille de molécules) peut entraîner une signature phénotypique caractéristique (voir la figure 1.4). L'étude du Mode d'Action de ces molécules antifongiques se fait sur des images et s'organise actuellement selon les étapes illustrées sur la figure 1.5. On cherche à reconnaître le phénotype présent à l'image afin de l'associer à la famille de molécules testée et si connu, à son Mode d'Action (voir la figure 1.6). À ce jour, la revue des images se fait manuellement, et compte-tenu du coût certain en ressources, il s'agit d'automatiser ce processus en développant une méthode d'analyse d'images robuste.

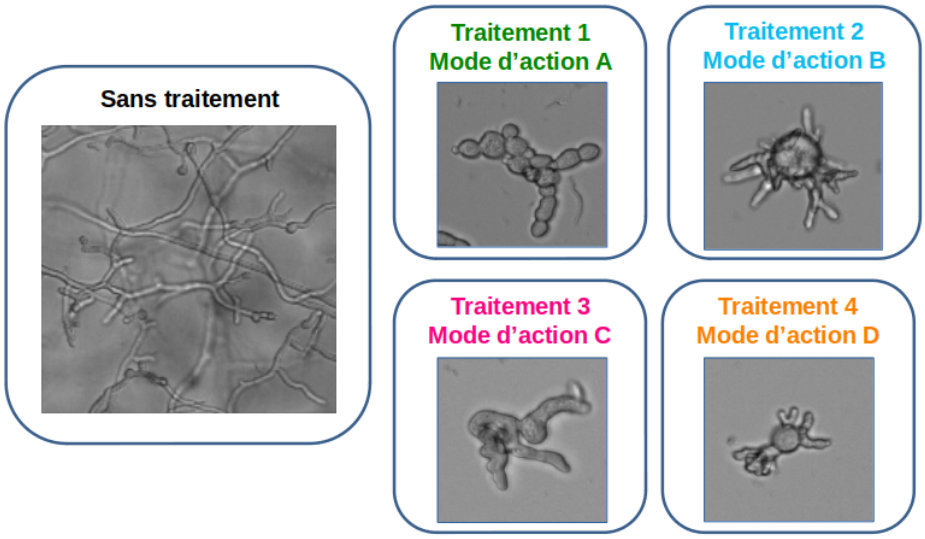


FIGURE 1.4 – Signature phénotypique caractéristique du traitement chimique utilisé : images en microscopie à lumière transmise, Microscope ImageXpress, Objectif x10.

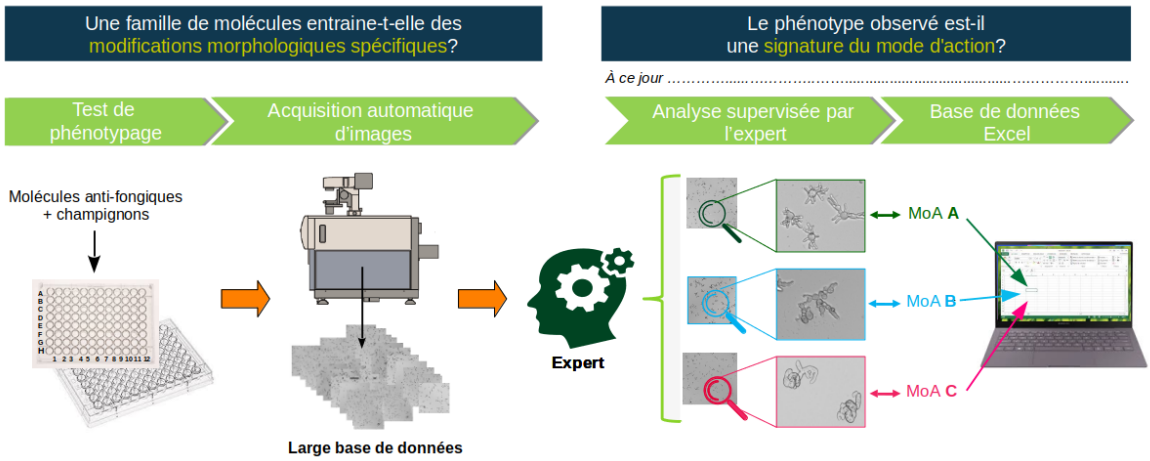


FIGURE 1.5 – Schéma simplifié de l'étude du Mode d'Action de molécules antifongiques par microscopie : les molécules sont testées en présence du champignon sur des microplaques. Un grand nombre d'images est généré via un microscope automatisé. Un expert procède à la reconnaissance des phénotypes sur les images et il annote ses conclusions à la main dans un tableau Excel.

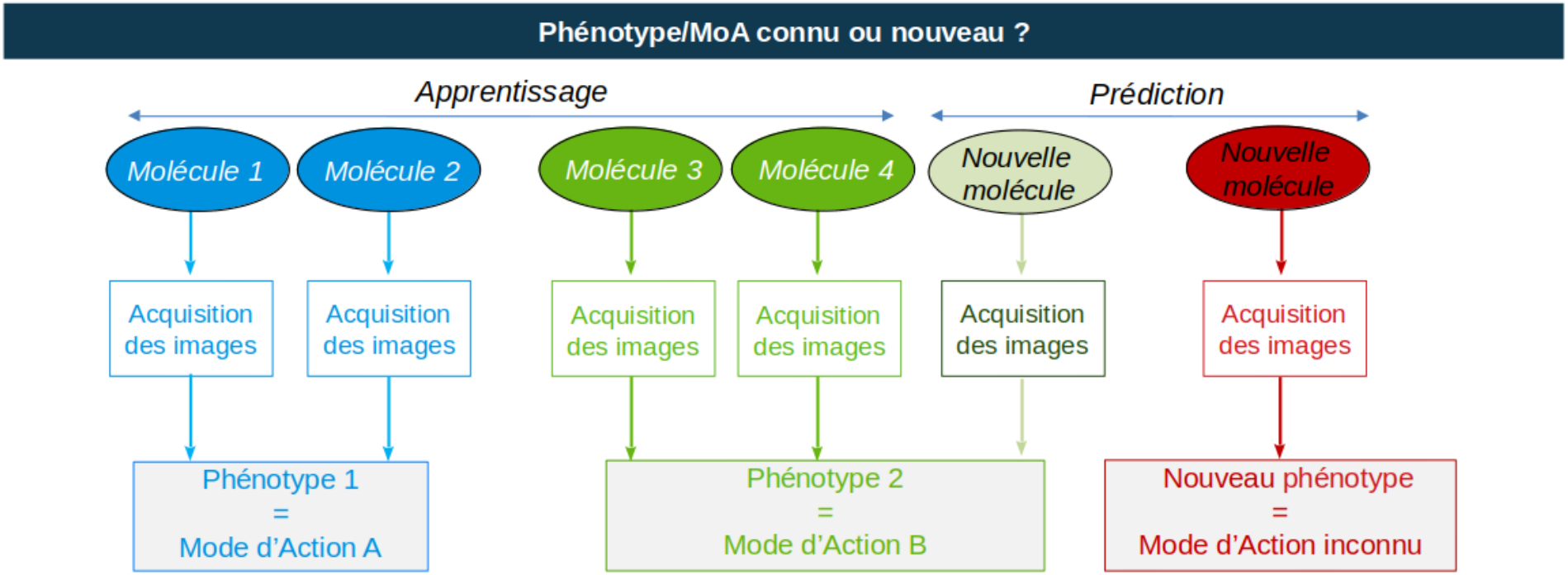


FIGURE 1.6 – Partie apprentissage : Les familles de molécules chimiques aux modes d'action connus sont testées et associées aux phénotypes de champignons obtenus. Partie prédiction : De nouvelles molécules sont testées et une hypothèse sur leurs modes d'action est obtenue en comparant les phénotypes observés à ceux de la base d'apprentissage. Un phénotype inconnu indique une molécule avec un nouveau mode d'action.

1.4 Contributions

Les contributions majeures de cette thèse sont :

— La caractérisation des différents phénotypes du champignon *Botrytis cinerea via* une analyse automatique des images en microscopie comprenant des étapes de traitement d'images et d'extraction de paramètres morphométriques.

— Le développement d'une nouvelle méthode de classification avec rejet établie dans un schéma générique de classification selon le type de classifieur qu'est le réseau de neurones. Dans un contexte supervisé, cette méthode suit une stratégie générale fondée sur trois étapes principales : apprentissage d'un modèle indépendamment pour chaque classe, apprentissage d'un seuil par modèle fondé sur les interactions entre classes, et procédure de prédiction s'appuyant sur les réponses des modèles par rapport à leur seuil. Cette approche est appliquée dans le cadre de notre problématique de reconnaissance de phénotypes connus et nouveaux de *Botrytis cinerea*.

— Mise en place d'un "système expert" correspondant à un protocole de détermination des Modes d'Action des molécules antifongiques. Une hypothèse de MoA est donnée sur la base de règles de prédiction appliquées aux images correspondant au test de molécules sur une gamme de plusieurs concentrations. Outre la conclusion sur le mécanisme d'action, cette procédure permet d'obtenir une analyse de la molécule testée, notamment des indications quant à son niveau d'activité (degré d'efficacité).

— Le développement d'une approche novatrice de classification fondée sur l'estimation de la fonction de densité de probabilité de la distribution d'une population. Cette approche s'appuie sur le domaine du transport optimal.

— La création de modèles de croissances calibrés à partir de données réelles. Puis, nous avons simulé la croissance des champignons suivant les traitements testés, pour des phénotypes donnés. Le modèle construit est un processus stochastique à temps discret utilisant des lois discrètes et continues pour piloter les différents événements (croissance, création d'une branche, ...) et leur ampleur.

Mes publications sont les suivantes :

— Sarah Laroui (et al.), Machine-Learning assisted phenotyping : from fungal morphology to Mode Of Action hypothesis., IUPAC International Congress 2019 (LAROUI et collab. [2019])

— Sarah Laroui (et al.), How to define a rejection class based on model learning? ICPR International Conference on Pattern Recognition 2020 (LAROUI et collab. [2021a])

1.5 Organisation du manuscrit

Cette thèse se découpe en cinq chapitres. Leur contenu est décrit ci-après :

1.5.1 Chapitre 1

Comme vu ci-dessus, ce premier chapitre est consacré à la description du modèle biologique étudié, le champignon pythopathogène des plantes *Botrytis cinerea* et à la définition d'un fongicide. Nous y exposons également la manière dont les experts au laboratoire élucident les Modes d'Actions des molécules antifongiques avant que la tâche ne soit en routine (automatisée).

1.5.2 Chapitre 2

Dans ce chapitre nous présentons toutes les clés nécessaires au développement de notre outil de reconnaissance des phénotypes connus de *Botrytis c.* sur des images en microcopie. Nous commençons par une introduction à l'analyse et au traitement d'images dans laquelle nous détaillons les outils de ce domaine dont nous nous sommes servi au cours de ce projet. Nous nous intéressons également à l'un des champs d'étude de l'intelligence artificielle, l'apprentissage machine (ou Machine Learning en anglais). Des méthodes très connues, dont certaines appliquées à notre étude, y seront détaillées. Nous y abordons également l'origine et la nature de nos données et décrivons les protocoles d'expérimentation biologique ainsi que les conditions d'acquisition de nos images par microscopie. Enfin, les résultats obtenus par le biais de deux méthodes de classification appliquées à nos données sont détaillés, analysés et discutés. Une conclusion est émise quant à la méthode retenue pour nos études.

1.5.3 Chapitre 3

La solution que nous visions à développer doit pouvoir alerter l'utilisateur lors de l'identification d'un nouveau phénotype, c'est-à-dire ne faisant pas partie de l'ensemble d'apprentissage. Or, en classification supervisée, les échantillons à classer ne peuvent être prédits que comme appartenant à l'une des classes sur lesquelles le classifieur a été entraîné. Ce chapitre est dédié à la description de notre méthode de classification avec une classe dite de rejet. En premier lieu, une synthèse des méthodes existantes et proposant une classe de rejet dans la littérature est réalisée. Les résultats obtenus avec l'une d'elles sont comparés à ceux obtenus avec notre méthode. Ce travail a fait l'objet d'une publication dans la conférence internationale ICPR2020 LAROUI et collab. [2021b] (événement qui a finalement eu lieu en 2021 à cause du covid). Nous décrivons en détail l'application transmise à l'entreprise qui y est à ce jour utilisé par les techniciens et les chercheurs.

1.5.4 Chapitre 4

Une approche alternative, utilisant le transport optimal (TO) est décrite, testée et comparée. Cette approche de classification est fondée sur l'estimation de la fonction de densité de probabilité de la distribution d'une population. L'une des motivations de cette approche est de s'affranchir de certaines contraintes liées à la première méthode. En effet, l'abscence de connaissance a priori des distributions des données limite l'utilisation de modèles paramétriques qui peuvent être problématiques si les données ne suivent pas la distribution supposée. La différence majeure séparant ces deux méthodes est le type de modèles appris. La complexité du modèle est "transférée" dans le transport (vers un modèle simple).

1.5.5 Chapitre 5

Il s'agit dans ce chapitre d'observer et de caractériser la croissance du champignon au cours du temps. Nous voulons caractériser les différents phénotypes ainsi que l'effet des molécules correspondantes et proposer un modèle discret de croissance. L'idée est de mesurer des caractéristiques à chaque temps et de calibrer les paramètres du modèle en fonction de ces mesures. Le but est de comprendre et de prédire l'apparition des phénotypes au cours du temps ainsi que de comparer l'efficacité de plusieurs molécules ayant le même mécanisme d'action. Pour cela, un protocole a été mis en place impliquant une

procédure biologique particulière et la génération de séquences d'images temporelles (ou TimeLapse) en microscopie. Ces TimeLapses sont générés avec une concentration de spores et un intervalle de temps donnés sur une sélection de phénotypes.

1.6 Références

«Artificial intelligence vs machine learning vs deep learning», https://medium.datadriveninvestor.com/artificial-intelligence-vs-machine-learning-vs-deep-learning Accessed : 2020-03-27. 4

«FRAC classification of fungicides», https://www.frac.info/docs/default-source/public_ations/frac-mode-of-action-poster/frac-moa-poster-2021.pdf?sfvrsn=a6f6499a_2. Accessed : 2021. x, 8

DEAN, R., J. A. VAN KAN, Z. A. PRETORIUS, K. E. HAMMOND-KOSACK, A. DI PIETRO, P. D. SPANU, J. J. RUDD, M. DICKMAN, R. KAHMANN, J. ELLIS et collab.. 2012, «The top 10 fungal pathogens in molecular plant pathology», *Molecular plant pathology*, vol. 13, nº 4, p. 414–430. 5

ELAD, Y. et A. STEWART. 2007, «Microbial control of botrytis spp», *Botrytis : biology, pathology and control*, p. 223–241. 7

ELMER, P. A. et T. J. MICHAILIDES. 2007, «Epidemiology of botrytis cinerea in orchard and vine crops», dans *Botrytis : biology, pathology and control*, Springer, p. 243–272. 6

GONZÁLEZ-FERNÁNDEZ, R., J. VALERO-GALVÁN, F. J. GÓMEZ-GÁLVEZ et J. V. JORRÍN-NOVO. 2015, «Unraveling the in vitro secretome of the phytopathogen botrytis cinerea to understand the interaction with its hosts», *Frontiers in plant science*, vol. 6, p. 839. 6

GOVRIN, E. M. et A. LEVINE. 2000, «The hypersensitive response facilitates plant infection by the necrotrophic pathogen botrytis cinerea», *Current biology*, vol. 10, nº 13, p. 751–757. 5

GROVES, J. W. et C. A. LOVELAND. 1953, «The connection between botryotinia fuckeliana and botrytis cinerea», *Mycologia*, vol. 45, nº 3, p. 415–425. 5

VAN KAN, J. A. 2006, «Licensed to kill : the lifestyle of a necrotrophic plant pathogen», *Trends in plant science*, vol. 11, nº 5, p. 247–253. 6

KOECK, M., A. R. HARDHAM et P. N. DODDS. 2011, «The role of effectors of biotrophic and hemibiotrophic fungi in infection», *Cellular microbiology*, vol. 13, nº 12, p. 1849–1857. 4

LAROUI, S. *Méthode de classification automatique de phénotypes de cellules de champignons pythopathogènes : rapport de stage de Master 2. Master2 Biologie, Informatique et Mathématiques. Nice Sophia Antipolis : Université de Nice Sophia Antipolis, 2017, 54 p.* 2

LAROUI, S., E. DEBREUVE, X. DESCOMBES, F. VILLALBA, F. VILLIERS et A. VERNAY. 2019, «Machine-learning assisted phenotyping : From fungal morphology to mode of action hypothesis», dans *IUPAC - 14th International Congress of Crop Protection Chemistry*, Ghent, Belgium. URL https://hal.archives-ouvertes.fr/hal-02403936. 12

LAROUI, S., X. DESCOMBES, A. VERNAY, F. VILLIERS, F. VILLALBA et E. DEBREUVE. 2021a, «How to define a rejection class based on model learning?», dans *ICPR2020 - International Conference on Pattern Recognition*, Milan / Virtual, Italy. URL https://hal.archives-ouvertes.fr/hal-02963115. 12

LAROUI, S., X. DESCOMBES, A. VERNAY, F. VILLIERS, F. VILLALBA et E. DEBREUVE. 2021b, «How to define a rejection class based on model learning?», dans *ICPR2020-International Conference on Pattern Recognition*. 13

LEROUX, P. 2007, «Chemical control of botrytis and its resistance to chemical fungicides», dans *Botrytis : Biology, pathology and control*, Springer, p. 195–222. 5, 7

LIU, S., L. FU, J. CHEN, S. WANG, J. JIANG, Y. ZHANG, Z. CHE, Y. TIAN et G. CHEN. 2019, «Baseline sensitivity of botrytis cinerea to fluazinam and cross-resistance», *Journal of Phytopathology*, vol. 167, nº 6, p. 344–350. 7

MARTINEZ, F., B. DUBOS et M. FERMAUD. 2005, «The role of saprotrophy and virulence in the population dynamics of botrytis cinerea in vineyards», *Phytopathology*, vol. 95, nº 6, p. 692–700. 6

MORPHEME. «Equipe-projet MORPHEME : Morphologie et images», https://www.inria.fr/fr/morpheme. Accessed : 2020-03-27. 3

WALKER, A. S. 2013, *Diversité et adaptation aux fongicides des populations de Botrytis cinerea, agent de la pourriture grise*, thèse de doctorat, Université Louis Pasteur (Strasbourg 1). 7

WANG, Z.-N., J. COLEY-SMITH et P. WAREING. 1986, «Dicarboximide resistance in botrytis cinerea in protected lettuce», *Plant Pathology*, vol. 35, nº 4, p. 427–433. 6

WILCOX, W. et R. SEEM. 1994, «Relationship between strawberry gray mold incidence, environmental variables, and fungicide applications during different periods of the fruiting season», *Phytopathology*, vol. 84, nº 3, p. 264–270. 6

WILLIAMSON, B., B. TUDZYNSKI, P. TUDZYNSKI et J. A. VAN KAN. 2007, «Botrytis cinerea : the cause of grey mould disease», *Molecular plant pathology*, vol. 8, nº 5, p. 561–580. 5

Chapitre 2

Classification de phénotypes de *Botrytis cinerea*

Dans ce chapitre, l'objectif du travail effectué est de définir la méthode de classification la plus adaptée à notre problématique. Le choix de cette méthode de classification de phénotypes de *Botrytis cinerea* repose sur la comparaison de méthodes distinctes qui présentent des approches différentes. Les éléments nécessaires à ces approches sont décrits dans plusieurs paragraphes. Le type de données correspondant à des images, un premier paragraphe est dédié au domaine du traitement d'images dans lequel nous définissons ce qu'est une image numérique et comment la manipuler à l'aide d'outils mathématiques. Puis la notion de classification est introduite dans une seconde section avec notamment la description de plusieurs méthodes très populaires dans la littérature. Enfin, une dernière section aborde l'origine et la nature de nos données, avec la description des protocoles d'expérimentation biologique, d'acquisition de nos images par microscopie et d'extraction des descripteurs nécessaires à la classification. Pour finir, les résultats obtenus avec deux méthodes de classification sont comparés et une conclusion sur la méthode retenue pour la suite du projet est émise et justifiée.

2.1 Quelques outils de traitement d'image

Le traitement d'image, discipline de l'informatique et des mathématiques appliquées, est un ensemble de méthodes et techniques appliquées aux images permettant d'en améliorer l'aspect visuel et d'en extraire des informations pertinentes.

2.1.1 Définition d'une image

Dans le but de comprendre comment le traitement d'image fonctionne, il nous faut définir ce qu'est une image numérique. Un capteur enregistre dans un premier temps la lumière réfléchie par les objets de la scène observée puis la convertit en tension électrique afin que l'image soit décrite par une suite de données binaires, c'est-à-dire, une suite de 0 et de 1. Enfin, ces données sont transférées sur une carte mémoire et codées dans le but d'obtenir un tableau de valeurs affectant à chaque cellule appelée pixel une valeur représentant une information lumineuse. Une image numérique est donc une image dont le domaine spatial et la luminosité ont été discrétisés et qui peut être assimilée à une fonction de deux variables u(x, y). Ainsi, le traitement d'image pourra alors appliquer des transformations mathématiques sur des matrices permettant d'aller vers une interprétation de celles-ci.

Topologie d'une image : voisinages

Un objet est dit connexe s'il est constitué d'une seule structure ou groupe de pixels qui sont connexes (adjacents). Sur des images 2D, deux types de connexité sont généralement utilisés : la connexité d'ordre 4 et la connexité d'ordre 8, c'est à dire le regroupement du pixel central avec 4 ou 8 de ces voisins (voir la figure 2.1). En traitement d'images, la notion de voisinage est centrale pour la définition d'un objet.

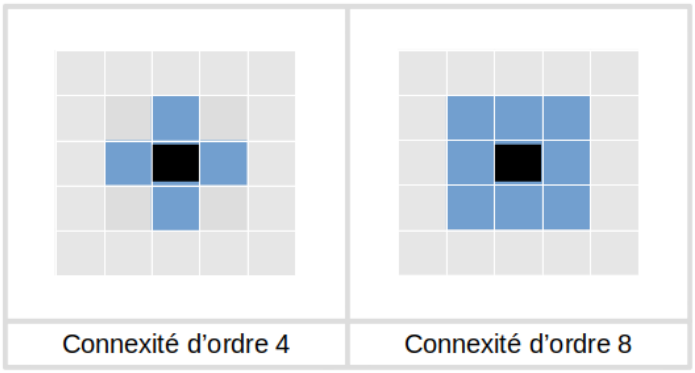


FIGURE 2.1 – Possibilités de connexité pour une image en 2D.

2.1.2 Amélioration de la qualité des images

Les images acquises nécessitent souvent des traitements afin d'améliorer leur qualité car de cette qualité découlera l'exactitude des informations extraites. Le bruit correspond à toute fluctuation parasite de l'image. Ce bruit peut provenir d'une dégradation de l'image due à la qualité des capteurs utilisés lors de l'acquisition des images, la nature des environnements ou les caractéristiques du signal. Dans ce cas, il s'agit de filtrer l'image initiale pour limiter ces informations parasites qui influencent son analyse. Ce bruit s'il est considéré au sens large du terme, peut également correspondre à des petits agrégats résiduels résultants de la binarisation objets/fond d'une image.
De plus, les masques des objets détectés peuvent également présenter certaines irrégularités (trous, ..). Pour pallier cela, nous utilisons le traitement d'images. On retrouve ainsi dans cette discipline différentes méthodes d'amélioration de l'image comme le rehaussement de contraste (Égalisation d'histogramme et masque flou ou "Unsharp masking" (USM) en anglais) et des techniques d'atténuation de bruits comme le filtrage LUISIER et collab. [2009] et la déconvolution TRISTAN-VEGA et collab. [2012], permettant de réduire les artefacts susceptibles de nuire à la description de l'information pertinente.

2.1.3 Morphologie mathématique

Développée par J.Serra et G.Matheron dans les années 60 en France HARALICK et collab. [1987]; MATHERON et SERRA [2002]; SERRA [1968], la morphologie mathématique est une méthodologie permettant d'avoir accès à un nombre important d'outils précieux pour le traitement des images. Contrairement au traitement linéaire des images qui s'appuie sur le traitement du signal, la morphologie mathématique est une approche qui repose sur des concepts ensemblistes. En effet, cette approche se concentre sur la structure géométrique présente dans l'image, la considérant comme un ensemble d'objets géométriques pouvant être manipulés selon la théorie des ensembles. On retrouve la morphologie mathématique dans de nombreux aspects du traitement d'image MARAGOS [1987]

comme la détection de contours CHANDA et collab. [1998]; SONG et NEUVO [1993], la segmentation GAUCH [1999]; SALEMBIER [1994] et l'amélioration des images MARAGOS [2005]. Les méthodes reposant sur la morphologie mathématique peuvent s'appliquer tant aux images binaires qu'à celles en niveaux de gris. Dans ce paragraphe les outils seront présentés dans le cas d'images binaires (cas dans lequel la morphologie mathématique est appliquée dans ce projet).

La morphologie mathématique correspond à des opérations logiques locales sur des ensembles de pixels qui reposent sur un élément structurant. Cet élément structurant (un ensemble de géométrie, de forme et de taille connue) va servir à sonder l'image binaire en se positionnant tour à tour sur chaque pixel de l'objet *via* une opération de translation. Une fois l'élément structurant centré en un point de référence, il va définir le type de voisinage que l'on souhaite considérer. Ces opérations sont répétées en balayant toute l'image. Les opérations de base sont la dilatation et l'érosion. L'érosion est relative à l'inclusion et la dilation à un test d'intersection (voir les figures 2.2, 2.3 et 2.4).

— L'érosion amincit la surface d'un composant. L'ensemble érodé correspond aux parties dans lesquelles l'élément structurant est inclus.

— Dans le cas de la dilation, permettant d'augmenter la surface d'un composant, les parties de l'ensemble qui intersectent l'élément structurant forment l'ensemble dilaté.

De leur combinaison successive, existent deux nouvelles opérations :

— L'ouverture : Pour un élément structurant e, l'image d'origine est d'abord érodée par e puis dilatée par e . L'ouverture de A par e (voir la figure 2.2) est obtenue en prenant l'union de tous les translatés de e entièrement inclus dans A. Une ouverture va faire disparaître les petites structures et détails fins lors de la phase d'érosion et ils ne pourront donc pas être recréés lors de l'étape suivante de dilatation. Cette opération est également intéressante puisqu'elle permet de lisser les contours.

— La fermeture de A par e (voir la figure 2.2) s'effectue *via* une dilatation de A suivie d'une érosion par e. L'intérêt d'une fermeture est de boucher les petits trous.

Dans le cas d'une image binaire bruitée, le bruit correspond à des petits agrégats. Certains pixels de l'image appartenant à l'objet d'intérêt son attribués au fond, ces pixels correspondent à des faux négatifs, et les pixels dans la situation contraire sont des faux positifs. Une ouverture avec un élément structurant suffisamment grand filtrera les faux positifs, tandis qu'une fermeture filtrera les faux négatifs. Ces deux opérateurs sont appliqués de façon successive pour filtrer le bruit (filtre séquentiel alterné).

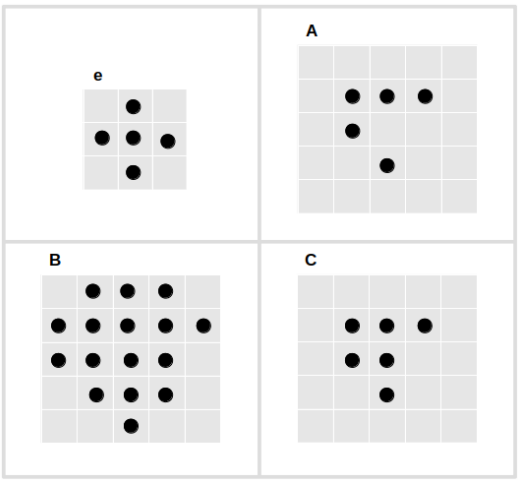


FIGURE 2.2 – La dilatation de l'image A par l'élément structurant e donne l'image B. L'érosion de l'image B par e donne l'image C.

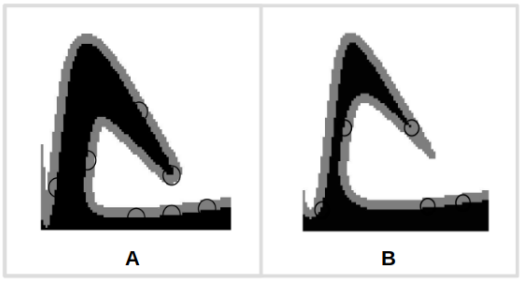


FIGURE 2.3 – Principe de la dilatation (A) et de l'érosion (B) : élément structurant symétrique (disque) centré en l'origine.

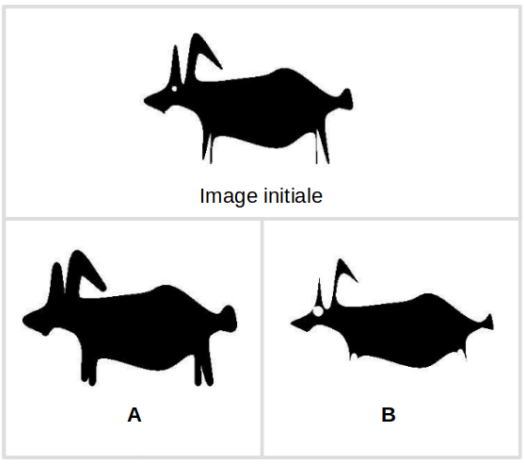


FIGURE 2.4 – Exemple de résultats de dilatation (A) et d'érosion (B) sur des images synthétiques : élément structurant symétrique centré en l'origine.

2.1.4 Segmentation

La segmentation d'images consiste en un partitionnement de l'image en régions ensembles de pixels homogènes. Cette partition, très fréquentes en analyse d'image sert de base à la classification des régions de l'image. Si le nombre de types de région est égal à deux, cette opération est appelée binarisation BLAYVAS et collab. [2006], c'est le cas lorsqu'on cherche par exemple à séparer les objets du fond de l'image. Plusieurs méthodes de segmentation existent ANJNA et ER [2017] et diffèrent selon que la segmentation est fondée sur les régions, les contours, le seuillage des pixels en fonction de leur valeur ou la coopération des trois. Le seuillage consiste à affecter la valeur 255 (ou 1) aux pixels dont la valeur est supérieure à un seuil S et 0 aux autres. On retrouve également la segmentation par couleurs mais aussi la segmentation par textures lorsque le contenu fréquentiel de l'image est utilisé. Un exemple connu de segmentation fondée sur les régions est la segmentation par ligne de partage des eaux qui est une méthode issue de la morphologie mathématique. On la retrouve notamment dans le domaine médical. Dans GRAU et collab. [2004] , l'utilisation de cet algorithme a deux applications : la segmentation d'images du cartilage du genou et celle d'images de cerveaux permettant la séparation de la matière blanche et de la matière grise. Dans AKHTAR et collab., il s'agit de segmenter des images couleurs de feuilles d'arbre afin d'en identifier l'espèce.

La détection des contours est l'une des composantes les plus importantes en traitement d'image, de la vision par ordinateur et de la vision industrielle BENSON et collab. [2003]; JIA [2010]; SELVAKUMAR et HARIGANESH [2016]. Parmi les méthodes de segmentation par contours, on peut citer les approches par convolution. Pour rappel, la convolution est une opération mathématique qui prend deux signaux s_1 et s_2 et qui renvoie un signal s, tel que : $s = s_1 * s_2$
Le plus souvent, en détection de contour, ces méthodes reposent sur la dérivation. En effet, la dérivée permet l'étude de la variation d'un signal. Ainsi, la dérivée permet la détection des variations d'intensité dans une image et les contours correspondent à des maxima locaux de la dérivée première. En 2D, la dérivée première correspond au gradient (2.1) dont les composantes du vecteur sont la dérivée partielle par rapport à x et y de l'image.

$$\nabla \mathrm{I} = \left[\frac{\partial \mathrm{I}}{\partial x}, \frac{\partial \mathrm{I}}{\partial y} \right] \tag{2.1}$$

Parmi les détecteurs de contours très utilisés en traitement d'image on retrouve le détecteur de Sobel proposé en 1970 JAIN et collab. [1995] qui utilise deux masques de convolution pour approximer numériquement ces dérivées par rapport à x (2.2) et y (2.3).

$$\frac{\partial \mathrm{I}}{\partial x} = \begin{pmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{pmatrix} \tag{2.2}$$

$$\frac{\partial \mathrm{I}}{\partial y} = \begin{pmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{pmatrix} \tag{2.3}$$

La norme du contour est obtenu en calculant le module du gradient pour chaque pixel de l'image. Le filtre de Canny CANNY [1986] suit plutôt une approche de segmentation par filtrage optimal. La différence avec les méthodes classiques est la non multiplicité des

maxima locaux en considérant un filtre à réponse impulsionnelle finie. L'auteur l'a conçu pour être optimal suivant trois critères clairement explicités :

— bonne détection (rapport signal sur bruit)

— bonne localisation, c'est-à-dire que les points détectés doivent être aussi proches que possible des points de contours réels.

— réponse unique, autrement dit, un point du contour ne doit être détecté qu'une seule fois par le filtre

L'approche de Deriche DERICHE [1990] a été de développer un filtre optimal à réponse impulsionnelle infinie sous forme d'un filtre récursif, permettant la détection de ces contours.

2.2 Apprentissage automatique (ou "Machine Learning")

La révolution numérique à laquelle nous assistons ces dernières années a radicalement changé la façon dont nous générons et consommons des données. Cette constante expansion de la quantité de données touche différents secteurs de la société, notamment la santé, l'agriculture ou encore le divertissement. L'intérêt du développement de nouveaux outils permettant des interprétations automatiques et robustes de ces données est donc évident. Les algorithmes de "Machine Learning" nous permettent de faire cela de façon plus ou moins simple, rapide et robuste. En effet, le but du "Machine Learning" est d'apprendre à l'algorithme ce qu'on veut qu'il fasse et qu'une fois entraîné, il soit capable d'effectuer sa tâche de façon automatique. Les algorithmes d'apprentissage automatique ont été appliqués à divers domaines, tels que le traitement du langage, la reconnaissance d'écriture manuscrite, la robotique, la fouille de données, les moteurs de recherche sur Internet, le diagnostic médical et la bioinformatique. Les techniques d'apprentissage jouent un rôle crucial dans des applications qui vont de la mise au point de traitements médicamenteux à l'analyse de grands réseaux de télécommunication. Bien que les systèmes développés soient très loin d'égaler les performances de l'œil et du cerveau humain, des avancées considérables sont effectuées au fil des années et permettent déjà d'effectuer de nombreuses tâches.

Certaines solutions de "Machine Learning" sont décrites, de façon très générale, ci-dessous et sont regroupées par types d'apprentissage DEY [2016]. Avant de décrire ces méthodes, quelques précisions :

En classification automatique, deux types de modèles statistiques existent, le modèle génératif (naïve bayésienne, modèles de Markov cachés et de mélange Gaussien, auto-encodeurs, ...) et le modèle discriminatif (régression logistique, arbres de décision, séparateurs à vaste marge, ...). Le modèle génératif apprend la distribution de probabilité conjointe $P(X,Y)$ et calculent leurs prédictions $P(Y|X)$ en utilisant les règles de Bayes. Le modèle discriminatif cherche à décrire directement la distribution de probabilité conditionnelle $P(Y|X)$. X et Y étant respectivement les variables donnée et classe et la formule de Bayes étant définie par :

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)} \tag{2.4}$$

où A et B sont deux évènements et $P(A)$ et $P(B)$ sont leurs probabilités. $P(A|B)$ correspond à la probabilité que l'évènement A se réalise sachant que l'évènement B s'est réalisé et

P(B|A) correspond à la probabilité que l'évènement B se réalise sachant que l'évènement A s'est réalisé.

D'autre part, étant donné que le travail effectué au cours de cette thèse se situe dans le cadre de l'apprentissage supervisé, les algorithmes des arbres de décision et de séparateurs à vaste marge sont décrits dans le paragraphe 2.2.2. Ce paragraphe est en effet destiné a aborder la classification supervisée de façon plus détailée.

1. Apprentissage supervisé (ou "supervised learning") : On parle d'apprentissage supervisé dans le cas où les classes sont prédéterminées et les exemples connus, le système apprend à classer selon un modèle de classement. Les classes sont définies par un expert et permettent de calculer un indicateur de performance du système de classification (taux de mal classé) en les comparant aux prédictions obtenues. On cherche alors à trouver ou du moins, approximer au mieux, la fonction qui permet d'affecter la bonne classe à ces échantillons.

   - **Naïve bayésienne** (ou "Naive Bayes") RISH et collab. [2001] Il s'agit d'un modèle statistique génératif, qui a pour but d'obtenir la probabilité conditionnelle P(Y|X) (X et Y étant respectivement les variables donnée et classe). L'a priori P(Y) et la vraisemblance P(X|Y) sont estimés à partir des données d'apprentissage en utilisant la formule de Bayes 2.4 pour calculer P(Y|X). L'algorithme cherche à décrire les classes et à déduire, pour chaque échantillon, la probabilité d'appartenance à la classe correspondante. La particularité du modèle Naive Bayes est qu'il simplifie le problème en imposant une forte indépendance (dite naïve) des hypothèses. En effet, il suppose que l'existence d'une caractéristique pour une classe est indépendante de l'existence d'autres caractéristiques. Le modèle est couplé à une règle de décision. Celle-ci peut par exemple correspondre à l'hypothèse la plus probable, c'est-à-dire à la règle du maximum à posteriori. La distribution est définie comme :

$$P(Y = 1|X) = \frac{P(y = 1).P(x_1|y = 1)\dots P(x_n|y = 1)}{P(X)} \tag{2.5}$$

   - **Arbres de décision** (ou "Decision tree") Voir le paragraphe 2.2.2.
   - **Séparateurs à vaste marge** (ou "Support vector machine") Voir le paragraphe 2.2.2.

2. Apprentissage non-supervisé (ou "unsupervised learning") : On parle d'apprentissage non-supervisé (ou clustering) dans le cas où les classes des données ainsi que le nombre de classe et leur nature sont inconnues. L'algorithme doit découvrir par lui-même la structure plus ou moins cachée des données. La méthode consiste à répartir un ensemble d'individus en groupes homogènes selon des critères de classification. Ces critères regroupent les échantillons s'ils sont similaires et les séparent dans le cas contraire.

   - **Analyse en composantes principales** (ou "Principal component analysis") WOLD et collab. [1987] : C'est une méthode statistique qui permet d'extraire et de visualiser les informations importantes contenues dans des données dites multivariées (données avec plusieurs variables). L'ACP identifie les directions, également appelées axes principaux ou composantes principales, le long desquelles la variation des données est maximale. Les informations importantes sont synthétisées en nouvelles variables (composantes principales) qui correspondent à des combinaisons linéaires des variables originelles. On cherche ici à réduire la dimension tout en conservant un maximum d'informations.

- **K-moyennes** (ou "K-means") KANUNGO et collab. [2002] : Le partitionnement en k-moyennes est un problème d'optimisation combinatoire qui cherche à diviser les points en k ensembles (avec $k \leq n$) en minimisant la distance entre les points à l'intérieur de chaque groupe.

3. Apprentissage semi-supervisé (ou "semi-supervised learning") : Effectué de manière probabiliste ou non, il vise à faire apparaître la distribution sous-jacente des échantillons dans leur espace de description. Ce type d'apprentissage est mis en œuvre lorsque des données ou la classe de certaines données manquent.

   - **L'auto-apprentissage** (ou "self-training") TRIGUERO et collab. [2015] Cette méthode consiste à apprendre un modèle sur les données d'apprentissage étiquetées puis à prédire les classes des données non-étiquetées. Les prédictions avec un haut degré de confiance sont ajoutées aux données d'apprentissage et le modèle est ré-entrainé et la procédure est répétée jusqu'à satisfaire un critère d'arrêt.
   - **Séparateur Semi-Supervisé à Vaste Marge** (S3VM) BENNETT et collab. [1999] Dans cette approche, deux contraintes sont ajoutées au problème quadratique des SVM. Ces contraintes sont définies pour maintenir les données non-étiquetées à l'extérieur de la marge tout en minimisant l'erreur de classification.
   - **SVM Vecteur de support transductif** (TSVM) JOACHIMS et collab. [1999] Il s'agit également d'une extension des SVM. L'approche consiste à étiqueter les données non-étiquetées de sorte que la marge soit maximum entre les données étiquetées et non étiquetées.

4. Apprentissage par renforcement (ou "reinforcement learning") : L'algorithme apprend un comportement sachant une observation. Chacune de ces actions sur l'environnement produit une valeur de retour qui guide l'algorithme.
5. Apprentissage ensembliste (ou "ensemble learning") YANG et collab. [2010]
   Le principe de ce type d'apprentissage est d'utiliser plusieurs algorithmes d'apprentissage dans le but d'obtenir de meilleurs résultats de prédictions.

   - **Bagging** : L'ensemble d'entraînement est scindé en sous-ensembles par échantillonnage uniforme avec remise. Plusieurs modèles sont entraînés sur ces sous-ensembles et la prédiction finale est obtenue en effectuant la moyenne ou par vote majoritaire des prédictions des modèles.
   - **Boosting** : Cette technique consiste à combiner un grand nombre d'algorithmes avec de faibles performances individuelles pour en générer un qui soit beaucoup plus efficace. Un des algorithmes les plus utilisés en boosting s'appelle AdaBoost (introduit en 1996 par Yoav Freund et Rob Shapire). Chaque classifieur est entraîné pour corriger les erreurs des classifieurs précédents. L'algorithme fonctionne en donnant plus d'importance aux observations difficiles à prédire et le poids de ces éléments mal classés va augmenter à chaque itération. La combinaison linéaire des classifieurs construits, au fur et à mesure, donne le classifieur final.

6. Réseau de neurones (ou "neural network")
   La conception d'un réseau de neurones est inspirée des fonctions d'un neurone dans le cerveau humain. Il est composé d'une succession de couches, chacune composée de N_i neurones, prenant leurs entrées sur les N_{i-1} neurones de la couche précédente. Un neurone est un modèle qui se caractérise par un état interne, des

signaux d'entrée $x_1, ..., x_p$ et une fonction d'activation (ou fonction de seuillage, ou de transfert). Les différents types de neurones se distinguent par la nature de leur fonction d'activation, fonction qui détermine le niveau d'activation du neurone en fonction des signaux qu'il reçoit. Pour cela, la valeur de sortie (voir la figure 2.5) est comparée à un seuil pour déterminer l'état du neurone (actif ou inactif).

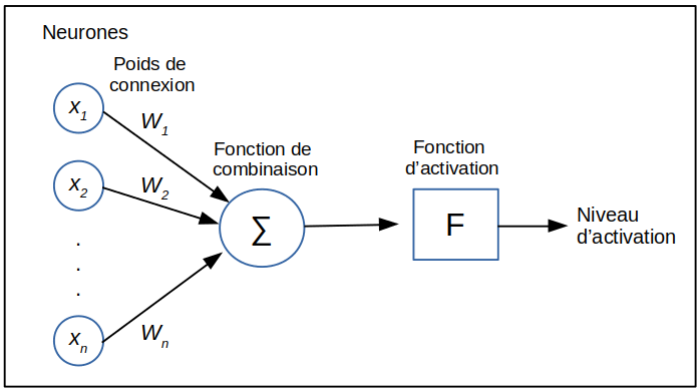


FIGURE 2.5 – Structure d'un neurone : la fonction d'activation prend la somme de ses entrées x, pondérée par les poids w pour calculer sa sortie.

Les réseaux se distinguent par leur architecture (nombre de couches, type de connexion, ...), par le type des neurones (leurs fonctions d'activation), par leur niveau de complexité (nombre de neurones, présence ou non de boucles de rétroaction), et enfin par l'objectif visé : apprentissage supervisé ou non, optimisation, systèmes dynamiques...

- **Réseau supervisé** (ou "Supervised neural network") Les poids sont modifiés dans le but de minimiser l 'erreur (entre sortie souhaitée et sortie obtenue) sur la base d'apprentissage.
- **Réseau non-supervisé** (ou "Unsupervised neural network") Le réseau catégorise les données selon certaines similitudes, il vérifie la corrélation entre les différentes entrées et les regroupe. Les poids sont donc déterminés par rapport à des critères de conformité.
- **Réseau renforcé** (ou "Reinforced neural network") Une information extérieure est fournie au réseau indiquant si la décision prise est bonne ou mauvaise. Si la décision est bonne, les connexions responsables sont renforcées et sont affaiblies dans le cas contraire.

7. Apprentissage fondé sur une instance (ou "instance-based learning") AHA et collab. [1991] Se base sur la similarité entre l'échantillon et ses voisins les plus proches dans les données d'entraînement pour émettre une prédiction.

    - **k plus poches voisins** (ou "k-Nearest Neighbor) : Pour prédire la classe d'un nouvel échantillon, l'algorithme va chercher ses K voisins les plus proches en se basant sur le calcul d'une distance (euclidienne, ou autres) et choisira la classe majoritaire de ces voisins.

Le choix de l'algorithme d'apprentissage automatique ainsi que la qualité du travail effectué dépendent de certains facteurs liés à la base de données. Le nombre d'échantillons est par exemple un de ces facteurs. En effet, une faible base de données rend l'analyse difficile et non robuste. Au contraire, une large base de données est préférable mais plus le nombre d'échantillons est élevé plus l'analyse est longue et le besoin en mémoire informatique important. Le nombre et la qualité des labels (étiquettes), le pourcentage de données renseignées et/ou manquantes ainsi que la présence de bruit (valeurs non conformes au modèle de distribution générale des échantillons sur leur espace de distribution) sont également des facteurs déterminants.

2.2.1 Introduction sur la classification

La classification est une méthode permettant de catégoriser un ensemble d'individus en **classes** (appelées aussi étiquettes ou catégories) à partir de divers types de données, dits **descripteurs**. Depuis ces dernières décennies de nombreuses méthodes de classification ont été proposées et ont été appliquées à de nombreux domaines (biologie, chimie, robotique, ..). Deux phases régissent la classification, l'apprentissage qui n'est réalisé qu'une fois à partir de données connues, puis la prédiction c'est-à-dire l'application du système de classification appris précédemment sur de nouvelles données afin de deviner leurs catégories. Afin d'optimiser ce système, la cohorte de données est divisée en deux, une cohorte d'apprentissage et une de test. La cohorte d'apprentissage est utilisée pour apprendre le système de classification et la cohorte test permet quant à elle d'estimer son efficacité de classification des données. L'extraction des valeurs de descripteurs discriminants permet l'obtention d'une représentation de l'observation qui correspond à une description de l'information pertinente qu'il contient.

La classification automatique d'images

Le principe de cette classification automatique correspond à un système permettant d'effectuer une tâche d'expertise non-influencée par le volume important de données d'images à traiter contrairement à un humain, pour qui ce même travail sera couteux en énergie (concentration, fatigue) et en temps. De plus, une classification automatique des images pourrait potentiellement mettre en évidence des caractéristiques qui ne semblent pas pertinentes pour l'œil humain, voire qui sont même ignorées comme par exemple du bruit. Cela est dû à deux facteurs : premièrement, une machine n'est pas biaisée par des attentes a priori et par des expériences passées, deuxièmement, la machine adapte sa façon d'apprendre aux données, tandis que les humains essaient d'intégrer les données dans leurs structures d'apprentissage déjà largement pré-formées. La classification d'images a de nombreuses applications concrètes telles que la classification d'objets, de documents, de textures, de scènes, la reconnaissance de visages et d'empreintes digitales. Dans le domaine médical on retrouve des applications à plusieurs niveaux, cellulaires, tissulaires ou à l'échelle de l'organe. On retrouve par exemple la classification de globules blancs LIPPEVELD et collab. [2020]., la classification de tissus IRSHAD et collab. [2013]; SUDHARSHAN et collab. [2019] et d'organes DOLZ [2016] dans le but d'identifier les tissus cancéreux. Des méthodes de classification utilisant comme données des images sont également utilisées dans les domaines tels que l'agronomie (reconnaissance de différents types de champignons, d'herbes, type de sols, de grains et de pollen) CHEN et collab. [2019], la géographie avec la classification des images de la couverture terrestre MA et collab. [2017], la classification d'images satellites ABBURU et GOLLA [2015]. On les retrouve également dans le domaine urbain comme la reconnaissance et le suivi de piétons LI et collab. [2006] ainsi que dans la reconnaissance automatique des panneaux de signalisation STALLKAMP et collab. [2011].

Toutes ces applications suivent un protocole permettant de fournir à partir d'images en entrée, des catégories pour chacune d'elles en sortie. Ce protocole correspond au système de classification dont les méthodes doivent être optimisées suivant la problématique de départ dans le but d'obtenir une performance globale optimale du système de classification. Il comprend à la fois les étapes de nettoyage des données, d'extraction de paramètres discriminants sous formes de vecteurs numériques et l'étape d'apprentissage du classifieur ou modèle.

2.2.2 Quelques méthodes de classification

L'ensemble d'apprentissage S est constitué de descripteurs x associés à leur classe respective y par un expert, et peut être écrit sous la forme :

— Classification binaire

$$S = \left\{(x^{(i)}, y^{(i)}) | x^{(i)} \in \mathbf{R}^d, y^{(i)} \in \{-1, +1\}, 1 \leq i \leq m\right\} \tag{2.6}$$

— Classification multi-classe

$$S = \left\{(x^{(i)}, y^{(i)}) | x^{(i)} \in \mathbf{R}^d, y^{(i)} \in \{1, ..., n\}, 1 \leq i \leq m\right\} \tag{2.7}$$

où m est le nombre de signatures et $n > 2$ le nombre de classes.

Il existe de nombreuses méthodes de classification linéaire ou non-linéaire, binaire ou multi-classe. Nous nous concentrerons en particulier sur trois méthodes qui sont très populaires en apprentissage automatique (Machine Learning). Ce paragraphe abordera d'une part les "Support Vector Machines" (ou séparateurs à vaste marge) qui sont une méthode de classification binaire linéaire pouvant être étendue à des cas non-linéaires et multi-classe (1-vs-1, 1-vs-all, DAGSVM ...), puis nous présenterons d'autre part deux méthodes multi-classe, la méthode des forêts aléatoires fondée sur l'apprentissage d'arbres de décision et une méthode de Deep Learning appelée réseaux de neurones convolutifs.

SVM Nous pouvons citer des exemples d'utilisation des SVM en médecine VAROL et collab. [2012], en reconnaissance de visages GUO et collab. [2001], de chiffres et de manuscrits FENG et MANMATHA [2005]. Nous recherchons l'hyperplan (si on se place en deux dimensions, l'hyperplan est une droite) qui sépare les deux classes en maximisant la distance (appelée marge) aux échantillons les plus proches appelés vecteurs supports (voir la figure 2.6). Nous recherchons donc l'hyperplan offrant la plus grande marge. Ainsi, l'équation de l'hyperplan de marge maximale ne dépend que des vecteurs de support, qui représentent en général un petit nombre d'échantillons de l'ensemble d'apprentissage. L'algorithme est construit de façon à obtenir un compromis entre la maximisation de la marge (pouvoir de généralisation) et le contrôle des erreurs (résultats de classification sur l'ensemble d'apprentissage). L'équation de l'hyperplan est une combinaison linéaire du vecteur d'entrée $x = (x_1, ..., x_N)^T$, avec un vecteur de poids $w = (w_1, ..., w_N)^T$. Afin de déterminer cette équation, nous devons estimer w et w_0 définissant l'hyperplan (2.8) :

$$h(x) = w^T x + w_0 = 0 \tag{2.8}$$

En optimisant : $\arg\min \frac{1}{2}||w||^2$
Sous la contrainte de : $\forall(\vec{x}, y) \in E, y * (w^T x + w_0) = 1$

Dans le cas non-linéaire, le produit scalaire entre les éléments de l'espace d'observation X est remplacé par le produit scalaire entre les images de ces éléments dans l'espace transformé, autrement dit par une fonction noyau. C'est le noyau K qui code une transformation des données permettant la recherche de surfaces séparatrices non linéaires (cas non linéaire illustré sur la figure 2.7).

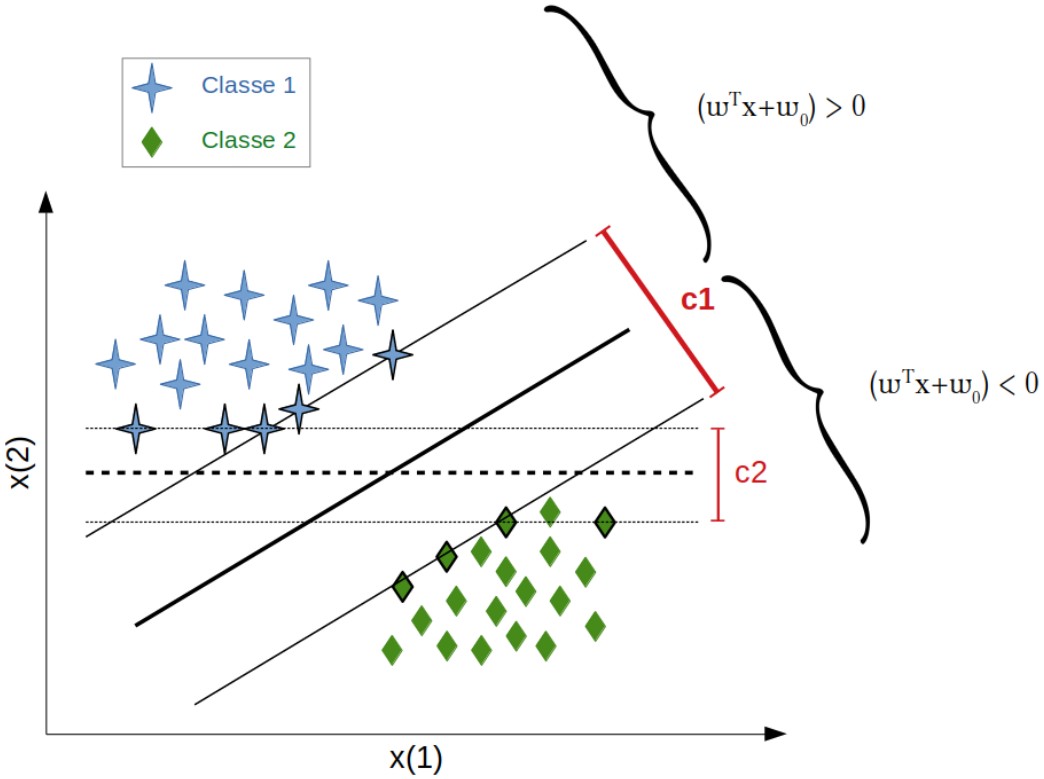


FIGURE 2.6 – Hyperplan séparateur de marge maximale : deux exemples d'hyperplans possibles 1 et 2, l'hyperplan 1 à la marge c1 la plus forte. L'hyperplan sépare les échantillons. Les vecteurs de supports sont les observations situées sur les droites frontières.

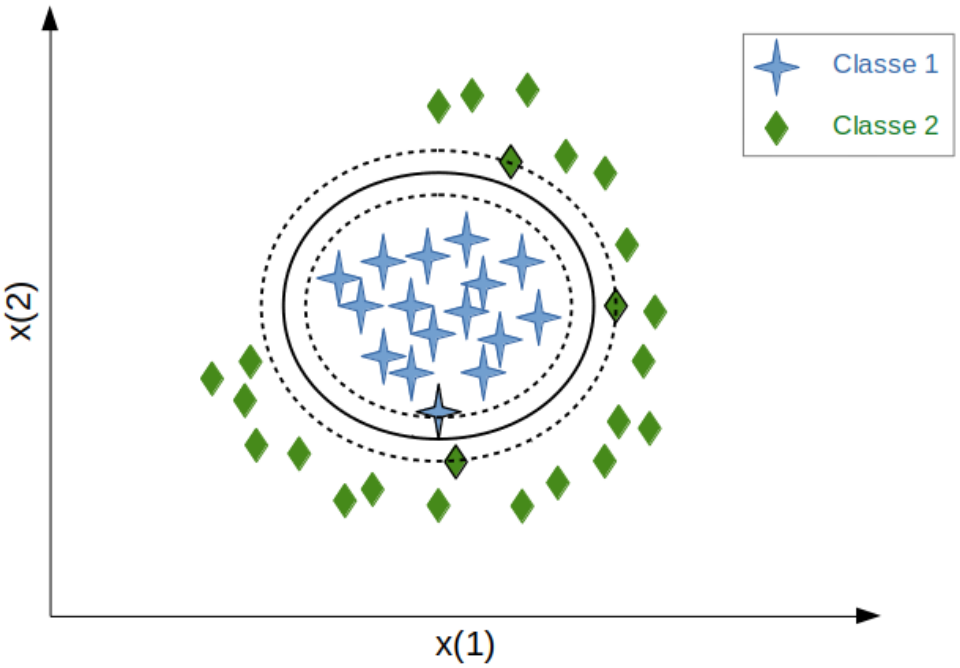


FIGURE 2.7 – Cas dans lequel la séparation linéaire n'est pas adaptée ; l'aide d'une fonction noyaux est alors nécessaire pour obtenir une fonction de séparation, celle-ci est non-linéaire. Par exemple ici : $\rho = \sqrt{x_1^2 + x_2^2}$ et $\theta = \arctan(x_2/x_1)$

Au-delà de deux classes, on parle de classifieur multi-classes. Parmi les méthodes de classification multi-classes les plus populaires, on retrouve le SVM multiclasses WANG et XUE [2014] mais également les arbres de décision QUINLAN [1986] et les forêts aléatoires. Le principe général de ces arbres est la décomposition du problème de classification en une suite de tests. Ces tests correspondent à une partition de l'espace des données en sous-régions. Ces sous-régions présentant une homogénéité par classe. Ces tests s'effectuent aux nœuds de l'arbre allant de la racine à une feuille finale indiquant la classe prédite. L'un des algorithmes les plus populaires et les plus puissants, fondé sur l'apprentissage de multiples arbres de décision est la méthode des forêts aléatoires HO [1995]. C'est un cas particulier de bagging (bootstrap aggregating) appliqué aux arbres de décision de type CART (Classification And Regression Trees) BREIMAN et IHAKA [1984].

Forêts aléatoires Le principe général d'une forêt aléatoire est de construire une collection d'arbres de classification (dits prédicteurs), pour ensuite agréger l'ensemble de leurs prédictions. La racine de l'arbre contient tous les échantillons de l'ensemble d'apprentissage. La première étape consiste à découper au mieux cette racine en deux nœuds fils. L'algorithme CART sélectionne alors la meilleure découpe, c'est-à-dire le couple paramètre / valeur du seuil sur ce paramètre p_n, qui minimise une certaine fonction de coût. L'indice de Gini, mesure statistique (entre 0 et 1) reflétant la répartition d'une variable au sein d'une population donnée permet de mesurer l'homogénéité des noeuds fils. Un fort indice indique une forte inégalité. Un nœud est dit homogène s'il ne contient que des observations de la même classe. Dans le but d'augmenter l'homogénéité des nœuds obtenus, on cherche à minimiser la fonction d'impureté de Gini; $h(p_1, p_2, ..., p_J) = \sum_{i \neq j} p_i p_j$. Toutes les observations de l'ensemble E_n avec une valeur de la variable p_n plus petite que le seuil s_{p_n} vont dans le nœud fils de gauche, et les autres vont dans celui de droite (voir la figure 2.8). On suit ainsi le même procédé en découpant à chaque fois chacun des nœuds fils en deux nouveaux nœuds. Le découpage de la partition dans l'espace des paramètres p_n est construite suivant la même logique. L'ensemble de départ E_1 est divisé en deux sous-ensembles E_2 et E_3. La partition est donc scindée en deux verticalement en p_1 suivant l'axe X_1. Les observations dont la variable est inférieur à p_1 sont placées dans le sous-ensembles E_2 (à gauche) et dans le cas contraire, dans E_3 (à droite). L'opération est répétée jusqu'au dernier paramètre de l'arbre.

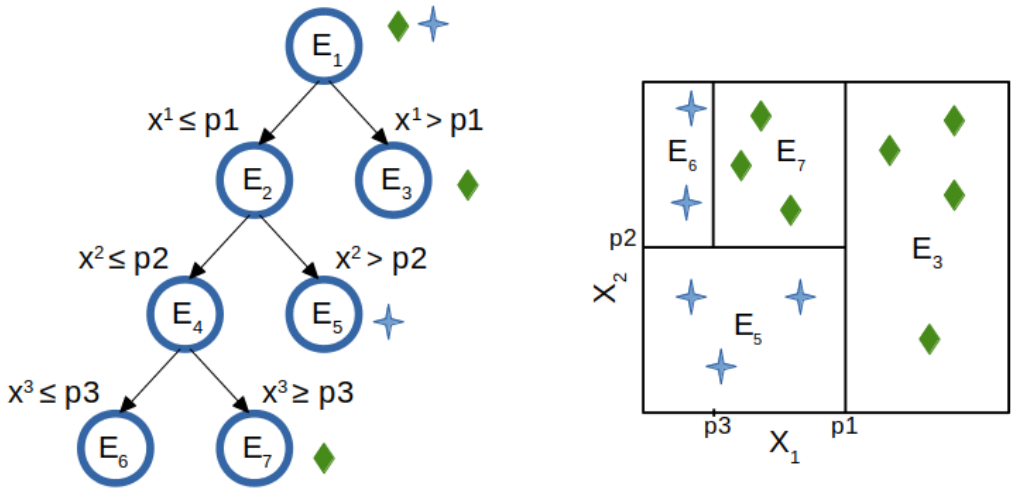


FIGURE 2.8 – À gauche : Arbre de décisions permettant la prédiction de la classe de l'observation x (gauche). À droite : la partition associée dans l'espace des paramètres p_n.

L'apprentissage des arbres de décisions s'effectue sur des sous-ensembles de données différents. La méthode de forêt aléatoire, dont le principe est décrit *via* la figure 2.9, consiste donc à construire et combiner plusieurs arbres. Cette combinaison est importante car les arbres de prédiction seuls présentent une sensibilité élevée aux fluctuations d'échantillonnage entraînant une qualité prédictive inférieure aux autres méthodes et notamment une variance assez élevée. Les arbres sont construits à partir de ré-échantillonnages boostrap du jeu de données d'origine. Le bootstrap consiste à créer de "nouveaux échantillons" statistiques en tirant avec remise à partir de l'ensemble d'échantillons initial. Cela permet d'avoir des arbres tous différents mais qui restent précis de manière unique. Cette technique permet de construire un modèle de prédiction de variance plus faible sans augmenter le biais. Afin de réaliser une prédiction, une classe est attribuée par chaque arbre de la forêt, à l'échantillon étudié. Pour cela, les arbres sont parcourus suivant les valeurs de l'observation puis arriver à une feuille, l'observation est étiquetée avec le label de la classe correspondant à cette feuille. L'étiquette finale de l'observation est attribuée par le résultat majoritaire. Deux paramètres sont importants à ajuster : le nombre d'arbre à construire et le nombre de caractéristiques sélectionnées aléatoirement (m). L'algorithme va créer une partition en ne prenant en compte que m variables lors de la formation de chaque nœud. Cela permet d'éviter que les variables les plus discriminantes apparaissent dans la majorité des arbres de la forêt et de conserver l'information contenue dans les autres variables, la méthode limite l'influence de ces variables.

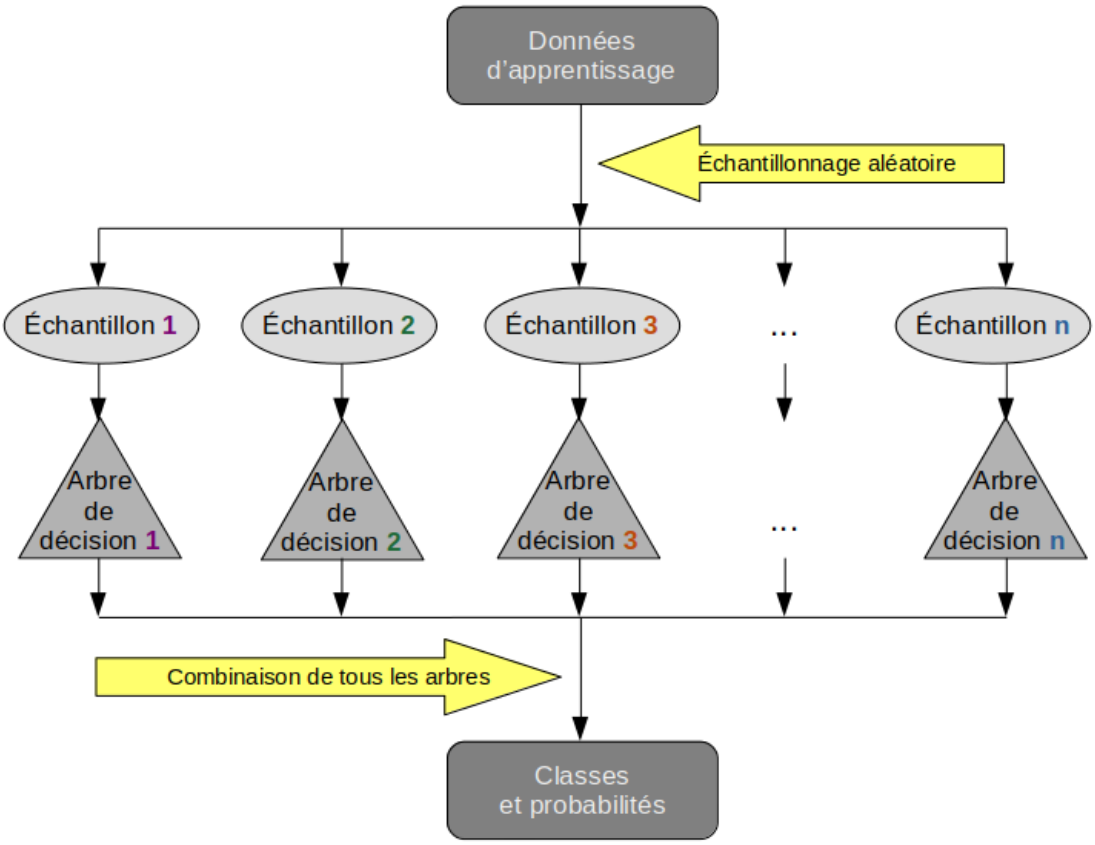


FIGURE 2.9 – Principe des forêts aléatoires : combinaison de plusieurs arbres de classification entraînés sur n sous-ensembles de données différents tirés selon un tirage aléatoire des données d'apprentissage.

Une dernière catégorie de classifieurs multi-classes et non des moindres, ayant beaucoup fait parler d'elle ces dernières années, est la méthode des réseaux de neurones. Cette méthode est inspirée de l'organisation des neurones dans le cerveau animal lors du traitement d'un signal visuel. Nous parlerons ici d'une sous-catégorie des réseaux de neurones conçus spécialement pour traiter des images, les réseaux de neurones convolutifs (CNN ou ConvNet). Un exemple d'un réseau de convolutions est présenté sur la figure 2.10. L'engouement pour ces méthodes est apparu en 2012 lors de la compétition annuelle de vision par ordinateur ILSVRC, durant laquelle AlexNet surpasse les autres méthodes de classification.

CNN La méthodologie des CNN est semblable à celle des méthodes traditionnelles d'apprentissage supervisé. De façon concise, les étapes sont les mêmes ; récupérer en entrée des images, identifier les caractéristiques discriminantes et apprendre un classifieur sur la base de ces paramètres. La différence est que dans ce cas précis, la détection des descripteurs est automatique et fait partie de l'étape d'apprentissage du classifieur. En effet, l'erreur de classification est minimisée dans le but d'optimiser à la fois les paramètres du classifieur mais également les caractéristiques. C'est justement pour cela que la construction de tels modèles pour la classification d'images connait un tel succès. Un CNN, de par son architecture peut donc être utilisé pour extraire progressivement des représentations de complexité toujours plus élevée du contenu de l'image. Le travail fastidieux et parfois très complexe du choix des descripteurs les plus pertinents pour une étude, est ici fait de façon automatique.
En ce qui concerne l'architecture du CNN, elle est composée de deux blocs. Le premier est spécifique de ce type de réseaux puisqu'il correspond au fameux extracteur automatique de caractéristiques (décrit dans le paragraphe 2.2.4). Cette partie du CNN fournit un vecteur de description pour chaque échantillon, ces vecteurs constituent l'entrée du deuxième bloc. Celui-ci est quant à lui retrouvé à la fin de tous les réseaux de neurones utilisés pour la classification. Il est composé de plusieurs couches de neurones connectées entre elles appelées couches "fully-connected", la sortie d'une couche correspond donc à l'entrée de la suivante. Chacune de ces couches reçoit en entrée des données à partir desquelles elle va calculer une combinaison linéaire dont les coefficients définissent les poids de la couche. La dernière couche "fully-connected" permet de classifier l'image en entrée du réseau et contient autant de neurones que de classes. Enfin, une fonction de perte est calculée, associée à la couche Softmax (après la dernière couche du CNN). En général, pour des problèmes de classification, on utilisera plutôt une fonction d'entropie croisée JAMIN et HUMEAU-HEURTIER [2020]. L'idée est de modifier itérativement les poids afin de minimiser la fonction de perte par rapport à la sortie attendue (classes réelles des échantillons de l'ensemble d'apprentissage). En se plaçant dans l'espace des paramètres/poids des neurones, la fonction de perte (voir la figure 2.11) possède des minima et maxima locaux, en fonction des valeurs de ces paramètres. L'initialisation aléatoire des paramètres définit une position dans l'espace des paramètres à partir de laquelle on suit l'opposé du gradient le long de la plus "forte" pente de la fonction de perte. L'algorithme de descente de gradient permet aux poids d'être mis à jour *via* la rétropropagation du gradient. Cette méthode permet le calcul du gradient de l'erreur pour chaque neurone du réseau de la dernière à la première couche. Des variantes de l'algorithme de descente de gradient existent tels que la méthode utilisant le gradient conjugué HESTENES et collab. [1952] et certains diffèrent de par la fréquence des mises à jour des poids (utilisant par exemple le "Dropout", c'est-à-dire en désactivant de manière aléatoire certains neurones).

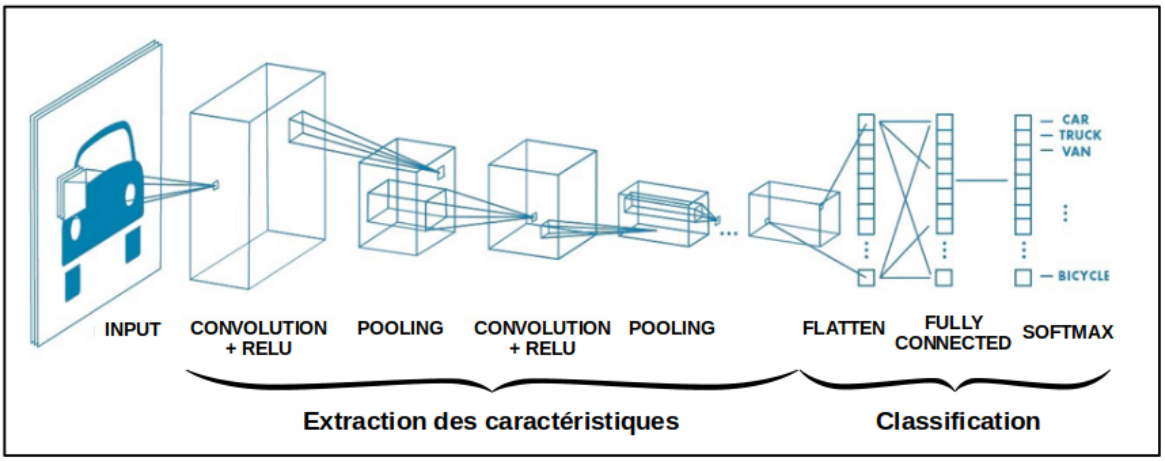


FIGURE 2.10 – Exemple d'un réseau de convolutions avec deux couches de convolution et de sous-échantillonnage représentées et connectées à deux couches entièrement connectées.
Partie "Extraction des caractéristiques" :
CONVOLUTION : Un filtre passe sur l'image en calculant le produit de convolution entre la caractéristique et chaque portion de l'image balayée.
RELU : fonction d'activation qui normalise les cartes de caractéristiques.
POOLING : couche permettant la réduction de la taille des images en la découpant en grille régulière et en gardant au sein de chaque cellule qu'une seule valeur (la valeur maximale ou autre).
Partie "Classification" :
FLATTEN : les valeurs des dernières cartes de caractéristiques sont misent dans un vecteur formant un descripteur.
FULLY CONNECTED : couches de neurones connectées entre elles.
SOFTMAX : fonction exponentielle normalisée qui modifie les valeurs des sorties en leur attribuant des valeurs positives dont la somme fait 1.
Schéma provenant de la page MathWorks.

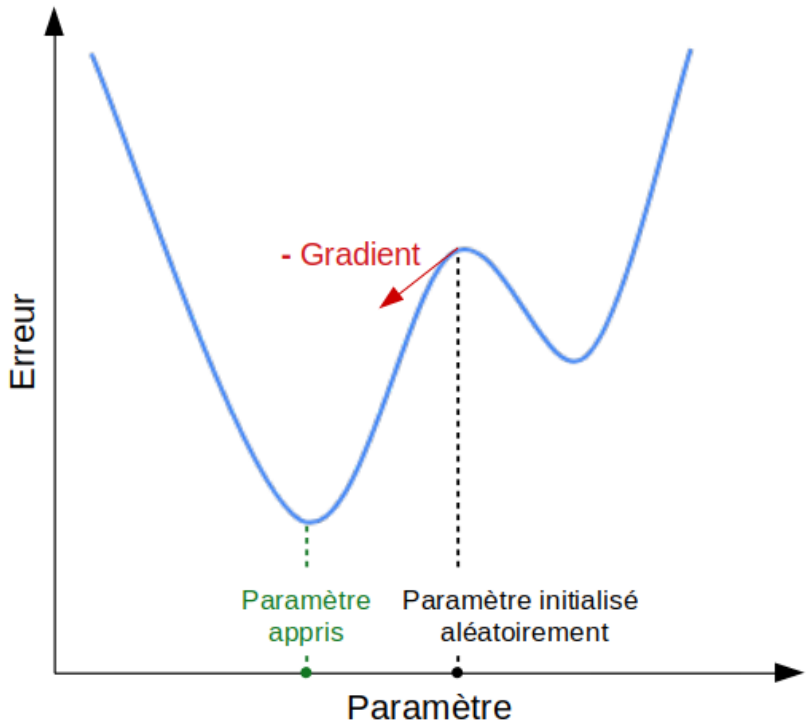


FIGURE 2.11 – Exemple de descente de gradient en une dimension (un paramètre à optimiser). La fonction de perte est ici représentée, à partir d'une valeur du paramètre (initialisée de manière aléatoire), on va optimiser la valeur de ce paramètre en cherchant à minimiser l'erreur.

Ainsi, si on prend un réseau à N couches, W^n les poids associés à la $n^{ième}$ couche, α le taux d'apprentissage et L la fonction d'erreur, les poids sont mis à jour suivant l'équation suivante :

$$W^n_{i,j} = W^n_{i,j} - \frac{\alpha \partial L}{\partial W^n_{i,j}} \tag{2.9}$$

2.2.3 Évaluation de la précision d'un classifieur

Évaluer les performances d'un système de classification est un enjeu de grande importance permettant entre autre l'optimisation des valeurs des hyperparamètres du classifieur. Ainsi, dans le but de mesurer la qualité d'un système de classification, le classifieur est testé sur des observations dont on connait la classe réelle et qui ne font pas partie de l'ensemble d'apprentissage sur lequel le classifieur a été appris. Ces échantillons constituent l'ensemble de test. Une fois les prédictions émises, un pourcentage par classe des observations correctement prédites par le classifieur est calculé. Les résultats sont généralement regroupés sous la forme d'une matrice de confusion. Chaque ligne de cette matrice correspond à une classe réelle et chaque colonne correspond à une classe prédite. Ainsi, dans l'exemple présenté dans le tableau 2.1, la cellule de la ligne C, colonne A contient le nombre d'éléments de la classe réelle C prédits comme appartenant à la classe A. L'intérêt de présenter les résultats ainsi est de voir rapidement si le classifieur est efficace dans la prédiction des observations des différentes classes. En effet, la diagonale de la matrice indique le score obtenu pour chaque classe. Les éléments de la classe A sont correctement prédits à 100%, ceux de B à 78% et ceux de C à 63%. Les autres valeurs correspondent aux pourcentages d'erreur. Si on prend comme classe de référence la classe C ; 63% des échantillons de cette classe sont correctement prédits comme appartenant à la classe C et 37% sont mal prédits (33% et 4% prédits respectivement comme appartenant aux classes A et B). On voit également que 22% des observations de la classe B sont prédits à tort comme appartenant à cette classe C. Ainsi, en plus de montrer clairement l'efficacité du classifieur au travers des scores sur la diagonale, son analyse permet de mettre en valeur le niveau de confusion entre les classes.

	A	B	C
A	1	0	0
B	0	0.78	0.22
C	0.33	0.04	0.63

TABLEAU 2.1 – Exemple de matrice de confusion illustrant les résultats de classification normalisés d'échantillons de trois classes (A, B et C). Chaque ligne de cette matrice correspond à une classe réelle et chaque colonne correspond à une classe prédite. Cette matrice indique pour chaque classe une valeur entre 0 et 1, indiquant le rapport des observations prédites comme appartenant à chacune des classes (la somme de ces rapports fait 1).

Dans ce paragraphe, nous venons de décrire trois méthodes de classification et d'expliquer comment évaluer l'efficacité d'un classifieur. Il n'existe pas de méthode universelle, le choix de l'algorithme dépend de nombreux facteurs tels que la taille, la nature, la qualité des données, le but de la classification et le temps d'analyse qui pourra être dédié.

2.2.4 Extraction de caractéristiques

En classification, la construction du modèle est généralement fondée sur l'apprentissage d'un ensemble de critères numériques décrivant l'observation, appelées caractéristiques ou descripteurs. Chaque donnée, peu importe sa nature (une image, une vidéo, ...), est donc représentée synthétiquement par un vecteur appelé descripteur pouvant appartenir à un espace de grande dimension, et dont les composantes sont les caractéristiques. Ce vecteur constitue en réalité une signature de l'observation qui nous permet de travailler dans un espace numérique. Comme évoqué dans le paragraphe 2.2.2, les réseaux de neurones convolutifs présentent la capacité de calculer de façon automatique ces descripteurs pendant l'étape d'apprentissage du classifieur. Cette faculté est en revanche propre à cette méthode des CNN. En effet, pour l'ensemble des autres méthodes, l'extraction des descripteurs représente une tâche à part entière. Cette tâche est généralement précédée d'une phase de pré-traitement des données. Une fois les images nettoyées et corrigées, nous devons choisir avec soin les descripteurs qui nous permettrons de caractériser les observations. Le choix de ces descripteurs nous placera dans un espace de description nous permettant de catégoriser les observations à distinguer.

