université PARIS-SACLAY

Multiple instance learning for cytometry data analysis and tumoral cells detection

Multiple instance learning pour l'analyse de données de cytométrie appliquée à la détection de cellules tumorales

Thèse de doctorat de l'Université Paris-Saclay

École doctorale n°574 Mathématiques Hadamard (EDMH)
Spécialité de doctorat : Mathématiques appliquées

Graduate school: Mathématiques. Référent : Faculté des sciences d'Orsay

Thèse préparée au Laboratoire de Mathématiques d'Orsay (Université Paris-Saclay, CNRS), sous la co-direction de
Christine KERIBIN, professeure et Pascal MASSART, professeur

Thèse soutenue à Paris-Saclay, le 22 Juin 2026, par

Pierre-André MIKEM

Composition du Jury

Membres du jury avec voix délibérative

Frédéric CHAZAL Directeur de recherche, INRIA Saclay	Président
Anne GEGOUT-PETIT Professeure, Université de Lorraine	Rapporteur & Examinatrice
Vincent VANDEWALLE Professeur, Université Côte d'Azur	Rapporteur & Examinateur
Guillermo DURAND Maître de conférences, Université Paris-Saclay	Examinateur
Baptiste LABARTHE Data scientist, Metafora biosystems	Examinateur
Cathy MAUGIS-RABUSSEAU Professeure, Institut de Mathématiques de Toulouse	Examinatrice

Thèse de doctorat
NNT : 2026UPASM022

Titre : Multiple instance learning pour l'analyse de données de cytométrie appliquée à la détection de cellules tumorales

Mots clés : Apprentissage supervisée, Modèles de mélange, Réseaux de neurones, Transport optimal numérique, Statistique non paramétrique

Résumé : Les technologies biomédicales modernes permettent d'observer les systèmes biologiques avec un niveau de détail sans précédent. En cytométrie en flux, des millions de cellules sont mesurées individuellement, chacune étant décrite par un ensemble de variables biologiques. Pourtant, les décisions cliniques ne sont pas prises au niveau des cellules, mais au niveau des patients. Cet écart entre des observations fines et des décisions globales soulève une question fondamentale : comment les différentes caractéristiques des cellules peuvent-elles être traduites en informations cliniques pertinentes ? L'une des principales difficultés réside dans le fait que les signaux pathologiques sont souvent subtils et portés par de petits sous-ensembles de cellules, sans annotations fiables. Dans ce contexte, le cadre du Multiple Instance Learning offre une manière naturelle de modéliser ces données, en représentant chaque patient comme un ensemble d'observations et en apprenant à partir d'étiquettes globales. Au-delà de ce cadre spécifique, ce travail met en lumière un paradigme plus large : l'apprentissage à partir de données structurées et distributionnelles. En interprétant les patients comme des distributions de cellules, il devient possible de capturer une variabilité complexe et d'identifier des motifs significatifs qui resteraient invisibles avec des approches traditionnelles. Ces idées ouvrent de nouvelles perspectives pour l'apprentissage automatique dans des applications où les données peuvent être hétérogènes, faiblement annotées et structurées à plusieurs échelles.

Title: Multiple instance learning for cytometry data analysis and tumoral cells detection

Keywords: Supervised learning, Mixture models, Neural networks, Computational optimal transport, Nonparametric statistics

Abstract: Modern biomedical technologies make it possible to observe biological systems at an unprecedented level of detail. In flow cytometry, millions of cells are measured individually, each described by a rich set of biological features. Yet, in clinical practice, decisions are not made at the level of cells, but at the level of patients. This gap between fine-grained observations and global decisions raises a fundamental question: how can cellular heterogeneity be translated into meaningful clinical insight? One of the main difficulties lies in the fact that pathological signals are often subtle and carried by small subsets of cells, without reliable annotations. In this context, the Multiple Instance Learning framework offers a natural way to model such data, by representing each patient as a set of observations and learning from global labels. Beyond this specific setting, this work highlights a broader paradigm: learning from structured and distributional data. By interpreting patients as distributions of cells, it becomes possible to capture complex variability and identify meaningful patterns that would remain hidden under traditional approaches. These ideas open new perspectives for machine learning in applications where data can be heterogeneous, weakly annotated, and structured across multiple scales.

Université Paris-Saclay
Institut de Mathématiques d'Orsay
Bâtiment 307 / Rue Michel Magat / 91400 Orsay, France

Remerciements

Au terme de ces années de recherche, je souhaite adresser mes remerciements à toutes les personnes qui ont contribué, de près ou de loin, à la réalisation de cette thèse.

Je remercie tout d'abord les membres de mon jury : Fédéric Chazal, Anne Gégout-Petit, Vincent Vandewalle, Guillermo Durand, Baptiste Labarthe et Cathy Maugis-Rabusseau. Je leur suis reconnaissant d'avoir accepté d'évaluer ce travail et de consacrer du temps à sa lecture. Je les remercie également pour l'intérêt qu'ils ont porté à mes recherches ainsi que pour les échanges et les remarques qui ont contribué à enrichir cette thèse.

Je tiens également à remercier les membres du laboratoire qui m'ont accueilli durant ces années. J'adresse une reconnaissance particulière à Clotilde D'epenoux, dont l'aide a été précieuse tout au long de la préparation de ma soutenance. Sa disponibilité, sa réactivité et son soutien dans les nombreuses démarches ont facilité cette étape importante de mon parcours doctoral.

J'adresse mes plus sincères remerciements à mes encadrants académiques pour leur accompagnement tout au long de cette thèse. Je souhaite tout particulièrement remercier Christine qui me soutient depuis mes années de master. Son accompagnement au fil des années, sa disponibilité et ses encouragements ont joué un rôle déterminant dans mon parcours académique et dans l'aboutissement de ce travail. Je remercie également Pascal pour ses conseils avisés tout au long de la thèse. Son intuition mathématique, sa capacité à prendre du recul sur les problèmes rencontrés et son optimisme ont souvent permis d'éclairer les discussions et d'ouvrir de nouvelles perspectives. Enfin, je tiens à remercier Guillermo qui a rejoint le projet en cours de route et dont la contribution a été précieuse. Son aide dans la rédaction du manuscrit a contribué à améliorer la qualité de ce travail. Je lui suis également reconnaissant pour son aide sur les aspects pratiques liés à l'implémentation des algorithmes, ainsi que pour sa disponibilité et ses conseils

toujours pertinents face aux difficultés rencontrées.

Cette thèse ayant été réalisée dans le cadre d'une convention Cifre, je souhaite remercier l'ensemble de mes collègues au sein de Metafora biosystems pour leur accueil et les échanges durant ces années de travail. J'adresse tout particulièrement mes remerciements à Baptiste qui m'a accompagné tout au long du projet. Son aide dans la mise à disposition et la compréhension des données, ainsi que le soutien qu'il m'a apporté dans les différentes étapes de la thèse, ont été précieux. Je remercie également Heng pour son aide sur les aspects techniques liés aux réseaux de neurones. Nos nombreuses discussions m'ont permis de progresser sur plusieurs problématiques de ce travail. Je tiens aussi à remercier Vincent pour la confiance qu'il m'a accordée en me confiant ce projet doctoral et en me donnant l'opportunité de mener cette thèse au sein de l'entreprise. Malgré les défis et les difficultés rencontrés au cours de ces années, j'ai particulièrement apprécié cette expérience à l'interface entre recherche académique et monde industriel. Enfin, je souhaite remercier l'ensemble de mes collègues de l'équipe biologique. Grâce à eux, j'ai découvert l'univers de la cytométrie en flux et pu mieux appréhender les enjeux expérimentaux associés à ce domaine.

Je remercie chaleureusement mes amis, et plus particulièrement ceux qui m'accompagnent depuis les années de licence et de master. Merci pour votre soutien, votre présence dans les moments de doute comme dans les réussites, ainsi que pour tous les moments de partage qui ont rendu cette aventure plus agréable.

Je souhaite adresser mes remerciements les plus sincères à ma famille. Merci pour votre affection, votre confiance et votre soutien constant tout au long de mon parcours. Vous avez toujours été présents pour m'encourager et me permettre d'avancer, même lorsque le chemin était difficile.

Enfin, une mention particulière à ma mère, une source inépuisable d'affection et de dévouement. Elle a toujours exprimé de la fierté à l'égard de mon travail. Sa présence m'a donné de la force pour aller jusqu'au bout de cette aventure.

Contents

Introduction **5**

1 Contexte applicatif 6
2 Cytométrie en flux 7
2.1 Historique et principe de la cytométrie en flux 7
2.2 Analyse de données de cytométrie 10
3 Multiple instance learning 14
4 Contributions 16

1 Exploratory data analysis **18**

1 Description of datasets 18
1.1 FlowCAP AML 18
1.2 Multiple Myeloma 19
2 Principal component analysis 20
3 Understanding variables independence structure 22
3.1 Independence Structure Density Estimation 22
3.2 FlowCAP-AML data analysis 24

2 A Gaussian mixture-based approach **27**

1 Overview of Multiple Instance Learning methods 27
1.1 Problem formulation and label ambiguity 27
1.2 MIL assumptions and categorization of methods 28
1.3 Theoretical analysis of MIL and a probabilistic perspective 30
2 A baseline model with Gaussian mixture 30
2.1 Diverse density 31
2.2 A MIL method by combining classifier predictions 33
2.3 Our baseline model 38
3 Numerical experiments 40
3.1 Visualizing bags with covariance ellipses 41

3.2 Experiment on synthetic data 42
3.3 Experiment on flow cytometry data 46

3 MIL with neural networks 50
1 Learning from bags considered as single objects 50
2 Deep learning on sets of observations 52
2.1 From classical rule-based MIL to bags representation learning . . . 52
2.2 Set-input neural networks and Attention-based MIL 53
2.3 Neural MIL for computer aided diagnosis: medical image analysis and an extension towards flow cytometry data analysis 56
3 Combining Autoencoders and Mixture of Experts 58
3.1 A survey on Autoencoders 58
3.2 A local MIL framework based on Autoencoders and Mixture of Experts 60
4 Application to flow cytometry data 62

4 Data augmentation with Wasserstein barycenters 67
1 Distribution interpolation for data augmentation 67
2 Barycenter notion in the Wassserstein space 69
2.1 Why Wasserstein barycenters are geometrically attractive? 69
2.2 Existence, characterization and computation in particular cases . . 69
3 Application to bags generation for MIL 72
3.1 Computing Wasserstein barycenter in the discrete setting 72
3.2 The proposed method for MIL data augmentation 74
4 Numerical experiments 76
5 Beyond discrete Wasserstein barycenters 80
5.1 From discrete to continuous optimal transport: stochastic and neural approaches 81
5.2 Generative Wasserstein barycenters: continuous estimation and sampling 82
6 Broader perspectives on Wasserstein barycenters 84

5 Bayesian classification in MIL 86
1 Bayes classifier 86
2 A Bayes Classifier for MIL 88
2.1 Intuition 88
2.2 The proposed method 89
3 Numerical experiments 91

Conclusion **95**

1 Conclusions on the research work 95

2 Perspectives 96

Introduction

Cette thèse s'inscrit dans le cadre d'une collaboration entre le Laboratoire de Mathématiques d'Orsay et l'entreprise Metafora Biosystems. Elle vise à développer des méthodes d'apprentissage automatique pour classer des patients selon leur état de santé à partir de données de cytométrie en flux, tout en identifiant des schémas explicatifs au niveau cellulaire. L'objectif applicatif est le développement d'outils de diagnostic non invasifs, fondés sur l'analyse de prélèvements biologiques tels que le sang ou les tissus d'organes. La cytométrie en flux permet de caractériser des millions de cellules individuellement selon de multiples variables, offrant ainsi une description fine de l'état biologique d'un patient. Toutefois, l'exploitation de ces données à des fins diagnostiques pose des défis méthodologiques importants, liés à leur grande dimension, à leur variabilité, et à l'absence fréquente d'annotations fiables au niveau des cellules tumorales. Dans ce contexte, cette thèse explore des approches d'apprentissage automatique permettant de relier des observations au niveau cellulaire à des décisions au niveau du patient. Plus précisément, nous cherchons à répondre aux questions suivantes :

- Peut-on discriminer l'état de santé de patients à partir de populations cellulaires hétérogènes ?
- Peut-on identifier des sous-ensembles de cellules caractéristiques d'un état pathologique ?
- Ces schémas locaux (au niveau cellulaire) peuvent-ils être interprétés comme des signatures biologiques pertinentes ?

Cette introduction est organisée comme suit. Nous présentons tout d'abord le contexte applicatif, le principe de la cytométrie en flux ainsi que son utilisation pour le diagnostic médical. Nous décrivons ensuite les spécificités des données associées et les principaux défis méthodologiques liés à leur analyse. Dans ce contexte, nous mettons en évidence les limites des approches existantes et introduisons le paradigme du Multiple Instance

Learning, qui constitue le cadre méthodologique principal de cette thèse. Enfin, nous détaillons les contributions proposées, en mettant en évidence la progression adoptée.

1 Contexte applicatif

En cytologie des cancers, il n'existe pas de définition universelle et stable d'une cellule tumorale basée uniquement sur des marqueurs observables. Les caractéristiques discriminantes varient selon les pathologies, les individus et les contextes biologiques. Cette variabilité rend difficile l'identification directe de cellules anormales à partir de critères fixes. À la question de savoir ce qu'est une cellule cancéreuse, il est difficile de répondre de façon franche et succinte. D'après une citation de [Sha05] : "It is sometimes this, and sometimes that, and not this, and the other thing, or the third thing, in the presence of a fourth ...".

Pour illustrer ces phénomènes, considérons une cellule que nous appellerons Néo. Dans un état normal, elle évolue dans un environnement régulé, où elle produit de l'énergie, se divise de manière contrôlée et interagit avec les autres cellules. Son métabolisme est stable et adapté à ses fonctions. Cependant, sous l'effet de perturbations — mutations, stress ou signaux anormaux — cet équilibre peut être rompu. La cellule modifie alors son métabolisme, par exemple en favorisant la glycolyse même en présence d'oxygène (effet Warburg) [LL16], afin de soutenir une prolifération rapide. Elle échappe progressivement aux mécanismes de régulation, se divise de manière incontrôlée et résiste à la mort cellulaire. Ce passage d'un état normal à un état pathologique ne repose pas sur une signature unique et stable, mais sur des transformations progressives et variables selon les cellules et les contextes. En pratique, les cellules pathologiques ne se distinguent pas toujours individuellement de manière évidente, mais plutôt à travers des schémas émergents au sein de populations cellulaires hétérogènes.

La cytométrie en flux permet d'observer indirectement ces phénomènes via des signatures multiparamétriques mesurées à l'échelle de la cellule. L'enjeu est alors d'identifier, au sein de populations cellulaires complexes, des schémas récurrents associés à des états pathologiques, sans disposer nécessairement d'annotations fiables au niveau individuel.

2 Cytométrie en flux

Dans cette section, nous faisons d'abord une présentation rapide de l'historique de la cytométrie (origines, principe et usage pour du diagnostic médical). Enfin, nous parlons des données de cytométrie, des méthodes d'analyse actuellement utilisées et de leurs limites.

2.1 Historique et principe de la cytométrie en flux

Analyser des cellules individuellement est un concept qui a des racines anciennes, notamment via des colorations microscopiques. Il faut compter à partir du XXe siècle avec l'apparition de technologies optiques et électroniques pour pouvoir analyser des échantillons de cellules à grande échelle. Le terme "cytométrie en flux" n'a été imposé qu'en 1976, huit ans après l'introduction du premier cytomètre en flux basé sur la fluorescence. Les premiers cytomètres avaient une utilisation assez simple. Ils permettaient juste de compter avec précision le nombre de cellules dans un échantillon sans les caractériser explicitement. Depuis la fin des années 90 et le début des années 2000, la cytométrie a bénéficié d'avancées rapides en optique, en électronique, en chimie et en informatique pour devenir ce qu'elle est aujourd'hui. L'usage clinique s'est également largement déployé, grâce aux consortiums de standardisation (ex. EuroFlow[1]).

Lumière et matière sont probablement les deux mots qui résument le mieux la cytométrie en flux. Derrière ces deux mots, se cache un concept assez technique qui nous permet de répondre à plusieurs questions à propos des cellules. Une suspension de cellules est placée dans un fluide (tampon). Ce flux de suspension passe dans un faisceau dit *sheath fluid*. Les cellules défilent à très grande vitesse à travers le faisceau laser. Grâce à la quantité de lumière diffusée, émise et absorbée par les cellules, plusieurs caractéristiques sont mesurées.

Lorsqu'une cellule passe à travers le faisceau laser (voir figure 1), la lumière diffusée permet de mesurer sa morphologie et sa structure interne. En cytométrie, ce phénomène est appelé *scattering*. La lumière diffusée sous un petit angle est essentiellement liée à la taille ou au diamètre de la cellule : on parle de *Forward Scatter* (FSC). La lumière diffusée à un angle large ($\approx$ 90 degrés) renseigne sur la granularité et la complexité interne de la cellule : dans ce cas, on parle de *Side Scatter* (SSC).

[1] consortium fondé en 2005 dont le but est de standardiser la cytométrie en flux afin de faire progresser le diagnostic de certaines maladies.

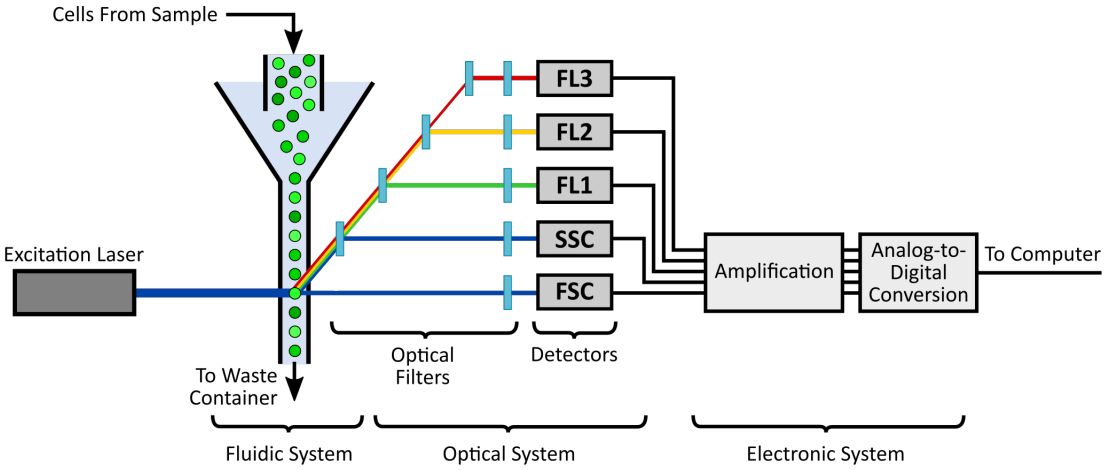


Figure 1: **Principe de la cytométrie en flux** [Bio25b]. Une suspension de cellules est injectée dans un flux liquide et traverse un faisceau laser une par une. Lors du passage dans le faisceau, chaque cellule diffuse et émet de la lumière, qui est mesurée par des détecteurs. La lumière diffusée sous un faible angle (Forward Scatter, FSC) fournit une information liée à la taille de la cellule, tandis que la diffusion à angle large (Side Scatter, SSC) renseigne sur sa structure interne et sa granularité. En parallèle, des signaux de fluorescence peuvent être mesurés lorsque les cellules sont marquées avec des fluorochromes. Chaque cellule est ainsi décrite par un vecteur de caractéristiques, correspondant aux intensités mesurées dans différents canaux.

Il est également possible de marquer chaque cellule avec un fluorochrome (anti-corps ou molécules fluorescentes). Chaque fluorochrome absorbe à une longueur d'onde (lumière excitatrice) et émet à une longueur d'onde plus longue. En cytométrie en flux, on utilise des anticorps monoclonaux (ou parfois polyclonaux) dirigés contre des antigènes spécifiques exprimés à la surface (ou parfois à l'intérieur) des cellules (voir figure 2) - par exemple, des marqueurs CD (*cluster of differentiation*) dans le sang ou le système immunitaire. Ces anticorps sont conjugués à des molécules fluorescentes (fluorochromes) : on parle d'une molécule fluorescente fixée covalentement sur l'anti-corps. Lorsqu'une cellule exprime l'antigène cible, l'anticorps marqué s'y fixe. Cela permet de l'identifier et de quantifier l'intensité de l'expression de ce marqueur via la fluorescence.

Un fluorochrome, après excitation par un laser à une certaine longueur d'onde, émet de la lumière à une longueur d'onde plus grande. Pour faire des analyses multi-paramétriques (en marquant plusieurs antigènes simultanément sur une même cellule), on choisit plusieurs fluorochromes compatibles. Ces fluorochromes sont excités par les lasers disponibles sur le cytomètre et leurs émissions sont détectables dans des canaux distincts (voir figure 1). Chaque fluorochrome absorbe à une longueur d'onde (lumière excitatrice) et émet à une longueur d'onde plus longue. Cette fluorescence est captée par des photodétecteurs (photomultiplicateurs, PMT) via des filtres spectraux. Le terme

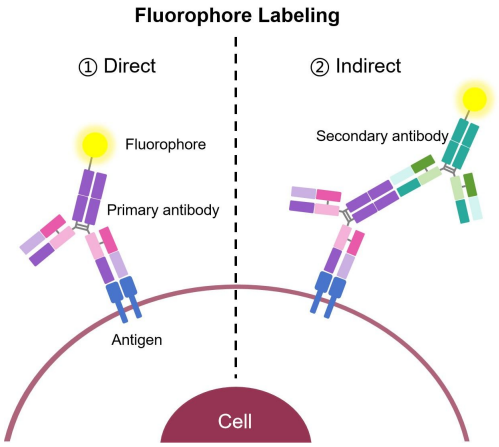


Figure 2: **Principe de l'immunofluorescence** [Bio25a]. Des anticorps spécifiques de marqueurs biologiques sont conjugués à des fluorochromes et se fixent sur les cellules exprimant ces marqueurs. Lorsqu'ils sont excités par un laser à une longueur d'onde donnée, les fluorochromes émettent de la lumière à une longueur d'onde plus grande. Cette fluorescence est détectée dans différents canaux, chacun correspondant à un marqueur particulier. L'intensité mesurée reflète le niveau d'expression du marqueur sur la cellule, permettant de caractériser finement les populations cellulaires.

Photomultiplier Tube (PMT) désigne un détecteur optique extrêmement sensible, capable de convertir des photons en un signal électrique amplifié. L'intensité de fluorescence lue dans un canal est donc directement liée à la quantité de photons émis par la cellule pour ce canal. Chaque canal PMT correspond à une caractérisque mesurée de cellule.

Usage clinique de la cytométrie. La perspective de développer des méthodes automatiques d'identification de cellules pour l'hématologie et la cytologie des cancers a donné de l'élan au financement des premiers cytomètres. Ces méthodes nécessitent qu'un schéma d'identification des cellules d'intérêt soit bien établi et validé. La cytométrie en flux est aujourd'hui une brique essentielle de nombreux diagnostics, particulièrement en hématologie [WB01], immunologie [BKA+19], oncologie [Sal+19]. Dès les débuts, l'un des usages majeurs a été l'immunophénotypage : identifier les populations de cellules immunitaires (lymphocytes T, B, monocytes, cellules NK, etc.) en fonction de marqueurs de fluorescence. En hématologie, la cytométrie a permis de caractériser les leucémies et lymphomes en déterminant les profils des cellules malignes (expression anormale de marqueurs fluorescents). Le consortium EuroFlow, mis en place pour standardiser les panels et les procédures, a joué un rôle décisif dans l'uniformisation des méthodes de diagnostic par cytométrie en hématologie. Le suivi de la maladie résiduelle minimale, dit MRD (Minimal Residual Disease), est un autre domaine où la cytométrie permet de détecter des cellules tumorales très rares après traitement, ce qui informe sur le pronostic et la réponse thérapeutique.

2.2 Analyse de données de cytométrie

D'un point de vue statistique, les données de cytométrie en flux se présentent sous la forme d'ensembles de cellules, chacune décrite par un vecteur de mesures correspondant à des caractéristiques morphologiques (FSC, SSC) et à des intensités de fluorescence associées à différents marqueurs biologiques. Un échantillon peut contenir jusqu'à plusieurs millions de cellules charactérisées par des dizaines de variables.

L'analyse de ces données repose traditionnellement sur une succession d'étapes de prétraitement et d'exploration réalisées par des experts. Une première phase consiste à corriger les artefacts liés à l'acquisition et à filtrer les événements non pertinents. Cela inclut notamment l'élimination des doublets qui au passage simultané de plusieurs cellules dans le faisceau laser, ainsi que des cellules mortes, susceptibles de biaiser les mesures. Un autre enjeu technique central est celui de la compensation spectrale. En effet, les spectres d'émission des fluorochromes peuvent se chevaucher, ce qui entraîne des interférences entre canaux de mesure. Afin de corriger ces effets de recouvrement (spillover), les biologistes construisent une matrice de compensation permettant d'isoler au mieux le signal propre à chaque marqueur. Après compensation, les données sont généralement transformées afin d'en faciliter l'analyse et la visualisation. Les intensités de fluorescence peuvent en effet s'étendre sur plusieurs ordres de grandeur, rendant leur interprétation difficile. Des transformations non linéaires, telles que la transformation arcsinh ou bi-exponentielle, sont alors appliquées pour stabiliser la variance et mieux séparer les populations cellulaires. L'étape suivante consiste en un gating manuel, c'est-à-dire la sélection successive de populations cellulaires à l'aide de fenêtres définies dans des projections bidimensionnelles des données. Comme illustré dans la figure 3, cette procédure repose fortement sur l'expertise métier et sur des connaissances biologiques a priori. Bien qu'elle soit largement utilisée en pratique, elle présente plusieurs limitations: elle est subjective, difficilement reproductible, et peu adaptée à l'analyse de données de grande dimension.

La standardisation entre laboratoires est un défi (reproductibilité, choix des marqueurs, compensation, etc.). Le bruit technique, la variabilité des machines, des lots de réactifs, et le *batch effect* entre acquisitions sont des obstacles à l'analyse comparée des données. L'interprétation via gating, est souvent subjective ou dépendante de l'expertise. Ces limitations ont motivé le développement de méthodes computationnelles automatisées. Des études comparatives, notamment dans le cadre du challenge Flow Cytometry: Critical Assessment of Population Identification Methods (FlowCAP) [Agh+13], ont montré que certaines approches algorithmiques pouvaient rivaliser avec le gating manuel pour identifier des populations cellulaires. Ces méthodes visent à rendre

les analyses plus reproductibles, plus rapides et mieux adaptées à des volumes de données importants.

Parmi d'autres approches proposées, les méthodes de réduction de dimension jouent un rôle central. Par exemple, la méthode viSNE introduite par [Ami+13], basée sur t-SNE [Wan+21], permet de projeter des données de grande dimension dans un espace de dimension réduite tout en préservant leur structure locale. Cela facilite la visualisation et permet à un expert de retrouver des populations cellulaires. D'autres travaux se concentrent sur l'automatisation de l'identification des populations cellulaires à l'aide de méthodes de clustering. Le pipeline proposé par [Now+19] combine des techniques de prétraitement, de réduction de dimension et de clustering, notamment via l'algorithme FlowSOM [Van+15]. De même, [Sam+16] introduisent la méthode X-shift, qui permet de détecter automatiquement des populations cellulaires.

Dans une perspective applicative, plusieurs outils industriels ont également commencé à émerger pour faciliter l'analyse des données de cytométrie en flux. En particulier, l'entreprise partenaire de cette thèse, Metafora Biosystems, développe une plateforme reposant notamment sur l'algorithme de clustering ToMATo [Cha+13]. Cette approche permet d'identifier automatiquement des populations cellulaires à partir de la structure topologique des données, offrant ainsi une alternative aux méthodes de clustering plus classiques et contribuant à l'automatisation des workflows d'analyse.

Malgré ces avancées, une tension fondamentale persiste dans l'analyse des données de cytométrie en flux. Comme le soulignent plusieurs travaux récents [Ng+24; Eck+20; Sal+19], les approches traditionnelles restent fortement dépendantes de l'expertise humaine. Les méthodes classiques d'apprentissage automatique telles que le clustering permettent d'accélérer et de rendre moins fastidieuses les analyses faites manuellement par les experts. Ces méthodes tendent ainsi à reproduire des schémas connus plutôt qu'à découvrir de nouvelles structures biologiques qui peuvent être associées aux pathologies étudiées. Ce constat soulève une question centrale : comment passer d'une automatisation des stratégies expertes à une véritable découverte de nouvelles signatures biologiques à partir des données ? C'est précisément dans cette perspective que s'inscrit notre travail.

Les limitations des approches actuelles ne tiennent pas uniquement à la complexité des données, mais aussi à leur structure. En effet, les observations sont disponibles au niveau des cellules, tandis que les décisions cliniques sont formulées au niveau des patients. Cette dissociation rend difficile l'application des méthodes d'apprentissage supervisé classiques, qui supposent généralement la disponibilité de labels au niveau de chaque observation

(cellulle). Une stratégie naïve consisterait à agréger les cellules pour obtenir une représentation globale de chaque patient. Cependant, une telle approche risque de masquer des signaux faibles portés par des sous-populations cellulaires rares, pourtant déterminantes dans le diagnostic. À l'inverse, une approche purement locale nécessiterait des annotations fiables au niveau cellulaire, qui sont en très souvent incertaines ou manquantes sur les cellules tumorales. Cette tension entre information locale et décision globale suggère de repenser le cadre d'apprentissage afin de prendre en compte explicitement la structure des données.

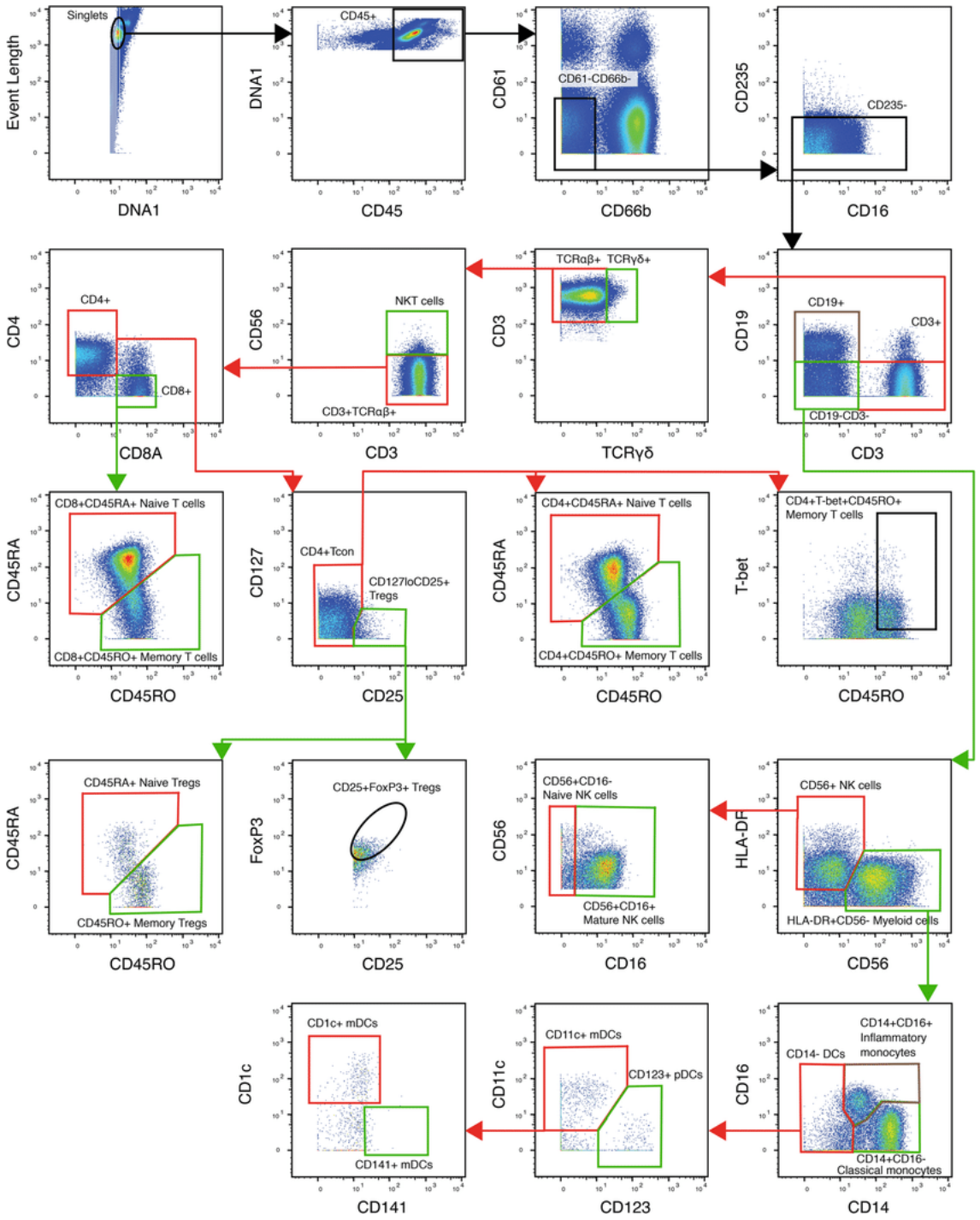


Figure 3: **Illustration d'une stratégie de gating manuel en cytométrie en flux** [Rus+19]. Les populations cellulaires sont sélectionnées de manière séquentielle à l'aide de projections bidimensionnelles des données. À chaque étape, une région est définie manuellement afin d'isoler un sous-ensemble de cellules, qui est ensuite utilisé comme entrée pour l'étape suivante. Les flèches indiquent la progression du processus de sélection, où chaque décision dépend des choix effectués précédemment. Cette procédure repose fortement sur l'expertise de l'utilisateur et peut conduire à des résultats subjectifs, sensibles aux paramètres choisis et difficilement reproductibles.

3 Multiple instance learning

Les données de cytométrie en flux peuvent présenter une structure particulière si l'on considère qu'elles proviennent d'études cliniques réalisées sur différents patients. Chaque patient est décrit par un ensemble de cellules mesurées individuellement, mais seule une information globale correspondant à son état de santé, est disponible. Autrement dit, nous considérons que les observations sont naturellement organisées en ensembles, sans annotation au niveau des cellules. Cette structure rend difficile l'application directe des méthodes d'apprentissage supervisé classiques et appelle un cadre d'apprentissage adapté aux données organisées en ensembles d'observations.

Le paradigme du Multiple Instance Learning (MIL) fournit un cadre naturel pour modéliser des données organisées en ensembles d'observations. Dans ce cadre, les observations sont regroupées en bags, chacun constitué de plusieurs instances, et les labels sont disponibles uniquement au niveau des bags (figure 4). Dans notre contexte, un patient correspond à un bag, et les cellules qui le composent en sont les instances. L'hypothèse centrale du MIL est que l'information pertinente pour la prédiction d'u bag peut être portée par un sous-ensemble d'instances. Ainsi, l'état de santé d'un patient peut être expliqué par la présence de certaines sous-populations cellulaires spécifiques, plutôt que par l'ensemble de ses cellules. Cette propriété est particulièrement adaptée aux données de cytométrie, où les cellules pathologiques peuvent représenter une fraction minoritaire de l'échantillon.

Le Multiple Instance Learning a été introduit par [DLL97] dans le cadre de la prédiction de l'activité de molécules, où chaque molécule peut adopter plusieurs conformations, sans que l'on sache laquelle est responsable de l'activité observée. Cette situation présente une analogie directe avec notre cadre : de la même manière qu'une molécule est active si au moins une de ses conformations l'est, un patient malade peut être caractérisé par la présence de certaines cellules spécifiques. Ce point de vue fait également écho au processus de diagnostic médical, où la décision repose rarement sur un signal unique, mais sur la combinaison de divers signaux dont l'interprétation conjointe permet de caractériser une pathologie.

Dans le contexte de la cytométrie en flux, plusieurs travaux récents abordent des problématiques proches des nôtres, bien que sous des angles différents. Par exemple,[Laz+23] proposent une méthode basée sur la géométrie de l'information (FINE), dans laquelle les échantillons de différents patients sont plongés dans un espace de plus faible dimension afin de faciliter la classification des patients. Ce type d'approche permet de capturer des différences globales entre patients, mais ne vise pas à identifier explicitement les

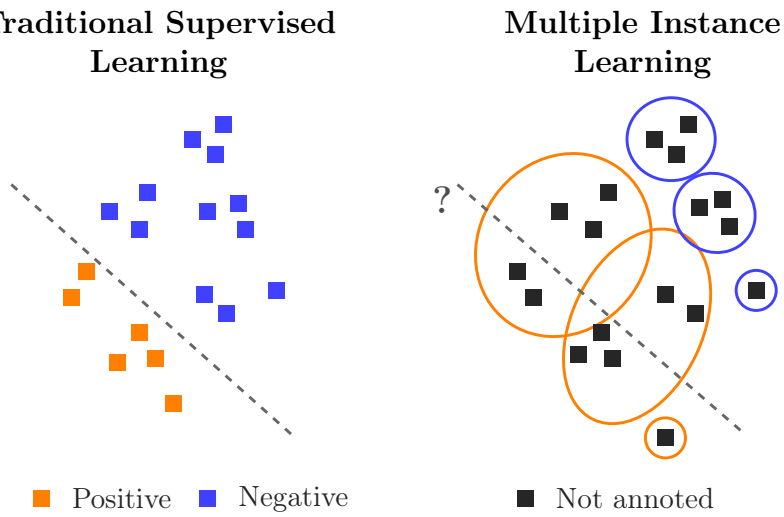


Figure 4: **Comparaison entre apprentissage supervisé classique et Multiple Instance Learning** [Mai21]. Dans le cadre supervisé classique, chaque observation individuelle (instance) est associée à un label, et le modèle apprend directement une relation entre les caractéristiques de chaque instance et sa classe. À l'inverse, dans le cadre du Multiple Instance Learning, les données sont organisées en ensembles (bags) composés de plusieurs instances, mais seuls des labels globaux sont disponibles au niveau des bags. Le modèle doit alors exploiter la structure interne de chaque ensemble pour relier des motifs locaux, portés par certaines instances, à une décision globale au niveau du bag.

sous-populations cellulaires responsables de ces différences. D'autres travaux tels que [GBC24] s'intéressent à la détection de populations rares ou anormales au sein des échantillons. Les auteurs proposent une méthode supervisée pour détecter la présence d'une faible proportion de cellules issues d'une distribution différente, à l'aide d'un test basé sur des rapports de densité. Leur méthode repose sur la disponibilité d'exemples annotés au niveau cellulaire, ce qui limite leur applicabilité dans des contextes faiblement supervisés. Par ailleurs, certains auteurs comme [HBL23] adoptent un point de vue intermédiaire en proposant un cadre d'apprentissage sur des mesures de probabilité. Dans ce cadre, les mesures font l'objet d'une tâche de classification. En considérant les bags aussi comme des mesures empiriques, ce formalisme est assez proche du cadre du Multiple Instance Learning. Toutefois, ces approches se concentrent principalement sur la classification au niveau des distributions, sans nécessairement fournir une interprétation fine au niveau des instances individuelles.

Ces différentes approches mettent en évidence deux stratégies principales : soit une modélisation globale des patients, au détriment de l'interprétabilité locale, soit une identification de motifs nécessitant des annotations précises au niveau des cellules. Notre approche vise à combiner ces deux perspectives en intégrant, à la fois une modélisation au niveau du patient et l'identification de schémas explicatifs au niveau cellulaire. L'objectif

est ainsi de relier classification au niveau du patient et interprétation biologique, en mettant en évidence des sous-populations cellulaires associées à l'état de santé.

4 Contributions

Cette thèse s'inscrit dans le cadre de l'analyse de données de cytométrie en flux, avec pour objectif de développer des méthodes d'apprentissage capables à la fois de classer les patients et d'identifier des schémas au niveau cellulaire associés à des états pathologiques. Dans ce contexte, les contributions de cette thèse s'articulent autour de plusieurs axes, chacun répondant à une limitation spécifique des approches existantes.

Avant toute modélisation, il est nécessaire de comprendre la structure statistique des données de cytométrie. Dans le chapitre 1, nous proposons une analyse exploratoire détaillée des données spécifiques à cette thèse, visant à caractériser les distributions cellulaires au niveau des patients ainsi que les structures de dépendance entre variables. Cette étape met en évidence des propriétés clés des données et fournit une base empirique pour orienter les choix de modélisation développés dans les chapitres suivants.

Dans le chapitre 2, nous proposons une approche probabiliste basée sur des modèles de mélanges gaussiens dans le cadre du MIL. Le modèle permet d'identifier explicitement des sous-populations cellulaires latentes associées à la pathologie, tout en conservant une interprétabilité directe via les composantes du mélange. Le modèle est estimé par un algorithme EM adapté.

Les approches modernes de MIL, en particulier en deep learning, reposent majoritairement sur des représentations globales des bags (via pooling ou mécanismes d'attention), ce qui limite leur capacité à fournir des explications au niveau local (des instances). Dans le chapitre 3, nous proposons un cadre de Multiple Instance Learning local, qui rompt avec ces approches globales. Notre méthode apprend des représentations au niveau des instances à l'aide d'un autoencodeur orthogonal, structure l'espace latent en sous-régions correspondant à des sous-populations cellulaires, et associe à chacune un expert spécialisé via un modèle de type *mixture of experts*.

Les approches développées dans les chapitres 2 et 3 ont fait l'objet d'une communication [MKM25] à la 56ème édition des Journées de Statistique.

Limités à différentes reprises par le nombre de patients disponibles dans les données (fournies par Metafora), nous avons exploré des méthodes pour générer des données.

Alors que l'augmentation de données est largement étudiée en apprentissage supervisé classique, elle reste encore peu explorée dans le cadre du Multiple Instance Learning, où les données sont naturellement structurées en ensembles. Dans le chapitre 4, nous proposons une méthode d'augmentation de données directement dans l'espace des distributions, en modélisant chaque bag comme une mesure empirique et en générant de nouvelles observations à l'aide de barycentres de Wasserstein. Cette approche fournit un cadre pour enrichir les jeux de données au niveau des bags et un outil intéressant dans des contextes où le nombre de patients est limité.

Les méthodes et résultats des chapitres 2, 3 et 4 ont été présentés sous forme de poster [Mik+25] dans le cadre de l'école d'automne *ML at Aussois* consacrée aux avancées récentes en apprentissage automatique.