Les caractéristiques peuvent être définies comme bas-niveau lorsqu'elles utilisent par exemple les informations au niveau local dans l'image (comme par exemple la texture ou la couleur) et de plus haut niveau si elles utilisent entre autres le squelette des objets de l'image comme paramètres. Pour rappel, la texture se définit par la répétition de motifs, de formes et d'orientations. Dans le cas de la classification d'images, les descripteurs sont très nombreux et comportent de nombreuses variantes. Parmi les descripteurs les plus utilisés on retrouve par exemple les descripteurs d'images colorimétriques, les filtres (Sobel, Canny, Gabor, ..), les descripteurs statistiques, les descripteurs fondés sur une transformation intégrale (transformée de Fourier rapide ou Fast Fourier Transform (FFT) (1965) entre autres). Cette description peut être globale (histogrammes de couleur et de textures ALI et collab. [2017] par exemple) ou locale suivant si elle tient compte de la totalité des informations présentes dans l'image MIKOLAJCZYK et SCHMID [2005]; SCHMID et collab. [2000]. Les descripteurs locaux comme par exemple SIFT LOWE [2004] ou SURF BAY et collab. [2008], ont l'avantage de prendre en compte la disposition spatiale du contenu visuel dans l'image. Si l'algorithme SIFT permet la détection et l'identification d'éléments similaires entre différentes images. L'algorithme SURF se concentre entre autres sur la détection d'objets. Ces descripteurs, suivant le domaine dans lequel on se trouve, peuvent avoir ou non un sens physique et correspondent à la quantification de caractéristiques morphologiques décrivant les objets d'intérêts (cellules, organes, ..), c'est le cas lors de l'étude d'images médicales et biologiques par exemple LASSOUAOUI et collab. [2005].

La morphologie se réfère à l'analyse quantitative de la forme d'une observation et les données mesurables d'une morphologie sont appelés caractéristiques morphologiques. Ces paramètres permettent donc de caractériser les objets mais également de les différencier d'autre types d'objets. Nous allons dans le paragraphe suivant nous concentrer sur l'opération morphologique de squelettisation qui définit le procédé selon lequel le squelette est extrait d'une forme binaire. Elle a suscité un grand intérêt dans le cadre de l'analyse d'images. Les méthodes présentes dans la littérature décrivent le squelette comme devant vérifier un certain nombre de propriétés :

— La reconstructibilité : le squelette doit être réversible, c'est-à-dire que l'on doit pou-

voir reconstruire l'objet à partir du squelette.

— L'épaisseur unitaire : le squelette doit posséder un seul pixel d'épaisseur.

— L'homotopie : le squelette doit être centré dans l'objet et posséder la même topologie que celui-ci. Il possède donc le même nombre de composantes connexes.

— La préservation de la géométrie : le squelette doit rendre compte de la forme globale de l'objet initial et de sa géométrie.

Ces propriétés permettent ainsi de décrire synthétiquement à la fois la forme, mais aussi des propriétés mathématiques des objets, tel que la longueur. Le puissant pouvoir descriptif du squelette justifie sa large utilisation dans le domaine de l'analyse d'image et de la reconnaissance de formes. Les approches issues de la morphologie mathématique permettant d'extraire un squelette sont populaires de par leur propriété d'homotopie. L'une de ces méthodes s'appuie sur le principe d'amincissements répétés consistant en une succession de passage visant à supprimer les pixels du bord de l'objet jusqu'à atteindre l'axe médian de celui-ci et être réduit à un seul pixel d'épaisseur ZHANG et SUEN [1984]. Dans ce cas de figure, une des limitations est la présence de bruit dans l'image. En effet, même si ces approches vérifient la propriété d'homotopie, sans nettoyage des images bruitées, le squelette obtenu risque d'être inexploitable.

Extraction automatique des caractéristiques par un CNN

Les CNN sont composés architecturalement de deux parties (voir la figure 2.10). La première constituant un extracteur automatique de caractéristiques est celle qui nous intéresse dans ce paragraphe (la deuxième ayant déjà été décrite dans le paragraphe 2.2.2). Le calcul des noyaux de convolution lors de la phase d'apprentissage fait que les couches de convolution, qui composent ce premier bloc, permettent la reconnaissance de motifs propres au problème à résoudre. Les couches de convolution filtrent l'image avec plusieurs noyaux de convolution (correspondant aux poids des neurones de la couche de convolution). Les deux premières dimensions de ce noyau correspondent à la largeur et la hauteur du filtre de convolution et la troisième indique le canal sur lequel la convolution est appliquée (s'il s'agit du traitement d'images multi-canaux). Ce canal correspond au canal couleur ou spectral. En considérant K un noyau (avec $\mathrm{K} \in \{0...w\} \times \{0..h\} \times \{0..d\}$, dont w la largeur, h la hauteur et d le canal) et une image $\mathrm{I} \in \mathbb{Z}$, la notation $\mathrm{I} \star \mathrm{K}$ correspondant à la convolution de l'image I par K, et est définie par l'équation (2.10) :

$$(\mathrm{I} * \mathrm{K})_{x,y,c} = \sum_{x'=0}^{w} \sum_{y'=0}^{h} \sum_{c'=0}^{d} \mathrm{K}_{x',y',c'} . \mathrm{I}_{x+x'-\frac{w}{2}, y+y'-\frac{h}{2}, c+c'-\frac{d}{2}} \tag{2.10}$$

Le calcul de la convolution de chacune des couches avec chaque filtre renvoie des cartes d'activation ou cartes de caractéristiques, qui sont ensuite normalisées à l'aide d'une fonction d'activation et/ou redimensionnées. Deux exemples de fonction d'activation sont les fonctions sigmoïde et l'unité linéaire rectifiée ou ReLU BREIMAN et IHAKA [1984] (voir la figure 2.12). Dans le premier cas, la variable z est fixée à la somme pondérée des entrées, puis est transmise à la fonction sigmoïde.

$$z = b + \sum_i w_i x_i \tag{2.11}$$

$$\sigma(z) = \frac{1}{1+e^z} \tag{2.12}$$

e est la constante exponentielle estimée à une valeur arrondie de 2.72. L'utilisation de cette fonction d'activation présente néanmoins un problème de gradient évanescent qui impacte négativement la formation des réseaux de neurones avec beaucoup de couches. La plupart des réseaux neuronaux actuels utilisent ReLU ou l'une de ses variantes. Elle est définie comme R(z)=max(0,z). Toutes les valeurs positives restent inchangées et les valeurs négatives sont misent à 0.

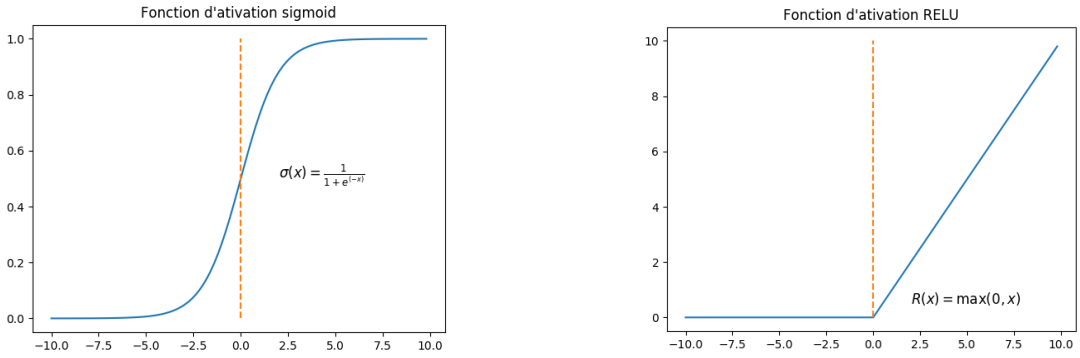


FIGURE 2.12 – Deux exemples de fonctions d'activation. La variable z est fixée à la somme pondérée des entrées (équation 2.11). À gauche, la variable z est transmise à la fonction sigmoïde (voir l'équation 2.12). À droite, la fonction RELU est définie comme R(z)=max(0,z) c'est-à-dire que les valeurs positives restent inchangées et les valeurs négatives sont mises à 0.

Le signal x_i arrivé à la dernière couche est en conséquence transformé suivant la fonction d'activation qui diffère selon que l'on se place dans un cas de classification binaire ou multi-classe. Dans le cas binaire, il s'agit d'une fonction logistique et dans l'autre une fonction softmax. Concernant la réduction de la taille des images, elle est possible *via* une couche de pooling généralement placée entre deux couches de convolution. Réduire la taille des images permet ainsi de diminuer la quantité de paramètres et de calcul dans le réseau tout en préservant les plus pertinentes de leurs caractéristiques. Cette couche découpe l'image en grille régulière, puis la valeur maximale ou autre (puisque d'autres fonctions de pooling que le maximum existent et peuvent être appliquées) est conservée au sein de chaque cellule. À la fin de l'apprentissage, les valeurs des dernières cartes de caractéristiques sont mises dans un vecteur formant un descripteur pertinent pour le problème donné. Cette méthode de calcul automatique des caractéristiques sur la base des échantillons d'apprentissage a permis des prouesses en termes de résultats de classification avec notamment le gagnant du challenge Image-Net (2014), le réseau GoogLeNet SZEGEDY et collab. [2015] présentant 22 couches de convolutions. Les CNN les plus connus autre que GoogLeNet sont VGG SIMONYAN et ZISSERMAN [2014] et ResNet HE et collab. [2016] avec respectivement 19 et 152 couches de convolutions.

L'extracteur automatique de caractéristiques du CNN permet donc l'obtention d'une représentation robuste des observations ATHIWARATKUN et KANG [2015]; SHARIF RAZAVIAN et collab. [2014]. Cet outil donne une représentation qui peut être utilisé comme descripteurs dans n'importe quelle méthode de classification LU et collab. [2014]. Dans certain cas, remplacer la seconde partie du CNN par la méthode des forêts aléatoires WANG et collab. [2019] ou un SVM NIU et SUEN [2012] permet d'obtenir des résultats de classification plus élevés.

2.3 Quelques outils d'analyse d'images et de classification utilisés pour le phénotypage

Les avancées en microscopie automatisée à haut débit ont considérablement augmenté la demande de méthodes de calcul pour l'analyse de données à grande échelle appliquées aux images biologiques GRYS et collab. [2017]. Ainsi, on retrouve d'une part, des approches de vision par ordinateur dédiées à la segmentation cellulaire et à l'extraction de caractéristiques et d'autre part, des approches d'apprentissage automatique permettant une classification phénotypique à partir d'images biologiques. L'analyse d'images phénotypiques est fréquemment utilisée pour la recherche de nouveaux médicaments ou de connaissances biologiques fondamentales. Les outils dédiés à cette tâche sont de plus en plus puissants et de nombreux logiciels commerciaux et open source ont été développés. Des revues telles que SMITH et collab. [2018] permettent d'avoir une idée globale des divers logiciels existants mais fournissent également une description de leurs différences (en termes de convivialité, de fonctionnalités, d'interface et de performances) ainsi qu'un résumé des fonctionnalités qu'ils possèdent. La revue GRYS et collab. [2017] fait, quant à elle, état des méthodes utilisées pour extraire des informations biologiques cellulaires quantitatives pour le profilage phénotypique. Cette revue se termine par une ouverture sur les méthodes fondées sur les réseaux de neurones profonds dont l'utilisation permettrait de dépasser certaines limites associées aux pipelines d'analyse conventionnels. Elle met notamment l'accent sur la possibilité d'automatiser l'ensemble du processus d'analyse pour la classification des phénotypes cellulaires (extraction des caractéristiques et apprentissage du modèle).
Parmi ces outils, nous pouvons citer les travaux qui ont inspiré mon projet de thèse NEGISHI et collab. [2009], dans lequel la microscopie permet l'identification de phénotypes de levures grâce à la quantification de leur morphologie cellulaire via le logiciel de traitement d'images, Calmorph OHTANI et collab. [2004]; OHYA et collab. [2005]. Les données extraites comprennent des mesures relatives à la forme des cellules de levure et à celle des bourgeons ainsi que la forme et l'emplacement des noyaux et la distribution de l'actine.
L'un des logiciels les plus couramment utilisés est CellProfiler CARPENTER et collab. [2006]. CellProfiler Analyst 2.0 DAO et collab. [2016], une extension du logiciel permet d'effectuer un apprentissage supervisé à partir de caractéristiques extraites pour reconnaître le phénotype d'une cellule présente à l'image. Il est écrit en Python et fonctionne avec plusieurs méthodes d'apprentissage automatique. Dans TOTH et collab. [2018], des images de cellules cancéreuses du sein, traitées avec différents médicaments et des images de coupes de tissus de vessie cancéreuse sont analysées par ce logiciel d'analyse d'images et d'apprentissage automatique. Leur méthode de phénotypage fondée sur l'apprentissage automatique, combine des caractéristiques locales et de voisinage pour identifier les phénotypes. CellProfiler leur permet de segmenter les images et de séparer les caractéristiques cellulaires (noyau, cytoplasme, ...). Puis, la méthode des k plus poches voisins (méthode décrite dans le paragraphe 2.2) ainsi qu'une approche fondée sur la distance sont utilisées pour calculer les caractéristiques du voisinage, en s'appuyant sur celles extraites des différents composants (texture, intensité, forme, ...). Enfin, ces caractéristiques servent dans des méthodes de classification (Forêts aléatoires, Naïve bayésienne, réseau de neurones, ...) pour différencier les classes phénotypiques. On retrouve également parmi les logiciels permettant de différencier des phénotypes, Advanced Cell Classifier (ACC) PICCININI et collab. [2017]. ACC v2.0 donne à l'utilisateur l'accès à une grande variété d'algorithmes d'apprentissage automatique. Ce logiciel permet également d'identifier de nouveaux phénotypes sans que cela nécessite de connaissances a priori sur l'ensemble de

données sous-jacent. Pour ce faire, un SVM à une classe est utilisé pour identifier les cellules trop différentes des types cellulaires connus. Ces cellules sont tout d'abord considérées comme leurs propres clusters puis un clustering hiérarchique est réalisé en fusionnant les clusters en fonction de leur similarités, jusqu'à tous les regrouper en un cluster unique. La métrique de similarité définie correspond à la distance euclidienne entre la moyenne des vecteurs de caractéristiques des cellules de chaque cluster. L'utilisateur a la capacité de créer manuellement, via l'interface graphique du logiciel, une nouvelle classe en annotant les échantillons "nouveaux" qui sont les plus similaires.

Il existe donc un certain nombre de logiciels qui permettent l'obtention de bons résultats d'analyse d'images, extraction de caractéristiques et classification de cellules ou ensemble de cellules. Néanmoins, deux raisons expliquent la nécessité de créer un outil spécifique à nos images et types de phénotypes. Ces raisons sont les suivantes :

— Nous devons prendre en compte la différence entre travailler en fluorescence (GFP par exemple), qui permet d'avoir un aspect quantitatif des observations et exploiter des images en lumière transmise (ou contraste de phase) permettant d'avoir un aspect qualitatif de ces observations. En effet, en fluorescence, nous avons accès à des informations morphométriques et radiométriques. En revanche, en lumière transmise, seule l'étude et l'analyse de la géométrie des objets, autrement dit la morphométrie est possible, ce qui explique la nécessité de développer une méthode spécifique qui ne s'appuie que sur cette information.

— La principale limite à l'utilisation de ces logiciels sur nos données est la complexité de nos phénotypes. En effet, les champignons sont constitués de plusieurs cellules et dans certains cas présentent des tubes germinatifs.

2.4 Origine et nature des données à classer

Dans cette partie du document, une description du modèle biologique choisi est effectuée et les protocoles d'expérimentation biologique et d'acquisition des images prises en microscopie sont abordés. Les classes propres à notre problématique y sont définies et leur choix justifié. Puis, la méthode d'extraction de caractéristiques décrivant les échantillons de chacune de ces classes est expliquée. Ces caractéristiques permettent de les distinguer et leur utilisation comme descripteurs dans une méthode de classification sera exposée dans le paragraphe 2.5.2.

2.4.1 Protocole d'expérimentation biologique

Le champignon *Botrytis cinerea* est un des modèles d'étude pour le laboratoire Bayer, de par la connaissance acquise sur ce pathogène filamenteux des plantes mais aussi du fait de sa facilité à être cultivé. Il a la capacité de croître rapidement et l'intégralité de son cycle biologique peut être reproduit. Le champignon est entretenu sur boîtes de Pétri dans un milieu nutritif gélosé contenant de l'agar (Potato Dextrose agar ou PDA). Au moment de l'inoculation (temps = 0 de culture), un petit morceau d'agar contenant du mycélium est déposé au centre de la boite de Pétri. Les boîtes sont ensuite incubées à 21 °C afin de permettre au champignon de se développer. On observe une croissance radiale pour finalement aboutir, dès 5 jours, à une boîte recouverte de mycélium (voir la figure 2.13).

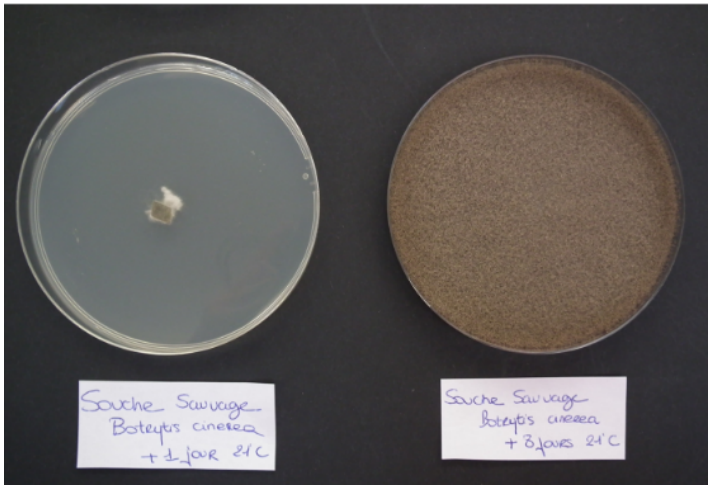


FIGURE 2.13 – Culture du champignon *Botrytis cinerea* en milieu solide : Développement selon une croissance radiale sur une boite de Pétri, milieu Potato Dextrose Agar. Observation à +1 et +8 jours d'incubation à 21 °C

Le mycélium produit des conidiophores semblables à un feutrage grisâtre qui porte les macroconidies ou spores. La manipulation du champignon se fait dans un laboratoire de type L1 et sous une hotte à flux laminaire afin d'éviter la dispersion de ces spores dans l'air.

Les études de phénotypes sont réalisées en milieu liquide en microplaques de 96 puits. Les plaques utilisées présentent un fond transparent essentiel aux applications de microscopie inversée. Le travail en microplaque et la miniaturisation des essais permet de réduire leur coût et d'augmenter le nombre de molécules testées par expérience. Une plaque contient un total de 8 molécules, testées chacune à 10 doses, de 100 à 0,005 µM (voir le paragraphe A.4 en annexe). Afin d'obtenir ces différentes concentrations en molécule, une dilution en série est effectuée, c'est-à-dire une dilution répétée de la solution originale. Les deux premiers puits de chaque ligne correspondent aux contrôles négatifs avec et sans DMSO (voir la figure 2.14). Le but est de visualiser l'effet de différentes molécules à différentes concentrations sur des spores de *Botrytis cinerea*. En effet, nous ne connaissons pas à l'avance le niveau d'efficacité de la molécule testées et il existe une dépendance entre la concentration de la molécule et le phénotype étudié. Par exemple, une molécule A pourra conduire au phénotype 1 sur une certaine gamme de doses donnée tandis qu'une molécule B pourra conduire au même phénotype 1 mais sur une gamme de doses différente. Ainsi le niveau d'efficacité d'une molécule est en corrélation avec l'impact de sa signature phénotypique. De plus, pour certaines familles de molécules, c'est la différence d'évolution du phénotype sur une même gamme de doses, qui pourra permettre de conclure qu'il s'agit d'un phénotype x ou y selon cette évolution.

Afin de préparer la solution de spores nécessaire à un essai, un volume d'eau stérile (5 ml) est déposé sur la boîte de Pétri contenant un mycélium âgé de 6 à 8 jours, de manière à assurer un taux de sporulation suffisant. À l'aide d'un grattoir passé sur le mycélium une solution d'eau concentrée en spores est récoltée. Cette solution est filtrée sur tamis de différentes tailles afin d'éliminer les déchets et de présenter uniquement des spores d'une taille moyenne de 10µm (filtration *via* deux tamis de 100 puis 40µm). Afin de calibrer la concentration de la solution en spores/ml, un comptage des spores est effectué au moyen d'une cellule de Thomas. Pour ce faire, une goutte de la solution de spores

est déposée entre la cellule de Thomas et la lamelle puis le comptage est réalisé sous un microscope à lumière transmise. Une solution fille est ensuite préparée et calibrée à la concentration de spores nécessaire à l'étude. Dans le cadre du projet, la concentration est calibrée à 2×10^4spores/ml.
A l'aide d'un robot, un volume de 200µL de la solution de spores et un de 2µL de la solution mère de molécules solubilisées dans le DMSO à la concentration souhaitée sont distribués dans chaque puits de la micro-plaque. Après incubation pendant 24h à 21 °C dans le noir, les microplaques sont placées dans le microscope de la plateforme de « High Content Analysis » pour une acquisition automatique des images.

2.4.2 Images par microscopie à lumière transmise

Les images sont acquises de façon automatique par un microscope à lumière transmise, Image Xpress de Molecular Devices, plateforme de type High Content Analysis. Ce microscope est doté d'un système de laser permettant le réajustement de la distance focale pour chaque position au sein de la plaque. Un protocole d'acquisition est ainsi établi, plusieurs images par puits sont réalisées avec l'objectif 10X. Le microscope automatisé représente un avantage considérable de gain de temps par rapport à un microscope classique manuel. En effet, une plaque correspond à 576 images, avec 6 images par puits, générées (voir la figure 2.14), ce qui représente une taille conséquente de données à traiter. Les images de la plaque sont produites en seulement 10 minutes par le microscope automatisé. Les images générées pour ce projet sont des images en lumière transmise à deux dimensions et en niveau de gris. Elles représentent une matrice de 2160 lignes et 2160 colonnes. La valeur d'un pixel peut varier en 2^n niveaux de gris, de 0 (noir) à $2^n - 1$ (blanc). Ainsi, les images étant codées en n = 16 bits ; les valeurs des pixels de nos images varient de 0 à 65535 ($2^{16} - 1$). Le vocabulaire associé à la mise en place d'une plaque 96 puits et de l'acquisition des images par microscopie est illustré sur la figure 2.15.

2.4.3 Notre première problématique : la classification des phénotypes

Comme décrit ci-avant (chapitre 1), un fongicide est une substance capable de limiter le développement des champignons. Les fongicides sont classés selon leur Mode d'Action. L'utilisation accrue de fongicides à fait apparaître de nouvelles souches pathogènes résistantes. L'apparition de ces souches résistantes conduit à l'urgent besoin d'identifier de nouvelles solutions de protection et l'identification de nouvelles molécules avec des nouveaux modes d'actions. Certaines familles de molécules données entraînent l'apparition de signatures phénotypiques caractéristiques. Un grand nombre de nouvelles molécules sont générées au laboratoire de chimie et doivent être soumises à des tests qui permettent l'identification de leurs mécanismes d'action, tests effectués au département de biochimie. Au sein de l'équipe Excellence biochimie, la microscopie est l'une des techniques utilisées pour étudier le mode d'action d'une molécule. Il s'agit de visualiser directement l'effet de celle-ci sur le développement du champignon, à des stades précoces *via* des observations entre 0 et 24 heures. La revue des images se fait actuellement de façon manuelle par un expert qui, en fonction de la forme du champignon, va émettre une hypothèse de mode d'action de la molécule testée. Le travail effectué, décrit ci-après, a pour but de développer un outil permettant la bonne reconnaissance du phénotype sur l'image afin d'émettre de façon automatique une hypothèse sur le mode d'action.

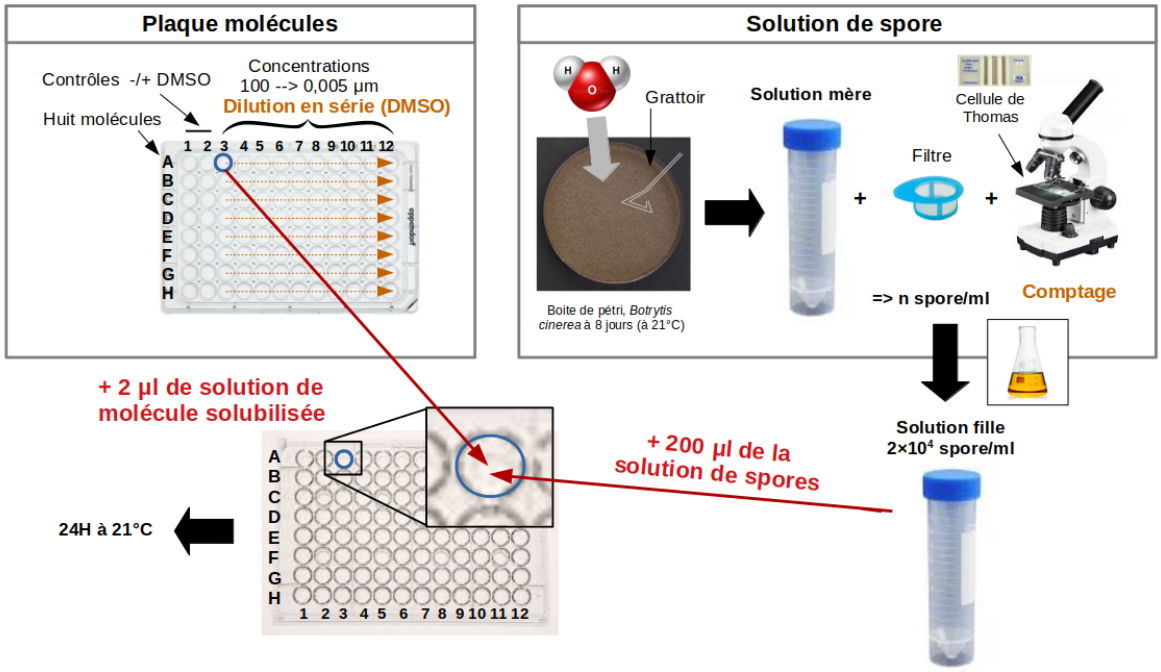


FIGURE 2.14 – Protocole d'expérimentation biologique : préparation des solutions de spores (mère et fille) et des différentes concentrations des molécules par dilution en série.

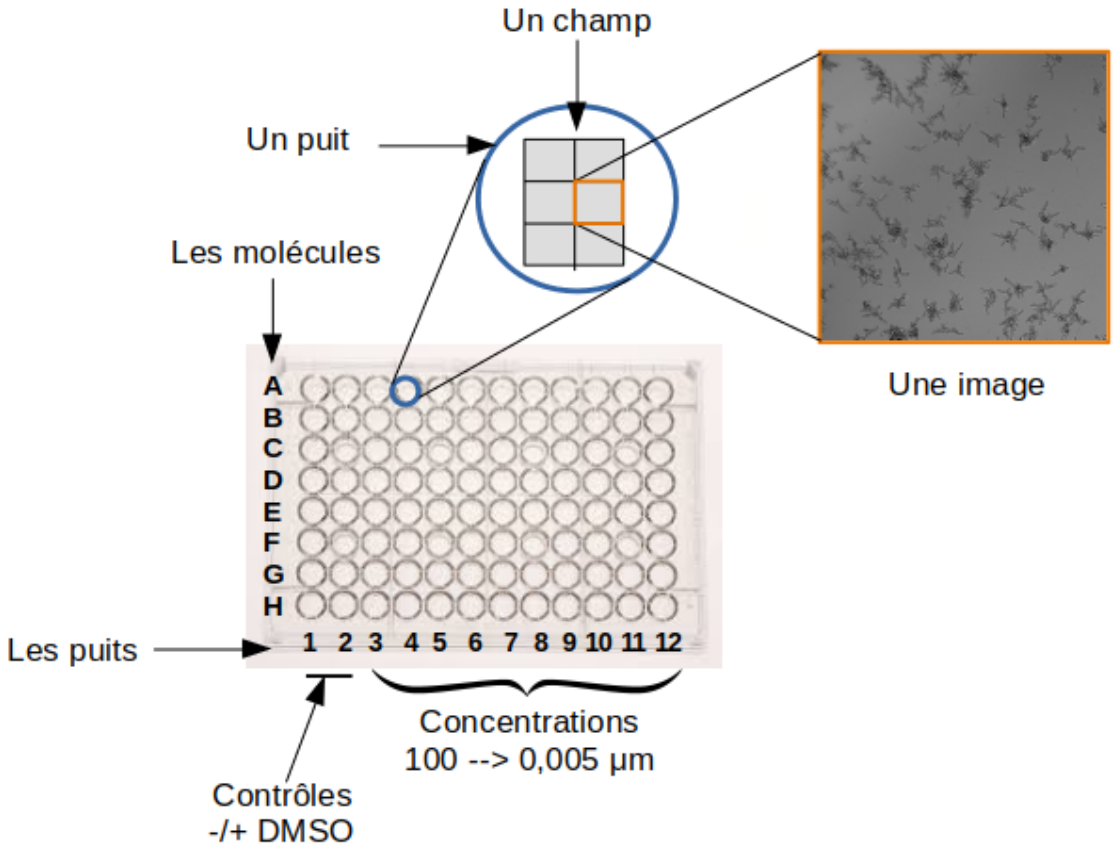


FIGURE 2.15 – Vocabulaire associé à la mise en place d'une plaque 96 puits et de l'acquisition des images par microscopie.

Nombre de classes (phénotypes connus / MoA)

Dans mon projet de thèse je me suis focalisée dans un premier temps sur quatre MoA connus selon la méthode agile WIKIPÉDIA [2021]. Ce sont en effet, à ce jour les phénotypes les plus représentés au laboratoire (voir la figure 2.16, ligne du bas). Il s'agira ensuite d'élargir à d'autres cas cette reconnaissance de phénotype.
Suivant le niveau d'efficacité de la molécule, le phénotype du champignon peut correspondre à l'un des quatre phénotypes remarquables mais il peut également correspondre à l'une des formes non remarquables du champignon. Ces formes non remarquables représentent les deux états de base du champignon, à savoir son état initial, le spore et l'état le plus développé, le mycelium (voir la figure 2.16, ligne du haut). Bien que ces formes ne donnent dans ce cas, aucune information sur le MoA de la molécule, des spores indiquent une forte efficacité de la molécule à bloquer le développement du champignon à la concentration donnée. Le mycelium permet, quant à lui, de donner l'information contraire. En effet, une molécule inactive ou peu active à une certaine concentration ne sera pas en mesure de bloquer le développement du champignon traité et permettra l'apparition d'un mycélium.

Définition des classes

Le système développé est entraîné à distinguer sept classes différentes (présentées sur la figure 2.16) divisées en deux catégories, les images "phénotypes" et les images dites "non phénotypes". Les images "phénotypes" regroupent les quatre phénotypes remarquables énoncés plus haut que l'on appellera les classes phénotypes numéro 1, 2, 3 et 4. Les images "non phénotypes" sont divisées en trois classes distinctes. Lorsqu'une molécule n'a pas ou plus d'effet sur le champignon, les images de mycélium se regroupent dans la classe mycelium. Dans le cas d'une inhibition de la croissance polarisée du champignon, les images sont rattachées à la classe Germination Inhibition. Enfin, certaines molécules peuvent présenter des problèmes de solubilité dans le solvant utilisé. Il y a alors présence de cristaux caractéristiques reconnaissables à l'image (voir le paragraphe A.5 en annexe) et qui constitueront la dernière classe appelée classe Crystal.

Nombre d'images par classes

Les molécules sont classées selon leur mode d'action et chaque molécule est testée à dix concentrations puis plusieurs images par puits sont générées afin d'augmenter la représentation statistique d'un phénotype par puits. Nous nous sommes concentrés, comme expliqué plus haut, sur quatre mécanismes d'action différents de molécule chimique antifongique. Afin d'entraîner notre système de classification, nous disposons d'images de phénotypes remarquables 1, 2, 3 et 4, respectivement : 224, 225, 297, 243 images. C'est en analysant la signature phénotypique présente sur ces images que l'on peut émettre une hypothèse de mode d'action. En ce qui concerne les images non phénotypes, nous bénéficions de 441 images Germination Inhibition, 5715 images mycelium et 360 images Crystal. Le grand nombre d'images mycelium s'explique par le fait que certaines molécules sont inactives et que chaque molécule testée présente au moins un puits où son efficacité est réduite, conduisant à la formation de mycélium. De plus, deux puits contrôles pour chaque test de molécule sont mis en place, des puits sans molécule dans lesquels, si l'expérimentation biologique s'est correctement déroulée, le champignon se développe en condition normale et forme du mycélium (voir la figure 2.14).

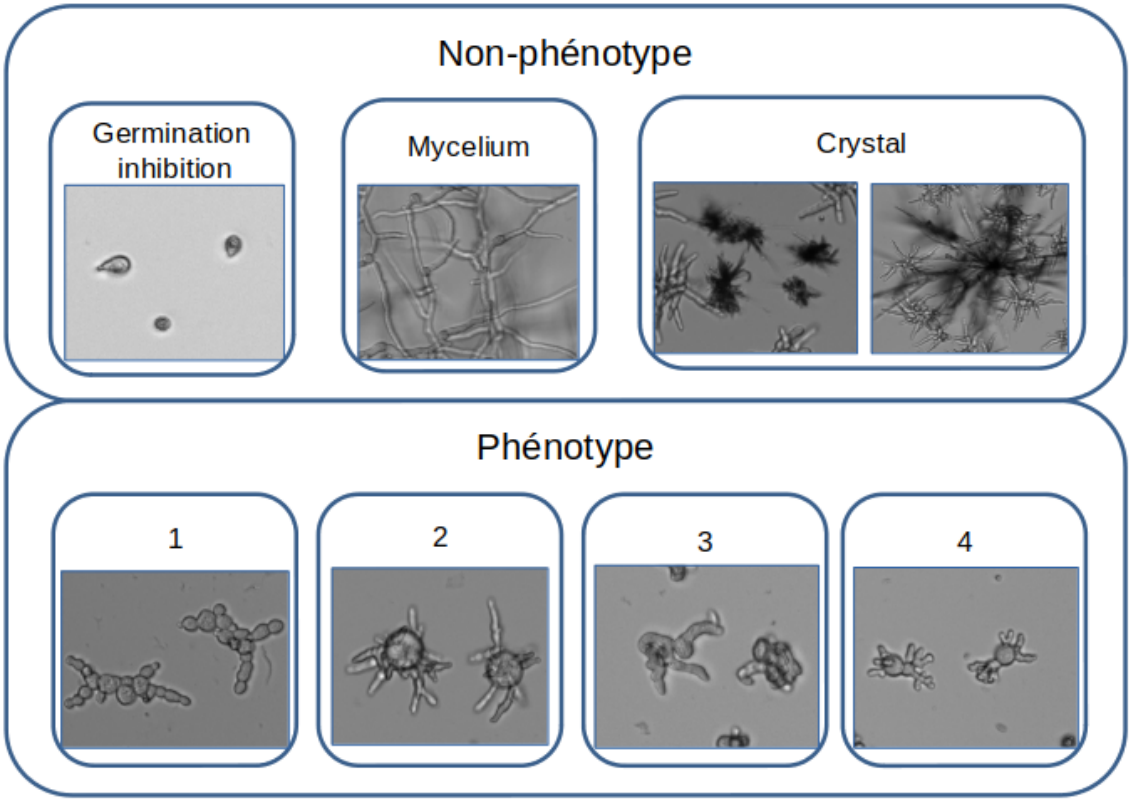


FIGURE 2.16 – Sept classes : six différentes formes du champignon et la classe crystal qui regroupe les images présentant des artefacts liés à l'agrégation de la molécule antifongique utilisée. Les quatre phénotypes de la deuxième case correspondent à des signatures phénotypiques caractéristiques du traitement chimique utilisé. Ces classes sont scindées en deux sous-classes : "Phénotype" et "Non-phénotype". Images en microscopie à lumière transmise, Microscope ImageXpress, Objectif x10.

2.5 Caractérisation et classification des phénotypes

2.5.1 Description des phénotypes

Nous avons insisté plus haut dans le paragraphe 2.2.4, sur l'importance du choix des descripteurs dans l'apprentissage d'un système de classification. Ainsi, dans le but de choisir au mieux les paramètres morphologiques qui serviront de représentation aux champignons des différentes classes, nous avons dans un premier temps établi une ontologie simple des différents phénotypes présentée ici sous forme de tableau de description (voir le tableau 2.2). Le tableau est scindé en trois, le portrait de la spore initiale du champignon, la description des branches (forme, composition et degrés des branches) et le caractère homogène ou non de l'effet de la molécule sur les champignons d'un puits.

Détection des objets de l'image

Afin de détecter les champignons de nos images, nous utilisons une opération de traitement d'images, la segmentation. Il s'agit plus précisément d'une binarisation. Le but de cette binarisation est de conserver les pixels pertinents appartenant aux champignons de l'image (pixels mis à 1) et d'éliminer les pixels du fond de l'image en mettant leur valeur à 0.

Phénotype		1	2
Spore initiale	Présence	Non identifiable	Visible
	Forme	–	Ovoïde
Branches	Taille	–	Plus grosse que les autres cellules
	Forme	Allongées de même épaisseur	Fines et de même épaisseur
	Composition	Plusieurs cellules	Une cellule allongée
	Nombre de branches principales	5	4-5 autour de la spore initiale
	Branches secondaires	Absence	Présence
Effet de la molécule antifongique		Homogène	Homogène
Phénotype		3	4
Spore initiale	Présence	Non identifiable	Visible
Spore initiale	Forme	–	Ovoïde
Spore initiale	Taille	–	Plus petite que celle du Phénotype 2
Branches	Forme	Épaisseur et longueur aléatoire	Fines de la même épaisseur
Branches	Composition	Cellules difformes	Une cellule allongée
Branches	Nombre de branches principales	Non identifiable	Nombre important
Branches	Branches secondaires	Absence	Présence
Effet de la molécule antifongique		Homogène	Hétérogène

TABLEAU 2.2 – Caractéristiques morphologiques identifiées décrivant les Phénotypes 1, 2, 3 et 4. Informations relatives à la spore initiale du champignon, aux branches (forme, composition et degrés des branches) et à l'homogénéité de l'effet de la molécule sur les champignons.

La première étape est la normalisation de nos images de sorte que toutes les images présentent la même plage de valeurs d'intensité des pixels. Ce processus permet aux méthodes appliquées sur des images différentes d'agir de la même manière. Dans ce projet, l'algorithme de normalisation permet de passer les images de 16 bits (valeurs de pixels de 0 à 65535) à 8 bits (valeurs de pixels de 0 à 255), de mettre les pixels du fond de l'image à un niveau 127 et d'étirer l'histogramme des valeurs des pixels. Les valeurs minimales et maximales sont misent vers 0 et 255 tout en conservant l'écart entre ces valeurs et 127. Pour cela nous nous sommes concentrés sur les valeurs de gradient de l'image. Comme le fond présente un caractère plutôt homogène, son gradient est globalement faible, contrairement à celui des champignons. Afin de calculer la valeur seuil nous permettant de distinguer les champignons du fond, nous calculons la dérivée de l'histogramme des valeurs de gradient à partir de son maximum et récupérons le point de pente maximal (qui est, idéalement, un point d'inflexion). L'intersection de la pente en ce point avec l'axe des abscisses correspond à la valeur seuil. Une fois la valeur seuil calculée, l'équation de normalisation est la suivant :

$$\mathrm{N}ew\mathrm{I}mage_{gradient} = \mathrm{I}mage_{gradient} - \mathrm{V}aleur_{seuil}$$

$$\mathrm{M}in = |\min_{\mathrm{N}ew\mathrm{I}mage_{gradient}}|$$

$$\mathrm{M}ax = \max_{\mathrm{N}ew\mathrm{I}mage_{gradient}}$$

$$\mathrm{I}mage_{normalisée} = 127 + (\frac{(\mathrm{N}ew\mathrm{I}mage_{gradient}) \times 127}{\max_{(Min,Max)}})$$

Nous nous sommes heurtés à deux principales difficultés lors du développement de la méthode de segmentation des objets de nos images. La première est que les images sont des images en lumière transmise et du fait de leur petites tailles, les champignons ne sont pas bien définis (voir la figure 2.17) et les niveaux de gris présents à l'intérieur des objets ne sont pas très différents de ceux du fond (voir la figure 2.18). L'application d'un ou plusieurs seuils sur la valeur des pixels de nos images ne permet donc pas une détection optimale des objets. La deuxième difficulté est l'hétérogénéité des différentes formes de champignon, qui se reflète dans les textures des objets (voir la figure 2.19). En effet un champignon qui a une forme compacte, présentera des valeurs de pixels globalement moins élevées que le fond. Un champignon dont la forme est plus linéique présentera des valeurs de pixels plus élevées que le fond, notamment aux extrémités de ces branches. Il est donc compliqué de développer une méthode automatique permettant une segmentation optimisée pour toutes ces formes. En revanche un des avantages de nos images est que les contours des objets bien que pas forcément continus, sont plutôt bien définis (voir les figures 2.17 et 2.19).

Deux méthodes de segmentation sont développées. Elles sont dénommées "Segmentation 1" et "Segmentation 2" et leurs étapes sont décrites ci-après. Pour la "Segmentation 1" : deux convolutions de l'image sont effectuées par un filtre gaussien puis par un filtre médian de façon successive afin de la lisser. Ces opérations nous permettent le calcul d'un seuil adaptatif par région de l'image. L'application de ce seuil, nous permet d'obtenir un premier masque binaire. Concernant la "Segmentation 2", dont les résultats obtenus à chaque étape sont illustrés sur la figure 2.20, elle repose sur l'utilisation du filtre de Canny sur nos images normalisées et dont les contrastes sont rehaussés afin de détecter les contours des objets. Un masque binaire des contours est donc obtenu.

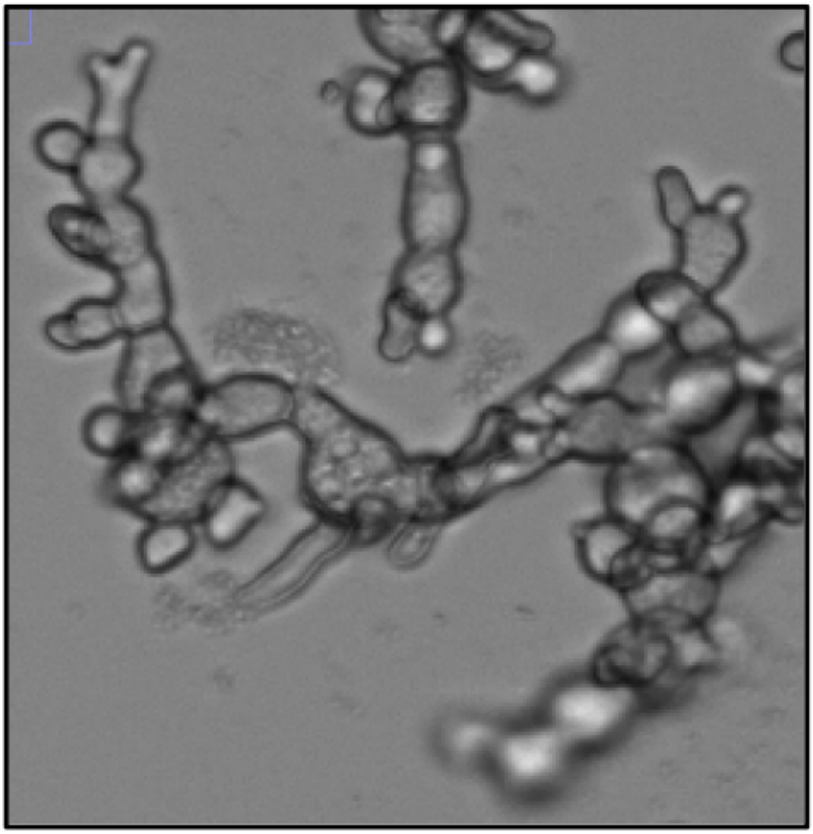


FIGURE 2.17 – Zoom sur les champignons d'une image. (Visualisation avec le logiciel ImageJ)

La présence des structures filiformes du mycelium formant un maillage qui lorsqu'elles se chevauchent se retrouve hors-focus (voir la figure 2.21) implique la nécessité de développer une approche particulière. L'identification des images mycelium est faite en appliquant un seuil sur la longueur totale du squelette des objets du masque binaire issu de la "Segmentation 1". S'il s'agit effectivement d'une image mycelium, le masque binaire issu de la "Segmentation 1" n'est pas conservé. Des opérations de morphologie mathématique (ouverture et fermeture) sont appliquées au masque binaire résultant de la "Segmentation 2". La taille et la géométrie des éléments structurants peuvent être fixées du fait de l'épaisseur constante des branches de cette forme du champignon. Ces opérations suffisent à reboucher les parties manquantes des objets et lisser les contours du masque final ainsi obtenu.
Concernant les formes de champignon autres que le mycelium, l'application du filtre de Canny nous aide à avoir une précision élevée de la détection des contours des objets. En revanche, ces différentes formes de champignon présentent des épaisseurs non constantes. L'utilisation des opérations de morphologie mathématiques pour le remplissage des contours est donc non adapté. En effet, le résultat est dépendant de la taille et la forme de l'élément structurant choisi. De plus, lorsque les branches d'un champignon (ou plusieurs) se chevauchent cela crée des parties de fond à l'intérieur d'un objet. Ainsi, dans le but d'obtenir une segmentation finale de nos objets qui soit la plus précise possible, nous avons décidé de combiner les deux approches de segmentation. La figure 2.22 illustre l'intérêt de cette combinaison. D'une part, l'intégralité de l'intérieur des contours détectés grâce à la "Segmentation 2" est comblé, nous donnant un premier masque binaire (colonne 2 de la figure 2.22) et d'autre part nous détenons le masque binaire produit par la "Segmentation 1" (colonne 1 de la figure 2.22). L'un présente des contours précis de

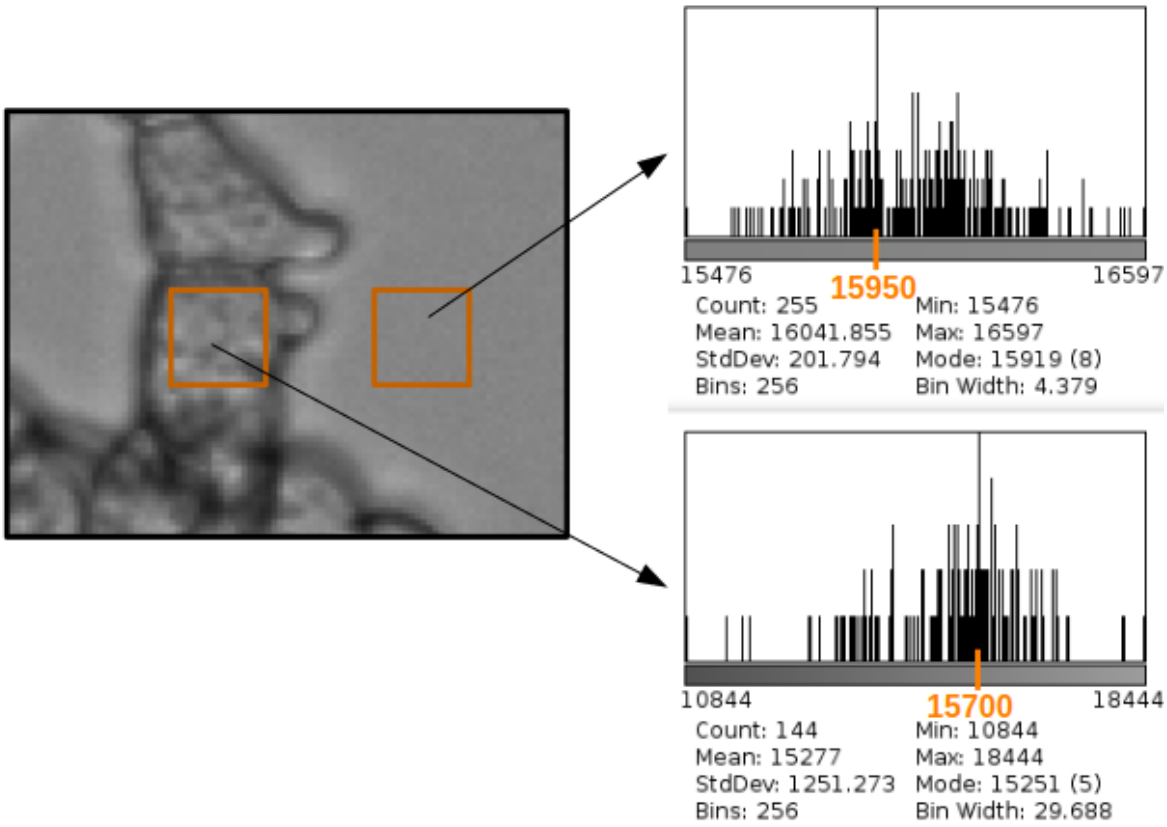


FIGURE 2.18 – Histogrammes de deux zones de l'image (zoomée x5) : plage à l'intérieur et à l'extérieur du champignon. Les histogrammes sont générés *via* le logiciel ImageJ.

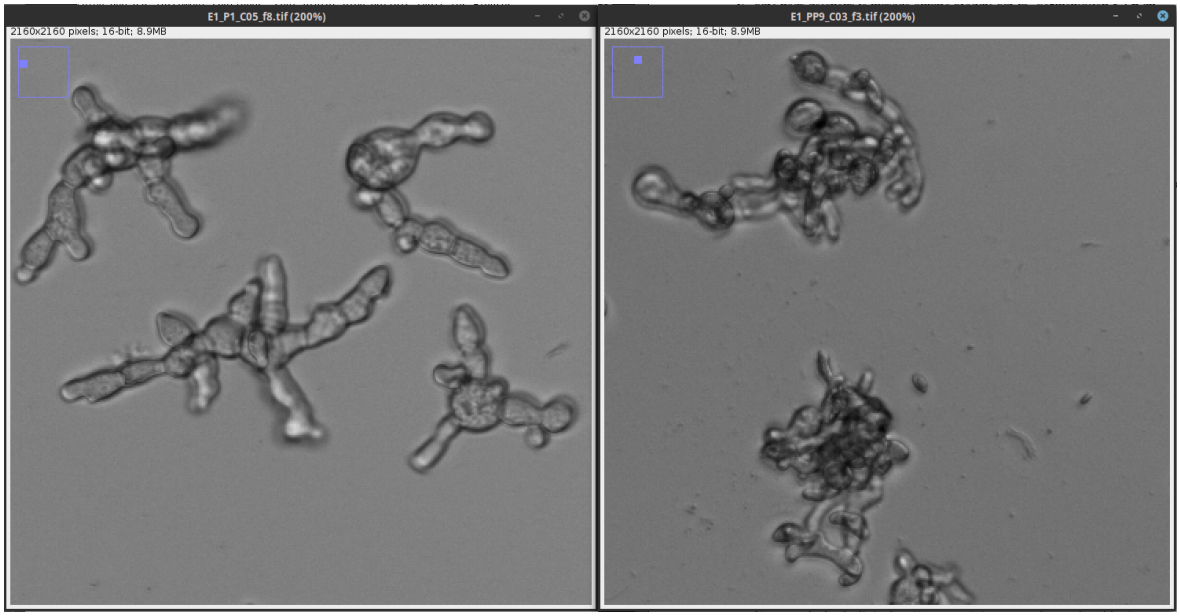


FIGURE 2.19 – Zoom sur deux différents phénotypes (x5) : A gauche les champignons présentent une forme moins compacte avec des valeurs globalement plus élevées que celles de l'image de droite (Visualisation avec le logiciel ImageJ).

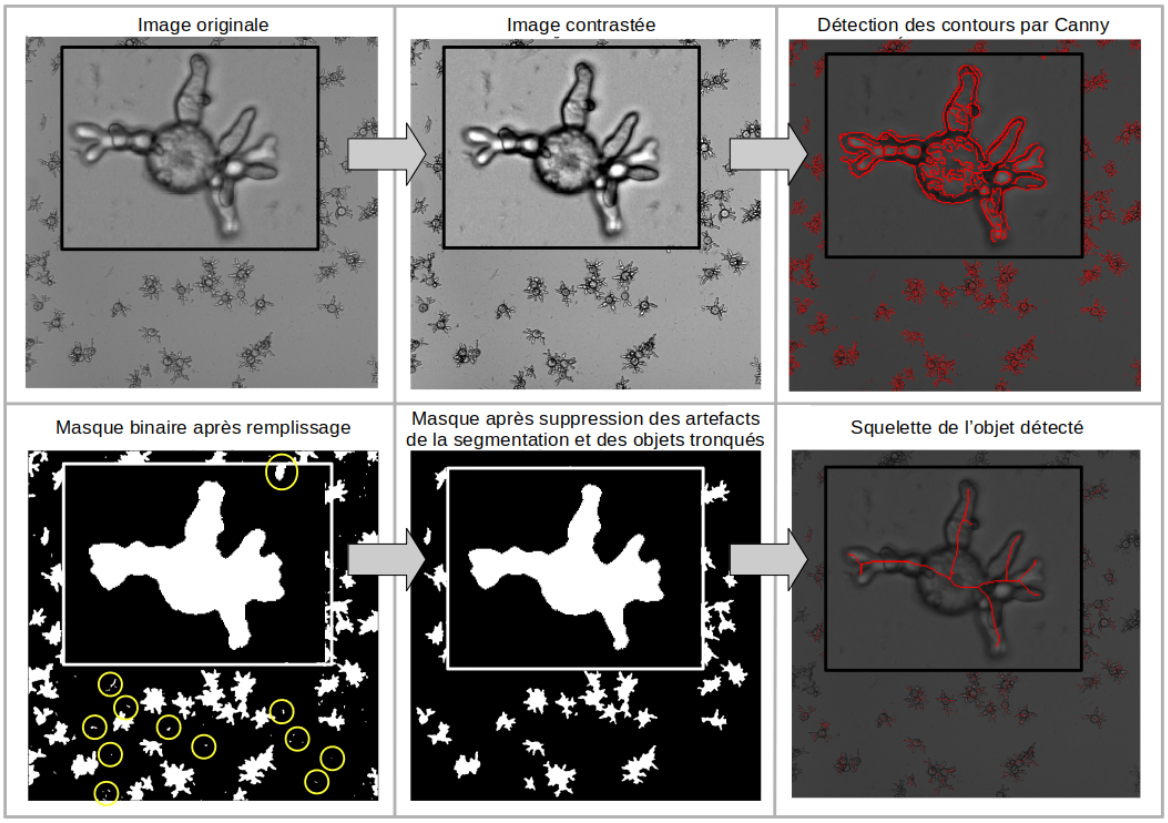


FIGURE 2.20 – Résultats issus des étapes de la méthode "segmentation 2" et du squelette calculé sur le masque binaire ainsi obtenu : les contours du champignon sont identifiés en appliquant le filtre de Canny sur nos images normalisées et dont les contrastes sont rehaussés. L'intérieur de l'objet est rempli et le masque est nettoyé en utilisant des opérations de morphologie mathématique. Les objets au bord de l'image sont également supprimés. Le squelette est calculé sur ce masque final.

nos objets et l'autre une détection des zones de fond ("trous") présents dans les objets.

La multiplication de ces deux masques nous permet l'obtention de la segmentation finale de nos objets (colonne 3 de la figure 2.22). Un exemple de résultats de cette segmentation pour chacun des phénotypes remarquables est présenté dans la ligne du haut de la figure 2.23.

A chaque étape les masques obtenus sont nettoyés avec des opérations de morphologie mathématique et les objets présents aux bords des images sont supprimés (sauf pour le cas mycelium qui forme quasiment un seul objet sur l'ensemble de l'image). En effet, la détection des objets permet le calcul des paramètres morphologiques qui serviront de descripteurs des différents phénotypes du champignon, ainsi, des champignons tronqués induiront des biais dans la mesure de ces paramètres. Le protocole de segmentation d'une image est illustré sur le schéma de la figure 2.24 suivant.

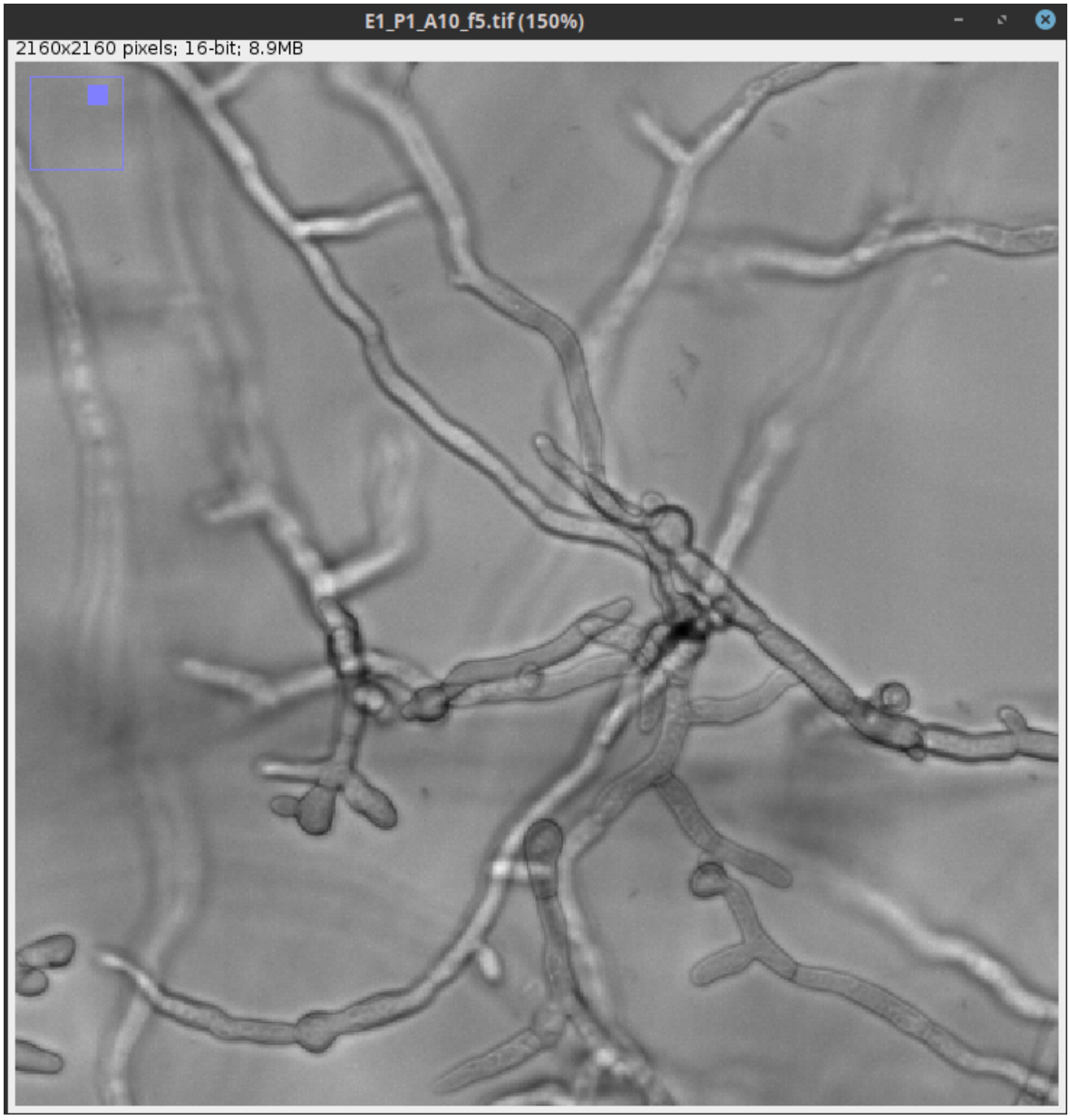


FIGURE 2.21 – Zoom (x4) sur une image de mycelium (Visualisation avec le logiciel ImageJ).

Extraction des paramètres morphologiques

La morphométrie, comme décrit paragraphe 2.2.4, est une discipline employée pour étudier et analyser la géométrie d'objets. Les données qu'elle permet d'extraire ont l'avantage de pouvoir faire l'objet de statistiques. Dans ce projet, la morphométrie est employée pour comparer différentes formes (phénotypes de champignon) entre elles de manière quantitative.

Squelettisation Les spores de *Botrytis cinerea* qui se développent en conidiophores sont caractérisées par des ramifications arborescentes. En observant les images de cellules de champignons obtenues après traitement antifongique, on remarque que le caractère arborescent est présent pour tous les phénotypes avec certaines particularités propre à chacun. Les cellules du phénotype 3 présentent par exemple assez clairement un défaut de polarité (voir la figure 2.16). Dans ce contexte, l'étude des squelettes des champignons nous a paru adaptée au regard de leur forme (la ligne du bas de la figure 2.23 montre les squelettes de champignons des phénotypes 1, 2, 3 et 4).

Transformation du squelette en graphe Le squelette (décrit paragraphe 2.2.4) permet d'extraire certains paramètres morphométriques. Dans cette optique, l'idée est de créer les graphes qui correspondent au squelette afin d'obtenir une forme plus structurée du squelette de l'objet et ainsi récupérer d'autres paramètres de formes. Le squelette est utilisé dans plusieurs travaux fondés sur des graphes. La construction du graphe passe par la classification des pixels de squelette en trois classes : les pixels de jonction (branchement), les pixels des branches et les pixels d'extrémité de branche.
Ces ensembles de points sont définis comme suit :

- Un pixel de jonction a au moins trois voisins dans le squelette. On note qu'un pixel de jonction fait souvent partie d'un petit groupe de pixels de jonction définissant une jonction entre trois branches ou plus.
- Un pixel de branche a exactement deux voisins dans le squelette.
- Un pixel d'extrémité a exactement un voisin dans le squelette.

Ainsi, dans un graphe de squelette G = (V, E), les jonctions et points d'extrémité forment l'ensemble des nœuds V et l'ensemble des points entre deux jonctions ou une jonction et un nœud d'extrémité correspondent aux arêtes du graphe dont l'ensemble est noté E. De plus, le graphe est enrichi par d'autres informations comme la position des nœuds, la longueur des branches. Ces informations seront utilisées lors du calcul des paramètres d'intérêts sur les squelettes.

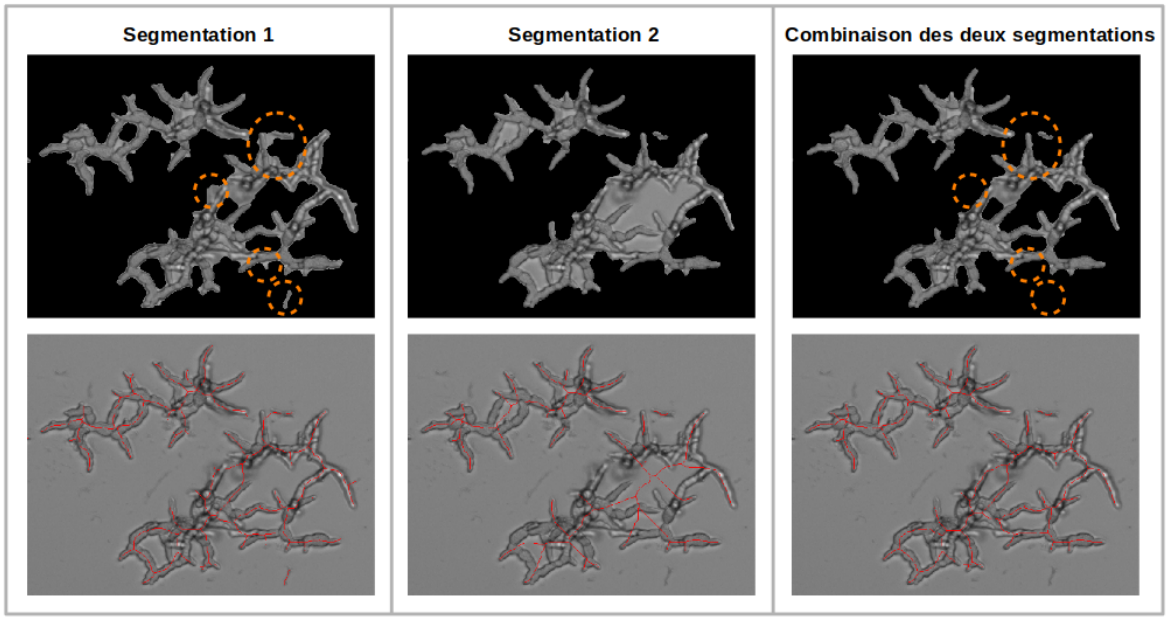


FIGURE 2.22 – Ligne du haut : résultat de la détection obtenu *via* les méthodes "Segmentation 1", "Segmentation 2" et la combinaison des deux. Ligne du bas : résultat de la squelettisation des objets dans chacun des cas. Les cercles en pointillés orange entourent des exemples d'amélioration du contour des objets.

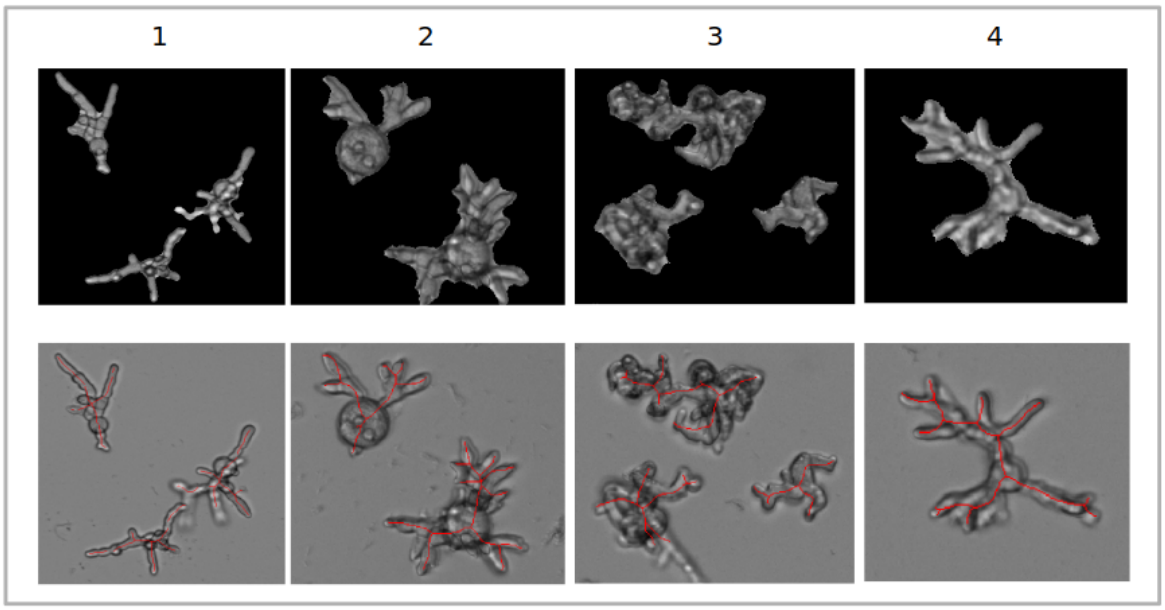


FIGURE 2.23 – Résultat de la détection (ligne du haut) et de la squelettisation des champignons (ligne du bas) de phénotypes remarquables (1, 2, 3 et 4).

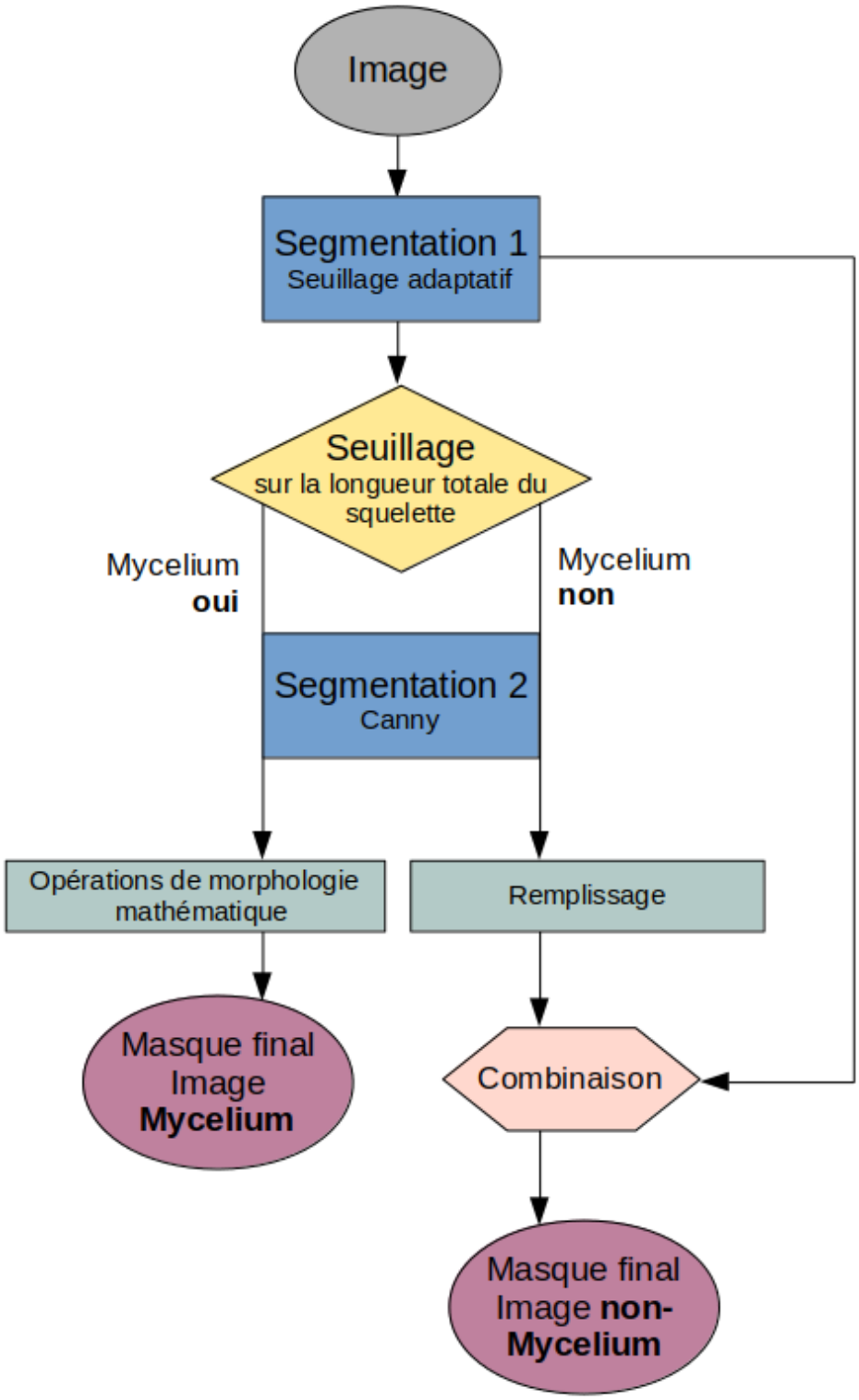


FIGURE 2.24 – Organigramme du procédé de segmentation d'une image. La méthode "Segmentation 1" est appliquée à l'image, suivant la valeur de la longueur totale du squelette qui en résulte, l'image est caractérisée comme appartenant à la classe mycelium ou non. La méthode "Segmentation 2" est également effectuée sur l'image afin d'obtenir un masque binaire des champignons qui y sont présents. Dans le cas d'une image caractérisée mycelium, ce masque binaire correspond à la segmentation finale de l'image. Dans le cas contraire, les deux masques résultants des segmentations "Segmentation 1" et "Segmentation 2" sont combinés pour obtenir le résultat final.

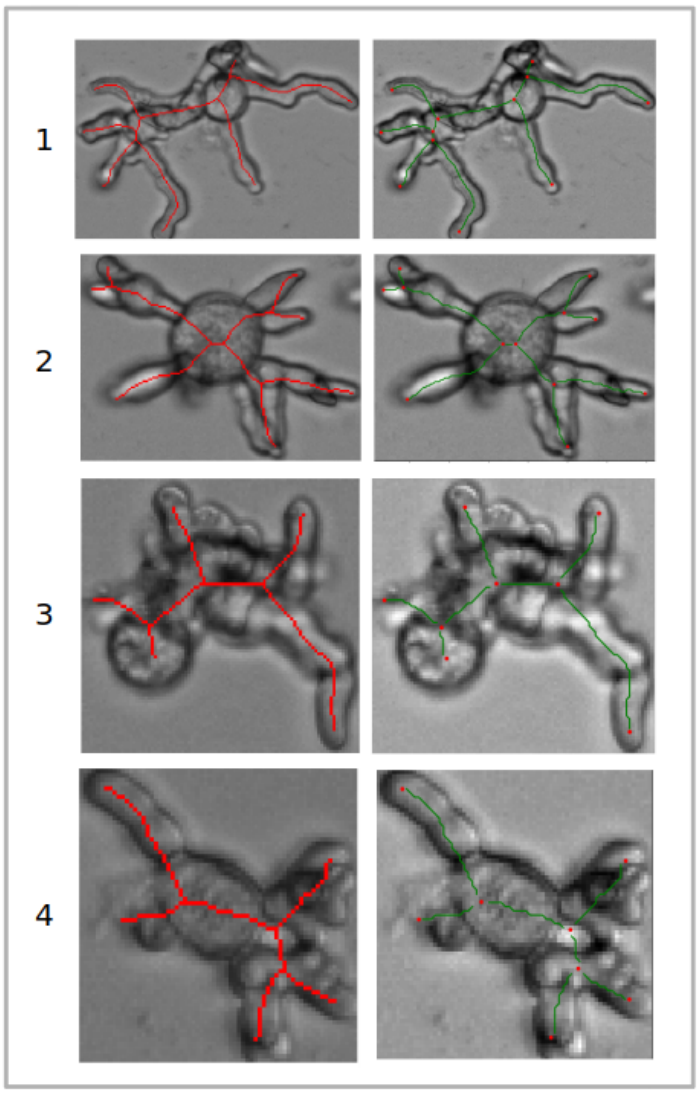


FIGURE 2.25 – Résultat des squelettes (première colonne) et graphes de ces squelettes (deuxième colonne) sur quatre champignons de différents phénotypes remarquables (1, 2, 3 et 4).

Paramètres morphologiques Pour apprendre un bon classifieur, il faut extraire des paramètres discriminants des objets à classer (ici les champignons). Il n'y a pas de paramètres universels. Nous avons imaginé ces paramètres à partir des connaissances des experts du domaine biologique et des experts en traitement d'images. Deux types de paramètres morphométriques sont retenus. D'une part des statistiques (au nombre de trois paramètres) calculées sur les paramètres de tous les objets d'une image tels que le nombre d'objets détectés et la longueur totale des squelettes présents à l'image. D'autre part sont extraits quatorze paramètres mesurés sur chacun des objets, comme par exemple, la longueur ainsi que le nombre de branches du squelette. Les paramètres extraits sont ceux présentés sur la figure 2.26. La figure 2.27 illustre sur un champignon schématique les notions de longueur, distance au bord et longueur pondérée par la distance aux bords d'une branche. Ces notions peuvent être étendues à l'ensemble du squelette. Cette figure présente également ce qu'est un nœud de jonctions et d'extrémités. Le paramètre nombre de nœuds correspond au nombre total des nœuds de jonctions et d'extrémités du graphe.

Image	Masque binaire	1) Nombre d'objets détectés à l'image,
	Squelette	2) Variance des longueurs des squelettes, 3) Longueur totale des squelettes,
Objet	Masque binaire	4) Aire de l'objet,
	Squelette	5) Longueur du squelette 6) Longueur du squelette pondérée par la largeur de l'objet 7) Moyenne de la distance aux bords de l'objet 8) Médiane de la distance aux bords de l'objet 9) Variance de la distance aux bords de l'objet
	Graphe	10) Nombre de nœuds 11) Nombre de nœuds d'extrémité 12) Nombre de nœuds de jonction 13) Longueur totale des branches pondérée par leurs largeurs 14) Médiane de la longueur des branches pondérée par leurs largeurs 15) Variance de la longueur des branches pondérée par leurs largeurs 16) Longueur de la branche la plus longue 17) Nombre de branches plus courte que la longueur moyenne des branches

FIGURE 2.26 – Liste des paramètres morphométriques calculés sur les trois formes des objets détectés à l'image : le masque binaire, le squelette et le graphe.

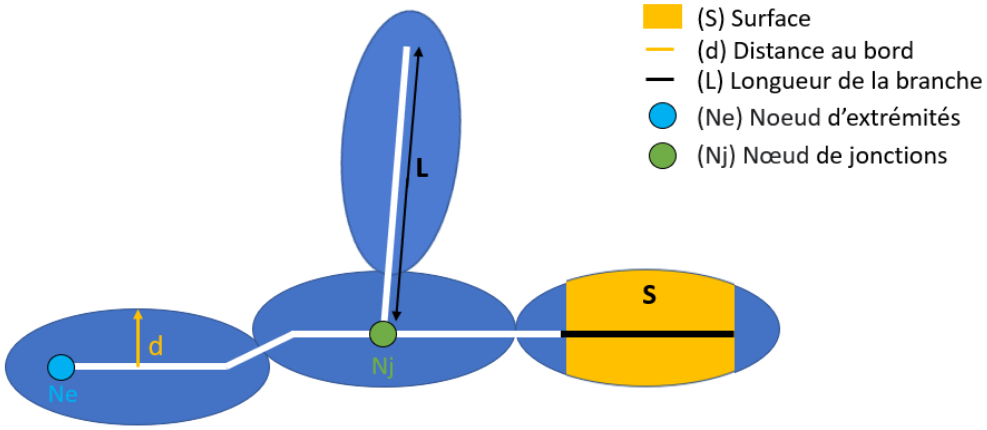


FIGURE 2.27 – Illustration sur un champignon schématique des notions de longueur (L), distance au bord (d), longueur pondérée par la distance au bord (S) d'une branche et des notions de nœuds de jonctions (Nj) et d'extrémités (Ne).

En travaillant sur le squelette des objets et sur le graphe correspondant (forme plus structurée du squelette), j'ai pu extraire des paramètres discriminants caractérisant les objets étudiés. Ces paramètres contiennent un grand nombre d'informations dont l'aire, la longueur et la largeur des objets, ainsi :

— La longueur du squelette pondérée par la distance aux bords de l'objet est une caractéristique très représentative de la surface de l'objet.

— Le calcul des moyennes des distances au bord le long des branches permet d'obtenir une mesure de la demi-épaisseur moyenne des branches du champignon.

— La variance des moyennes des distances au bord le long de chaque branche correspond à une mesure de la variation des demi-épaisseurs moyennes des branches du champignon.

Ces paramètres morphométriques sont utilisés comme descripteurs dans une méthode de classification supervisée permettant d'identifier les phénotypes d'une image dont la classe est inconnue.

Mise en place d'une base de données Dans ce projet, l'étude de chaque champignon de façon indépendante entraîne une grande quantité de données à stocker. Ainsi, une fois les images traitées, les champignons correctement détectés et les paramètres morphologiques les caractérisant extraits, nous avons décidé de créer une base de données regroupant toutes ces informations. Une base de données relationnelle permet de stocker un ensemble d'informations de manière structurée et avec le moins de redondance possible. La table est l'unité de base de cette structure. Chaque entrée d'une table est caractérisée par plusieurs renseignements distincts, appelés champs (ou attributs). De nombreux logiciels de gestion de bases de données existent tel que postgresql, mysql ou sqlite, c'est ce dernier que nous avons choisi d'utiliser. La base de données est structurée en deux tables, l'une contenant les caractéristiques de l'image entière et l'autre concernant les valeurs des paramètres de chacun des objets des images. Les deux tables sont reliées entre elles *via* l'identifiant de l'image correspondante.