Dans le chapitre 5, nous proposons une analyse discriminante en Multiple Instance Learning dans laquelle les bags sont modélisés comme des distributions. Contrairement aux approches naïves supposant l'indépendance des variables, notre modèle exploite la structure de dépendance entre variables mise en évidence dans le chapitre 1, afin de capturer des interactions pertinentes. Ce cadre permet de comparer directement les patients à travers leurs distributions et de fournir des contributions interprétables au niveau des cellules.

Chapter 1

Exploratory data analysis

This chapter investigates the structure of flow cytometry data through exploratory analysis. We begin with a descriptive and multivariate analysis to examine the distribution of cells at the patient level. We then use the ISDE algorithm [Puj22] to estimate the independence structure between biological variables. This analysis provides key insights and establish a foundation for the modeling strategies developed in the following chapters.

1 Description of datasets

This section introduces the datasets used in this work. We describe their structure, the type of measurements they contain, and the main challenges they present for analysis.

1.1 FlowCAP AML

Acute Myeloid Leukemia (AML) is a cancer that originates in immature blast cells produced in the bone marrow, which normally develop into blood cells. Detecting abnormalities in these cells is central to diagnosis.

This dataset is a benchmark dataset introduced in [Agh+13] as part of *Flow Cytometry: Critical Assessment of Population Identification Methods* (FlowCAP) challenge. It contains samples from 359 patients, including 43 diagnosed with AML and 316 healthy individuals. The objective associated with this dataset is to classify patients as AML or non-AML based on flow cytometry measurements. For each patient, data is represented as a matrix $X \in \mathbb{R}^d$, where m is the number of cells (ranging from 10,000 to 30,000) and $d = 7$ is the number of measured variables. Each row corresponds to a single cell, and each column to a biological marker. The variables are: FSC (Forward Scatter), SSC (Side Scatter), which describe cell size and granularity, and HLA-DR, CD117, CD45,

CD34, CD38 which are the fluorescence markers.

An important characteristic of this dataset is the absence of cell-level annotations. Labels are only available at the patient level (AML diagnosed vs healthy), which makes the problem inherently weakly supervised. From a biological perspective, no single variable is sufficient to discriminate between healthy and tumor cells. Instead, diagnosis relies on the interaction between multiple markers. This motivates the need to study dependencies between variables and to analyze the joint distribution of cell populations.

To explore these distributions, 2D histograms were constructed for each patient (see see figure 1.1). These visualizations highlight a strong variability across patients. Nevertheless, a notable pattern emerges on the pair (FSC, SSC): healthy patients often exhibit bimodal distributions, whereas AML patients tend to show a single mode. This difference may reflect the presence of abnormal cell populations and could be related to disease progression.

1.2 Multiple Myeloma

Multiple Myeloma (MM) is a cancer that affects plasma cells, a type of white blood cell produced in the bone marrow. The disease is characterized by the accumulation of abnormal plasma cells, which can form multiple lesions in different locations of the body.

The dataset used in this work originates from an ongoing clinical study conducted in collaboration with Metafora Biosystems. It includes data collected from more than 50 patients. Flow cytometry measurements provide detailed information on plasma cells, allowing the analysis of cellular abnormalities associated with the disease.

Due to experimental constraints and the evolving nature of the clinical study, the number of measured biological markers is not constant across patients. In addition, some patients are considered more relevant than others from a clinical perspective, depending on the stage and severity of the disease. As a result, the dataset has been accessed in several versions over time, all derived from the same study and involving the same underlying group of patients. To avoid ambiguity, we adopt the following approach throughout this thesis: for each numerical experiment, we explicitly specify the subset of patients used, the variables used, and the number of cells per patient.

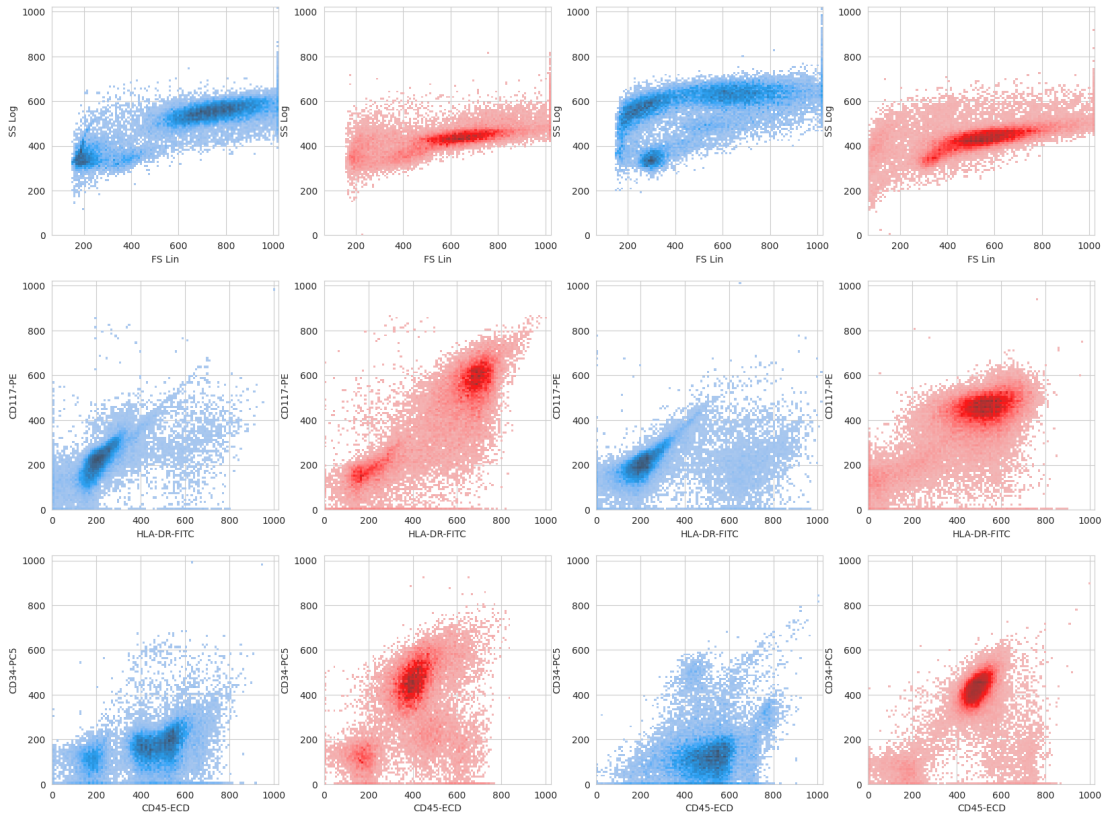


Figure 1.1: 2D histograms of cell distributions for four representative patients (two AML and two healthy). Each column corresponds to one patient, with AML patients shown in red and healthy patients in blue. Rows display three pairs of variables: (FSC, SSC), (HLA-DR, CD117), and (CD45, CD34). Each plot represents the distribution of individual cells in the corresponding 2D feature space. Notably, the (FSC, SSC) projections reveal a bimodal structure for healthy patients, whereas AML patients tend to exhibit a single dominant mode.

2 Principal component analysis

This section explores the structure of the FlowCAP AML dataset using Principal Component Analysis (PCA). The objective is to identify the main sources of variability in the data and to analyze the relationships between variables. By comparing results across patients and groups of patients, we aim to better understand inter-patient variability.

PCA on all patients. To begin, PCA was performed on all patients combined (concatenation of all datasets). The results are illustrated on figure 1.2. On the graph of individuals, each patient is represented by the center of gravity (*i.e.*, the mean of its feature values) of its point cloud. The healthy patients are represented in blue and ill patients are represented in orange. We can notice that the first component of PCA is insufficient to separate healthy patients from ill patients. The first component encodes

more than half of the information (52.45%) on the variance of the point cloud. On the graph of variables and on the first PCA plane, we notice that almost all variables are strongly contributing to the first axis.

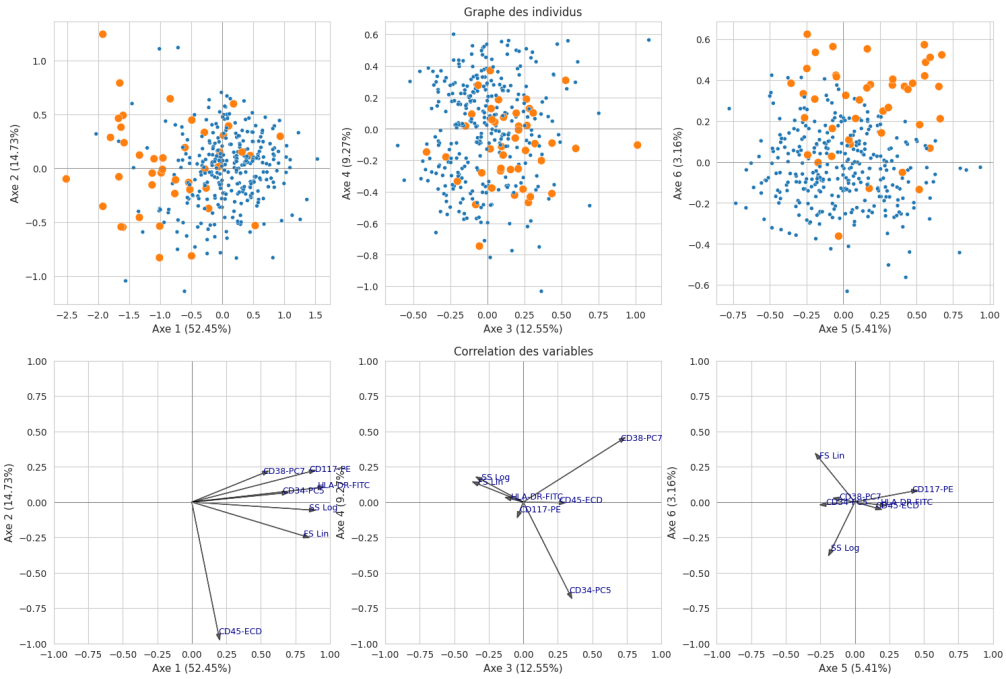


Figure 1.2: Principal Component Analysis (PCA) performed on the concatenated dataset of all patients. Top row: projection of patients onto selected principal component planes, where each point represents a patient summarized by the mean of its cell measurements. Bottom row: showing the contribution of each biological marker to the principal components. Healthy patients are shown in blue and AML patients in orange. The first principal component captures 52.45% of the variance but does not clearly separate the two groups.

PCA on healthy patients and ill patients separately. Next, two distinct PCAs were performed: one on all the healthy patients concatenated and the other on all the ill patients concatenated. The projection of variables on principal components have been superposed for each group. We notice (see figure 1.3) similar linear forms in the two projections. This similarity suggests that the structure of variables does not depend on patient status. In other words, the interactions between variables seem to maintain a relatively constant structure, independently of patient status. However, this analysis is only a first step. In the following, we will do another analysis to confirm that the interaction of variables does not depend on the health of patients.

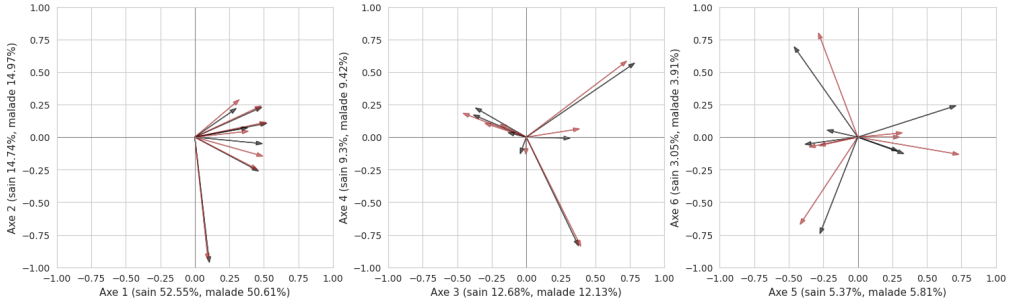


Figure 1.3: Comparison of the results obtained from PCA performed separately on healthy patients and AML patients. The projections of variables onto the principal component space are shown for each group, with healthy patients in black and AML patients in brown. The similarity between the two configurations suggests that the dependency structure between variables is preserved across patient groups.

3 Understanding variables independence structure

This section investigates the independence structure between variables in flow cytometry data. We rely on the ISDE method [Puj22] to identify groups of variables that can be considered independent and to construct density estimators based on this decomposition. This approach is motivated by the need to handle moderately high-dimensional data and to mitigate the curse of dimensionality.

3.1 Independence Structure Density Estimation

In flow cytometry, it is very common to work with dozens of variables (biological markers). In this sense, when analyzing flow cytometry data, it is possible to face various issues related related to the curse of dimensionality.

The curse of dimensionality describes various phenomena that happen when analyzing data in high dimensional spaces and that do not occur in low-dimensional settings. In statistics [Gir21] and machine learning, it refers to the exponential increase of the amount of data required n and computation time to train models as the number d of dimensions grows ($d \gg n$). The symptoms of the curse of dimensionality are diverse. In statistics, a common sign is the high bias and the high variance of estimators. This is particularly the case in non-parametric estimation when building multivariate density estimators.

When analyzing data with a number of features varying from a few ones to a few dozens, the curse of dimensionality can still occur. Indeed, as the dimension increases, data points become far apart, distances lose meaning, making it difficult to estimate densities or find clusters. This setting in which we are working is called “the moderately

high-dimensional setting". It lies between the low-dimensional and the high-dimensional settings. Making a distinction between these settings is important as emphasized in [Gir21]. Algorithms with an acceptable computation cost can be developed in the moderately high-dimensional setting. In higher dimensions, it is more complicated to achieve.

In [Puj22], a method is proposed to estimate multivariate densities in the moderately high dimension setting. The method consists in decomposing the density as a product of low-dimensional marginals by defining a partition of the original features.

More formally, let f be a density function over $\mathbb{R}^d$. Consider $\{1, \dots, d\}$ the set of features indices. We denote by $\mathcal{P}$ a partition of $\{1, \dots, d\}$ and by f_S the marginal density of f over $S \subset \{1, \dots, d\}$. We use the shorthand notation $f_S(x_S)$ to highlight the fact that f_S depends only on $x_S = (x_i)_{i \in S}$. Estimating f as a product of marginals, consists in assuming the following:

$$f(x) = \prod_{S \in \mathcal{P}} f_S(x_S) \ \forall x \in \mathbb{R}^d, \quad \text{where } f_S(x_S) = \int f(x) \prod_{i \notin S} dx_i.$$

With this formalism, Independence Structure Density Estimation (ISDE) is proposed. By decomposing the density as a product of marginals that maximizes the empirical log-likelihood, a partition of variables is then computed. This partition reflects an independence structure between groups of variables. ISDE can be used with any density estimation technique whether it is parametric or not. Here, we use a Gaussian kernel density estimator (KDE). For a given bandwidth $h > 0$ and observations $X_1, \dots, X_N$, the estimator $\hat{f}_h$ is defined as:

$$\hat{f}_h(x) = \frac{1}{N} \sum_{i=1}^{N} \frac{\exp\left(-\frac{(X_i - x)^T (X_i - x)}{2h^2}\right)}{(2\pi)^{(d/2)} h^d} \tag{1.1}$$

ISDE separates variables into independent blocks. A constraint can be set on the size of the blocks. By setting a maximal number of variables per block, the density can be decomposed as a product of low-dimensional marginals, which helps ISDE to tackle the curse of dimensionality. The resulting partition might be sub-optimal but it saves some computation time and avoid performing an exhaustive search. Let $k \in \{1, \dots, d\}$ be the maximal number of variables per block. We denote Set_d^k the set of all subsets of $\{1, \dots, d\}$ with cardinal lower than or equal to k by Part_d^k the collection of all partitions of $\{1, \dots, d\}$ constructed with elements in Set_d^k. ISDE outputs a partition of variables from Part_d^k and aims to build a density estimator as a product of marginal estimators :

$$\hat{f}_{h_{\mathcal{P}}}(x) = \prod_{S \in \mathcal{P}} \hat{f}_{S,h_S}$$

where $\mathcal{P}$ is a partition from Part_d^k, $h_{\mathcal{P}} = (h_S)_{S \in \mathcal{P}}$ is a list of bandwidths. $\hat{f}_{S,h_S}$ is an estimator of the form (1.1) but applied to S.

In order to build the best estimator $\hat{f}_{h_{\mathcal{P}}}$, the best partition is found. In [Puj22], a score is used to evaluate partitions. The score of a partition $\mathcal{P}$ is defined as the empirical log-likelihood :

$$\mathscr{L}_n(\mathcal{P}) = \sum_{S \in \mathcal{P}} \mathscr{L}_n(S) \quad \text{where} \quad \mathscr{L}_n(S) = \frac{1}{n} \sum_{i=1}^{n} \log(\hat{f}_{S,h_S}(X_i))$$

The best partition can then be found by solving an optimization problem :

$$\max_{\mathcal{P} \in \text{Part}_d^k} \mathscr{L}_n(\mathcal{P}) = \max_{\mathcal{P} \in \text{Part}_d^k} \sum_{S \in \mathcal{P}} \mathscr{L}_n(S) \tag{1.2}$$

The brute-force approach to solve the above problem is to compute the score of every partition from Part_d^k. [Puj22] proposes instead to reformulate the problem as a linear program. The solution of the linear problem is the exact of solution of (1.2). The research of the best partition in ISDE can then be done in an acceptable amount of time and it scales well with the dimension as seen in the experiments of [Puj22].

3.2 FlowCAP-AML data analysis

To better understand the interactions between biological variables in this case, we conducted an experiment based on the algorithm ISDE. This experiment consists in applying ISDE individually to each patient. By performing this operation on each patient, we can extract groups of variables reflecting the independence structure specific to each case. Comparing those groups will help us identify similarities and/or differences existing between patients.

The experiment goes as follows:

1. For each patient, creation of train and test sets of sizes respectively 5000 and 2000 cells.

2. ISDE is executed in order to perform a density estimation by partitioning variables. No constraint is imposed on the size of groups of variables ($d = 7$). An exhaustive search is then made on the set of all variables partitions. In order to construct

different estimators, the bandwidth is chosen by cross-validation (among 200 values tested between 0.001 and 1).

3. For each patient (359 overall), the partition with the best score on the test set is selected.

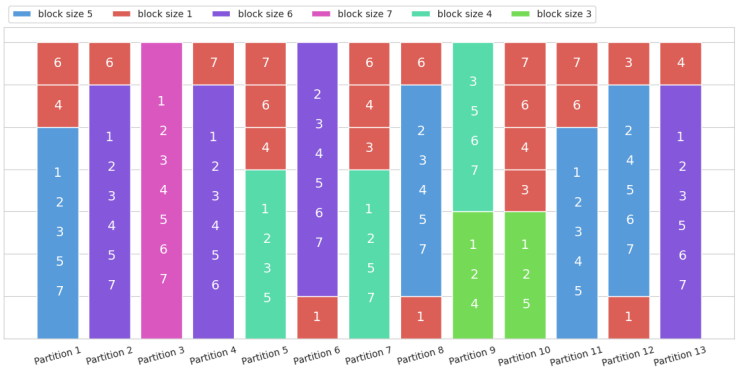
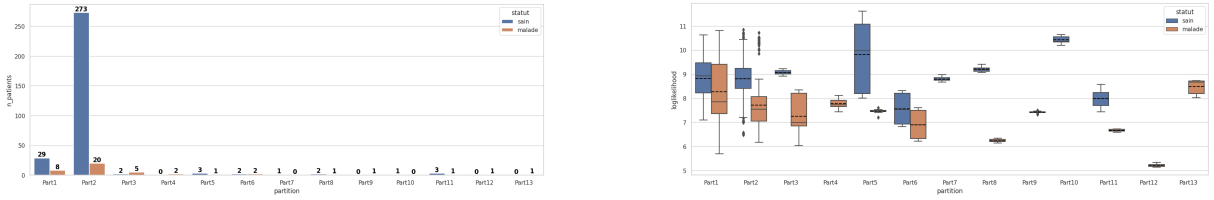


Figure 1.4: Variable partitions identified by the ISDE algorithm across all patients. Each partition groups variables into blocks assumed to be mutually independent, with colors indicating different blocks. 13 distinct partition structures are observed.

All patients combined, 13 different partitions of variables have been obtained (see figure 1.4). We notice that the second partition was the most selected as the best-performing partition (see figure 1.5a). Whether patients are healthy or diseased, partition 2 seems to reflect the best independence structure between variables. In the previous analysis with PCA (section 2), we noticed that variables structure on healthy patients is similar to their structure on ill patients. So this experiment with ISDE confirms this finding. In other words, if we want to create a probabilistic model, the assumptions on variables can be made regardless of patient status. This conclusion holds only for the FlowCAP-AML data.



(a) (b)

Figure 1.5: Evaluation of ISDE partitions on the FlowCAP AML dataset. (a) Number of patients for which each partition is selected as the best according to the likelihood criterion. (b) Log-likelihood scores of the partitions evaluated using cross-validation. Partition 2 is most frequently selected and achieves strong performance across validation sets, suggesting a consistent independence structure across patients.

Conclusion of FlowCAP data analysis

The ISDE analysis on the FlowCAP AML dataset reveals a consistent pattern: despite strong variability in cell distributions across patients, the independence structure between variables remains stable. A limited number of partition structures are identified, with one dominant partition selected for most patients, regardless of their status. This result suggests that differences between patients are mainly driven by changes in the distribution of cells rather than by changes in the relationships between variables. The dependency structure between markers can therefore be considered stable across individuals. This observation motivates a modeling approach in which patient data are represented through distributions of cells, while relying on a shared dependency structure between variables. This perspective is developed later in chapter 5, where we introduce a Bayesian classification method in a Multiple Instance Learning.

Chapter 2

A Gaussian mixture-based approach

Using flow cytometry data, each patient can be represented by a collection of cells, while labels are available only at the patient level through clinical diagnoses. The main difficulty is that tumor cells do not follow a single, well-defined pattern: their behavior varies across patients, and disease complexity makes their identification challenging. This raises the following question: by comparing healthy and ill patients, can we identify cell-level patterns that explain the disease and support patient classification? This setting naturally fits within the Multiple Instance Learning (MIL) framework, where patients are modeled as bags of instances. In this chapter, we propose a probabilistic approach based on Gaussian mixture models, in which latent components aim to capture underlying cell populations associated with the disease. The model is estimated using an Expectation-Maximization (EM) algorithm. Before introducing our method, we provide an overview of existing MIL approaches.

1 Overview of Multiple Instance Learning methods

1.1 Problem formulation and label ambiguity

In many real-world applications, observations can be organized into sets, while labels are only available at the level of the set. Multiple Instance Learning (MIL), introduced by [DLL97], provides a framework for learning in this context. In MIL, each observation is called an *instance* and a set of instances is a *bag*. Formally, we consider a dataset $\{(\mathcal{X}_i, Y_i);\ 1 \leq i \leq n\}$, where a bag $\mathcal{X}_i = \{X_j^{(i)} \in \mathbb{R}^d;\ 1 \leq j \leq m_i\}$, and $Y_i \in \{0, 1\}$. The goal of a MIL model is to predict the label of a bag with the least possible error.

Multiple Instance Learning differs from traditional supervised learning [RC05; Alp+15] by relaxing the instance–label correspondence. While traditional supervised

learning assumes that each instance is associated with a known label, MIL provides supervision only at the bag level, leaving instance labels latent. This induces an inherent ambiguity because we don't know which instances are responsible for generating of the bag label. As a result, classifying bags is both a classification and a latent structure discovery problem.

The bag label ambiguity is one of the main characteristics of a MIL problem and has been emphasized by many authors[Car+18; Amo13; Bab08]. A bag label typically indicates the presence of a discriminative pattern at the instance level without specifying which instances are responsible for it. For example, in [CDP19], this situation can occur if the radiologist only labeled an entire CT scan (medical imaging) as containing tumor regions or not, but has not indicated its locations. Most patches of the image are normal, only a few regions are diseased and we don't know which ones. In such settings, the relationship between individual observations and bag labels depends a lot on the context and the application.

1.2 MIL assumptions and categorization of methods

A central component of any MIL model is the assumption linking instances to the bag label. Let us assume that instances $X_j^{(i)}$ have unknow labels $Z_j^{(i)}$. The most common assumption is the standard MIL assumption introduced in [DLL97], where a bag is positive if at least one of its instances is positive:

$$Y_i = \begin{cases} 1 & \text{if} \quad \exists\, j,\ Z_j^{(i)} = 1 \\ 0 & \text{else} \end{cases}$$

The relationship between instances and bag label is generally defined through an aggregation function Ψ. The standard assumption corresponds to a max operator: $Y_i = \Psi(Z_j^{(1)}, \ldots, Z_j^{(m_i)}) = \max_j Z_j^{(i)}$. While simple and intuitive, it may be too restrictive for many applications. Alternative assumptions have been proposed [FF10]: Soft or probabilistic assumptions, where instances contribute with different individual scores (e.g., noisy-OR [ML97]), and collective assumptions [Xu03], where the bag label depends on the overall distribution of instances rather than a single one. These assumptions reflect different views on how information is distributed within bags. The choice of the aggregation function Ψ has then a direct impact on model behavior.

MIL methods can be categorized according to how they model the relationship between instances and bags.There are two principal types of MIL methods in the literature. First, methods that create a classifier in the space of instances. Second, methods that

directly operate in the bags space. In the literature, these approaches are often referred to as instance-level and bag-level methods. In this thesis, we use the terms local and global methods, respectively, as they better reflect the modeling perspective adopted throughout.

The idea behind local or instance-level methods, is to discriminate instances first and then deduce a decision rule for classifying bags. Formally, A decision function f is first defined on instances. An aggregation function Φ is then used to combine instances scores. The final bag decision function F is a composition of Φ and f. A MIL model is said to be local if there exist $f : \mathbb{R}^d \to \mathbb{R}$ and $\Phi : \mathbb{R}^m \to \{0, 1\}$ such that for any bag $\mathcal{X} = \{x_1, \cdots, x_m\}$, $F(\mathcal{X}) = \Phi(f(x_1), \cdots, f(x_m))$.

In practice, many local methods adapt standard supervised learning techniques by treating instances as if they were labeled, often through some form of bag label inheritance or iterative refinement. The key idea is that bag classification relies on detecting informative instances. Many classical supervised methods have been adapted within this framework. For example, mi-SVM [ATH02], a MIL version of Support Vector Machines, iteratively refines instance labels during training. Boosting-based approaches [Bab+08] have also been proposed, by modifying the traditional objective function to account for bag structure, and logistic regression models have been extended to the MIL setting [XF04]. There are also methods [ZG01] falling into this category, that introduce latent variables to model instance labels and optimize the model iteratively. Local methods are particularly relevant when the goal is to identify localized (instance level) patterns within bags, which is the perspective adopted in this work.

In contrast, global methods operate directly at the bag level without explicitly modeling instance-level decisions. In the local approach, the bag decision function breaks down into instance-wise evaluation but global approach does not admit such factorization. A MIL model is said to be global if its decision function cannot be decomposed in the above form (local method), and instead operates directly on bags, typically via a representation $F(\mathcal{X}) = g(\Psi(\mathcal{X}))$ where Ψ is defined in the bag space. Typically, global methods define distances between bags or use embeddings to represent bags in a feature space in order to compare them. In [WZ00] for example, authors proposed a version of kNN (k-Nearest Neighbors) for MIL by computing the Hausdorff distance between bags. In [CTL15], bags are represented as vector of dissimilarities and SVM is used to classify them. We will discuss this family of methods in more detail later in chapter 2.

1.3 Theoretical analysis of MIL and a probabilistic perspective

Despite the wide range of MIL methods, there are very few theoretical studies of MIL in the literature. MIL models have inherently a two-level complexity: How much do they perform with respect to the number of bags? How many instances per bag do we need to reach a satisfying robustness? This complexity makes it hard to make a theoretical analysis. [BK98] shows that learning from multiple-instance examples is reducible to probably approximately correct (PAC) learning with two sided noise. In [ST12], the authors proved that the sample complexity of MIL is mildly (poly-logarithmically) dependent on the size of the bag for any kind of hypothesis on bag-instances relationship. In [DR14], the authors made a theoretical analysis of a MIL version of SVM. In [DR16], a unifying theoretical framework is proposed for MIL methods. In this framework, bags are treated as latent distributions from which instances are observed. Finally, more recently, the authors of [JK24] proposed a theoretical framework based on PAC learning theory to assess if MIL methods combined with deep learning, can learn instance-level patterns.

In this thesis, we are more focused on the modeling aspects than the theoretical aspects of MIL. Especially, the approach taken in this chapter is probabilistic and aims to model posterior probabilities of bags. Formally, the bag decision function is defined as:

$$F(\mathcal{X}_i) = P(Y_i = 1 \mid \mathcal{X}_i) = P(Y_i = 1 \mid X_1^{(i)}, \ldots, X_{m_i}^{(i)})$$

This formulation can offer several advantages. It can provide a natural way to handle the bag label ambiguity, by decomposing the bag probability and modeling the uncertainty over instances. The model can then reason in terms of probabilities. It also connects MIL to classical statistical modeling, in particular mixture models and latent variable methods, for which well-established inference techniques exist. Finally, it can provide an interpretable framework which is particularly important for the tumor cells detection problem we are trying to solve.

2 A baseline model with Gaussian mixture

We propose to model the distribution of cells with a two-component Gaussian mixture. By considering that a healthy patient have only healthy cells and ill patients have a mixture of healthy and tumor cells, each component of the mixture represents cells that have the same state. Healthy patients are then modeled with a single Gaussian distribution, and ill patients are modeled with two Gaussians distributions. In this baseline model, we assume that the distribution of healthy cells is the same across all the patients. It is a strong assumption because there is actually a variability at the patient

level. The simplicity of this model makes it fit as a baseline.

Two MIL methods [ZG01; Li+13] proposed in the literature have a similar approach to the baseline model. In [ZG01], an EM algorithm is proposed to optimize a function called the *diverse density*. [Li+13] defines a two-component mixture model, uses a supervised classifier to estimate posterior probabilities of instances and proposes decision rules for bags. Contrary to the baseline model we are proposing, these methods do not explicitly modeled the joint distribution $\mathbf{P}(X, Y)$, which distinguishes it from a fully generative approach.

In this section, I present the methods [ZG01; Li+13] and introduce a MIL model based on a Gaussian mixture.

2.1 Diverse density

The term *diverse density* was first mentioned and defined in [ML97]. The authors reformulated the MIL classification problem as follows: finding a concept that is common to all positive bags. A concept is a pattern at the instance level that plays an important role in the discrimination of bags. The goal was to find a discriminative pattern by searching for regions where positive bags intersect. To achieve that, it is necessary to quantify how much a point is close to positive bags and how far it is from negative bags. In order to define a model, it is also important to give the nature of the desired concept: is it a probability distribution over instances or something more specific like a vector? [ML97] made the assumption that the concept is a region of the space centered in a target point t. The closer an instance is to t, the more likely it belongs to the concept. From there, standard assumption is proposed: a bag $\mathcal{X}_i$ is classified positive ($Y_i = 1$) if at least one of its instances $X_j^{(i)}$ is close enough to t. The authors proposed a function called *diverse density* to measure how well a point t represents the desired concept. It is defined as follows:

$$\mathrm{DD}(t) = \prod_{i=1}^{n} \left(1 - \prod_{j}(1 - f(t; X_j^{(i)})\right)^{Y_i} \left(\prod_{j}(1 - f(t; X_j^{(i)}))\right)^{(1-Y_i)} \tag{2.1}$$

where $f(t; X_j^{(i)}) = \exp(-\|X_j^{(i)} - t\|_2^2)$ is a similarity function.

The above formulation is based on the Noisy-OR model [ZMP16], which uses an approximation of the logical-OR operator. For positive bags, if at least one instance $X_j^{(i)}$ is close to t, *i.e* $f(t; X_j^{(i)}) \approx 1$, then $1 - \prod_j (1 - f(t; X_j^{(i)}) \approx 1$.

The naming *diverse density* may suggest that a density estimation is performed but there is no such estimation in the proposed method. The authors used a classical gradient descent for the optimization part. They added a scaling parameter s to the similarity function $f(t, s; X_j^{(i)}) = \exp(-\|X_j^{(i)} - t\|_2^2/s^2)$.

In [Doo+02], the authors proposed an extension of diverse density for real valued labels. They were trying to solve the drug activity problem where the goal is to predict whether a molecule can bind to a specific protein or not. Using real valued labels in this case, allows to quantify the binding strength. Formally, for bag labels $Y_i \in [0, 1]$, the diverse density becomes:

$$\mathrm{DD}_2(t) = \prod_{i=1}^{n}(1 - |Y_i - F(t; \mathcal{X}_i)|) \tag{2.2}$$

where $F(t; \mathcal{X}_i) = \max_j \exp(-\|X_j^{(i)} - t\|_2^2)$.

In the continuity of [Doo+02], [ZG01] proposed an EM algorithm to optimize the log-diverse density defined below:

$$L(t; \mathcal{X}, Y) = \sum_{i=1}^{n} \log(1 - |Y_i - F(t; \mathcal{X}_i)|)$$

Although not originally formulated this way, it is possible to rewrite the above function by assigning a latent variable to each instance $X_j^{(i)}$. The deciding factor for a bag $\mathcal{X}_i$ is how close its instances are to t. We can consider the closest instance $X_j^{(i)}$ as a representative of the bag. Let us denote by $Z_j^{(i)}$ the latent variable indicating that an instance is the representative. The log-diverse density can be rewritten as follows:

$$L(t; \mathcal{X}, Y, \mathcal{Z}) = \sum_{i=1}^{n}\left(\sum_j Z_j^{(i)} \log(1 - |Y_i - f(t; X_j^{(i)})|)\right)$$

$$Z_j^{(i)} = \begin{cases} 1 & \text{if} \quad j = \operatorname{argmin}_{X_l^{(i)} \in \mathcal{X}_i} \|X_l^{(i)} - t\|_2 \\ 0 & \text{else} \end{cases}$$

where $\mathcal{Z}$ is the set of latent variables.

The EM algorithm proposed by [ZG01] has two steps:

- E-step: For each bag, choose a representative

$$Z_j^{(i)} = \mathbf{1}_{\{j = \operatorname{argmin}_l \|X_l^{(i)} - t\|_2\}}$$

- M-step:

$$t = \text{argmax}_t \sum_{i=1}^{n} \sum_{j} Z_j^{(i)} \log \left(1 - \left|Y_i - \exp(-\|X_j^{(i)} - t\|^2)\right|\right)$$

The prediction step is done by thresholding $F(t; \mathcal{X}_i)$. The authors proposed by default: $\hat{Y}_i = \mathbf{1}_{\{F(t;\mathcal{X}_i)>0.5\}}$.

2.2 A MIL method by combining classifier predictions

In this section, I present the MIL method proposed by [Li+13]. Inspired by the work of [Kit+02], the authors proposed a way to reformulate the problem of classifying bags as a combining problem using the predictions of a supervised classifier. Their approach is based on the assumption that the instances come from the mixture of a concept distribution and a non-concept distribution. A concept distribution refers to the distribution of instances that allows to discriminate positive bags from negative bags. Unlike the geometric notion of a concept used in diverse density, here the concept is modeled as a probability distribution.

Let's consider a set of annotated bags $\{(\mathcal{X}_i, Y_i), 1 \leq i \leq n\}$ where $Y_i \in \{0, 1\}$. Each bag is a set of instances: $\mathcal{X}_i = \{X_1^{(i)}, \ldots, X_{m_i}^{(i)};\ X_j^{(i)} \in \mathbb{R}^d\}$. Let's suppose that the instances are drawn independently from a law of density f_0 for $Y_i = 0$ and from a law of density f_1 for $Y_i = 1$. We have:

$$\begin{cases} \mathbf{P}(\mathcal{X}_i|Y_i = 1) &= \prod_{j=1}^{m_i} f_1(X_j^{(i)}) \\ \mathbf{P}(\mathcal{X}_i|Y_i = 0) &= \prod_{j=1}^{m_i} f_0(X_j^{(i)}) \end{cases}$$

Next, let's suppose that the instances allowing to discriminate bags are from a certain concept. Let's denote by Z the variable indicating if the instances belong to a concept. These instances are drawn from a law of density f^C. The instances that do not belong to the concept are drawn from a probability distribution with density $f^{\overline{C}}$. Let's denote by α the proportion of instances belonging the concept in a positive bag. We have:

$$\begin{cases} f_0(x) = f^{\overline{C}}(x) \\ f_1(x) = \alpha f^C(x) + (1 - \alpha) f^{\overline{C}}(x); \quad 0 < \alpha < 1 \end{cases} \tag{2.3}$$

The expression of the density f^C can be written like this:

$$f^C(x) = \frac{f_1(x) - (1 - \alpha) f_0(x)}{\alpha} \tag{2.4}$$

Let β be the proportion of positive bags. We have:

$$\begin{cases} \mathbf{P}(Y_i = 1) & = \beta \\ \mathbf{P}(Z_j^{(i)} = 1) & = \alpha\beta \end{cases}$$

[Li+13] proposed some decision rules to classify bags. The whole idea is to make $\mathbf{P}(Y_i = 1|X_j^{(i)})$ appear in the formulas. $\mathbf{P}(Y_i = 1|X_j^{(i)})$ can be estimated by a supervised classifier after assigning a bag label Y_i to its instances $X_j^{(i)}$.

Product rule

Since a bag $\mathcal{X}_i$ is a set of instances, we have $\mathbf{P}(Y_i|\mathcal{X}_i) = \mathbf{P}(Y_i|X_1^{(i)}, \ldots, X_{m_i}^{(i)})$. By using Bayes formula, $\mathbf{P}(Y_i = 1|\mathcal{X}_i) \geq \mathbf{P}(Y_i = 0|\mathcal{X}_i)$ is then equivalent to:

$$\mathbf{P}(X_1^{(i)}, \ldots, X_{m_i}^{(i)}|Y_i = 1)\mathbf{P}(Y_i = 1) \geq \mathbf{P}(X_1^{(i)}, \ldots, X_{m_i}^{(i)}|Y_i = 0)\mathbf{P}(Y_i = 0)$$

Next, we can use the independence between each instance of a bag to factorize the conditional probabilities, and by applying Bayes' formula one time again to each $\mathbf{P}(X_j^{(i)}|Y_i)$, we obtain the expression below:

$$\frac{\prod_{j=1}^{m_i} \mathbf{P}(Y_i = 1|X_j^{(i)})\mathbf{P}(X_j^{(i)})}{\mathbf{P}(Y_i = 1)^{m_i}}\mathbf{P}(Y_i = 1) \geq \frac{\prod_{j=1}^{m_i} \mathbf{P}(Y_i = 0|X_j^{(i)})\mathbf{P}(X_j^{(i)})}{\mathbf{P}(Y_i = 0)^{m_i}}\mathbf{P}(Y_i = 0)$$

Finally, we can simplify the expression and replace $\mathbf{P}(Y_i = 1)$ by β. The product rule is then obtained:

$$\mathrm{H}_1(\mathcal{X}_i) = \begin{cases} 1 & \text{if } \beta^{(1-m_i)} \prod_j \mathbf{P}(Y_i = 1|X_j^{(i)}) \geq (1-\beta)^{(1-m_i)} \prod_j \mathbf{P}(Y_i = 0|X_j^{(i)}) \\ 0 & \text{else} \end{cases} \tag{2.5}$$

To apply this rule in practice it is necessary to estimate $\mathbf{P}(Y_i = 1|X_j^{(i)})$. In [Li+13], instances inherit the labels of their bag. In other words, for a positive bag, all the instances are considered positive (and the other way around for a negative bag). This creates instances labels with an asymmetric noise. A supervised classifier is then trained with the pairs $(X_j^{(i)}, Y_i)$. Finally, to predict the label of a bag, the product rule can be applied by combining the predictions outputted by the classifier. Unfortunately, this rule is very sensitive to estimation errors.

Sum rule

Let us assume that $X_j^{(i)}$ are so noisy that the posteriors are close enough to priors.

Let $\delta_{ij} > 0$ such that $\delta_{ij} \ll 1$ and $\mathbf{P}(Y_i|X_j^{(i)}) = (1+\delta_{ij})\mathbf{P}(Y_i)$. By replacing $\mathbf{P}(Y_i|X_j^{(i)})$ in the expression of the product of probabilities, we have:

$$\begin{aligned}\prod_{j\geq 1}\mathbf{P}(Y_i|X_j^{(i)}) &= \prod_{j\geq 1}[(1+\delta_{ij})\mathbf{P}(Y_i)] \\ &= \mathbf{P}(Y_i)^{m_i}\left(1+\sum_{j\geq 1}\delta_{ij}+\cdots\right)\end{aligned}$$

By neglecting the terms of order greater or equal to 2, we obtain the following approximation: $\prod_{j\geq 1}\mathbf{P}(Y_i|X_j^{(i)}) \approx \mathbf{P}(Y_i)^{m_i} + \mathbf{P}(Y_i)^{m_i}\sum_{j\geq 1}\delta_{ij}$. By applying this approximation in the expression of the production rule, we obtain:

$$\begin{aligned}\mathbf{P}(Y_i)^{(1-m_i)}\prod_{j\geq 1}\mathbf{P}(Y^{(i)}|X_j^{(i)}) &= \mathbf{P}(Y_i)+\mathbf{P}(Y_i)\sum_{j=1}^{m_i}\delta_{ij} \\ &= \mathbf{P}(Y_i)+\sum_{j=1}^{m_i}\mathbf{P}(Y_i|X_j^{(i)})-\sum_{j=1}^{m_i}\mathbf{P}(Y_i) \\ &= (1-m_i)\mathbf{P}(Y_i)+\sum_{j=1}^{m_i}\mathbf{P}(Y_i|X_j^{(i)})\end{aligned}$$

We deduce then the sum rule as an approximation of the product rule:

$$\mathrm{H}_2(\mathcal{X}_i) = \begin{cases} 1 & \text{if } \beta(1-m_i)+\sum_{j=1}^{m_i}\mathbf{P}(Y_i=1|X_j^{(i)}) \geq \\ & \quad (1-\beta)(1-m_i)+\sum_{j=1}^{m_i}\mathbf{P}(Y_i=0|X_j^{(i)}) \\ 0 & \text{else}\end{cases} \tag{2.6}$$

The sum rule is used in practice in the same manner as the product rule by training a supervised classifier.