Résultats, Analyses, Discussion

La démarche adoptée pour choisir ces caractéristiques repose sur le fait que les experts savent reconnaître les différents phénotypes. Les experts nous donnent alors des indications sur la manière dont ils les reconnaissent, des indications qui sont sous la forme de critères visuels. Puis, nous traduisons ces critères en caractéristiques morphométriques. Nous notons en revanche qu'une part importante de leur expérience est difficile a exprimer sous forme précise et objective. Afin de combler ce manque de critères objectifs, nous avons complété les caractéristiques morphométriques précédentes par des mesures géométriques. Cette démarche implique la vérification de certains points. Nous nous sommes alors posées les questions suivantes : La traduction des critères visuels donnés par les experts en caractéristiques morphométriques est-elle correcte ? Les mesures géométriques supplémentaires sont-elles pertinentes pour la classification des différentes formes de champignons ?

— **Évaluation de la pertinence des caractéristiques choisies**
La figure 2.28 permet d'avoir une idée des longueurs des squelettes des objets des différentes classes. En classant les phénotypes en fonction des médianes des longueurs des squelettes des moins au plus élevées, on retrouve dans l'ordre, Germination Inhibition puis les Phénotypes 3, 4, 1 et 2 et enfin la classe Crystal puis la

classe mycelium.
La figure 2.29 illustre le nombre de nœuds d'extrémités des squelettes des objets. Les champignons de la classe Germination-Inhibition présentent une médiane de deux nœuds d'extrémités et une faible longueur de squelette d'après la figure 2.28 indiquant une forme ovoïde des cellules de cette classe. De plus, ce paramètre est généralement lié avec la longueur du squelette. Dans le cas des phénotypes remarquables, nous pouvons noter que c'est le phénotype 2 qui présente globalement un nombre de nœuds d'extrémités plus important, suivit du phénotype 1 puis du phénotype 3 et enfin du 4. Cette figure met en évidence dans le cas du phénotype 2 de nombreuses valeurs aberrantes. Après vérification, il s'avère que les champignons de ce phénotype ont tendance à se chevaucher, ce qui se traduit par moins d'objets indépendants qui ont plus de nœuds d'extrémités. Nous pouvons également noter que certains objets de la classe phénotypes 4 ne présentent aucune branche (un seul nœud). Cela est explicable par la présence de cellules qui sont peu ou pas développées (certaines sont restés à l'état de spore), des formes souvent observées sur les images de cette classe.
La figure 2.30 représente les aires des objets de chacune des classes. Cette figure permet de nous rendre compte de l'hétérogénéité de l'aire des objets au sein d'une même classe, surtout dans le cas du phénotype 3. En effet la boîte à moustache pour ce paramètre affiche des valeurs aberrantes contrairement à celles des deux autres paramètres. C'est également le cas pour la classe Germination-Inhibition puisque ce paramètre s'étale quasiment sur l'intervalle qui s'étend de la valeur la plus basse à la valeur la plus élevée toutes classes confondues. Au contraire, le phénotype 2 ne présente qu'une seule valeur aberrante parmi la distribution des valeurs d'aire des objets.

En ce qui concerne la classe mycelium, l'ensemble de ces figures indique la présence d'objets de très grande, moyenne et petite taille. Les observations avec des valeurs de paramètres élevées reflètent en fait le réseau filamenteux correspondant aux cellules développées qui se chevauchent. Au contraire, les objets avec de faibles valeurs de paramètres correspondent à des artefacts de la segmentation. On remarque néanmoins une aire plus faible que celle des objets des classes phénotypes remarquables.
Dans le cas de la classe crystal, nous pouvons facilement nous rendre compte de l'hétérogénéité importante des valeurs des paramètres. Cette classe regroupant l'ensemble des formes de cristaux présents dans notre base de données, les résultats ne sont pas surprenants (voir le paragraphe A.5 en annexe).
La figure 2.31 représente le nombre d'objets détectés en fonction de la longueur totale des squelettes pour l'ensemble des classes exceptée la classe Crystal. Pour rappels, le nombre de spores dans les puits est calibré. Sans chevauchement, on devrait donc retrouver plus ou moins le même nombre d'objets dans chaque champ de chaque puits. De ce fait, cette figure montre, entre autres, la présence de chevauchements du fait du nombre très variable d'objets d'une image au sein des classes (par exemple ce nombre pour la classe Germination Inhibition varie entre moins de 40 et plus de 175). Pour illustrer ce problème, nous pouvons nous référer à la figure 2.22 : les champignons présents à l'image se chevauchent et après segmentation correspondent à deux objets seulement, alors qu'un expert en recense plus (au moins cinq mais il est difficile de connaître le nombre exact). Nous devinons donc clairement que les valeurs des paramètres morphologiques sont, dans ce cas-ci, erronées.

Globalement les valeurs de ces paramètres morphologiques extraits semblent cohérentes avec la réalité des observations des différentes formes de champignons. Néanmoins, il apparait au travers de ces mesures que le nombre de cellules par puits est, dans certain cas, trop élevé. Ce nombre favorise le chevauchement des cellules qui rend quasiment impossible la séparation des objets et biaise les valeurs des paramètres extraits. Nous pouvons également dire que, à la lumière de ces trois paramètres (voir les figures 2.28, 2.29 et 2.30) les échantillons des classes phénotypes 1 et 2 sont difficilement distinguables.

— **Importance des caractéristiques choisies**
La performance d'un système de classification dépend fortement des caractéristiques utilisées dans la tâche d'apprentissage. Elles peuvent être catégorisées comme étant pertinentes ou non pertinentes. Comme expliqué dans le paragraphe 2.2.2, les arbres d'une forêt aléatoire sont construits à partir d'échantillons bootstrap du jeux de données d'origine. Les échantillons de chaque arbre qui ne sont pas retenues par les bootstrap (dits Out-Of-Bag (OOB)) sont utilisés pour mesurer l'importance des variables. En effet, une variable est considérée comme importante si en cassant son lien avec la classe, l'erreur de prédiction augmente. Dans BREIMAN [2001], casser le lien entre la variable et la classe se traduit par permuter aléatoirement les valeurs de cette variable dans l'ensemble des échantillons OOB. L'importance de la variable correspond à l'augmentation moyenne de l'erreur de prédiction sur l'ensemble des arbres. Les caractéristiques peuvent donc être classées selon cette erreur de prédiction.

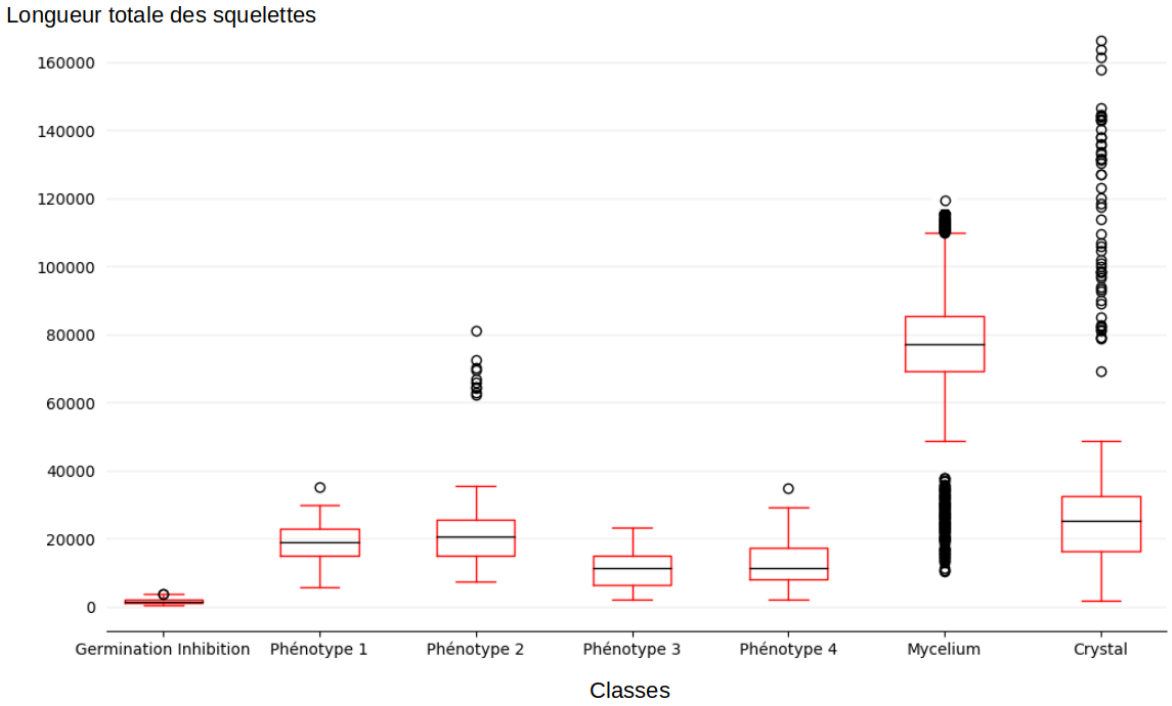


FIGURE 2.28 – Boîtes à moustaches représentant la distribution des longueurs totales des squelettes des objets détectés dans les images des différentes classes. Les cercles correspondent aux valeurs aberrantes. La médiane est représentée par la ligne dans la boîte.

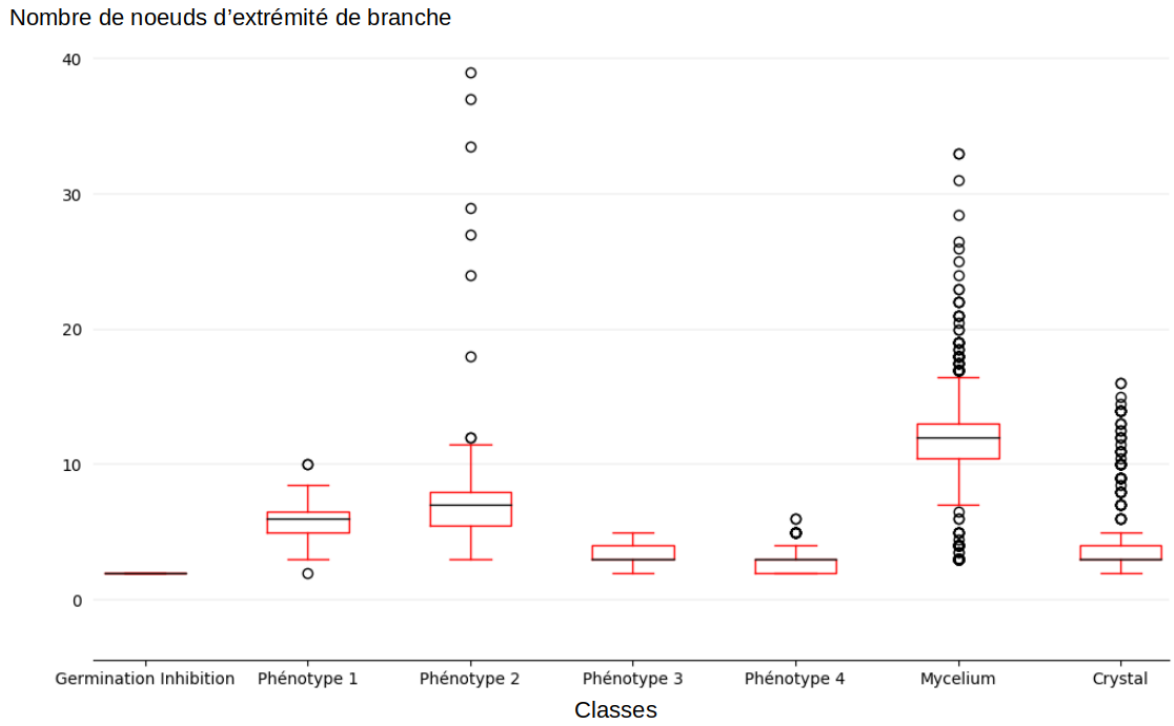


FIGURE 2.29 – Boîtes à moustaches présentant la dispersion du nombre de nœuds d'extrémités des squelettes des objets détectés dans les images des différentes classes. Les cercles correspondent aux valeurs aberrantes et la médiane est représentée par la ligne dans la boîte.

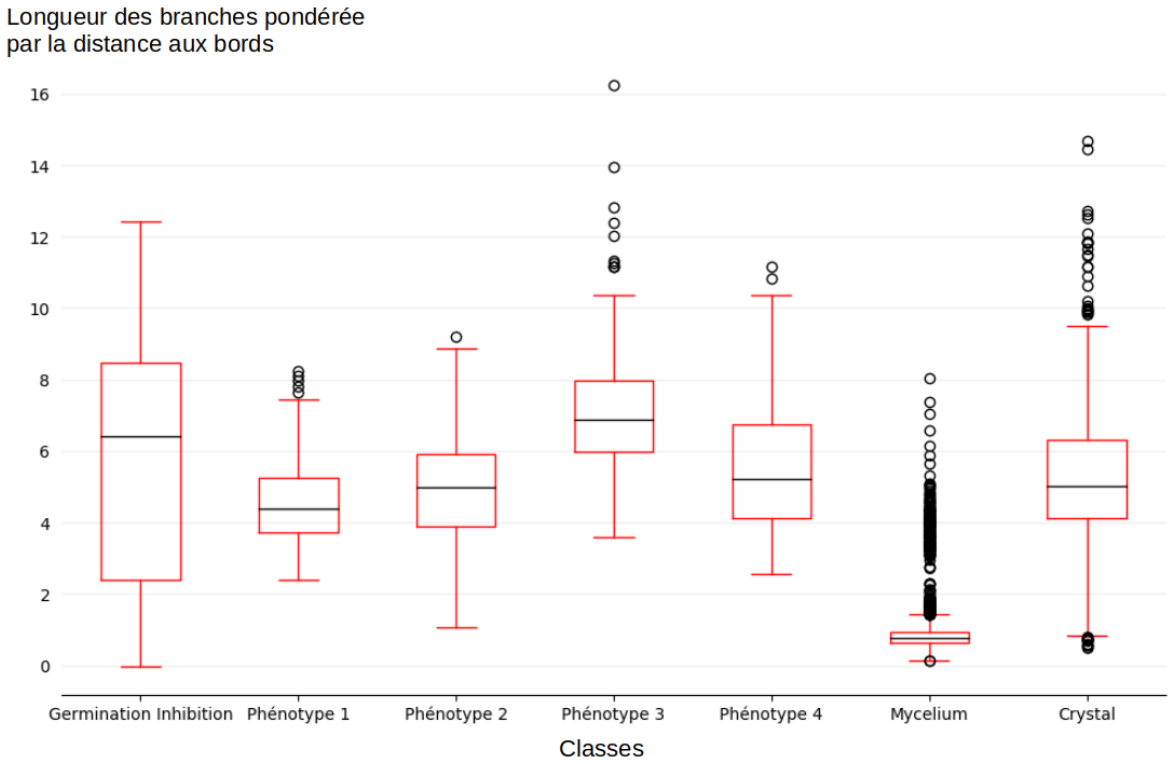


FIGURE 2.30 – Boîtes à moustaches illustrant la distribution des aires des objets détectés dans les images des différentes classes. Les cercles correspondent aux valeurs aberrantes. La médiane est représentée par la ligne dans la boîte.

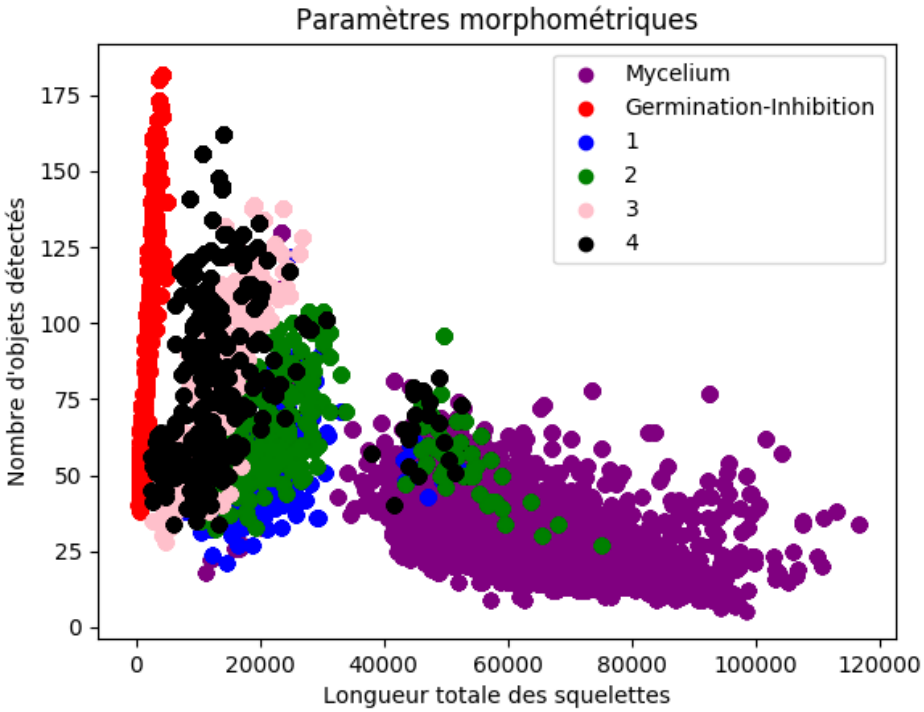


FIGURE 2.31 – Nombre d'objets détectés en fonction de la longueur totale du squelette des objets. Les échantillons correspondent aux images des classes Germination-Inhibition, Phénotypes 1, 2, 3 et 4 ainsi que celles de la classe Mycelium.

En pratique, nous avons utilisé l'attribut "feature importances" de la fonction RandomForestClassifier du module sklearn.ensemble de la librairie Scikit-learn de Python. Cette fonction propose une valeur par défaut du nombre de caractéristiques sélectionnées aléatoirement qui correspond à la racine carrée du nombre de variables. Les paramètres fixés sont : quatre caractéristiques ($\sqrt{17}$) par arbre et 500 arbres constituant la forêt. La figure 2.32 représente les poids/importances des paramètres morphologiques dans la bonne classification des images des sept classes. Cette représentation nous permet de nous rendre compte que parmi les dix-sept paramètres, trois présentent une pertinence bien plus élevée. Ces paramètres correspondent dans l'ordre du plus au moins pertinent, la "Longueur totale des squelettes" (2), le "Nombre d'objets détectés dans l'image" (0) et la "Variance des longueurs des squelettes" (1). Ces paramètres correspondent en fait aux trois paramètres statistiques calculés sur les caractéristiques de tous les objets d'une image. Sachant que le "Nombre d'objets détectés à l'image" reflète à la fois le nombre de champignons et leurs niveaux de chevauchements. Sur la figure 2.31, les échantillons de la classe Crystal ne sont pas affichés avec les autres, par souci de lisibilité. Cette figure présente clairement plusieurs clusters bien définis. Les quatre clusters identifiés correspondent aux classes Germination-Inhibition (cluster 1), aux classes 1 et 2 (cluster 2) et aux classes 3 et 4 (cluster 3) puis à la classe mycelium (cluster 4). Ces groupes sont différentiables sur la base de ces deux paramètres. Nous pouvons également voir apparaître des clusters intra-classes. En effet, on devine aisément que les classes 1,2 et 4 sont divisées en deux populations. Ces deux populations correspondent à deux stades de développement différents. La figure 2.33 donne un exemple de ces deux états pour chacune de ces classes. La ligne du haut correspond à l'état où les champignons sont moins développés et à l'inverse la ligne du bas illustre un état de développement plus avancé.

2.5.2 Deux méthodes de classification

Nous travaillons dans le cadre supervisé avec deux méthodes de classification, dont nous comparons les résultats. Dans cette partie, les méthodes des forêts aléatoires et de réseau de neurones convolutif sont décrites et les protocoles d'apprentissage et de prédiction de nos images sont exposés. Puis, le protocole permettant de conclure sur le mécanisme d'action des molécules testées à partir des prédictions des images y est expliqué. Enfin nous déduirons des résultats obtenus, le classifieur le plus adapté à notre problème.

Forêts aléatoires

Méthode La méthode des forêts aléatoires, comme décrite dans le paragraphe 2.2.2, est une méthode de classification supervisée qui effectue un apprentissage de multiples arbres de décision entraînés sur des sous-ensembles de données différents. Afin d'appliquer cette méthode sur nos images, les objets sont segmentés puis les paramètres morphologiques décrits dans le paragraphe 2.5.1 sont extraits des masques binaires, squelettes et graphes correspondants. Ces vecteurs de caractéristiques constituent les données représentatives de nos images qui sont utilisées pour construire les arbres de décision de la forêt aléatoire.

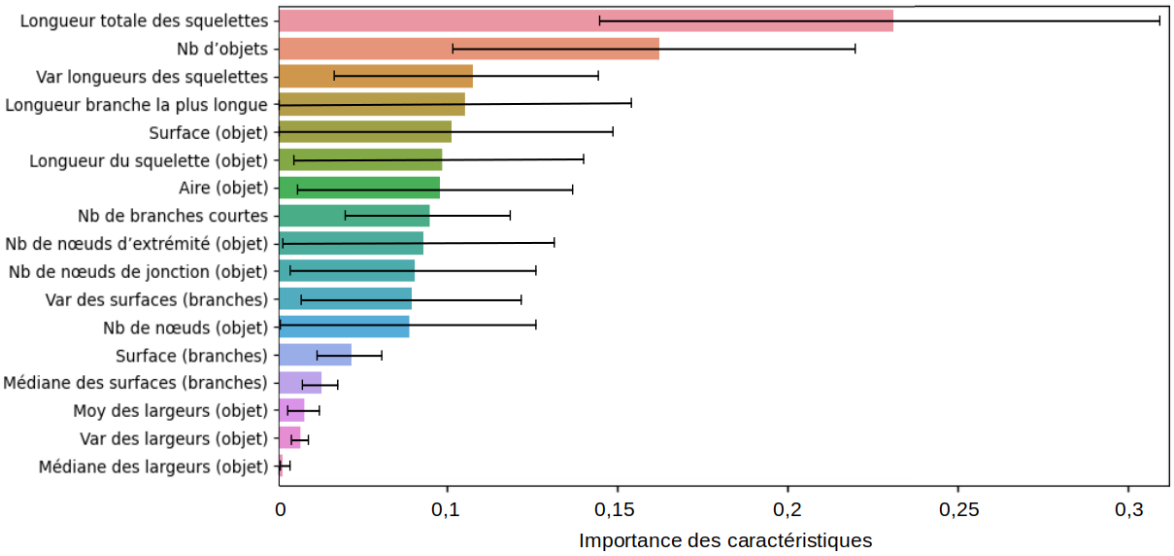


FIGURE 2.32 – Importance des paramètres morphométriques dans la classification des objets détectés des sept différentes classes. Les paramètres sont ordonnés du plus au moins important et leur variabilité inter-arbres (écart-type) est représentée par les barres noires. (*Annotations : Nb = Nombre, Var = Variance, Moy = Moyenne*)

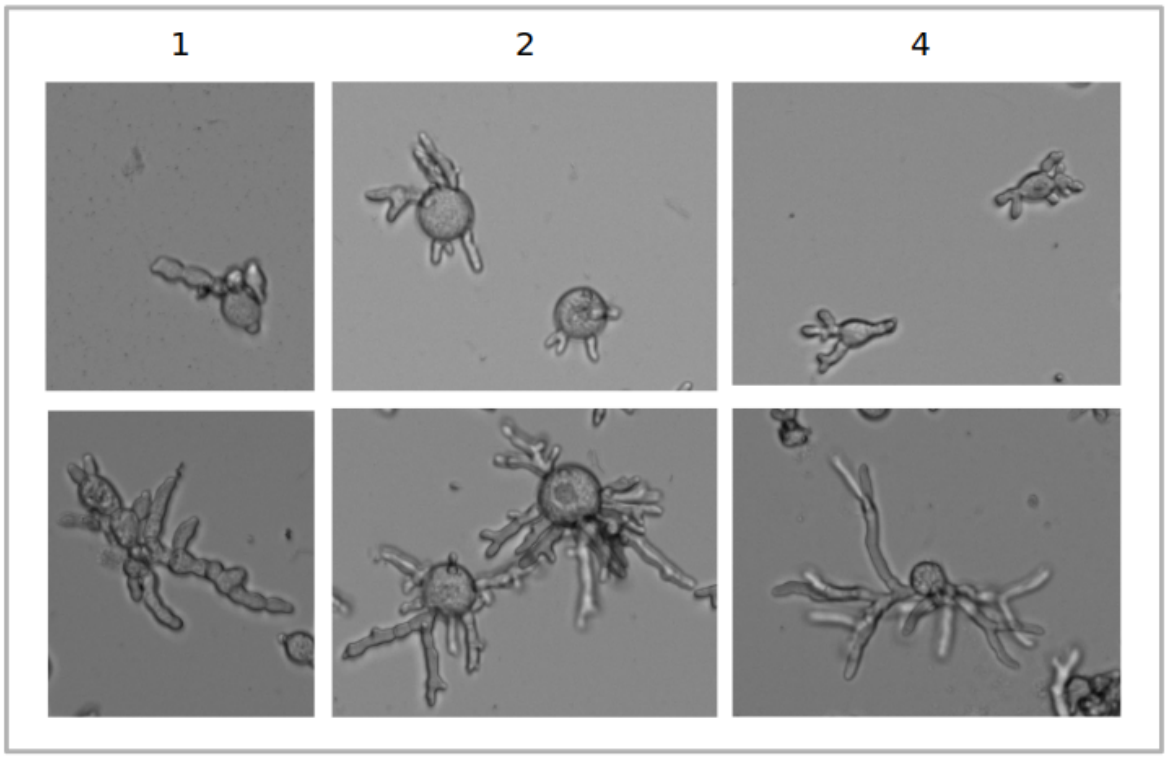


FIGURE 2.33 – Deux différents stades de développement des champignons des classes 1, 2 et 4.

Protocole de prédiction d'une image Le résultat de classification est obtenu selon deux méthodes que nous appellerons les méthodes de classification par "objets" et par "images". Dans l'approche par "objets", les échantillons à prédire sont les objets détectés à l'image. Les données correspondent donc à leurs vecteurs de caractéristiques. La prédiction finale de la classe d'une image est la classe majoritaire parmi les prédictions des objets qui y sont détectés. La deuxième approche, celle par "images", consiste à prédire directement la classe d'une image en prenant comme descripteur la médiane des valeurs des caractéristiques des objets qui y sont détectés. Les avantages de la méthode par "images" sont que l'on s'affranchit d'une part de l'hétérogénéité des phénotypes au sein d'une même image en la "moyennant" et d'autre part que les processus d'apprentissage et de prédiction sont moins longs en termes de temps de calcul. En revanche, la méthode par "objets" présente un intérêt certain en termes de prédiction beaucoup plus fine des différents phénotypes d'une même image.

Les questions que nous nous sommes posées ici sont les suivantes : les molécules peuvent impacter de façon homogène ou hétérogène les champignons d'un puits, ce critère peut-il influencer notre intérêt pour une molécule donnée? Est t-il possible qu'une molécule entraîne l'apparition d'un phénotype "hybride" ou même de deux différents phénotypes au sein du même puits?

Deux approches sont testées :

— *Première approche : Classification en une étape ("RF 1")*
L'idée dans ce cas, est de construire un système de classification capable de classifier les images en sept classes, les quatre phénotypes remarquables et les trois classes non-phénotypes, mycelium, Inhibition Germination et Crystal.

— *Deuxième approche : Classification en deux étapes ("RF 2")*
Cette approche est fondée sur la construction de deux classifieurs différents. Le premier est entrainé à distinguer les classes non-phénotypes et une seule classe appelée "Phénotype". La classe "Phénotype" regroupe toutes les images de phénotype remarquable sans distinction du type de phénotype (1, 2, 3 et 4). Le deuxième classifieur est entrainé à reconnaitre les images des classes phénotype 1, 2, 3 et 4. Le protocole de prédiction pour une image, dont la classe est à prédire, est de récupérer son vecteur de caractéristiques et d'appliquer dans un premier temps le premier classifieur. Dans le cas d'une réponse majoritaire de tous les arbres de la forêt correspondant à la classe "Phénotype", l'image passera par le deuxième classifieur dans le but d'affiner la classification et d'émettre une prédiction finale. Cette prédiction correspond à l'une des quatre classes phénotype remarquable. Concernant les images prédites Germination Inhibition, mycelium ou Crystal par le premier classifieur, la prédiction est définitive. Ainsi, toutes les images dont la classe inconnue est à prédire passeront par ce premier classifieur et seules les images prédites "Phénotype" passeront par le deuxième classifieur.

Les paramètres utilisés dans la construction de nos modèles de forêts aléatoires sont présentés dans le tableau 2.3.

Pour résumer, nous avons testé deux approches de classifications ("RF 1" et "RF 2") et en utilisant deux types d'échantillons ("objets" et "images"). Le but étant d'obtenir un score de classification des images de chaque classe dans chacun des cas afin d'en déduire la méthode la plus efficace.

	RF 1	RF 2
Nombre de caractéristiques	($\sqrt{17}$)	$\sqrt{3}$ (étape 1) et $\sqrt{14}$ (étape 2)
Nombre d'arbres de la forêt	500	200 (étape 1) et 500 (étape 2)

TABLEAU 2.3 – Valeurs des paramètres pour les classifications en une et deux étapes (décrites dans le paragraphe 2.5.2) : nombre d'arbres de la forêt et nombre de caractéristiques sélectionnées aléatoirement.

Indépendamment de la méthode, chaque prédiction s'accompagne d'un indice de confiance reflétant la probabilité d'appartenance de l'échantillon à la classe prédite. Cet indice sera un outil précieux pour l'expert scientifique afin d'évaluer la robustesse et comparer son expertise et la vérité terrain avec la prédiction de l'algorithme. Ainsi, dans le cas de la méthode des forêts décisionnelles, un indice de confiance est calculé en effectuant un pourcentage du nombre d'arbres ayant répondu la classe finale prédite majoritairement par la forêt. Dans le cas des prédictions de la classe des images par l'approche par "objets", l'indice de confiance d'une prédiction image correspond à la moyenne des indices de confiance des objets prédits comme appartenant à cette classe.

Réseau de neurone convolutif

Méthode Nous utilisons TensorFlow qui est une bibliothèque open source de Machine Learning, créée par Google, permettant de développer et d'exécuter des applications en Deep Learning. Cette bibliothèque permet notamment d'entraîner des réseaux de neurones pour, entre autres, la reconnaissance d'images. Nous avons testé plusieurs réseaux de neurones pré-entrainés sur la base d'images du concours ImageNet. L'évaluation de la précision de ces réseaux sur nos images, a permis de conclure que MobileNet HOWARD et collab. [2017] est plus efficace en termes de bonne prédiction des images des classes phénotypes 1, 2, 3 et 4.

Introduction à MobileNet :
Les réseaux de neurones et plus spécifiquement les CNN sont particulièrement appréciés pour la classification d'image, la détection d'objet et de visage. Cependant, comme décrit dans le paragraphe 2.2.2, ces réseaux de neurones effectuent des convolutions. L'application des filtres de convolution représente des opérations très coûteuses à la fois en terme de calcul et mais également en termes de mémoire. Du fait des contraintes matérielles, la classification d'image représente un enjeu de taille dans le domaine des systèmes embarqués. Ainsi, en avril 2017, Google a permis l'optimisation de la convolution *via* le développement de la Depthwise Separate Convolution. MobileNet HOWARD et collab. [2017] est un réseau spécialisé en vision par ordinateur dans lequel la Convolution est remplacée par une Depthwise Separable Convolution, rendant son exécution plus rapide et plus légère en termes de mémoire sans perte significative d'efficacité de prédiction. Cette nouvelle méthode de convolution est scindée en deux étapes (voir la figure 2.34), la Depthwise Convolution et la Pointwise convolution. La Depthwise Convolution, à la différence d'une convolution standard applique un filtre sur chaque canal plutôt que sur l'ensemble des canaux. La Pointwise Convolution combine les sorties de la Depthwise Convolution en utilisant N noyaux, obtenant ainsi une image de profondeur N.

Architecture de MobileNet :
Les MobileNets présentent une architecture composée de 28 couches dont 13 Depthwise Convolution et 13 Pointwise Convolution. Selon l'article HOWARD et collab. [2017], une "Depthwise Separate Convolution" de noyau 3×3 est environ 9 fois plus rapide qu'une opération de convolution régulière. Les couches constituants MobileNet sont, pour la plupart, des répétitions de "depthwise separable convolution", et 95% du calcul total est consacré à la "pointwise convolution" $1x1$ (voir la figure 2.34).
C'est un réseau de neurones pré-entrainé. Afin qu'il soit capable de classifier les images de nos différentes classes, la dernière couche est ré-entrainée sur nos données. La taille d'entrée du réseau est 224*224*3 (3 = couleur RGB) comme on peut le voir dans le tableau 2.35 et les images doivent être codées sur 8bits.

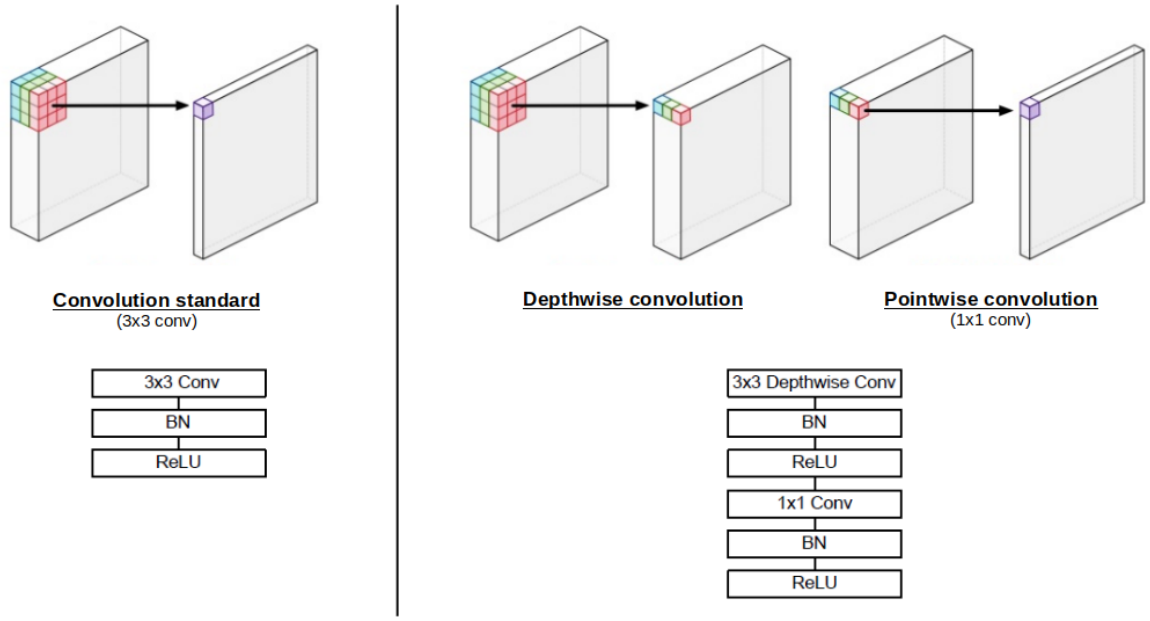


FIGURE 2.34 – Figure illustrant la différence entre une convolution standard dans les CNN et une Depthwise Separate Convolution : la depthwise et la pointwise convolution (Figures extraites de Mel [January 10th, 2018]), BN = "Batch Normalization" (Tableaux extraits de TSANG [Oct 14, 2018]).

Pré-traitement des images Dans l'optique d'adapter les images à la taille et au type de codage des images d'entrée du réseau de neurone MobileNet, nous avons fait un compromis entre le nombre de champignons conservés dans l'image et la perte d'information. En effet, redimensionner directement l'image de 2160*2160 en images de $224 * 224$ pixels entraînerait une perte conséquente d'informations. Ces informations étant essentielles aux calculs des descripteurs, nous avons préféré passer par la génération de plusieurs vignettes par image. Le traitement est effectué en trois étapes successives :

— Découper les images de $2160 * 2160$ en vignettes de $500 * 500$ pixels
— Redimensionner les images de $500 * 500$ à $224 * 224$ pixels.
— Convertir les vignettes en 8bits et répliquer les images de sorte à obtenir trois fois la même image en une, celle-ci sera alors d'une taille de $224 * 224 * 3$.

L'image après traitement correspond à 16 vignettes conformes aux normes d'entrée du réseau.

Type / Stride	Filter Shape	Input Size
Conv / s2	$3 \times 3 \times 3 \times 32$	$224 \times 224 \times 3$
Conv dw / s1	$3 \times 3 \times 32$ dw	$112 \times 112 \times 32$
Conv / s1	$1 \times 1 \times 32 \times 64$	$112 \times 112 \times 32$
Conv dw / s2	$3 \times 3 \times 64$ dw	$112 \times 112 \times 64$
Conv / s1	$1 \times 1 \times 64 \times 128$	$56 \times 56 \times 64$
Conv dw / s1	$3 \times 3 \times 128$ dw	$56 \times 56 \times 128$
Conv / s1	$1 \times 1 \times 128 \times 128$	$56 \times 56 \times 128$
Conv dw / s2	$3 \times 3 \times 128$ dw	$56 \times 56 \times 128$
Conv / s1	$1 \times 1 \times 128 \times 256$	$28 \times 28 \times 128$
Conv dw / s1	$3 \times 3 \times 256$ dw	$28 \times 28 \times 256$
Conv / s1	$1 \times 1 \times 256 \times 256$	$28 \times 28 \times 256$
Conv dw / s2	$3 \times 3 \times 256$ dw	$28 \times 28 \times 256$
Conv / s1	$1 \times 1 \times 256 \times 512$	$14 \times 14 \times 256$
$5\times$ Conv dw / s1	$3 \times 3 \times 512$ dw	$14 \times 14 \times 512$
Conv / s1	$1 \times 1 \times 512 \times 512$	$14 \times 14 \times 512$
Conv dw / s2	$3 \times 3 \times 512$ dw	$14 \times 14 \times 512$
Conv / s1	$1 \times 1 \times 512 \times 1024$	$7 \times 7 \times 512$
Conv dw / s2	$3 \times 3 \times 1024$ dw	$7 \times 7 \times 1024$
Conv / s1	$1 \times 1 \times 1024 \times 1024$	$7 \times 7 \times 1024$
Avg Pool / s1	Pool 7×7	$7 \times 7 \times 1024$
FC / s1	1024×1000	$1 \times 1 \times 1024$
Softmax / s1	Classifier	$1 \times 1 \times 1000$

FIGURE 2.35 – Architecture du réseau MobileNet (Tableau extrait de HOWARD et collab. [2017]).

Protocole de prédiction d'une image Une fois le réseau de neurones ré-entrainé sur les vignettes des sept classes (phénotypes et non-phénotypes), nous pouvons appliquer ce classifieur et récupérer pour chaque vignette testée, une prédiction. Dans le but d'attribuer une prédiction finale aux images, les prédictions de chacune des vignettes sont rassemblées. La classe prédite d'une image correspond à la réponse majoritaire des prédictions de toutes les vignettes correspondantes et son score de classification correspond à la moyenne des pourcentages d'appartenance des vignettes à la classe prédite par le réseau. On a donc un score sur la précision de classification des vignettes d'une part et d'autre part sur les images.

Règles de prédiction du MoA d'une molécule

Pour rappel, dans le but de prédire le mode d'action d'une molécule antifongique donnée, elle est testée à dix concentrations (de 100 à 0,05µM) et plusieurs images par puits sont acquises par microscopie à lumière transmise. Ces images, dans le cas de la classification suivant la méthode des forêts aléatoires, sont binarisées et les paramètres morphologiques des objets segmentés sont extraits et utilisés comme descripteurs. Donc dans le cas des forêts aléatoires, "l'unité" correspond à un objet et dans celui du CNN, "l'unité" correspond à une vignette. Dans les deux cas, les prédictions unitaires sont ramenées à des prédictions images. Ce protocole est décrit sur la figure 2.36.

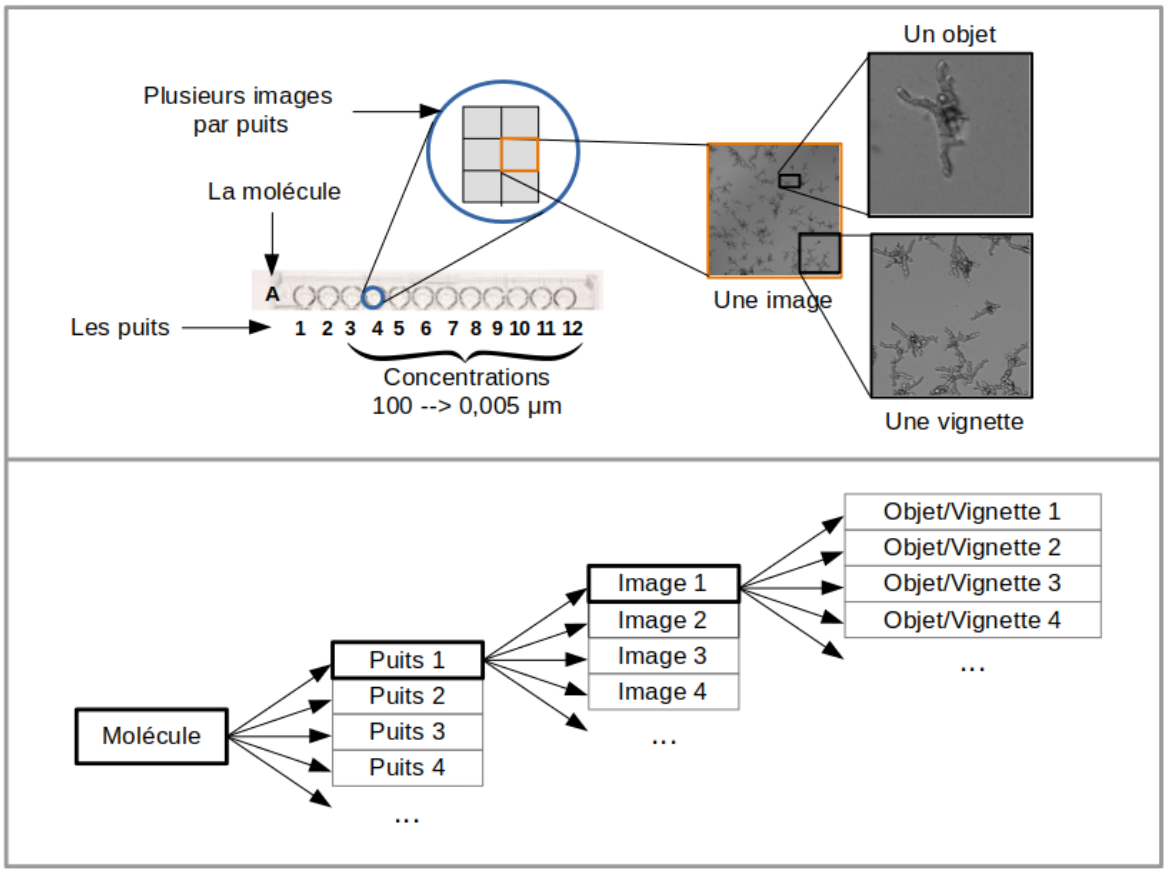


FIGURE 2.36 – Protocole de prédiction du MoA d'une molécule à partir des prédictions images, elles-mêmes obtenues à partir des prédictions objets ou vignettes, suivant la méthode de classification utilisée.

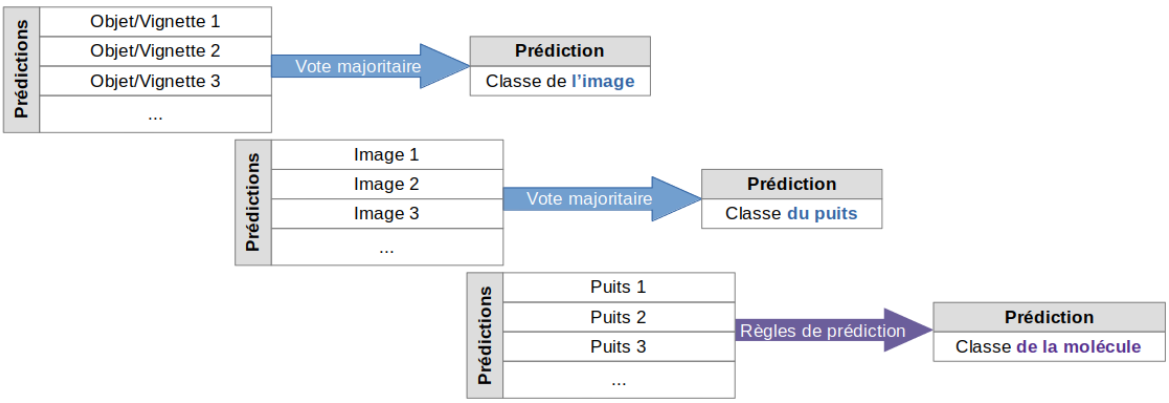


FIGURE 2.37 – Schéma de la cascade de prédictions.

La figure 2.37 permet de suivre les différents niveaux de prédiction. Pour chaque concentration, les prédictions des images du puits sont rassemblées et la réponse majoritaire correspond à la prédiction finale du puits. On obtient alors une classe prédite pour chacune des concentrations. Les prédictions correspondant à des classes non-phénotypes ne sont pas porteuses d'information permettant de conclure sur le MoA de la molécule testée. En revanche, les prédictions de l'une des classes Phénotype donne une idée du type de mécanisme d'action de la molécule. Ainsi, pour une molécule donnée :

- Dans le cas où tous les puits sont prédit mycelium, la conclusion à cette situation est que la molécule testée est inactive. En effet, des prédictions mycelium indiquent que la molécule n'a aucun effet sur le champignon.
- Dans le cas où les classes finales prédites correspondent uniquement à des classes Non-Phénotype, le mécanisme d'action de la molécule n'est pas identifiable mais des informations peuvent néanmoins être extraites. Les puits dont la prédiction finale est la classe Spore indiquent une forte activité de la molécule antifongique à ces concentrations. Les puits prédits Crystal signalent que la molécule à ces concentrations présentent des problèmes de solubilité dans le solvant utilisé.
- En ce qui concerne une molécule dont au moins un des puits est prédit Phénotype :
  - Dans le cas où les classes prédites sont les mêmes, le mode d'action de cette molécule correspond à cette classe de phénotype (exemple : Phénotype 1 est associé au MoA 1).
  - Dans le cas où les classes prédites sont différentes, les moyennes des scores de prédiction pour chacune des différentes classes sont calculées et comparées. La classe présentant le score moyen le plus élevé correspond à la prédiction finale du mode d'action de la molécule testée.

  Pour les molécules testées et prédites dans l'une des classes phénotype, l'indice de confiance de la prédiction finale du mode d'action équivaut à la moyenne des indices de confiance des puits prédisant la classe correspondante. Concernant les autres molécules (non prédites dans l'une des classes phénotype), comme aucune conclusion sur le mode d'action ne peut être émise, l'indice de confiance de la prédiction finale n'est donc pas calculé.

Vote majoritaire

Dans le cas d'une égalité entre deux ou plusieurs classes, lors d'un vote majoritaire sur un ensemble de prédictions, la moyenne des indices de confiance des prédictions de ces échantillons est calculée. La moyenne la plus élevée indique la classe finale prédite de l'échantillon.

2.5.3 Résultats de l'étude comparative des deux méthodes de classification

Les résultats énoncés ci-après sont les scores obtenus sur le résultat d'un test un contre tous. Les images sont regroupées en fonction de la molécule testée. L'ensemble d'apprentissage est constitué des images de toutes les molécules testées sauf une, les images acquises dans le cadre du test de cette molécule donnée composent l'ensemble de test. Les scores de classification sont normalisés par le nombre d'échantillons de chaque ensemble (Nombre d'échantillons bien prédit classe n * nombre d'échantillons classe n / nombre total d'échantillons).

Forêts aléatoires

Quatre types de résultats sont générés; les résultats provenant des classifications en une et deux étapes et les classifications dont l'unité à prédire est soit l'image soit l'objet.

- Classification "objets" : Comme décrit dans le paragraphe 2.5.1, chaque objet est représenté par le vecteur de paramètres morphologiques qui lui est propre. Les résultats sont présentés en termes de score de prédiction sur les images; la prédiction d'une image correspondant dans ce cas-ci au résultat d'un vote majoritaire sur les prédictions des classes des objets la constituant.

- Classification "images" : une image est représentée par la médiane des caractéristiques des objets qui y sont détectés.

Le tableau 2.4 présente les scores finaux calculés (moyennes des scores par classe) dans ces quatre situations. Ces scores nous permettent de conclure que la classification en termes d'images en prenant la médiane des caractéristiques comme descripteurs, permet une meilleure précision de classification. En effet, on obtient 81.4% et 77.3% (prédictions images) contre 71.8% et 66.5% (prédictions objets) pour la classification en une et deux étapes. Le fait de s'affranchir de l'hétérogénéité des phénotypes au sein d'une même image en la "moyennant" ainsi que du biais des mesures des caractéristiques dû aux chevauchements des objets permet une meilleure classification des différentes formes de champignons. De plus, les résultats démontrent que la classification des images en une seule étape permet une meilleure efficacité de prédiction avec une moyenne de 81.4% contre 77.3% pour celle en deux étapes.

Le tableau 2.5 présentent le détail des scores obtenus lors de la classification en une et deux étapes de la classification image. La première ligne correspond à la classification en quatre classes : Germination Inhibition, Crystal, mycelium (les classes non-phénotype) et la classe Phénotypes remarquables (incluant les phénotypes 1, 2, 3 et 4). Le score de chacune des classes y est inscrit. Ces résultats montrent que la classification en une seule étape permet l'obtention de meilleurs scores de prédiction des images pour chacune des sept différentes classes.

Enfin, le tableau 2.6 correspond à une matrice de confusion des pourcentages de classification obtenu dans chaque classe lors de la classification des images en une étape.

La comparaison de l'ensemble de ces résultats nous permet de conclure, dans le cas de la méthode des forêts aléatoires, que le meilleure score (81.3%) est obtenu en effectuant une classification des images en une étape.

Le tableau 2.7 illustre les résultats de classification des modes d'action des molécules testées. Ces résultats sont issus de l'application des règles de prédiction (voir le paragraphe 2.5.2) sur les classes prédites des images correspondantes. Ainsi, en utilisant la méthode des forêts aléatoires dans les conditions décrites plus haut, la moyenne d'efficacité de prédiction de ces différents modes d'action est de 84.5%.

Score de classification des images	Objets	Images
RF 1	71.8	81.4
RF 2	66.5	77.3

TABLEAU 2.4 – Résultats globaux de classification (en pourcentages) des images issus des classifications par "objets" (scores ramenés à un résultat de classification sur les images par vote majoritaire) et par "images" (médianes des valeurs des paramètres des objets).

Classes non-phénotype	G-I	Crystal	mycelium	Phénotypes 1, 2, 3 et 4
RF 2 ($1^{ère}$)	90	60	99	92.7
RF 1	99.3	60.5	99	–
Classes phénotype	1	2	3	4
RF 2 ($2^{ème}$)	73.7	57.3	86.5	74.9
RF 1	78.1	63.5	90.9	78.2

TABLEAU 2.5 – Résultats de classification en deux étapes (en pourcentages) des images. G-I correspond à Germination-Inhibition.

	1	2	3	4	G-I	Crystal	mycelium
1	78.1	15.2	1.3	1.8	0	3.5	0
2	14.6	63.5	4	8	0	3.1	6.7
3	1.3	0	90.9	4.3	0	3.4	0
4	0.8	5.76	7.6	78.2	0.4	6.6	2.4
G-I	0	0	0	0.7	99.3	0	0
Crystal	4.4	5	14.4	8.3	1.4	60.5	6.1
mycelium	0.14	0.33	0.1	0.22	0.09	0.1	99

TABLEAU 2.6 – Matrice de confusion des résultats de classification (en pourcentages) des images obtenus avec la méthode des forêts aléatoire. G-I correspond à Germination-Inhibition.

Mode d'action	1	2	3	4
Score	67.2	83.3	95	92.6

TABLEAU 2.7 – Résultats de classification (en pourcentages) des mécanismes d'action des molécules testées calculées suivant les règles de prédiction citées dans le paragraphe 2.5.2 *via* les résultats obtenues sur les images.

Réseau de neurone convolutif

Le tableau 2.8 présente les pourcentages de précisions de classification du réseau de neurones MobileNet pour les différentes classes. La ligne du haut correspond aux scores sur les vignettes et la ligne du bas aux scores des prédictions des classes des images. Comme décrit dans le paragraphe 2.5.2, les résultats sur les images sont obtenus par vote majoritaire des prédictions de la classe des seize vignettes correspondantes. Le score final de classification est d'environ 83% sur les vignettes ce qui amène donc à un score d'environ 94.5% sur les images. Ces valeurs démontrent que le vote majoritaire permet d'augmenter les résultats de classification.

Le tableau 2.9 correspond à une matrice de confusion des pourcentages de bonnes prédictions des classes des images.

Le tableau 2.10 indique les résultats de classification des modes d'action des molécules testées qui sont calculés *via* les règles de prédiction (le protocole est décrit dans le paragraphe 2.5.2). Cette classification *via* le réseau pré-entrainé MobileNet et ré-entrainé sur nos sept classes d'images, nous permet l'obtention d'une efficacité de prédiction de ces différents modes d'action de 100%.

	1	2	3	4	G-I	Crystal	mycelium
Vignette	78.5	86.3	73.2	73	96.3	73.9	99.5
Image	95	96	97.3	94.4	100	78.6	100

TABLEAU 2.8 – Résultats de classification en pourcentages par classes obtenus par le réseau de neurones MobileNet entrainé sur nos sept classes d'images. La première ligne correspond aux scores de prédiction sur la classe des vignettes. La seconde ligne fait état des résultats de classification des images qui sont obtenus par vote majoritaire des prédictions de la classe des vignettes correspondantes. (G-I correspond à Germination-Inhibition)

	1	2	3	4	G-I	Crystal	mycelium
1	95	0.5	0	0	0	4.5	0
2	0	96	0	2.2	0	1.8	0
3	2.7	0	97.3	2	0	0	0
4	0.8	2	1.6	94.4	1.2	0	0
G-I	0	0	0	0	100	0	0
Crystal	9.4	5	2.8	2.8	1.4	78.6	0
mycelium	0	0	0	0	0	0	100

TABLEAU 2.9 – Matrice de confusion des résultats de classification (en pourcentages) des images. (G-I correspond à Germination-Inhibition)

Mode d'action	1	2	3	4
Score	100	100	100	100

TABLEAU 2.10 – Résultats de classification (en pourcentages) des mécanismes d'action des molécules testées calculées suivant les règles de prédiction citées dans le paragraphe 2.5.2 *via* les résultats obtenues sur les images.

	1	2	3	4	G-I	Crystal	mycelium
RF	78.1	63.5	90.9	78.2	99.3	60.5	99
CNN	95	96	97.3	94.4	100	78.6	100

TABLEAU 2.11 – Récapitulatif des résultats de classification des images en pourcentages par classes obtenus par les deux méthodes de classification (RF et CNN).

2.5.4 Discussion

Au vu de ces résultats (voir le tableau 2.11), les deux méthodes sont capables de reconnaître les sept classes d'images, mais ceci avec des niveaux de précision différents. On obtient 81.4% (méthode des forêts aléatoires) contre 94.5% (méthode des réseaux de neurones) de bonne classification des images, soit une différence de 13.1% entre les deux méthodes. De plus, au vu des matrices de confusion 2.6 et 2.9, on se rend bien compte que dans le cas de la méthode des réseaux de neurones, les confusions se font entre deux ou trois autres classes, maximum, contrairement à celles observées dans le cas de la méthode des forêts aléatoires. En conclusion, outre une efficacité de prédiction supérieure, la méthode des réseaux de neurone semble engendrer des résultats plus homogènes.

De plus, le temps de calcul nécessaire à la prédiction d'une image est bien plus faible du fait de l'abscence de l'étape d'extraction des caractéristiques. En effet, la méthode des forêts aléatoires nécessite que l'image passe par une étape de segmentation, puis que les paramètres morphométriques de chacun des objets soient quantifiés ce qui induit un

temps de calcul plus ou moins conséquent en fonction du nombre et du volume des objets présents à l'image.

Un autre point à prendre en compte concerne les scores obtenus sur les images labellisées cristaux. Or les cristaux peuvent présenter des types très différents. Leurs tailles, leurs formes, leurs textures, les valeurs de niveau de gris les composant sont divers (voir le paragraphe A.5 en annexe). De ce fait, certains empêchent effectivement la bonne détection du phénotype à l'image en masquant les informations les concernant. D'autres en revanche, sont comparables à du bruit plus ou moins importants mais les champignons sont néanmoins toujours reconnaissables. La labellisation des images de cette classe ayant été réalisée suivant les limites de la méthode de segmentation nécessaire à la méthode des forêts aléatoires, toutes les images présentant des artefacts y ont été placées. L'analyse des résultats de prédictions de ces images, nous fait nous rendre compte de la précision du CNN. En effet, une bonne partie des mauvaises prédictions des images Crystal se sont avérées être en fait exactes au vu de la forme du champignon (voir le tableau 2.9). Au contraire, l'analyse des images labellisées Crystal et mal classées par la méthode des forêts aléatoires correspondent bien à des faux positifs. Une grande partie des images prédites phénotype 3 (14.4%) se rapporte en fait à des images phénotype 1 présentant des artéfacts. L'apprentissage de cette classe Crystal reste néanmoins nécessaire pour deux raisons. Cette classe permet d'une part d'éviter l'obtention de faux-positifs dans les autres classes et permet d'autre part à l'expert d'avoir une idée des puits où la molécule testée présente des problèmes de solubilité.

Pour finir, l'issue finale de ces prédictions images est d'émettre une hypothèse sur les modes d'action des molécules testées. Les résultats démontrent un écart de bonne précision de classification des MoAs de 15.5% (84.5% et 100% pour respectivement les forêts aléatoires et le réseau de neurone). Ainsi, pour toutes les raisons ici citées, la méthode conservée pour la suite du projet est la méthode des réseaux de neurones convolutionnels.

La méthode des forêts aléatoires, bien que moins efficace que le CNN, présente de bons scores de classifications avec 81.4% sur les images et 84.5% sur les MoAs. La différence de précision peut s'expliquer par la nature et le nombre des paramètres pris en compte dans la classification. Pour rappel, le nombre de ces paramètres est de dix-sept pour RF (contre mille-une pour le CNN) et de nature morphométrique. Ces paramètres ont un sens physique et correspondent à la quantification de caractéristiques morphologiques décrivant nos objets d'intérêts. Les masques binaires sur lesquels sont extraits et calculés les paramètres utilisés pour la discrimination des classes sont non-optimaux. En effet, comme expliqué précédemment, il est fréquent que les champignons se chevauchent et qu'ainsi les objets détectés correspondent à plusieurs champignons. Les valeurs extraites sont biaisées. De plus, l'ajout d'autres paramètres tels que des paramètres de texture par exemple ou des paramètres plus globaux pourraient nous permettre l'obtention d'un meilleur résultat de classification. L'écart entre les deux scores peut également s'expliquer par le fait que dans le cas du CNN la détection des descripteurs fait parti de l'étape d'apprentissage du classifieur. L'erreur de classification est minimisée dans le but d'optimiser entre autres les caractéristiques.

Dans le but d'améliorer ces masques binaires et les résultats de classification des phénotypes avec la méthode RF, nous avons mis en place deux stratégies :

— La première consiste à utiliser U-Net RONNEBERGER et collab. [2015], un réseau de neurones à convolution développé pour la segmentation d'images biomédicales afin de segmenter nos images. Nous avons travaillé sur l'optimisation des valeurs des paramètres utilisés par le réseau afin d'obtenir la segmentation la plus pré-

cise possible. L'ensemble d'apprentissage utilisé correspond aux images segmentées via notre méthode de segmentation mais sélectionnées pour leur exactitude. L'idée était d'entraîner le réseau en espérant que sur les champignons qui se chevauchent, la segmentation correspondrait à un objet par champignon. Les résultats n'ont malheureusement pas été ceux attendus.

— La deuxième méthode est une détection d'objets avec contraintes de forme (ellipses, bâtonnets) pour introduire de l'information a priori dans la segmentation des champignons (processus ponctuel marqué). Nous avons testé sur les images d'un phénotype (voir la figure 2.38) et nous essayons de trouver les meilleurs paramètres afin de détecter un maximum d'objets. Des informations pertinentes telles que le nombre de champignon par objet ainsi que des mesures relatives aux formes utilisées (comme par exemple le rayon des cercles) peuvent être extraites. L'idée est d'avoir d'autres paramètres décrivant nos différents phénotypes ainsi que corriger les paramètres initiaux (comme par exemple la longueur des objets qui sera divisée par le nombre de champignons détectés dans l'objet en question). À ce jour, l'algorithme développé n'a pu être généralisé à tous les phénotypes. En effet, il ne présente des résultats intéressants, de part la détection des spores initiaux, que pour le phénotype 2 (voir la figure 2.38). De plus, le temps de calcul trop élevé de l'algorithme nous a décidé à mettre de côté cette approche.

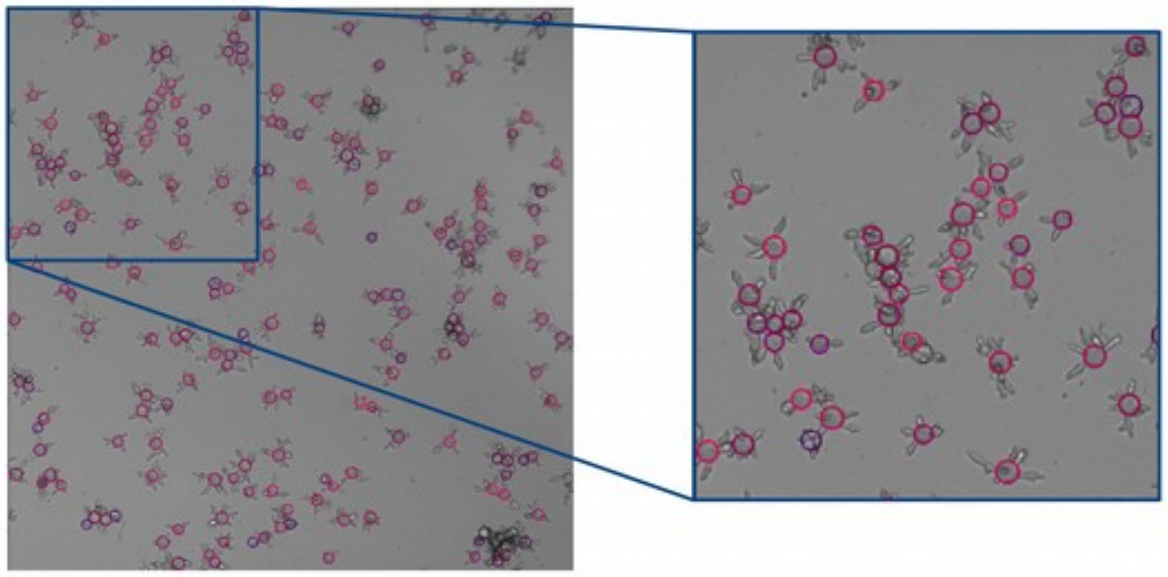


FIGURE 2.38 – Résultats préliminaires de shape matching : Étude du gradient de l'image afin de détecter les cercles de rayons prédéfinis correspondants aux spores initiaux de ce phénotype.

2.6 Références

January 10th, 2018, *Binary Classifier for Melanoma Using MobileNet*. URL https://jyu-theartofml.github.io/posts/melanoma. xiii, 64

ABBURU, S. et S. B. GOLLA. 2015, «Satellite image classification methods and techniques : A review», *International journal of computer applications*, vol. 119, n° 8. 26

AHA, D. W., D. KIBLER et M. K. ALBERT. 1991, «Instance-based learning algorithms», *Machine learning*, vol. 6, n° 1, p. 37–66. 25

AKHTAR, M. Z. S., S. A. PATEKAR et M. G. SOHANI. «A mobile application for quick classification of plant leaf based on color and shape», . 20

ALI, H., M. LALI, M. Z. NAWAZ, M. SHARIF et B. SALEEM. 2017, «Symptom based automated detection of citrus diseases using color histogram and textural descriptors», *Computers and Electronics in agriculture*, vol. 138, p. 92–104. 34

ANJNA, E. A. et R. K. ER. 2017, «Review of image segmentation technique», *International Journal of Advanced Research in Computer Science*, vol. 8, nº 4. 20

ATHIWARATKUN, B. et K. KANG. 2015, «Feature representation in convolutional neural networks», *arXiv preprint arXiv :1507.02313*. 36

BAY, H., A. ESS, T. TUYTELAARS et L. VAN GOOL. 2008, «Speeded-up robust features (surf)», *Computer vision and image understanding*, vol. 110, nº 3, p. 346–359. 34

BENNETT, K., A. DEMIRIZ et collab.. 1999, «Semi-supervised support vector machines», *Advances in Neural Information processing systems*, p. 368–374. 23

BENSON, E., J. REID et Q. ZHANG. 2003, «Machine vision-based guidance system for agricultural grain harvesters using cut-edge detection», *Biosystems Engineering*, vol. 86, nº 4, p. 389–398. 20

BLAYVAS, I., A. BRUCKSTEIN et R. KIMMEL. 2006, «Efficient computation of adaptive threshold surfaces for image binarization», *Pattern Recognition*, vol. 39, nº 1, p. 89–101. 20

BREIMAN, L. 2001, «Random forests», *Machine learning*, vol. 45, nº 1, p. 5–32. 57

BREIMAN, L. et R. IHAKA. 1984, *Nonlinear discriminant analysis via scaling and ACE*, Department of Statistics, University of California. 29, 35

CANNY, J. 1986, «A computational approach to edge detection», *IEEE Transactions on pattern analysis and machine intelligence*, , nº 6, p. 679–698. 20

CARPENTER, A. E., T. R. JONES, M. R. LAMPRECHT, C. CLARKE, I. H. KANG, O. FRIMAN, D. A. GUERTIN, J. H. CHANG, R. A. LINDQUIST, J. MOFFAT et collab.. 2006, «Cellprofiler : image analysis software for identifying and quantifying cell phenotypes», *Genome biology*, vol. 7, nº 10, p. 1–11. 37

CHANDA, B., M. K. KUNDU et Y. V. PADMAJA. 1998, «A multi-scale morphologic edge detector», *Pattern Recognition*, vol. 31, nº 10, p. 1469–1478. 18

CHEN, C.-H., H.-Y. KUNG et F.-J. HWANG. 2019, «Deep learning techniques for agronomy applications», . 26

DAO, D., A. N. FRASER, J. HUNG, V. LJOSA, S. SINGH et A. E. CARPENTER. 2016, «Cellprofiler analyst : interactive data exploration, analysis and classification of large biological image sets», *Bioinformatics*, vol. 32, nº 20, p. 3210–3212. 37

DERICHE, R. 1990, «Fast algorithms for low-level vision», *IEEE transactions on pattern analysis and machine intelligence*, vol. 12, nº 1, p. 78–87. 21

DEY, A. 2016, «Machine learning algorithms : a review», *International Journal of Computer Science and Information Technologies*, vol. 7, nº 3, p. 1174–1179. 21

DOLZ, J. 2016, *Towards automatic segmentation of the organs at risk in brain cancer context via a deep learning classification scheme*, thèse de doctorat. 26

FENG, S. et R. MANMATHA. 2005, «Classification models for historical manuscript recognition», dans *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*, IEEE, p. 528–532. 27

GAUCH, J. M. 1999, «Image segmentation and analysis via multiscale gradient watershed hierarchies», *IEEE transactions on image processing*, vol. 8, nº 1, p. 69–79. 18

GRAU, V., A. MEWES, M. ALCANIZ, R. KIKINIS et S. K. WARFIELD. 2004, «Improved watershed transform for medical image segmentation using prior information», *IEEE transactions on medical imaging*, vol. 23, nº 4, p. 447–458. 20

GRYS, B. T., D. S. LO, N. SAHIN, O. Z. KRAUS, Q. MORRIS, C. BOONE et B. J. ANDREWS. 2017, «Machine learning and computer vision approaches for phenotypic profiling», *Journal of Cell Biology*, vol. 216, nº 1, p. 65–71. 37

GUO, G., S. Z. LI et K. L. CHAN. 2001, «Support vector machines for face recognition», *Image and Vision computing*, vol. 19, nº 9-10, p. 631–638. 27

HARALICK, R. M., S. R. STERNBERG et X. ZHUANG. 1987, «Image analysis using mathematical morphology», *IEEE transactions on pattern analysis and machine intelligence*, , nº 4, p. 532–550. 17

HE, K., X. ZHANG, S. REN et J. SUN. 2016, «Deep residual learning for image recognition», dans *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 770–778. 36

HESTENES, M. R., E. STIEFEL et collab.. 1952, «Methods of conjugate gradients for solving linear systems», *Journal of research of the National Bureau of Standards*, vol. 49, nº 6, p. 409–436. 31

HO, T. K. 1995, «Random decision forests», dans *Proceedings of 3rd international conference on document analysis and recognition*, vol. 1, IEEE, p. 278–282. 29

HOWARD, A. G., M. ZHU, B. CHEN, D. KALENICHENKO, W. WANG, T. WEYAND, M. ANDREETTO et H. ADAM. 2017, «Mobilenets : Efficient convolutional neural networks for mobile vision applications», *arXiv preprint arXiv :1704.04861*. xiii, 63, 64, 65

IRSHAD, H., A. VEILLARD, L. ROUX et D. RACOCEANU. 2013, «Methods for nuclei detection, segmentation, and classification in digital histopathology : a review—current status and future potential», *IEEE reviews in biomedical engineering*, vol. 7, p. 97–114. 26

JAIN, R., R. KASTURI et B. G. SCHUNCK. 1995, *Machine vision*, vol. 5, McGraw-hill New York. 20

JAMIN, A. et A. HUMEAU-HEURTIER. 2020, «(multiscale) cross-entropy methods : A review», *Entropy*, vol. 22, nº 1, p. 45. 31

JIA, X. 2010, «Fabric defect detection based on open source computer vision library opencv», dans *2010 2nd International Conference on Signal Processing Systems*, vol. 1, IEEE, p. V1–342. 20

JOACHIMS, T. et collab.. 1999, «Transductive inference for text classification using support vector machines», dans *Icml*, vol. 99, p. 200–209. 23

KANUNGO, T., D. M. MOUNT, N. S. NETANYAHU, C. D. PIATKO, R. SILVERMAN et A. Y. WU. 2002, «An efficient k-means clustering algorithm : Analysis and implementation», *IEEE transactions on pattern analysis and machine intelligence*, vol. 24, nº 7, p. 881–892. 23

LASSOUAOUI, N., L. HAMAMI et N. NOUALI. 2005, «Morphological description of cervical cell images for the pathological recognition.», dans *WEC (5)*, Citeseer, p. 49–52. 34

LI, Z., K. WANG, L. LI et F.-Y. WANG. 2006, «A review on vision-based pedestrian detection for intelligent vehicles», dans *2006 IEEE International Conference on Vehicular Electronics and Safety*, IEEE, p. 57–62. 26

LIPPEVELD, M., C. KNILL, E. LADLOW, A. FULLER, L. J. MICHAELIS, Y. SAEYS, A. FILBY et D. PERALTA. 2020, «Classification of human white blood cells using machine learning for stain-free imaging flow cytometry», *Cytometry Part A*, vol. 97, nº 3, p. 308–319. 26

LOWE, D. G. 2004, «Object recognition from local scale-invariant features. int», *Journal of Computer Vision*, vol. 60, nº 2, p. 91–110. 34

LU, Y., L. ZHANG, B. WANG et J. YANG. 2014, «Feature ensemble learning based on sparse autoencoders for image classification», dans *2014 International Joint Conference on Neural Networks (IJCNN)*, IEEE, p. 1739–1745. 36

LUISIER, F., C. VONESCH, T. BLU et M. UNSER. 2009, «Fast haar-wavelet denoising of multidimensional fluorescence microscopy data», dans *2009 IEEE International Symposium on Biomedical Imaging : From Nano to Macro*, IEEE, p. 310–313. 17

MA, L., M. LI, X. MA, L. CHENG, P. DU et Y. LIU. 2017, «A review of supervised object-based land-cover image classification», *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 130, p. 277–293. 26

MARAGOS, P. 1987, «Tutorial on advances in morphological image processing and analysis», *Optical engineering*, vol. 26, nº 7, p. 267 623. 17

MARAGOS, P. 2005, «Morphological filtering for image enhancement and feature detection», *The Image and Video Processing Handbook*, p. 135–156. 18

MATHERON, G. et J. SERRA. 2002, «The birth of mathematical morphology», dans *Proc. 6th Intl. Symp. Mathematical Morphology*, Sydney, Australia, p. 1–16. 17

MIKOLAJCZYK, K. et C. SCHMID. 2005, «A performance evaluation of local descriptors», *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, nº 10, p. 1615–1630. 34

NEGISHI, T., S. NOGAMI et Y. OHYA. 2009, «Multidimensional quantification of subcellular morphology of saccharomyces cerevisiae using calmorph, the high-throughput image-processing program», *Journal of biotechnology*, vol. 141, nº 3-4, p. 109–117. 37

NIU, X.-X. et C. Y. SUEN. 2012, «A novel hybrid cnn–svm classifier for recognizing handwritten digits», *Pattern Recognition*, vol. 45, nº 4, p. 1318–1325. 36

OHTANI, M., A. SAKA, F. SANO, Y. OHYA et S. MORISHITA. 2004, «Development of image processing program for yeast cell morphology», *Journal of bioinformatics and computational biology*, vol. 1, nº 04, p. 695–709. 37

OHYA, Y., J. SESE, M. YUKAWA, F. SANO, Y. NAKATANI, T. L. SAITO, A. SAKA, T. FUKUDA, S. ISHIHARA, S. OKA et collab.. 2005, «High-dimensional and large-scale phenotyping of yeast mutants», *Proceedings of the National Academy of Sciences*, vol. 102, nº 52, p. 19 015–19 020. 37

PICCININI, F., T. BALASSA, A. SZKALISITY, C. MOLNAR, L. PAAVOLAINEN, K. KUJALA, K. BUZAS, M. SARAZOVA, V. PIETIAINEN, U. KUTAY et collab.. 2017, «Advanced cell classifier : user-friendly machine-learning-based software for discovering phenotypes in high-content imaging data», *Cell systems*, vol. 4, nº 6, p. 651–655. 37

QUINLAN, J. R. 1986, «Induction of decision trees», *Machine learning*, vol. 1, nº 1, p. 81–106. 29

RISH, I. et collab.. 2001, «An empirical study of the naive bayes classifier», dans *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, p. 41–46. 22

RONNEBERGER, O., P. FISCHER et T. BROX. 2015, «U-net : Convolutional networks for biomedical image segmentation», dans *International Conference on Medical image computing and computer-assisted intervention*, Springer, p. 234–241. 71

SALEMBIER, P. 1994, «Morphological multiscale segmentation for image coding», *Signal processing*, vol. 38, nº 3, p. 359–386. 18

SCHMID, C., R. MOHR et C. BAUCKHAGE. 2000, «Evaluation of interest point detectors», *International Journal of computer vision*, vol. 37, nº 2, p. 151–172. 34

SELVAKUMAR, P. et S. HARIGANESH. 2016, «The performance analysis of edge detection algorithms for image processing», dans *2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16)*, IEEE, p. 1–5. 20

SERRA, J. 1968, «Les structures gigognes : morphologie mathématique et interprétation métallogénique», *Mineralium Deposita*, vol. 3, nº 2, p. 135–154. 17

SHARIF RAZAVIAN, A., H. AZIZPOUR, J. SULLIVAN et S. CARLSSON. 2014, «Cnn features off-the-shelf : an astounding baseline for recognition», dans *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, p. 806–813. 36

SIMONYAN, K. et A. ZISSERMAN. 2014, «Very deep convolutional networks for large-scale image recognition», *arXiv preprint arXiv :1409.1556*. 36

SMITH, K., F. PICCININI, T. BALASSA, K. KOOS, T. DANKA, H. AZIZPOUR et P. HORVATH. 2018, «Phenotypic image analysis software tools for exploring and understanding big image data from cell-based assays», *Cell systems*, vol. 6, nº 6, p. 636–653. 37

SONG, X. et Y. NEUVO. 1993, «Robust edge detector based on morphological filters», *Pattern Recognition Letters*, vol. 14, nº 11, p. 889–894. 18

STALLKAMP, J., M. SCHLIPSING, J. SALMEN et C. IGEL. 2011, «The german traffic sign recognition benchmark : a multi-class classification competition», dans *The 2011 international joint conference on neural networks*, IEEE, p. 1453–1460. 26

SUDHARSHAN, P., C. PETITJEAN, F. SPANHOL, L. E. OLIVEIRA, L. HEUTTE et P. HONEINE. 2019, «Multiple instance learning for histopathological breast cancer image classification», *Expert Systems with Applications*, vol. 117, p. 103–111. 26

SZEGEDY, C., W. LIU, Y. JIA, P. SERMANET, S. REED, D. ANGUELOV, D. ERHAN, V. VANHOUCKE et A. RABINOVICH. 2015, «Going deeper with convolutions», dans *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 1–9. 36

TOTH, T., T. BALASSA, N. BARA, F. KOVACS, A. KRISTON, C. MOLNAR, L. HARACSKA, F. SUKOSD et P. HORVATH. 2018, «Environmental properties of cells improve machine learning-based phenotype recognition accuracy», *Scientific reports*, vol. 8, nº 1, p. 1–9. 37

TRIGUERO, I., S. GARCÍA et F. HERRERA. 2015, «Self-labeled techniques for semi-supervised learning : taxonomy, software and empirical study», *Knowledge and Information systems*, vol. 42, nº 2, p. 245–284. 23

TRISTAN-VEGA, A., S. AJA-FERNÁNDEZ et C.-F. WESTIN. 2012, «Deblurring of probabilistic odfs in quantitative diffusion mri», dans *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*, IEEE, p. 932–935. 17

TSANG, S.-H. Oct 14, 2018, *Review : MobileNetV1 — Depthwise Separable Convolution (Light Weight Model)*. URL https://towardsdatascience.com/review-mobilenetv1-depthwise-separable-convolution-light-weight-model-a382df364b6 xiii, 64

VAROL, E., B. GAONKAR, G. ERUS, R. SCHULTZ et C. DAVATZIKOS. 2012, «Feature ranking based nested support vector machine ensemble for medical image classification», dans *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*, IEEE, p. 146–149. 27

WANG, A., Y. WANG et Y. CHEN. 2019, «Hyperspectral image classification based on convolutional neural network and random forest», *Remote sensing letters*, vol. 10, nº 11, p. 1086–1094. 36

WANG, Z. et X. XUE. 2014, «Multi-class support vector machine», dans *Support vector machines applications*, Springer, p. 23–48. 29

WIKIPÉDIA. 2021, «Méthode agile — wikipédia, l'encyclopédie libre», URL http://fr.wikipedia.org/w/index.php?title=M%C3%A9thode_agile&oldid=183297963, [En ligne ; Page disponible le 27-mai-2021]. 42

WOLD, S., K. ESBENSEN et P. GELADI. 1987, «Principal component analysis», *Chemometrics and intelligent laboratory systems*, vol. 2, nº 1-3, p. 37–52. 22

YANG, P., Y. HWA YANG, B. B ZHOU et A. Y ZOMAYA. 2010, «A review of ensemble methods in bioinformatics», *Current Bioinformatics*, vol. 5, nº 4, p. 296–308. 23

ZHANG, T. et C. Y. SUEN. 1984, «A fast parallel algorithm for thinning digital patterns», *Communications of the ACM*, vol. 27, nº 3, p. 236–239. 35

Chapitre 3

Classification avec classe de rejet

Dans le chapitre 2, il était question de trouver la meilleure méthode de classification pour notre problème de reconnaissance des phénotypes connus du champignon *Botrytis cinerea*. Or, comme expliqué dans le paragraphe 1.2 du chapitre 1, l'identification de nouvelles molécules antifongiques avec des nouveaux Modes d'Action est essentielle pour la lutte contre ce champignon phytopathogène. Ainsi, la reconnaissance automatique de ces nouveaux phénotypes permettrait de rapidement détecter ces potentiels nouveaux MoAs. Dans ce chapitre, l'objectif est de développer une méthode de classification avec classe dite de rejet. En effet, les méthodes de classification supervisées présentent de grandes capacités de reconnaissance d'images. Le classifieur cherche à séparer l'espace des caractéristiques en optimisant les frontières entre les classes. Ces classes sont celles qui apparaissent dans l'ensemble d'apprentissage. Il s'agit uniquement des classes prédéfinies. Afin de pouvoir alerter l'utilisateur lors de l'identification d'un nouveau phénotype, c'est-à-dire ne faisant pas partie de l'ensemble d'apprentissage, nous avons développé une méthode de classification présentant une classe de rejet.
Le chapitre introduit dans un premier temps un état de l'art sur les méthodes de classification présentant une option de rejet qui existent dans la littérature. Puis, une description précise de la méthode avec classe de rejet, développée au cours de ce projet, est présentée. Cette méthode est appliquée sur des données synthétiques, permettant d'illustrer son fonctionnement et de la valider sur des cas "simples" de classification. Enfin, les résultats obtenus sur nos images de phénotypes connus et nouveaux de champignons, sont énoncés et discutés.