It is possible to rewrite the sum rule with respect to $\mathbf{P}(Z_j^{(i)}|X_j^{(i)})$. First, we find the relation between $\mathbf{P}(Y_i|X_j^{(i)})$ and $\mathbf{P}(Z_j^{(i)}|X_j^{(i)})$.

$$\begin{aligned}\mathbf{P}(Y_i=1|X_j^{(i)}) &= \frac{\mathbf{P}(Y_i=1)\mathbf{P}(X_j^{(i)}|Y_i=1)}{\mathbf{P}(X_j^{(i)})} \\ &= \frac{\mathbf{P}(Y_i=1)f_1(X_j^{(i)})}{\mathbf{P}(X_j^{(i)})}\end{aligned}$$

$$= \frac{\beta(\alpha f^{C}(X_j^{(i)}) + (1-\alpha) f^{\overline{C}}(X_j^{(i)}))}{\mathbf{P}(X_j^{(i)})}$$

The last equality is obtained by using (2.3). In the same manner, we have:

$$\mathbf{P}(Y_i = 0|X_j^{(i)}) = \frac{(1-\beta) f^{\overline{C}}(X_j^{(i)})}{\mathbf{P}(X_j^{(i)})}$$

In order to write the expression of $\mathbf{P}(Y_i|X_j^{(i)})$ with respect to $\mathbf{P}(Z_j^{(i)}|X_j^{(i)})$, we need to substitute f^C and $f^{\overline{C}}$ in the formulas above. To do so, we can notice that:

$$\mathbf{P}(Z_j^{(i)} = 1|X_j^{(i)}) = \frac{\mathbf{P}(X_j^{(i)}|Z_j^{(i)} = 1)\mathbf{P}(Z_j^{(i)} = 1)}{\mathbf{P}(X_j^{(i)})} = \alpha\beta \frac{f^{C}(X_j^{(i)})}{\mathbf{P}(X_j^{(i)})}$$

$$\mathbf{P}(Z_j^{(i)} = 0|X_j^{(i)}) = \frac{\mathbf{P}(X_j^{(i)}|Z_j^{(i)} = 0)\mathbf{P}(Z_j^{(i)} = 0)}{\mathbf{P}(X_j^{(i)})} = (1-\alpha\beta) \frac{f^{\overline{C}}(X_j^{(i)})}{\mathbf{P}(X_j^{(i)})}$$

We can finally write:

$$\begin{cases} \mathbf{P}(Y_i = 1|X_j^{(i)}) &= \mathbf{P}(Z_j^{(i)} = 1|X_j^{(i)}) + \frac{\beta-\alpha\beta}{1-\alpha\beta}\mathbf{P}(Z_j^{(i)} = 0|X_j^{(i)}) \\ \mathbf{P}(Y_i = 0|X_j^{(i)}) &= \frac{1-\beta}{1-\alpha\beta}\mathbf{P}(Z_j^{(i)} = 0|X_j^{(i)}) \end{cases}$$

By replacing $\mathbf{P}(Y_i|X_j^{(i)})$ in the expression of the sum rule (2.6), we can derive an equivalent rule:

$$\mathrm{H}_3(\mathcal{X}_i) = \begin{cases} 1 & \text{if } \beta(1-m_i) + \sum_j \mathbf{P}(Z_j^{(i)} = 1|X_j^{(i)}) + \frac{\beta-\alpha\beta}{1-\alpha\beta}\mathbf{P}(Z_j^{(i)} = 0|X_j^{(i)}) \geq \\ & \quad (1-\beta)(1-m_i) + \sum_j \frac{1-\beta}{1-\alpha\beta}\mathbf{P}(Z_j^{(i)} = 0|X_j^{(i)}) \\ 0 & \text{else} \end{cases} \tag{2.7}$$

The rule obtained above is especially interesting when we define a probabilistic model on the data and estimate $\mathbf{P}(Z_j^{(i)}|X_j^{(i)})$. We will see in the following (see section 2.3) that it can be applied to our baseline model.

Gamma rule

In [Li+13], no assumption is made on the concept distribution $\mathbf{P}(X_j^{(i)}|Z_j^{(i)} = 1)$. It is then difficult to estimate correctly the posteriors $\mathbf{P}(Z_j^{(i)}|X_j^{(i)})$. The authors proposed an alternative by supposing roughly that $\mathbf{P}(Z_j^{(i)} = 1|X_j^{(i)}) = 1$ if $Z_j^{(i)} = 1$. Despite this assumption, it is still necessary to have a predictive model on $Z_j^{(i)}$. As mentioned above, a supervised classifier is used to estimate $\mathbf{P}(Y_i|X_j^{(i)})$. From this point, the approach proposed to discriminate the concept from the non-concept (*i.e* predict $Z_j^{(i)}$) is the one below.

In theory, we predict $Z_j^{(i)} = 1$ if we have $\mathbf{P}(Z_j^{(i)} = 1|X_j^{(i)}) \geq \mathbf{P}(Z_j^{(i)} = 0|X_j^{(i)})$. By using Bayes formula, we obtain $\mathbf{P}(X_j^{(i)}|Z_j^{(i)} = 1)\mathbf{P}(Z_j^{(i)} = 1) \geq \mathbf{P}(X_j^{(i)}|Z_j^{(i)} = 0)\mathbf{P}(Z_j^{(i)} = 0)$ from which we can deduce the following inequality:

$$\alpha\beta f^C(X_j^{(i)}) \geq (1 - \alpha\beta) f^{\overline{C}}(X_j^{(i)})$$

By substituting f^C and $f^{\overline{C}}$ with the relations in (2.3) and (2.4), we obtain:

$$f_1(X_j^{(i)}) \geq (\frac{1}{\beta} + 1 - 2\alpha) f_0(X_j^{(i)})$$

Finally, by noticing that $f_1(X_j^{(i)}) = \mathbf{P}(X_j^{(i)}|Y_i = 1)$ and $f_0(X_j^{(i)}) = \mathbf{P}(X_j^{(i)}|Y_i = 0)$ and using Bayes' formula one last time, we deduce:

$$\mathbf{P}(Y_i = 1|X_j^{(i)}) \geq \frac{1 + \beta - 2\alpha\beta}{1 - \beta} \mathbf{P}(Y_i = 0|X_j^{(i)}) \tag{2.8}$$

The inequality above is the condition used to classify an instance $X_j^{(i)}$ as belonging to the concept (*i.e* $Z_j^{(i)} = 1$). Now that the method to classify instances is defined, we can go back to the classification of bags. As mentioned just before, the authors in [Li+13] supposed roughly that $\mathbf{P}(Z_j^{(i)} = 1|X_j^{(i)}) = 1$ when they predict $Z_j^{(i)} = 1$. It is then possible to replace $\mathbf{P}(Z_j^{(i)} = 1|X_j^{(i)})$ by $\mathbf{1}_{\{Z_j^{(i)}=1\}}$ in the expression of the sum rule. Let $\gamma_i = \frac{1}{m_i} \sum_j \mathbf{1}_{\{Z_j^{(i)}=1\}}$ be the proportion of instances belonging to the concept in a bag. The sum rule can be rewritten like this:

$$\mathrm{H}_4(\mathcal{X}_i) = \begin{cases} 1 & \text{if } \gamma_i \geq \frac{(1-\alpha\beta)(1-2\beta)}{2(1-\beta)m_i} + \alpha\beta \\ 0 & \text{else} \end{cases} \tag{2.9}$$

The rule obtained above is the γ-rule. The estimation of parameters α and β remains a problem to solve in order to obtain a robust decision rule. β can be easily estimated like the proportion of positive bags. As for α, it's complicated given that we don't have any prior information about the bags composition. The authors proposed to estimate α as below:

$$\begin{cases} \hat{\alpha} & = \frac{1}{m_i} \sum_{j=1}^{m_i} \mathbf{1}\{\mathbf{P}(Y_i = 1|X_j^{(i)}) > \tau\} \\ & \text{where } \tau = \frac{1}{2}(1 - \theta) + \frac{1}{2} \max_{i,j,Y_i=1} \mathbf{P}(Y_i = 1|X_j^{(i)}) \\ & \text{and } \theta = \frac{1}{n_0} \sum_{i,j,Y_i=0} \mathbf{P}(Y_i = 0|X_j^{(i)}) \end{cases}$$

where n_0 is the sum of the number of instances in negative bags.

These are the steps of the MIL algorithm proposed in [Li+13]:

- Assign to each instance the label of its bag
- Train a supervised classifier to obtain the posteriors $\mathbf{P}(Y_i|X_j^{(i)})$
- Estimate the parameters α and β
- Classify the instances with the rule (2.8)
- Classify the bags with the γ-rule (or sum rule or product rule)

2.3 Our baseline model

Let us consider that each patient is a bag $\mathcal{X}_i$ which cells are instances $X_j^{(i)}$ drawn from a mixture of two Gaussians, one representing the component of healthy cells, and the other the tumor cells. The components proportions vary from one bag to another. Each bag has a label Y_i which is the patient state. Instances labels (healthy or tumoral) are considered like latent variables. Negative bags are represented by a single Gaussian distribution and positive bags are represented as a 2-component Gaussian Mixture.

$$\begin{cases} X_j^{(i)} \sim f_0 & \text{if} \quad Y_i = 0 \\ X_j^{(i)} \sim (1-\pi_i)f_0 + \pi_i f_1 & \text{if} \quad Y_i = 1 \end{cases}$$

where f_0 and f_1 are the density functions of respectively $\mathcal{N}(\mu_0, \Sigma_0)$ and $\mathcal{N}(\mu_1, \Sigma_1)$, π_i is the proportion of tumor cells in a positive bag $\mathcal{X}_i$. By definition, $\pi_i = 0$ for negative bags.

Let's denote by $Z_j^{(i)}$ the latent variable for an instance $X_j^{(i)}$, that encodes the membership of an instance to a component of the mixture. The proportion of tumor cells in a bag is then $\mathbf{P}(Z_j^{(i)} = 1) = \pi_i$. We denote by $\mathcal{Z}$ the set of latent variables of all bags. The log-likelihood of the model is defined as:

$$\begin{aligned} L(\mathcal{X}, \mathcal{Z}; Y; \Theta) = \sum_{i=1}^{n}(1-Y_i)\left(\sum_{j=1}^{m_i}\log f_0(X_j^{(i)};\theta_0)\right) \\ + Y_i\left(\sum_{j=1}^{m_i} Z_j^{(i)}\log\Big(\pi_i f_1(X_j^{(i)};\theta_1)\Big) + (1-Z_j^{(i)})\log\Big((1-\pi_i)f_0(X_j^{(i)};\theta_0)\Big)\right) \end{aligned}$$

where $\theta_0 = (\mu_0, \Sigma_0)$; $\theta_1 = (\mu_1, \Sigma_1)$ and $\Theta = ((\pi_i)_{1\leq i\leq n}, \theta_0, \theta_1)$.

The EM algorithm [DLR77] is well adapted to estimate the parameters of the model. The principle is to optimize iteratively $Q(\Theta;\Theta^{(c)})$, the expectation of the complete log-likelihood conditionally to observations and for a current parameter. Hence the succession of two steps: compute the expectation (E-step) and maximize the expectation (M-step). In our case, we propose to write a new EM algorithm to estimate the parameters of the model.

$$\begin{aligned} Q(\Theta;\Theta^{(c)}) &= \mathbb{E}_{Z\sim p(\cdot|\mathcal{X};Y;\Theta^{(c)})}[L(\mathcal{X},\mathcal{Z};Y;\Theta)] \\ &= \sum_{i=1}^{n}(1-Y_i)\left(\sum_{j=1}^{m_i}\log f_0(X_j^{(i)};\theta_0)\right) \\ &\quad + Y_i\left(\sum_{j=1}^{m_i}T_j^{(i)}\log\Big(\pi_i f_1(X_j^{(i)};\theta_1)\Big) + (1-T_j^{(i)})\log\Big((1-\pi_i)f_0(X_j^{(i)};\theta_0)\Big)\right) \end{aligned}$$

$T_j^{(i)}$ the conditional membership probability, is defined as:

$$\begin{aligned} T_j^{(i)} &= \mathbf{P}(Z_j^{(i)}=1 \mid X_j^{(i)}) \\ &= \frac{\pi_i f_1(X_j^{(i)};\theta_1)}{\pi_i f_1(X_j^{(i)};\theta_1)+(1-\pi_i)f_0(X_j^{(i)};\theta_0)} \end{aligned}$$

E-step:

$$T_j^{(i)} = \begin{cases} \mathbf{P}(Z_j^{(i)}=1|X;\Theta^{(c)}) & \text{if} \quad Y_i=1 \\ 0 & \text{else} \quad Y_i=0 \end{cases}$$

M-step:

$$\pi_i^{(c+1)} = \begin{cases} \frac{1}{m_i}\sum_{j=1}^{m_i}T_j^{(i)} & \text{if} \quad Y_i=1 \\ 0 & \text{else} \quad Y_i=0 \end{cases}$$

$$\mu_0^{(c+1)} = \frac{\sum_{i=1}^{n}Y_i\Big(\sum_{j=1}^{m_i}(1-T_j^{(i)})X_j^{(i)}\Big)+(1-Y_i)\Big(\sum_{j=1}^{m_i}X_j^{(i)}\Big)}{\sum_{i=1}^{n}Y_i\Big(\sum_{j=1}^{m_i}1-T_j^{(i)}\Big)+(1-Y_i)m_i}$$

$$\mu_1^{(c+1)} = \frac{\sum_{i=1}^{n}Y_i\Big(\sum_{j=1}^{m_i}T_j^{(i)}X_j^{(i)}\Big)}{\sum_{i=1}^{n}Y_i\Big(\sum_{j=1}^{m_i}T_j^{(i)}\Big)}$$

For the covariance matrices ${\Sigma_0}^{(c)}$ and ${\Sigma_1}^{(c)}$, replace $X_j^{(i)}$ by $(X_j^{(i)} - {\mu_k}^{(c)})(X_j^{(i)} - {\mu_k}^{(c)})^\intercal$, with $k \in \{0, 1\}$, in the expression of ${\mu_0}^{(c)}$ and ${\mu_1}^{(c)}$.

$$\boldsymbol{\Sigma}_0^{(c+1)} = \frac{\sum_{i=1}^{n} Y_i \Big(\sum_{j=1}^{m_i} (1 - T_j^{(i)}) S_{ij}^0 \Big) + (1 - Y_i) \Big(\sum_{j=1}^{m_i} S_{ij}^1 \Big)}{\sum_{i=1}^{n} Y_i \Big(\sum_{j=1}^{m_i} 1 - T_j^{(i)} \Big) + (1 - Y_i) m_i}$$

$$\boldsymbol{\Sigma}_1^{(c+1)} = \frac{\sum_{i=1}^{n} Y_i \Big(\sum_{j=1}^{m_i} T_j^{(i)} S_{ij}^1 \Big)}{\sum_{i=1}^{n} Y_i \Big(\sum_{j=1}^{m_i} T_j^{(i)} \Big)}$$

where $S_{ij}^0 = (X_j^{(i)} - \mu_0^{(c+1)})(X_j^{(i)} - \mu_0^{(c+1)})^\intercal$ and $S_{ij}^1 = (X_j^{(i)} - \mu_1^{(c+1)})(X_j^{(i)} - \mu_1^{(c+1)})^\intercal$.

Note that μ_0 and Σ_0 depend on all the bags but μ_1 and Σ_1 depend only on positive bags. So the distribution of healthy cells is estimated with all the patients but the distribution of tumor cells is estimated only with sick patients. The updates formulas are similar in some way to a semi-supervised EM [HJ20].

$T_j^{(i)}$ are used to define a decision rule to predict the label of a bag. We use the decision rules proposed in [Li+13]. Although derived under a different modeling framework, these rules can be interpreted as aggregation heuristics when using posterior probabilities. More precisely, $T_j^{(i)}$ is plugged in the product (2.5) and sum (2.7) rules. Finally, at the prediction step, one iteration of the EM algorithm is done to estimate π_i individually on new bags.

3 Numerical experiments

In the previous sections, I presented two MIL methods [ZG01; Li+13] that share similar ideas with our baseline model. We designed some experiments to compare them, as well as the decision rules referred above in Section 2.2. Since all these methods try to classify bags, we propose a method to visualize bags, given a pair of variables. In the same way scatter plots allow to see how points are spread in a feature space $\mathbb{R}^d$, this method allows to visualize how bags are scattered and then identify some simple discriminative patterns.

3.1 Visualizing bags with covariance ellipses

In this research work, we manipulate bags that can be thought of as probability distributions in $\mathbb{R}^d$ and we seek to discriminate them. Scatter plots are basic visualization tools in statistics. When analyzing data, they help to discover some simple discriminative patterns and trends in the data. While histograms and density plots help to visualize data distributions, they are not convenient when comparing many distributions at the same time. Inspired by [Sch25], we propose to visualize bags on a pair of variables: each bag seen as a probability distribution, is represented by an ellipse. The center is the statistical mean and the radius of the ellipse depends on the standard deviations. This visualization method can be seen as an equivalent scatter plot for bags: the objects to compare are no more points in $\mathbb{R}^d$ but empirical measures. This allows also to localize the supports of distributions in the feature space.

Let us suppose that we have a set of n data points $\{(x_i, y_i) \in \mathbb{R}^2;\ 1 \leq i \leq n\}$. To draw an ellipse representing this data, it is necessary to identify in which directions the point cloud is the most spread. Principal Component Analysis is the kind of method that allows to do so by extracting where most of the variance of the data is kept. In other words, we need to do an eigenvalue decomposition of the covariance matrix. [Sch25] proposed a method that avoid computing eigenvalues, by just using the mean, the standard deviations and the Pearson correlation coefficient. For a given covariance matrix $\mathbf{\Sigma} \in \mathcal{M}_2(\mathbb{R})$, the steps to draw a two-dimensional covariance ellipse are the following:

1. Compute the covariance $\text{cov}_{x,y}$.
2. Compute the correlation coefficient $\rho_{x,y}$: normalize the covariance by the standard deviations σ_x and σ_y.
$$\rho_{x,y} = \frac{\text{cov}_{x,y}}{\sigma_x \sigma_y}$$
3. Draw an ellipse centered at the origin and with axis aligned with x-axis and y-axis. The radii (horizontal and vertical radius) are:
$$\begin{cases} a &= \sqrt{1 + \rho_{x,y}} \\ b &= \sqrt{1 - \rho_{x,y}} \end{cases}$$
4. Rotate the ellipse counter-clockwise by 45°.
5. Choose n_{std} to define the ellipse scale: the number of standard deviations from the mean.
6. Scale the ellipse horizontally and vertically by $2n_{\text{std}} \times \sigma_x$ and $2n_{\text{std}} \times \sigma_y$ respectively.

7. Shift the ellipse to the center that represents the statistical mean $\mu = (\mu_x, \mu_y)$.

In sum, to draw an ellipse of the data covariance matrix $\mathbf{\Sigma}$, we first draw an ellipse representing the normalized version of the covariance matrix which is defined as:

$$\mathbf{\Sigma}_{\text{norm}} = \begin{pmatrix} 1 & \rho_{x,y} \\ \rho_{x,y} & 1 \end{pmatrix}$$

The fact that the Pearson correlation coefficient is playing an important role is not that surprising. Indeed, the less correlated the two variables (x, y) are, the more round the point cloud will look, and the more correlated the variables are, the more elongated the point cloud will be.

As mentioned in [Sch25], the proposed method tries to avoid computing eigenvalues but actually for a normalized covariance matrix of size 2×2, there is a relation between its eigenvalues and the correlation coefficient. We can notice it by solving the equation $\det(\mathbf{\Sigma}_{\text{norm}} - \lambda \mathbf{I}) = 0$ from which we obtain:

$$\begin{cases} \lambda_1 &= 1 + \rho_{x,y} \\ \lambda_2 &= 1 - \rho_{x,y} \end{cases}$$

The eigenvalues of $\mathbf{\Sigma}_{\text{norm}}$ correspond then to the squared horizontal and vertical radius of the normalized ellipse. In the steps enumerated above to draw the covariance ellipse, we made a 45° rotation at some point. This is justified by the slopes of the eigen vectors of $\mathbf{\Sigma}_{\text{norm}}$. Indeed, by solving $(\mathbf{\Sigma}_{\text{norm}} - \lambda_i \mathbf{I})v_i = 0$ for λ_1 and λ_2, we obtain respectively:

$$v_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad v_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

The slopes of v_1 and v_2 are 1 and -1, so $\theta_1 = \pi/4$ and $\theta_2 = -\pi/4$ are the angles with respect to the x-axis. Finally, to transform $\mathbf{\Sigma}_{\text{norm}}$ ellipse into $\mathbf{\Sigma}$ ellipse, we scale horizontally by $2n_{\text{std}} \times \sigma_x$ and vertically by $2n_{\text{std}} \times \sigma_y$. The parameter n_{std} allows to control the scale of the ellipse. As you can see on the figure 2.1, the smallest ellipse (the red one) is one standard deviation from the center and the biggest one (the green) is at three standard deviations.