3.1 État de l'art

En classification supervisée, la précision de classification des observations dans les classes de l'ensemble d'apprentissage est optimisée lors de la phase d'apprentissage du classifieur. Ainsi, les échantillons à classer ne peuvent être prédits que comme appartenant à l'une des classes sur lesquelles on a entraîné le classifieur. Dans de nombreux cas pratiques, cette approche est suffisante car l'ensemble des classes possibles est connu et peut être représenté par un nombre suffisant d'échantillons permettant l'apprentissage des classes en question. Malgré tout, il existe des cas où seul un sous-ensemble des classes possibles est connu à un instant donné car le phénomène étudié peut évoluer et faire émerger de nouvelles classes. Il est alors nécessaire de pouvoir différencier les observations des classes connues et apprises de celles qui ne le sont pas. Par exemple, dans certains contextes médicaux, l'expert sait à l'avance que les données seront réparties entre la classe saine et un nombre donné de classes pathologiques SIEGISMUND et collab. [2018].

En revanche, dans certaines situations le processus d'apprentissage est sciemment exécuté en utilisant un ensemble d'apprentissage qui ne représente que les données qui ont été observées à ce jour parmi une variété d'échantillons possibles. Dans ce cas, il est important de reconnaître lors de la prédiction s'il s'agit d'un échantillon dont la classe est connue ou non.
Ce principe se retrouve représenté sous plusieurs appellations, dans la littérature telles que détection des données aberrantes, détection de nouveautés, classification avec stratégie ou option de rejet et détection de conditions anormales. Dans de nombreux travaux, repérer des cas dissemblables aux éléments de l'ensemble d'apprentissage est essentiel. L'ensemble des domaines concernés est très vaste avec par exemple, les domaines tels que la santé, la sécurité (informatique ou électronique), la conduite automobile, la robotique ou encore l'astronomie.

La notion de classe de rejet a été introduite en 1957 par Chow CHOW [1957]. Dans CHOW [1970], il optimise l'erreur dans un système de reconnaissance sur la base d'une règle de rejet optimale. Sur la base des probabilités a posteriori d'un système de reconnaissance optimal de Bayes. L'échantillon est rejeté si : $P(w_i) < t$ pour les classes w_1 à w_n où $P(w_i)$ est la probabilité d'appartenance à la classe w_i et t un seuil.

Dans le cas où l'ensemble des classes est connu et représenté, l'échantillon est appelé outlier. Dans le cas contraire, l'échantillon est considéré comme appartenant à une nouvelle classe. On parle alors de distance de rejet DUBUISSON et MASSON [1993] : les observations sont dites trop éloignées de l'ensemble d'apprentissage pour appartenir à l'une des classes connues. Cette première catégorie de méthodes de classification avec classe de rejet correspond aux modèles de prédiction classiques auxquels un mécanisme de sélection est ajouté. C'est à cette catégorie qu'appartient la méthode de Chow CHOW [1970] mais également les nombreuses méthodes fondées sur l'application d'un ou plusieurs seuils sur un système de classification GEIFMAN et EL-YANIV [2017]. L'une des difficultés majeures de ce type d'approches est la détermination du ou des seuils appropriés. Certaines approches peuvent également inclure des méthodes de clustering pour la détection des échantillons à rejeter TARASSENKO et collab. [2006]. Les algorithmes existants peuvent être globalement regroupés en cinq catégories, selon la revue PIMENTEL et collab. [2014], et diffèrent principalement de par les hypothèses faites sur la nature des données d'apprentissage. En effet, la performance de tels outils est très dépendante des propriétés statistiques des données d'entrées. On retrouve ainsi parmi les méthodes apparues dans la littérature :

— **Les méthodes fondées sur une distance :** l'hypothèse est que les données des classes connues sont regroupées alors que les échantillons à rejeter sont loin de leurs plus proches voisins issus de l'ensemble d'apprentissage. Parmi ces méthodes, on retrouve les approches qui s'appuient sur les voisins les plus proches (k-NN), des approches très couramment utilisées pour la détection de nouveauté HENRION et collab. [2013]. La méthode des k plus proches voisins consiste à prendre en compte les k échantillons d'apprentissage les plus proches d'un nouvel échantillon x, selon une métrique à définir afin d'émettre une hypothèse sur la classe potentielle de x. Dans ce cas de figure, l'ajout d'un mode de sélection se traduit par un seuil sur la distance de ce nouvel échantillon par rapport aux éléments de l'apprentissage. Ainsi les échantillons trop éloignés sont rejetés HELLMAN [1970]. La distance euclidienne est la métrique naturelle mais d'autres mesures existent. Comme par exemple la distance de Mahalanobis. Outre la distance au k-ème plus proche voisin SUGIYAMA et BORGWARDT [2013], on peut utiliser la distance à la moyenne des k plus proches

voisins. Dans le cas du SVM, il peut s'agir par exemple, d'un seuil sur la distance du nouvel échantillon à la frontière de marge maximale comme dans MUKHERJEE et collab. [1999].
Les méthodes fondées sur une distance incluent également les approches fondées sur le clustering tels que le clustering k-means TARASSENKO et collab. [2006]. Dans ce type de méthodes, les classes connues sont caractérisées par un petit nombre d'échantillons de référence dans l'espace des données. La distance minimale entre un échantillon à prédire et la référence la plus proche est souvent utilisée pour quantifier le degré d'appartenance de l'échantillon aux classes connues.

— **Les méthodes fondées sur un domaine :** Dans cette partie, seule une classe est considérée, la classe cible. Généralement insensibles à l'échantillonnage et à la densité spécifique de la classe cible, les méthodes fondées sur un domaine décrivent la limite de la classe cible (appelée aussi domaine). Dans ce cas, la position de l'échantillon par rapport à la limite définie au préalable définit son appartenance à la classe en question. Deux exemples très connus dans la littérature sont le SVM à une classe RATSCH et collab. [2002] et le SVDD ZHAO et collab. [2013] qui définit la limite comme étant l'hypersphère avec un volume minimum qui englobe un maximum des données d'apprentissage de la classe. L'extension à plusieurs classes est directe sauf si les domaines de deux classes s'intersectent. Dans ce cas, de multiples critères peuvent être utilisés pour prendre une décision.

— **Les approches issues d'une reconstruction :** les méthodes fondées sur la reconstruction peuvent modéliser les données sous-jacentes. L'erreur de reconstruction peut être calculée en effectuant la différence entre le vecteur d'entrée de l'échantillon à prédire et la sortie du modèle. Suivant la valeur de cette erreur de reconstruction, un échantillon peut être rejeté. Parmi les méthodes suivant cette approche, on peut citer des réseaux de neurones tels que les auto-encodeurs THOMPSON et collab. [2002]. Ce type de réseau présente un même nombre de neurones en entrée et en sortie et des couches cachées entre les deux. La première partie du réseau constitue l'encodeur, elle encode les données d'entrée en un vecteur de caractéristiques. La deuxième partie, appelée "décodeur", cherche à reproduire les données d'entrée à partir de ces caractéristiques (sortie de l'encodeur). Le but de ce type de réseau de neurones est donc de reproduire les échantillons d'entrée en sortie, en minimisant l'erreur de reconstruction. Une grosse erreur de reconstruction permet d'identifier les échantillons considérés comme des valeurs aberrantes. D'autres méthodes de cette catégorie, supposent qu'il existe un espace de dimension inférieure dans lequel, une fois les données projetées (en utilisant une ACP par exemple) les échantillons d'une même classe et ceux à rejeter sont différentiables. Dutta et coll. HOFFMANN [2007] décrivent un algorithme de rejet qui utilise des composantes principales. La dernière composante principale permet d'identifier des observations qui s'écartent significativement de la «structure de corrélation» des données d'apprentissage. Ainsi, une fois projetés sur le sous-espace puis reconstruits, les échantillons dont l'erreur de reconstruction est trop élevée sont rejetés.

— **Les approches probabilistes :** l'utilisation de méthodes probabilistes implique l'estimation de la fonction de densité de probabilité (PDF) des données des classes connues. Ces approches supposent que les zones à faible densité dans l'ensemble d'apprentissage indiquent des zones avec de faibles probabilités de contenir des observations de classes connues. La PDF estimée peut ensuite être seuillée dans le but de définir les limites de "normalité" dans l'espace de données. Sur la base de cette limite, de nouveaux échantillons (non vus dans la base d'apprentissage) pour-

ront être prédits comme appartenant ou non à la distribution. Ces approches sont paramétriques ou non-paramétriques. Parmi les méthodes paramétriques, on retrouve par exemple les modèles de mélange (gaussien, gamma THOM [1958], Poisson CONSUL et JAIN [1973], etc.) et les modèles espace-état. Le modèle de Markov caché ou HMM EDDY [2004] est un modèles espace-état massivement utilisé en reconnaissance de formes, en intelligence artificielle ou en traitement automatique du langage naturel. Ilonen et coll. ILONEN et collab. [2006] calculent une mesure de confiance sur des mélanges de gaussiennes pour estimer la fiabilité d'un résultat de classification lorsqu'une étiquette de classe est attribuée à une observation inconnue. Plus cette mesure est élevée plus la prédiction est considérée comme fiable. On retrouve par exemple cette approche dans PAALANEN et collab. [2006], une application de détection de visage.
Comme approche non paramétrique, on peut citer l'estimateur de densité par noyau KIM et SCOTT [2012] dans laquelle la fonction de densité de probabilité est estimée en utilisant un noyau par échantillon de l'ensemble d'apprentissage. L'estimation de la PDF repose sur les observations localisées dans un voisinage défini par le noyau. L'entraînement de l'estimateur de densité consiste à déterminer la variance des noyaux, qui contrôle la régularité de la distribution globale.

— **Les méthodes qui s'appuient sur la théorie de l'information :** Le principe est d'identifier des sous-ensembles d'échantillons à rejeter en calculant le contenu informationnel d'un ensemble de données (à l'aide de mesures telles que l'entropie ou la complexité de Kolmogorov). Généralement, le sous-ensemble d'échantillons qui induit la plus grande valeur de ces métriques, suite à leur retrait de l'ensemble de données, sont qualifiés d'échantillons à rejeter.

Parmi les méthodes les plus populaires de classification qui se sont vu ajouter une classe de rejet, on retrouve les réseaux de neurones CHANDOLA et collab. [2009]; MARKOU et SINGH [2003] . Dans le cas des réseaux de neurones convolutifs, la dernière couche du réseau est composée d'un nombre de neurones égal à celui du nombre de classes apprises par le réseau. Ainsi, pour chaque observation à prédire, la couche Softmax à la fin du réseau lui attribuera n valeurs entre 0 et 1 associées aux n classes possibles. Plus la valeur est proche de 1, plus la classe correspondante est probable. De ce fait, une prédiction finale émise avec un faible indice de confiance pourrait révéler un échantillon n'appartenant pas aux classes connues. Les travaux dans STEFANO et collab. [2000] définissent une option de rejet pouvant être appliquée à n'importe quel classifieur. L'option de rejet est fondée sur une fonction définie pour estimer la fiabilité de la classification d'un échantillon et d'un seuil appris sur cette fonction. La procédure de rejet est optimisée en fonction de trois coûts : les coûts d'une bonne prédiction C_c, d'une mauvaise prédiction C_e et d'un rejet d'échantillon C_r, qui est inférieur à celui d'une mauvaise prédiction. La procédure est optimale lorsqu'elle rejette le plus de mauvaises prédictions tout en gardant un maximum de bonnes prédictions. Ceci s'exprime comme l'optimisation d'une mesure de performance P par rapport à un seuil σ. En pratique, les auteurs proposent de mettre en œuvre leur approche pour un CNN et d'utiliser deux ensembles de fonction de fiabilité et de seuil en cascade. Ces seuils sont estimés à l'aide de l'ensemble d'apprentissage. La première fonction de fiabilité,Ψ_A, correspond au maximum de la couche de sortie. La deuxième fonction, Ψ_b, correspond à la différence entre les deux plus grandes valeurs de la couche de sortie. Un échantillon s est rejeté si $\Psi_a(s) < \sigma_a$, ou bien si $\Psi_b(s) < \sigma_b$ (voir la figure 3.1), où σ_a et σ_b sont optimisés en fonction de P sur l'ensemble de l'échantillon et sur les échantillons qui n'ont pas été rejetés par Ψ_a, respectivement. Cette optimisation

dépend d'un paramètre appelé le coût normalisé C_N :

$$C_N = \frac{(C_e - Cr)}{(C_r + C_c)} \tag{3.1}$$

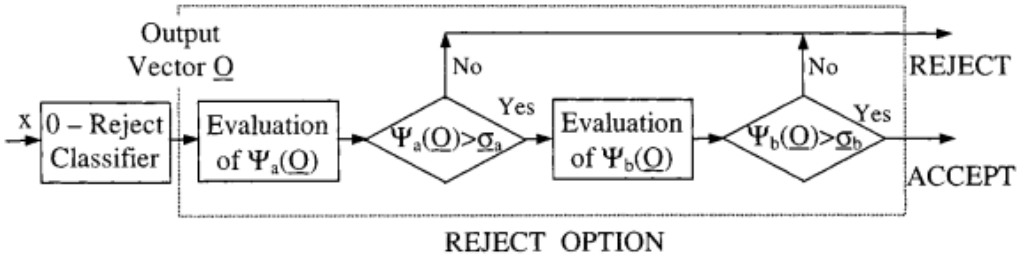


FIGURE 3.1 – L'option de rejet fonctionne sur la base des deux évaluateurs Ψ_a et Ψ_b. Les échantillons rejetés sont ceux dont la valeur de Ψ_a est inférieure à σ_a ou dont la valeur de Ψ_b est inférieure à σ_b. Cette figure correspond à la Figure 5 de 3.1.

Certaines de ces méthodes sont développées directement avec un concept de rejet et d'autres présentent une option de rejet ajoutées a posteriori. Or, il apparaît que l'ajout d'un mécanisme de sélection sur un modèle de prédiction n'est pas l'approche la plus efficace en termes de construction de classe de rejet WIENER et EL-YANIV [2011]. En effet, un classifieur se concentre sur l'optimisation des frontières entre les classes dans l'espace des paramètres sans prendre en compte l'espace *vide* entre et autour des classes qui ne doit manifestement être associé à aucune classe connue. Par conséquent, une meilleure approche est donc de tenir compte directement dans la conception du classifieur de la possibilité de rejeter éventuellement des échantillons. Une notion pour la première fois introduite dans CORTES et collab. [2016].

Dans ce cas, l'optimisation du mécanisme de sélection et du modèle de prédiction se font de façon simultanée et non l'une après l'autre. Dans HOMENDA et collab. [2016], plusieurs ensembles flexibles de classifieurs binaires sont construits, chaque classifieur prédisant une des classes connues ou le rejet de l'observation. Ces modèles de classification utilisent différentes configurations de cascades de SVM avec une ou plusieurs classes de rejet. Certaines études comme celle-ci, construisent effectivement leur système de classification en y intégrant une option de rejet. Néanmoins, la classe de rejet y est représentée par un ensemble d'échantillons. D'autres études, comme GRANDVALET et collab. [2009], sont fondées sur la construction d'un classifieur s'appuyant sur une fonction de coût, permettant le calcul du SVM utilisé pour rejeter ou non un échantillon. On peut également citer ZIYIN et collab. [2019] ou encore SelectiveNet GEIFMAN et EL-YANIV [2019] qui correspond à un réseau de neurones profond avec une option de rejet intégrée et dont les modèles de sélection et de prédiction sont entraînés mutuellement au sein du même réseau. La dernière couche de la partie de sélection du réseau correspond à un unique neurone régi par une fonction d'activation sigmoïde (voir la figure 3.2). La prédiction d'un échantillon est donnée uniquement dans le cas où la valeur du neurone est supérieure ou égale à 0.5. Le modèle sélectif optimal est défini en optimisant un risque sélectif, compte tenu d'une contrainte sur la couverture des données $0 < c \leq 1$. Ce paramètre est à spécifier pour chaque classe et la performance du modèle en dépend. En effet deux situations non optimales peuvent apparaître : la couverture réelle est significativement plus petite ou significativement plus grande que la couverture cible souhaitée.

La variété des méthodes existantes est directement liée à des facteurs tels que la disponibilité des données d'entraînement, des paramètres liés aux données comme le type,

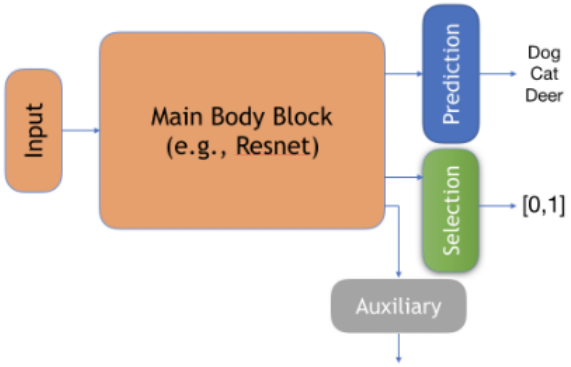


FIGURE 3.2 – Schéma de l'architecture de SelectiveNet. (Figure extraite de GEIFMAN et EL-YANIV [2019]).

la dimension, la continuité et le format KALINICHENKO et collab. [2014]. Le domaine d'application étudié a également une influence sur l'approche considérée. En conséquence de cela, il n'existe pas d'algorithme unique de classe de rejet universellement applicable.

3.2 Méthode proposée

L'idée est d'apprendre une description de la distribution des caractéristiques c'est-à-dire d'ajuster un modèle aux échantillons de l'ensemble d'apprentissage d'une classe, puis de calculer un seuil qui lui est propre, permettant de définir si un échantillon appartient ou non à la classe en question NAIRAC et collab. [1997] (voir la figure 3.3). L'originalité de notre approche repose sur le fait que la méthode de classification supervisée avec une classe de rejet ici proposée, tient compte de la proximité entre les classes connues au lieu de définir classiquement une procédure de rejet ad-hoc comme un ensemble de décisions indépendantes par classe. Les principales étapes sont les estimations du modèle pour chaque classe et l'apprentissage de seuils fondés sur les relations inter-classes.

La stratégie d'apprentissage proposée repose sur des modèles de classe où un modèle est une combinaison d'une fonction d'appartenance floue (FMF) sur l'espace des données et d'un seuil. Pour un échantillon s, si F est une FMF d'une classe, alors F(s) est une valeur réelle entre zéro et un indiquant la probabilité que s appartienne à la classe considérée. Les FMF sont apprises indépendamment pour chaque classe BREW et collab. [2007]. Cependant, les seuils sont appris en tenant compte du chevauchement entre les classes, déduisant ainsi la localisation possible de nouvelles classes en fonction de la distance par rapport aux classes voisines (voir la figure 3.4). Enfin, la prédiction d'une classe d'échantillons est faite en considérant les réponses de tous les modèles. La combinaison de ces réponses induit une prédiction finale selon laquelle l'échantillon appartient à l'une des classes utilisées pour l'apprentissage ou à la classe de rejet.

3.2.1 Notations et principe de la stratégie proposée

Soit S l'ensemble d'apprentissage :

$$S = \cup_{i=1}^{n_c} S_i \tag{3.2}$$

où S_i est le sous-ensemble d'échantillons de la classe C_i et n_c est le nombre de classes défini par l'expert du domaine.

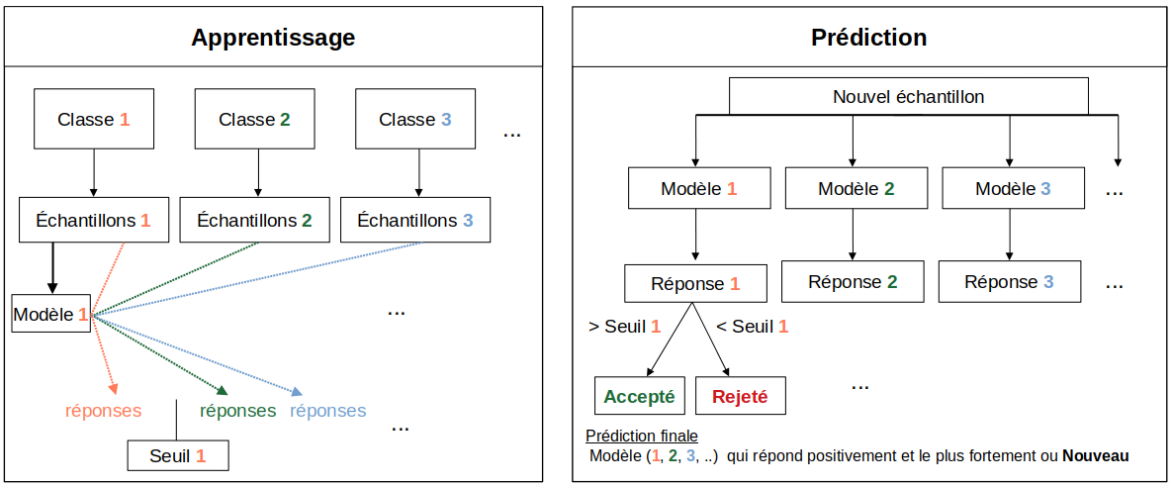


FIGURE 3.3 – Schéma de la procédure d'apprentissage et de prédiction de la méthode de classification avec classe de rejet. Lors de la phase d'apprentissage, les modèles et les seuils sont appris sur les données des différentes classes connues et prennent en compte les relations inter-classes. Dans la phase de prédiction, les nouveaux échantillons sont testés par chaque modèle et les seuils correspondants. Leurs réponses servent à prédire l'appartenance de ces échantillons à l'une de ces classes connues ou à une classe inconnue.

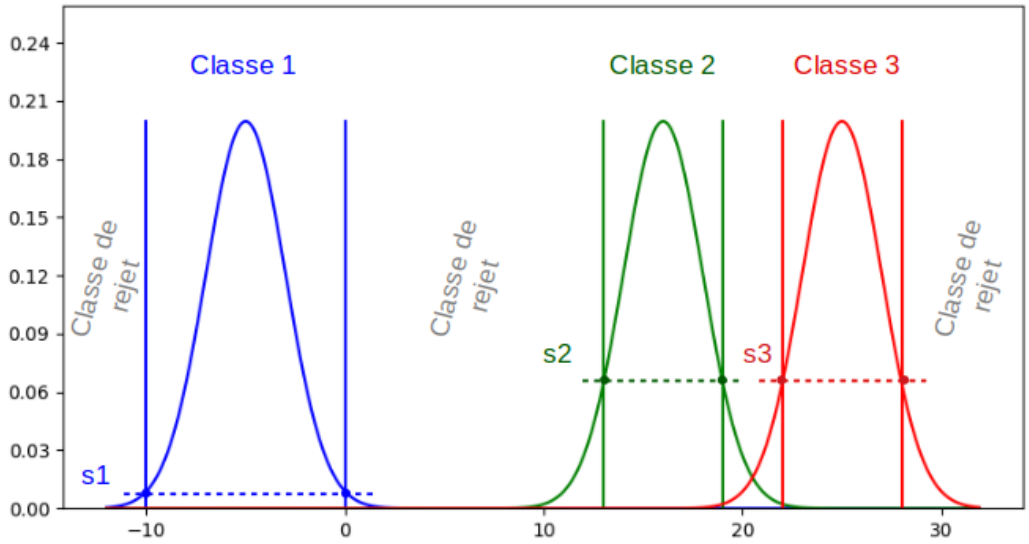


FIGURE 3.4 – Illustration via trois fonctions Gaussiennes de même variance de la dépendance des seuils FMF à la proximité entre les classes.

Nous choisissons de prendre comme FMF, la fonction de densité de probabilité (PDF). Pour définir une PDF M_i pour le phénotype C_i fondée sur le sous-ensemble S_i, un modèle est estimé avec les éléments de S_i. Nous apprenons un seuil T_i sur la PDF M_i de la classe C_i en analysant comment cette PDF répond à «son» jeu d'échantillons mais également à ceux des autres classes. Nous avons mis en place deux options et deux méthodes possibles pour le calcul de ce seuil :

Deux options :

— 1-vs-all : L'approche 1 contre tous, où T_i est directement appris de S_i et de tous les S_k, (avec k différent de i).

— 1-vs-1 : L'approche 1 contre 1, où on apprend un ensemble de seuils $T_{i,k}$ en considérant toutes les paires (S_i ; S_k avec k différent de i). Dans ce cas, le seuil T_i sera le maximum des seuils $T_{i,k}$.

Deux méthodes :

— La méthode « misclassification » utilisant le nombre d'erreurs de classification pour déterminer le seuil qui correspond ici, à la valeur de la PDF quand les nombres de faux positif et faux négatifs sont égaux. Si un échantillon s appartient à S_i, $T_\star$ doit être tel que $M_i(s) \geq T_\star$. Sinon, la prédiction de s est un faux négatif. Inversement, on doit avoir $M_i(s) < T_\star$ si s appartient aux autres ensembles d'échantillons. Sinon, la prédiction est un faux positif. Le seuil $T_\star$ peut alors être exprimé comme le/un minimiseur de :

$$G(T) = \Big|\#\{s \in S_i \text{ tel que } M_i(s) < T\} - \#\{s \in S_k, k \neq i \text{ tel que } M_i(s) \geq T\}\Big| \tag{3.3}$$

où #S est le cardinal de S. En pratique, G n'a pas forcément de minimiseur unique de sorte que $T_\star$ ne peut pas être défini comme $\arg\min_T G(T)$. Au lieu de cela, il existe généralement un intervalle $[T_{min}, T_{max}]$ tel que :

$$\forall T \in [T_{min}, T_{max}], G(T) = \min_U G(U). \tag{3.4}$$

Ensuite, nous proposons de définir $T_\star$ comme une fonction de T_{min} et T_{max} (voir la figure 3.5). Le choix typique est la fonction moyenne :

$$T_\star = (T_{min} + T_{max})/2. \tag{3.5}$$

— La régression logistique SPERANDEI [2014] permet de modéliser la relation entre la PDF M_i et le label 1 ou 0 des échantillons considérés suivant l'option choisie (1-vs-all/1-vs-1). Pour rappel, les échantillons de la classe actuellement traitée sont étiquetés 1 et tous les autres (1-vs-all) ou ceux d'une autre classe (1-vs-1) sont étiquetés 0. Le seuil T_i est dans ce cas égal à la valeur de l'antécédent de la valeur 0.5 par la fonction de régression R.

$$T_\star = R^{-1}(0.5). \tag{3.6}$$

Une fois le modèle et le seuil appris pour chaque classe lors de la phase d'apprentissage, il est possible de prédire la classe d'un échantillon nouveau. Les règles de prédiction sont les suivantes :

— Si seul un modèle répond positivement, l'échantillon est classé dans la classe correspondante.

— Si plusieurs modèles répondent positivement, l'échantillon s est prédit comme appartenant à la classe j correspondant au modèle ayant répondu le plus fortement (probabilité la plus élevée, voir l'équation 3.7).

$$j = \arg \max_{i | M_i(s) \geq T_i} M_i(s) \tag{3.7}$$

— Si aucun modèle ne répond favorablement, alors l'échantillon est classé dans la classe de rejet.

3.2.2 Sensibilité des méthodes de calcul des seuils aux outliers :

Un inconvénient de l'estimation du seuil $T_\star$ via la méthode fondée sur la régression logistique est sa sensibilité aux valeurs aberrantes. En effet, selon la régularisation appliquée au problème de régression, le passage de R de zéro à un pourrait être décalé en présence de valeurs aberrantes (échantillons de S_i avec une valeur très faible de M_i et échantillons de quelques $S_k, k \neq i$ avec une valeur très élevée de M_i). L'ampleur d'un tel déplacement dépendra de la position horizontale des valeurs aberrantes (voir la figure 3.5). Concernant la sensibilité de l'estimation (voir l'équation (3.5)) de $T_\star$ aux valeurs aberrantes dans le cas de la méthode misclassification, notez que tant qu'aucune valeur aberrante ne se déplace dans l'intervalle $[T_{min}, T_{max}]$ (c'est-à-dire que les valeurs de M_i pour les valeurs aberrantes restent en dehors de cet intervalle), la valeur estimée reste inchangée (voir la figure 3.5).

Nous avons décliné l'approche ci-dessus selon une stratégie de modélisation par mélange de gaussiennes (GMM).

3.2.3 Choix du modèle : GMM

Une manière classique d'apprendre une FMF pour la classe C_i à partir de S_i est d'ajuster un modèle de mélange gaussien aux échantillons de S_i. Un modèle statistique de mélange gaussien (GMM) permet d'estimer paramétriquement la distribution de variables en les modélisant comme une somme de plusieurs gaussiennes. Il faut déterminer le nombre de gaussiennes ainsi que les paramètres variance (matrice de covariance), moyenne et pondération de chaque gaussienne, traditionnellement, selon un critère de maximum de vraisemblance afin d'estimer au mieux la distribution recherchée.

Les PDF des classes sont donc estimées à l'aide de mélanges de Gaussiennes. Ajuster un modèle de mélange gaussien (GMM) aux échantillons implique qu'ils suivent une distribution relativement "lisse" et avec une décroissance suffisamment forte aux bords. De plus, selon la complexité du nuage d'échantillons, le mélange devra être composé d'un nombre important de composantes pour estimer au mieux les populations.

3.2.4 Paramètres à optimiser

— Le nombre de composantes du modèle de mélange gaussien.

— Le type de matrice de covariance (trois options : Diagonale constante (Sphérique), Diagonale, Pleine)

Lors du calcul du seuil :

— L'option : 1-vs-all / 1-vs-1

— La méthode : misclassification / régression logistique

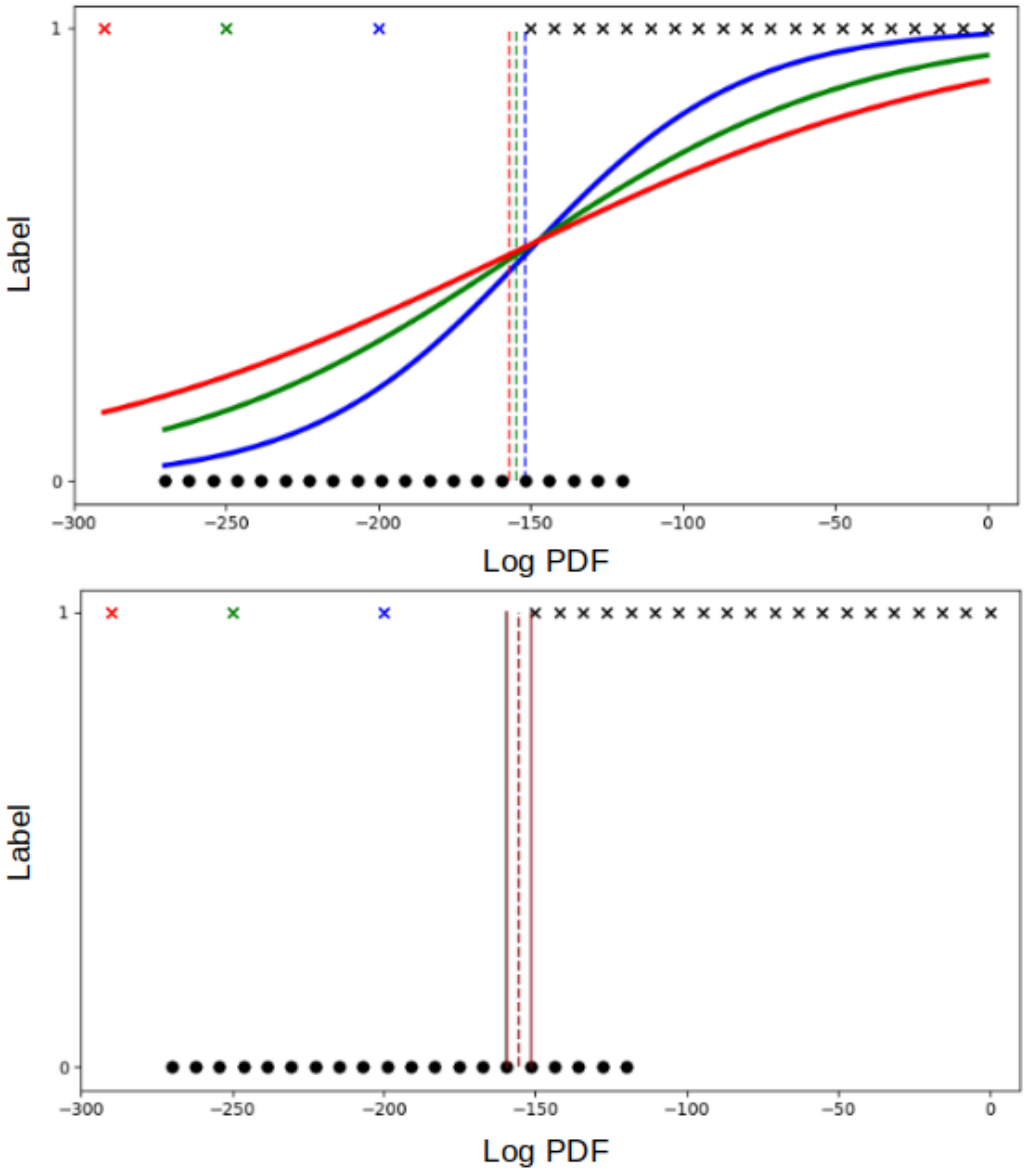


FIGURE 3.5 – Apprentissage du seuil PDF illustré sur des données synthétiques avec les deux options : fondé sur la régression (en haut) et fondé sur le nombre de mauvaises classifications (en bas). Chaque couleur de courbe et de ligne de seuillage en pointillés correspond à un résultat d'apprentissage avec l'ensemble d'apprentissage composé des échantillons de croix et de points noirs plus l'échantillon aberrant "croix" de la même couleur. On peut observer que le décalage de la position de l'échantillon "croix" aberrant déplace également le seuil dans le même sens lors de l'apprentissage avec régression, alors que le seuil appris en utilisant le critère de classifications erronées ne change pas (tant que la valeur aberrante n'entre pas dans le $[T_{min}, T_{max}]$ comme mentionné dans le texte). Notez que les valeurs aberrantes ont reçu une pondération de cinq afin de souligner leur effet.

3.3 Évaluation et validation de la méthode sur des données synthétiques

3.3.1 Description des données

Dans le but d'illustrer et d'évaluer la méthode, deux expériences ont été menées sur des données synthétiques, correspondant dans un premier temps à de simples gaussiennes puis à des distributions plus complexes.

Expérience A Afin d'étudier le comportement des modèles estimés par des mélanges gaussiens, nous avons créé plusieurs populations (de 1000 échantillons chacune) correspondant à des données gaussiennes 2D dont la moyenne est tirée de façon aléatoire et qui ont la même variance. Les modèles de deux à six classes sont chacun estimés par une seule gaussienne (voir la figure 3.7). Pour chacun d'eux, nous calculons le seuil d'appartenance sur la PDF. Afin de visualiser l'influence des modèles les uns sur les autres, nous prédisons la classe d'appartenance (celle pour laquelle la probabilité d'appartenance est la plus élevée) sur une grille régulière (voir la seconde colonne de la figure 3.7). L'intérêt d'ajouter une classe après l'autre est de visualiser l'influence d'une nouvelle population sur les seuils calculés (approche "1-vs-1") et sur la répartition des classes dans l'espace des paramètres au travers de la carte des prédictions.

Expérience B Nous avons testé la méthode de classification avec classe de rejet sur trois jeux de données (voir la figure 3.6) que l'on appellera dans la suite du manuscrit "Moon data", "Ring data" et "S data". Chaque ensemble de données présente deux classes d'échantillons étiquetés 1 et 2 dont le nombre par classe est indiqué dans le tableau 3.1. Les données "Moon data" et "Ring data" sont générés grâce à la bibliothèque libre Python Scikit-learn destinée à l'apprentissage automatique. Les échantillons sont tirés selon une gaussienne tronquée par leurs supports respectifs ("ring" ou "moon"). En ce qui concerne "Moon data" il s'agit de deux demi-cercles entrelacés et pour "Ring data", il s'agit d'un grand cercle contenant un cercle plus petit. Les données "S data", sont quant à elle construite de manière ad hoc. La classe 1 correspond à une gaussienne et la classe 2 correspond à des échantillons tirés de manière uniforme sur un support défini par le "S".
Nous avons également comparé les résultats à ceux de plusieurs classifieurs tels que la méthode des Random Forest BREIMAN [2001] (RF), et les séparateurs à vaste marge CHANG et LIN [2011] (SVM) toutes deux décrites dans le paragraphe 2.2.2, mais également la méthode des k-plus proches voisins ALTMAN [1992] (k-NN) et un algorithme de boosting, AdaBoost HASTIE et collab. [2009].
Des échantillons permettant l'évaluation de la classe de rejet sont générés de façon aléatoire tout autour des classes 1 et 2 de chaque jeu de données.

Nb. échan./Datasets	Moon	Ring	S
Classe 1	500	500	≈ 1000
Classe 2	500	500	≈ 1000
Nouveau	≈ 1200		

TABLEAU 3.1 – Nombre d'échantillons par classe dans les trois ensembles de données.

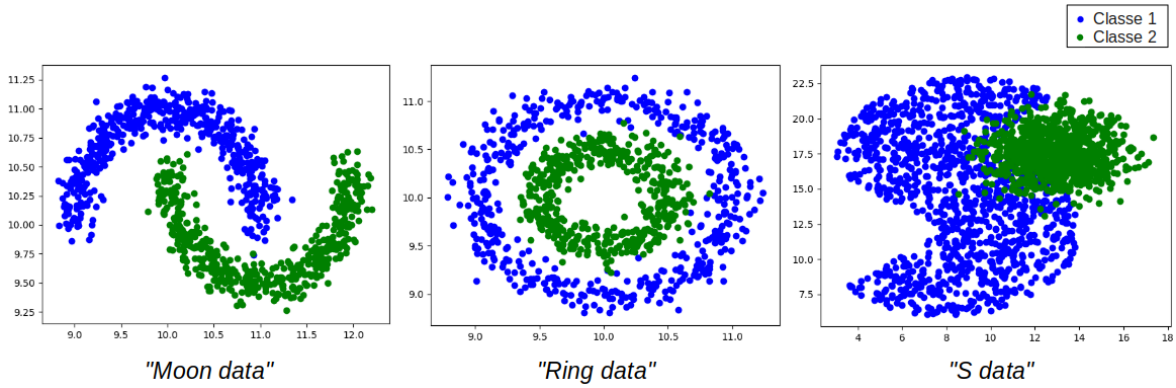


FIGURE 3.6 – Ensemble de données des jeux "Moon data", "Ring data" et "S data".

3.3.2 Résultats

Expérience A La première colonne de la figure 3.7 correspond aux données gaussiennes 2D et la seconde à des cartes de prédiction. Ces cartes sont le résultat des prédictions par les modèles et seuils appris (approche "1-vs-1"), pour les échantillons d'une grille régulière.
La phase d'apprentissage des seuils d'appartenance sur la probabilité, pour la classe 1, dans le cas du jeu de données à six classes (0 à 5) est illustré sur la figure 3.8 pour l'approche "1-vs-1" et sur la figure 3.9 pour l'approche "1-vs-all". Pour rappel, le label 1 ou 0 des échantillons considérés dépend de l'option choisie ("1-vs-1" ou "1-vs-all"). Les échantillons de la classe actuellement traitée (sur les figures il s'agit de la classe 1) sont étiquetés 1 et tous les autres (1-vs-all) ou ceux d'une autre classe (1-vs-1) sont étiquetés 0. Dans le cas de l'approche "1-vs-1", cinq seuils sont appris : un entre la classe 1 et la classe 0, la classe 1 et la classe 2, la classe 1 et la classe 4 et enfin entre la classe 1 et la classe 5 (voir la figure 3.8 A, B, C, D et E).

La première colonne correspond à l'évolution du nombre d'échantillons mal classés en fonction du seuil testé. Pour rappel, le seuil selon l'approche par misclassification correspond à la valeur pour laquelle le nombre de faux positifs et de faux négatifs sont égaux. Les figures de la deuxième colonne présentent les valeurs de PDFs renvoyées par le modèle (GMM) pour les échantillons de la classe 1 (étiquetés 1) et les échantillons étiquetés 0 ainsi que la régression logistique associée. Le seuil calculé par régression logistique est égal à la valeur de l'antécédent de la valeur 0.5 par la fonction de régression estimée sur les valeurs de PDFs. Les valeurs des seuils calculés par les deux méthodes (régression et misclassification) sont indiquées, nous permettant ainsi de les comparer.

Pour l'option "1-vs-all", l'unique seuil calculé concerne d'une part les échantillons de la classe 1 (label 1) et d'autre part, ceux de toutes les autres classes confondues (label 0). Contrairement à l'expérience suivante, celle-ci ne permet pas de définir la méthode la plus efficace.

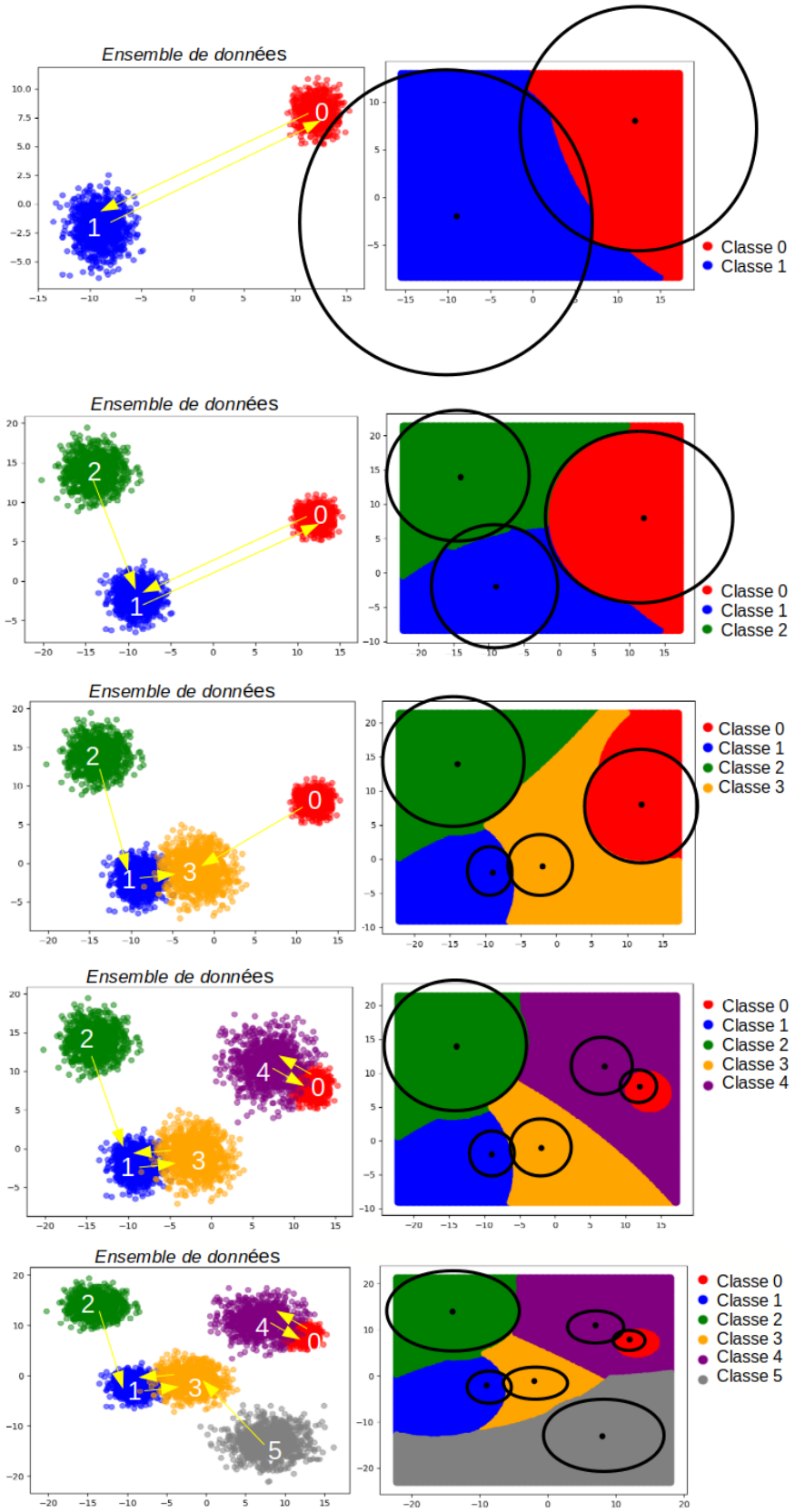
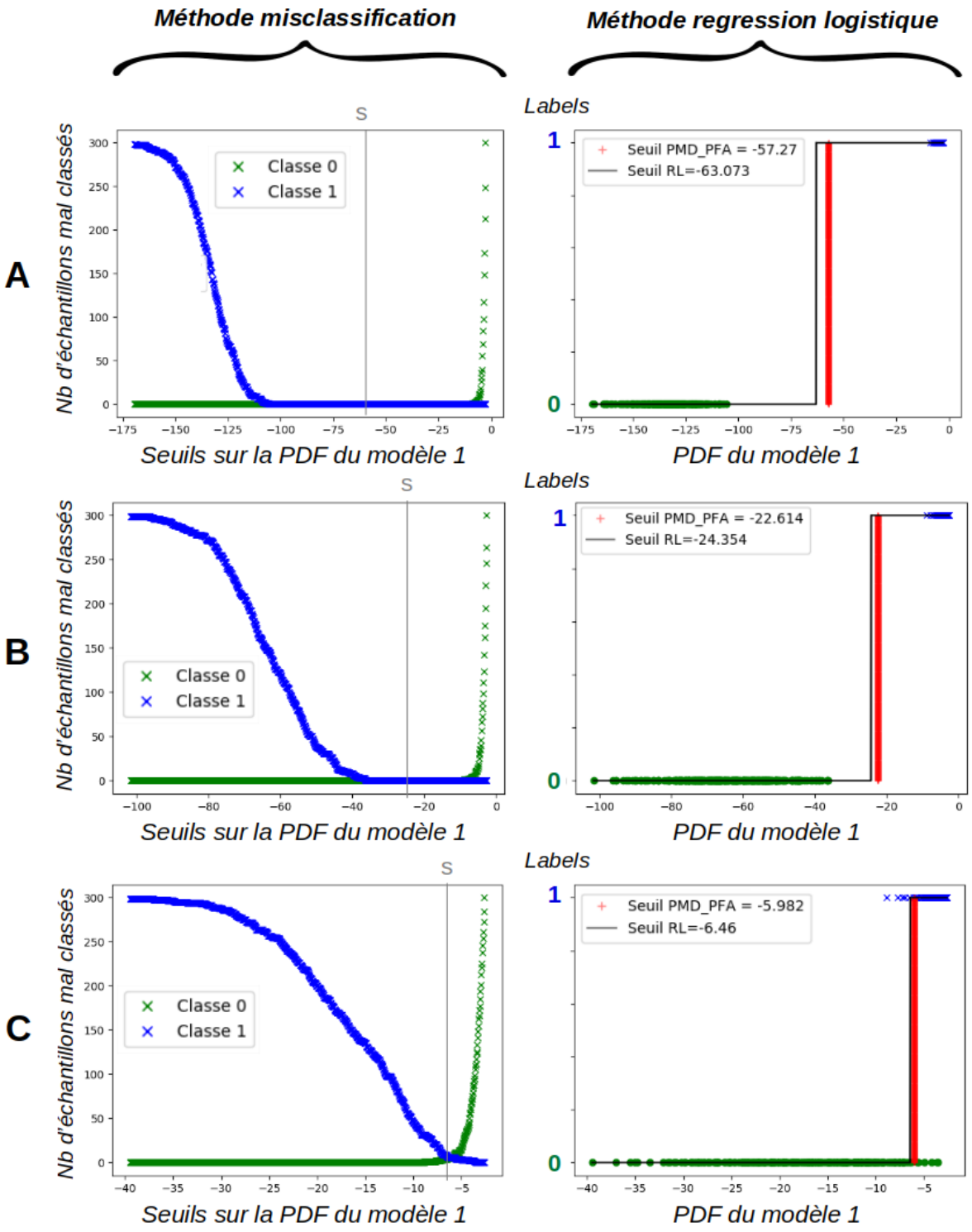


FIGURE 3.7 – Première colonne : données synthétiques modélisées par une gaussienne. Deuxième colonne : carte de prédictions. Un cercle par population est tracé, de rayon le seuil calculé en fonction des autres modèles et de centre la moyenne des échantillons. Les flèches jaunes indiquent le modèle le plus proche, celui qui permet de définir le seuil sur la probabilité d'appartenance.


Méthode misclassification
Méthode regression logistique
A
B
C
Nb d'échantillons mal classés
Classe 0
Classe 1
S
Seuils sur la PDF du modèle 1
Labels
Seuil PMD_PFA = -57.27
Seuil RL=-63.073
Seuil PMD_PFA = -22.614
Seuil RL=-24.354
Seuil PMD_PFA = -5.982
Seuil RL=-6.46
PDF du modèle 1

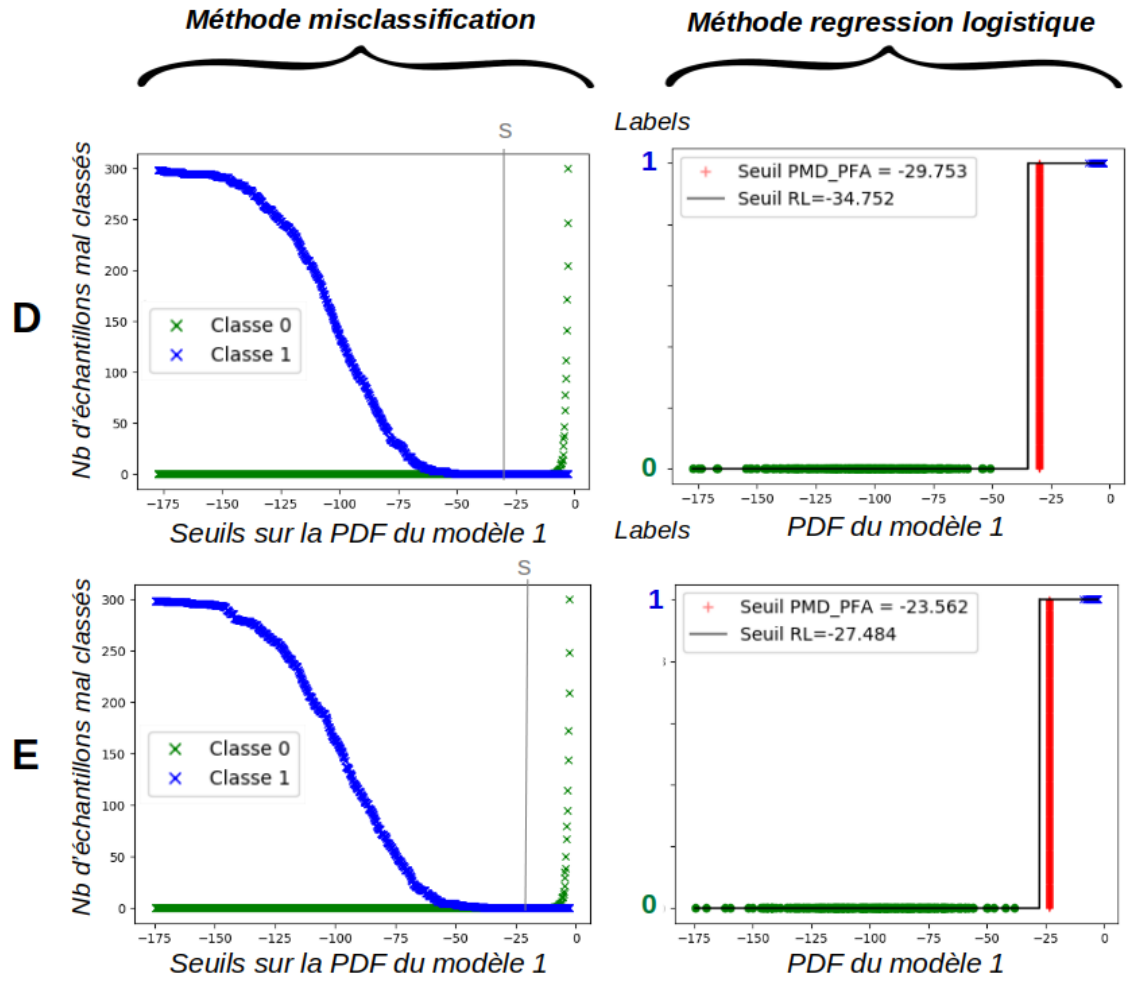


FIGURE 3.8 – Apprentissage des seuils du modèle de la classe 1 (dans le cas d'un ensemble de données à six populations) : Dans le cas de l'**approche "1-vs-1"**, cinq seuils s sur les PDFs (log) sont appris : un entre la classe 1 et la classe 0 (A), la classe 1 et la classe 2 (B), la classe 1 et la classe 3 (C), la classe 1 et la classe 4 (D) et enfin entre la classe 1 et la classe 5 (E). Pour chaque calcul, les échantillons de la classe 1 sont étiquetés 1 et ceux de l'autre classe sont étiquetés 0. Les figures de la première colonne illustrent le calcul du seuil par la méthode misclassification. La figure correspond au nombre d'échantillons mal classés en fonction du seuil testé. La seconde colonne présente le calcul du seuil par la méthode de régression. Le seuil calculé est égal à la valeur de l'antécédent de la valeur 0.5 par la fonction de régression estimée sur les valeurs de PDFs du modèle 1 (GMM) pour les échantillons de la classe 1 (label 1) et des échantillons labellisés 0. Le seuil trouvé par misclassification est également indiqué dans le but de comparer les deux possibilités.

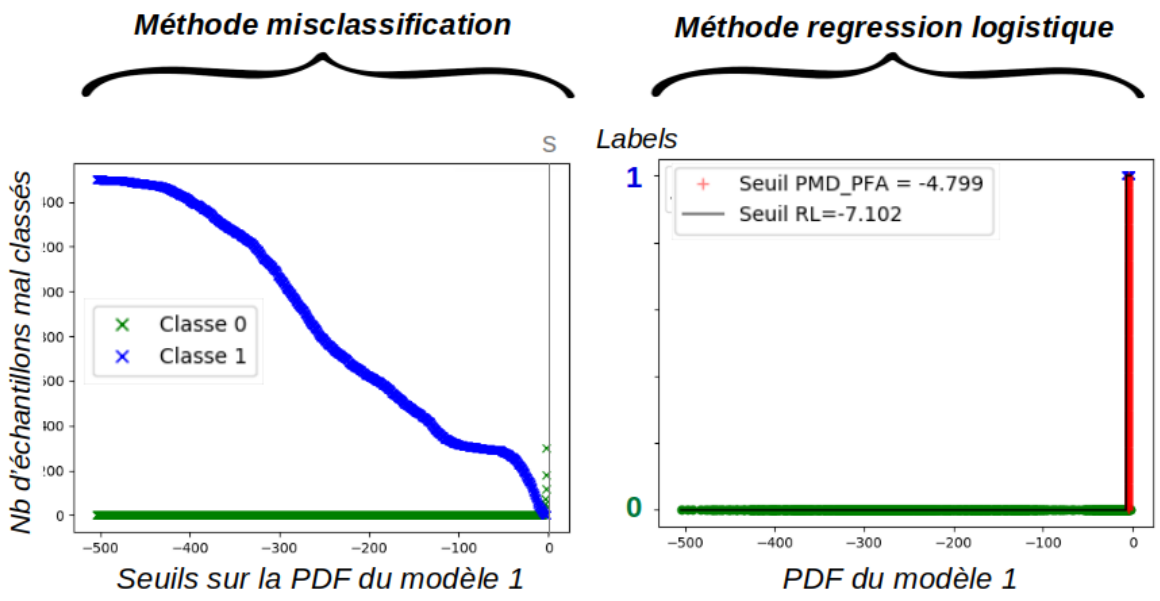


FIGURE 3.9 – Apprentissage des seuils du modèle de la classe 1 (dans le cas d'un ensemble de données à six populations) : Dans le cas de l'**approche "1-vs-all"**, un seul seuil s sur les PDFs (log) est appris. Les échantillons de la classe 1 sont étiquetés 1 et les échantillons de toutes les autres classes sont étiquetés 0. Les figures de la première colonne illustrent le calcul du seuil par la méthode misclassification. La figure correspond au nombre d'échantillons mal classés en fonction du seuil testé. La seconde colonne présente le calcul du seuil par la méthode de régression. Le seuil calculé est égal à la valeur de l'antécédent de la valeur 0.5 par la fonction de régression estimée sur les valeurs de PDFs du modèle 1 (GMM) pour les échantillons de la classe 1 (label 1) et des échantillons labellisés 0.

Expérience B Les modèles de deux classes sont estimés par un mélange gaussien à plusieurs composantes dont les paramètres nombre de composantes et type de matrice de covariance sont optimisés via le critère d'information bayésien ("Bayesian information criterion" en anglais ou BIC) SCHWARZ et collab. [1978]. Ce critère permet de sélectionner un modèle parmi un ensemble fini de modèles, le modèle retenu étant celui qui minimise le critère BIC. Une pénalité dépendant de la taille de l'échantillon et du nombre de paramètres est utilisée afin de mesurer la qualité du modèle statistique. Le critère BIC repose en partie sur la fonction de vraisemblance et est défini par :

$$\mathrm{BIC} = -2\ln(\mathrm{L}) + k\ln(\mathrm{N}) \tag{3.8}$$

avec L la vraisemblance des données selon le modèle estimé, N le nombre d'observations dans l'échantillon et k le nombre de paramètres du modèle.

Pour chacun des deux modèles de classe, un seuil sur les probabilités d'appartenance est déterminé par recherche exhaustive à l'aide d'une régression logistique (cas le plus intéressant car permettant l'obtention de meilleurs résultats). La figure 3.10 illustre la recherche du seuil dans le cas de la classe 1 avec les deux méthodes. Dans le cas d'une classification des échantillons de seulement deux classes, les contextes "1-vs-1" et "1-vs-all" sont équivalents.
Dans le but d'évaluer la méthode, les résultats expérimentaux ont été moyennés sur une validation croisée "3/4, 1/4". La totalité de la cohorte est divisée aléatoirement en quatre et la proportion initiale des deux classes est conservée dans chacun des ensembles. Ainsi,

pour chacune des quatre itérations, 3/4 des données est destiné à l'apprentissage et 1/4 au test. De cette manière chaque échantillon est utilisé exactement une fois en tant qu'échantillon de test. De plus, afin de visualiser l'influence des modèles les uns sur les autres, les points d'une grille régulière sont également prédits, nous permettant d'obtenir une carte de prédiction (voir les figures 3.11 et 3.12). Chaque classe est associée à une couleur et la valeur de PDF du modèle sélectionné définit la nuance de cette couleur : plus la PDF est forte plus la couleur est foncée. Une fois les valeurs de PDF récupérées et comparées, la couleur de l'échantillon correspond à la classe ayant répondu le plus fortement et la nuance de cette couleur à la valeur de la PDF.
D'autre part, la méthode de classification sans classe de rejet correspond au maximum de vraisemblance. La prédiction finale de la classe d'un échantillon correspond à la classe dont le modèle répond le plus fortement à l'échantillon.

Autres classifieurs : Comme expliqué dans le paragraphe 4.4.1, nous avons comparé les résultats de l'approche de classification avec et sans classe de rejet à ceux de plusieurs classifieurs ; RF, SVM, k-NN et AdaBoost. Les paramètres ont été optimisés par rapport au taux de bonne classification. La valeur des principaux paramètres sont :

- RF : Le nombre d'arbres de décision est fixé à 100 et la profondeur maximale de l'arbre à 20.
- SVM : Le paramètre de régularisation (ou paramètre de compromis de la marge souple) est fixé à 1 et l'écart-type à 2.
- k-NN : Le nombre de voisins pris en compte est de K = 3 et la métrique définie correspond à la distance euclidienne.
- AdaBoost : Le nombre d'arbres de décisions est fixé à 100.

L'ensemble des résultats obtenus sont regroupés dans les tableaux 3.2, 3.3 et 3.4. Dans le cas de l'évaluation de la méthode GMM avec classe de rejet, les classes des échantillons à rejeter (voir les figures 3.11 et 3.12) sont également prédites. Les matrices de confusions 3.2 correspondent aux résultats de classification des échantillons des classes 1 et 2 des bases de données "Moon data", "Ring data" et "S data" obtenus avec le modèle GMM sans la classe de rejet. Les matrices de confusions 3.3 correspondent quant à elles, aux résultats de classification des échantillons des classes 1 et 2 et nouveau de ces trois bases de données obtenus avec le modèle GMM avec la classe de rejet. Enfin, le tableau 3.4 correspond aux moyennes des scores de classification obtenus avec les différents classifieurs (RF, SVM, k-NN et AdaBoost) ainsi que la méthode GMM avec et sans la classe de rejet.

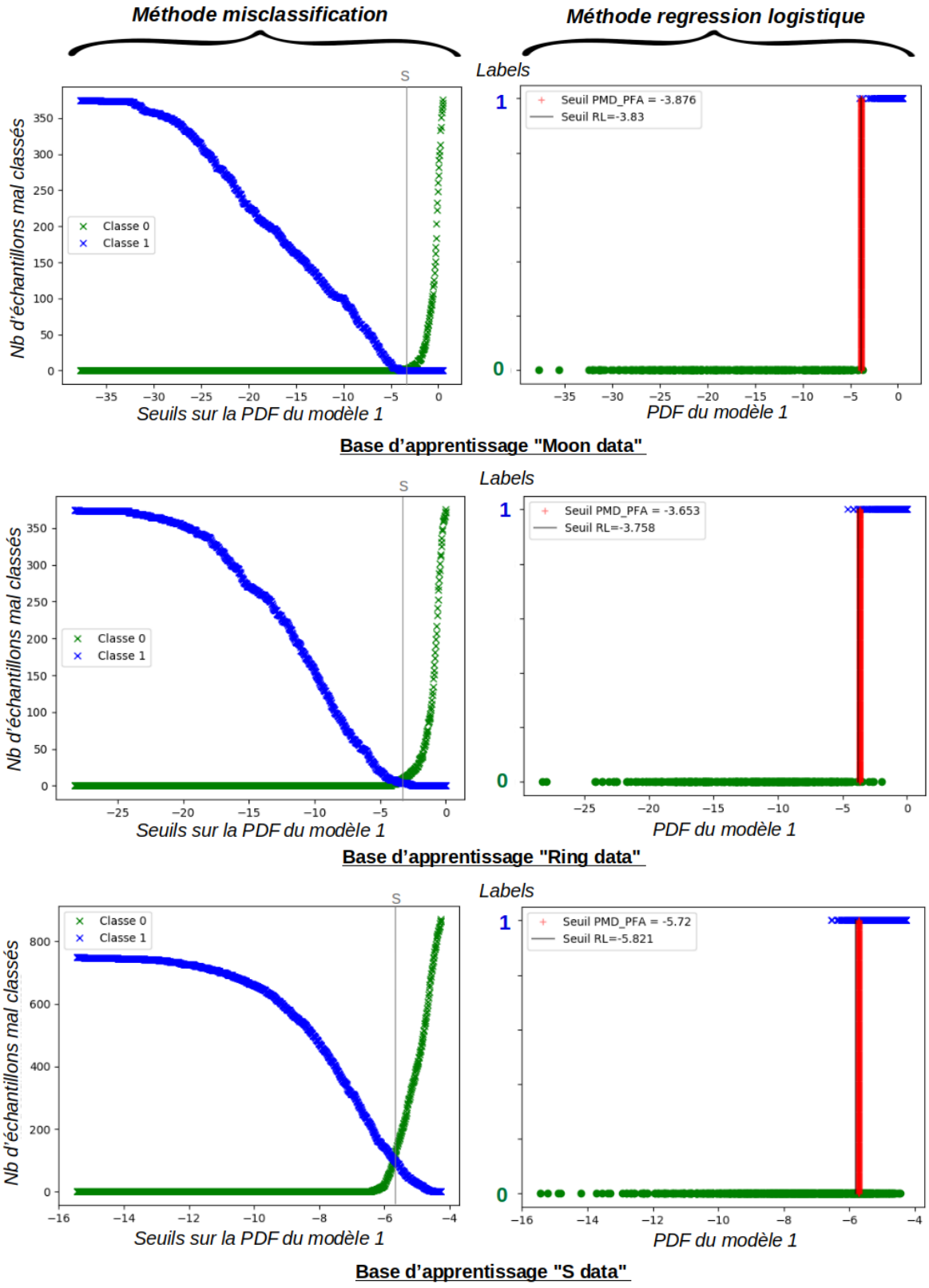


FIGURE 3.10 – Apprentissage des seuils du modèle 1 :
Les figures de la première colonne illustrent le calcul du seuil s sur les PDFs (log) par la méthode misclassification. Le nombre d'échantillons mal classés en fonction du seuil testé est calculé pour 1000 valeurs de seuils (dans l'intervalle de la valeur minimale à la valeur maximale des PDFs de l'ensemble d'apprentissage). Les figures de la deuxième colonne présentent le calcul du seuil par la méthode de régression, le seuil est égal à la valeur de l'antécédent de la valeur 0.5 par la fonction de régression estimée sur les valeurs de PDFs du modèle 1. Les échantillons des classes 1 et 2 sont respectivement étiquetés 1 et 0.

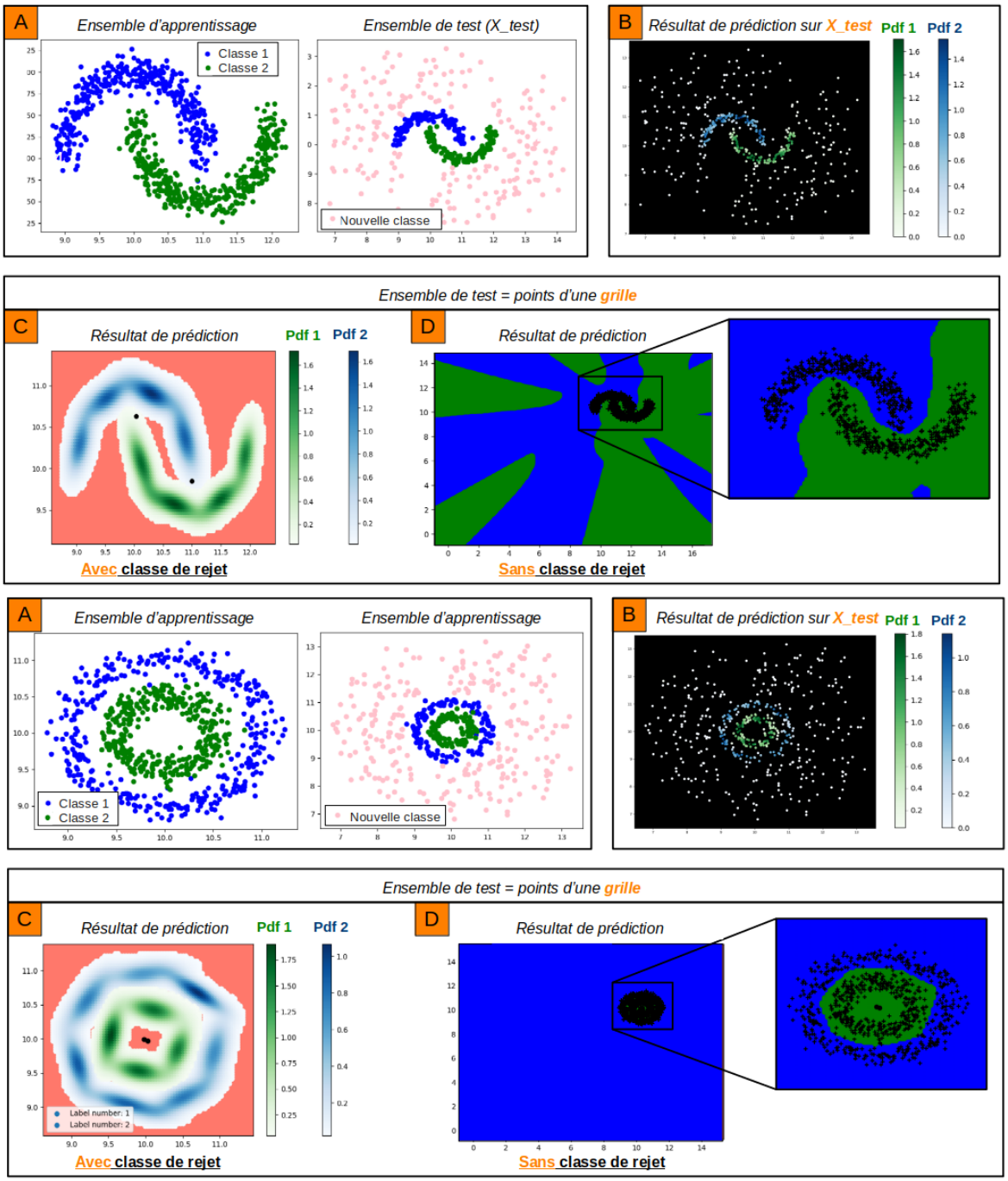


FIGURE 3.11 – Illustration des résultats de classification sur les bases de données "Moon data" et "Ring data". Les croix noires correspondent aux échantillons de l'ensemble d'apprentissage. (A) Ensembles d'apprentissage et de test. (B) Résultats de classification sans classe de rejet des échantillons de test. La couleur correspond aux valeurs des PDFs. (C) Résultats de classification sans classe de rejet des échantillons d'une grille régulière définie. La couleur d'un échantillon correspond à la classe prédite. (D) Résultats de classification avec classe de rejet des échantillons d'une grille régulière définie. La couleur correspond aux valeurs des PDFs et la couleur saumon indique la zone de rejet dans l'espace des caractéristiques.

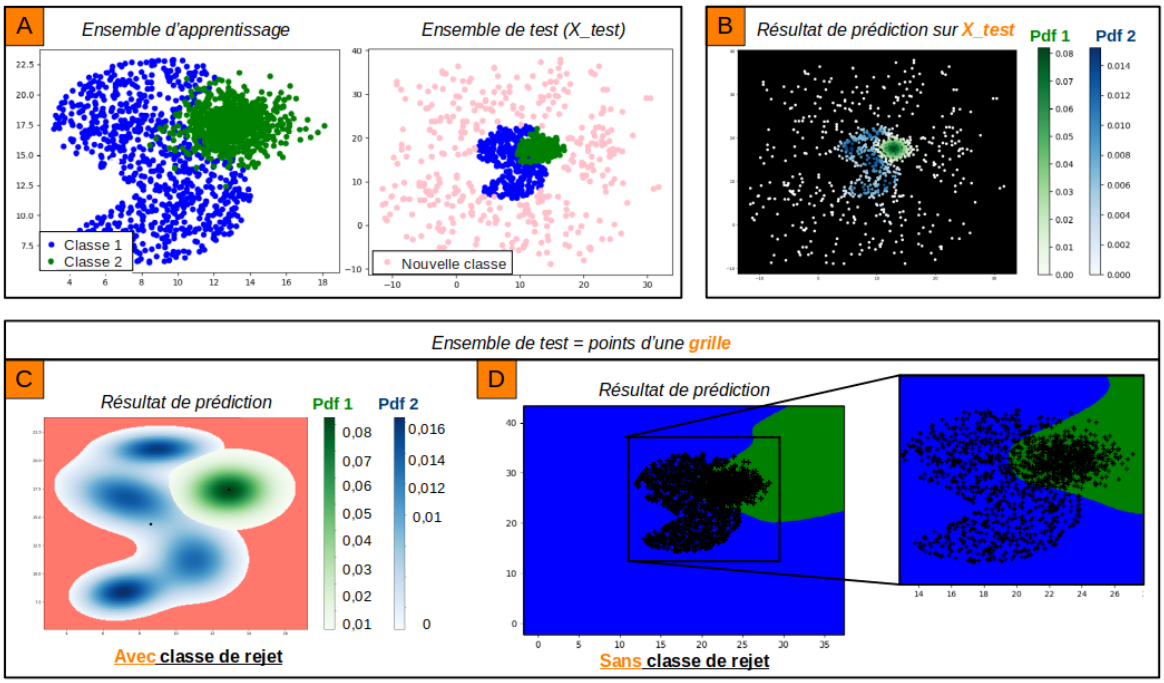


FIGURE 3.12 – Illustration des résultats de classification sur la base de données "S data". Les croix noires correspondent aux échantillons de l'ensemble d'apprentissage. (A) Ensembles d'apprentissage et de test. (B) Résultats de classification sans classe de rejet des échantillons de test. La couleur correspond aux valeurs des PDFs. (C) Résultats de classification sans classe de rejet des échantillons d'une grille régulière définie. La couleur d'un échantillon correspond à la classe prédite. (D) Résultats de classification avec classe de rejet des échantillons d'une grille régulière définie. La couleur correspond aux valeurs des PDFs et la couleur saumon indique la zone de rejet dans l'espace des caractéristiques.

Réelles / Prédites	Classe 1	Classe 2
Classe 1	99.6	0.4
Classe 2	0.6	99.4

Réelles / Prédites	Classe 1	Classe 2
Classe 1	99	2
Classe 2	3.6	98.4

Réelles / Prédites	Classe 1	Classe 2
Classe 1	94.9	5
Classe 2	3.6	96.4

TABLEAU 3.2 – Matrices de confusion des résultats de classification (en pourcentages) des échantillons des bases de données (dans l'ordre) "Moon data", "Ring data" et "S data" obtenus avec le modèle GMM sans la classe de rejet.

Réelles / Prédites	Classe 1	Classe 2	Classe de rejet
Classe 1	98.6	0.4	1
Classe 2	0.2	99	0.8
Classe de rejet	0.2	0.2	99.6

Réelles / Prédites	Classe 1	Classe 2	Classe de rejet
Classe 1	96.2	1.2	2.6
Classe 2	1.2	97.8	1
Classe de rejet	0.5	0	99.5

Réelles / Prédites	Classe 1	Classe 2	Classe de rejet
Classe 1	91.4	3.1	5.4
Classe 2	5.2	93.6	1.2
Classe de rejet	0.5	0	99.5

TABLEAU 3.3 – Matrices de confusion des résultats de classification (en pourcentages) obtenus avec le modèle GMM avec la classe de rejet. Les échantillons considérés sont ceux des bases de données "Moon data" (en haut), "Ring data" (au milieu) et "S data" (en bas) ainsi que dans chaque cas, des échantillons dits "nouveaux". Les seuils sur les PDF sont estimés à l'aide d'une régression logistique, dans un contexte "1-vs-1".

Classifieur / Datasets	Moon	Ring	S
RF	98.9	97.9	94.5
KNeighbors	98.8	98.8	94.75
SVM	98.8	**98.9**	**95.8**
AdaBoost	99.4	97.5	95.1
GMM (sans rejet)	**99.5**	98.7	95.65
GMM (avec rejet)			
Connu	98.8	97	92.5
Classe de rejet	99.6	99.5	99.5
Moyenne des deux	99.2	98.2	96

TABLEAU 3.4 – Scores moyens de classification (en pourcentages) des échantillons des bases de données "Moon", "Ring" et "S", obtenus avec différents classifieurs ; RF : Random Forest, KNeighbors : méthode des k-plus proches voisins, SVM : les machines à vecteurs de support, AdaBoost : algorithme de boosting et la méthode GMM sans la classe de rejet. Les moyennes des scores obtenus avec la méthode de classification avec classe de rejet sont indiquées de manière indépendante sur les classes connues et inconnue, puis une moyenne globale est également calculée.

3.3.3 Discussion

L'expérience A (voir les figures 3.7, 3.8 et 3.9) nous a permis de visualiser l'influence des modèles gaussiens les uns sur les autres mais également l'influence de l'ajout de nouvelles classes, c'est-à-dire la mise à jour des seuils en prenant en compte les nouveaux modèles.

L'expérience B a, quant à elle, permis d'évaluer la méthode de classification avec et sans classe de rejet dans le cadre de plusieurs problèmes de classification auxquels on a ajouté des échantillons n'appartenant à aucune classe connue (voir les figures 3.11 et 3.12), des échantillons "hors-classe". Pour chaque jeu de données les résultats de classification avec et sans classe de rejet pour chacune des classes sont présentés sous forme d'une matrice de confusion dans les tableaux 3.2 et 3.3. Les scores obtenus avec la méthode sans classe de rejet sont importants car ils représentent des valeurs de référence dans le sens où nous voulons conserver une bonne efficacité de prédiction des classes connues dans la méthode avec rejet. Enfin, les moyennes des précisions de classification obtenues dans chaque cas sont regroupées dans le tableau 3.4.

Concernant la partie classification sans classe de rejet, les résultats obtenus nous permettent de calculer une moyenne de 97.95% sur les trois expériences. Cette moyenne est supérieure à celles obtenues avec les autres classifiers (RF : 97.1%, SVM : 97.83%, k-NN : 97.45% et AdaBoost : 97.3%).

D'autre part, l'ajout de la classe de rejet entraîne une légère diminution des scores sur les classes connues avec une différence de 0.7% sur les données "Moon", 1.7% sur les données "Ring" et 3.15% sur les données "S", soit une diminution moyenne de 1.85%. Cette diminution peut s'expliquer par le fait que la frontière, définie par le seuil sur le modèle, est fixée par les contraintes qui existent dans les régions où il y a une "interaction" avec l'autre classe. Cependant, dans les régions de l'espace où les classes ne sont pas "en interaction", la frontière est la même. Certains échantillons de la classe peuvent donc être rejetés à tort.

De plus, les scores sur la classe de rejet sont très proches de 100% avec une moyenne d'environ 99.5% sur les trois jeux de données. Enfin, les moyennes des deux scores (classes connues et inconnue) sont calculées et présentées à la dernière ligne du tableau 3.4. En les comparant aux scores de classification sans classe de rejet, on observe une diminution des scores de 0.3% sur les données "Moon" et de 0.5% sur les données "Ring". Dans le cas des données "S", on obtient une très légère augmentation du score de classification (+0.35%) due au très bon score sur les échantillons "hors-classe". La différence moyenne est de −0.45%, soit une diminution proche de 0, ce qui est négligeable.