3.2 Experiment on synthetic data

[Li+13] proposed a MIL method that combines predictions of a supervised classifier. Similarly to our baseline model (section 2.3), they define a mixture of distributions. But contrary to us, the proportions of their mixture are the same across all the bags. We

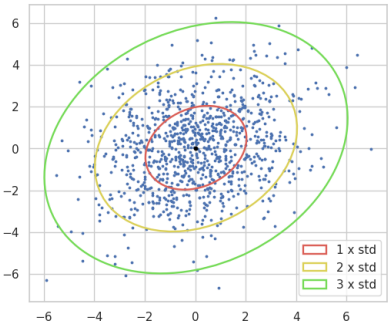
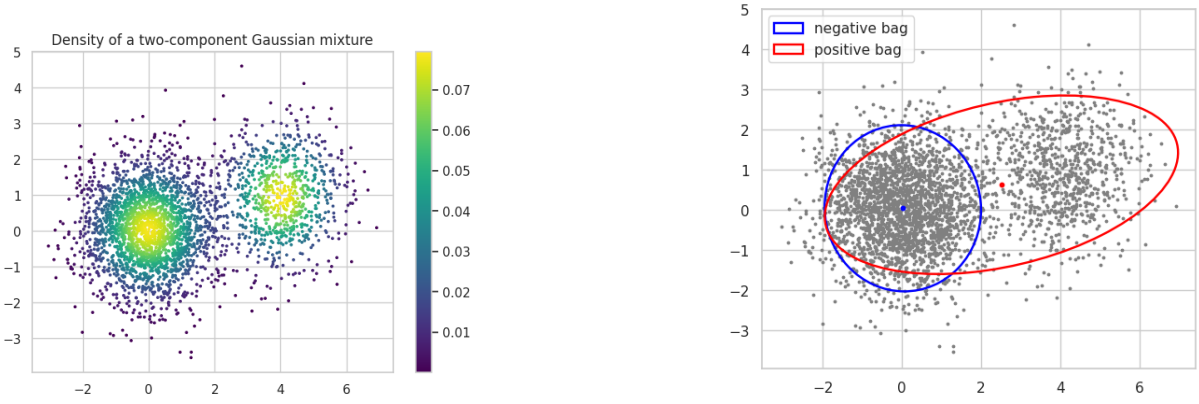


Figure 2.1: The covariance ellipse: It is a 2D representation of a covariance matrix of size 2×2. The center of the ellipses is the statistical mean. The red ellipse is at one standard deviation from the center. The yellow and the green are respectively at two and three standard deviations from the center.

designed an experiment to confront our approach to theirs, and compared some decision rules (see section 2.2). The diverse density method (section 2.1) is also included in the comparison.



(a) (b)

Figure 2.2: a) A mixture of two Gaussians. b) Bags creation with synthetic data: instances from negative bags (blue ellipse) are generated with the left Gaussian, instances from positive bags (red ellipse) are generated with a mixture of the two Gaussians. The mixture proportions are different from one bag to another. The experiment was repeated 10 times so the error bars represent the variability of the scores obtained.

Evaluation metrics for bag prediction performance. Throughout this thesis, I use classical binary classification metrics. Namely, the precision and the recall scores defined as:

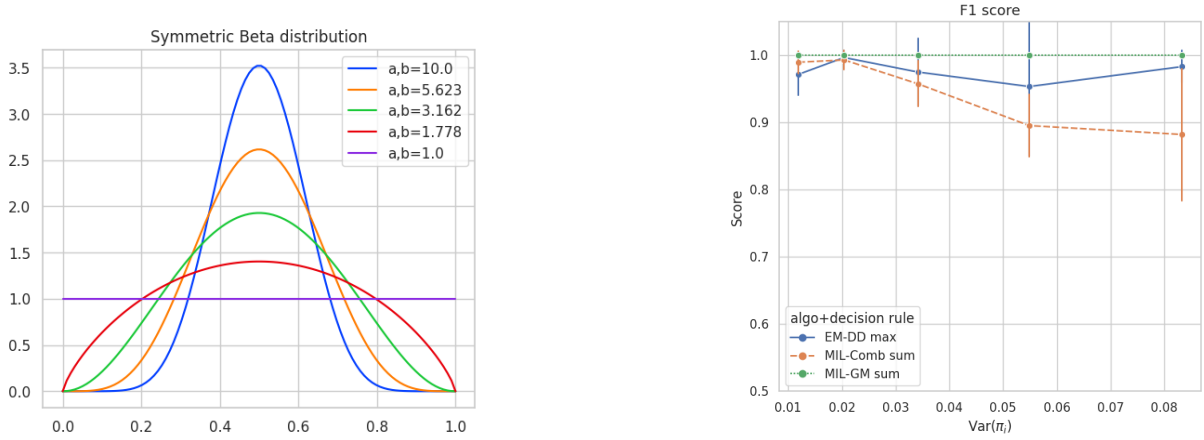
$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

where TP=True Positive, FP=False Positive, and FN=False Negative.

We also use the F_1 score defined as the harmonic mean between precision and recall:

$$\text{F1} = \frac{2}{\text{Precision}^{-1} + \text{Recall}^{-1}} = 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

For this experiment, we used a two-component Gaussian mixture (figure 2.2a) to generate the data. One of the two Gaussians represents the instances that belong to the concept (discriminative pattern for bags classification). As illustrated on figure 2.2b, instances from negative bags are generated with a single Gaussian, and instances from positive bags are generated with a mixture of two Gaussians. In order to simulate some variability across the bags, we randomly selected the mixture proportions with a Beta distribution Beta(a, b), where $a, b > 0$. So, for each positive bag, we draw $\pi_i \in (0, 1)$ from Beta(a, b). In order to ensure that having $\pi_i \geq 0.5$ is as likely as having $\pi_i < 0.5$, we used a symmetric Beta distribution by taking $a = b$ (see figure 2.3a).



(a) (b)

Figure 2.3: a) Symmetric Beta distributions: Law used to draw mixture proportion π_i for synthetic data generation. b) Results of experiment on synthetic data: EM-DD is diverse density + EM algorithm [ZG01], MIL-GM is our baseline model, MIL-Comb is a method combining classifier predictions [Li+13]. On the x-axis is the variance of the Beta distribution used to generate the mixture proportions π_i.

For the comparison of the MIL methods, it would be interesting to evaluate how much their performance is affected by the variability between bags. To control the variability of mixture proportion π_i, we have to control the variance of the Beta distribution. For a symmetric Beta(b, b), the variance is $\frac{1}{4(2b+1)}$. So, by decreasing b, the variance of Beta(b, b) increases.

The full steps to generate the data for this experiment are the following:

1. For each $b \in \{10, 10^{3/4}, 10^{1/2}, 10^{1/4}, 1\}$,

2. Define two Gaussians distributions: $f_0 \sim \mathcal{N}(\mu_0, \mathbf{\Sigma}_0)$ and $f_1 \sim \mathcal{N}(\mu_1, \mathbf{\Sigma}_1)$. Generate 30 bags (15 positive and 15 negative).

   - For a negative bag, generate 100 instances with f_0.
   - For a positive bag, draw first $\pi_i \sim \text{Beta}(b, b)$ and then generate 100 instances with the mixture $(1 - \pi_i)f_0 + \pi_i f_1$.

We repeated these steps 10 times in order to evaluate the variability of the results. We used different decision rules (see section 2.2) for the bags label prediction, evaluated them with the F1 score and compared them. As displayed in the table 2.1, the results for the classifier combining method, confirms that the product rule (2.5) is the least interesting rule as it yields the worst results. Indeed, by multiplying many probabilities that are poorly estimated, errors are accumulated on the overall product. This is even more true when the model is not generative ($\mathbf{P}(X, Y)$ is not modeled), which is the case here because we trained a traditional k-Nearest Neighbor classifier on instances and combined their predictions (as in [Li+13]). The sum rule (2.6) and the γ-rule (2.9) produced results that are significantly better than the product rule. The scores degrade when we increase the variability of π_i (decreasing b) across the bags. This was expected for the classifier combining method since all bags share the same proportion of instances belonging to the concept. The sum rule and the γ-rule have pretty close performances, which suggest that the γ-rule is a good alternative to the sum rule when the probabilities from instances cannot be estimated.

	$b = 10$	$b = 10^{3/4}$	$b = 10^{1/2}$	$b = 10^{1/4}$	1
Product rule	0.11 ± 0.12	0.21 ± 0.18	0.42 ± 0.15	0.42 ± 0.22	0.48 ± 0.14
Sum rule	$\mathbf{0.99 \pm 0.01}$	$\mathbf{0.98 \pm 0.02}$	$\mathbf{0.93 \pm 0.04}$	$\mathbf{0.94 \pm 0.04}$	$\mathbf{0.83 \pm 0.06}$
γ-rule	0.98 ± 0.03	0.98 ± 0.04	0.9 ± 0.08	0.9 ± 0.1	0.81 ± 0.12

Table 2.1: MIL-Comb [Li+13] results on the mixture of Gaussians: F1 scores obtained on the bags label prediction, b is the parameter of the symmetric Beta distribution used to draw the mixture proportions π_i.

Unlike the classifier combining method, our generative model provides calibrated probabilities, which makes the product rule more stable. We made indeed a Gaussian mixture assumption but more importantly we vary the mixture proportion π_i across the bags. As a result (see table 2.2), MIL-GM adapts better to the variability of π_i and since the model is probabilistic, the product rule yields almost perfect scores. Yet, the sum rule (2.7) is the best rule on the synthetic data. We did not use the γ-rule

for MIL-GM, which is not interesting when instances probabilities are correctly estimated.

	$b = 10$	$b = 10^{3/4}$	$b = 10^{1/2}$	$b = 10^{1/4}$	1
Product rule	0.99 ± 0.0	0.99 ± 0.02	0.99 ± 0.01	0.99 ± 0.01	0.99 ± 0.02
Sum rule	$\mathbf{1.0 \pm 0.0}$	$\mathbf{1.0 \pm 0.0}$	$\mathbf{1.0 \pm 0.0}$	$\mathbf{1.0 \pm 0.0}$	$\mathbf{1.0 \pm 0.0}$

Table 2.2: MIL-GM results on the mixture of Gaussians: F1 scores obtained on the bags label prediction, b is the parameter of the symmetric Beta distribution used to draw the mixture proportions π_i.

Finally, we compared the best rules for MIL-Comb and MIL-GM to the diverse density method EM-DD [ZG01]. EM-DD is particularly based on the MIL standard assumption and aims to find the region of the space where positive bags intersect. Naturally, the rule that fits with it, consists in classifying a bag positive if at least one if its instances belongs to this region. As illustrated on figure 2.3b, EM-DD produces results that are better than MIL-Comb. Put another way, EM-DD which aims to find at least one significant instance per bag, performs better then MIL-Comb which classifies all the instances individually before classifying bags. This raises an important problem in MIL, which is the relevance of instances for a learning task. This problem is somehow similar to the variables selection problem in machine learning. In MIL, finding where the relevant information at the observations (instances) level is that important. Can we derive meaningful representations for bags by identifying relevant instances for a task? How can we measure the importance of instances for a learning task? These are questions we will continue to discuss about in the following of this thesis and especially in chapters 3 and 5.

3.3 Experiment on flow cytometry data

In section 2.3, we proposed to represent patients as bags of cells and defined a two-component Gaussian mixture to model the distribution of cells. Naturally, we would like to know how this model performs on real data. In section 1 of chapter 1, we introduced the two datasets we worked with in this thesis: FlowCAP AML and Multiple Myeloma. We have access to cell population annotations (cell phenotypes) on Myeloma data but not on FlowCAP-AML data. One of the key challenges in flow cytometry data analysis is the presence of multiple cell populations with different behaviors in a biological sample. It makes the detection of tumor cells particularly difficult. Luckily, for Multiple Myeloma pathology, we know the target population is the plasmocytes. For a patient diagnosed with Myeloma, we then expect to have a mix of healthy plasmocytes and tumoral plasmocytes. This is the exact setup needed to test our baseline model

and it simplifies the original problem by reducing the search area for tumor cells. For this experiment, we then worked with Myeloma data and compared our baseline model MIL-GM to MIL-Comb [Li+13] and EM-DD [ZG01].

To perform the experiment, we used the data from 31 patients. Among those patients, 11 are at an early stage and 20 are diagnosed with Multiple Myeloma. We didn't have access to biological samples from healthy patients. The comparison of early stage and advanced stage patients is still interesting for different reasons. First, understanding the differences between these two groups of patients and characterizing the behavior of tumor plasmocytes, is currently one of the main research interests at Metafora biosystems. Second, at an early stage, patients have less tumor cells ($< 20\%$ of plasmocytes) than diagnosed patients ($> 80\%$ of plasmocytes). We illustrated this on figure 2.4 which shows on each subplot, biological samples from two early stage patients (blue) and two diagnosed patients (red). The variable CD19 displayed on the y-axis is the one used in practice by biologists to discriminate tumor plasmocytes from healthy ones. Due to the presence of cells taking negative values on CD19, the red dots are more spread over the y-axis than the blue dots. Yet, this simple criterion is not enough for discriminating patients. There is clearly a variability in the distribution of cells from one patient to another (even at the same stage). We illustrated this on figure 2.5 with covariance ellipses as introduced in section 3.1. Each ellipse represents the point cloud of a patient by the statistical mean and the two directions where it is the most spread. The shift between the centers and the different orientation from one ellipse to another, visually demonstrates the variability between patients.

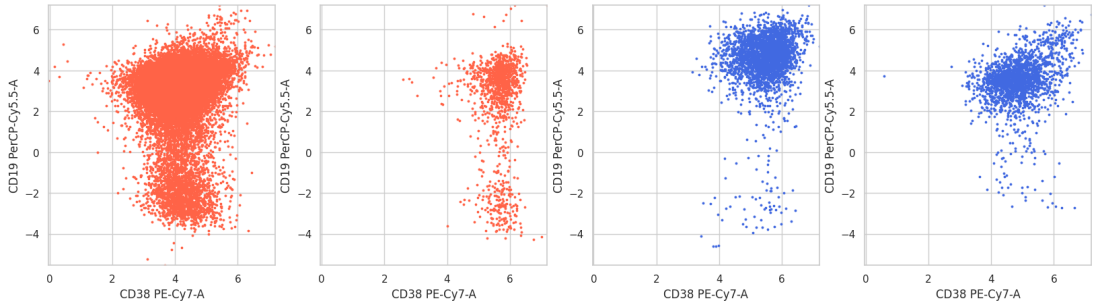


Figure 2.4: Plasmocytes populations from Multiple Myeloma data. Each subplot represents a patient. MGUS (early stage) patients are in blue and diagnosed patients are in red.

Including CD19 marker, we used $d = 7$ variables for this experiment. The number of plasmocytes varies from one patient to another but there is on average 3746 cells per patient. We performed a 3-fold cross-validation by keeping the same proportion of diagnosed patients in the train set and the validation set. The best decision rule

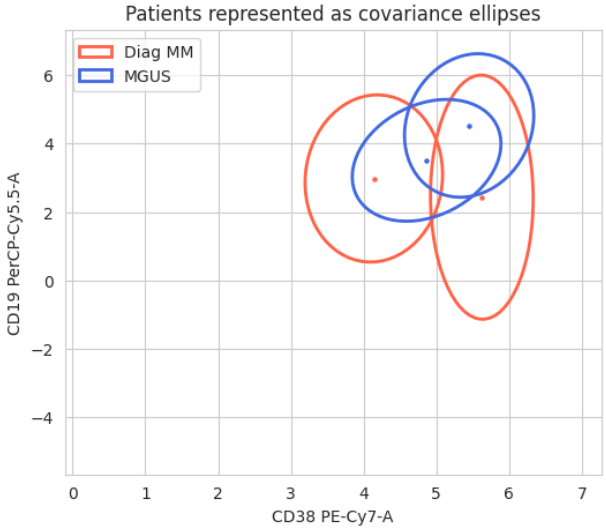


Figure 2.5: Multiple Myeloma patients are considered as bags of cells. We visualize them using covariance ellipses. MGUS = Early stage and Diag MM = Severe stage

(obtained after a comparison as in 3.2) for each MIL method was used. Namely, the max rule (MIL standard assumption) for the diverse density method EM-DD, the γ-rule (2.9) for the classifier combining method MIL-Comb and the sum rule (2.7) for our baseline model MIL-GM. The classifier used this time in MIL-Comb is a Gradient Boosting Tree. The AUC (Area Under ROC Curve) scores on table 2.3 shows that MIL-Comb produced the best results. Our baseline model produced the second best results overall but it is not close enough to MIL-Comb results. This is still fair since we made strong assumptions (Gaussian distributions and patient variability not modeled).

	Fold 1		**Fold 2**		**Fold 3**	
	Train	Validation	Train	Validation	Train	Validation
EM-DD	0.63	0.5	0.87	0.38	0.69	**0.48**
MIL-Comb γ-rule	**0.99**	**0.93**	**0.92**	**0.83**	**0.98**	0.29
MIL-GM Sum rule	**0.86**	**0.61**	0.76	0.71	**0.88**	0.24

Table 2.3: AUC scores from the cross-validation on Myeloma data. Validation mean AUC scores: 0.52 for MIL-GM, 0.68 for MIL-Comb, 0.45 for EM-DD.

Looking closer to the scores on table 2.3, for each fold, we notice a degradation between the train and the validation, with the biggest gap at Fold 3. There could be many interpretations to this phenomenon but the most relevant here is the fact we have few examples (31 bags or patients) for this experiment. The models seem to have difficulties to adapt to samples from the validation set, which would also explain the

scores variability across the folds. In order to evaluate correctly this variability, we performed a repeated 3- fold cross-validation. A repeated k-fold cross-validation consists in repeating the cross-validation procedure many times and by splitting each time, the data in k segments in a different way. After choosing the best decision thresholds for each model using the ROC curve, we used the F1 score to evaluate our models. The average F1 scores obtained at the validation steps are: 0.51 for EM-DD, 0.6 for MIL-Comb, 0.54 for MIL-GM. MIL-Comb is still the best but the performance is closer (from F1 score perspective). As shown on figure 2.6, we also compared the Precision and Recall scores obtained at the validation step. For each model, the precision score is on average higher than the recall score. The most striking detail is the interquartile range of the recall. The models are struggling to identify which patients are at an early stage (negative bags) and predicting the majority of patients as positive. This is probably due the imbalance (2/3 positive, 1/3 negative) at the patient level.

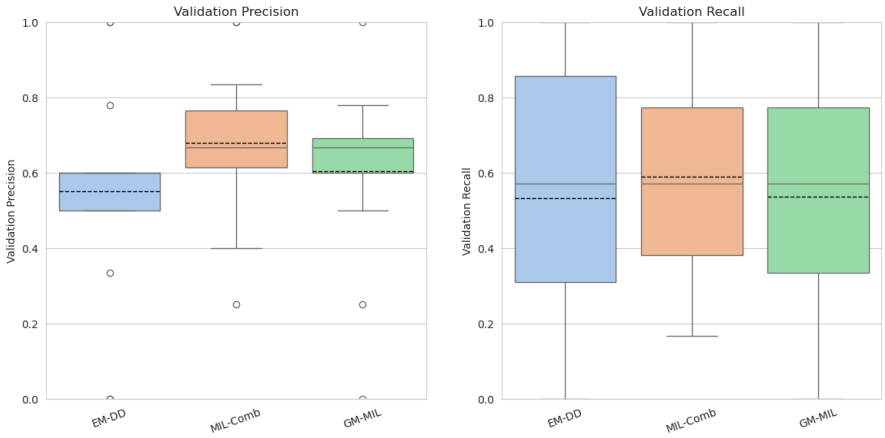


Figure 2.6: Precision and Recall validation scores on Multiple Myeloma data

This experiment allows us to observe some singular phenomena that might be related to the class imbalance and the lack of examples in the dataset. However, there is still a question: are we using the appropriate metrics to evaluate MIL models? Having for example, MIL-Comb with really good AUC scores (as shown on table 2.3) and mid F1 scores (with the best threshold selected), is pretty strange. While the AUC score translates how well a model is able to rank examples overall, the F1 score reflects at a fixed threshold, how well it makes binary decisions. MIL models learn soft probabilistic scores for bags by combining instances individual scores. So, by nature, they might be more suited to answer questions like "might this bag be positive or is this patient likely ill?" than making sharp binary decisions. In this thesis, we did not explore metrics more suitable for MIL models evaluation but it is definitely one of the main research directions to pursue in Multiple Instance Learning.

Chapter 3

Multiple instance learning with neural networks

In this chapter, we study Multiple Instance Learning (MIL) from the perspective of representation learning with neural networks, with a focus on patient-level classification from flow cytometry data. Our main contribution is a local MIL framework that departs from standard global approaches. Instead of representing each bag by a single embedding (vector representation), we learn a discriminative latent representation at the instance level using an Orthogonal Autoencoder, structure this space into regions corresponding to subpopulations, and associate a specialized classifier to each region through a Mixture of Experts. This approach aims to capture intra-bag heterogeneity more effectively and improve interpretability by linking predictions to meaningful substructures in the data. To position this contribution, we first review existing MIL methods that rely on global bag representations, including neural approaches based on pooling and attention mechanisms. We then present deep learning models designed for set-structured data, which provide a natural framework for MIL. Building on these foundations, we detail our proposed method and compare it to standard attention-based MIL on patient classification task. This comparison highlights the differences between global and local modeling strategies and provides insight into their respective strengths and limitations.

1 Learning from bags considered as single objects

Multiple Instance Learning (MIL) can be approached by considering each bag as a single entity, rather than focusing on individual instances. From this perspective, a bag can be viewed as a distribution of instances, which can be summarized using global statistics such as the mean, variance, or more complex descriptors. These summaries allow each bag to be represented as a fixed-length feature vector, enabling the use of standard

supervised learning methods. This family of approaches is commonly referred to as bag-space methods in the literature [Car+18; Amo13]. In this thesis, we adopt the term global methods to emphasize that bags are processed as whole objects, in contrast with the local, instance-based approach introduced in Chapter 2.

A key challenge in MIL lies in the ambiguity of the relationship between instances and bag labels. As highlighted in [Bab08], two main types of ambiguity can arise. The first, polymorphism ambiguity, refers to situations where relevant instances may exhibit significant variability (for example, different conformations of a molecule). The second, part–whole ambiguity, occurs when the label is assigned at the bag level, although only a subset of instances is actually informative (for instance, the presence of a specific object within an image). These ambiguities motivate the design of methods that rely on global representations rather than explicit instance labeling.

Several frameworks have been proposed to construct such representations. In this global approach to MIL, bags can be directly compared using distances or kernel functions defined between sets of instances. Classical approaches include extensions of the k-nearest neighbors algorithm based on the Hausdorff distance [WZ00], as well as multi-instance kernels [Gär+02], which allow kernel-based classifiers such as Support Vector Machines to operate on bags.

An alternative strategy is to map each bag to a vector representation before applying a classifier. For instance, the MILES method [CBW06] embeds bags by measuring their similarity to all instances in the training set, while YARDS [Fou08] constructs histogram-like representations based on similarities to selected prototypes. These approaches aim to capture the internal structure of bags while producing fixed-dimensional representations suitable for standard learning algorithms.

More recent contributions have further explored ways to encode bag-level information. For example, the Multiple Instance Dissimilarity (MInD) framework [CTL15] represents each bag through its dissimilarity to other bags, enabling the use of conventional classifiers such as SVM. Some methods such as CCE (Constructive Clustering based Ensemble) [ZZ07], group instances into clusters and describe bags through their cluster memberships, often combined within ensemble learning schemes. Other approaches rely on probabilistic models, such as mixture models [WYZ08], to capture the distribution of instances within each bag and define similarities accordingly.

Recent approaches leverage neural networks to learn representations directly from bags. Instead of relying on predefined similarities or summary statistics, these methods

aim to learn task-specific embeddings. In the following sections, we focus on MIL approaches based on neural networks, and in particular on methods that learn meaningful latent representations of instances before aggregating them at the bag level.

2 Deep learning on sets of observations: a framework compatible with Multiple instance learning

One of the main characteristics of MIL problems, and consequently one of the most important aspects in their modeling, is the intra-bag structure. The relationship between a bag and its instances fundamentally determines how the bag label is generated. Traditionally, modeling relies on explicit assumptions such as the standard MIL assumption, where a bag is positive if at least one of its instances is positive. Alternative formulations treat bags as probability distributions and derive decision rules by combining instance-level scores (see chapter 2, section 2.2). Although straightforward to implement, these approaches impose rigid structural constraints on the bag decision function. As a result, they restrict the expressiveness of the model and limit its ability to adapt to more complex data. Ideally, the decision function should not be predefined but more flexible, capable of adapting to diverse intra-bag relationships. Neural networks provide exactly such a framework. By composing nonlinear transformations, they approximate broad classes of functions (cf. uniform [Mak98] and universal approx [KL20] with neural networks, survey [Aug24]) and allow the aggregation rule itself to be learned rather than imposed.

2.1 From classical rule-based MIL to bags representation learning

The first important step toward neural MIL was proposed by [RDK00]. In the classical MIL formulation, the bag probability is typically computed using a max operator over instance probabilities, reflecting the standard assumption that a single positive instance is sufficient to make the bag positive. However, the max operator is not differentiable, which prevents direct optimization through backpropagation. To address this limitation, [RDK00] MINNs (Multiple Instance Neural Networks) replaced the max operator with the log-sum-exp function, a smooth and convex approximation of the maximum. This substitution made the objective function differentiable and allowed MIL models to be trained using gradient-based methods. Interestingly, their approach did not aim to learn a bag representation by combining instance features. Instead, each instance was first assigned a probability, and the log-sum-exp operator acted purely as a pooling

mechanism over instance probabilities. The model therefore remained conceptually close to the traditional MIL assumption: instances were evaluated independently, and aggregation occurred only at the level of predicted scores.

A significant conceptual shift occurred later with [Wan+18], who revisited Multiple Instance Neural Networks (MINNs). Rather than focusing on estimating instance labels and then aggregating them, they proposed architectures that learn bag representations directly. In their framework, instance features are transformed through neural layers and aggregated before the final classification step. This moves the emphasis from instance-level prediction toward representation learning under weak supervision.

This evolution marks an important transition in neural MIL: from approximating the max operator with a differentiable surrogate, to learning expressive bag-level representations in an end-to-end manner. Once aggregation becomes differentiable and embedded within a neural architecture, the entire pipeline, from instance embedding to bag decision, can be optimized jointly. This opens the door to more flexible and data-driven modeling of intra-bag structure.

2.2 Set-input neural networks and Attention-based MIL

A central property of MIL, and more generally of learning problems defined on sets, is permutation invariance: the prediction should not depend on the ordering of instances within a bag. Unlike sequences, elements of a set are unordered, and any architecture operating on them must respect this structural constraint. [Zah+17] formalized this requirement in their framework called Deep Sets. Note that, the authors did not propose a MIL model but since their model allow to process set-structured data, it actually fits with MIL. They provided necessary and sufficient conditions for a function defined on a set to be permutation invariant. Specifically, they showed that for inputs drawn from a countable set of arbitrary size, any permutation-invariant function f can be decomposed as:

$$f(\mathcal{X}) = \rho\left(\sum_{x\in\mathcal{X}} \phi(x)\right),$$

where ϕ maps individual elements to a latent representation and ρ maps the aggregated representation to the final output. This result is fundamental as it characterizes an entire class of permutation-invariant functions and provides a constructive recipe for designing neural architectures operating on sets. Note that, their framework connects to classical theoretical results such as *de Finetti* theorem [Lau07] and representation results in kernel

methods.

The theorem proposed by [Zah+17] has direct implications for MIL. It suggests that any valid MIL model must consist of three components: a transformation of instances into latent embeddings, a symmetric aggregation operator, and a final transformation producing the bag-level prediction. Within this framework, Deep Sets can be used in MIL in two conceptually distinct ways. The first possibility is a local (instance-level) approach. Each instance is independently transformed into a score or probability, and these instance-level outputs are then combined using a symmetric operator. This corresponds closely to the spirit of early neural MIL models by [RDK00]. The second possibility is a global (bag-level or embedding) approach. Instead of aggregating instance scores, instances are first mapped into a latent space, and the aggregation is performed on these learned embeddings. The aggregated representation is then used as a feature vector describing the bag. This approach emphasizes representation learning under weak supervision and enables richer modeling of intra-bag structure.

Transforming instances into embeddings and combining them through a symmetric operator to produce a bag representation is commonly referred as MIL pooling. Early neural MIL models, including [Wan+18], typically relied on predefined pooling operators such as the max or the mean. [ITW18] addressed this limitation by introducing an attention-based pooling mechanism within the Deep Sets formulation. Instead of assigning equal importance (mean pooling) or focusing only on the most extreme instance (max pooling), their model learns attention weights for each instance:

$$z = \sum_{i=1}^{n} a_i h_i,$$

where h_i are instance embeddings and a_i are learnable attention coefficients.

Attention mechanisms were already widely used in deep learning, particularly in natural language processing and computer vision, to allow models to focus selectively on relevant parts of the input. In the context of MIL, attention plays a two roles. First, it increases flexibility. The aggregation rule is no longer predefined but learned from data. The model can assign varying degrees of importance to instances depending on their contribution to the bag label. Second, it introduces interpretability. The attention weights provide a direct measure of instance contribution in the weighted average, offering insight into which elements of the bag drive the prediction. In this sense, attention-based MIL can be viewed as a natural and powerful application of the Deep Sets theorem: the instance transformation ϕ, the symmetric aggregation (now weighted and learned), and the final transformation ρ are all parameterized by neural networks and optimized

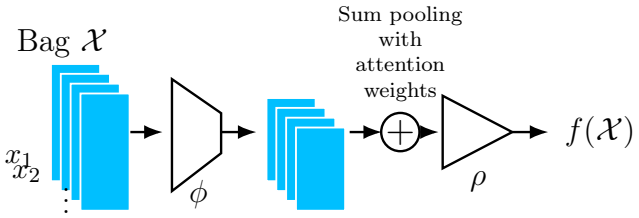


Figure 3.1: **Architecture of a neural network with attention mechanism for bags classification.** The original architecture comes from the Deep Sets [Zah+17] framework. Based on that, [ITW18] proposed an attention-based MIL framework. ϕ maps individual elements to a latent representation and ρ maps the aggregated representation to the final output.

jointly.

Further extending set-based architectures, [Lee+19] introduced the Set Transformer, a self-attention-based architecture designed specifically for unordered input-sets. Unlike the original Transformer [Vas+17], which models sequences, the Set Transformer captures pairwise interactions among elements in a permutation-invariant manner. This allows modeling richer intra-set dependencies beyond independent instance embeddings followed by pooling.

Set-input neural networks are particularly fascinating because of their remarkable versatility. Although some problems are naturally defined on sets, such as point clouds or collections of measurements, many traditional supervised learning problems can also be reformulated from a set-based perspective. In practice, the objects manipulated in machine learning are often decomposable. Images, for example, can be segmented into patches or pixels. Even when the original task is not explicitly defined on unordered collections, segmentation allows us to reinterpret the data as a set of meaningful components. This reformulation is especially useful when global annotations are available but fine-grained labels are not.

Deep Sets and related architectures have been successfully applied to a wide variety of tasks, including point cloud classification, population statistics estimation, anomaly detection, and set expansion. Within this landscape, neural MIL occupies a particularly important position. It offers a structured way to handle weak supervision in problems where only group-level labels are available. Recently, one of the most active application domains has been medical image analysis. In the following section, we focus on these developments in medical imaging, which represent one of the main application areas of neural MIL in the recent literature. We then extend this perspective to flow cytometry data analysis across multiple patients.

2.3 Neural MIL for computer aided diagnosis: medical image analysis and an extension towards flow cytometry data analysis

MIL for medical image analysis

Multiple Instance Learning has found a natural application domain in medical imaging. Processing entire high-resolution medical images, often consisting of billions of pixels, is computationally prohibitive. Moreover, obtaining pixel-level annotations is extremely costly and requires expert knowledge. MIL offers a practical compromise: images are divided into smaller patches, and the full image is treated as a bag with a single label [Que+17]. In computational histopathology [Xu+14], patches may correspond to cellular regions potentially indicative of malignancy. This formulation enables training with weak image-level annotations.

[CDP19] provide a comprehensive survey of MIL in medical image analysis, highlighting applications in various areas. They addressed a central challenge in medical image analysis: Lack of annotated data, despite the increasing size of imaging datasets. Manual annotation of medical images (pixel-level segmentation, lesion localization, etc.) is time-consuming, expensive and often unavailable in routine clinical practice. This motivates approaches that learn with less or weaker supervision.

Early applications directly adopted the standard MIL assumption: a bag is positive if at least one instance is positive. Images were divided into patches, and the objective was to predict disease presence from image-level labels. Representative works include abnormality detection [Vur+06], diabetic retinopathy [Que+12], histopathology cancer detection [KH15], and COPD (chronic obstructive pulmonary disease) classification [Che+14]. These works demonstrated that learning from weak image-level annotations was feasible and sometimes competitive with fully supervised approaches. However, most methods relied on simple assumptions and treated patches independently, ignoring structural relationships within images.

Subsequent work emphasized the distinction between local or instance-level methods, and global or bag-level methods. Local methods inferred latent patch labels and aggregated them, while global approaches modeled the image as a structured entity [KH15]. Global models often achieved stronger bag classification performance because they implicitly accounted for correlations among patches. This marked a conceptual

shift: MIL was no longer only a weak supervision workaround but a framework for modeling intra-image structure. Yet, a trade-off became apparent. Global methods provided limited localization, while local methods offered interpretability but sometimes underperformed for the classfication of bags (images).

To address this tension, later works aimed to combine global diagnosis with local detection. Applications included tuberculosis detection [Mel+14], and histopathology segmentation [Xu+14]. However, empirical findings revealed a fundamental limitation: optimizing bag-level accuracy does not guarantee reliable instance-level detection. For example,[Che+15] observed instability in instance outputs of strong bag classifiers.

Attention-based MIL [ITW18] replaced fixed pooling operators with learnable, permutation invariant aggregation. Instead of relying on standard and restrictive assumptions, instances are weighted and aggregated through an attention mechanism. This provided a greater flexibility than predefined pooling operators and allowed to model collectively instance contributions. Attention weights can help to improve interpretability. Although attention does not fully resolve the problem, it represents the most advanced and widely adopted paradigm in recent medical image analysis.

Adapting neural MIL to flow cytometry data analysis

The evolution of MIL in medical imaging naturally motivates its extension to other diagnostic domains involving observationsorganized into sets. In medical imaging, MIL has been used to assist diagnosis when only global labels are available and fine-grained annotations are costly. The ultimate objective remains computer-aided diagnosis: detecting disease reliably, efficiently, and as early as possible. However, imaging techniques such as MRI or CT scans are relatively expensive and, in some cases, invasive or burdensome for patients. A long-term objective in clinical research is therefore to design diagnostic techniques that are less invasive while remaining reliable and informative. In this context, the analysis of single-cell data offers a promising alternative. Subtle changes at the cellular level may reveal pathological processes at very early stages and potentially before visible evidence appears in medical imaging.

With this perspective in mind, it becomes natural to ask how would these MIL methods for medical image analysis perform when applied to flow cytometry data. Both settings share key characteristics: a patient is represented as a set of elements (image patches or cells), only patient-level labels are available, the goal is diagnosis under weak supervision.

Attention-based MIL is particularly appealing in this context because it combines permutation invariance, learnable aggregation, and interpretability through instance weighting. In medical imaging, it allows identification of relevant image regions. In cytometry, it could potentially highlight diagnostically relevant cellular subpopulations. However, transferring these models from image analysis to flow cytometry is not straightforward. Architectures proposed in the literature are primarily designed for convolutional encoders operating on image patches. Applying neural MIL in this new context requires methodological adaptation.

In this thesis, our contribution is therefore not merely to apply an existing framework, but to adapt attention-based MIL [ITW18] for patient-level classification using flow cytometry data. This involves: replacing convolutional instance encoders with fully connected architectures suited to structured cellular measurements, designing aggregation mechanisms stable under large bag sizes, and ensuring that attention-based pooling remains computationally tractable and biologically interpretable. By grounding our approach in advances from medical image analysis while adapting them to single-cell data, we aim to explore whether neural MIL can contribute to the development of less invasive diagnostic strategies. To our knowledge, attention-based MIL has not previously been adapted for flow cytometry–based patient classification.

3 Combining Autoencoders and Mixture of Experts

In this section, we propose a MIL framework that combines autoencoders and mixture of experts. Prior to that, I present a survey on autoencoders.

3.1 A survey on Autoencoders

An autoencoder is a neural network designed to encode input data into a compressed representation and then decode it to reconstruct the input. The objective is to make the output the most similar possible to the input. The network is composed principally of: the encoder which compresses the input data and the decoder which reconstructs it. The autoencoders are generally trained in a non supervised fashion, which allows them to learn informative representations without needing labels.

A vanilla autoencoder is composed of an input layer, one or several hidden layers and an output layer. During the decoding step, the input X is transformed into a code Z (in general in lower dimension) by using an encoding function f_θ. For the decoding step, the autoencoder reconstructs the input data by transforming the compressed data

Z and by using a decoding function $g_{\theta'}$. The weights of the network are computed by minimizing the reconstruction error between the input and the output. Once the autoencoder is trained, the encoding function can create representations of reduced dimension of new input data, whereas the decoding function can reconstruct the data from these representations. Training an autoencoder consists generally in solving:

$$\min_{\theta,\theta'} \sum_{i=1}^{n} \mathcal{L}(x_i, g_{\theta'}(f_\theta(x_i)))$$

where x_i is the i-th observation and $\mathcal{L}$ is the reconstruction error between the input and the output. $\mathcal{L}$ is the quadratic error: $\mathcal{L}(x, x') = \|x - x'\|_2^2$. The functions of encoding and decoding f_θ and $g_{\theta'}$ are two neural networks.

In the literature [Ber+24; BKG23], there exists different kinds of autoencoders depending on the network architecture used and the underlying model. We can distinguish principally: regularized autoencoders and generative autoencoders.

A regularized autoencoder is an autoencoder that aims to extract a compressed representation of the input data by applying some constraints (through regularization). The regularization is necessary for avoiding overfitting. This is particularly important when the size of the hidden layer is equal or higher the input data. By incorporating various regularization technique, autoencoders can learn characteristics of the data and represent them with particular properties (sparsity for example). The representations learned are then used for other learning tasks. Some authors [LLY18] for example, use these representations to do supervised classification.

The objective function of a regularized autoencoder is generally defined as follows:

$$\sum_{i=1}^{n} \mathcal{L}(x_i, g_\theta(f_\theta(x_i))) + \lambda \sum_{i=1}^{n} \mathcal{R}(f_\theta(x_i))$$

where $\lambda > 0$ is the regularization parameter and $\mathcal{R}$ is a penalty function.

A generative autoencoder is a kind of autoencoder that concentrates on learning the probability distribution of the input data. Contrary to traditional autoencoders, that aim principally to reduce the dimension, they are conceived to generate new data samples that are similar to training data. The generative autoencoders learn to encode new input data in a latent space, and to decode them for generating new instances. The most common example is the variational autoencoder [KW+19]. It is particularly efficient for modeling complex distributions. It was already used in the detection of cancer types [Pra+21].

The autoencoders are principally used to learn the compressed and relevant representations of the data, which can be beneficial for tasks such as clustering and classification. Thanks to their versatility in various domains: images processing, natural language processing, anomaly detection, recommendation systems.

To detect anomalies, the autoencoders are trained on data in order to learn recurrent and discriminative patterns. They can then efficiently reconstruct these data when they are presented with similar inputs, which is important to identify anomalies. The reconstruction error serves as a key indicator to detect anomalies. The critical step is to define a suitable threshold for the error reconstruction [MSC18]. As a general rule, this threshold is determined based on the errors distribution coming from the data. Any observation that produces a reconstruction error exceeding the threshold, is considered like an anomaly. This method is useful in various fields and more particularly in the health one [Pra+21].

3.2 A local MIL framework based on Autoencoders and Mixture of Experts

Most recent neural MIL approaches adopt a global representation strategy: instance embeddings are aggregated (via pooling or attention mechanisms) into a single bag-level representation, which is then used for classification. While effective, this strategy suffers from two important limitations:

- Loss of instance-level information: local cellular heterogeneity is compressed into a single representation.
- Limited interpretability: it becomes difficult to identify which cellular subpopulations explain the prediction.

Our approach proposes a different perspective: instead of learning a global bag embedding, we construct a discriminant latent representation at the instance level and structure this space into regions, each associated with a specialized expert classifier. This introduces a locally structured MIL model, combining: unsupervised representation learning (autoencoder), probabilistic clustering (Gaussian mixture), and a mixture of experts classifier. The key idea is that patient status emerges from the interaction between interpretable cellular subpopulations, rather than from a single bag embedding.

Latent Representation via Orthogonal Autoencoder. To capture complex nonlinear cellular structures, we first learn a latent representation. Let $f_\theta : \mathbb{R}^d \to \mathbb{R}^q$ be the

encoder, $g_{\theta'} : \mathbb{R}^q \to \mathbb{R}^d$ be the decoder. For each instance: $H_j^{(i)} = f_\theta(X_j^{(i)})$. The network is trained to minimize the orthogonal autoencoder objective:

$$\min_{\theta,\theta'} \|X - \hat{X}\|_F^2 + \lambda \|H^\intercal H - \mathbf{I}\|_F^2$$

where $\hat{X} = g_{\theta'}(H)$, the second term enforces approximate orthogonality of latent components, λ controls the strength of the orthogonality constraint. The orthogonality is applied to the embedding, not the network weights.

The main idea of the orthogonal autoencoder [Wan+19] is to improve clustering performance by forcing the learned features (embeddings) to be orthogonal inside an autoencoder based deep clustering framework. Traditional clustering (like K-means) struggles when data is high-dimensional (e.g., images), and features are not well-separated. Deep learning models like autoencoders can reduce dimensionality and extract features but, the features learned by vanilla autoencoders are not always ideal for clustering [Yan+17]. The idea of using an orthogonality constraint comes from the fact that methods like PCA and spectral clustering work well partly because they enforce orthogonality in embeddings, and orthogonal features are more distinct and less redundant.

Latent Space Segmentation via Gaussian Mixture. Once the latent representations H_{ij} are obtained, we model their distribution using a Gaussian Mixture Model (GMM) with K components. This step allows to partition the latent space into K subregions corresponding to cellular subpopulations. The mixture weights w_k represent the global importance of each subpopulation. Importantly, the segmentation is performed in a learned discriminative space rather than in the original feature space. This significantly increases flexibility compared to our baseline model based on a two-component Gaussian mixture.