3.3.4 Conclusion

En conclusion, l'ajout de la classe de rejet ne diminue que très peu les précisions de classification sur les classes connues et obtient d'excellents scores de prédiction sur la classe de rejet. Les observations effectuées et les excellents résultats obtenus lors de ces deux expériences nous ont permis de valider notre méthode de classification avec classe de rejet sur des données synthétiques. Dans le cadre de ma thèse, le développement de cette méthode est destiné à la reconnaissance des phénotypes connus et nouveaux de champignons. La partie suivante du manuscrit est consacrée aux travaux effectués dans ce cadre précis.

3.4 Test de la méthode sur les images des phénotypes connus et inconnus de *Botrytis cinerea*

3.4.1 Description des données

Dans cette première partie du chapitre, nous nous concentrons sur les images des classes phénotypes remarquables connus (1-4) et inconnus, en mettant de côté les classes mycelium, MiniGermTube et Crystal. Cette décision est justifiée dans le paragraphe 3.5.1. Les images de phénotypes des classes connues sont les mêmes que celles décrites dans le paragraphe 2.4.3. Dans le but d'évaluer la méthode de classification avec classe de rejet proposée, des images de nouveaux phénotypes (voir la figure 3.14) sont ajoutées à notre ensemble de données (voir la figure 3.13). Nous disposons ainsi de 647 images de nouveaux phénotypes.

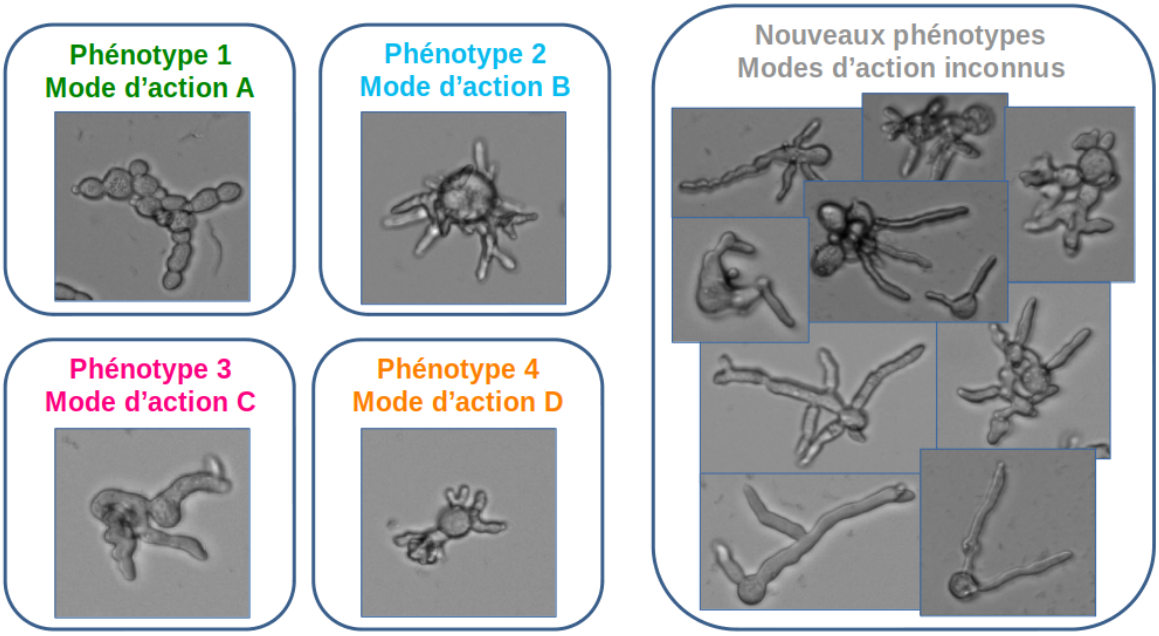


FIGURE 3.13 – Phénotypes connus (1-4) correspondant à des signatures phénotypiques caractéristiques du traitement chimique utilisé et quelques exemples de nouveaux phénotypes. Images en microscopie à lumière transmise, Microscope ImageXpress, Objectif x10.

3.4.2 Description des caractéristiques

Nous avons choisi d'utiliser comme représentation de nos images deux types de caractéristiques. Ces caractéristiques proviennent d'un réseau de neurones, type de classifieur qui peut être utilisé comme extracteur de caractéristiques. Les caractéristiques sont calculées de façon automatique pendant l'étape d'apprentissage (voir le paragraphe 2.2.4), comme il s'agit de la sortie de la couche communément appelée Bottleneck, on parlera de caractéristiques Bottlenecks. Le second type de caractéristiques correspond à la sortie de la dernière couche du réseau ; on parlera de caractéristiques Output. Nous avons donc utilisé le CNN MobileNet une fois entraîné sur les sous-images des phénotypes connus (sept classes, voir la figure 2.16) pour récupérer les vecteurs caractéristiques [1] (Bottlenecks et Output) des images données en entrée. Pour rappel (voir le paragraphe 2.5.2) afin de nous

1. Listes des valeurs des caractéristiques des échantillons.

adapter à la taille d'entrée des images du réseau de neurones MobileNet, nous avons découpé les images de $2160 * 2160$ en seize vignettes de $500 * 500$ pixels, chacune redimensionnées à $224 * 224$ pixels. Dans le chapitre 2, paragraphe 2.5.2, le réseau de neurones est utilisé comme classifieur où les classes de chacune des vignettes d'une image sont prédites puis fusionnées dans le but de lui attribuer une prédiction finale. Contrairement au chapitre 2, dans cette partie du projet le réseau de neurones est utilisé comme extracteur de caractéristiques. Ainsi, afin de s'affranchir des vecteurs caractéristiques de sous-images contenant trop peu d'information relative au phénotype (voir la figure 3.15), nous prenons la médiane de chacune des composantes de ces vecteurs comme représentation d'une image (16 sous-images par image).

— Output : Chaque composante du vecteur reflète un indice de confiance correspondant à l'appartenance d'un échantillon pour cette classe. Les valeurs de ces composantes étant bornées (elles appartiennent à l'intervalle [0,1]) (la somme de toutes les composantes est normalisée à un), nous avons décidé de les symétriser pour permettre une meilleure optimisation des modèles utilisant les fonctions de base symétriques que sont les gaussiennes. Pour cela chaque échantillon sera dupliqué en 2^d échantillons (où d représente la dimension de la représentation). Considérons un échantillon s dont les composantes sont s_i. Nous partons du principe que la plus haute valeur du vecteur, indique la classe d'appartenance de l'échantillon de par sa position (la composante i). Ainsi, un nouvel échantillon est créé avec le même vecteur de caractéristiques mais dont la composante i aura pour valeur, "2− la valeur initiale". Dans le cas des autres composantes, leur valeur sera de "− la valeur initiale" (voir la figure 3.16). Autrement dit, la valeur la plus élevée du vecteur est symétrisée par rapport à 1 et les autres par rapport à 0.

— Bottlenecks : Lorsque les données sont extraites, la dimension de 1001 caractéristiques (nombre de neurones de la couche de caractéristiques du réseau MobileNet) est trop élevée pour estimer correctement les modèles. Nous avons donc testé deux méthodes de réduction de dimension. La première passe par l'apprentissage des arbres de décision d'un modèle Random Forest. Puis, la mesure classique d'importance des variables (calculée selon l'indice moyen d'impureté de Gini sur chaque arbre), proposée par l'implémentation de sklearn nous a permis de sélectionner un sous-ensemble de caractéristiques. La deuxième méthode de réduction de dimension est l'analyse en composantes principales (ACP) [2]. Le nombre de composantes et le nombre des caractéristiques les plus importantes à conserver doit être optimisé.

3.4.3 Classe de rejet fondée sur un CNN

Du fait de l'utilisation d'un réseau de neurones dans l'application de notre méthode de classification avec classe de rejet, nous avons choisi de comparer notre approche à celle de l'article STEFANO et collab. [2000], décrite dans la partie état de l'art du chapitre 3.1 et que nous appellerons la méthode CNNro. Nous avons également comparé notre approche avec deux méthodes simples de rejet d'un échantillon avec un CNN : les méthodes sont appelées CNNa et CNNr.

2. Approche géométrique et statistique consistant à transformer des variables corrélées en nouvelles variables décorrélées les unes des autres (appelées composantes principales).

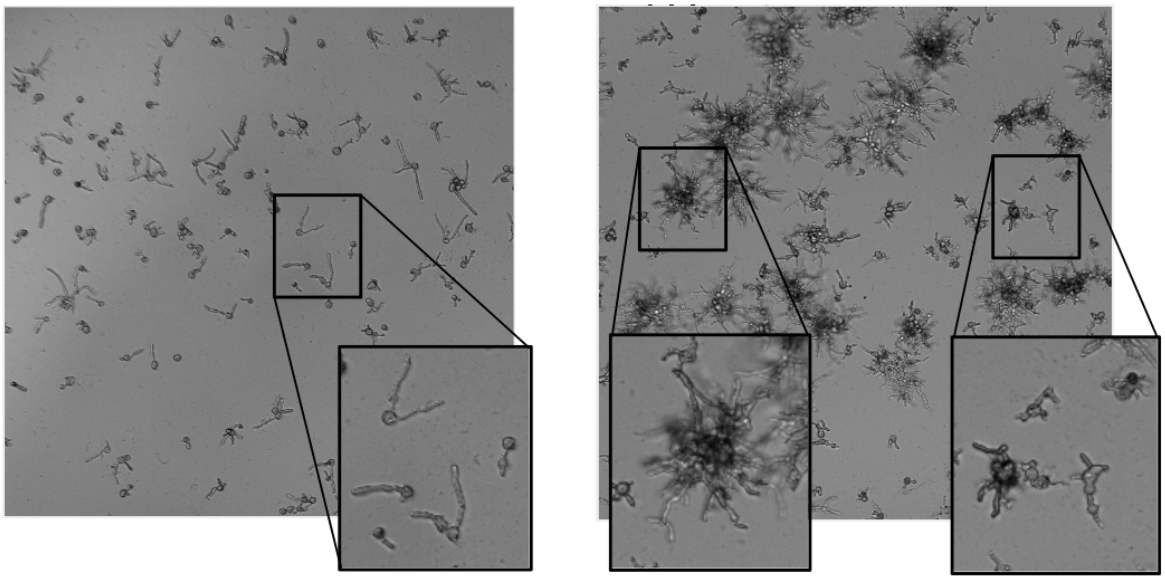


FIGURE 3.14 – Exemple de deux phénotypes différents considérés comme nouveaux phénotypes car ne faisant pas partie des sept classes sur lesquelles le CNN a été entraîné.

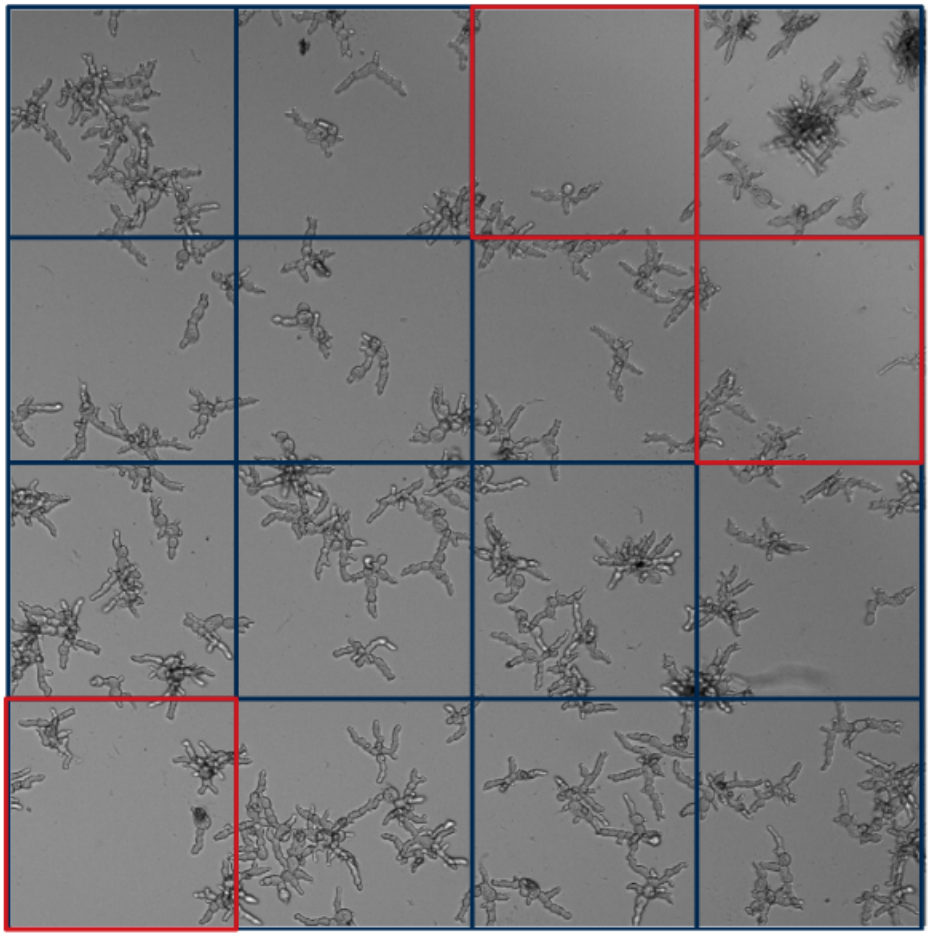


FIGURE 3.15 – Illustration du découpage en vignettes d'une image dont l'information relative aux champignons est répartie de manière hétérogène entraînant des sous-images quasiment vides (vignettes encadrées en rouge).

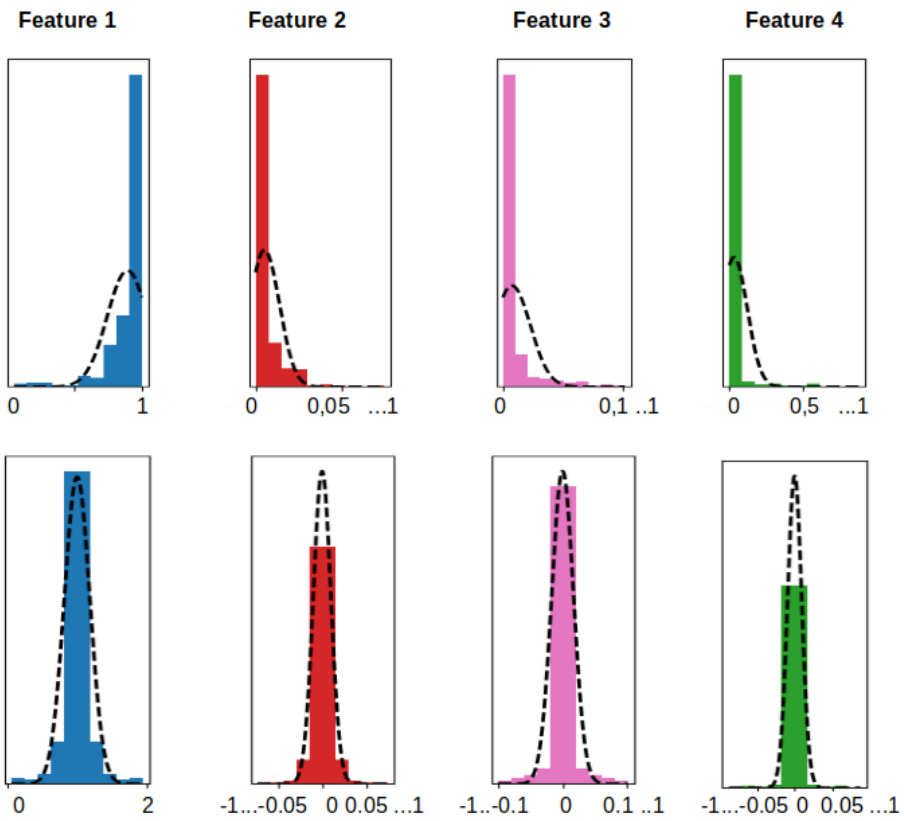


FIGURE 3.16 – Exemple de symétrisation de données Output à quatre dimensions (données illustrées ici seulement de manière indépendante par dimension). Estimation de gaussiennes sur les histogrammes des quatre caractéristiques des échantillons d'une classe donnée : la première et deuxième ligne représentent respectivement les données originales et les données symétrisées. Les Gaussiennes se rapprochent mieux des données lorsqu'elles sont estimées sur les données symétrisées.

Méthode CNNro

Pour rappel, l'option de rejet est fondée sur une fonction définie pour estimer la fiabilité de classification d'un échantillon. Un seuil est optimisé via cette fonction afin de rejeter un maximum d'échantillons mal classés tout en conservant le plus d'échantillons correctement classés. Cela correspond à l'optimisation d'une mesure de performance P et d'un seuil σ. En pratique, les auteurs utilisent un ensemble de deux fonctions de fiabilité, appliquées en cascade et associées à deux seuils. La première fonction, Ψ_A, correspond à la valeur maximale du vecteur de sortie de la dernière couche du réseau. La seconde fonction, Ψ_B, correspond à la différence entre les deux plus grandes valeurs de ce vecteur. Un échantillon est rejeté si $\Psi_A(s) < \sigma_A$, et sinon si $\Psi_B(s) < \sigma_B$. Le seuil σ_A est optimisé en fonction de P sur la totalité de l'ensemble d'apprentissage et le seuil σ_B sur les échantillons non rejetés par Ψ_A. Cette optimisation dépend du paramètre appelé coût normalisé C_N (voir le paragraphe 3.4.4).

Méthodes CNNa et CNNr

— Seuillage absolu (CNNa) : Si la plus grande valeur du vecteur de sortie de la dernière couche du réseau est inférieure à un certain seuil, l'échantillon est rejeté. Dans le cas contraire l'échantillon est classé comme appartenant à la classe correspondant à la plus haute valeur du vecteur output.

— Seuillage relatif (CNNr) : Même principe que le seuillage absolu mais appliqué à la différence entre les deux plus hautes valeurs du vecteur.

Les échantillons de la classe "nouveaux phénotypes" ne sont pas utilisés pour l'apprentissage. Au lieu de cela, chacune des quatre classes connues est utilisée à tour de rôle comme nouvelle. Par exemple si la classe 4 est considérée comme la classe contenant de nouveaux échantillons, les trois autres (1, 2 et 3) sont les classes connues. Pour chaque combinaison, un classifieur CNN doit être optimisé pour la reconnaissance des trois classes connues. Pour cela, nous avons fait du "transfert learning" et une fois le réseau optimisé, un seuil $T_{CNN\star}(1,2,3||4)$ est calculé de façon à maximiser le score moyen de classification sur d'une part les classes connues et d'autre part la classe inconnue. Au final, quatre seuils ont été appris : $T_{CNN\star}(\{1,2,3,4\} \setminus \{i\}||i)$ où $i \in [1..4]$. Le seuil final $T_{CNN\star}$ est calculé en tant que fonction du $T_{CNN\star}(\cdots||\cdot)$. Dans notre application, la médiane conduit aux meilleurs résultats.
Les méthodes CNNa et CNNr correspondent à des simplifications triviales de CNNro.

3.4.4 Résultats

Méthode proposée

Les résultats expérimentaux ont été moyennés sur une validation croisée de sous-échantillonnages aléatoires répétée de 20 fois. Pour rappel, les images sont représentées par les caractéristiques Bottlenecks et Output. En ce qui concerne les Bottlenecks, une réduction de dimension a été effectuée en utilisant l'importance des variables ou une ACP. La dimension réduite optimale a été sélectionnée dans l'intervalle [10, 300]. Pour les Outputs, les données sont symétrisées ou non. Concernant le modèle GMM, nous avons testé différents types de paramètres de covariance et nous avons conservé ceux qui donnent les meilleurs résultats. Notez que le nombre de composantes gaussiennes a été optimisé globalement pour toutes les classes (dans l'intervalle [1, 11]). Les seuils sur les PDFs ont été estimés à l'aide d'une régression logistique ou de la misclassification, dans un contexte "1-vs-all" ou "1-vs-1" (voir le tableau 3.5).

Méthode CNNro

La méthode proposée est comparée à la méthode CNNro (voir le paragraphe 3.4.3). Pour définir les deux seuils σ_A et σ_B, nous avons testé une plage de valeurs pour le coût normalisé C_N, qui pour rappel est donné par :

$$C_N = \frac{C_e}{C_r} - 1 \tag{3.9}$$

où C_e est le coût d'une mauvaise prédiction et C_r le coût du rejet d'un échantillon. À l'aide d'une validation croisée sur 50 folds, nous avons conclu que les meilleurs résultats de classification étaient obtenus pour C_N égal à 25. Les valeurs optimales pour σ_A et σ_B sont respectivement 0.94 et 0.91.

Méthodes CNNr et CNNa

La méthode proposée est également comparée aux méthodes CNNr et CNNa (voir le paragraphe 3.4.3). Lors de la phase d'apprentissage des seuils de ces deux méthodes. Les seuils calculés sont les suivants : $\{0.70, 0.72, 0.84, 0.98\}$ pour $T_{CNNa}(\cdots|\cdot)$, et $\{0.49, 0.52, 0.73, 0.96\}$

pour $T_{CNNr}(\cdots|\cdot)$. Dans les deux cas, on observe une grande variabilité entre les seuils, reflétant la différence de chevauchement et l'écart entre les quatre classes. En effet, dans le cas où la classe de rejet est éloignée des trois autres classes, les valeurs du vecteur de la dernière couche du réseau sont logiquement assez faibles, cela expliquant les faibles valeurs de seuils. Au contraire, si la classe de rejet chevauche de manière significative les classes connues, la sortie du CNN pour ses échantillons sera certainement plus proche des sorties des échantillons de ces classes. C'est-à-dire que la couche de sortie présentera un neurone avec une valeur proche de un et les autres avec une valeur proche de zéro. Dans ce cas, la procédure de rejet est "forcée" d'appliquer un seuil haut. En pratique, la prise de la médiane de ces seuils conduit aux meilleurs résultats. Nous avons donc utilisé $T_{CNNa} = 0.81$ et $T_{CNNr} = 0.675$.

Methods	**Caractéristiques**		**Résultats**	**Apprentissage du seuil**			
				Régression		*Misclassif.*	
				`1-vs-all`	`1-vs-1`	`1-vs-all`	`1-vs-1`
Proposée	Bottleneck (GMM matrice de covariance sphérique)	Importance var.	Connu	94	88	**91**	86
			Nouveau	77	87	**98**	91
			Dimension	90	90	**90**	90
			Nb composents	10	6	**10**	6
		ACP	Connu	93	92	88	90
			Nouveau	82	85	90	89
			Dimension	30	30	30	30
			Nb composents	10	11	11	10
	Output (GMM matrice de covariance diagonale)	Originale	Connu	97	97	78	95
			Nouveau	72	73	75	71
			Dimension	7			
			Nb composents	6	11	11	10
		Symétrique	Connu	97	**96**	52	73
			Nouveau	71	**76**	91	79
			Dimension	**7**			
			Nb composents	7	**10**	9	8
CNNro	N/A		Connu	65			
			Nouveau	100			
CNNa			Connu	65			
			Nouveau	81			
CNNr			Connu	70			
			Nouveau	75			
CNNmax			Connu	91			

TABLEAU 3.5 – Précisions de classification en pourcentages pour les différents paramètres expérimentaux. Les valeurs réelles ont été arrondies à l'entier le plus proche.

3.4.5 Discussion

Avec la méthode CNNro, les scores sont d'environ 65% sur les classes connues et 100% sur la classe de rejet. Notre méthode de classification avec classe de rejet obtient un bien meilleur score de classification sur les classes connues avec une moyenne de 91% pour le choix optimal parmi les variantes, contre 65% pour la méthode CNNro, soit une différence de plus de 26%. En revanche sur la classe de rejet, avec 100% contre 98% la méthode CNNro classe mieux les éléments à rejeter. Étant donné que la méthode proposée est également plus équilibrée (7% d'écart entre classes connues et classe de rejet, contre 35%

pour CNNro), nous pouvons conclure que notre méthode est plus efficace en termes de classification avec classe de rejet.
Avec la méthode CNNa et CNNr, les résultats de classification sont respectivement de 65% et 70% sur les classes connues et de 81% et 75% sur la classe de rejet. Au regard de ces scores, nous pouvons attester que notre méthode est plus efficace que ces deux approches simples pour définir une classe de rejet avec un classificateur CNN. En effet, elle offre des précisions supérieures de 10 à 25% pour les classes connues et de rejet. Nous pouvons également relever l'égalité obtenue avec le score du CNN classiquement appris sur les classes connues (désigné par CNNmax) avec une précision de 91% (voir le tableau 3.5).

3.4.6 Conclusion

Nous avons proposé une méthode de classification avec classe de rejet fondée sur l'apprentissage de modèles dans un contexte supervisé. Elle suit une stratégie générale fondée sur trois étapes principales : apprentissage de modèle indépendamment pour chaque classe, apprentissage d'un seuil fondé sur les interactions de classes et procédure de prédiction. Dans l'application étudiée, la méthode proposée a fourni de très bonnes précisions de classification pour la classe de rejet et les classes connues. Les meilleurs résultats sont obtenus avec les données Bottlenecks dont la dimension est réduite à 100 caractéristiques avec la méthode de réduction de dimension selon l'importance des variables. Les paramètres du mélange de gaussiennes sont optimisés avec une matrice de covariance sphérique, dix composantes et la méthode misclassification avec l'option "1-vs-all" pour l'apprentissage des seuils.
Concernant les différences pratiques entre les deux types de caractéristiques, avec les Bottlenecks, une réduction de la dimension sera souvent nécessaire. Cependant, si une nouvelle classe est ajoutée, ces caractéristiques n'auront pas à être réapprises. Au contraire, dans le cas des données Output, la dimension est raisonnable mais le CNN doit être obligatoirement ré-entrainé lors de l'ajout de nouvelles classes.

3.5 De la vignette à la molécule : procédure globale de classification

3.5.1 Une classification en deux étapes

Les résultats obtenus dans le chapitre 2, présentés dans le tableau 2.11, démontrent que le CNN, est assez efficace pour bien distinguer les images des classes mycelium, MiniGermTube et Crystal. Nous partons du principe que si l'image correspond à un nouveau phénotype remarquable (voir la figure 3.14), le CNN ne le classera pas dans l'une de ces classes. Ainsi, nous avons décidé d'effectuer la classification des images en deux étapes (voir la figure 3.18). Dans la première étape, les images sont classées dans quatre classes sur la base de la sortie du CNN (voir la figure 3.17). Trois classes dites non-phénotypes (Germination-Inhibition, Crystal et Mycelium) ainsi qu'une classe Phénotypes, regroupant sans disctinction les phénotypes connus (1-4) et nouveau. Pour cela un seuil est fixé expérimentalement et les prédictions dans les classes Germination-Inhibition, Crystal et mycelium, avec un indice de confiance supérieur à 0.8 constituent les prédictions finales de ces images. L'ensemble des autres images sont prédites dans la classe Phénotypes et seules ces images passeront par la deuxième étape de classification. Les résultats de la première étape de classification sont regroupés dans le tableau 3.6.

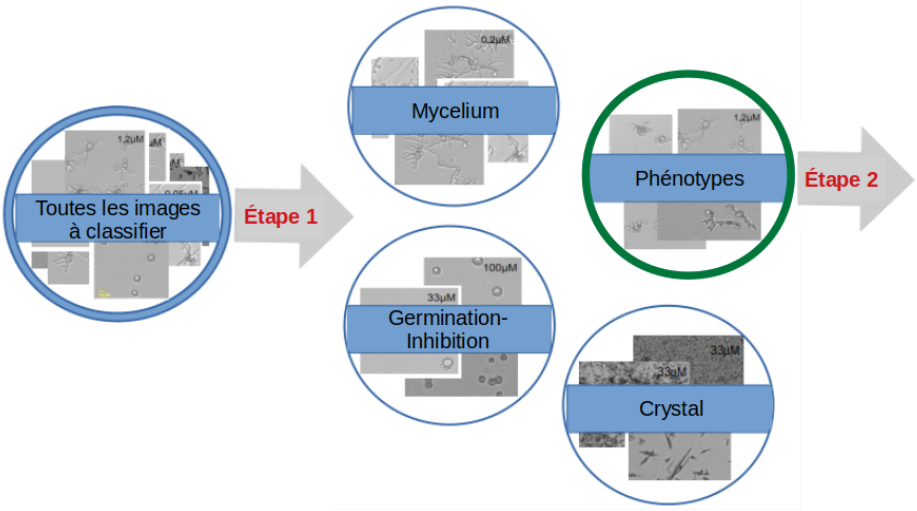


FIGURE 3.17 – Classification en deux étapes. Étape 1 : classification des images dans les classes Germination-Inhibition, Crystal, mycelium et Phénotypes en fonction des scores obtenus par le CNN. Étape 2 : classification des images dans les classes Phénotypes 1-4 et Nouveau en fonction des réponses des modèles GMM et des seuils appris.

	G-I	Crystal	mycelium	Phénotypes
G-I	100	0	0	0
Crystal	3	65	5	27
mycelium	0	0	100	0
Phénotypes	0	0	0	100

TABLEAU 3.6 – Matrice de confusion des résultats de classification (en pourcentages) des images. (G-I correspond à Germination-Inhibition)

Ces résultats montrent une très bonne efficacité de prédiction des images des classes Germination-Inhibition, mycelium et Phénotypes. Concernant la classe Crystal, les résultats sont satisfaisants au vu de la complexité et de l'hétérogénéité des images de cette classe (voir le paragraphe A.5 en annexe). Les images de cristaux sont discutées plus en détails dans le paragraphe 2.5.4.
Les résultats de la ligne "Phénotypes" du tableau 3.6 (qui correspond à la prédiction d'un ensemble de test composé de phénotypes 1 à 4 mais aussi de phénotypes nouveaux), montrent que le CNN a effectivement tout mis dans la classe Phénotypes. L'hypothèse mentionnée au début, qui était que si l'image correspond à un nouveau phénotype remarquable, le CNN ne la classera pas dans l'une des classes non-phénotypes, est donc confirmée.

3.5.2 Règles de prédiction du MoA d'une molécule

Concernant les classes Germination-Inhibition, Crystal, Mycelium et Phénotypes 1-4, les règles de prédiction du MoA sont décrites dans le paragraphe 2.5.2. Dans cette partie du manuscrit, nous allons apporter les modifications nécessaires pour prendre en compte les nouveaux phénotypes.

Pour rappel, chaque molécule antifongique étudiée est testée à dix concentrations, donc 10 puits (de 100 à 0,005µM) et plusieurs images par puits sont acquises par microscopie à lumière transmise (voir les figures 2.14 et 2.15). Ainsi, une fois l'ensemble des

images prédites et dans le but d'émettre une hypothèse de mode d'action de la molécule testée, les prédictions des images sont rassemblées et sont soumises à une cascade de prédiction (voir la figure 2.37). Ainsi, pour chaque concentration, la prédiction finale du puits correspond à la réponse majoritaire parmi les prédictions des images du puits. Un indice de confiance est par ailleurs attribué à cette prédiction. Cet indice correspond au nombre d'occurrences de cette réponse sur le nombre total d'images du puits. Les prédictions finales des images des puits contrôles permettent uniquement de vérifier que l'expérience biologique s'est correctement déroulée. Dans le cas où les prédictions indiquent une autre classe que mycelium, le programme prévient l'utilisateur d'une éventuelle erreur de manipulation lors de l'élaboration de la plaque test. Concernant les autres puits, les prédictions finales sont soumises aux mêmes règles de prédiction que celles décrites dans le paragraphe 2.5.2 à l'exception des cas dans lesquels on retrouve à la fois des prédictions phénotype connu et nouveau. Ainsi, les règles appliquées sont les suivantes :

- Si aucun phénotype n'est prédit à aucune concentration, la molécule est prédite « No Answer » (NA)car on ne peut conclure sur le MoA de la molécule, avec la particularité que si toutes les prédictions sont « Mycelium » (aucun effet de la molécule), la molécule est prédite « No molecule effect ».

- Dans le cas contraire, seules les prédictions phénotypes (1,2,3,4 et nouveau) sont prises en compte dans la prédiction du MoA de la molécule :
- Si la/les prédiction(s) indiquent la classe nouveau phénotype alors le MoA prédit sera « nouveau MoA ».
- Si la/les prédiction(s) sont des phénotypes connus alors le MoA prédit sera le plus représentés de ces phénotypes connus (exemple, 3 puits Phénotype 1 et 1 puits Phénotype 2, alors le 1 l'emporte).
- S'il y a des prédictions phénotype connu et nouveau alors on se réfère aux indices de confiance. Dans ce cas, l'indice de confiance moyen ic des prédictions du phénotype connu est comparé à un seuil s fixé expérimentalement :
  - $ic \geq s$: Prédiction finale = phénotype connu et le MoA = le MoA correspondant
  - $ic < s$: Prédiction finale = nouveau phénotype et le MoA = « nouveau MoA »

3.5.3 Résultats (Hypothèse de MoA)

Pour rappel, les meilleurs résultats de la deuxième étape de classification sont obtenus avec les données Bottlenecks dont la dimension est réduite à 100 caractéristiques (voir le tableau 3.5). Ce sont ces prédictions qui sont prises en compte dans l'élucidation des MoA des molécules testées. En appliquant les règles de prédictions décrites dans le paragraphe 3.5.2 et en les comparant à la vérité terrain, nous obtenons les scores présentés dans le tableau 3.7 suivant :

Mode d'action	1	2	3	4	Nouveau	NA
Score	100	100	100	100	95	99

TABLEAU 3.7 – Résultats de classification (en pourcentages) des modes d'action des molécules testées obtenus suivant les règles de prédiction citées dans le paragraphe 3.5.2. Les six classes de MoAs sont les quatre MoA connus (1-4), la classe nouveau MoA et la classe NA (« No Answer ») qui contient les molécules dont les images ne permettent pas d'émettre une hypothèse de MoA.

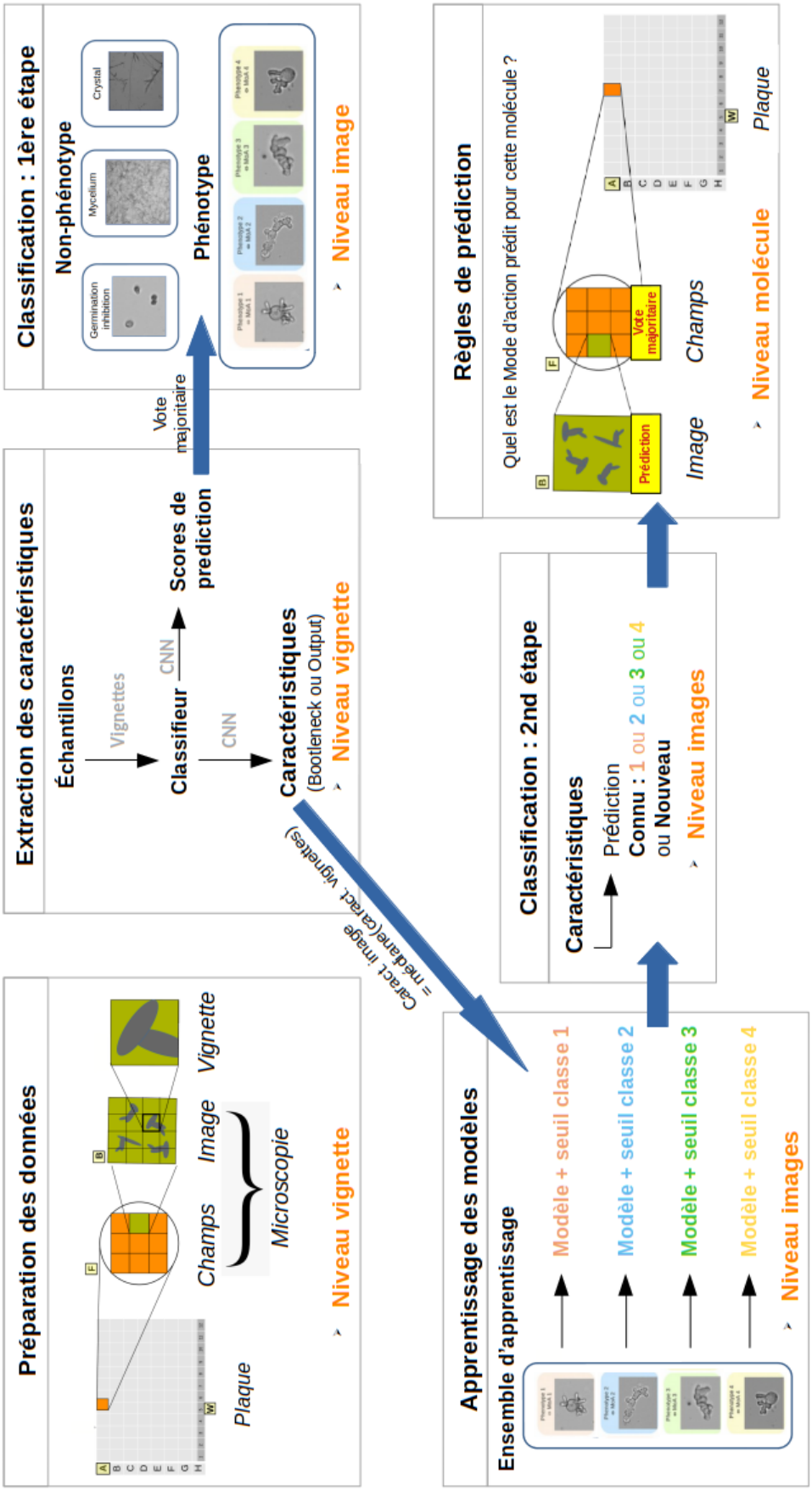


FIGURE 3.18 – Schéma général de classification.

3.5.4 Discussions et conclusions

D'une part la classification en deux étapes nous permet d'obtenir de très bons scores de classification sur les images des sept différentes classes. D'autre part, les précisions sur la prédiction du MoA des molécules testées sont très élevées avec une moyenne de 100% sur les MoAs connus, 95% sur les nouveaux MoAs et 99% sur les molécules dont on ne peut élucider le MoA *via* les images dont nous disposons (NA).

En conclusion, la méthode proposée de classification avec classe de rejet associée aux règles de prédiction établies permet l'identification des MoAs des molécules anti-fongiques sur le modèle biologique *Botrytis cinerea* avec une grande efficacité.

3.6 Références

ALTMAN, N. S. 1992, «An introduction to kernel and nearest-neighbor nonparametric regression», *The American Statistician*, vol. 46, n° 3, p. 175–185. 88

BREIMAN, L. 2001, «Random forests», *Machine learning*, vol. 45, n° 1, p. 5–32. 88

BREW, A., M. GRIMALDI et P. CUNNINGHAM. 2007, «An evaluation of one-class classification techniques for speaker verification», *Artificial Intelligence Review*, vol. 27, n° 4, p. 295–307. 83

CHANDOLA, V., A. BANERJEE et V. KUMAR. 2009, «Anomaly detection : A survey», *ACM computing surveys (CSUR)*, vol. 41, n° 3, p. 1–58. 81

CHANG, C.-C. et C.-J. LIN. 2011, «Libsvm : a library for support vector machines», *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, n° 3, p. 1–27. 88

CHOW, C. 1970, «On optimum recognition error and reject tradeoff», *IEEE Transactions on information theory*, vol. 16, n° 1, p. 41–46. 79

CHOW, C.-K. 1957, «An optimum character recognition system using decision functions», *IRE Transactions on Electronic Computers*, , n° 4, p. 247–254. 79

CONSUL, P. C. et G. C. JAIN. 1973, «A generalization of the poisson distribution», *Technometrics*, vol. 15, n° 4, p. 791–799. 81

CORTES, C., G. DESALVO et M. MOHRI. 2016, «Learning with rejection», dans *International Conference on Algorithmic Learning Theory*, Springer, p. 67–82. 82

DUBUISSON, B. et M. MASSON. 1993, «A statistical decision rule with incomplete knowledge about classes», *Pattern recognition*, vol. 26, n° 1, p. 155–165. 79

EDDY, S. R. 2004, «What is a hidden markov model?», *Nature biotechnology*, vol. 22, n° 10, p. 1315–1316. 81

GEIFMAN, Y. et R. EL-YANIV. 2017, «Selective classification for deep neural networks», *arXiv preprint arXiv :1705.08500*. 79

GEIFMAN, Y. et R. EL-YANIV. 2019, «Selectivenet : A deep neural network with an integrated reject option», *arXiv preprint arXiv :1901.09192*. xiii, 82, 83

GRANDVALET, Y., A. RAKOTOMAMONJY, J. KESHET et S. CANU. 2009, «Support vector machines with a reject option», dans *Advances in neural information processing systems*, p. 537–544. 82

HASTIE, T., S. ROSSET, J. ZHU et H. ZOU. 2009, «Multi-class adaboost», *Statistics and its Interface*, vol. 2, nº 3, p. 349–360. 88

HELLMAN, M. E. 1970, «The nearest neighbor classification rule with a reject option», *IEEE Transactions on Systems Science and Cybernetics*, vol. 6, nº 3, p. 179–185. 79

HENRION, M., D. J. HAND, A. GANDY et D. J. MORTLOCK. 2013, «Casos : a subspace method for anomaly detection in high dimensional astronomical databases», *Statistical Analysis and Data Mining : The ASA Data Science Journal*, vol. 6, nº 1, p. 53–72. 79

HOFFMANN, H. 2007, «Kernel pca for novelty detection», *Pattern recognition*, vol. 40, nº 3, p. 863–874. 80

HOMENDA, W., M. LUCKNER et W. PEDRYCZ. 2016, «Classification with rejection : concepts and evaluations», dans *Knowledge, Information and Creativity Support Systems : Recent Trends, Advances and Solutions*, Springer, p. 413–425. 82

ILONEN, J., P. PAALANEN, J.-K. KAMARAINEN et H. KALVIAINEN. 2006, «Gaussian mixture pdf in one-class classification : computing and utilizing confidence values», dans *18th International Conference on Pattern Recognition (ICPR'06)*, vol. 2, IEEE, p. 577–580. 81

KALINICHENKO, L., I. SHANIN et I. TARABAN. 2014, «Methods for anomaly detection : A survey», dans *CEUR Workshop Proceedings*, vol. 1297, p. 2025. 83

KIM, J. et C. D. SCOTT. 2012, «Robust kernel density estimation», *The Journal of Machine Learning Research*, vol. 13, nº 1, p. 2529–2565. 81

MARKOU, M. et S. SINGH. 2003, «Novelty detection : a review—part 2 : : neural network based approaches», *Signal processing*, vol. 83, nº 12, p. 2499–2521. 81

MUKHERJEE, S., P. TAMAYO, D. SLONIM, A. VERRI, T. GOLUB, J. MESIROV et T. POGGIO. 1999, «Support vector machine classification of microarray data», cahier de recherche, AI Memo 1677, Massachusetts Institute of Technology. 80

NAIRAC, A., T. A. CORBETT-CLARK, R. RIPLEY, N. W. TOWNSEND et L. TARASSENKO. 1997, «Choosing an appropriate model for novelty detection», . 83

PAALANEN, P., J.-K. KAMARAINEN, J. ILONEN et H. KÄLVIÄINEN. 2006, «Feature representation and discrimination based on gaussian mixture model probability densities—practices and algorithms», *Pattern Recognition*, vol. 39, nº 7, p. 1346–1358. 81

PIMENTEL, M. A., D. A. CLIFTON, L. CLIFTON et L. TARASSENKO. 2014, «A review of novelty detection», *Signal Processing*, vol. 99, p. 215–249. 79

RATSCH, G., S. MIKA, B. SCHOLKOPF et K.-R. MULLER. 2002, «Constructing boosting algorithms from svms : An application to one-class classification», *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, nº 9, p. 1184–1199. 80

SCHWARZ, G. et collab.. 1978, «Estimating the dimension of a model», *Annals of statistics*, vol. 6, nº 2, p. 461–464. 93

SIEGISMUND, D., V. TOLKACHEV, S. HEYSE, B. SICK, O. DUERR et S. STEIGELE. 2018, «Developing deep learning applications for life science and pharma industry», *Drug research*, vol. 68, nº 06, p. 305–310. 78

SPERANDEI, S. 2014, «Understanding logistic regression analysis», *Biochemia medica : Biochemia medica*, vol. 24, nº 1, p. 12–18. 85

STEFANO, C. D., C. SANSONE et M. VENTO. 2000, «To reject or not to reject : that is the question-an answer in case of neural classifiers», *IEEE Trans. Systems, Man, and Cybernetics, Part C*, vol. 30, p. 84–94. 81, 102

SUGIYAMA, M. et K. BORGWARDT. 2013, «Rapid distance-based outlier detection via sampling», *Advances in Neural Information Processing Systems*, vol. 26, p. 467–475. 79

TARASSENKO, L., A. HANN et D. YOUNG. 2006, «Integrated monitoring and analysis for early warning of patient deterioration», *BJA : British Journal of Anaesthesia*, vol. 97, nº 1, p. 64–68. 79, 80

THOM, H. C. 1958, «A note on the gamma distribution», *Monthly Weather Review*, vol. 86, nº 4, p. 117–122. 81

THOMPSON, B. B., R. J. MARKS, J. J. CHOI, M. A. EL-SHARKAWI, M.-Y. HUANG et C. BUNJE. 2002, «Implicit learning in autoencoder novelty assessment», dans *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No. 02CH37290)*, vol. 3, IEEE, p. 2878–2883. 80

WIENER, Y. et R. EL-YANIV. 2011, «Agnostic selective classification», dans *Advances in neural information processing systems*, p. 1665–1673. 82

ZHAO, Y., S. WANG et F. XIAO. 2013, «Pattern recognition-based chillers fault detection method using support vector data description (svdd)», *Applied Energy*, vol. 112, p. 1041–1048. 80

ZIYIN, L., Z. WANG, P. P. LIANG, R. SALAKHUTDINOV, L.-P. MORENCY et M. UEDA. 2019, «Deep gamblers : Learning to abstain with portfolio theory», *arXiv preprint arXiv :1907.00208*. 82

Chapitre 4

Une approche de classification par transport optimal

4.1 État de l'art

La théorie du Transport Optimal (OT) commence avec un problème formulé au 18ème siècle par Gaspard Monge qui se demandait comment déplacer un tas de sable vers un trou en fournissant le moins d'effort possible (théorie des déblais et des remblais). Le transport optimal permet de définir une notion de distance entre deux distributions de probabilités et permet de transformer une distribution en une autre à moindre coût. Dans ce problème, le coût engendré par le transport d'une unité de masse d'un point x vers un point y est égal à la distance euclidienne entre ces points. Nous définissons deux espaces

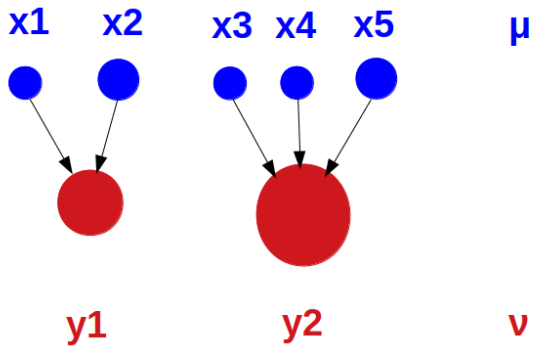


FIGURE 4.1 – Cas d'un transport d'une distribution discrète μ (points x_i) vers une distribution ν (points y_j). Les masses des points sont représentées via leurs aires et le transport est tel que $T(x1) = T(x2) = y1$ et $T(x3) = T(x4) = T(x5) = y2$.

X et Y, et nous considérons μ et ν deux mesures de probabilité sur chacun de ces espaces, respectivement. Ainsi T transporte μ sur ν si pour tout ensemble mesurable y de Y on a :

$$\mu(T^{-1}(y)) = \nu(y) \tag{4.1}$$

et on note $T\#\mu = \nu$.

Dans ce cadre, on se pose le problème de trouver la transformation T qui minimise un cout de transport entre μ et ν :

$$T \in \arg\min_{T\#\mu=\nu} \left\{ \int_X c(x, T^\star(x)) d\mu(x) \right\} \tag{4.2}$$

où $c(x, T(x))$ représente le coût du transport de x vers $T(x)$.

Par exemple, le coût du transport T d'une unité de masse x à $y = T(x)$ peut être donné par la distance euclidienne entre x et y. Ainsi, le coût total du transport est :

$$C(T) = \int_X |x - T(x)| d\mu(x) \tag{4.3}$$

On cherche ainsi à déterminer :

$$C(\mu, \nu) = \inf\{C(T); T : X \to Y, T\#\mu = \nu\} \tag{4.4}$$

La notion de distance entre des distributions de probabilité induite par le transport optimal constitue l'un des principaux intérêts de son utilisation.

Plusieurs points font du problème de Monge, un problème extrêmement difficile à résoudre :

— Nature combinatoire du problème,

— Questions d'existence et d'unicité de la solution (Problème non convexe).

Voici deux configurations qui posent problème :

— Si on a deux droites perpendiculaires, on trouve deux solutions (même effort calculé),

— Si on a une seule masse dans l'espace source et deux dans l'espace cible, il n'existe pas de solution.

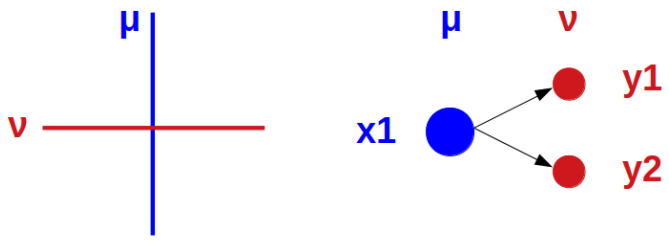


FIGURE 4.2 – A gauche : Cas de deux droites perpendiculaires. À droite : Cas d'un nombre de masses dans l'espace source inférieur à celui dans l'espace cible.

Il fallut attendre près de 200 ans pour qu'une nouvelle formulation apparaisse dans les années 40 et révolutionne la théorie du transport optimal : la formulation relaxée de Kantorovich. C'est durant la seconde guerre mondiale qu'un mathématicien russe, Leonid Kantorovitch, expert en optimisation d'allocation de ressources propose une solution à ce problème permettant un calcul efficace du transport optimal T. Ainsi, comme décrit plus haut, Monge exige que la matière au point x d'une distribution μ doit être transportée entièrement vers $T(x)$, un point dans la distribution cible ν. Kantorovich de son côté,

propose d'autoriser la distribution de cette matière sur différents points dans la cible. Cette étude lui a valu le prix Nobel d'économie.

Le plan de transport correspond à une distribution de probabilités conjointe $\gamma(x, y)$ qui indique la proportion de matière au voisinage de x dans μ que l'on va transporter au voisinage de y dans ν. Ce plan doit vérifier la contrainte selon laquelle toute la masse provenant de X et arrivant en Y doit être égale à la masse de ν(Y) de Y. L'inverse doit également être vérifié : toute la masse arrivant en Y et provenant de X doit être égale à la masse de μ(X) de X. On cherche donc le plan de transport γ :

$$\gamma \in \arg\min \left\{ \int_{x \in X} \int_{y \in Y} c(x, y) d\gamma(x, y) \right\} \tag{4.5}$$

Sous condition que : $\gamma \in P = \left\{ \gamma \geq 0, \int_Y \gamma(x, y) dy = \mu, \int_X \gamma(x, y) dx = \nu \right\}$

γ est donc une distribution de probabilité jointe avec les marginales μ et ν. Ce plan de transport constitue la solution du problème. L'avantage principal de la formulation de Kantorovich, est qu'il existe toujours un couplage optimal bien que pas forcément unique NATHAEL GOZLAN [16 mars 2018].

Une deuxième avancée, apparue à la même époque et faite par George Dantzig DANTZIG [1951] a permis de rendre le TO applicable à des problèmes de grande taille. En effet, Dantzig propose l'algorithme du simplexe qui permet de résoudre efficacement des problèmes d'optimisation linéaire (convexe) consistant à minimiser une fonction linéaire de variables réelles, ce qui est le cas du problème de Kantorovitch. On note P_n l'ensemble des $n!$ matrices de permutation de taille $n \times n$. Kantorovitch permet la relaxation de la contrainte suivant laquelle les matrices sont contraintes à être dans $\{0, 1\}^{n \times n}$ en choisissant les entrées de P comme étant entre 0 et 1. L'ensemble B_n de matrices bistochastiques remplace donc l'ensemble plus petit P_n (voir l'équation (4.6)). Dans l'équation (4.7) à résoudre, on doit payer un coût $C_{x,y}$ à chaque fois que l'on transfert une unité de masse entre x et y.

$$B_n := \left\{ P \in [0, 1]^{n \times n} : \forall x, \sum_y P_{x,y} = 1, \forall y, \sum_x P_{x,y} = 1 \right\} \tag{4.6}$$

$$W(\mu, \nu) := \sum_{x,y} P_{x,y} C_{x,y} \tag{4.7}$$

Des applications du TO pour la résolution de problèmes appliqués en science des données, notamment des problématiques liées aux domaines du traitement d'image et de l'apprentissage machine, émergent depuis peu. L'une des utilisations du TO consiste à comparer les histogrammes colorimétriques de plusieurs images. Un des but étant par exemple de changer les couleurs de la première image en lui imposant la palette de la deuxième image (voir la figure 4.3). Les images peuvent également être triées selon leur similarité via le calcul de la distance de Wasserstein entre les histogrammes associés. Une autre utilisation est la manipulation de l'histogramme comme dans le cas illustré sur la figure 4.4, dans lequel plus la valeur du paramètre t tend vers 1 plus l'histogramme tend vers une égalisation. Le but étant, ici, de créer une image contrastée.

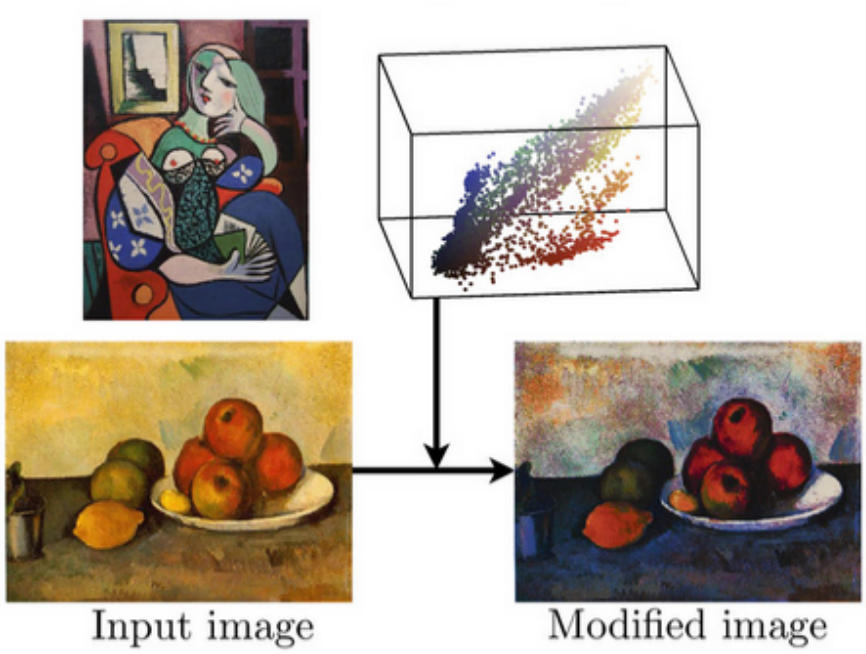


FIGURE 4.3 – Transfert de la palette de couleurs de Picasso sur un tableau de Cézanne : Le transport T est calculé en utilisant un coût $C_{x,y}$ dont la valeur dépend de la similarité entre les couleurs p_x et p_y des deux pixels. Moins les couleurs sont proches plus $C_{x,y}$ est élevé. La valeur du pixel p_x de l'image résultante est remplacée par celle de p_y. – source PEYRÉ [19 janvier 2017]

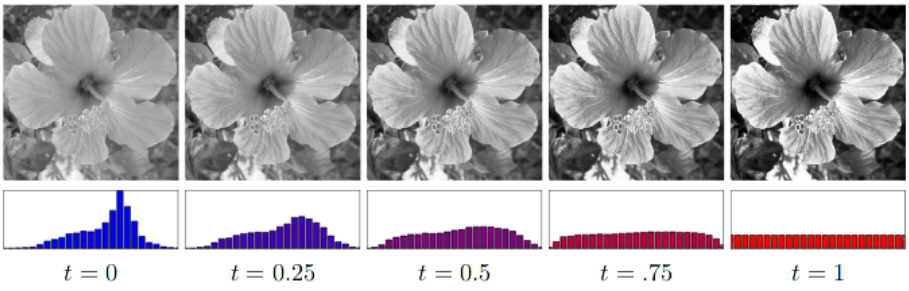


FIGURE 4.4 – Égalisation d'histogramme dans le but de créer une image contrastée : les niveaux de gris sont redistribués par TO en fonction de t qui paramètre l'interpolation du déplacement entre les histogrammes. Ligne du haut : Évolution des images, Ligne du bas : évolution de l'histogramme des niveaux de gris des images correspondantes – source COT [2019].

Le transport optimal permet donc de définir une distance entre deux distributions qui représente un coût minimal pour transformer une distribution en une autre. C'est un outil puissant pour la comparaison de distributions (continues ou discrètes), ou directement de nuages de points. On le retrouve dans des applications en apprentissage machine tel que dans le cas de la classification multi-labels FROGNER et collab. [2015] ou encore le problème d'adaptation de domaine COURTY et collab. [2016]. Dans le premier cas, la fonction de perte de Wasserstein remplace la fonction de perte classique (Hinge ou fonction perte quadratique ou cross-entropy en Deep Learning), l'avantage étant que cette fonction de Wasserstein encode les relations spécifiques entre les différents labels. Dans le second cas, lorsque les bases de données d'apprentissage et de test sont trop différentes et/ou ne sont pas dans le même espace, le classifieur appris ne peut pas être utilisé directement. Calculer le TO entre données sources et cibles et déplacer les points d'apprentissage dans l'espace test, permet le calcul d'un classifieur efficace sur les données test (voir la figure 4.5). Parmi les tâches dans lesquelles les propriétés du transport optimal se sont révélées utiles on retrouve également le recalage d'images HAKER et TANNENBAUM [2001], le traitement du signal KOLOURI et collab. [2017], l'équité en machine learning GORDALIZA et collab. [2019], ou encore en biologie SCHIEBINGER et collab. [2019].

Un autre exemple très populaire de l'application du TO en science des données est la méthode Word Mover Distance (WMD) KUSNER et collab. [2015] très utilisée en traitement du langage qui permet via le TO de comparer des phrases. Dans ce cas, une fonction de coût qui traduit la similitude sémantique entre les mots est définie. Pour cela on peut par exemple utiliser le plongement de mots "Words embedding" qui permet de représenter chaque mot d'un dictionnaire par un vecteur de nombres réels. La distance entre les deux phrases est calculée en trouvant la meilleure mise en correspondance entre chacun des mots dans un premier temps puis en faisant la somme de toutes les distances.

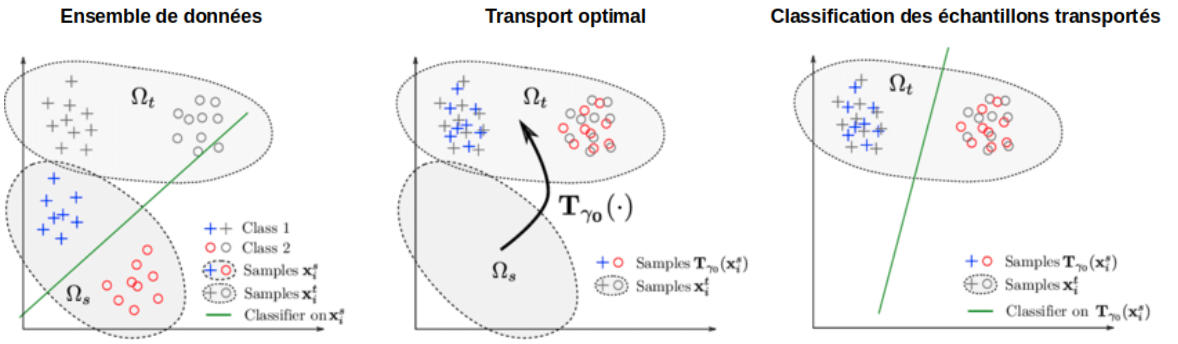


FIGURE 4.5 – Le transport optimal pour l'adaptation du domaine. À gauche : les données d'apprentissage dans le domaine source, et les données tests dans le domaine cible. Le classifieur estimé ne permet pas la classification des données cibles. Au milieu : transport des observations d'apprentissage dans le domaine cible via une carte de transport T ∗ γ0 . La transformation n'est généralement pas linéaire. À droite : calcul d'un classifieur efficace dans le domaine cible. – source COURTY et collab. [2016].

Pour calculer la différence entre deux distributions de probabilités X et Y, nous utilisons la distance de p-Wasserstein définie comme la racine p^e (p-ième) du coût de transport solution du problème de Kantorovitch lorsque $C_{x,y}$ est une distance d mise à la puissance p.

Dans le cas où $X = Y = R^d$, le coût de transport est donné par le carré de la norme euclidienne (voir l'équation (4.8)) et la quantité $W(\mu,\nu)^{1/2}$ est une distance entre les distributions.

$$c(x,y) = ||x-y||^2 \tag{4.8}$$

Afin de pouvoir appliquer le transport à des problèmes pratiques, la quantité $W(\mu,\nu)^{1/2}$ doit vérifier les axiomes d'une distance. En effet, $W(\mu,\nu)$ vérifie que $W(\mu,\nu) = 0$ si et seulement si $\mu = \nu$. De plus, $W(\mu,\nu)^{1/2}$ doit également respecter la symétrie : $W(\mu,\nu) = W(\nu,\mu)$. Enfin, W(A, B) doit vérifier l'inégalité triangulaire $W(A,B) \leq W(A,C) + W(C,B)$. Dans R^n, cela se traduit par : pour aller d'un point A à un point B, il est plus court d'aller tout droit en parcourant le segment [A, B] plutôt que de passer par un point intermédiaire C. L'intérêt de cette distance de p-Wasserstein par rapport à des distances classiques euclidienne ou L_1, est l'évolution de la valeur de cette distance lorsque le support des deux distributions est disjoint. La distance est constante dans le cas de ces deux distances classiques contrairement à la distance de Wasserstein qui est croissante avec l'écart entre les supports des distributions (voir la figure 4.6). Afin d'avoir des algorithmes plus efficaces

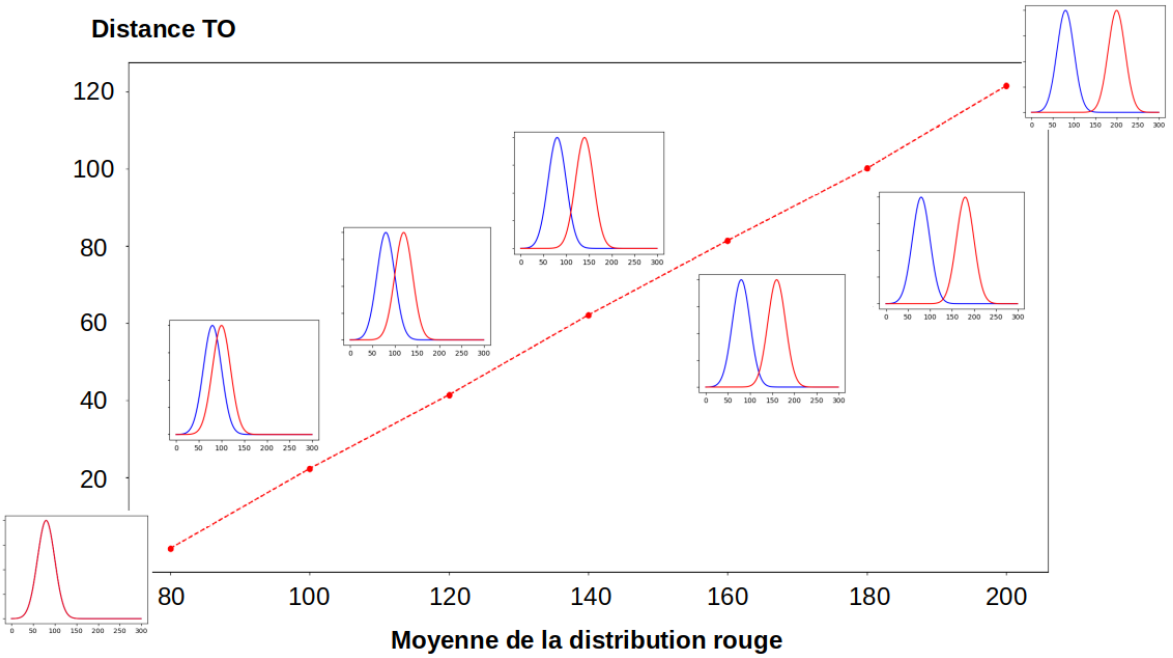


FIGURE 4.6 – Les distances définies par le TO (wasserstein) entre deux distributions de probabilité 1D bleue et rouge. Les moyennes des distributions sont initialisées à 80 puis la distribution rouge se décale.

et plus rapides encore pour résoudre le problème de TO, la technique est d'approcher le problème de départ en le régularisant. C'est-à-dire, qu'on ajoute une contrainte avec une certaine force λ sur la nature du couplage. Il existe plusieurs techniques telles que la régularisation entropique CUTURI [2013], le Group Lasso COURTY et collab. [2016] et la β-divergences DESSEIN et collab. [2018] pour faire cela. C'est la découverte récente d'algorithmes performants qui explique le foisonnement d'applications. L'une des avancées majeures est le développement de l'algorithme de Sinkhorn qui permet de trouver le plan

de TO régularisé $\gamma * (\lambda)$ en utilisant la régularisation entropique CUTURI [2013]. Dans ce cas, on essaye de contrôler l'entropie de ce couplage.

Le problème de transport régularisé est :

$$\gamma \in \arg\min \left\{ \int_{x \in X} \int_{y \in Y} c(x,y) d\gamma(x,y) \right\} + \lambda \Omega(\gamma) \tag{4.9}$$

La régularisation entropique est notée $\Omega(\gamma)$ et vaut :

$$\Omega(\gamma) = \sum_{i,j} \gamma(i,j)(\log \gamma(i,j) - 1) \tag{4.10}$$

γ est une distribution de probabilité jointe, et on peut autoriser que l'entropie de cette distribution augmente si cela favorise la recherche de solution. Sur la Figure 4.7, on voit que pour une très faible valeur du paramètre de régularisation entropique on obtient un couplage très parcimonieux et quand on laisse cette valeur augmenter, la solution du problème s'étale. L'idée n'est pas de trouver la solution exacte au problème de transport. En effet, on ne veut pas être trop attaché aux données pour deux raisons principales. La première est la présence d'outliers et de bruit d'observation. La deuxième est que dans le domaine d'apprentissage machine dans lequel nous sommes, le but est d'appliquer le résultat à de nouvelles données jamais vues. Ainsi, la régularisation lisse effectivement le plan de transport pour une meilleure généralisation à de nouvelles données. Cela a un intérêt surtout lorsque les données manipulées sont en grandes dimensions. Dans le cas de l'algorithme de Sinkhorn par exemple, la complexité théorique passe de $n^3 * \log(n)$ à n^2. Plus on augmente la dimension de nos données plus le transport est instable et plus on doit augmenter cette régularisation.

La solution du couplage (dont l'entropie est contrôlée) est écrite directement comme un nouveau problème qui fait intervenir la matrice de coût :

$$\gamma_0^\lambda = \text{diag}(\mu) \exp^{(\frac{-C}{\lambda})} \text{diag}(\nu) \tag{4.11}$$

On note diag(x) la matrice carrée diagonale dont la diagonale contient les valeurs du vecteur x.

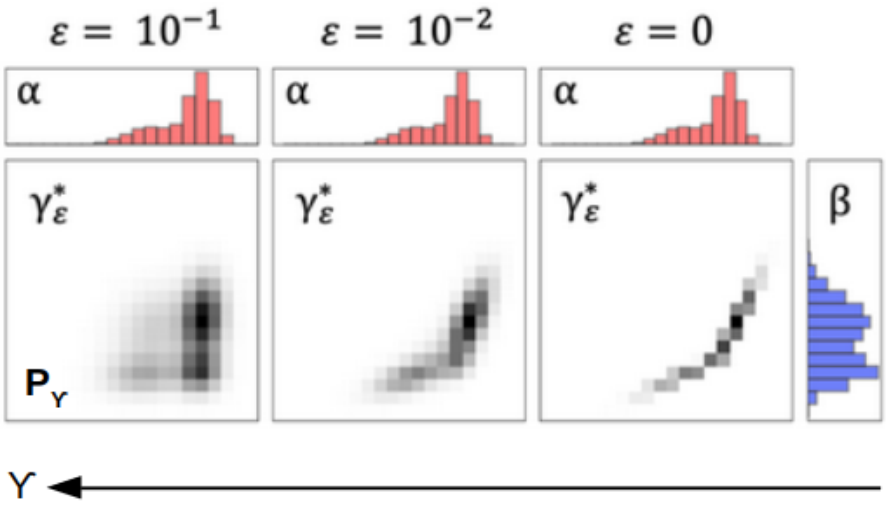


FIGURE 4.7 – Influence du paramètre de régularisation entropique sur la dispersion de la masse des distributions α à β sur le plan de TO régularisé $\gamma * (\lambda)$. Le plan de TO indique la proportion de matière au voisinage de x dans α que l'on va transporter au voisinage de y dans β. À mesure que $\epsilon > 0$ croit la solution $\gamma * (\lambda)$ "s'étale" de plus en plus. – source FLAMARY [novembre 2017] où λ le terme de régularisation est noté ϵ dans le cas de la régularisation par entropie.

4.2 Motivations

Comme alternative au mélange de gaussiennes (GMM) qui cherche à "coller à la complexité" de la distribution des données, nous proposons de transformer les données de sorte qu'elles suivent un modèle simple (en pratique, gaussien), la complexité des données étant alors "cachée" dans la transformation induite par le transport. L'une des motivations de cette approche est de s'affranchir de certaines contraintes liées à la première méthode. En effet, l'absence de connaissance a priori sur les distributions des données limite l'utilisation d'approches paramétriques qui peuvent être problématiques si les données ne suivent pas la distribution supposée.

4.3 Description de la méthode

Cette méthode repose sur le calcul d'un plan de TO (matrice de couplage) entre les données d'une classe et des échantillons générés selon une loi cible gaussienne de mêmes moyenne et variance que celles des échantillons de la classe. Les PDFs des échantillons de chaque classe peuvent être estimées par interpolation de valeurs obtenues via un transport vers la distribution cible gaussienne.

Le transport optimal a la capacité de fournir des correspondances entre des ensembles de points. Cette mise en correspondance est aussi appelée couplage. Il induit une notion de distance entre des distributions de probabilité. Ce transport ou couplage peut être calculé selon différentes fonctions objectifs. Dans la suite des travaux énoncés, nous avons utilisé la fonction objectif liée à la distance de Wasserstein (Emd) et Sinkhorn qui présente une fonction objective de la même famille mais qui intègre un terme de régularisation. Cette distance est aussi appelée Kantorovich-Rubinstein ou "Earth mover's distance" en anglais. Il s'agit du problème classique de Kantorovitch : on cherche à minimiser le coût du transport c'est-à-dire à minimiser la somme des distances entre les couples de points sources et cibles mis en correspondance. Nous parlerons de transport par "EMD" ("Earth

mover's distance") pour désigner la recherche du couplage γ selon ce problème PEYRÉ et collab. [2019]. On considère un ensemble de points source x_i munis de poids a_i et un ensemble de points cibles y_j munis poids b_j. Pour rappel, la recherche du transport optimal revient à résoudre le problème d'optimisation suivant :

$$\hat{\gamma} = \arg\min_{\gamma} < \gamma, M >_F \tag{4.12}$$

$$\sum_j \gamma_{i,j} = a_i \tag{4.13}$$

$$\sum_i \gamma_{i,j} = b_j \tag{4.14}$$

$$\gamma \geq 0 \tag{4.15}$$

où M est la matrice de coût (ou matrice de distance), a et b sont les points sources et cibles et F la métrique. Pour plus de détails sur cette équation, se référer au paragraphe 4.1 ou à l'article BONNEEL et collab. [2011].

Dans le cas d'un transport par "Sinkhorn" CUTURI [2013], il s'agit, comme décrit ci-avant dans le paragraphe 4.1, de transporter la masse d'un point source vers un ou plusieurs points cibles en utilisant un terme de régularisation par entropie (voir les équations (4.9) et (4.10)).

Le transport du point x_i est alors donné par : $T(x_i) = \sum_j \gamma_{i,j} y_j$. A noter que si le nombre de points cibles est supérieur à celui dans l'espace source, plusieurs solutions de couplage pourront être trouvées, même avec le transport par "EMD". Dans ce cas, l'ensemble des réponses est pris en considération pour définir le point transporté correspondant.

Les étapes de la procédure sont les suivantes :

— Apprentissage :

1. Une population cible par classe est définie : nous avons choisi de générer des échantillons dont la distribution suit une loi normale, de mêmes moyenne et variance que les données sources. Le nombre des échantillons de cette population cible peut être variable. L'idée est que l'on laisse "plus de choix" à la méthode de transport, ce qui permet d'obtenir des mises en correspondance significativement meilleures qu'avec un nombre d'échantillons cibles égal au nombre d'échantillons sources. Par suite, les scores de classification s'en trouve améliorés.
2. Un couplage γ est appris qui minimise le transport des points sources vers les points cibles (voir l'équation (4.12)).
3. Les échantillons sources sont transportés dans l'espace cible selon le couplage appris.
4. Des valeurs de PDFs sont affectées aux échantillons sources selon la position des échantillons transportés correspondant par rapport à la gaussienne choisie comme cible à l'étape 1. La valeur de PDF d'un échantillon dépend des deux paramètres de la gaussienne cible, sa moyenne μ et son écart type σ.

5. Définir un espace de validité : pour cela, l'enveloppe convexe de la population est calculée et dilatée. L'ensemble des points présents à l'intérieur du masque sont susceptibles d'appartenir à la classe en question.
6. Choisir une méthode d'interpolation puis l'appliquer en prenant les valeurs des PDFs des échantillons de l'ensemble d'apprentissage (PDF source). La carte des PDFs est uniquement calculée dans l'espace de validité définie.

— Prédiction :

1. En fonction de la position de l'échantillon à prédire, récupérer les n PDFs renvoyées par les n modèles de classes (valeur de PDF sur la carte des PDFs de la classe correspondante).
2. Comparer ces n PDFs. La valeur maximale indique la classe d'appartenance la plus probable et donc la classe prédite.

Pour résumer, il faut choisir une fonction objectif, le nombre de points cibles et une méthode d'interpolation.

4.3.1 Paramètres à optimiser

Seul le paramètre λ pour l'approche par TO avec régularisation par entropie est à fixer. La valeur de ce paramètre est optimisée de façon à obtenir le meilleur score de classification.

4.4 Évaluation et validation de la méthode sur des données synthétiques

L'hypothèse ayant motivée l'élaboration de cette approche est que la distribution des points que nous souhaitons modéliser ne peut pas être représentée correctement par un mélange de gaussiennes. C'est ce qui nous a amené dans un premier temps à vouloir traiter des jeux de données qui obéissent à ce schéma précis. Pour mettre en pratique cette idée, nous avons commencé par travailler sur des données synthétiques.

4.4.1 Description des données

Ainsi, dans le but d'évaluer la méthode, nous avons généré des données 2D de deux classes distinctes dont les lignes de niveau varient de celles d'une distribution gaussienne. En effet, un autre avantage de la génération de données de synthèse est la connaissance de la distribution théorique des données (voir les figures 4.8 et 4.10). Chaque classe présente 600 échantillons (voir la figure 4.9). Concernant l'approche par ajustement d'un mélange de gaussiennes, le nombre de composantes du mélange est estimé à une composante par l'indice BIC. Nous parlerons alors de "Gaussienne" plutôt que de GMM dans la suite du document.

4.4.2 Méthodes d'évaluation

Comme expliqué ci-avant, la classe finale prédite d'un échantillon est issue de la comparaison des PDFs estimées par l'interpolation des valeurs de PDFs des échantillons de l'ensemble d'apprentissage dans l'espace de validité défini. La valeur de PDFs maximale indique la classe d'appartenance la plus probable de l'échantillon en question.

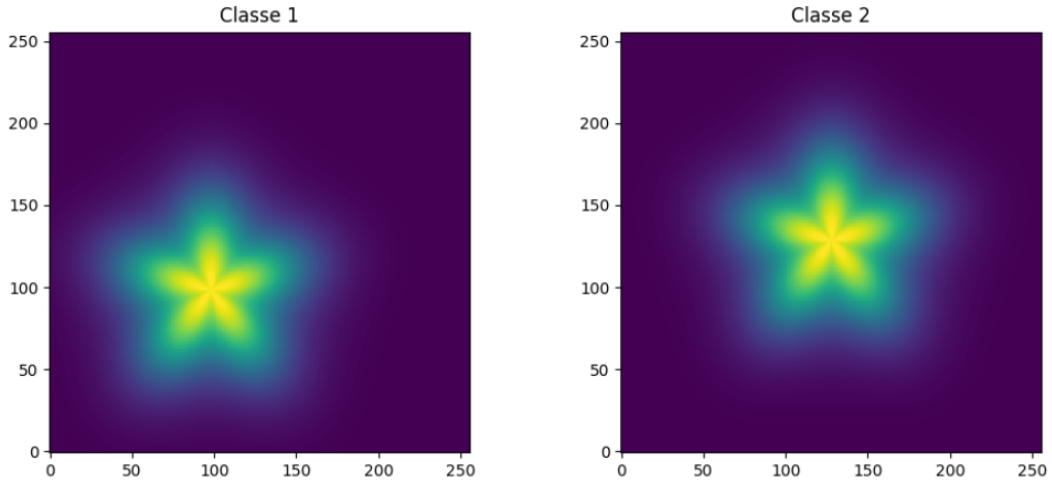


FIGURE 4.8 – Lignes de niveau des PDFs théoriques des distributions des deux classes de données synthétiques.

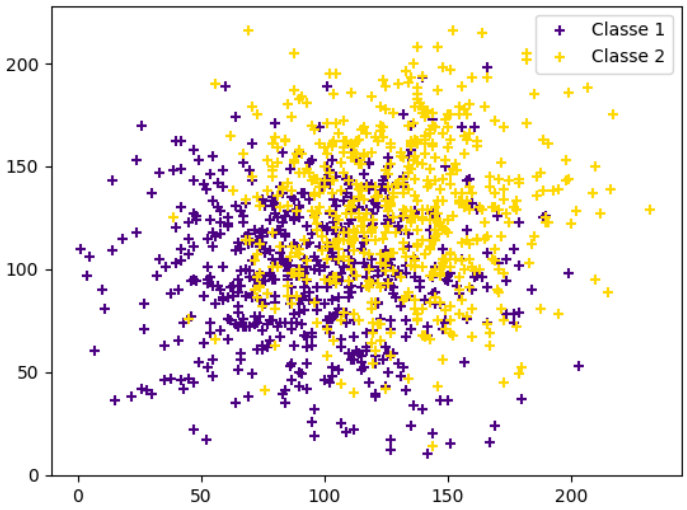


FIGURE 4.9 – Échantillons des deux classes correspondant à notre ensemble de données synthétiques.

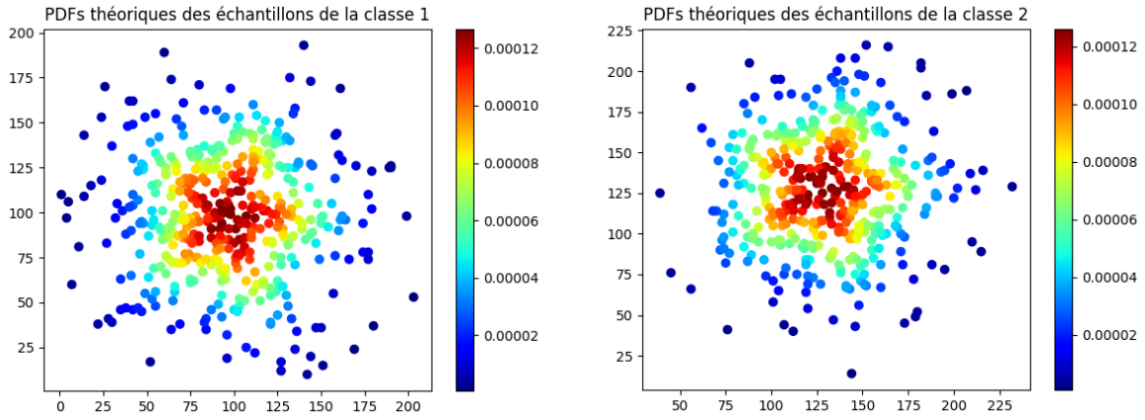


FIGURE 4.10 – PDFs théoriques des échantillons des deux classes 1 et 2.

Dans le but d'évaluer l'efficacité de notre approche, nous avons utilisé deux méthodes de calcul des scores de classification.

— La première méthode correspond à une validation croisée : l'ensemble de données est divisé en quatre. Ainsi, pour chacune des quatre itérations, 3/4 des données sont destinés à l'apprentissage et 1/4 au test. De cette manière chaque échantillon est utilisé exactement une fois en tant qu'échantillon de test. Le score final correspond à la moyenne des quatre scores obtenus à chaque tirage.

— Dans le cas de la seconde méthode, l'ensemble des données est utilisé comme échantillons de l'ensemble d'apprentissage. Les modèles sont testés sur un ensemble de points d'une grille régulière. L'idée est de prédire la classe de ces échantillons sur la base de la valeur de PDF estimée grâce au transport optimal. Ayant les PDFs théoriques de ces échantillons, nous pouvons suivant le même schéma, construire une matrice de confusion "idéale" nous permettant d'avoir la notion de score de classification maximal pouvant être atteint (score inférieur à 100% dans le cas où les populations des deux classes se chevauchent). Il s'agit d'une information que nous ne possédons pas avec la première méthode d'évaluation. Pour rappel, la valeur maximale entre les n PDFs renvoyées par les n modèles de classes, indique la classe d'appartenance la plus probable et donc la classe prédite. Ainsi, une carte de prédiction est construite selon les valeurs théoriques de la PDF des échantillons d'apprentissage et comparée à celle issue des valeurs estimées de la PDF de ces mêmes échantillons.

4.4.3 Résultats

Partie apprentissage

Pour chaque classe, une distribution cible suivant une loi normale de mêmes variance et moyenne que la distribution source est générée. Nous avons défini le nombre d'échantillons cible à huit fois celui des échantillons sources (voir la figure 4.11). Ce nombre a été fixé de façon à obtenir les meilleurs scores de classification. À l'aide des deux matrices de coût (voir la figure 4.12) correspondant aux distances euclidiennes calculées point à point et normalisées par la plus grande distance, pour les deux distributions (source et cible), un couplage est appris. Le résultat de couplage dépend de la méthode utilisée. Dans notre cas, nous avons appliqué les algorithmes de "Sinkhorn" et "EMD" permettant le calcul du transport optimal entre nos deux distributions, avec et sans régularisation. Les matrices résultantes sont illustrées sur les figures 4.13 et 4.14. Enfin, la figure 4.15 présente les échantillons transportés issus des deux couplages appris.

La figure 4.15 illustre le transport des échantillons de la classe 1 via les deux transports appris. On peut remarquer qu'avec "EMD" les points sont plus étalés qu'avec "Sinkhorn". Dans le cas du transport par "EMD", les variances en x et en y des échantillons transportés sont les mêmes que celles de la gaussienne cible générée ([1207.3, 1143.8]). Au contraire, les variances des échantillons transportés par "Sinkhorn" sont moins élevées([764.6, 882.3]).

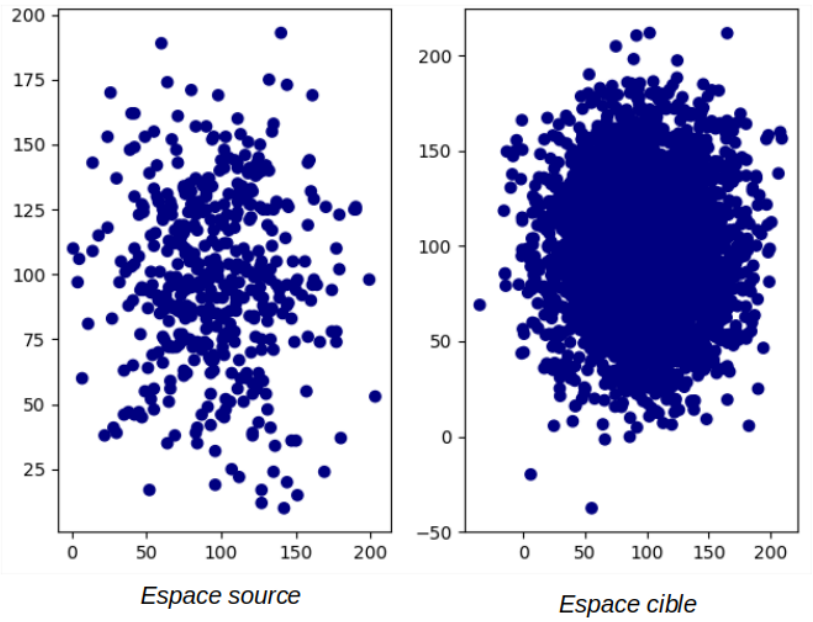


FIGURE 4.11 – A gauche : les échantillons de la classe 1 (décrits dans le paragraphe 4.4.1) dans l'espace source. À droite : les échantillons générés suivant une loi normale, de mêmes moyenne et variance correspondants aux données cibles.

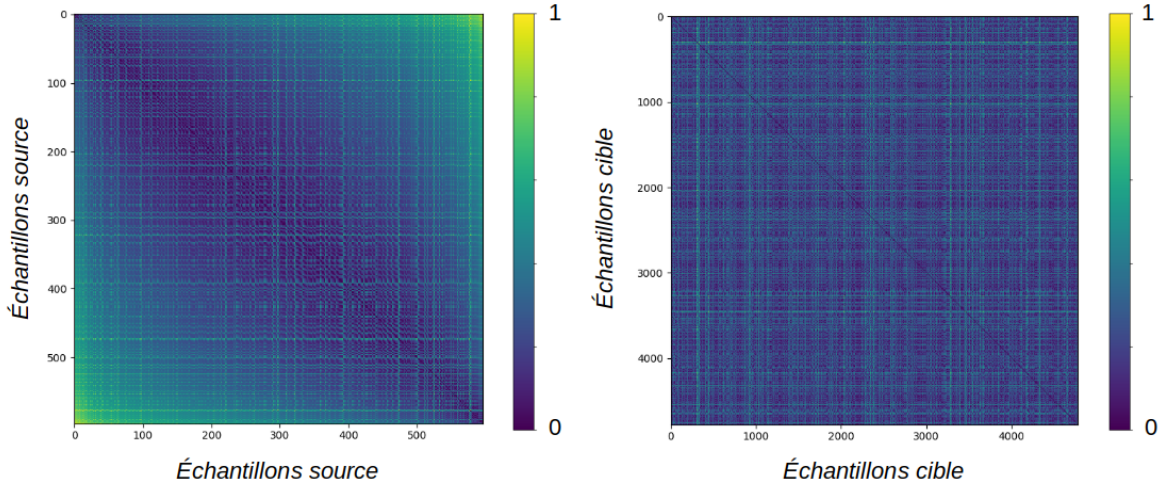


FIGURE 4.12 – A gauche : la matrice de coût des échantillons source de la classe 1. À droite : la matrice de coût des échantillons cible générés.

Une fois les échantillons transportés, des valeurs de PDFs sont affectées aux échantillons sources selon la position des échantillons transportés correspondant. Ainsi, pour chaque échantillon transporté, une valeur est attribuée à l'échantillon source correspondant. La figure 4.16 correspond aux échantillons source et transportés selon "EMD", colorés en fonction de la valeur de PDF calculée. Enfin, comme décrit dans les "étapes de la procédure" du paragraphe 4.3, l'espace de validité et l'interpolation des PDFs sur une grille sont générés pour chaque classe. Un exemple de ces deux images (pour la classe 1) est présenté sur la figure 4.17.

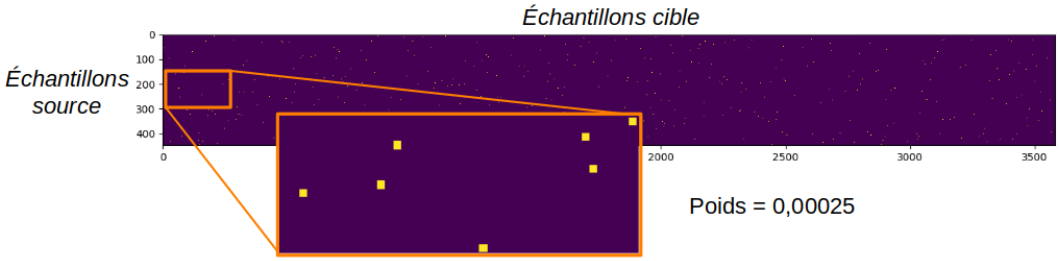


FIGURE 4.13 – Plan de TO des échantillons source et cible de la classe 1 obtenue avec l'approche par transport optimal sans régularisation ("EMD"). Cette matrice indique la proportion de matière de x dans la distribution source que l'on va transporter sur un point y de la distribution cible. La matière est envoyée vers un unique point cible, le poids d'un point est fixé à 0.00025.

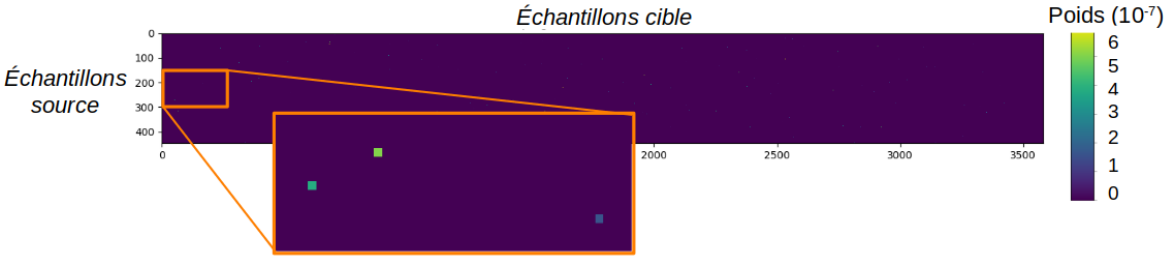


FIGURE 4.14 – Plan de TO des échantillons source et cible de la classe 1 obtenue avec l'approche par transport optimal avec régularisation ("Sinkhorn"). Cette matrice indique la proportion de matière de x dans la distribution source que l'on va transporter sur plusieurs points de la distribution cible. La couleur du point indique la valeur du poids de l'échantillon source transporté vers le point cible.

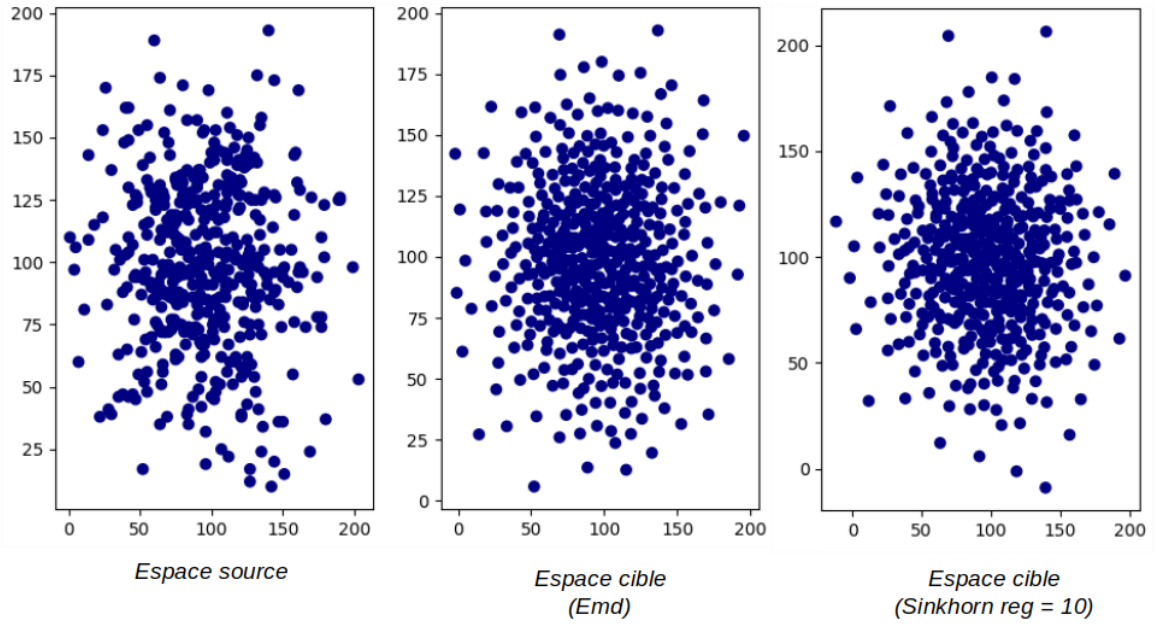


FIGURE 4.15 – A gauche : les échantillons de la classe 1 (décrits dans le paragraphe 4.4.1) dans l'espace source. Au milieu et à droite : les échantillons dits transportés issus du résultat du couplage par "EMD" et "Sinkhorn".

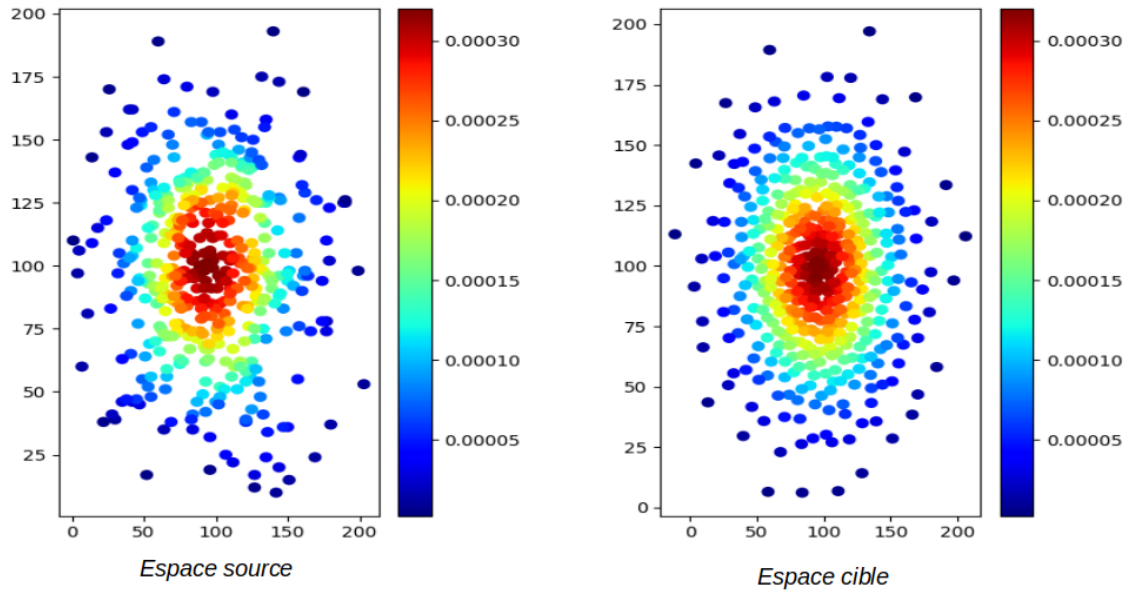


FIGURE 4.16 – A gauche : les PDFs des échantillons de la classe 1 dans l'espace source. À droite : les PDFs des échantillons transportés correspondants.

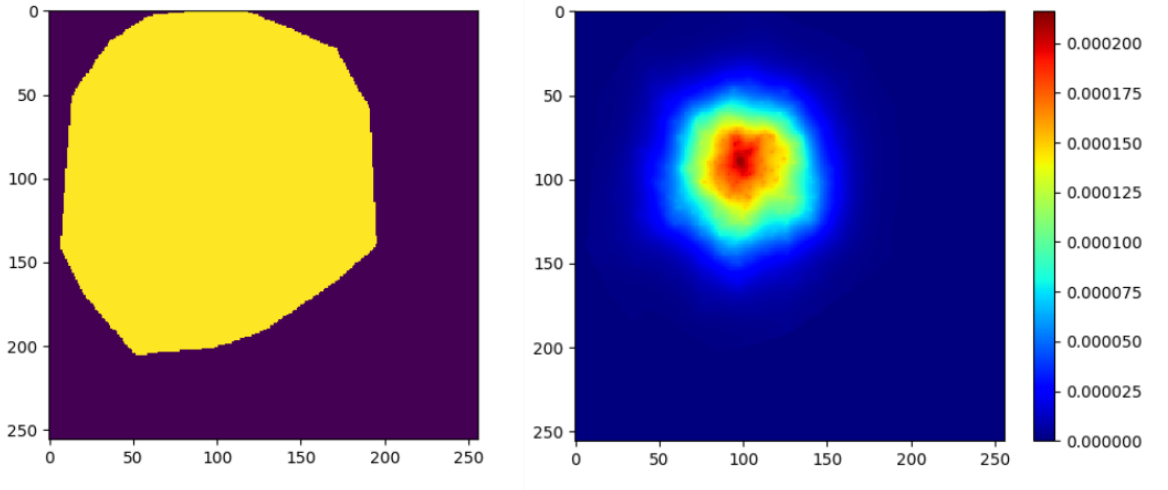


FIGURE 4.17 – A gauche : masque binaire délimitant la zone de validité des populations de la classe 1. À droite : résultat de l'interpolation des valeurs de PDFs des échantillons de l'ensemble d'apprentissage de la classe 1 pour les points présents à l'intérieur de la zone de validité.

Partie prédiction

Une fois la carte des PDFs apprise pour chaque classe, la position $[i, j]$ de l'échantillon de test s sur ces cartes permet la récupération de deux valeurs de PDFs (pour deux classes : $\text{PDF}^1_{i,j}$ et $\text{PDF}^2_{i,j}$). La PDF la plus élevée indique la classe prédite de l'échantillon s.

1. **Validation croisée :**

   Le tableau 4.1 présente les pourcentages de précisions de classification dans le cas d'estimation des PDFs des échantillons des classes par trois méthodes. Ces valeurs de PDFs sont d'une part calculées en ajustant une gaussienne aux données et d'autre part en utilisant deux approches par transport optimal. Ces approches appelées "Sinkhorn" et "EMD" calculent respectivement le transport avec et sans régularisation par entropie.

Méthode de classification	Score moyen
Gaussienne	73.1
TO "EMD"	**74.1**
TO "Sinkhorn" ($\lambda = 1\,e-3$)	73.6

TABLEAU 4.1 – Scores moyens de classification issus de la validation croisée (en pourcentages) des échantillons de l'ensemble de données de synthèses, obtenus avec une Gaussienne et une approche par transport optimal en utilisant deux algorithmes ("Sinkhorn" et "EMD").

2. **Matrice de confusion sur les échantillons de la grille :**

   Étant donné que l'ensemble des points d'une grille régulière est ici considéré comme échantillons de test, une carte de prédiction peut être générée suivant les résultats obtenus. Les figures 4.19 et 4.18 présentent en A la carte des prédictions obtenues en utilisant les valeurs de PDFs théoriques. Puis les images annotées B, correspondent en premier lieu à la carte résultante de l'approche par transport optimal sans régularisation et en second lieu à celle obtenue par ajustement d'une gaussienne aux données d'apprentissage. Enfin, les images annotées C correspondent, dans les deux cas, aux échantillons de la grille dont la prédiction est erronée.

   Le tableau 4.2 rassemble les pourcentages de précisions de classification moyens obtenus d'une part avec les valeurs de PDFs théoriques et d'autre part, avec celles estimées par les trois modèles.

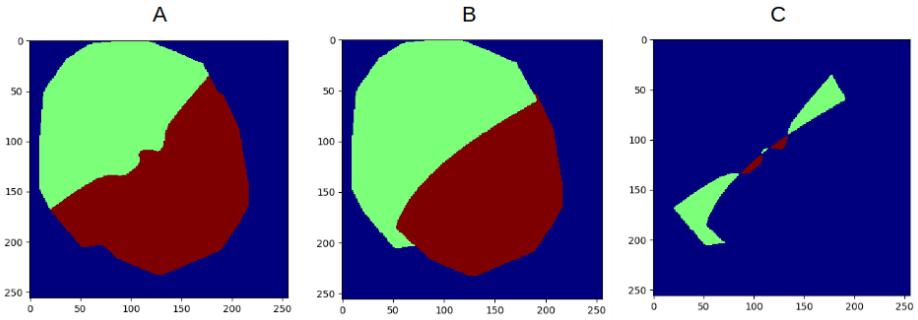


FIGURE 4.18 – Cartes de prédiction obtenues avec : Figure A : les PDFs théoriques et Figure B : les PDFs estimées par l'ajustement d'une gaussienne directement sur les données. La figure C correspond aux points de la grille dont la prédiction est différente de celle de la carte figure A.

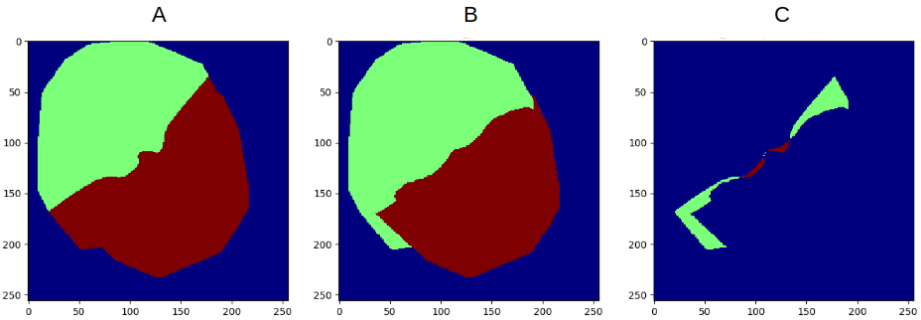


FIGURE 4.19 – Cartes de prédiction obtenues avec : Figure A : les PDFs théoriques et Figure B : les PDFs estimées par la méthode utilisant le transport optimal ("EMD"). La figure C correspond aux points de la grille dont la prédiction est différente de celle de la carte figure A. La carte dans le cas de l'approche "Sinkhorn" n'est pas affichée car très proche de celle-ci.

Méthode de classification	Score moyen
Théorique	72.85
Gaussienne	71
TO "EMD"	**71.5**
TO "Sinkhorn" ($\lambda = 1e-3$)	71.2

TABLEAU 4.2 – Scores de classification (en pourcentages) des échantillons de la grille régulière, obtenus en utilisant les valeurs de PDFs théoriques et celles estimées par une gaussienne ajustée aux données ou par les approches par transport optimal (appris respectivement avec "EMD" et "Sinkhorn").

4.4.4 Discussions et conclusions

Les résultats de la validation croisée, présentés dans le tableau 4.1, permettent de dire que les approches par transport optimal obtiennent de meilleurs résultats de classification qu'en utilisant une gaussienne avec +1% et +0.2% pour respectivement les méthodes "EMD" et "Sinkhorn". De plus, les résultats montrent qu'entre les deux méthodes utilisant le transport optimal, la méthode "EMD" (74.1%) est plus efficace que la méthode "Sinkhorn" (73.6%).

Les scores moyens présentés dans le tableau 4.2 nous permettent de comparer les résultats obtenus avec les méthodes "Gaussienne", "EMD" et "Sinkhorn" entre eux et avec le score maximum "Théorique". Nous pouvons conclure que les méthodes testées sont efficaces de par la faible différence avec le maximum théorique (une différence moyenne d'environ 1.4% et 1.65% pour le TO et 1.9 % pour la gaussienne). Nous pouvons également souligner de meilleurs scores de classification obtenus avec les approches par transport optimal (71.5% et 71.2%) qu'avec la méthode d'ajustement d'une gaussienne (71%).

La méthode "EMD" permet l'obtention de meilleurs résultats sur ces données que ceux obtenus avec la gaussienne. La méthode "Sinkhorn", quoique plus "complexe", est moins efficace que "EMD" et n'apporte qu'une amélioration négligeable par rapport à la gaussienne.

Expériences menées dans le but d'améliorer les résultats

Plusieurs approches (TO) ont été testées afin d'améliorer les résultats mais n'ont à ce jour pas abouti. Certaines sont décrites ci-après :

— Aider à apprendre le transport en étiquetant les échantillons afin de forcer le couplage entre les échantillons source et cible de même label. Pour cela, deux méthodes de labellisation ont été implémentées.

1. La première repose sur l'utilisation de la notion d'Alpha-shape EDELSBRUNNER et collab. [1983]. Elle est tout simplement appelée méthode "Alpha-shape" ci-après et consiste à étiqueter les échantillons en fonction de la "couche" de points dans laquelle il se trouve (voir la figure 4.20 **A**).
2. La deuxième méthode correspond à un étiquetage des échantillons en fonction de la valeur de PDF estimée par un mélange de gaussienne ajustée aux données et dont les paramètres sont estimés via l'indice BIC. Cette méthode est désignée sous l'appellation "GMM pdf". Les étiquettes sont affectées par intervalle de valeurs de PDF du GMM, le nombre d'intervalles étant un paramètre à fixer (voir la figure 4.20 **B**).

— Seuiller la matrice de coût en ne conservant que les distances aux k plus proches voisins des échantillons.

— Changer de métrique. Lorsque la notion de distance entre un échantillon source et un échantillon cible ne peut pas être facilement définie, une solution intéressante est de remplacer l'EMD ou Sinkhorn par la distance de Gromov-Wasserstein (MÉMOLI [2011], TITOUAN et collab. [2019]). Le transport optimal correspond au transport minimisant la somme des différences des distances deux à deux des points sources et cibles mis en correspondance (voir la figure 4.21).

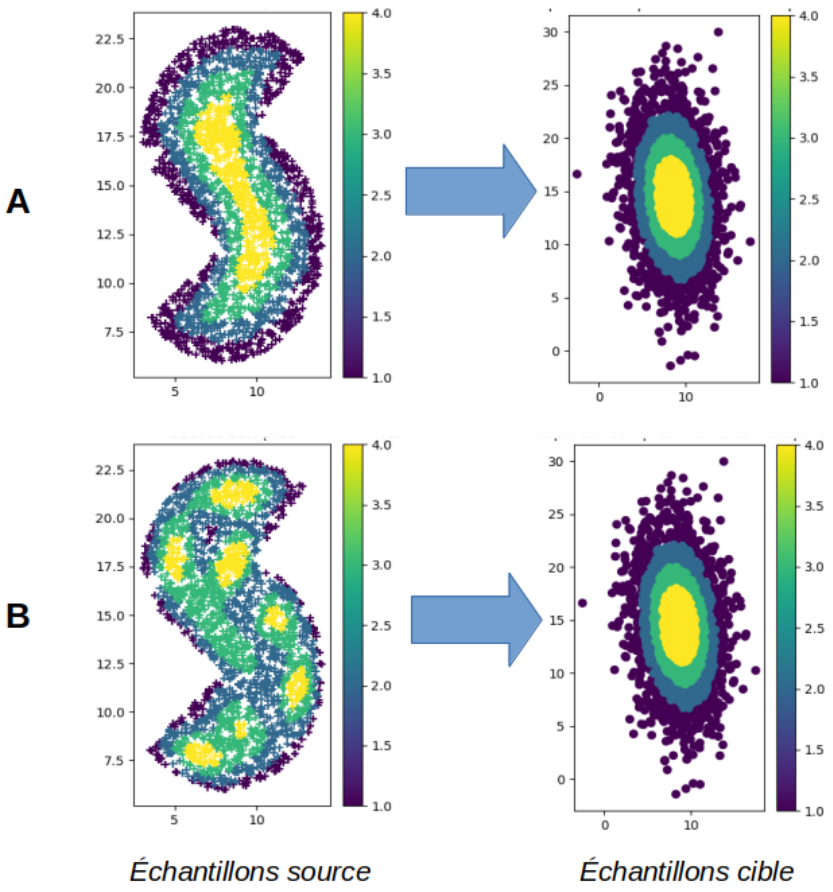


FIGURE 4.20 – Exemple de distributions étiquetées en 4 couches. **A :** Résultat obtenu avec la méthode "Alpha-shape". **B :** Résultat obtenu avec la méthode "GMM pdf". Les couplages sont contraints de se faire entre couches source et cible de même label et le nombre de couches est un paramètre à fixer.

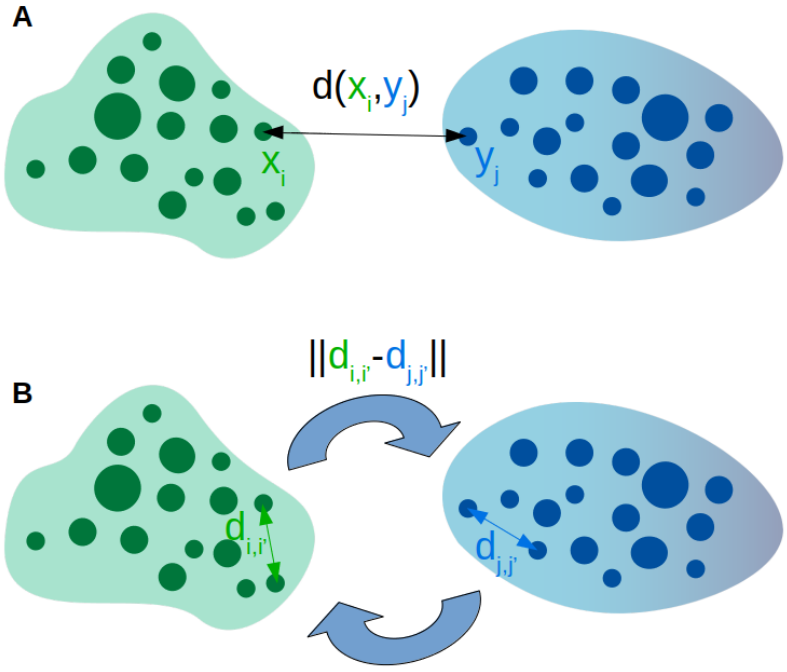


FIGURE 4.21 – **(A) :** Distance de Wasserstein. **(B) :** Distance de Gromov.

4.5 Test de la méthode sur les images des phénotypes connus et inconnus de *Botrytis cinerea*

4.5.1 Description des données

Les données ici utilisées sont celles du chapitre 3 décrites dans le paragraphe 3.4.2. Pour rappel, nous avons choisi deux types de représentation de nos images, toutes deux extraites d'un réseau de neurones. Les caractéristiques Bottleneck sont calculées de façon automatique pendant l'étape d'apprentissage (voir le paragraphe 2.2.2) et le second type de caractéristiques correspond aux données Output, c'est-à-dire à la sortie de la dernière couche du réseau.

Toujours dans le but de s'affranchir des vecteurs caractéristiques aberrants correspondant à des sous-images atypiques (typiquement, sans champignon, voir la figure 3.15), nous prenons la médiane de ces vecteurs comme caractéristiques d'image (16 sous-images par image).

A la différence du travail effectué dans le paragraphe 3.4.2, une analyse en composantes principales (ACP) [1] est appliquée aux données (Output et Bottleneck). Le nombre de dimension est fixé à deux en utilisant l'ACP pour (classiquement) réduire la dimension des données en ne retenant que les composantes de plus forte variation (voir la figure 4.22).

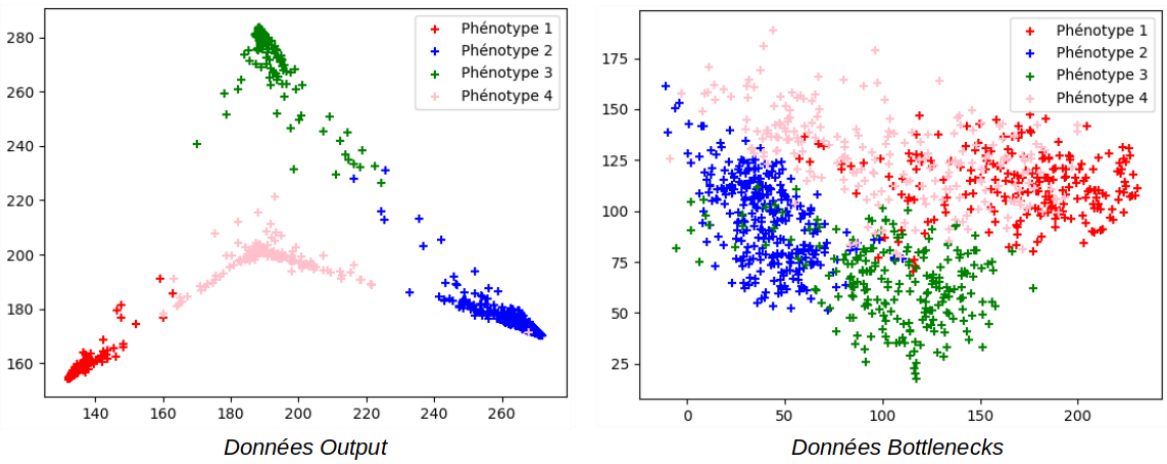


FIGURE 4.22 – Données Output et Bottlenecks dont la dimension est réduite par ACP à 2 dimensions.

Résultats

Le tableau 4.4 présente les scores de classification moyens obtenus en estimant la PDF avec d'une part un mélange de gaussiennes et d'autre part les approches par transport optimal ("EMD" et "Sinkhorn"). Concernant les résultats de l'approche par ajustement d'un mélange de gaussiennes, le nombre de composantes des mélanges par classe (estimé par l'indice BIC), suivant le type de données considéré, est indiqué dans le tableau 4.3 suivant :

1. Approche géométrique et statistique consistant à transformer des variables corrélées en nouvelles variables décorrélées les unes des autres (appelées composantes principales).

Type de données / Classe	1	2	3	4
Bottlenecks	3	1	3	4
Output	7	9	8	8

TABLEAU 4.3 – Nombre de composantes des mélanges de gaussiennes (estimé par l'indice BIC), par classe et par type de données considéré (Bottlenecks et Output).

Type de données	Modèles	Scores
Bottlenecks	GMM	80.6
	TO "EMD"	**82.2**
	TO "Sinkhorn" ($\lambda = 1e-2$)	81.9
Output	GMM	98.5
	TO "EMD"	**98.6**
	TO "Sinkhorn" ($\lambda = 1e-3$)	**98.6**

TABLEAU 4.4 – Scores moyens de classification (en pourcentages) des échantillons Output et Bottlenecks, obtenus avec les GMM puis les approches par transport optimal, "EMD" et "Sinkhorn" permettant le calcul du transport optimal avec et sans régularisation.

Remarque : Pour rappel, nous utilisons l'indice BIC pour estimer les paramètres des composantes des mélanges de gaussiennnes, dont, leur nombre nc. Nous nous sommes néanmoins posé la question suivante : comment évoluent les scores avec la méthode GMM si on multiplie les valeurs nc estimées par 1.5 ou 2 ? Après avoir testé cela, les résultats obtenus indiquent qu'augmenter le nombre de composantes des mélanges n'entraîne pas d'amélioration notable des scores de classification.

Discussions

Les données Output ont globalement besoin de plus de composantes que les données Bottleneck (se référer au tableau 4.3). En effet, sur la figure 4.22, on voit que les données Bottleneck se répartissent de manière plus douce. Elle se prêtent donc mieux à une représentation par un mélange de peu de gaussiennes, voire par une gaussienne unique (comme dans le cas des données synthétiques).

Les données Output sont beaucoup plus faciles à classer. On le sait d'avance par leur définition, et on le comprend aussi visuellement sur la figure 4.22. Par contre, les distributions sont plus difficiles à modéliser. Malgré cela, avec un nombre suffisamment important de composantes, le GMM s'en sort très bien. Concernant les données Bottleneck, la classification est plus difficile (du fait de gros chevauchements des classes), mais les classes sont clairement plus faciles à modéliser et à transporter vers des gaussiennes.

Les scores moyens de classifications en considérant les données Bottlenecks comme représentation de nos images sont de 80.6% avec les GMM contre 82.2% pour le TO "EMD" et 81.9% pour le TO "Sinkhorn". Ceux sur les données Output sont de 98.5% avec la GMM contre 98.6% pour le TO "EMD" et "Sinkhorn". Les résultats de classification sur ces données (Bottlenecks et Output) sont donc quasiment égaux avec un score très légèrement supérieur (+0.1%) pour les approches par transport optimal.

Ces résultats peuvent être comparés à ceux obtenus avec l'approche par transport optimal décrite dans le paragraphe A.6 de l'annexe et reposant sur un système d'apprentissage simultané d'un couplage et d'un mappage en utilisant l'algorithme de "Sinkhorn". La méthode implique l'apprentissage d'un modèle indépendamment pour chaque classe et l'apprentissage d'un seuil fondé sur les interactions de classes. Nous avions décidé de réduire le nombre de dimension des données Bottlenecks à celui des données Output, autrement dit à sept via une ACP. L'idée générale était d'apprendre pour chaque classe, le transport optimal entre les échantillons (de l'ensemble d'apprentissage) de l'espace source et ceux de l'espace cible puis de transporter les échantillons à prédire via les différentes transformations associées aux transports et de récupérer les PDFs pour chaque modèle. Les scores moyens de classification en considérant les données Bottlenecks comme représentation de nos images étaient de 72% pour le TO et ceux sur les données Output étaient de 92%. Il semblerait donc que la méthode actuelle fonctionne mieux sur les deux types de données avec une augmentation de la précision de classification de 10.2% pour les données Bottlenecks et de 6.6% pour les données Output.

Conclusion

Les résultats obtenus sur les données synthétiques d'une part et réelles d'autre part, nous permettent de conclure que les approches par transport optimal présentent une efficacité de classification équivalente ou légèrement supérieure à celle de la méthode GMM. De plus, en comparant la méthode "EMD" et "Sinkhorn", on se rend compte que le calcul du transport optimal avec régularisation ne permet pas toujours d'améliorer les résultats obtenus par la méthode sans régularisation. Dans le cas des données synthétiques et des données Bottlenecks, les méthodes par transport optimal obtiennent des résultats très proches (82.2% pour "EMD" contre 81.9% pour Sinkhorn") et sur les données Output, le calcul du transport avec régularisation ne montre aucune amélioration.

4.5.2 Perspectives

Au lieu de faire une interpolation, nous avons pensé à utiliser une approche non-paramétrique afin d'estimer les PDFs dans l'espace des données, l'estimateur de densité par noyau, en prenant comme dans l'article PEHERSTORFER et collab. [2014], un noyau par échantillon d'une grille et en optimisant leurs poids de façon à minimiser l'erreur d'estimation de PDF sur les points d'apprentissage.

4.6 Classification avec rejet

La méthode proposée de l'estimation d'un mélange de gaussiennes et d'apprentissage d'un seuil sur les PDFs est décrite dans le chapitre 3. Concernant le TO en 2D présenté ici, le rejet d'un échantillon est décidé de par sa position : s'il se trouve en dehors de l'espace de validité (masque binaire), l'échantillon est rejeté. La figure 4.23 montre clairement que peu importe la méthode (GMM ou TO) avec classe de rejet, les données Bottleneck dont la dimension est réduite à 2 dimensions ne correspondent pas à des caractéristiques assez discriminantes pour distinguer les échantillons à rejeter. Dans le cas des données Output, on observe surtout une forte confusion entre les échantillons de la classe phénotype 4 et les nouveaux phénotypes. Il apparaît alors nécessaire de conserver certaines informations, par exemple les 7 dimensions des vecteurs Output afin de correctement identifier les quatre phénotypes connus et inconnus. Il faut rester en dimension plus grande,

comme en témoignent les résultats de la méthode GMM avec rejet sur les données Output en 7D et les données Bottlenecks en grandes dimensions (90D).

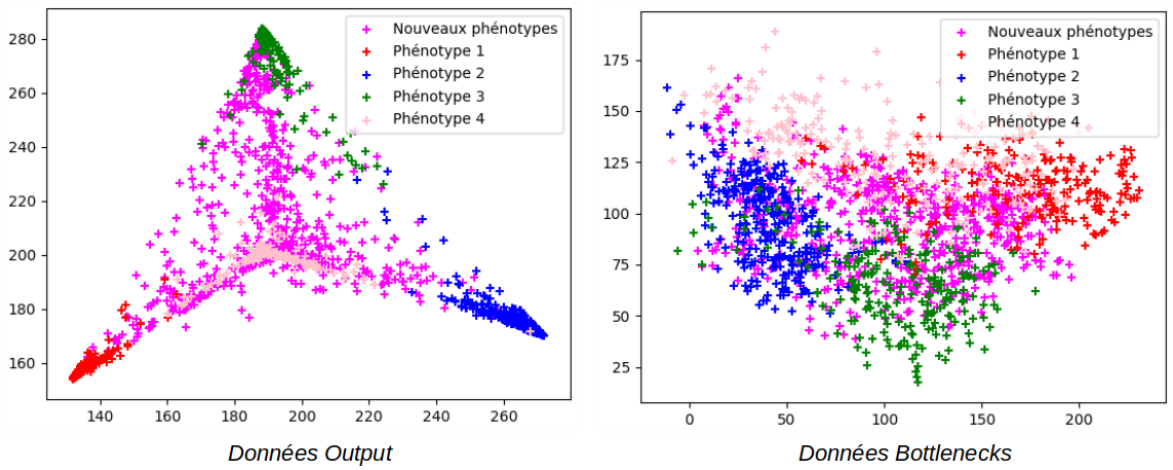


FIGURE 4.23 – Données Output et Bottlenecks dont la dimension est réduite par ACP à 2 dimensions. Les échantillons appartiennent aux classes connues (1-4) ou correspondent à des nouveaux phénotypes.

4.7 Références

2019, «Computational optimal transport», *Foundations and Trends in Machine Learning*, vol. 11, nº 5-6, p. 355–607. xvi, 117

BONNEEL, N., M. VAN DE PANNE, S. PARIS et W. HEIDRICH. 2011, «Displacement interpolation using lagrangian mass transport», dans *Proceedings of the 2011 SIGGRAPH Asia Conference*, p. 1–12. 122

COURTY, N., R. FLAMARY, D. TUIA et A. RAKOTOMAMONJY. 2016, «Optimal transport for domain adaptation», *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, nº 9, p. 1853–1865. xvi, 118, 119

CUTURI, M. 2013, «Sinkhorn distances : Lightspeed computation of optimal transport», dans *Advances in neural information processing systems*, p. 2292–2300. 119, 120, 122

DANTZIG, G. B. 1951, «Application of the simplex method to a transportation problem. activity analysis of production and allocation», *Koopmans, T.C., Ed., John Wiley and Sons*, p. 359–373. 116

DESSEIN, A., N. PAPADAKIS et J.-L. ROUAS. 2018, «Regularized optimal transport and the rot mover's distance», *The Journal of Machine Learning Research*, vol. 19, nº 1, p. 590–642. 119

EDELSBRUNNER, H., D. KIRKPATRICK et R. SEIDEL. 1983, «On the shape of a set of points in the plane», *IEEE Transactions on information theory*, vol. 29, nº 4, p. 551–559. 131

FLAMARY, R. novembre 2017, *Optimal transport for machine learning*. URL https://remi.flamary.com/pres/OTML_ISIS_2017.pdf. xvi, 121

FROGNER, C., C. ZHANG, H. MOBAHI, M. ARAYA et T. A. POGGIO. 2015, «Learning with a wasserstein loss», *Advances in neural information processing systems*, vol. 28, p. 2053–2061. 118

GORDALIZA, P., E. DEL BARRIO, G. FABRICE et J.-M. LOUBES. 2019, «Obtaining fairness using optimal transport theory», dans *International Conference on Machine Learning*, PMLR, p. 2357–2365. 118

HAKER, S. et A. TANNENBAUM. 2001, «Optimal mass transport and image registration», dans *Proceedings IEEE Workshop on Variational and Level Set Methods in Computer Vision*, IEEE, p. 29–36. 118

KOLOURI, S., S. R. PARK, M. THORPE, D. SLEPCEV et G. K. ROHDE. 2017, «Optimal mass transport : Signal processing and machine-learning applications», *IEEE signal processing magazine*, vol. 34, nº 4, p. 43–59. 118

KUSNER, M., Y. SUN, N. KOLKIN et K. WEINBERGER. 2015, «From word embeddings to document distances», dans *International conference on machine learning*, p. 957–966. 118

MÉMOLI, F. 2011, «Gromov–wasserstein distances and the metric approach to object matching», *Foundations of computational mathematics*, vol. 11, nº 4, p. 417–487. 131

NATHAEL GOZLAN, P.-A. Z., PAUL-MARIE SAMSON. 16 mars 2018, *Notes de cours sur le Transport Optimal*. URL https://perso.math.u-pem.fr/samson.paul-marie/pdf/coursM2transport.pdf. 116

PEHERSTORFER, B., D. PFLÜGE et H.-J. BUNGARTZ. 2014, «Density estimation with adaptive sparse grids for large data sets», dans *Proceedings of the 2014 SIAM international conference on data mining*, SIAM, p. 443–451. 135

PEYRÉ, G., M. CUTURI et collab.. 2019, «Computational optimal transport : With applications to data science», *Foundations and Trends® in Machine Learning*, vol. 11, nº 5-6, p. 355–607. 122

PEYRÉ, G. 19 janvier 2017, *Numerical Optimal Transport and Applications*. URL https://www.irit.fr/cimi-machine-learning/sites/irit.fr.CMS-DRUPAL7.cimi_ml/files/pictures/PeyreWS1.pdf. xvi, 117

SCHIEBINGER, G., J. SHU, M. TABAKA, B. CLEARY, V. SUBRAMANIAN, A. SOLOMON, J. GOULD, S. LIU, S. LIN, P. BERUBE et collab.. 2019, «Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming», *Cell*, vol. 176, nº 4, p. 928–943. 118

TITOUAN, V., N. COURTY, R. TAVENARD et R. FLAMARY. 2019, «Optimal transport for structured data with application on graphs», dans *International Conference on Machine Learning*, PMLR, p. 6275–6284. 131

Chapitre 5

Modélisation de la croissance du champignon *Botrytis cinerea* au cours du temps

La croissance du champignon *Botrytis cinerea* rappelle celle d'une plante de part la possibilité de la modéliser (topologiquement) par un système ramifié qui évolue au cours du temps. La spore pour notre champignon est ce qu'est la graine pour la plante. Dans notre cas de figure, le processus de ramification fait partie des stratégies d'exploitation des ressources. En effet, il permet aux plantes ainsi qu'aux champignons, d'optimiser la surface d'échange avec le milieu extérieur. Tous deux peuvent être vus comme des objets dynamiques qui changent de taille et de forme (nombre de branches) au fil du temps. En fonction de certains mécanismes de croissance, l'objet grossit, les branches peuvent s'allonger et de nouvelles peuvent être créées. Un phénotype de champignon comme de plante, présente des processus de croissance et de ramification qui dépendent entre autres, de paramètres extérieurs. Dans le cas des plantes on retrouve par exemple : la température, la luminosité et le niveau d'humidité. Ces paramètres peuvent être utilisés comme paramètres dans les modèles dynamiques afin de reproduire au mieux le phénotype de la condition choisie. Nous pouvons dans notre cas apparenter ces conditions au type de traitement testé et à la concentration en molécule.

Dans ce chapitre, dédié à la caractérisation des phénotypes de *Botrytis cinerea* au cours du temps, nous proposons un modèle discret de croissance inspiré de ceux introduits dans le domaine de la botanique. La morphologie sera ici considérée au sens large du terme, incluant des notions de topologie.

5.1 État de l'art

Parmi les approches classiques en modélisation de croissance des plantes, celles qui nous intéressent sont celles amenant à la construction de modèles morphologiques De Reffye et Blaise [1993]. Ces modèles sont fondés sur les connaissances a priori en architecture végétale et sont mis en œuvre *via* des algorithmes de construction d'arborescence. Un végétal contient des informations qualitatives et quantitatives aux niveaux élémentaires (nœuds, feuilles, ...) et global (stratégie de croissance, ...) qui sont utilisées comme caractéristiques dans le modèle morphologique afin de coller au mieux à la réalité De Reffye et Blaise [1993].

Parmi les techniques connues pour simuler le développement des plantes en temps discret, nous pouvons citer le L-système LINDENMAYER [1968] et ses variantes AONO et KUNII [1984], ainsi que le logiciel AMAP DE REFFYE et collab. [1988] LECOUSTRE et DE REFFYE [1993].

En 1968, au travers de ses travaux de recherche LINDENMAYER [1968], Lindenmayer introduit le L-système ou système de Lindenmayer, un langage formel permettant, par le biais d'un ensemble de règles intuitives, une description de la structure statique de la plante mais également de sa dynamique. En effet, chaque plante est représentée par une répétition de modèles simples. Ainsi, en élucidant la façon dont ces modèles sont répétés, nous obtenons les règles régissant la croissance de la plante en question PRUSINKIEWICZ et LINDENMAYER [1990]. Un L-système présente une chaîne de départ et une ou plusieurs règles de réécriture décrivant comment faire évoluer (par réécriture) la chaîne de départ pour en obtenir une nouvelle. Interprétées comme des instructions de dessin, ces chaînes peuvent permettre la génération d'images. Un exemple simple de cette approche permet la génération de la courbe de Koch Lsy : En prenant "F" comme chaine initiale et "F-F++F-F" comme règle de réécriture ainsi que les instructions de dessin suivantes :

— F : dessiner un trait dans la direction courante,

— -/+ : tourner à gauche/droite de 60°

on obtient alors les instructions suivantes : Dessiner un trait, tourner à gauche de 60°, dessiner un trait, tourner deux fois à droite de 60°, dessiner un trait puis tourner à gauche de 60°et enfin, dessiner un trait. Les dessins qui en découlent sont présentés sur la figure 5.1.

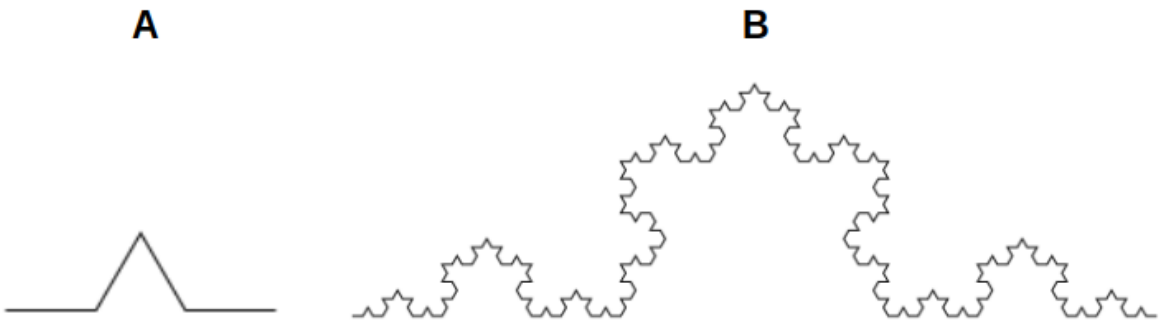


FIGURE 5.1 – Courbe de Koch générée via l'approche du L-système, après un nombre de 1 (**A**) et 4 (**B**) itérations.

Prusinkiewiez a étendu l'approche en dimension 3 et a développé un système informatique permettant d'en visualiser les résultats PRUSINKIEWICZ [1986], PRUSINKIEWICZ et LINDENMAYER [1990]. De nombreux modèles morphologiques sont fondés sur le L-système PRUSINKIEWICZ [1998] dont par exemple ceux issus du projet Virtual Plants PRUSINKIEWICZ [2002] et du logiciel GroIMP HEMMERLING et collab. [2008]. Malgré un cadre général correspondant bien au contexte de notre étude, la notion de réécriture n'est pas le type de croissance dont nous nous sommes inspirés.

Dans le cadre de nos travaux, nous avons décidé de mettre l'accent sur une analyse morphologique d'un point de vue topologique. En effet, parmi les domaines mathématiques permettant l'extraction de mesures quantitatives de la morphologie des plantes, nous nous sommes intéressés à l'approche qui nous paraissait la plus intuitive. Cette approche permet une représentation de la morphologie ramifiée des plantes via un gra-

phique mathématique BUCKSCH et collab. [2017] fondé sur l'extraction de descripteurs squelettiques PRUSINKIEWICZ et LINDENMAYER [2012]CONN et collab. [2017] (voir la figure 5.2 **A**). Ces graphiques sont issus de mesures manuelles ou de données d'imagerie. Les mesures quantitatives (longueurs, diamètres, angles) peuvent être mesurées à un moment donné mais également au fil du temps afin de capturer la dynamique de croissance.

Les graphes arborescents multi-échelles caractérisent la topologie d'une structure ramifiée à différentes échelles et niveaux de détails (voir la figure 5.2 **B**). Ces graphes connexes sans cycle, constituent la base d'un langage de codage implémenté dans le logiciel AMAPmod, un programme interactif d'analyse de la structure topologique des plantes GODIN et collab. [1997]. Dans GODIN et collab. [1999] les auteurs décrivent la création et l'analyse statistique d'une base de données de mesures relatives à la morphologie de pommiers hybrides à partir de données réelles. La procédure est illustrée sur la figure 5.3. La représentation multi-échelle de l'architecture végétale ainsi que ces avantages sont par ailleurs discutés par Remphrey et Prusinkiewicz (1997) dans MICHALEWICZ [1997].

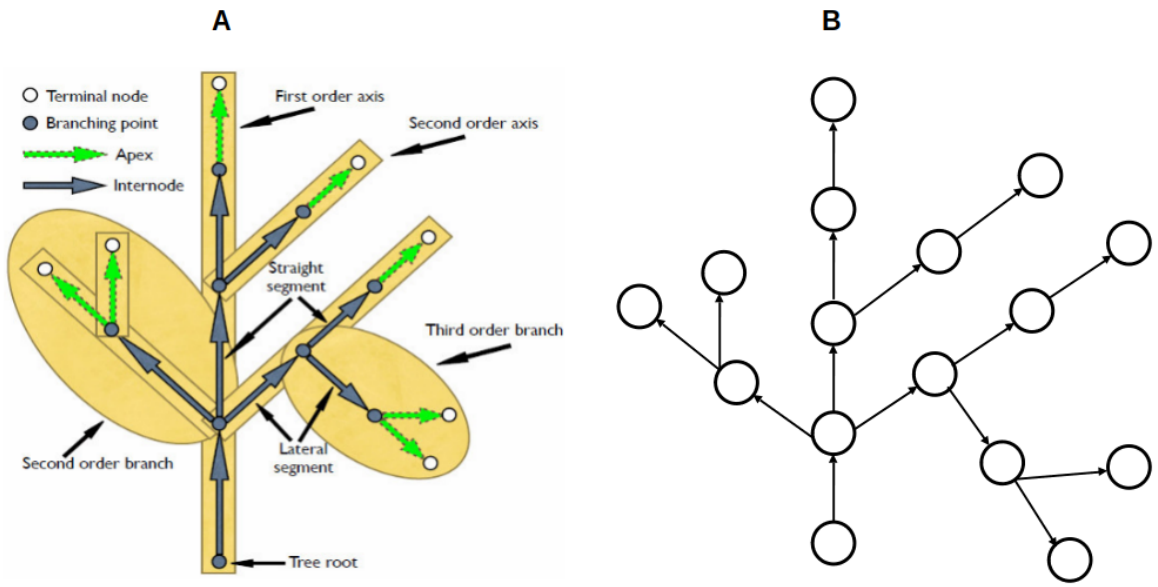


FIGURE 5.2 – A : Illustration de l'architecture végétale sous la forme d'un graphe arborescent multi-échelles (MTG). La figure est extraite de l'article GODIN et CARAGLIO [1998]. B : Graphe arborescent correspondant.

Les simulations informatiques utilisent des principes de la théorie des graphes. Parmi ces principes, on retrouve la réécriture de graphes qui ajoute successivement à un graphe des nœuds et des arêtes dans le but de modéliser la morphologie des plantes au cours du temps BUCKSCH et collab. [2017]. Les règles régissant la construction de ces graphes reflètent les différences entre les phénotypes de végétaux existants.

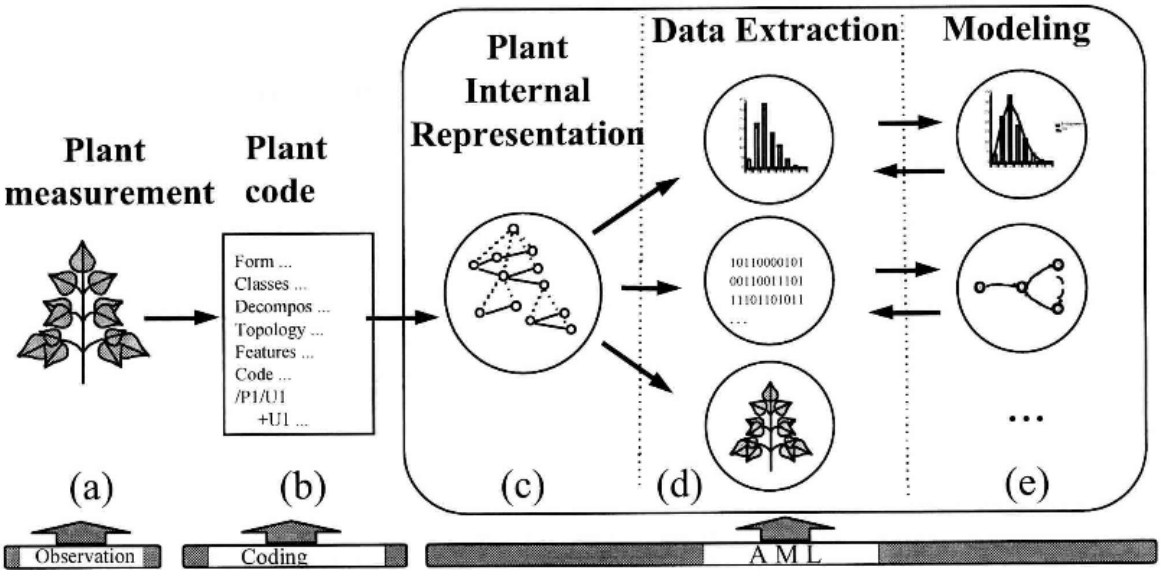


FIGURE 5.3 – Illustration du procédé du logiciel AMAPmod. La figure est extraite de l'article GODIN et collab. [1999] : (a) Mesures réelles issues d'observations sur le terrain, (b) Code informatique décrivant la topologie de la plante à différentes échelles (observations faites en (a)), (c) construction d'une représentation interne de l'architecture de la plante (graphes arborescents multi-échelles) par le logiciel à partir du code. (d) extraction d'informations à partir de la représentation interne, (e) analyses statistiques de ces données *via* des modèles probabilistes ou stochastiques. **AML :** langage de modélisation AMAP

5.2 Méthode proposée

Dans ce chapitre, nous proposons un modèle discret pour modéliser la croissance du champignon *Botrytis cinerea* au cours du temps selon différents phénotypes. Les travaux menés dans le cadre de cette étude suivent les étapes illustrées sur la figure 5.4. Nous choisissons dans un premier temps les phénotypes d'intérêts puis nous sélectionnons les molécules antifongiques dont le mode d'action mène à l'apparition de ces phénotypes (voir la figure 5.4. **A**). Une fois le protocole appliqué et les plaques préparées, des acquisitions en microscopie sont effectuées pour générer des séquences temporelles (voir la figure 5.4 **B**). Puis des méthodes de traitement d'images sont appliquées aux images acquises afin d'en améliorer le contenu et de détecter les objets qui s'y trouvent (voir la figure 5.4 **C**). Pour chaque objet, des paramètres morphométriques et topologiques sont extraits à chaque temps d'acquisition. Ces paramètres sont qualifiés de paramètres phénotypiques dans le reste du manuscrit. Des analyses statistiques sont effectuées dans le but de comprendre et caractériser l'évolution de ces paramètres phénotypiques au cours du temps pour chaque condition (voir la figure 5.4 **D**). Les paramètres du modèle sont extraits des analyses de ces mesures. Suivant la condition (nom de la molécule antifongique choisie, phénotype, et concentration en molécules plus ou moins forte), les valeurs de ces paramètres sont différentes (voir la figure 5.4 **E**). Une fois les paramètres calculés, le modèle permet, suivant les conditions indiquées, de reproduire la croissance du phénotype de champignon correspondant au cours du temps (voir la figure 5.4 **F**).

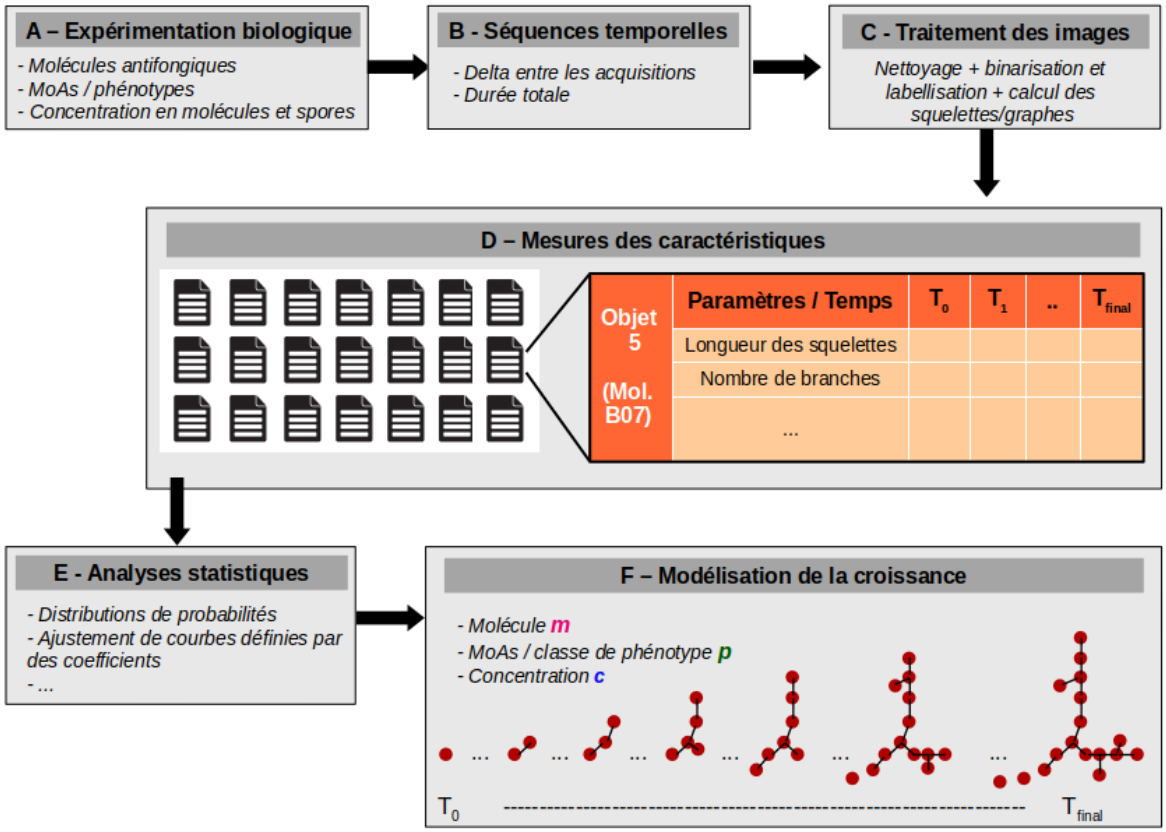


FIGURE 5.4 – **A :** Expérimentation biologique, **B :** Acquisition des images par microscopie, **C :** Traitement des images, **D :** Mesures des caractéristiques, **E :** Analyse des données, **F :** Mise en place des modèles et simulation de la croissance d'un phénotype remarquable suivant une molécule et une concentration.

L'ensemble de cette procédure ainsi que les résultats, discussions et conclusions sont présentés dans ce chapitre sous la forme de six sections :

1. Présentation des TimeLapses
2. Segmentation et squelettisation avec cohérence temporelle
3. Extraction des paramètres phénotypiques
4. Étude de l'influence de la concentration en molécule sur le phénotype 1
5. Analyses complémentaires
6. Développement d'un modèle de croissance

5.3 Présentation des TimeLapses

5.3.1 Nature des données

Nous avons acquis des séquences temporelles (TimeLapses) générées avec une concentration de spores et un intervalle de temps donnés. Dans le cas de la plaque étudiée ci-après, une image est réalisée toutes les 15 minutes sur un intervalle de 0 à 48 heures sur une sélection de trois phénotypes. Le protocole d'expérimentation biologique est le même que décrit dans le paragraphe 2.4.1 (voir la figure 2.14), à l'exception de la concentration en cellules de champignon. En effet, à la suite des travaux menés sur la segmentation de nos images (détaillés dans la partie 2.5.1 du chapitre 2), nous nous étions rendu

compte de la difficulté de segmentation de nos champignons, une difficulté en grande partie due aux chevauchements de certains d'entre eux. Afin de contourner le problème, faciliter la tâche de binarisation et réduire les biais dues aux chevauchements des cellules, nous avons diminué leur concentration. Pour déterminer ce paramètre, nous avons testé cinq concentrations en spores (voir le tableau 5.1) et observé leur densité sur l'image (voir la figure 5.5). La concentration C4 est retenue du fait d'un nombre minimum d'objets se chevauchant mais avec un nombre d'objets suffisant pour effectuer une étude statistique.

	C1	C2	C3	C4	C5
Concentrations	10^4 sp/mL	$0{,}5.10^4$ sp/mL	10^3 sp/mL	$0{,}5.10^3$ sp/mL	10^2 sp/mL

TABLEAU 5.1 – Concentrations en spores testées dans le cadre de notre étude.

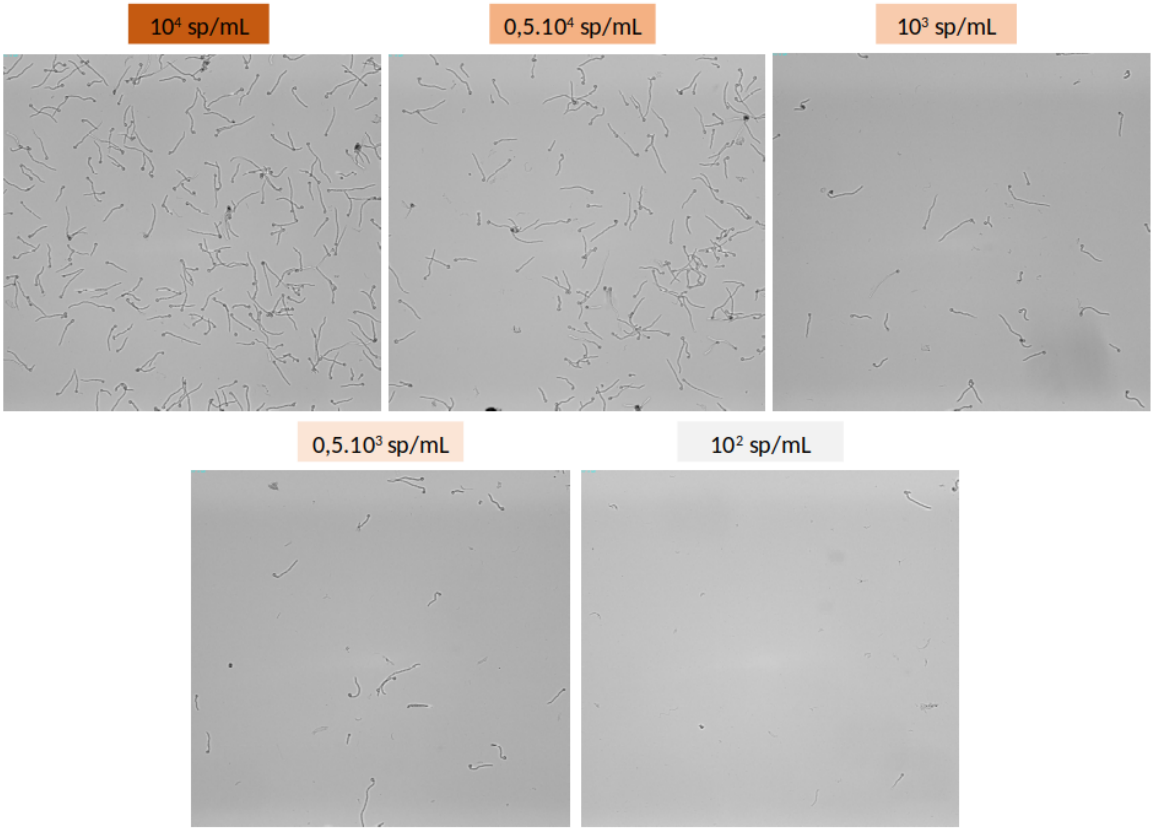


FIGURE 5.5 – Images à 12 heures après inoculation des spores.

5.3.2 Analyse des images du TimeLapse

En regardant les images des séquences temporelles, nous nous sommes rendu compte qu'avant un certain laps de temps, les cellules de champignons ne sont pas encore toutes agrippées au fond du puits (voir la figure 5.6). En effet, le mouvement occasionné par le robot, lors de la prise de la plaque dans la colonne et son basculement dans le microscope, entraîne le déplacement des spores dans le puits. Ainsi, un recalage des objets serait nécessaire afin de pouvoir suivre leur évolution au cours du temps. Toutefois ces objets sont encore à l'état de spore, c'est-à-dire, des cellules rondes de champignons qui ont

à peine commencé à se développer. Nous avons donc pris la décision d'ignorer, dans un premier temps, ces images dans notre analyse. En revanche, l'identification automatique du temps au-delà duquel les cellules ne bougent plus est primordiale pour la mise en correspondance des champignons d'une image à la suivante. Nous avons donc développé un algorithme calculant le rapport du total des aires des objets segmentés (la méthode de segmentation est décrite ci-après dans le paragraphe 2.5.1) du temps t et au temps $t+1$ ainsi que le nombre d'objets détectés à ces deux temps. Un seuil de 95% a été fixé empiriquement. Le temps T_0 est défini comme le premier temps obtenant un rapport supérieur à ce seuil et dont la différence du nombre d'objets du temps t au temps $t+1$ est de un au maximum. Pour l'ensemble des séquences temporelles acquises, nous avons obtenu des temps T_0 autour de la $30^{ème}$ prise (timepoint 30), autrement dit après 7.5 heures d'incubation.

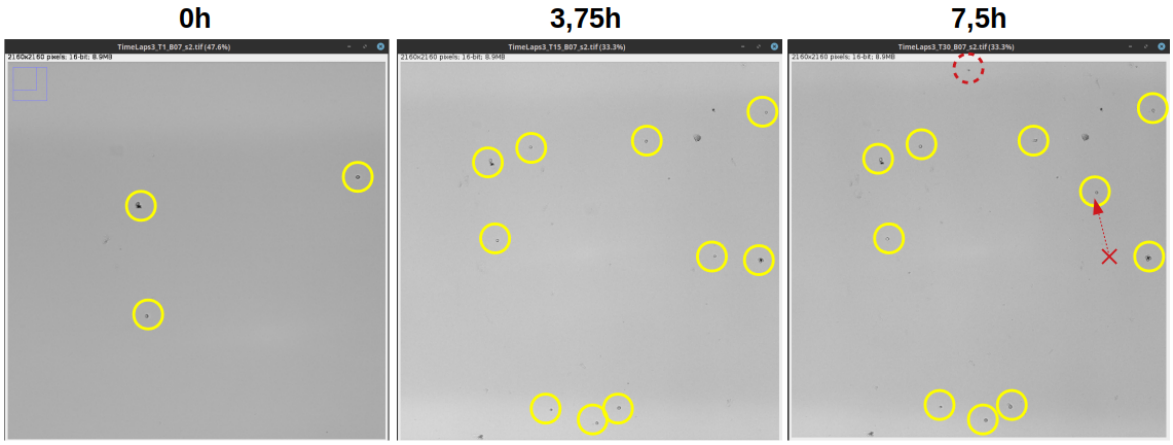


FIGURE 5.6 – Image 2 du champ pris dans le puits 7 aux temps 0, 3.75 et 7.5 heures, illustrant le changement de place des spores dans le milieu. Entourés d'un cercle jaune : les spores. À 7.5 heures, la flèche rouge montre le déplacement de la spore entre les temps 3.75 et 7.5 heures. Le cercle en pointillés rouges entoure une spore absente du champ de l'image à 3.75 heures.

D'autre part, nous avons également observé que certaines cellules de champignons éclataient (voir la figure 5.7), dans le cas du phénotype 1, après la $130^{ème}$ prise (timepoint 130). Ce phénomène entraîne la propagation de débris intracellulaires dans le milieu, visibles dans le fond de l'image. Ces particules sont responsables d'une dégradation de la segmentation des objets de l'image. Néanmoins, des cellules qui éclatent ne présentent plus vraiment d'intérêt. De plus, concernant d'autres phénotypes, après ce temps là, les branches ne s'allongent plus mais se superposent (voir la figure 5.8), ce qui a posteriori biaise les mesures effectuées.
Nous avons donc décidé d'étudier uniquement les images entre les temps 30 et 129 (T_{final}), soit un total de 100 images par séquence. Ainsi, lorsque nous parlerons de T_0 dans la suite du manuscrit, il s'agira du timepoint 30 et donc du temps 7.5 heures et T_{100} ou T_{final} correspond au timepoint 129, c'est-à-dire à la prise d'une image au temps 32.25 heures.

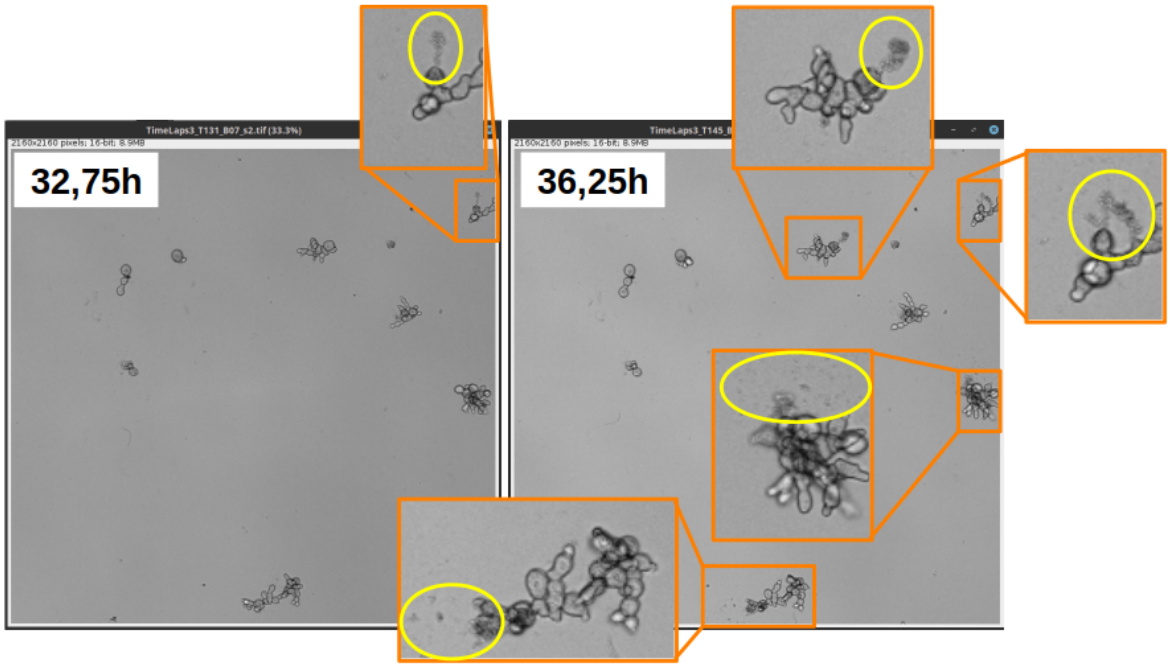


FIGURE 5.7 – Image 2 du champ pris dans le puits 7 aux temps 32.75 et 36.25 heures, illustrant l'explosion des cellules de champignon.

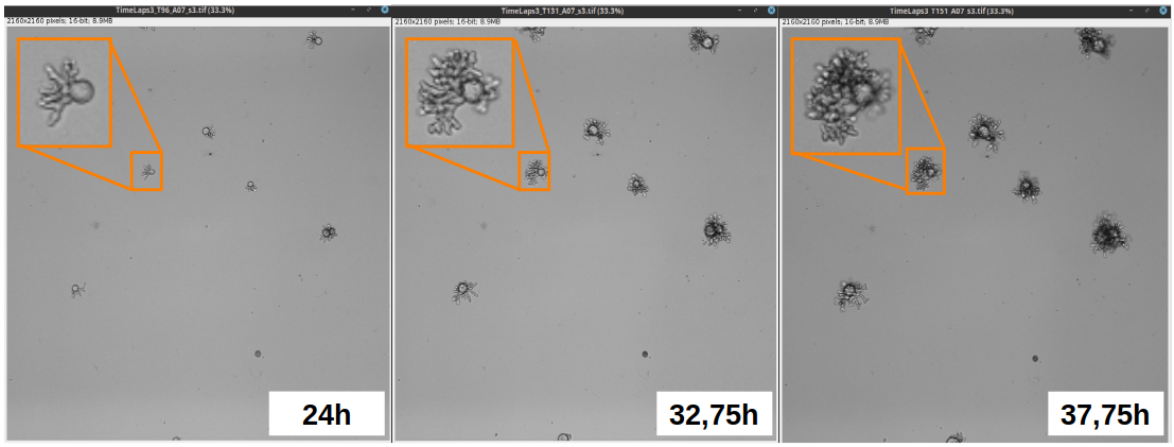


FIGURE 5.8 – Image 3 du champ pris dans le puits 7 aux temps 24, 32.75 et 37.75 heures, illustrant le développement des cellules de champignon de phénotype 2.

5.4 Segmentation et squelettisation avec cohérence temporelle

Pour le traitement des images considérées, nous procédons selon les trois étapes suivantes :

1. Nous générons et corrigeons les masques binaires des objets détectés suivant une hypothèse d'inclusion des champignons d'un temps à l'autre.
2. Nous découpons les objets tangents grâce à la méthode Watershed.
3. Les squelettes peuvent fortement varier d'un temps à l'autre malgré l'étape 1, nous corrigeons donc les squelettes de sorte que :
   Squelette à T_n = Squelette à T_{n-1} + branches apparues entre les deux temps.

Ces étapes sont décrites et discutées plus en détails dans cette partie du chapitre.

5.4.1 Segmentation et division des champignons tangents

Concernant la méthode de segmentation, il s'agit de celle utilisée pour binariser les objets des images afin d'en extraire des paramètres phénotypiques et d'utiliser ces caractéristiques comme descripteurs dans la méthode de classification des forêts aléatoires. Cette méthode est décrite dans le paragraphe 2.5.1 du chapitre 2.
Nous nous sommes fondés sur l'hypothèse que le champignon (des classes phénotypes 1,2,3, mycelium et spore) ne bouge pas au cours du temps mais grossit et crée des branches. Ainsi, il est raisonnable de faire l'hypothèse qu'une segmentation correcte d'un champignon au temps T_n contient nécessairement celle au temps T_{n-1}. Pour chaque champignon segmenté à T_n par la méthode du paragraphe 2.5.1, la segmentation est corrigée pour rendre l'hypothèse valide (voir la figure 5.9). Un exemple du résultat de la segmentation des images aux temps T_0 et T_{100} est présenté sur la figure 5.10.

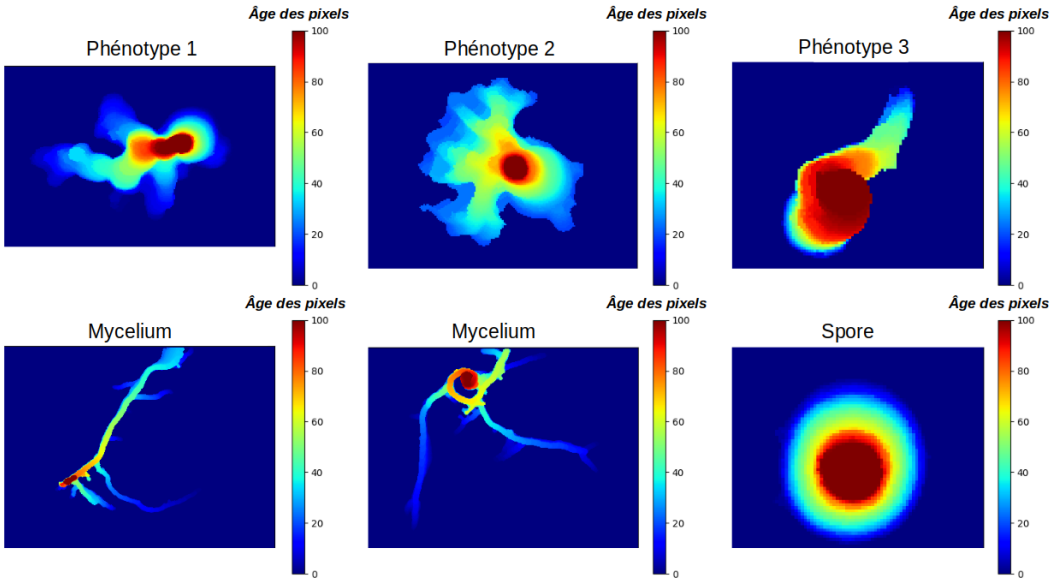


FIGURE 5.9 – Somme des masques binaires au cours du temps de certains objets (exemples de phénotypes 1,2,3, mycelium et spore) du temps 0 au temps 100. Lorsque la couleur du pixel tend vers le rouge, le pixel est présent dans un nombre croissant de masques, et inversement lorsque sa couleur tend vers le bleu.

Tout comme relevé dans le paragraphe 2.5.1, les masques binaires des champignons sur lesquels sont extraits les paramètres ne sont pas optimaux. En effet, il est fréquent que certains champignons se chevauchent et qu'ainsi les objets détectés correspondent à plusieurs champignons au lieu d'un (voir l'objet numéroté 8 sur la figure 5.10 qui correspond à deux champignons). L'avantage de travailler avec des séquences temporelles est de ne pas être limité à une seule image prise à 24 heures. L'utilisation des images des champignons à des temps en amont, nous permet la séparation des champignons d'un objet. Pour cela, nous utilisons un algorithme classique de segmentation par ligne de partage des eaux, le Watershed en anglais. Cette méthode simule l'inondation d'une image considérée comme un relief topographique. Afin de l'utiliser sur un objet, des "graines" doivent être renseignées. Ces graines sont des marqueurs utilisés comme initialisations spatiales des régions produites par le processus de segmentation. Un marqueur par champignon est donc nécessaire. Au temps T_0, les champignons sont à l'état de spore. Autrement dit les objets détectés présentent une très forte probabilité de correspondre à un unique champignon. Le masque à ce temps-ci correspond au masque référence, indiquant notamment le nombre de champignon détectés. La correction de la labellisation des pixels est donc effectuée de façon successive d'un temps à l'autre (sauf au temps T_0) et les marqueurs au temps T_n sont calculés sur les masques labellisés corrigés du temps précédent T_{n-1}. L'objet initialement labellisé 8, correspondant à deux champignons, est scindé en deux par le Watershed (voir la figure 5.12). La figure 5.13 présente la labellisation de ces deux champignons au temps T_{88} avant leur chevauchement ainsi que le résultat de re-labellisation des pixels en utilisant le Watershed, aux temps T_{89}, T_{90} et T_{91} où ils se chevauchent. Une fois le Watershed appliqué nous obtenons donc deux objets labellisés 8 et 9 (pour correspondre aux mêmes labels qu'au temps précédent). Appliqué à l'ensemble de la séquence temporelle, cet algorithme permet l'obtention de 100 masques dont les pixels sont labellisés de 1 au nombre d'objets détectés au temps T_0 (voir la figure 5.14).

5.4.2 Squelettisation

À partir de ces masques finaux, les squelettes des objets sont extraits. L'hypothèse selon laquelle le champignon grossit et crée des branches au fil du temps est illustrée sur la figure 5.9. Cette hypothèse implique que les squelettes soient corrigés d'un temps à l'autre afin de se correspondre. Ainsi, le squelette au cours du temps T_n doit correspondre à celui au temps T_{n-1} et présenter ou non des branches en plus. Pour ce faire, l'ensemble des squelettes sont tous d'abord calculés. Le squelette au temps T_0 est conservé tel quel. Puis, dans l'ordre chronologique, le squelette S1 au temps T_{n-1} est dilaté et retiré du squelette S2 au temps T_n. Le squelette résultant S3 correspond aux branches apparues entre les deux temps. Enfin, le squelette final S4 correspond à S1 + S3. Une étape de dilatation et de re-squelettisation permet de reconnecter le tout. Une fois les squelettes des objets corrigés (voir les figures 5.15 et 5.16), les graphes correspondants sont construits.
Cette procédure peut donner dans certains cas des résultats non optimaux. En effet, cette tâche s'est avérée assez complexe et l'approche actuelle de correction des squelettes mérite d'être améliorée.

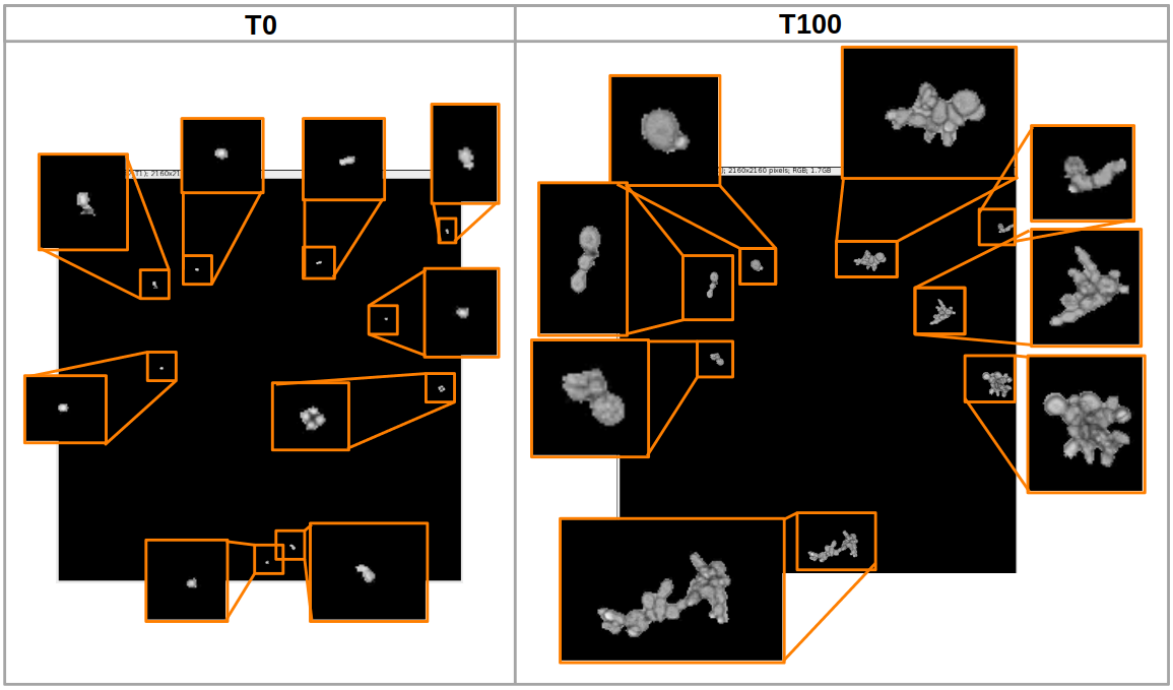


FIGURE 5.10 – Résultats de segmentation de deux images d'un même TimeLapse du phénotype 1, images prises dans le puits 7 de la plaque (concentration en molécule de 1.23µm) : à T_0 (à gauche) et T_{100} (à droite). Les masques binaires sont multipliés par les images originales permettant une meilleure vérification de la segmentation des objets.

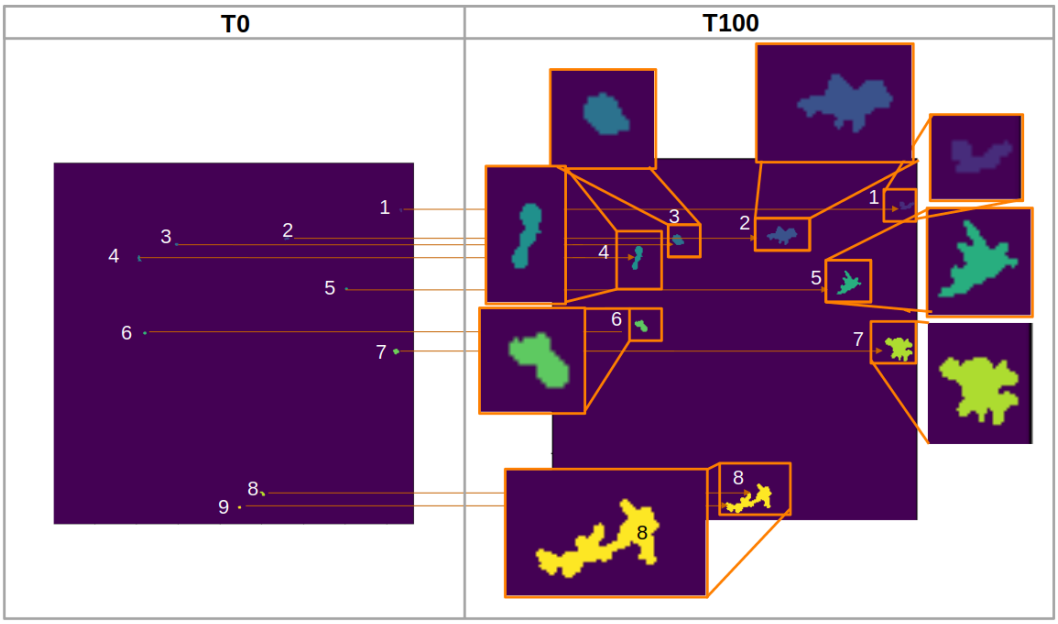


FIGURE 5.11 – Résultats de labellisation des objets de deux images d'un même TimeLapse du phénotype 1, images prises dans le puits 7 (concentration en molécule de 1.23µm) : à T_0 (à gauche) et T_{100} (à droite). Les objets sont mis en correspondance d'un temps à l'autre.

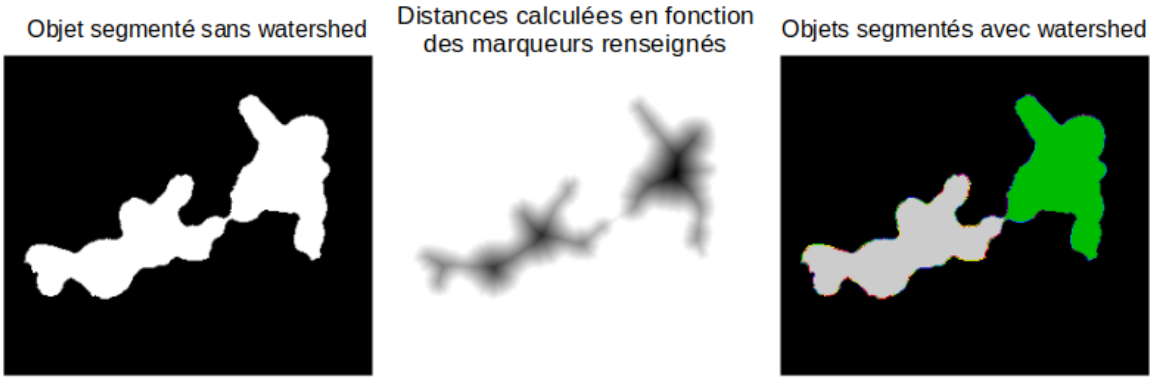


FIGURE 5.12 – Résultat de l'utilisation de l'algorithme de segmentation Watershed pour la séparation des champignons d'un même objet au temps T_{89}.

FIGURE 5.13 – Résultats de la correction de labellisation des pixels (objets aux temps T_{88}, T_{89}, T_{90} et T_{91}) par l'algorithme de Watershed dans le cas de la segmentation des champignons initialement numérotés 8 et 9.

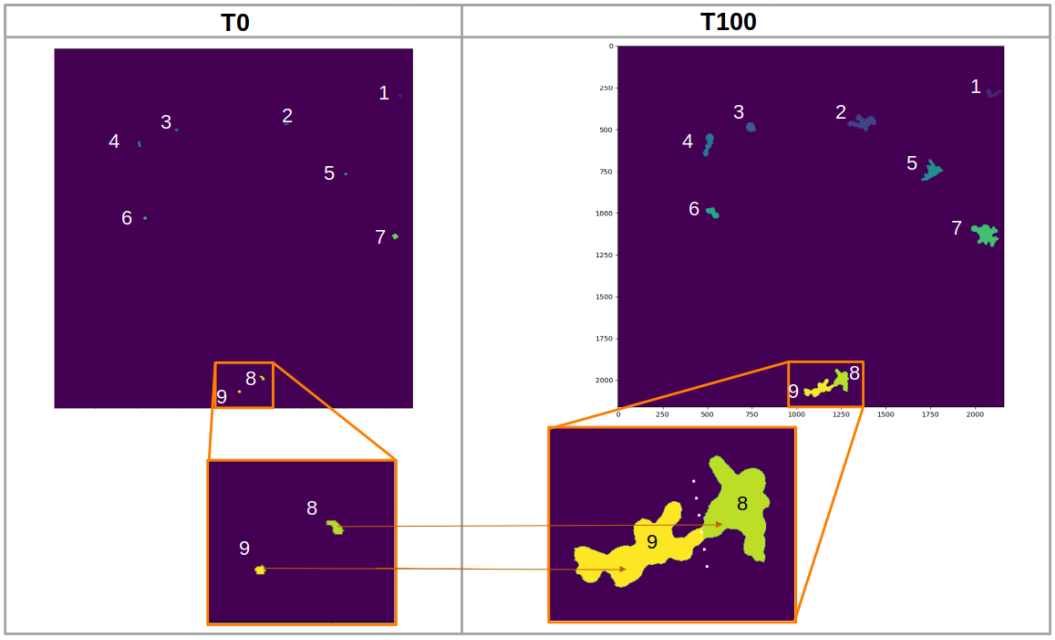


FIGURE 5.14 – Résultats de labellisation des objets des images du TimeLapse : à T_0 et T_{100} . Les objets sont mis en correspondance d'un temps à l'autre. La labellisation est corrigée grâce à l'algorithme de Watershed.

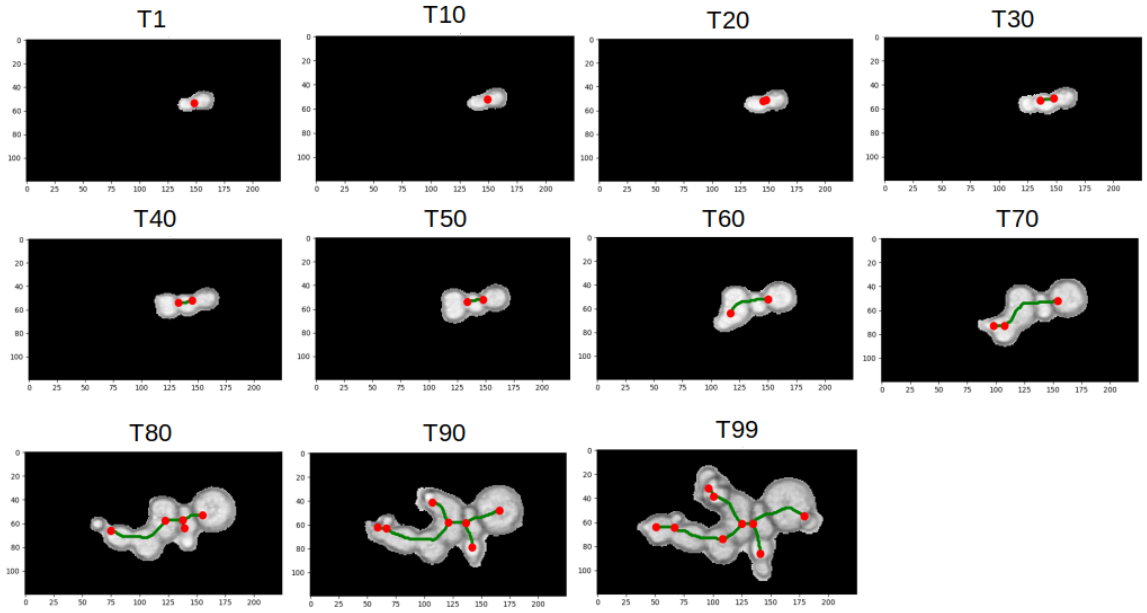


FIGURE 5.15 – Résultats sur onze temps, de la superposition du graphe obtenu, sur le masque binaire du champignon détecté (de phénotype 1).

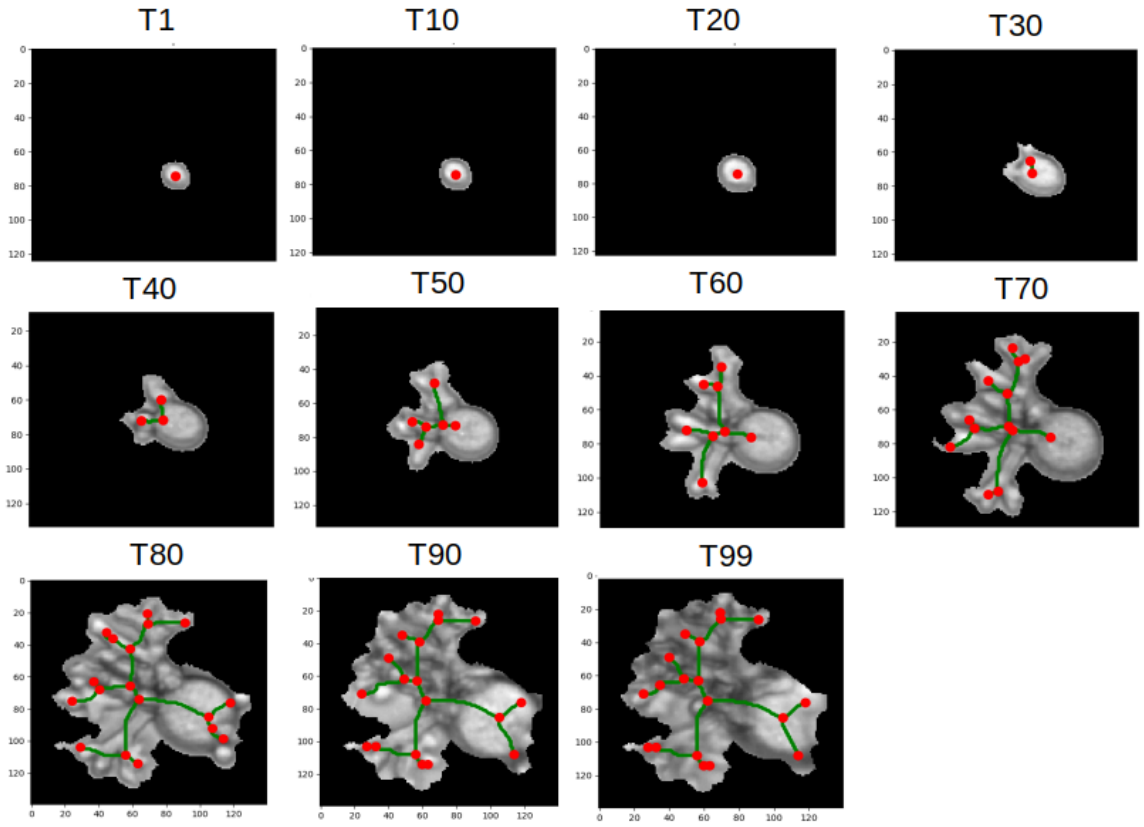


FIGURE 5.16 – Résultats sur onze temps, de la superposition du graphe obtenu, sur le masque binaire du champignon détecté (de phénotype 2).

5.5 Extraction des paramètres morphométriques et topologiques : les paramètres phénotypiques

L'idée ici est d'extraire des mesures de paramètres afin de modéliser leur évolution au cours du temps via des fonctions paramétriques. Les paramètres que l'on a choisi d'extraire sont présentés dans le tableau 5.2 :

Extrait de :	Mesure
Masque binaire	Aire des objets
Squelette	Longueur totale
Graphe	Nombre total de branches
	Nombre de branches incidentes à la spore initiale (dites branches primaires)

TABLEAU 5.2 – Paramètres phénotypiques calculés sur les trois formes des objets détectés à l'image : le masque binaire, le squelette et le graphe.

Lors de la visualisation des séquences temporelles, nous nous sommes rendu compte de l'hétérogénéité intra-classe de certains phénotypes (voir la figure 5.17) d'une part mais également de la présence de spores ne s'étant pas du tout développées. Nous avons donc décidé d'étudier la distribution des paramètres des échantillons et d'appliquer un seuil sur l'aire des objets afin d'identifier et de retirer les spores de l'ensemble des objets détectés (voir la figure 5.18). D'autre part, nous avons identifié les valeurs aberrantes de nos données en utilisant les propriétés du graphique appelé "boîte à moustaches". La boîte à moustaches permet de résumer une variable de manière simple et visuelle. En effet, ce graphique permet de comprendre la répartition des observations et d'identifier les valeurs extrêmes (outliers) ainsi que de comparer un même caractère dans plusieurs distributions. Les échantillons situés en dehors des moustaches au-dessus du 3e quartile (Q3) et en-dessous du 1er quartile (Q1) sont écartés de l'étude (voir la figure 5.19).

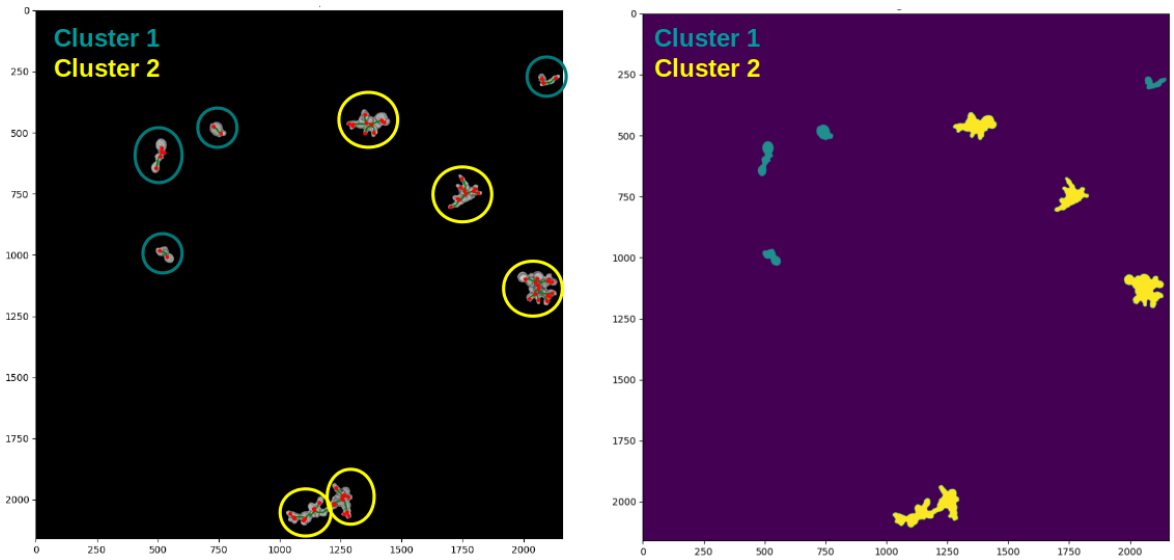


FIGURE 5.17 – Deux clusters de phénotype de champignons. Les champignons sont regroupés suivant leurs valeurs de paramètres phénotypiques, par la méthode de clustering k-Means. À gauche : les champignons et leurs graphes correspondants. À droite : masque binaire correspondant.

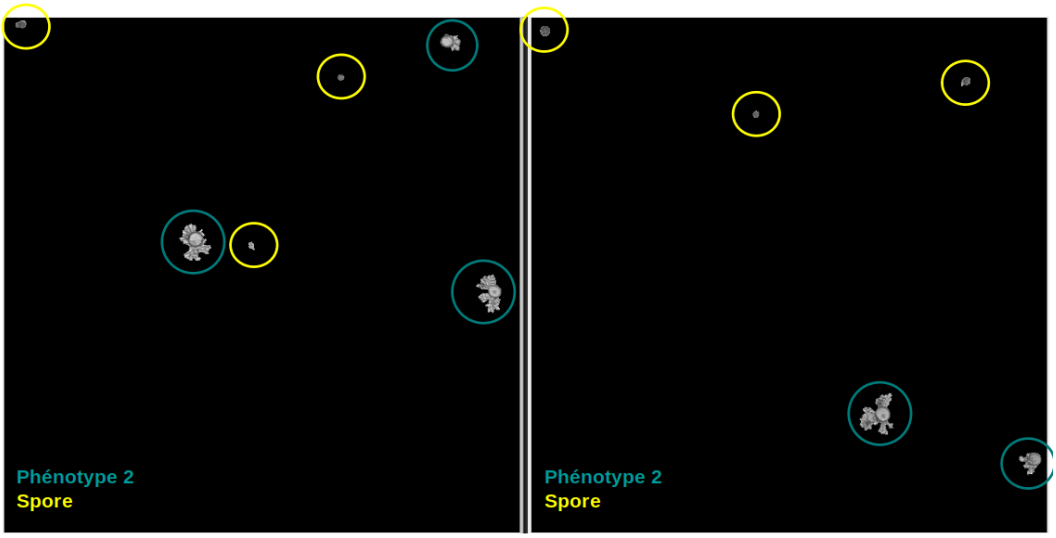


FIGURE 5.18 – Deux images de champignons traités avec la molécule A qui présentent pour certains le phénotype 2 (cercles bleu) et pour d'autres, la forme spore (cercles jaune).

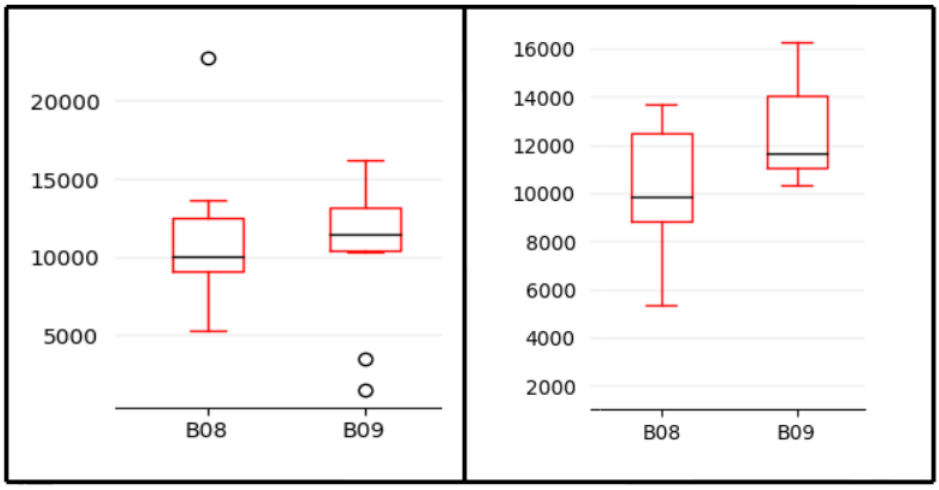


FIGURE 5.19 – Boîtes à moustaches avant (à gauche) et après (à droite) suppression des valeurs aberrantes dans le cas des champignons traités avec la molécule B aux concentrations 08 et 09. La barre noire correspond à la médiane des observations et le rectangle de la boîte va du premier quartile au troisième quartile. La longueur des moustaches vaut 1,5 fois l'intervalle inter-quartile.

5.6 Étude de l'influence de la concentration en molécule sur le phénotype 1

Dans le reste du manuscrit, nous nous sommes focalisés sur le phénotype 1.
Nous avons dans un premier temps étudié l'évolution des paramètres phénotypiques au cours du temps, pour les différentes concentrations en molécule permettant l'obtention d'un phénotype remarquable plus ou moins marqué. Plus la concentration en molécule est élevée plus le développement du champignon est influencé. L'effet peut être plus ou moins prononcé et son étude donne une information relative à l'efficacité de la molécule.

La figure 5.20 présente l'évolution, par concentration, de la moyenne des paramètres phénotypiques des champignons des phénotypes 1, 2 et 3. Un exemple de champignon détecté à chaque concentration est également affiché.

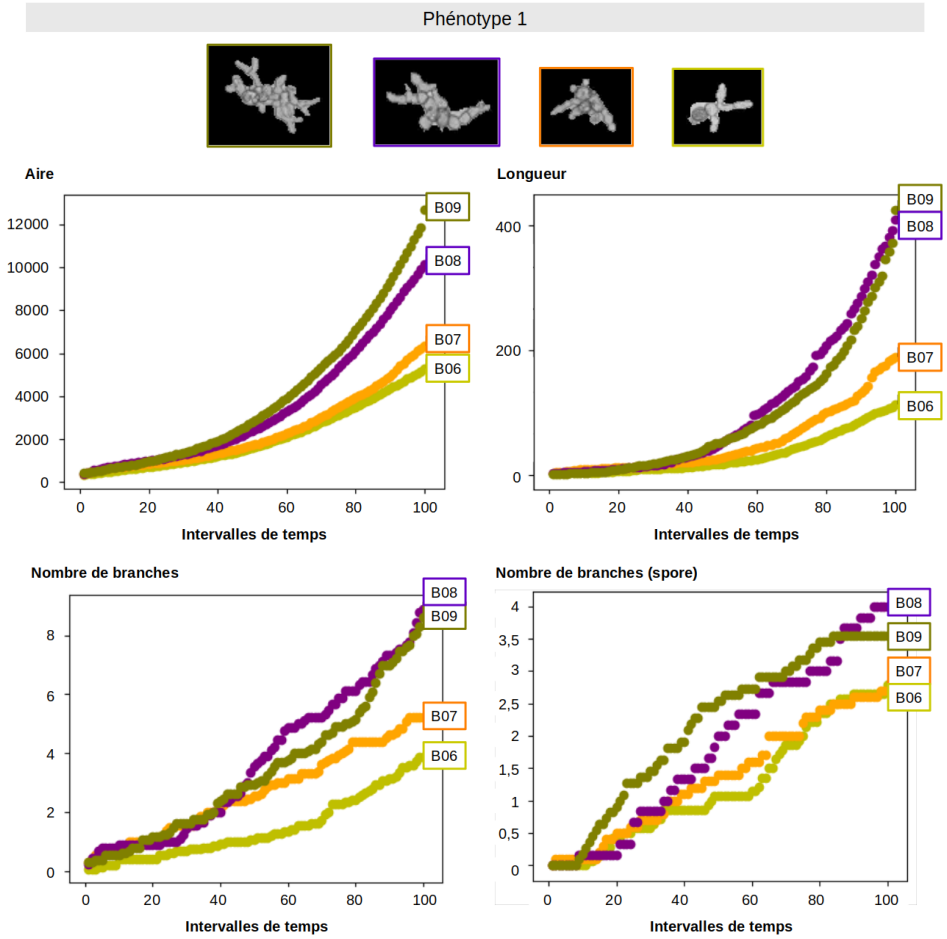


FIGURE 5.20 – Évolution des paramètres phénotypiques (moyenne sur les champignons d'une même concentration) du phénotype 1 au cours du temps. **B :** molécule dont le MoA entraîne le phénotype 1. **06 à 09 :** concentrations décroissantes en molécule.

Nous avons également observé la distribution des valeurs des quatre paramètres phénotypiques, pour l'ensemble des objets de la classe phénotype 1. Les boîtes à moustaches (présentées sur la figure 5.21) permettent de nous rendre compte de la répartition des différentes distributions. Une description des résultats obtenus ainsi que des informations extraites à partir des figures est effectuée ci-après :

— Aire :

La médiane est située plus haut que la moitié de la boîte pour les concentrations les plus élevées en molécule (06 et 07) contrairement aux concentrations 08 et 09. Aux concentrations 06, 07 et 09, 25% des plus petits objets présentent une faible variabilité d'aire contrairement aux 25% plus gros objets. Plus précisément, dans le cas de la condition 09, cette faible variabilité d'aire s'applique aux 50% des plus petits objets. À la concentration 08, les 25% des plus gros objets et les 25% des plus petits présentent respectivement une faible et une forte variabilité de ce paramètre.
Nous observons donc, à toutes les conditions, que la moitié des observations présente une variabilité plus faible que l'autre moitié. Cela fait écho aux observations effectuées sur les images, à savoir la présence de champignons plus et moins développés (séparables en deux groupes, voir la figure 5.17).

— Longueur :

Dans le cas de ce paramètre, les données aux concentrations les plus élevées (06, 07 et 08) sont présentées par des boîtes asymétriques d'environ la même taille et dont la médiane est située plus haut que la moitié de la boîte. Au contraire, la boîte de la condition 09 est symétrique et la médiane est située au milieu du 1er et 3ème quartiles.
Certains objets à la concentration la plus faible (09) présentent des longueurs de squelette moins élevées qu'à des concentrations plus fortes en molécule, phénomène que nous n'observons pas dans le cas du paramètre aire. Ainsi, une aire qui augmente ne veut pas toujours dire longueur qui augmente également, certains objets semblent grossir mais pas s'allonger. Cela reflète une fois de plus la présence de champignons plus et moins développés, une hétérogénéité du phénotype qui semble plus importante à mesure que la concentration en molécule diminue.

— Nombre de branches :

Nous pouvons remarquer que les boîtes des conditions 06 et 07, présentent quasiment le même intervalle de valeurs, avec des objets ne présentant aucune branche pour la concentration la plus élevée (06). On devine ainsi qu'à cette concentration certains champignons sont restés à l'état de spore.
La boîte à la concentration 09 présente environ le même intervalle de valeurs (9 branches) que les concentrations 06 et 07. En reprenant la constatation émise dans le cas de la distribution des longueurs pour cette condition, c'est-à-dire, que certains objets semblent grossir mais ne pas s'allonger, nous pouvons supposer que ces objets présentent un nombre de branches faible (3 à 6 branches). En effet, nous partons du principe que plus un objet a de branches plus la longueur totale de son squelette augmente.

— Nombre de branches à partir de la spore principale :

La médiane aux concentrations 06 et 07 est de trois branches et celle aux concentrations 08 et 09 est de quatre. Nous pouvons en déduire que la concentration en molécule a moins d'influence sur le nombre de branches primaires que sur le nombre de branches terminales.

Globalement, les figures 5.20 et 5.21 nous permettent de voir que les valeurs des paramètres sont plus élevées à mesure que la concentration en molécule diminue. Les mesures effectuées sur les squelettes et graphes des objets détectés sur les images suivent ainsi la logique de l'efficacité d'un traitement bloquant le développement du champignon. Plus la molécule est efficace, plus le champignon tend vers une forme spore et moins elle l'est, plus le champignon tend vers l'état mycelium. Nous pouvons noter que les phénotypes aux concentrations les plus élevées semblent plus proches que ceux aux concentrations les plus faibles et inversement. Néanmoins, à la concentration la plus faible, la molécule ici considérée présente également des objets qui grossissent mais qui font peu de branches. Ces phénotypes particuliers ainsi que la variabilité des paramètres phénotypiques à chaque condition indiquent l'importance de prendre en compte l'hétérogénéité des champignons du phénotype 1 dans la conception de notre modèle de simulation.

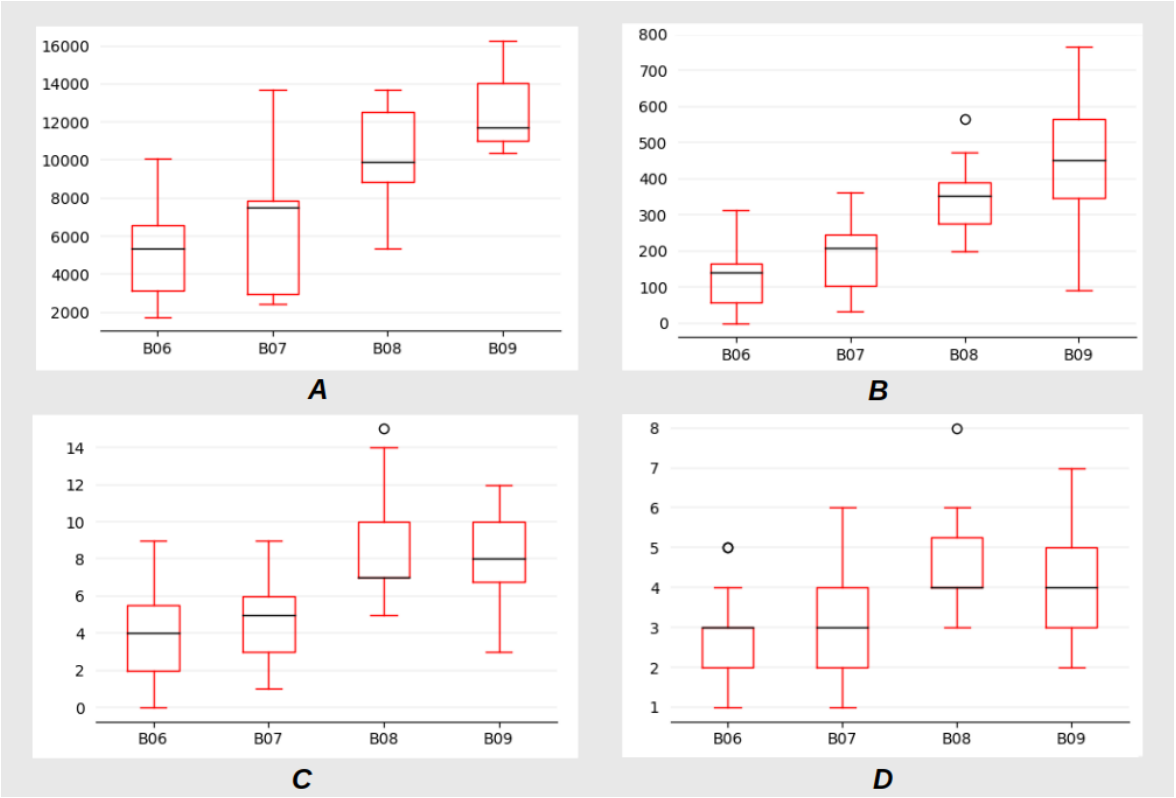


FIGURE 5.21 – Distribution des valeurs (à T_{100}) des quatre paramètres phénotypiques pour les concentrations où le phénotype 1 est observable. **A :** aire, **B :** longueur, **C :** nombre de branches et **D :** nombre de branches partants de la spore initiale. Les cercles correspondent aux outliers. La médiane est représentée par la ligne noire dans la boîte.

5.7 Analyses complémentaires

5.7.1 Carte de trajectoires : profil de la molécule B

Le concept de carte de trajectoires repose sur l'évolution de la valeur des paramètres phénotypiques des champignons en fonction du temps. Elle apporte une possibilité supplémentaire de classer le phénotype des champignons d'une image sur la base de sa carte de trajectoires et d'évaluer l'efficacité du composé dans le temps (propriétés cinétiques du composé antifongique) par l'allure de son profil. Les paramètres phénotypiques sont pris deux à deux pour construire les profils des classes. La figure 5.22 illustre un exemple de champignon pour chaque concentration. La figure 5.23 correspond à la carte de trajectoires construite sur la base des paramètres aire et longueur des objets détectés sur les images du test de la molécule B aux concentrations 02, 03, 04, 05, 06, 07, 08, 09, 10, 11 et 12. Pour rappel, la concentration 02 correspond à 0 µL de molécule, il s'agit de la condition contrôle, les cellules de champignons se développent en l'absence de molécule.
La carte de trajectoires (voir la figure 5.23) nous permet d'évaluer l'effet de la molécule B sur les champignons sur une gamme de concentrations. Ainsi, nous observons que les profils des différentes formes de champignons (spore en bleu, phénotype remarquable 1 en rouge et mycelium en vert) sont distinguables. Cette figure permet également de montrer que la molécule à la concentration 12 présente encore un léger impact sur le champignon car le profil n'a pas atteint celui de la condition contrôle (02). Enfin, le profil pour les conditions 10 et 11 correspond plus à du mycelium intermédiaire, une forme intermédiaire entre phénotype remarquable et mycelium.

La figure 5.23 nous permet de nous rendre compte que pour l'ensemble des phénotypes, la carte de trajectoires donne des indications sur la morphogénèse du champignon. En effet, nous pouvons remarquer que les profils correspondent globalement à des courbes droites, ce qui nous permet de dire que la croissance s'effectue essentiellement suivant la longueur et à épaisseur constante.

5.7.2 Comparaison des profils de différentes molécules

Nous avons généré deux TimeLapses. Le premier correspond aux images décrites dans le paragraphe 5.3.1 et dont la concentration en spores a été fixée à $0,5.10^3$ spores/ml. Concernant le deuxième TimeLapse, cette concentration en spores est de $0,5.10^4$ spores/ml soit un nombre dix fois plus important en cellules de champignons (voir la figure 5.5). Pour rappel, plus la concentration en champignon est élevée plus le risque de chevauchements et par conséquent de biais dans les mesures des paramètres augmente.

Le premier TimeLapse comprend le test de trois molécules avec trois Modes d'Action différents, les molécules A, B et C qui donnent les phénotypes 2, 1 et 3. Concernant le deuxième TimeLapse, cinq molécules différentes sont testées. La molécule D donne le phénotype 2 et les autres le phénotype 1. Les molécules G, I et K sont des réplicats des molécules F, H et J. L'ensemble de ces informations est regroupé dans le tableau 5.3.

TimeLapse	Code molécule	Phénotype
1	A	2
1	B	1
1	C	3
2	D	2
2	E	1
2	F	1
2	G	1
2	H	1
2	I	1
2	J	1
2	K	1

TABLEAU 5.3 – Tableau indiquant les molécules testées dans le cadre de cette étude. Les couleurs regroupent les molécules testées en réplicats.

Plusieurs études sont effectuées à partir de ces expériences. Les trois principales sur lesquelles nous nous sommes penchés sont les suivantes :

— TimeLapse 1 : Comparaison des trois molécules avec trois Mode d'Action différents 1, 2 et 3 (A, B et C).

  La figure 5.24 représente les profils des molécules A, B et C aux mêmes concentrations (06 et 07). Dans le paragraphe 2.5.4 du chapitre 2, nous avions relevé que les paramètres longueur et aire des objets (extraits sur des images à 24 heures) ne sont pas assez discriminants pour différencier les phénotypes 1 et 2. En prenant en compte l'évolution de ces paramètres au cours du temps, il semble que ce soit toujours le cas. Malgré une longueur plus importante et une aire moins élevée pour le

phénotype 2 au cours du temps, les profils ne sont pas assez discriminants. En effet, nous observons que la dynamique s'approche d'une croissance linéaire et donc que la loi est proportionnelle entre les deux paramètres. En revanche, nous constatons que le profil du phénotype 3 est différent du fait d'une accélération de la croissance de la longueur par rapport à l'aire.

— TimeLapse 2 :
  — Vérification des mesures pour les paires de réplicats (F/G, H/I et J/K).

    La figure 5.26 illustre les profils des molécules F, H et J ainsi que ceux de leurs réplicats G, I et K. Cette figure permet de vérifier que les profils de ces molécules et de leurs réplicats sont très proches et ainsi valider les mesures extraites.
  — Comparaison de quatre molécules différentes ayant le même Mode d'Action 1 (E, F, H et J).

    Les profils de l'ensemble des molécules testées ayant le Mode d'Action 1 et entraînant l'apparition du phénotype 1 sont tracés sur la figure 5.27 à l'exception des réplicats. En effet, dans un souci de lisibilité de la figure et ayant validé *via* la figure 5.26 que les profils entre molécules et réplicats sont très proches, nous avons décidé de ne pas afficher leurs profils. La figure 5.27 nous permet donc de comparer le profil de molécules différentes mais ayant le même Mode d'Action 1.

De par la comparaison de ces profils, nous pouvons émettre les conclusions suivantes :

— Contrairement aux autres molécules, la molécule H est encore très efficace aux concentrations 07, 08 et 09. Elle est donc la molécule la plus efficace.
— La molécule J est la moins efficace car elle nécessite une forte concentration (04, 05) pour bloquer la croissance du champignon.
— Les molécules E et F semblent avoir des profils très similaires pour les concentrations 07 et 08. En revanche à la concentration 09, nous pouvons remarquer, au vu de son profil, que les champignons sont plus développés en présence de la molécule F. La molécule E est donc légèrement plus efficace que la molécule F.
— En comparant les molécules B et F, on se rend compte que la dynamique de F07 se trouve entre celles de B06 et B07, celle de F08 se trouve entre celles de B07 et B08 et que celles de F09 et B09 sont très proches. Nous pouvons en conclure que la molécule B est légèrement moins efficace que la F.

Une bonne molécule est une molécule efficace à faible concentration. Ainsi, sur la base des profils (phénotype/MoA 1) obtenus et des conclusions émises, nous pouvons établir le classement suivant :
H > E > F > B > J

— TimeLapses 1 et 2 : Comparaison de deux différentes molécules ayant le même Mode d'Action 2 (molécules A et D).

La figure 5.25 illustre les profils des molécules A et D, deux molécules différentes avec le même Mode d'Action 2. Au vu des résultats obtenus, il semble que la molécule D soit plus efficace que la A. En effet, nous pouvons noter que les profils A06 et D09 ainsi que A07 et D10 sont très proches. Ainsi, pour une concentration bien moins élevée (06 : 4µM, 07 : 1.2µM, 09 : 0.2µM et 10 : 0.05µM) la molécule D semble tout aussi efficace que la molécule A.

5.7.3 Conclusions

La construction de ces cartes de trajectoires peut constituer un outil utile dans l'élucidation du mode d'action de nouvelles molécules mais également un outil permettant l'obtention d'informations sur leurs propriétés cinétiques. En effet, dans le cas d'une nouvelle molécule :

1. La molécule est testée à plusieurs concentrations et une séquence temporelle est générée.
2. Les profils pour chaque concentration sont générés et comparés à des profils "types" des différentes classes (Spore, phénotype (1-4) et Mycelium).

Cette comparaison nous apporte des indications sur les propriétés de cette nouvelle molécule. Les concentrations où le phénotype apparaît sont identifiées et les profils à ces concentrations sont étudiés. En fonction de ces indications, la molécule pourra être considérées comme intéressante ou non pour la suite des recherches.

5.7.4 Perspectives

Ces travaux nécessitent un grand nombre de données. Afin de renforcer et d'approfondir cette méthode de nouvelles séquences temporelles doivent être générées avec un protocole fixé. En effet, il est difficile de conclure sur deux molécules dont les profils sont construits sur la base d'expérimentations différentes (différentes concentrations en spores). D'autre part, il est nécessaire comme pour toute expérience scientifique de générer des réplicats afin d'attester de la robustesse des résultats obtenus. Enfin, les cartes de trajectoires sont ici construites en ne prenant en compte que deux paramètres phénotypiques mais nous pouvons imaginer des profils 3D ou plus permettant une meilleure distinction des différents phénotypes.

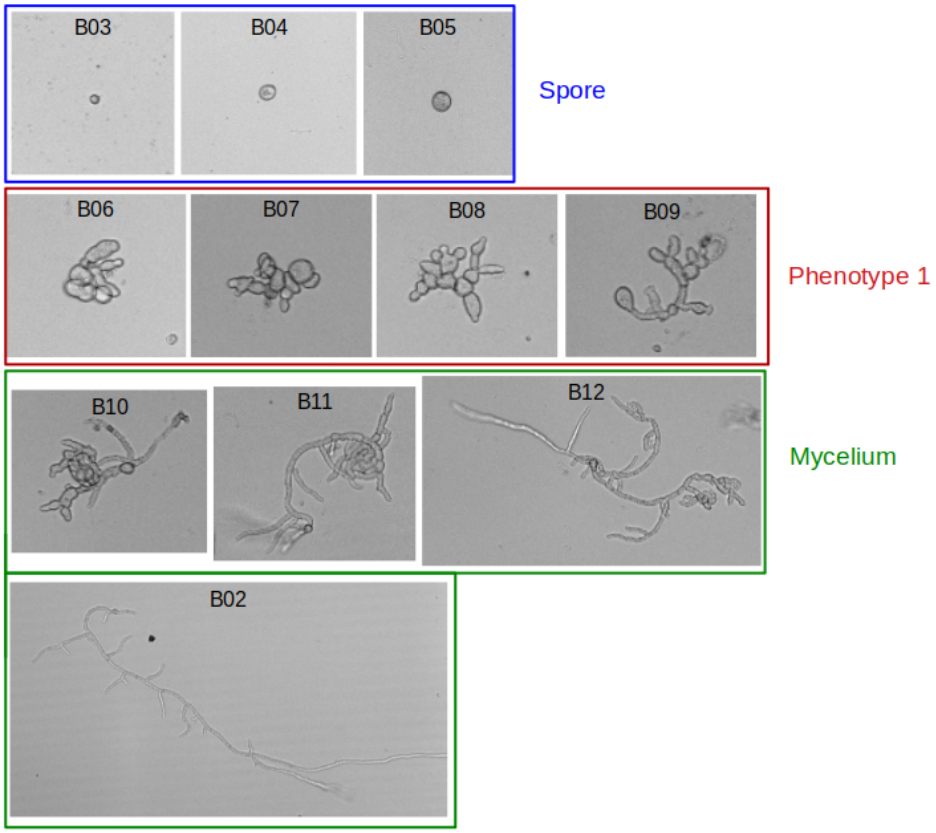


FIGURE 5.22 – Exemples de champignons présents sur les images, issues du test de la molécule B, aux concentrations 02, 03, 04, 05, 06, 07, 08, 09, 10, 11 et 12.

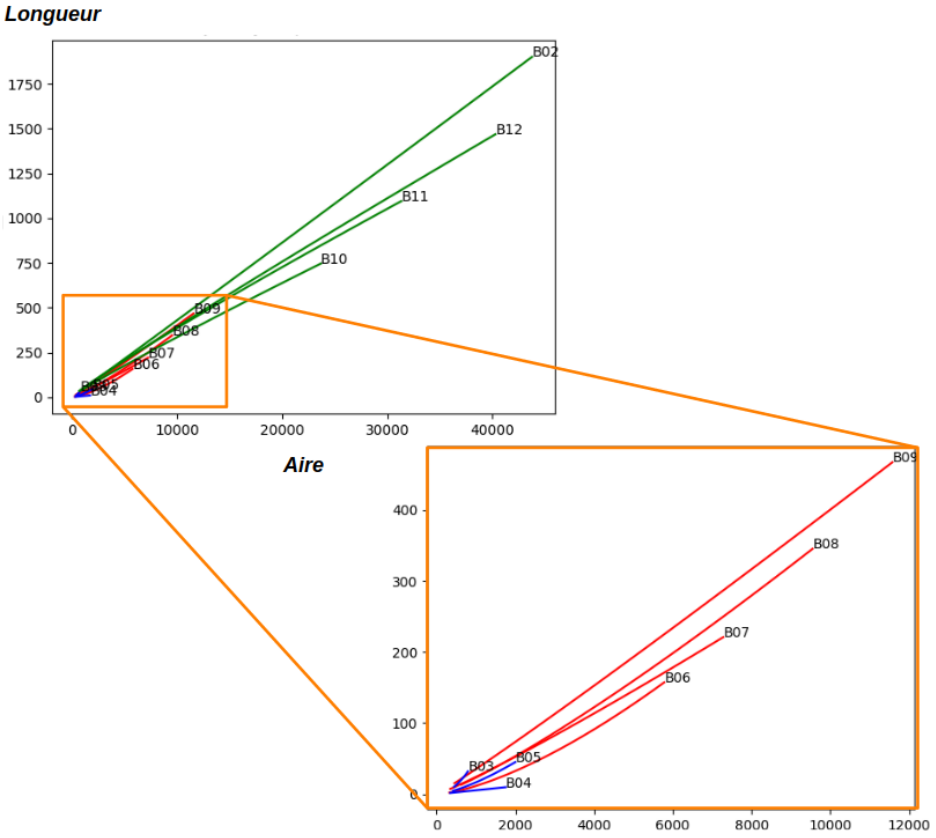


FIGURE 5.23 – Carte de trajectoires des valeurs moyennes des paramètres aire et longueur des objets détectés dans les images issues du test de la molécule B. Les courbes bleues correspondent au phénotype spore (concentrations 03, 04 et 05), les rouges au phénotype 1 (concentrations 06, 07, 08 et 09) et les vertes à la forme mycelium (concentrations 02, 10, 11 et 12).

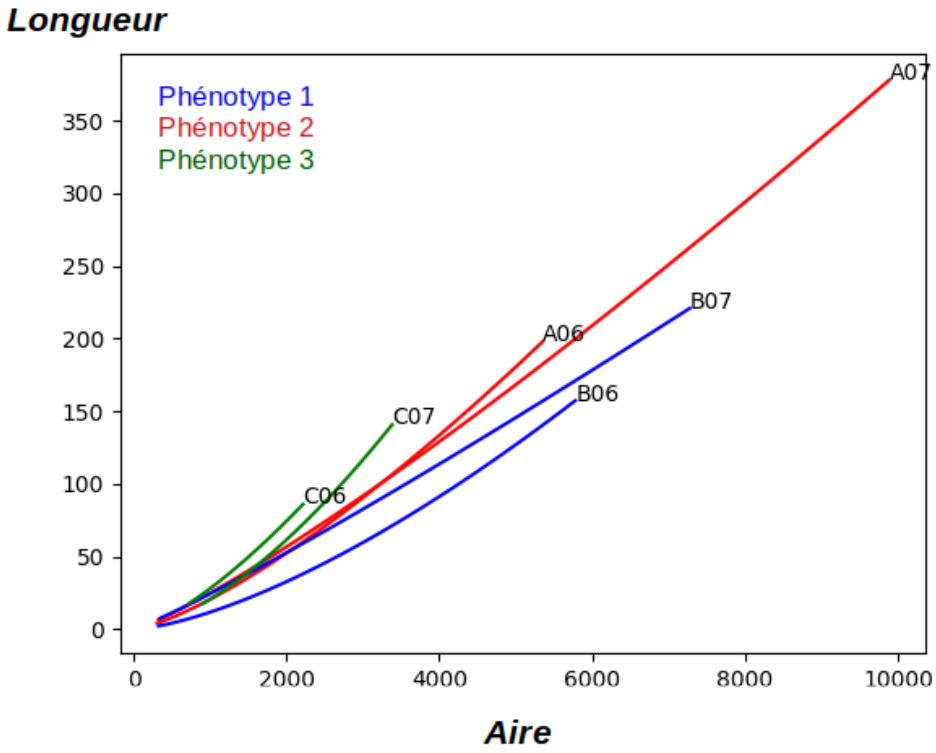


FIGURE 5.24 – Carte de trajectoires des valeurs moyennes des paramètres aire et longueur des objets détectés dans les images issues du test des molécules A (courbes rouges), B (courbes bleues) et C (courbes vertes) aux concentrations 06 et 07.

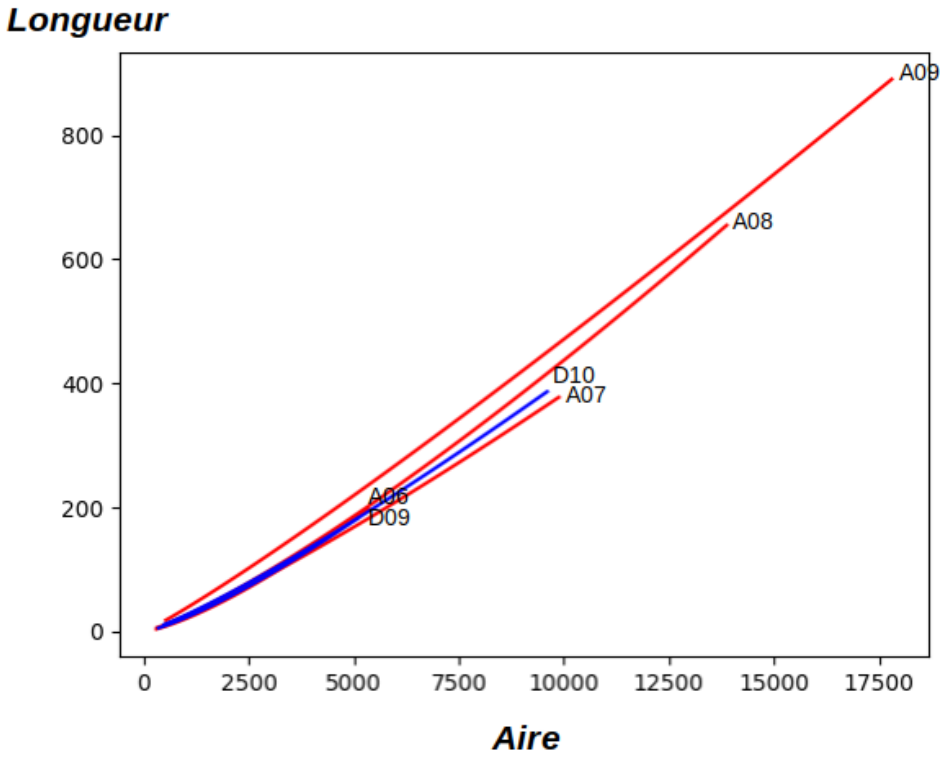


FIGURE 5.25 – Carte de trajectoires des valeurs moyennes des paramètres aire et longueur des objets détectés dans les images issues du test des molécules A (courbes rouges) et D (courbes bleues) aux concentrations où le phénotype 2 apparaît.

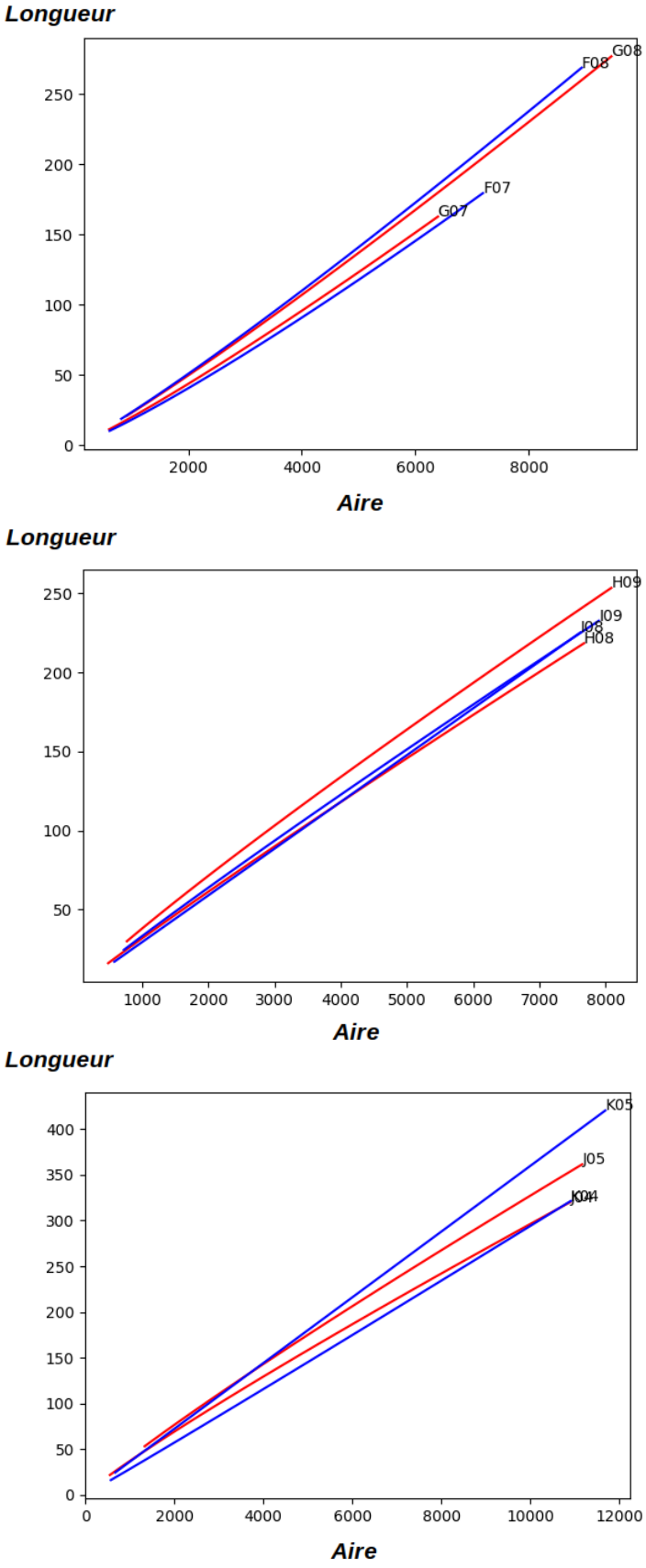


FIGURE 5.26 – Carte de trajectoires des valeurs moyennes des paramètres aire et longueur des objets détectés dans les images issues du test des molécules et de leurs réplicats. Dans l'ordre de haut en bas : cas des molécules G et F, puis des molécules H et I et enfin des molécules K et J.

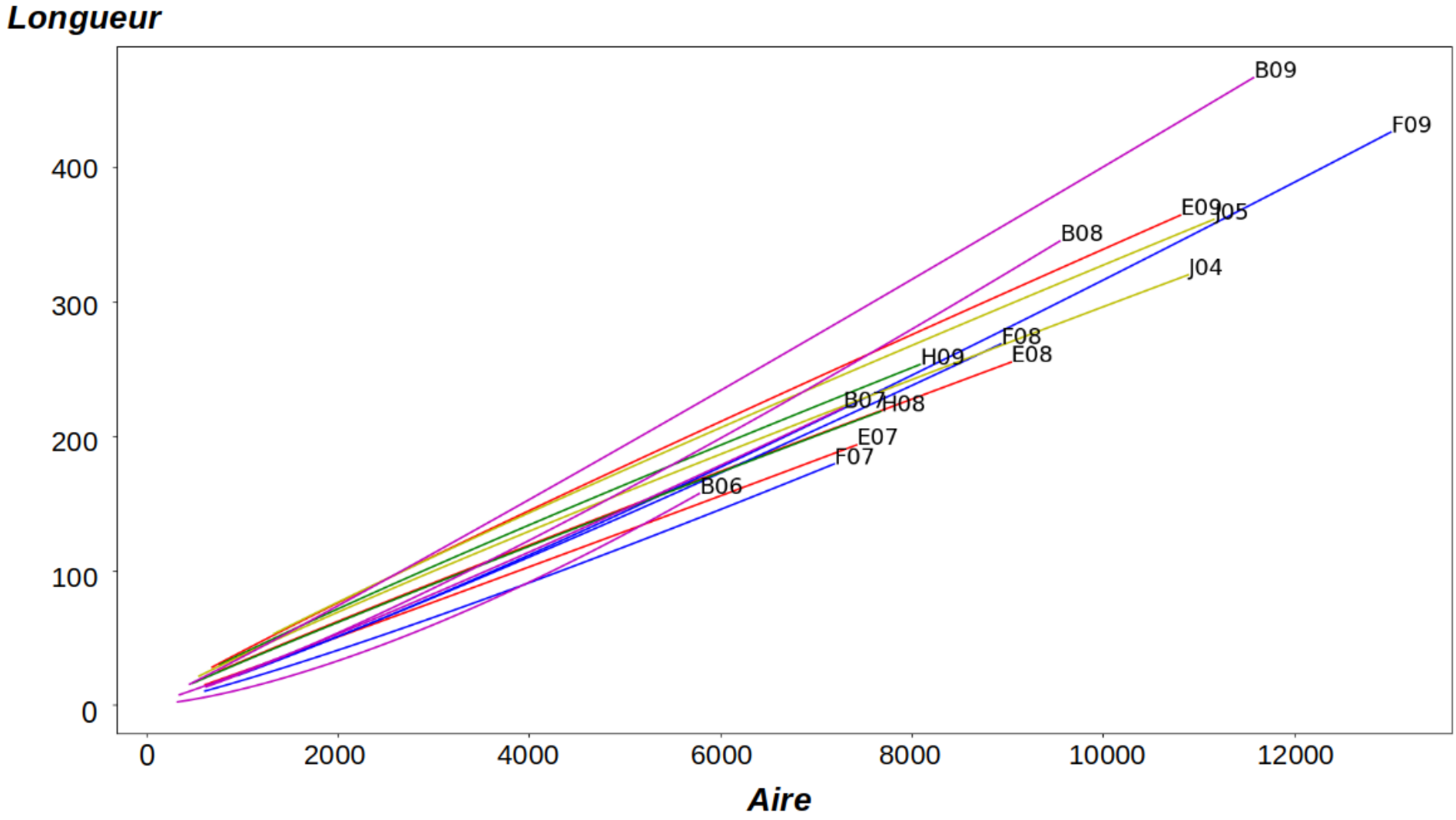


FIGURE 5.27 – Carte de trajectoires des valeurs moyennes des paramètres aire et longueur des objets détectés dans les images issues du test des molécules B (courbes roses), E (courbes rouges) F (courbes bleues), H (courbes moutardes), et J (courbes vertes) aux concentrations où le phénotype 1 apparaît.

Ces profils simples permettent d'obtenir des informations sur les composés antifongiques testés. L'élaboration d'un modèle discret de croissance permet d'aller plus loin dans l'étude de ces dynamiques. Le paragraphe suivant est consacré à la procédure de mise en place du modèle.

5.8 Développement d'un modèle de croissance

Au vu des courbes représentant l'évolution des paramètres aire et longueur (voir la figure 5.20), nous avons choisi une fonction exponentielle pour modéliser ces courbes. Afin de conserver l'information d'hétérogénéité (voir la figure 5.17)du phénotype, les coefficients de la fonction $f(x) = a * e^{b.x}$ sont estimés de façon indépendante pour chacun des champignons. La figure 5.28 présente les données réelles ainsi que les courbes obtenues après estimation de ces coefficients par la méthode des moindres carrés.

5.8.1 Paramètres des modèles

La distribution des coefficients estimés, à chaque concentration, est utilisée comme une fonction de densité de probabilité (PDF). Ainsi, nous avons deux PDFs (a et b) par concentration. Ces PDFs sont présentées sur la figure 5.29. Nous avons donc le choix d'un niveau d'efficacité, et que pour le niveau choisi, la fonction d'évolution de chaque caractéristique pour la croissance d'un champignon est définie par un tirage des coefficients a et b. Pour chaque nouvelle croissance, de nouvelles paires (a, b) sont tirées pour les différentes caractéristiques.

Sur la figure 5.29 , nous pouvons faire les constatations générales suivantes :

— Concernant les paramètres **aire** : plus la concentration en molécule diminue plus b augmente, à l'inverse le coefficient a augmente à mesure que la concentration diminue.

— Concernant le paramètre **longueur** : plus la concentration en molécule diminue plus la variabilité de la distribution du coefficient b diminue, la PDF est plus piquée. Le coefficient a augmente à mesure que la concentration en molécule diminue. Aux concentrations les plus élevée (06 et 07), on observe des objets présentant à la fois des longueurs plus et moins élevées qu'aux concentrations plus faibles.

— En comparant les coefficients a des deux paramètres, nous pouvons remarquer qu'à T_0 (T_0 = 30 ème acquisition) les spores ont commencé à germer et d'autres non. En effet, à T_0, $f(x) = a$. Or, nous pouvons noter que les distributions pour le paramètre **aire**, ne démarre pas à 0 (spores qui ont juste grossi mais n'ont pas fait d'extensions). Nous remarquons également, qu'aux concentrations les moins élevées en molécules (conditions 08 et 09) le paramètre **longueur** ne démarre pas non plus à 0 (les cellules de champignon ont commencé à se développer au temps T_0).

— Le support des distributions des coefficients b, est plus étalé vers des valeurs plus élevées pour la **longueur** que pour l'**aire**. Nous en déduisons que la longueur semble augmenter plus rapidement que l'air. Une hypothèse est que les branches des champignons, en s'allongeant, s'affinent.

Concernant les paramètres **nombre de branches** et **nombre de branches à partir de la spore initiale**, nous avons tout d'abord étudié le phénomène d'apparition de branches au cours du temps. La figure 5.30 représente le nombre de branches créées à chaque temps

sur l'ensemble des champignons du puits. Cette figure nous a permis de conclure que cet événement ne suit pas de tendance notable. Autrement dit, nous ne remarquons aucun intervalle de temps pour lequel la probabilité de créer des branches est plus élevée. Cette étude nous a décidé à considérer la création de branche comme un phénomène aléatoire uniforme au cours du temps. Nous calculons donc pour chaque concentration, les probabilités P suivantes :

$$P = \frac{\sum_{i=1}^{N}(B_i)}{N * T} \tag{5.1}$$

Avec T la durée jusqu'à l'acquisition finale considérée, N, le nombre d'objets considérés à la concentration étudiée et B_i le nombre de branches, soit primaires soit terminales qui sera précisé plus tard, du i-ème champignon au dernier temps.
Il s'agit de la probabilité pour un champignon de créer une branche à un instant donné de la séquence. Cette modélisation correspond à une épreuve de Bernouilli, ou binomial, de paramètre P. Cela implique qu'il ne peut pas y avoir création "simultanée" de plusieurs branches. Si l'échantillonnage temporel de la séquence est très faible (très longue pause entre les acquisitions), cette contrainte conduira à des croissances non réalistes (pas assez de branches). Une modification naturelle serait alors de remplacer l'épreuve de Bernouilli par un schéma de Bernoulli de paramètres n et P où n est le nombre d'épreuves indépendantes de Bernoulli et dépendrait de l'échantillonnage temporel (plus l'échantillonnage est faible, plus n doit être grand).

5.8.2 Élaboration du modèle

Un champignon est représenté par un graphe arborescent dont le premier nœud correspond à la spore initiale. Une branche est composée d'une ou plusieurs arêtes. La croissance du champignon est générée par des créations successives d'arêtes sans notion d'épaisseur.

Pour rappel, le temps T_0 correspond à la 30 ème prise de la séquence temporelle, autrement dit après 7.5 heures d'incubation. La raison est expliquée dans le paragraphe 5.3.2. Néanmoins, les cellules de champignons ont commencé à se développer de 0 à 7.5 heures. Afin de rendre notre simulation plus réaliste, nous avons décidé de faire croître à chaque itération une branche à partir de la spore dont les arêtes sont d'une longueur constante égale à $\frac{a}{30}$, a étant l'expression de la longueur à T_0 (voir le paragraphe A.7 en annexe). Puis, le nombre d'arêtes à chaque temps discret et leurs longueurs sont définis selon leurs PDFs respectives (loi uniforme et PDFs des coefficients a et b). Les arêtes peuvent être créées à partir de la spore ou d'une extrémité. On note P_{sb} la probabilité de créer une arête à partir du nœud spore et P_b la probabilité d'en créer une à partir d'un nœud d'extrémité.P_b et P_{sb} sont définies selon l'équation (5.1) avec respectivement B_i le nombre de branches terminales et B_i le nombre de branches primaires du i-ème champignon au dernier temps. Le nombre de nœuds d'extrémité est noté nb.

Le modèle de croissance nécessite au préalable la définition des paramètres suivants :

— Le type de molécule.

— La concentration.

— La valeur qui définit l'intervalle d'angles possibles lors de la création d'une nouvelle arête. On note θ_{sb} la valeur d'angle maximale concernant le nœud spore et θ_b celle concernant les nœuds d'extrémité.

L'algorithme de croissance du champignon *Botrytis c.* proposé est décrit ci-après :

- Pour la longueur : Tirer les coefficients a et b de la fonction exponentielle selon la PDF estimée.
- $l = \frac{a}{30}$
- Initialisation :
  - Tirer aléatoirement un angle o dans $[0 : 360[$ selon une loi uniforme.
  - Ajouter un nœud à une distance l du nœud spore et d'angle o (voir la figure 5.31).
  - Créer l'arête entre le nœud spore et le nœud créé.
- Répéter 29 fois :
  - Tirer aléatoirement un angle o dans $[-\theta_b : \theta_b]$
  - Ajouter un nœud à une distance l du dernier nœud et d'angle o (voir la figure 5.31).
  - Créer l'arête entre le nœud père et le nœud créé.
- $nb = 1$
- $nb_{sb} = 0$
- $nb_{eb} = 0$
- Pour t de 0 à 99 :
  - Calculer $\mathrm{L} = a * e^{b.(t+1)} - a * e^{b.(t)}$.
  - Tirer p_{sb} et p_b selon une loi uniforme entre 0 et 1.
    - Si $p_{sb} < \mathrm{P}_{sb}$:
      - $nb_{sb} = 1$
      - $nb = nb + nb_{sb}$
    - Si $p_b < \mathrm{P}_b$:
      - $nb_{eb} = 1$
      - $nb = nb + nb_{eb}$
  - Calculer $l = \mathrm{L}/nb$.
  - Si $nb_{sb} = 1$:
    - Tirer aléatoirement un angle o dans $[-\theta_{sb} : \theta_{sb}]$ selon une loi uniforme.
    - Ajouter un nœud à une distance l du nœud spore et d'angle o (voir la figure 5.31).
    - Créer l'arête entre le nœud spore (nœud extrémité) et le nœud créé.
  - Si $nb_{eb} = 1$:
    - Tirer aléatoirement un nœud Ne parmi les nœuds d'extrémité du graphe.
  - Pour chaque nœud d'extrémité :
    - Si le nœud est Ne :
      - k = 2
    - Sinon :
      - k = 1

— Répéter k fois :
  — Tirer un angle o dans $[-\theta_b : \theta_b]$
  — Ajouter un nœud à une distance l du nœud extrémité et d'angle o (voir la figure 5.31).
  — Créer l'arête entre le nœud père (nœud extrémité) et le nœud créé.

- Pour l'aire : Tirer les coefficients a et b de la fonction exponentielle selon la PDF estimée.

Le graphe obtenu au temps final est ensuite converti en squelette sur une image 2160*2160 (taille de nos images réelles). Enfin, le squelette est dilaté de façon uniforme jusqu'à l'obtention d'un objet d'aire la valeur estimée au temps T_{final}.

5.8.3 Exemples de simulation

La figure 5.32 illustre le graphe, le squelette ainsi que le masque binaire de l'objet correspondant obtenu avec notre modèle de croissance pour le phénotype 1 à la concentration 07. La figure 43 présentée dans le paragraphe A.8 en annexe, illustre l'évolution de ce graphe au cours du temps. Le résultat de la simulation de la croissance de neuf autres champignons est illustré sur la figure 5.33 avec les squelettes en **A** et les objets correspondants en **B**.

La figure 5.34 permet quant à elle de vérifier que les courbes moyennes des paramètres des champignons résultant de la simulation (**B**) suivent bien les mêmes dynamiques que celles issues des données réelles (**A**). Les moyennes sont calculées sur 20 croissances simulées. Bien que la longueur moyenne pour la concentration 09 soit légèrement surestimée et que celle pour la concentration 06 soit légèrement sous-estimée, les valeurs de longueurs moyennes pour les quatre concentrations évoluent de manière très similaire à celles issues de nos données réelles. Concernant les paramètres nombre de branches primaires et terminales, on observe que les dynamiques des simulations sont moins fidèles aux données que pour le paramètre longueur (voir la figure 5.34). On observe notamment que les dynamiques simulées sont globalement linéaires tandis que les dynamiques réelles du nombre de branches terminales montrent une légère accélération. Cette constatation n'est pas étonnante car le principe de création de branches mise en place dans notre modèle dépend uniquement du nombre de branches observées à la fin de la séquence temporelle. Nous pouvons également remarquer que le nombre de branches terminales est surestimé, d'autant plus pour les fortes concentrations (B06 et B07). À l'inverse, le nombre de branches primaires est sous-estimé dans le cas de la concentration faible 09.

Au vu des résultats de la simulation (squelettes) et de la figure 5.34, nous pouvons conclure que le modèle de croissance calibré sur les données réelles est capable de simuler des squelettes suffisamment réalistes de champignons traités avec la molécule B, aux concentrations où le phénotype 1 est observable (06, 07, 08 et 09). Des améliorations sont toutefois à étudier.

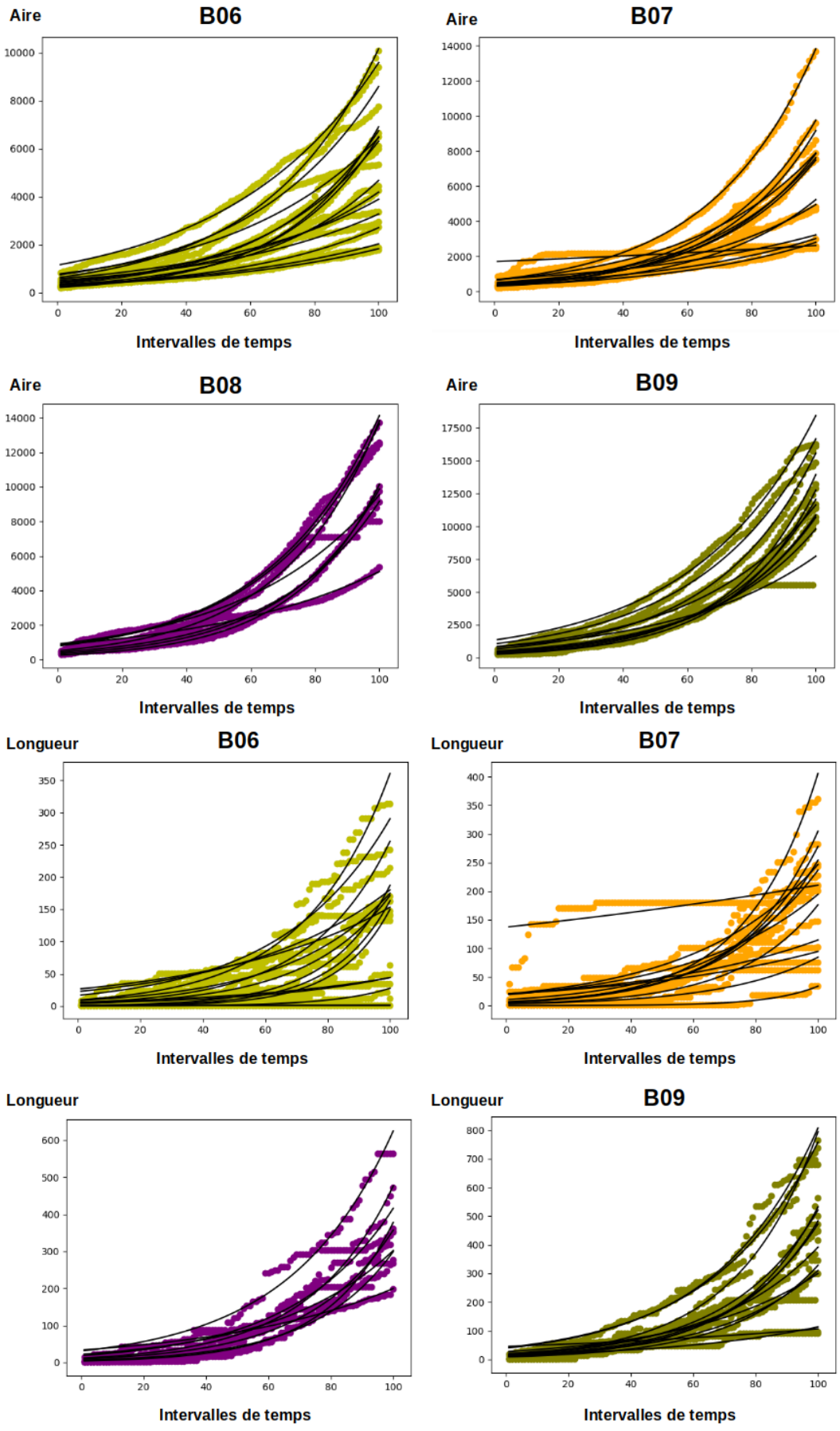


FIGURE 5.28 – Fonctions exponentielles finales (courbes noires) dont les coefficients sont estimés à partir des valeurs des paramètres phénotypiques (points en couleur).

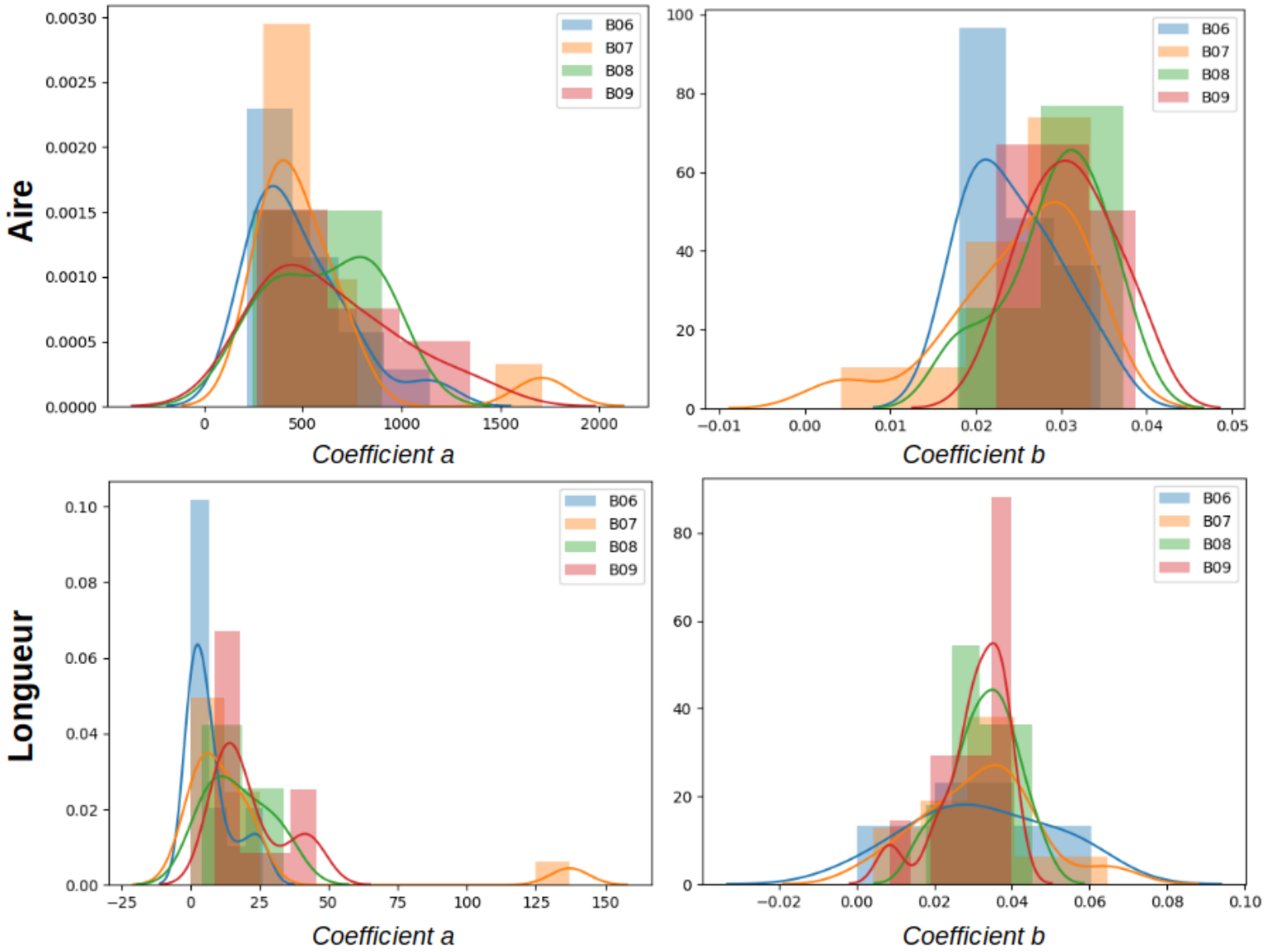


FIGURE 5.29 – Distributions des coefficients a et b des fonctions exponentielles ajustées sur l'évolution des paramètres aire et longueur pour les différentes concentrations.

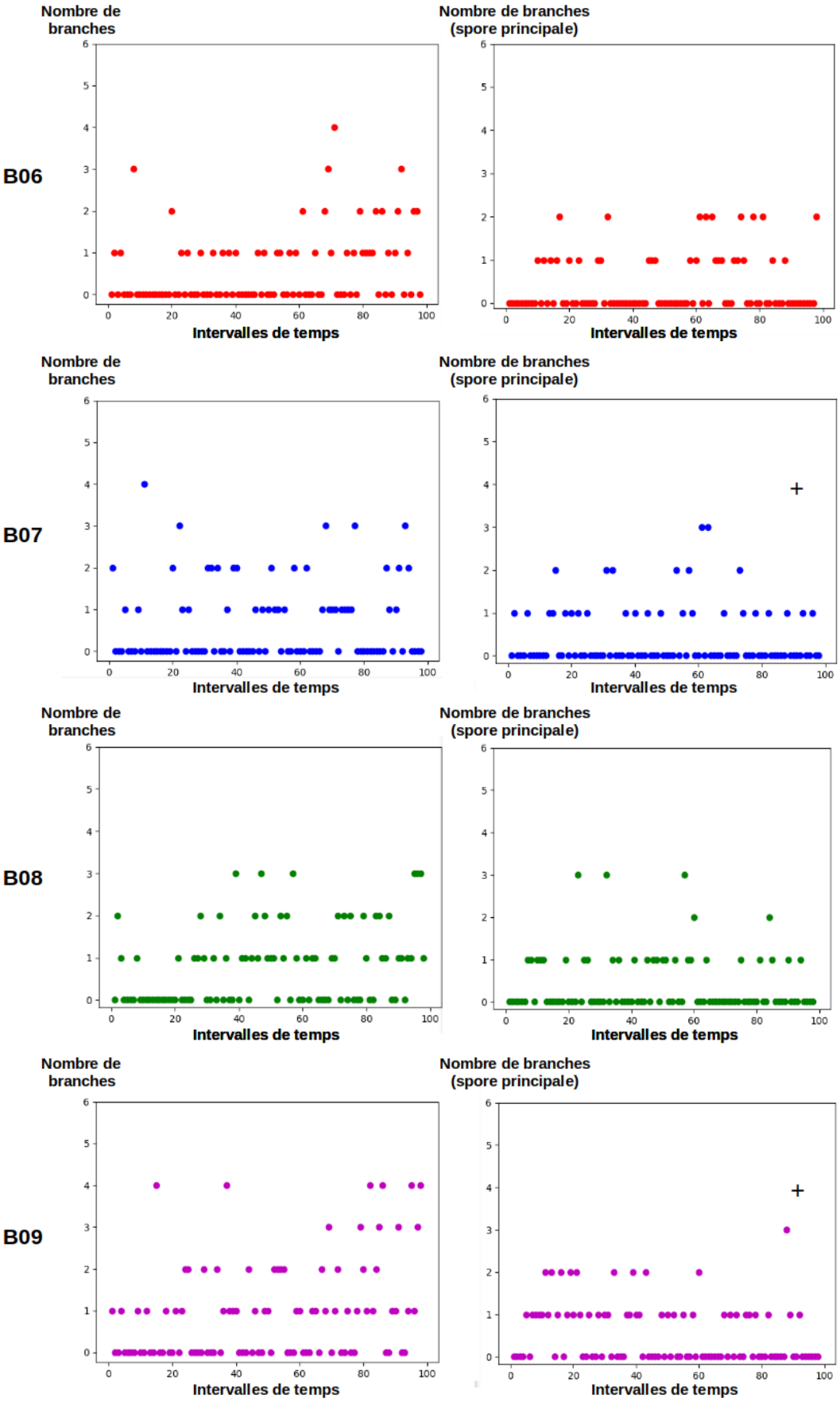


FIGURE 5.30 – Nombre de branches créées à chaque intervalle de temps en considérant l'ensemble des objets à chaque concentration.

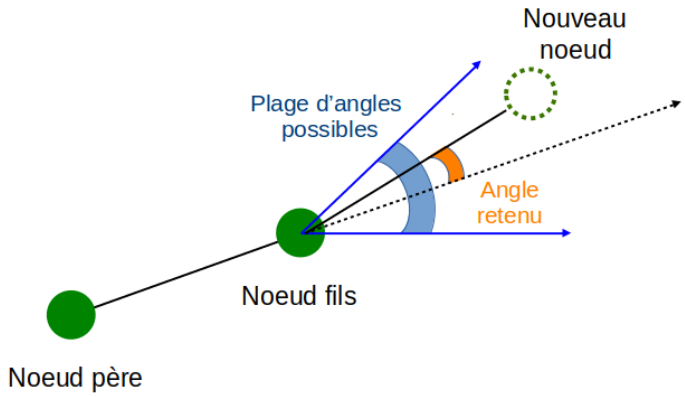


FIGURE 5.31 – Illustration de la recherche des coordonnées d'un nouveau nœud et d'une nouvelle branche en fonction de la longueur calculée, et de l'angle tiré aléatoirement dans un intervalle donné.

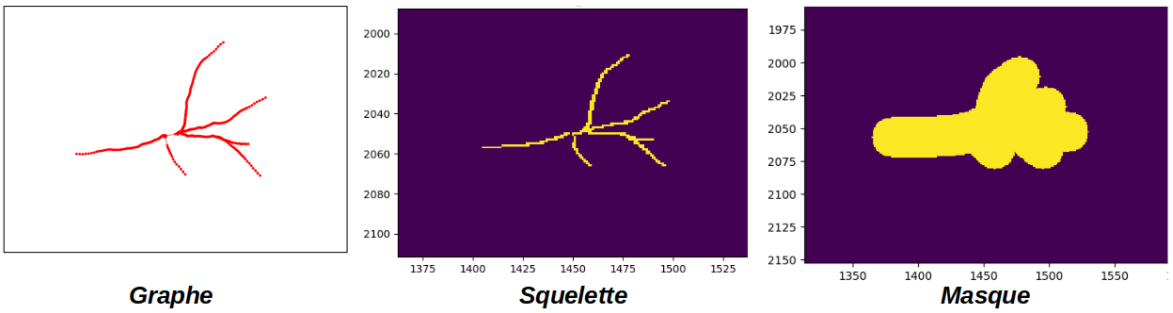


FIGURE 5.32 – Exemple de résultat de la simulation, au temps T_{final}, d'un champignon du phénotype 1 (molécule B à la concentration 07) *via* notre modèle de croissance : Le graphe, le squelette et l'objet correspondant.

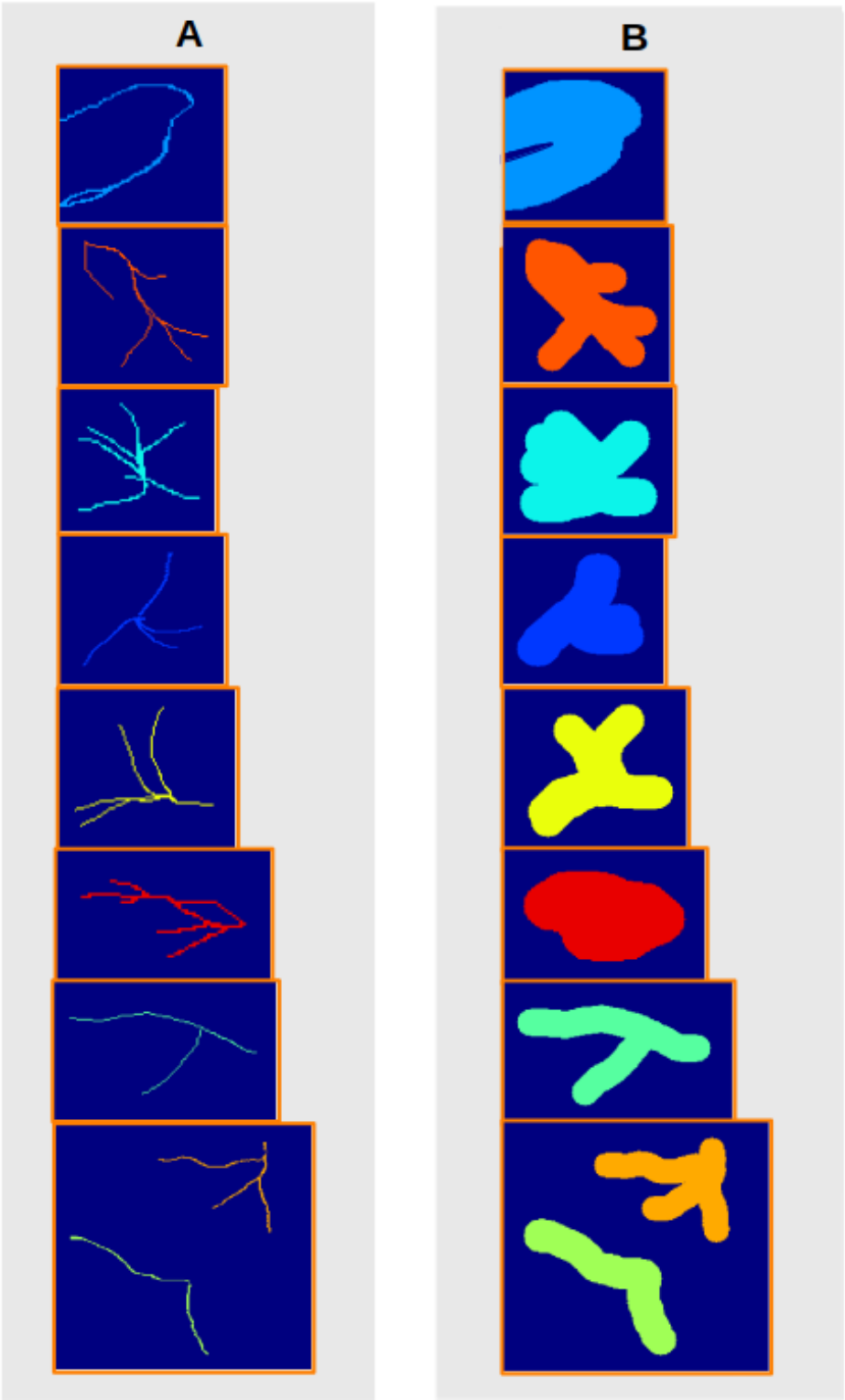


FIGURE 5.33 – Exemple de résultats, au temps 100, de la simulation *via* notre modèle de croissance de neuf champignons du phénotype 1 (molécule B à la concentration 07). **A :** les squelettes et **B :** les objets correspondants.

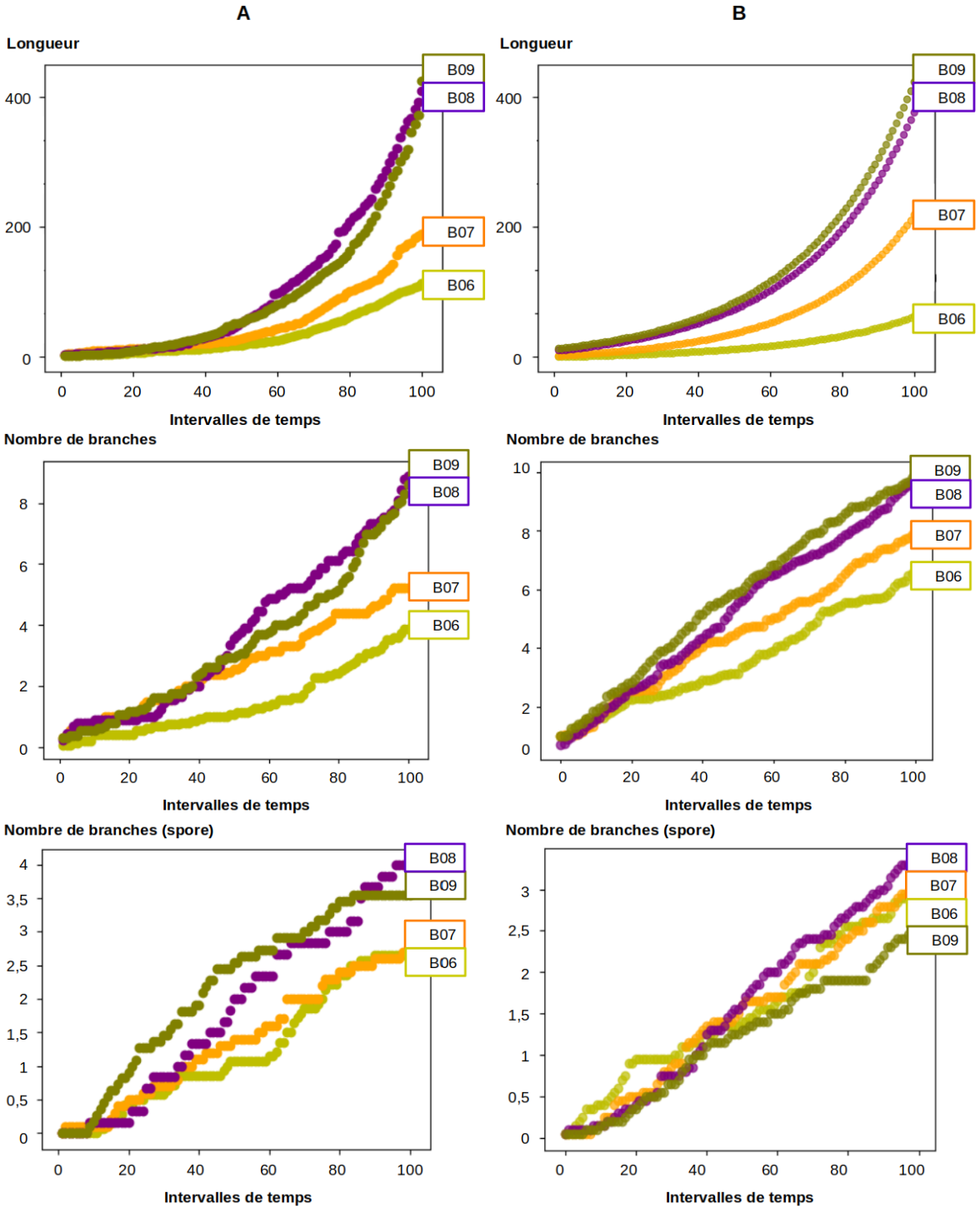


FIGURE 5.34 – Évolution des paramètres longueur, nombre de branches terminales et nombre de branches primaires (moyenne sur les champignons d'une même concentration) du phénotype 1 au cours du temps. B : molécule dont le MoA entraîne le phénotype 1. 06 à 09 : concentrations décroissantes en molécule. **A :** Données réelles et **B :** Données issues de la simulation (20 simulations par concentration).

5.9 Conclusion

Le travail effectué dans cette partie du projet a permis d'acquérir une meilleure compréhension de la croissance de *Botrytis cinerea* en présence d'une molécule antifongique en fonction de son mode d'action et de son niveau d'efficacité. Pour cela, un protocole expérimental particulier incluant une concentration spécifique en cellules de champignon a été mis en place et des séquences temporelles ont été générées. Les champignons dans les images sont segmentés puis des caractéristiques sont mesurées sur les objets détectés, à chaque temps. Nous avons étudié l'évolution des caractéristiques au cours du temps afin de mettre en place notre modèle de simulation de la croissance du champignon. En effet, ces caractéristiques sont utilisées pour calibrer les paramètres du modèle. Le modèle construit est un processus stochastique à temps discret utilisant des lois discrètes et continues pour piloter les différents événements (croissance, création d'une branche...) et leur ampleur. Nous avons également créé des profils relatifs au type de composé antifongique et à sa concentration en utilisant l'évolution des paramètres aire et longueur au court du temps. Ces profils apportent une possibilité supplémentaire de classer le phénotype des champignons d'une image et d'évaluer l'efficacité de la molécule dans le temps (comparaison des différents profils).

5.10 Références

«MS Windows NT l-systèmes», https://cs108.epfl.ch/archive/19/e/LSYS/LSYS.html. 139

AONO, M. et T. L. KUNII. 1984, «Botanical tree image generation», *IEEE computer graphics and applications*, vol. 4, nº 5, p. 10–34. 139

BUCKSCH, A., A. ATTA-BOATENG, A. F. AZIHOU, D. BATTOGTOKH, A. BAUMGARTNER, B. M. BINDER, S. A. BRAYBROOK, C. CHANG, V. CONEVA, T. J. DEWITT et collab.. 2017, «Morphological plant modeling : unleashing geometric and topological potential within the plant sciences», *Frontiers in plant science*, vol. 8, p. 900. 140

CONN, A., U. V. PEDMALE, J. CHORY et S. NAVLAKHA. 2017, «High-resolution laser scanning reveals plant architectures that reflect universal network design principles», *Cell systems*, vol. 5, nº 1, p. 53–62. 140

DE REFFYE, P. et F. BLAISE. 1993, «Modélisation de l'architecture des arbres. applications forestières et paysagères», *Revue Forestière Française*. 138

DE REFFYE, P., C. EDELIN, J. FRANÇON, M. JAEGER et C. PUECH. 1988, «Plant models faithful to botanical structure and development», *ACM Siggraph Computer Graphics*, vol. 22, nº 4, p. 151–158. 139

GODIN, C. et Y. CARAGLIO. 1998, «A multiscale model of plant topological structures», *Journal of theoretical biology*, vol. 191, nº 1, p. 1–46. xvii, 140

GODIN, C., Y. GUÉDON et E. COSTES. 1999, «Exploration of a plant architecture database with the amapmod software illustrated on an apple tree hybrid family», *Agronomie*, vol. 19, nº 3-4, p. 163–184. xviii, 140, 141

GODIN, C., Y. GUÉDON, E. COSTES et Y. CARAGLIO. 1997, «Measuring and analysing plants with the amapmod software», . 140

HEMMERLING, R., O. KNIEMEYER, D. LANWERT, W. KURTH et G. BUCK-SORLIN. 2008, «The rule-based language xl and the modelling environment groimp illustrated with simulated tree competition», *Functional plant biology*, vol. 35, nº 10, p. 739–750. 139

LECOUSTRE, R. et P. DE REFFYE. 1993, «Amap : un modeleur de végétaux, un ensemble de logiciels de cao/dao à l'usage des professionnels de l'aménagement et des paysagistes», . 139

LINDENMAYER, A. 1968, «Mathematical models for cellular interactions in development i. filaments with one-sided inputs», *Journal of theoretical biology*, vol. 18, nº 3, p. 280–299. 139

MICHALEWICZ, M. T. 1997, *Advances in computational life sciences*, CSIRO Publishing. 140

PRUSINKIEWICZ, P. 1986, «Graphical applications of l-systems», dans *Proceedings of graphics interface*, vol. 86, p. 247–253. 139

PRUSINKIEWICZ, P. 1998, «Modeling of spatial structure and development of plants : a review», *Scientia Horticulturae*, vol. 74, nº 1-2, p. 113–149. 139

PRUSINKIEWICZ, P. 2002, «Art and science of life : designing and growing virtual plants with l-systems», dans *XXVI International Horticultural Congress : Nursery Crops ; Development, Evaluation, Production and Use 630*, p. 15–28. 139

PRUSINKIEWICZ, P. et A. LINDENMAYER. 1990, «Graphical modeling using l-systems», dans *The Algorithmic Beauty of Plants*, Springer, p. 1–50. 139

PRUSINKIEWICZ, P. et A. LINDENMAYER. 2012, *The algorithmic beauty of plants*, Springer Science & Business Media. 140

Conclusion générale

Le but de ce projet est le développement d'un outil de reconnaissance des phénotypes connus et inconnus de *Botrytis c.* sur des images en microcopie ainsi que la modélisation de sa croissance au cours du temps.
Nous avons dans un premier temps choisi une méthode adaptée à la classification de nos images. Pour cela, nous avons travaillé dans un cadre supervisé en comparant les résultats obtenus par deux méthodes de classification : les forêt aléatoires (RF) et les réseaux de neurones convolutifs (CNN). La méthode des forêts aléatoires a nécessité la mise en place d'une analyse automatique des images de microscopie. Elle comprend des étapes de traitement d'images et d'extraction de paramètres morphométriques. Pour cela, les objets sont segmentés puis squelettisés et enfin, convertis en graphe. Des paramètres morphométriques extraits de ces trois formes sont utilisés comme caractéristiques. Quatre types de résultats ont été générés et comparés : les résultats provenant des classifications directes ou en cascade et les classifications dont l'unité à prédire est soit l'objet soit l'image. Les résultats obtenus nous ont permis de conclure que les deux méthodes (RF et CNN) sont capables de reconnaître les sept classes d'images, mais avec des niveaux de précision différents (81.4% pour la méthode des forêts aléatoires contre 94.5% pour les réseaux de neurones).
Notre second objectif était de pouvoir alerter l'utilisateur lors de l'identification d'un nouveau phénotype. Pour cela, nous avons proposé une nouvelle méthode de classification avec classe de rejet qui suit une stratégie générale fondée sur trois étapes principales : apprentissage d'un modèle indépendamment pour chaque classe, apprentissage d'un seuil fondé sur les interactions entre classes et procédure de prédiction. Dans l'application étudiée, nous avons choisi d'utiliser comme représentation de nos images deux types de caractéristiques provenant d'un réseau de neurones (les données Bottlenecks et Output). Les modèles sont des mélanges de gaussiennes (GMM) optimisés avec une matrice de covariance sphérique. Les seuils associés aux modèles peuvent être estimés à l'aide d'une régression logistique ou de l'écart entre le nombre de faux positifs et de faux négatifs, dans un contexte "1-vs-all" ou "1-vs-1". Les meilleurs résultats sont obtenus en utilisant l'écart entre le nombre de faux positifs et de faux négatifs avec l'option "1-vs-all" pour l'apprentissage des seuils, sur les données Bottlenecks dont la dimension est réduite à 100 caractéristiques avec une réduction de dimension selon l'importance des variables. La méthode proposée a fourni de très bonnes précisions de classification (91% en moyenne) pour la classe de rejet et les classes connues des phénotypes de *Botrytis cinerea*.
L'objectif final est l'identification des modes d'action (MoAs) connus et non-connus des molécules antifongiques. Pour ce faire, notre approche correspond à une classification des images en deux étapes se fondant sur deux étapes de prédiction : par un réseau de neurones d'abord, puis par la méthode de classification avec rejet proposée. Enfin, des règles de prédiction sont appliquées aux résultats de classification des images aux différentes concentrations afin de proposer un MoA des molécules testées. Ce protocole d'élucidation des MoAs mis en place nous a permis d'obtenir de très bons scores de classifica-

tion (100% sur les MoAs connus et 95% sur les nouveaux MoAs).
Au cours des travaux de cette thèse, une approche de classification alternative, utilisant le transport optimal (TO) est testée sur des données synthétiques d'une part et sur les données de notre étude sur *Botrytis c.* (réduites à la dimension 2 par ACP) d'autre part. Cette approche novatrice est fondée sur l'estimation d'une fonction de transport pour mettre en correspondance la distribution des données acquises avec une distribution connue prédéfinie. L'avantage de cette approche est que la complexité du modèle (GMM) est "transférée" dans le transport vers un modèle simple (une gaussienne par exemple). Les résultats obtenus nous ont permis de conclure que l'approche par transport optimal présentent une efficacité de classification équivalente ou légèrement supérieure à celle de la méthode GMM.
Le dernier objectif de mes travaux a été de caractériser et modéliser la croissance des différents phénotypes au cours du temps. Pour cela, les champignons dans les images de séquences temporelles sont segmentés puis les caractéristiques mesurées sur les objets détectés, à chaque temps, sont utilisées pour calibrer les paramètres du modèle. Nous avons dans un premier temps étudié l'évolution des caractéristiques au cours du temps, mis en place notre modèle mais également créé des profils relatifs au type de composés antifongiques et à sa concentration. Ces profils apportent une possibilité supplémentaire de classer le phénotype des champignons d'une image et d'évaluer l'efficacité de la molécule dans le temps.

Perspectives

Comme perspectives de ces travaux, nous proposons les idées suivantes :

Les données : Dans le cadre de notre objectif de modélisation de la croissance du champignon *Botrytis cinerea* au cours du temps, nous avons généré deux TimeLapses. Or, ces deux TimeLapses diffèrent de par la concentration en spores définie lors de l'expérimentation biologique ($0,5.10^3$ spores/mL et $0,5.10^4$ spores/ml). Ces concentrations présentent toutes deux un avantage et un inconvénient. En effet, à une forte concentration est associé un nombre conséquent d'échantillons et par conséquent la possibilité d'analyser moins d'images pour une analyse statistique. Cependant, une limite à cette forte concentration est qu'elle favorise le chevauchement des cellules. Le chevauchement complexifie voire rend impossible la séparation des objets et biaise les valeurs des paramètres extraits. Au contraire, une concentration faible en spore permet de réduire ce risque de chevauchement mais implique en revanche la nécessité de générer et analyser plus de séquences temporelles afin de disposer de suffisamment d'échantillons.
Ainsi, nous proposons comme perspective de générer de nouvelles séquences temporelles avec une concentration en spores fixe, dans l'idéal $0,5.10^3$ spores/ml et en nombre suffisant pour une analyse statistique plus robuste. D'autre part, il est nécessaire de générer des réplicats des molécules testées afin d'attester de la pertinence des résultats obtenus, ainsi que d'accroître le nombre d'échantillons.

La caractérisation des phénotypes : Considérer les champignons comme des objets dont on peut calculer le squelette caractérisé par leur topologie est une solution à notre problématique mais pas la seule. Une autre manière de procéder est de considérer chaque objet comme un ensemble de plusieurs formes prédéfinies telles que des ellipses ou des bâtonnets. Un premier essai a été effectué avec une méthode de détection d'objets avec contraintes de forme (processus ponctuel marqué Eldin et collab. [2012]), comme évoqué dans le paragraphe 2.5.4. Ainsi, détecter un objet dans une image revient à trouver une position p associée à un ensemble de paramètres m dont la mesure de qualité est supérieure à un certain seuil De Graeve et collab. [2019]. Cet ensemble de paramètres peut être utilisé comme caractéristiques dans une méthode de classification et injecté dans un modèle de simulation de la croissance des différentes formes de champignon.
A cette fin, nous avons encadré une étudiante en dernière année d'école d'ingénieur, Laguili Oumaima pour son stage de fin d'étude. Ces travaux ont pour but le développement d'un réseau de neurone permettant la détection de ces différentes formes sur des images dont celles de *Botrytis cinerea*.

La classe de rejet : Dans le chapitre 3 de ce manuscrit nous décrivons la méthode de classification avec classe de rejet que nous proposons. Cette approche est fondée sur l'apprentissage de modèles et de seuils dans un contexte supervisé. Bien que l'apprentissage

des modèles se fasse indépendamment pour chaque classe, l'apprentissage des seuils en revanche est fondé sur les interactions entre les classes. Le calcul du seuil d'une classe dépend donc de la répartition des autres classes. On peut noter que dans les régions de l'espace où la classe n'est pas "en interaction" avec d'autres classes, sa frontière (définie par le seuil sur le modèle) est tout de même fixée par les contraintes qui existent dans les régions où il y a effectivement interaction avec d'autres classes. Ce phénomène peut par exemple "empêcher" la frontière de s'étirer librement dans ces régions sans interactions alors que la densité d'échantillons n'y est pas négligeable.
La méthode actuelle pourrait être améliorée en rendant le seuil local de sorte à s'adapter à une intensité d'interaction (à définir en tout point de l'espace) avec les autres classes. Dans les régions à forte interaction, le seuil serait proche du seuil global obtenu par la méthode actuelle. Là où les interactions sont faibles, le seuil pourrait tendre vers une valeur ne dépendant que des échantillons de la classe et de son modèle, par exemple permettant d'englober une proportion donnée d'échantillons. L'intensité d'interaction en un point pourrait être une fonction décroissante de la distance minimale aux autres classes, distance qui s'exprimerait en fonction des échantillons des autres classes ou de leur modèle.

L'approche par transport optimal : Dans le chapitre 4 une méthode de classification utilisant le transport optimal (TO) qui permet l'estimation de la fonction de densité de probabilité de la distribution d'une population est décrite et testée. Les PDFs des échantillons de chaque classe sont estimées par interpolation des valeurs obtenues via leur transport vers la distribution cible gaussienne. Cette étape d'interpolation est essentielle à la méthode. Néanmoins d'autre approches permettant l'estimation des PDFs dans l'espace des données existent.
Ainsi, nous pensions reprendre une approche non-paramétrique telle que l'estimateur de densité par noyau. L'article PEHERSTORFER et collab. [2014] décrit l'utilisation d'un noyau par échantillon d'une grille afin d'estimer leurs PDFs. Pour faire cela, les poids des noyaux sont optimisés de sorte à minimiser l'erreur d'estimation de PDF sur les points d'apprentissage dont les valeurs de PDF sont définies par le transport.

Les profils des molécules antifongiques : Dans le dernier chapitre de cette thèse (chapitre 5), des cartes de trajectoires sont construites en prenant en compte l'évolution de la valeur de deux paramètres morphométriques (l'aire et la longueur des champignons) en fonction du temps. Ces cartes sont considérées comme les profils des molécules testées aux concentrations correspondantes. L'allure de ces profils apporte une possibilité supplémentaire de déterminer le phénotype des champignons d'une image ainsi que d'évaluer l'efficacité du composé dans le temps (propriétés cinétiques du composé antifongique).
Une perspective d'amélioration de ces cartes serait de prendre en considération l'évolution au cours du temps de plusieurs autres paramètres (courbes en 3D ou plus) afin d'obtenir des profils plus marqués, plus spécifiques des différents phénotypes. En revanche, l'analyse visuelle devient plus délicate voire inappropriée des lors que l'on dépasse deux dimensions (nécessité de réduire la dimension à deux ou trois).
Enfin, il est possible d'utiliser des méthodes à noyau telles que les SVMs (avec un noyau spécifique aux courbes) ou les LSTM KARIM et collab. [2017] afin d'effectuer une classification de courbes (fondée sur l'information dynamique).

Modèle de simulation : À ce jour, le modèle obtient de très bons résultats de simulation de la croissance de champignon du phénotype 1. Néanmoins, lors de la création d'une nouvelle arête, l'angle relatif est tiré aléatoirement dans un intervalle donné. Cet intervalle est fixé expérimentalement.
Une amélioration possible du modèle serait d'estimer sur les données réelles la distribution de ces angles pour chaque phénotype (et possiblement à chaque concentration si cela s'avérait être un facteur déterminant).

6.11 Références

DE GRAEVE, F., E. DEBREUVE, S. RAHMOUN, S. ECSEDI, A. BAHRI, A. HUBSTENBERGER, X. DESCOMBES et F. BESSE. 2019, «Detecting and quantifying stress granules in tissues of multicellular organisms with the obj. mpp analysis tool», *Traffic*, vol. 20, nº 9, p. 697–711. 177

ELDIN, A. G., X. DESCOMBES, G. CHARPIAT et J. ZERUBIA. 2012, «Multiple birth and cut algorithm for multiple object detection», *Journal of Multimedia Processing and Technologies*. 177

KARIM, F., S. MAJUMDAR, H. DARABI et S. CHEN. 2017, «Lstm fully convolutional networks for time series classification», *IEEE access*, vol. 6, p. 1662–1669. 178

PEHERSTORFER, B., D. PFLÜGE et H.-J. BUNGARTZ. 2014, «Density estimation with adaptive sparse grids for large data sets», dans *Proceedings of the 2014 SIAM international conference on data mining*, SIAM, p. 443–451. 178

Annexes

Annexes

A.1 Les grandes étapes de R&D au sein de Bayer CropScience

Ce paragraphe est issu du rapport de stage de Master 2 LAROUI (2017).

Le site de la Dargoire est organisé en différents départements :

— Le département Excellence de la Chimie dans lequel les chercheurs chimistes conçoivent et synthétisent des familles de molécules en s'appuyant sur une connaissance approfondie de l'agrochimie.

— Le département Excellence Biologie et Agro-Kinetic : Une fois synthétisées, les molécules font l'objet d'évaluations biologiques : leurs effets sont ainsi testés in vitro sur les champignons parasites seuls ainsi qu'in vivo en présence de la plante hôte.

— Le département Excellence Biochimie, dont le rôle est de découvrir le mode d'action d'une molécule, c'est-à-dire identifier la façon dont les composés actifs bloquent les activités cellulaires du champignon.

— Le département Analyse de Résidus (ou Human Safety) réalisant des études complémentaires de quantification des résidus de molécule fongicide dans les fruits et légumes après traitement afin de d'assurer que l'exposition potentielle des consommateurs à la molécule est inférieure aux seuils toxicologiques. Ces analyses de résidus combinées à des essais en plein champ sont nécessaires à la préparation des dossiers d'homologation de nouvelles substances actives.

A.2 Cycle asexué chez *Botrytis cinerea*

Ce paragraphe est issu du rapport de stage de Master 2 LAROUI (2017).

Botrytis cinerea se développe selon une croissance polarisée. Il passe d'un état de conidies (spores) à la forme filamenteuse appelée mycélium DUBOS [2002].

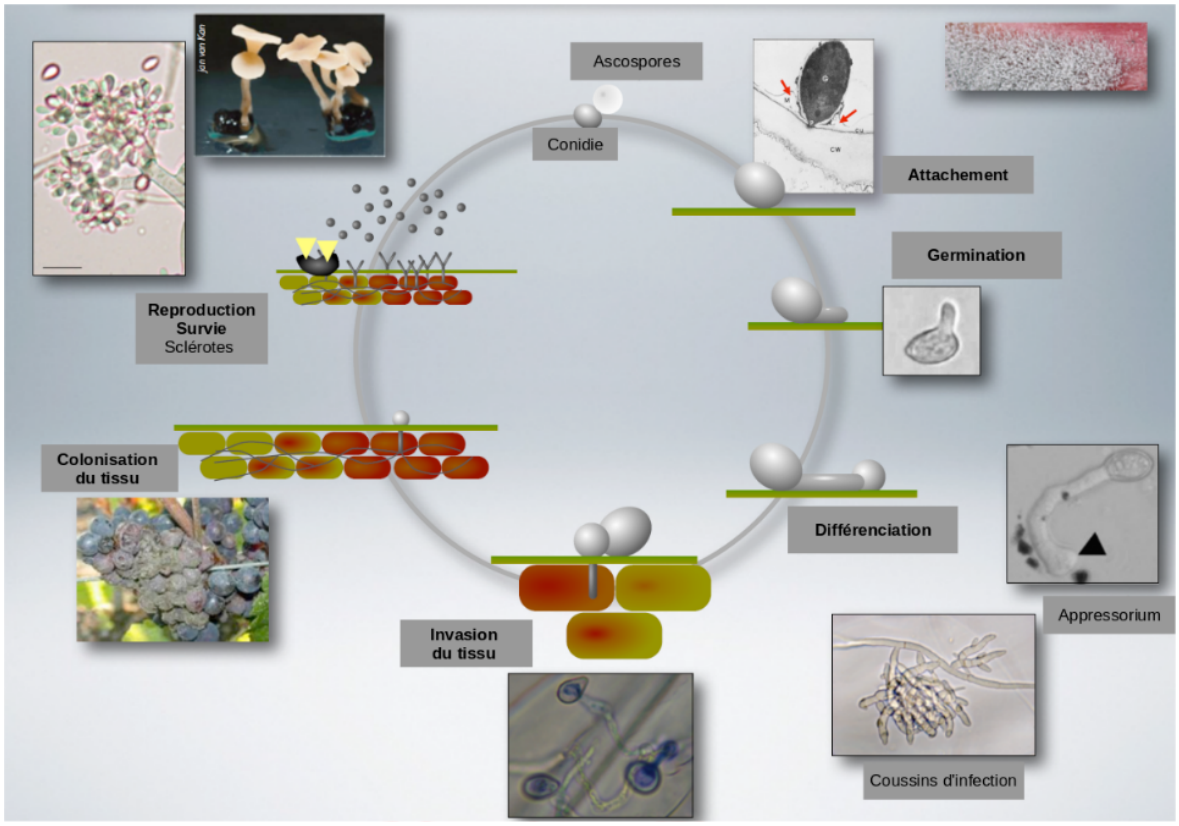
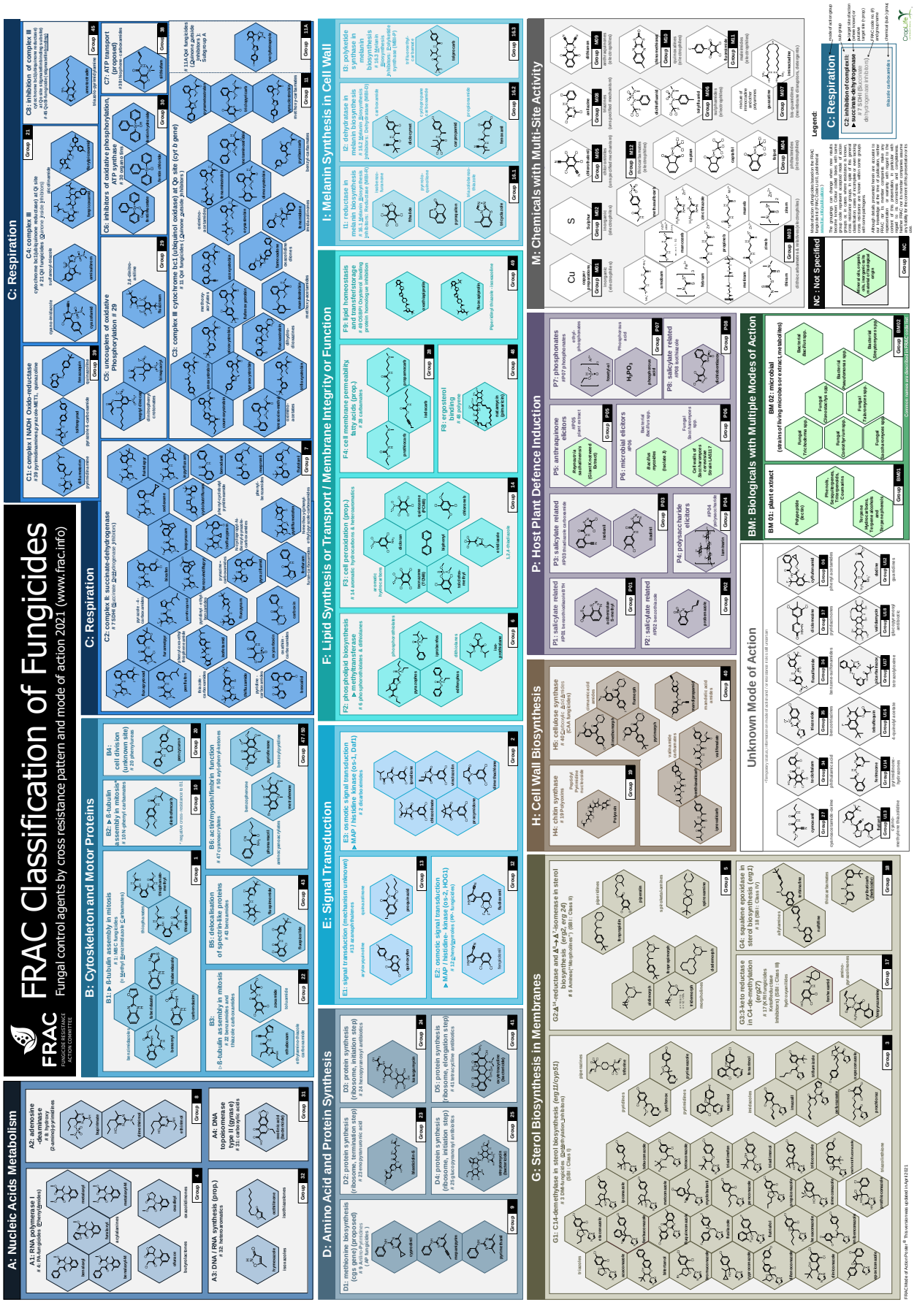


FIGURE 35 – Cycle asexué de Botrytis : Les sclérotes germent et pénètrent dans le tissu végétal de l'hôte quand les conditions de température et d'humidité sont favorables. Le développement aboutit à la formation d'un mycélium portant des conidiophores ; hyphe aérienne portant une ou plusieurs conidies (Ascospore). Après la sporulation, la spore va adhérer à la surface de la plante et va germer. Cela mène à la formation d'un appressorium, qui en exerçant une pression mécanique (pression de turgescence) va passer la paroi pecto-cellulosique et s'accoler à la membrane plasmique de la cellule hôte.

Le mycélium est la forme filamenteuse du champignon. Il est constitué d'hyphes cloisonnées et utilise des armes chimiques dans le but de détruire et de se nourrir de ses hôtes. Lorsque les conditions deviennent défavorables à son développement, de petites pustules noires apparaissent. Elles sont appelées sclérotes et sont formées d'un cortex rigide renfermant une masse mycélienne dense. Lorsque les conditions de température et d'humidité redeviennent favorables, les sclérotes germent et pénètrent dans le tissu végétal. Cela donne un mycélium qui produit des conidiophores (feutrage grisâtre) qui portent des macroconidies (conidies de grande taille ou spores) MARTINEZ et collab. [2003]. Les macroconidies sont dispersables par le vent et la pluie et sont la source principale d'infection. Elles assurent la formation de nouvelles colonies.

A.3 Classification des Mode d'Action de molécules antifongiques selon le FRAC (2021). Figure extraite de FRA.



A.4 Concentrations

Chaque molécule est testée à dix concentrations, de la colonne 3 à 12 sur la microplaque (les deux premières colonnes étant les contrôles DMSO). Le tableau 4 indique la quantité en molécule (en µM) par puits.

Puits	Concentrations
3	100
4	33
5	11
6	4
7	1
8	0.4
9	0.1
10	0.05
11	0.02
12	0.01

TABLEAU 4 – Concentrations par puits (en µM).

A.5 Cristaux

Les molécules présentant des problèmes de solubilité dans le solvant utilisé entraînent la formation de cristaux caractéristiques reconnaissables à l'image. Ces agrégats peuvent prendre plusieurs formes, la figure 36 illustre certains de ces cas.

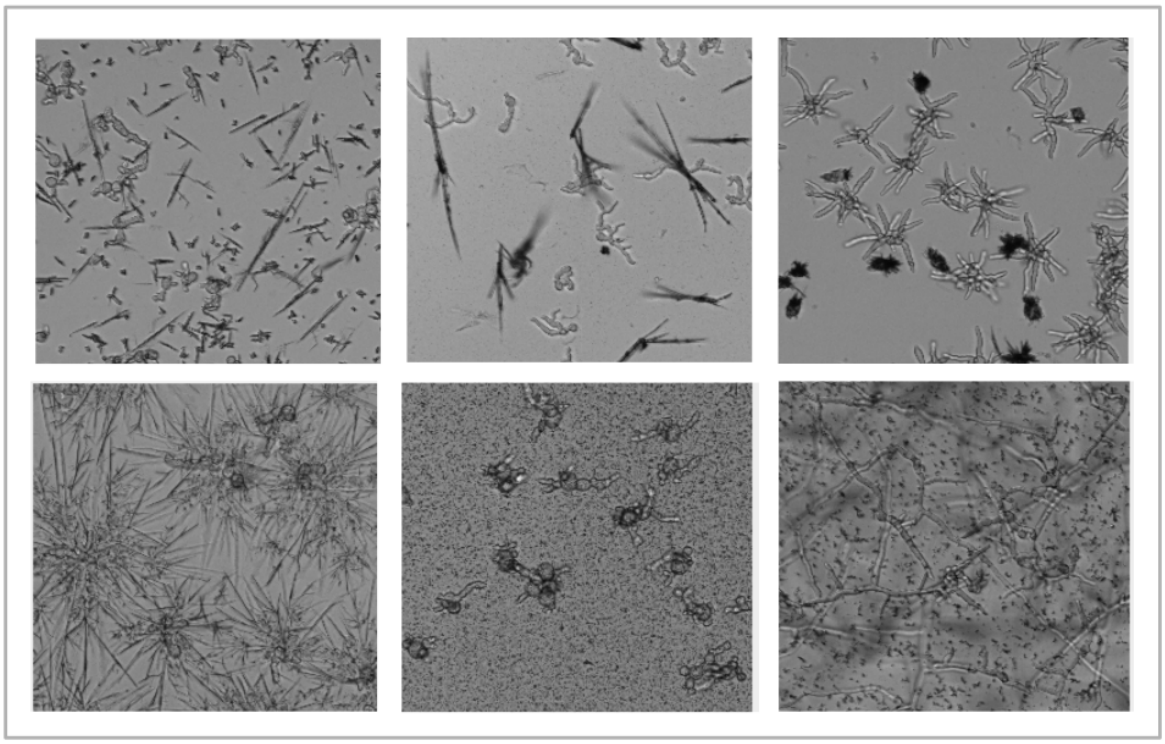


FIGURE 36 – Six exemples de cristaux obtenus en conséquence des problèmes de solubilité de certaines molécules dans le solvant utilisé.

A.6 Approche de classification par transport optimal : Couplage et "mappage" simultanés

A.6.1 Principe de la méthode

La méthode MappingTransport a été développée dans le cadre du problème d'adaptation de domaine COURTY et collab. [2016] utilisant l'algorithme de Sinkhorn CUTURI [2013] pour le résoudre. Cette approche a été la première approche envisagée pour notre problème de classification du fait d'une estimation simultanée des étapes de couplage et de mappage. En effet, l'idée générale est d'apprendre une fonction de transport et de pouvoir appliquer le transport appris à d'autres échantillons que ceux de l'ensemble d'apprentissage. Outre la capacité de travailler en grande dimension, un autre avantage certain de cette méthode est la possibilité d'apprendre un transport multi-labels, c'est-à-dire contraint par des étiquettes associées aux échantillons.

La méthode proposée selon cette approche par transport optimal est décrite ci-après :

- Première étape : Apprentissage d'un transport optimal par classe
  1. Choisir et générer une distribution cible X_t qui suit une loi normale de même moyenne que la population X_s dans l'espace source et dont la matrice de covariance est obtenue par normalisation de celle des données de façon à avoir un déterminant de 1 (voir la figure 37).
  2. Apprendre le transport optimal entre les échantillons de l'espace source X_s et ceux de l'espace cible X_t (voir la figure 37).

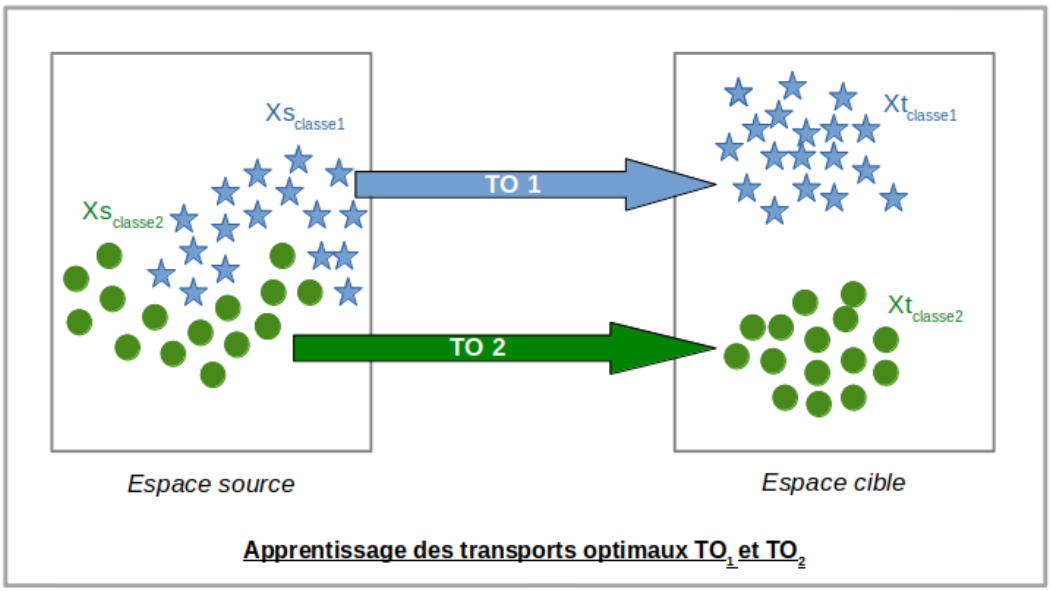


FIGURE 37 – Illustration de l'apprentissage des transports des échantillons de chaque classe de l'espace source vers les échantillons générés dans l'espace cible selon la loi correspondante.

- Deuxième étape : Apprentissage d'un seuil par classe (voir la figure 39)
  1. Transporter les échantillons de la classe dite de référence, et de toutes les autres classe de l'espace source dans l'espace cible (voir la figure 38).
  2. Récupérer les valeurs de la PDF cible aux échantillons transportés. En principe, nous sommes censés obtenir des valeurs hautes pour les échantillons appartenant à la classe de référence, et de faibles valeurs pour les autres.

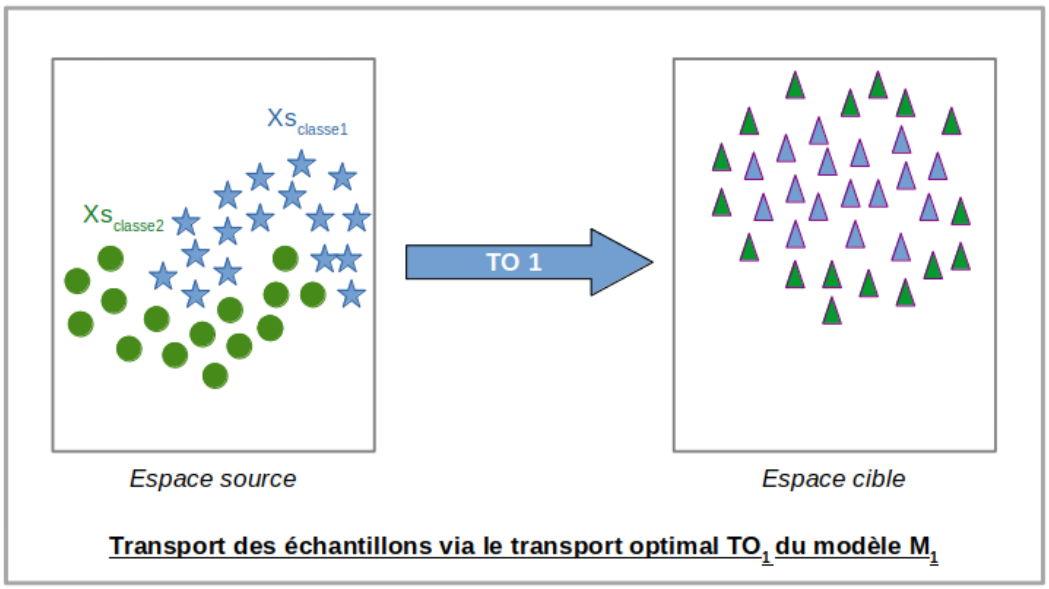


FIGURE 38 – Illustration du transport de l'ensemble des échantillons de l'ensemble d'apprentissage via le transport optimal appris T1 (Seul le modèle de la classe 1 est ici considéré).

3. Calculer le seuil en fonction des valeurs de PDF. Comme expliqué plus haut, nous analysons comment la PDF répond à «son» jeu d'échantillons mais également à ceux des autres classes. Pour définir ce seuil, nous pouvons utiliser les options "1-vs-all" et "1-vs-1" ainsi que les deux méthodes de calcul que sont la régression logistique et l'écart entre le nombre de faux positifs et faux négatifs. (ces deux méthodes sont décrites dans le paragraphe 3.2 du chapitre 3).

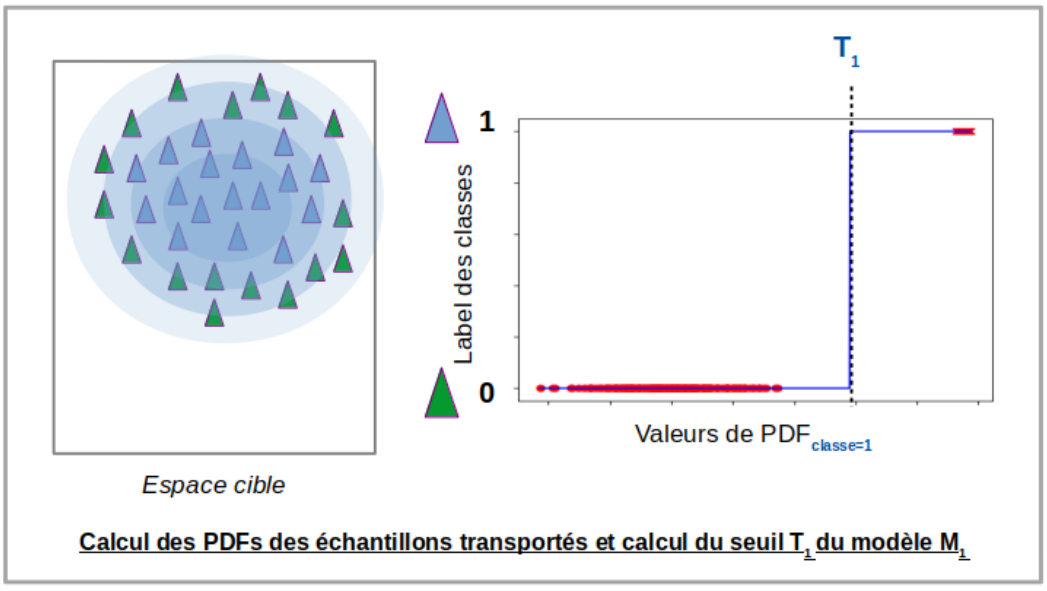


FIGURE 39 – Illustration de l'apprentissage du seuil sur des données de synthèse (Seul le modèle de la classe 1 est ici concidéré). Les échantillons de la classe 1 reçoivent l'étiquette 1. Les autres (soit les échantillons de toutes les autres classes en "1-vs-all", soit ceux des autres classes à tour de rôle en "1-vs-1") reçoivent l'étiquette 0.

Une fois le modèle et le seuil appris pour chaque classe de l'ensemble d'apprentissage, il est possible d'émettre une prédiction de la classe de nouveaux échantillons (non vus lors

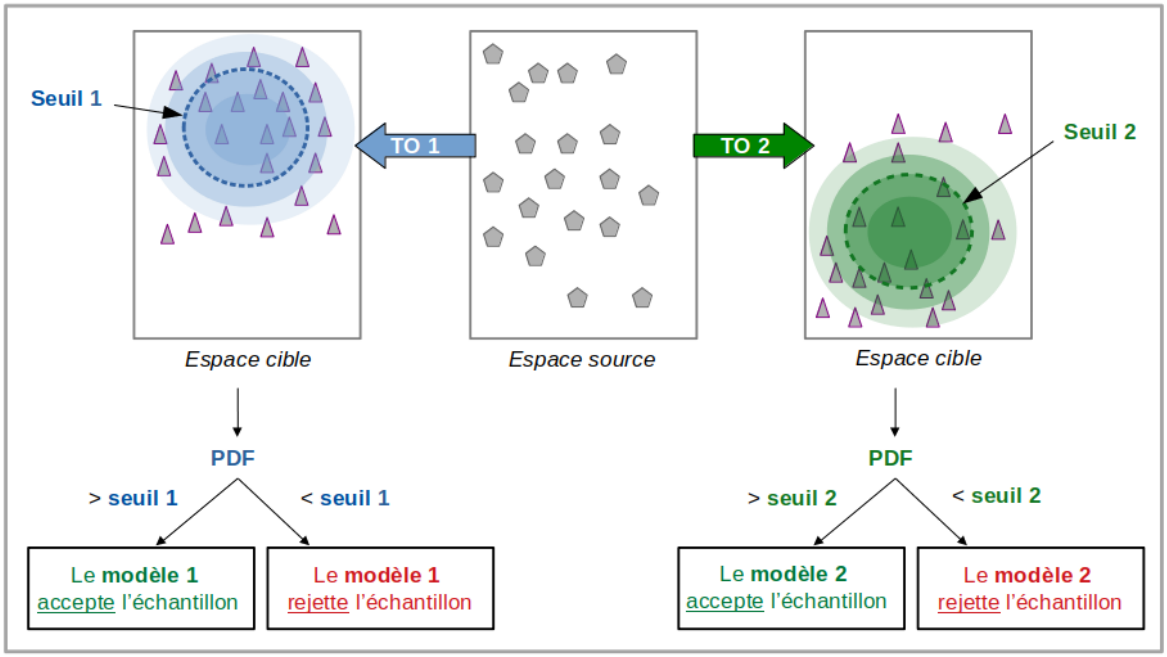


FIGURE 40 – Illustration du protocole de prédiction dans le cas de deux classes connues (1 et 2) pour lesquels le TO et le seuil sont optimisés lors de la phase d'apprentissage.

de la phase d'apprentissage).

La procédure est la suivante : pour chaque classe,

1. Transporter les échantillons à prédire dans l'espace cible avec le TO appris sur l'ensemble d'apprentissage.
2. Récupérer les valeurs de PDF des échantillons transportés en fonction de leur position dans l'espace cible.
3. Comparer les valeurs au seuil défini pour la classe afin d'obtenir la réponse du modèle aux échantillons (voir la figure 40).

Une fois la réponse de chaque modèle obtenue, on se réfère aux règles de prédiction (voir le paragraphe 3.2.1.

Dans certains cas, comme lorsque le support de la distribution des données est non convexe, estimer un TO linéaire n'est pas adapté. Pour cela, un TO non linéaire associé à une fonction noyau, typiquement gaussienne, peut être considéré PERROT et collab. [2016].

Néanmoins, nous nous sommes rendu compte que dans le cas du transport avec noyau gaussien, le transport d'échantillons trop éloignés de l'ensemble d'apprentissage se déroulait mal, les envoyant tous à l'origine.

Nous avons donc modifié notre approche dans le cas du transport gaussien en prenant en compte cette limite. D'une part, un seuil par classe est défini sur la distance de l'échantillon à la moyenne de la classe correspondante. Ce seuil permet la définition d'un espace de validité du transport. Si un échantillon est au-delà de ce seuil, il est directement rejeté par la classe sans être transporté. D'autre part, dans le but d'aider le transport des échantillons autour de la population de la classe considérée, le transport est appris avec une population synthétique additionnelle qui "borde" la classe. Le transport devient un transport multi-label car appris entre les échantillons de la classe et ceux de synthèse

dans l'espace source d'une part, et ceux d'une distribution gaussienne et d'une couronne d'échantillons (de même moyenne).

A.6.2 Paramètres à optimiser

— Le type de transport (linéaire ou à noyau)

— Le protocole de génération des distributions dans l'espace cible pour l'apprentissage du transport (moyennes et matrices de covariance)

— Le poids du terme de régularisation entropique

Dans le cas d'un transport non-linéaire à noyau gaussien :

— La largeur du noyau gaussien (sigma)

— Le protocole de génération des données synthétiques autour de la population de la classe considérée lors de l'apprentissage du transport

— Le seuil sur la distance lors de la définition de l'espace de validité de transport

Lors du calcul du seuil :

— L'option : "1-vs-all" / "1-vs-1"

— La méthode : misclassification / régression logistique

A.6.3 Résultats

Classification avec classe de rejet *via* le TO : Dans le cas du transport optimal, la nature des données permet de statuer quant à l'approche à utiliser.Les données Output sont de petite dimension (la dimension est le nombre de classes, soit 4 dans notre cas). De plus, la distribution des échantillons d'une classe donnée peut être considérée comme ayant un support convexe puisqu'ils se répartissent autour d'un des sommets du simplexe de probabilité. En effet, les composantes de ces caractéristiques sont bornées par l'intervalle [0,1], leur somme est égale à un, et une des composantes a en principe une valeur très proche de un. Dans ce cas, notre approche de classification avec classe de rejet utilisant le transport optimal peut être directement appliquée sur les données Output en calculant les transports dans le cas des transformations linéaires. Dans le cas des données Bottlenecks comme représentation de nos images, l'absence d'hypothèse sur la convexité des données fait que nous sommes contraints de considérer les transformations non linéaires en utilisant un noyau gaussien pour le calcul du transport. Le paramètre Sigma est calculé en divisant la moyenne des distances au plus proche voisin de chaque échantillon au sein de la population d'une même classe par une constante. Cette constante empirique est optimisée dans le but d'obtenir le meilleur score de classification. Le seuil sur la distance est fixé pour chaque classe à 1.5 fois le rayon de la plus petite sphère englobante de la population. Le transport est multi-label et les échantillons de la population synthétique sont générés de façon aléatoire autour de la plus petite sphère englobante dans un espace qui s'étend jusqu'à deux fois la variance de la population de la classe correspondante]. Pour une question d'occupation mémoire sur ordinateur, nous avons décidé de réduire le nombre de dimensions à celui des données Output, autrement dit sept via une ACP.

Les meilleurs résultats sont obtenus avec les données Output, en calculant un transport linéaire par classe, et en utilisant la méthode régression logistique pour l'apprentis-

sage des seuils. Les résultats de classification obtenus avec les deux options "1-vs-all" / "1-vs-1" sont très proches. Les résultats sur les images des phénotypes connus et nouveaux obtenus avec la méthode GMM (approche décrite dans le paragraphe 3.2 du chapitre 3) et TO sont présentés dans le tableau :

Type de données	Modèle	Classes connues	Classe de rejet
Bottlenecks	GMM	91	98
	TO	72	100
Output	GMM	96	76
	TO	92	83

TABLEAU 5 – Précisions de classification en pourcentage pour les deux méthodes de classification avec classe de rejet (GMM et TO) et les deux types de données (Bottlenecks et Output). Les valeurs réelles ont été arrondies à l'entier le plus proche.

Les scores moyens de classification en considérant les données Bottlenecks comme représentation de nos images sont de 94.5% avec les GMM contre 86% pour le TO. Ceux sur les données Output sont de 86% avec les GMM contre 87.5% pour le TO.

Limite de la méthode : La figure 41 illustre les limites de la méthode couplage et "mappage" simultanés sur des données de synthèse. En effet, nous nous sommes rendu compte d'un "effet de rebond" lors du transport d'échantillons non vus dans l'ensemble d'apprentissage et qui sont trop éloignés d'un des échantillons utilisés pour apprendre le transport. En effet, un échantillon au cœur de la population est transporté au cœur de la PDF cible. En revanche, c'est également le cas pour ceux situés hors de la population. En regardant de l'intérieur vers l'extérieur de la population, les échantillons présentent (suite à leur transport) une valeur de PDF forte (échantillon noir) puis une PDF faible (orange), très faible (jaune) puis à nouveau faible (orange) puis forte (noir). Ce phénomène de rebond explique les résultats obtenus sur nos données.

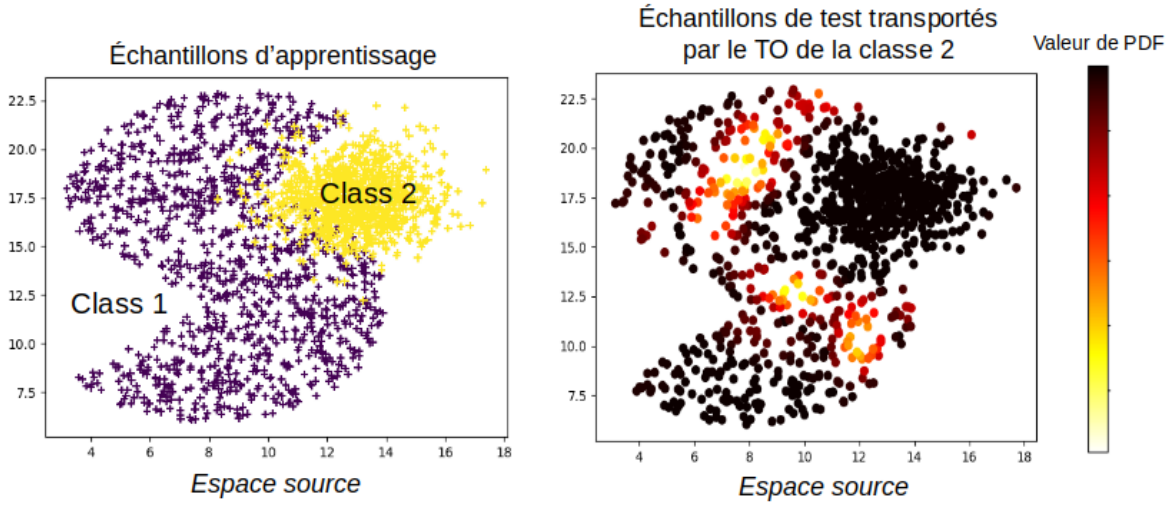


FIGURE 41 – À gauche : Les échantillons de l'ensemble d'apprentissage des deux classes. À droite : Les échantillons de l'ensemble de test des deux classes colorés en fonction de la valeur de la PDF cible aux échantillons transportés par le transport optimal appris pour la classe 2.

A.6.4 Conclusion

Les méthodes GMM et TO de classification avec classe de rejet suivent toute deux la même stratégie générale fondée sur l'apprentissage d'un modèle ou d'un transport indépendamment pour chaque classe, l'apprentissage d'un seuil fondé sur les interactions de classes et une procédure de prédiction. La différence réside dans le fait que les échantillons dont la classe est à prédire devront être transportés dans l'espace cible avant d'être classés. Dans l'application étudiée, les méthodes proposées fournissent de bonnes précisions de classification pour la classe de rejet et les classes connues. Plusieurs points restent néanmoins à améliorer. Informatiquement, le grand nombre de données entraîne un problème de mémoire lorsque l'on dépasse une certaine dimension ou encore lors de la génération des échantillons de synthèse (ceux autour de la population de la classe considérée). Par ailleurs, le phénomène de rebond n'est pas négligeable. Pour le supprimer ou l'atténuer, il est nécessaire de développer une fonction de mapping plus adaptée à notre approche, ce qui constitue une question difficile.

A.7 Évolution des paramètres longueur et aire des cellules de champignons avant la 30 ème acquisition

Les images avant la 30 ème acquisition (T_{30}) des séquences temporelles ne sont pas prises en compte dans les travaux présentés dans le chapitre 5 de ce manuscrit (création de modèles de croissance de *Botritis cinerea*). Afin d'estimer le niveau de développement des champignons des images avant le temps T_{30}, la distribution des valeurs des paramètres aire et longueur est étudiée. La figure 42 illustre ces distributions pour plusieurs temps entre T_1 et T_{29}.

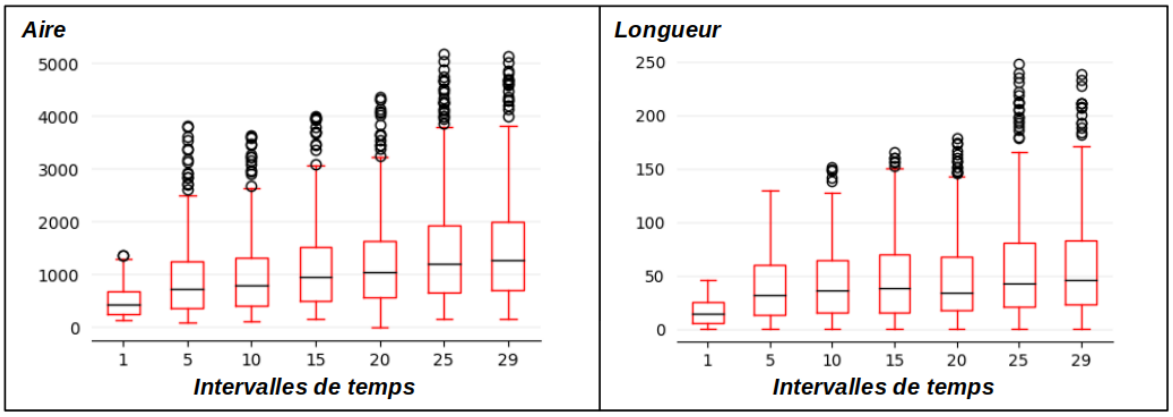


FIGURE 42 – Boîtes à moustaches des distributions des valeurs des paramètres aire et longueur des squelettes des objets détectés sur les images aux temps T_1, T_5, T_{10}, T_{15}, T_{20}, T_{25} et T_{29}. Les cercles correspondent aux valeurs aberrantes et la ligne noire dans chaque boîte représente la médiane.

Pour rappel, les images avant le temps T_{30} ne sont pas considérées dans le calcul des paramètres du modèle (voir le paragraphe 5.3.2). Or la figure 42 permet de nous rendre compte que les cellules de champignons ont commencé à se développer. En effet, les boîtes à moustaches indiquent une augmentation linéaire des valeurs des paramètres aire et longueur.

Ainsi, du fait du caractère linéaire de l'évolution de la longueur avant le temps T_{30}, nous avons choisi de faire croître à chaque itération une branche à partir du spore dont les arêtes sont d'une longueur constante égale à $\frac{a}{30}$, avec a l'expression de la longueur à T_0(voir le paragraphe 5.8.2).

A.8 Simulation de la croissance d'un champignon

La figure 43 illustre l'évolution du graphe (présenté dans le paragraphe 5.8.3 du chapitre 5) au cours du temps. Les paramètres (longueurs et nombre de branches) sont définis selon leurs PDFs respectives (loi uniforme et PDFs des coefficients a et b).

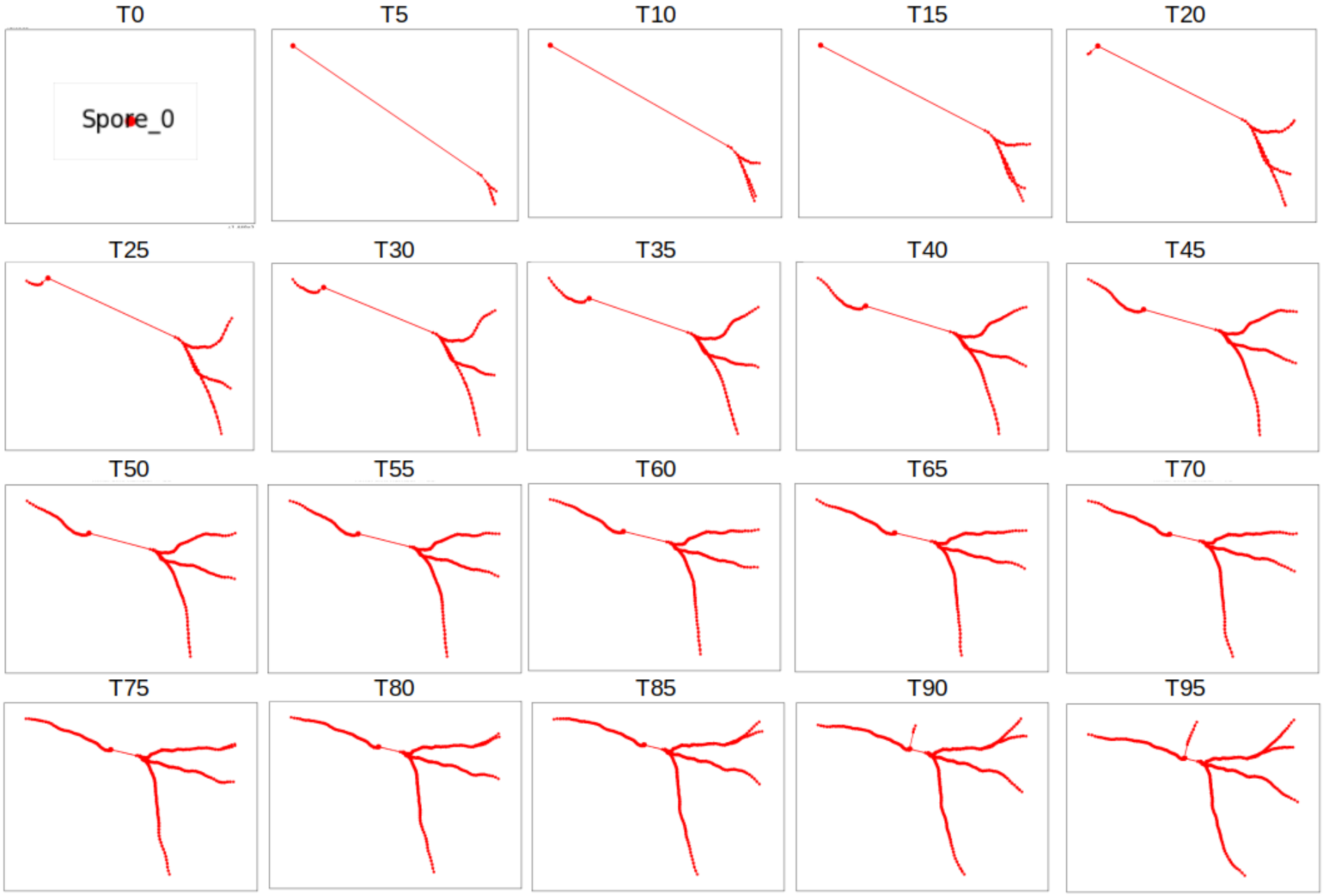


FIGURE 43 – Évolution du graphe au cours du temps.

A.9 Références

«Frac classification of fungicides», https://www.frac.info/docs/default-source/publications/frac-mode-of-action-poster/frac-moa-poster-2021.pdf?sfvrsn=a6f6499a_2. Accessed : 2021. ix, III

COURTY, N., R. FLAMARY, D. TUIA et A. RAKOTOMAMONJY. 2016, «Optimal transport for domain adaptation», *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, nº 9, p. 1853–1865. VI

CUTURI, M. 2013, «Sinkhorn distances : Lightspeed computation of optimal transport», dans *Advances in neural information processing systems*, p. 2292–2300. VI

DUBOS, B. 2002, *Les maladies cryptogamiques de la vigne : les champignons parasites des organes herbacés et du bois de la vigne.*, Éditions Féret. II

LAROUI, S. *Méthode de classification automatique de phénotypes de cellules de champignons pythopathogènes : rapport de stage de Master 2. Master2 Biologie, Informatique et Mathématiques. Nice Sophia Antipolis : Université de Nice Sophia Antipolis, 2017, 54 p.* I, II

MARTINEZ, F., D. BLANCARD, P. LECOMTE, C. LEVIS, B. DUBOS et M. FERMAUD. 2003, «Phenotypic differences between vacuma and transposa subpopulations of botrytis cinerea», *European Journal of Plant Pathology*, vol. 109, nº 5, p. 479–488. II

PERROT, M., N. COURTY, R. FLAMARY et A. HABRARD. 2016, «Mapping estimation for discrete optimal transport», *Advances in Neural Information Processing Systems*, vol. 29, p. 4197–4205. VIII