Mixture of Experts on regions of the latent space. Each cluster k is associated with a specialized classifier (expert) Φ_k. We adopt the following supervision strategy: the bag label is assigned to its instances. Each expert Φ_k is trained on instances associated to the corresponding cluster in the latent space. The final instance-level prediction is:

$$\Phi^{\text{MoE}}(X_j^{(i)}) = \sum_{k=1}^{K} w_k \Phi_k(H_j^{(i)})$$

Unlike standard neural Mixture of Experts [GF19] where the weight of each expert is learned, here the weights are induced by the clustering in the latent space.

Bag-level prediction. Once instance-level scores are computed, we aggregate them into a bag-level decision using the average over instances predictions. Figure 3.2 summarizes the full pipeline of our proposed method.

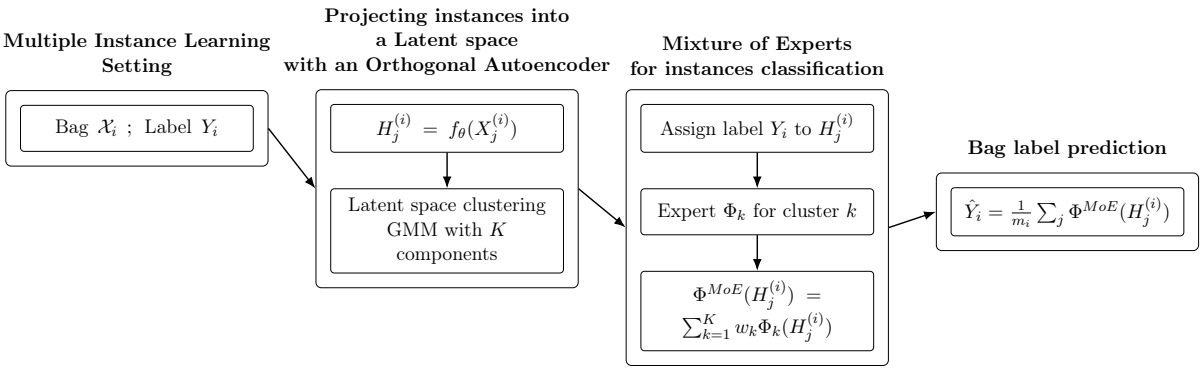


Figure 3.2: The pipeline of our local MIL method based on Orthogonal autoencoder wand Mixture of Experts

4 Application to flow cytometry data

We propose to evaluate the local MIL framework that we introduced just above. For the same reasons and motivation mentioned in section 3.3 of chapter 2, we worked with Multiple Myeloma data. We would like to compare our approach to the global approach taken in Attention-based MIL [ITW18]. This comparison could contribute to answer one of they key questions in MIL: does classifying instances and aggregating their predictions, yield better results than classifying bags as whole objects? To assess the relevance of both approaches, we included our baseline model MIL-GM (2.3) in the comparison.

Data and experimental setup. For this experiment, we used 24 patients including 10 MGUS patients (early stage condition) and 14 diagnosed Multiple Myeloma patients (Diag MM). The classification task consists in discriminating: MGUS vs Diag MM. Each patient constitutes one bag, and each plasmocyte (target cell population) corresponds to an instance. Most patients in this dataset contain approximately 2000 plasmocytes. Each plasmocyte is characterized by 7 biological variables: morphology-related markers (FSC, SSC), and fluorescence markers (CD38, CD138, CD19, CD45, H2). We adopted a repeated cross-validation strategy to evaluate the variability of the performance estimates. More precisely, we performed 5 times a 2-fold cross-validation and averaged the scores obtained across repetitions. Before discussing the results reported in table 3.5, we first detail the implementation choices and optimization procedures used for each

model to ensure reproducibility.

Implementation details. For the global MIL approach, we implemented[1] an attention-based neural network architecture inspired by [ITW18]. While the original model was designed for image classification using convolutional layers, our data consists of tabular cytometry measurements. We therefore replaced convolutional layers with fully connected layers. The network consists of five fully connected layers, as reported in table 3.1. The attention mechanism is applied after the instances embedding layers to compute attention weights, which are then used to form bag representation before the final classification layer. This specific architecture was selected after a validation procedure (details omitted here for conciseness) where we varied the number of neurons per layer. Optimization details are provided in table 3.2.

Layer	Type
1	Full connect (dim=256) + ReLU activation
2	Full connect (dim=128) + ReLU activation
3	Full connect (dim=64) + ReLU activation
4	Full connect (dim=64) + Attention weights + Average Pooling
5	Full connect (dim=1) + Sigmoid activation

Table 3.1: Architecture AttentionMIL (Dropout after each ReLU)

Optimizer	Learning rate	Weight decay	Epochs	Stopping criteria
Adam	10^{-4}	5×10^{-3}	150	Best validation accuracy and loss

Table 3.2: Optimization procedure AttentionMIL. Initialization of weights with Glorot uniform [GB10].

The training of our local MIL framework (section 3.2) was performed in two successive stages. We first trained an Orthogonal Autoencoder (OAE) to embed instances into a 3-dimensional latent space. The orthogonality regularization parameter was fixed as reported in table 3.4. The architecture (table 3.3) and optimization parameters were chosen through trial and error. The only validation criterion was qualitative: the ability to visually distinguish structured groups of cells in the latent space. The final embedding is illustrated in figure 3.3, where distinct cellular structures emerge. Once the embedding was learned, we applied a Gaussian Mixture Model with three components to cluster

[1]The full implementation and evaluation pipeline is available at https://github.com/pamikem/deep-learning-on-sets.git

the latent representations. The choice of three components was motivated by empirical observation of the latent space structure. After clustering, we assigned one expert per cluster. Each expert is a supervised classifier trained on the instances belonging to its cluster. For this experiment, we used Gradient Boosting Trees [NK13] as expert models. The final instance-level prediction is obtained by combining the predictions of the three experts according to their cluster memberships, before aggregation at the bag level.

Layer	Encoder	Decoder
1	Full connect (dim=16) + ReLU	Full connect (dim=8) + ReLU
2	Full connect (dim=8) + ReLU	Full connect (dim=16) + ReLU
3	Full connect (dim=3) + ReLU	Full connect (dim=7) + ReLU

Table 3.3: Architecture Orthogonal Autoencoder

Optimizer	Learning rate	Epochs	Stopping criteria
Adam	10^{-3}	150	Best validation loss

Table 3.4: Optimization procedure Orthogonal Autoencoder. Regularization parameter $\lambda = 0.0005$.

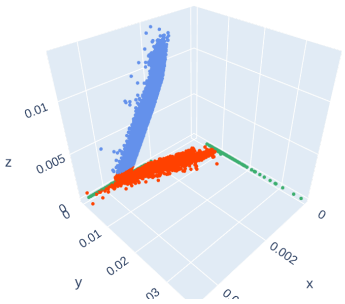


Figure 3.3: Visualization of the 3-dimensional latent space learned by the Orthogonal Autoencoder. Each point corresponds to a cell, and colors represent the clusters obtained with a Gaussian Mixture Model (K = 3).

For the baseline model MIL-GM, we followed exactly the experimental protocol described in the chapter 2. In particular, bag-level predictions are obtained using the

sum rule (2.6) for aggregation. No additional tuning was performed for this experiment.

Results and discussion. According to the scores on table 3.5, both AttentionMIL and OAE + MoE (our local MIL framework) exhibit very high recall (> 0.8) and mid precision (≈ 0.6). This suggests a strong tendency to predict positive labels. Crucially, the AUC score of OAE + MoE is close 0.5, which is the performance of a random guessing decision rule. The AUC of our baseline model MIL-GM is similarly low. Only AttentionMIL shows moderate discrimination (AUC > 0.7).

Models	Accuracy	AUC	F1	Precision	Recall
AttentionMIL	0.675 ± 0.1	0.743 ± 0.14	0.772 ± 0.05	0.681 ± 0.11	0.928 ± 0.12
OAE + MoE	0.608 ± 0.1	0.594 ± 0.19	0.726 ± 0.06	0.645 ± 0.15	0.886 ± 0.16
MIL-GM	0.492 ± 0.11	0.614 ± 0.11	0.366 ± 0.26	0.563 ± 0.34	0.329 ± 0.3
NaiveMIL	0.583 ± 0.0	0.5 ± 0.0	0.737 ± 0.0	0.583 ± 0.0	1.0 ± 0.0

Table 3.5: AttentionMIL [ITW18], OAE + MoE our local MIL framework, MIL-GM our baseline model (2.3), NaiveMIL is a supervised SVM (Support Vector Machines) trained at the instance level, after assigning the bag label to instances.

Due to unexpected performance patterns, we introduced an additional baseline model: NaiveMIL. It consists in a supervised SVM (Support Vector Machines) trained at the instance level. All instances inherit the label of their bag. At prediction time, the standard assumption is applied for bag classification. This method deliberately ignores the bag label ambiguity and treats the problem as fully supervised at the instance level. It serves as a diagnostic baseline to evaluate whether more sophisticated MIL mechanisms truly learn discriminative patterns.

NaiveMIL achieves perfect recall by predicting nearly all bags as positive. It also achieves the second best F1 score (0.737) after AttentionMIL. A high F1 score can be obtained without learning meaningful discriminative structure, simply by exploiting class imbalance and predicting every one positive. This suggests all the models (AttentionMIL, OAE + MoE, MIL-GM) did not learn discriminative patterns as we would like. This can be attributed to the insufficient sample size (few patients) and the class imbalance (14 positive bags and 10 negative bags). And more specifically, OAE+MoE also fails to learn discriminative patterns because the latent space representation and segmentation may not align with with disease-specific patterns.

In summary, the limited performance of all models indicates that there is a challenge

related to the class imbalance and sample size at the bag level. While cell-level information is abundant, the number of patients remains too small to support reliable generalization. This highlights a key limitation and motivates the need for data augmentation at the bag level. In the next chapter, we address this issue by proposing a novel framework to generate synthetic bags using Wasserstein barycenters.

Chapter 4

Data augmentation with Wasserstein barycenters

In Multiple Instance Learning (MIL), data are organized as bags of instances, yet in many real-world applications the number of available bags, such as patients in medical datasets, is severely limited. Our numerical experiments in previous chapters highlight this bottleneck: performance is constrained not by the number of instances, but by the scarcity of bags. This makes data augmentation at the bag level a critical, yet largely unexplored, direction. In this chapter, we introduce a novel framework for generating synthetic bags by operating directly in the space of probability measures. We model each bag as an empirical distribution and construct artificial patients through interpolation using Wasserstein barycenters. To the best of our knowledge, data augmentation remains underdeveloped in the MIL literature, and no existing method leverages Wasserstein barycenters for this purpose. Our approach addresses this gap by providing a principled, distribution-level augmentation strategy grounded in optimal transport.

1 Distribution interpolation for data augmentation

Data augmentation refers to a set of techniques used to increase the size and diversity of a training dataset by generating synthetic samples. Its main objectives are to improve generalization, reduce overfitting, and mitigate data scarcity or class imbalance. Common approaches include transformation-based methods (rotations or noise injection), feature-space interpolation like SMOTE (Synthetic minority oversampling technique) [Cha+02], and generative models such as GANs (Generative Adversarial Networks) [Goo+14; Gul+17] or Variational autoencoders [KW+19].

In multiple instance learning (MIL), data are organized into bags of instances, where

each bag $\mathcal{X} = \{X_1, \cdots, X_m\}$ can be naturally interpreted as an empirical distribution:

$$\nu = \frac{1}{m} \sum_{j=1}^{m} \delta_{X_j}$$

This distributional perspective is particularly relevant in real-world applications, where each bag can be seen as a finite sample drawn from an underlying probability measure. In practice, MIL datasets often suffer from a limited number of labeled bags and strong class imbalance. Standard data augmentation methods, such as SMOTE [Cha+02], operate at the level of individual instances and do not account for the distributional structure of bags. To address this limitation, we propose to perform data augmentation directly at the distribution level.

We extend the interpolation principle from individual observations to entire distributions. If a bag $\mathcal{X}$ is interpreted as an empirical measure μ, it is then possible to generate a synthetic bag by constructing an interpolated measure. By varying an interpolation parameter $\lambda \in [0, 1]$, we can simulate a continuum of intermediate distributions whose supports partially overlap with those of the original bags. In order to construct this kind of interpolation, we propose to use *Wasserstein barycenters.*

Let (Ω, D) be a metric space, $\mathcal{M}_+^1(\Omega)$ the set of probability measures on Ω. The p-Wasserstein distance between μ and ν is defined as:

$$\mathcal{W}_p(\mu, \nu) := \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\Omega \times \Omega} D(x, y)^p d\pi(x, y) \right)^{1/p}$$

where $\Pi(\mu, \nu)$ is the set of all probability measures on $\Omega \times \Omega$ that have marginals μ and ν.

A Wasserstein barycenter of $\nu_1, \ldots, \nu_N$ is defined as:

$$\mu^\star = \text{argmin}_\mu \sum_{i=1}^{N} \lambda_i \mathcal{W}_p^p(\mu, \nu_i)$$

where $\sum_i \lambda_i = 1$.

2 Barycenter notion in the Wasserstein space

2.1 Why Wasserstein barycenters are geometrically attractive?

When interpolating probability distributions, the choice of distance or divergence is crucial. Several alternatives to Wasserstein distances have been proposed, including Kullback–Leibler (KL) divergence and more general Bregman divergences [NN07; GMP23]. However, these measures present important limitations in our setting. In particular, KL divergence requires absolute continuity between distributions and becomes infinite when their supports do not overlap. This is problematic when working with empirical measures, which are often discrete and may have disjoint supports. Moreover, KL divergence compares distributions pointwise and does not take into account the geometry of the underlying space. In contrast, Wasserstein distances, arising from optimal transport, explicitly incorporate the geometry of the feature space by measuring the minimal cost of transporting mass between distributions. They remain finite even when supports do not overlap and define meaningful notions of interpolation between measures. This makes them particularly well suited for comparing and interpolating empirical distributions.

A key distinction must also be made between simple mixtures of distributions and Wasserstein barycenters. A mixture corresponds to a convex combination of measures and effectively aggregates their supports without modifying their locations. As a result, it often produces multimodal distributions that do not reflect intermediate structures. By contrast, Wasserstein barycenters rely on transport geometry: mass is displaced toward intermediate locations, and the support evolves continuously between the input distributions. This leads to smoother interpolations and avoids artificial multimodality. This difference is illustrated in figure 4.1, where the mixture retains the union of supports, while the Wasserstein barycenter generates new intermediate support locations. These geometric properties make Wasserstein barycenters particularly attractive for our setting, where bags are represented as empirical measures and interpolation should reflect meaningful variations in the underlying data distribution.

2.2 Existence, characterization and computation in particular cases

The notion of Wasserstein barycenters was first introduced by [AC11]. They provide existence, uniqueness, characterizations and regularity of the barycenter, and relate it to the multi-marginal optimal transport problem. They also provide examples and solve the gaussian case.

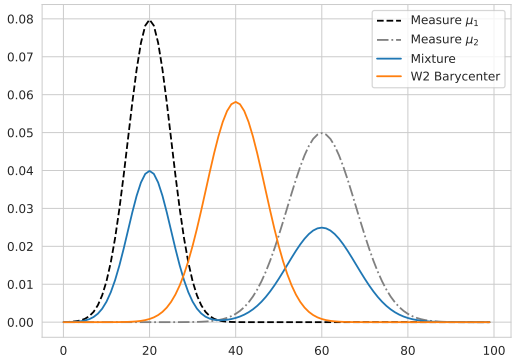


Figure 4.1: A comparison of a mixture of distributions (in blue) and a Wasserstein barycenter (in orange). The two distributions used as inputs are in black. This example is inspired from Python Optimal Transport toolbox POT [Fla+21].

We consider the 2-Wasserstein distance between probability measures defined $\mathbb{R}^d$:

$$\mathrm{W}_2(\mu,\nu) := \left(\inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|^2 d\pi(x,y)\right)^{1/2}$$

Given a set of measures $\{\nu_1,\dots,\nu_N\}$ defined on an arbitary space Ω, the barycenter problem is defined as:

$$\min_{\mu} \sum_{i=1}^{N} \lambda_i \,\mathrm{W}_2^2(\mu,\nu_i) \tag{4.1}$$

where the λ_i are positive weights summing to 1.

The support of the barycenter $\mu^\star = \mathrm{Bar}((\nu_i,\lambda_i)_i)$ is located in the convex hull of the support of $(\nu_i)_i$. Note that, this barycenter concept is a generalization of the Fréchet mean in a metric space to arbitrary measures. For two measures, this construction reduces to McCann's interpolation [McC97], which is fundamental in optimal transport and gradient flow theory. [AC11] generalizes the notion of Wasserstein barycenter to more than two measures. The authors showed that a barycenter always exists and is unique under mild conditions (for example, if one of the measures is regular enough).

[AC11] prove that if $\nu_1,\dots,\nu_N$ are in $\mathcal{P}_2(\mathbb{R}^d)$, *i.e* probability measures in $\mathbb{R}^d$ with finite second moment then the barycenter problem has always a minimizer (4.1). Uniqueness does not hold in full generality, but they prove uniqueness under a mild condition. If at least one input measure ν_i is absolutely continuous with respect to Lebesgue measure, then the barycenter is unique. They also prove an important density regularity result. If one measure has a density bounded in L^∞, then the barycenter also has a bounded

density. More precisely, $\|\mu^\star\|_{L^\infty} \leq \frac{1}{\lambda_i^d}\|\nu_i\|_{L^\infty}$. In other words, the barycenter inherits regularity from the input distributions.

[AC11] derive a structural equation describing the barycenter using optimal transport maps. The barycenter is the measure whose mass is mapped to the inputs so that the weighted average of the transport maps equals the identity.

A transport map $T : \mathbb{R}^d \to \mathbb{R}^d$ is a function that allows to move one point from a measurable space to another. The push-forward operator $T_\#$ is an extension of T that can move an entire probability measure on $\mathbb{R}^d$ toward a new probability measure on $\mathbb{R}^d$. The operator $T_\#$ pushes forward each elementary mass of a measure ν on $\mathbb{R}^d$ by applying the map T to obtain then an elementary mass in $\mathbb{R}^d$.

In the case of an Euclidean space like $\mathbb{R}^d$, it is possible to characterize a Wasserstein barycenter using transport maps. As shown in [AC11] and emphasized in Remark 9.4 of [PC19], when $\mu = \mathrm{Bar}((\nu_i, \lambda_i)_i)$ has a density, we can define optimal transport maps T_i between μ and ν_i such that $T_{i,\#}\mu = \nu_i$. The average map

$$\sum_{i=1}^{N} \lambda_i T_i(x) = x \tag{4.2}$$

This barycenter characterization can be equivalently formulated as follows: If $\mu^\star$ is solution of the barycenter problem (4.1) then there exists convex potentials $\psi_\star$ such that:

$$\sum_{i=1}^{N} \lambda_i \nabla\psi_i^\star = \mathrm{Id}$$

where $\nabla\psi_i^\star$ are Brenier's maps transporting ν_i to $\mu^\star$.

Solving the barycenter problem is not straightforward in general but there are some special cases for which solutions are explicit.

Case $d = 1$. As demonstrated in [AC11], the barycenter can be computed almost in closed form in the unidimensional case. Let μ be the barycenter of some measures $\nu_1, \ldots, \nu_N$ on the real line with finite second moments. Let $\lambda_1, \ldots, \lambda_N$ be positive reals summing to 1. We denote the barycenter $\mu = \mathrm{Bar}(\nu_i, \lambda_i)$ which can be characterized like this:

$$\mu = \left(\sum_{i=1}^{N} \lambda_i T_i^l\right)_{\#} \nu_l \quad \forall l \in \{1, \ldots, N\} \tag{4.3}$$

where T_i^l is the optimal transport between ν_i and ν_l.

According to Brenier's theorem [Vil+09], T_i^l exists and is unique. Moreover, we have that $T_i^l := F_i^{-1} \circ F_l$, where F_i is the cumulative function of ν_i i.e. $F_i(\alpha) = \nu_i(]-\infty, \alpha])$, and F_i^{-1} denotes the generalized inverse of F_i, $F_i^{-1}(t) := \inf\{\alpha : F_i(\alpha) \geq t\}$. Note that, this way of characterizing the barycenter with transports maps is also possible in a multidimensional case. But using cumulative functions is specific to the case $d = 1$.

Gaussian case. As proved by [AC11], for Gaussian inputs, the barycenter remains Gaussian. Closed-form covariance updates can be used. If all ν_i are centered Gaussian with covariance matrix S_i i.e. $\nu_i = \mathcal{N}(0, S_i)$ then barycenter is a Gaussian $\mathcal{N}(0, S)$ where S is positive definite and the unique root of the following matrix equation:

$$\sum_{i=1}^{N} \lambda_i (S^{1/2} S_i S^{1/2})^{1/2} = S$$

In other words, the root $S^\star$ is solution of the fixed point map equation $S = \Phi(S)$, where $\Phi(S) = \sum_{i=1}^{N} \lambda_i (S^{1/2} S_i S^{1/2})^{1/2}$. As emphasized in [PC19], iterating $S^{(l+1)} = \Phi(S^{(l)})$ allows to converge in practice to the solution $S^\star$.

3 Application to the generation of bags for Multiple instance learning

We propose to generate synthetic bags by interpolating between empirical measures using Wasserstein barycenters [AC11]. Given two bags represented as empirical distributions, we construct new bags by computing intermediate distributions along the Wasserstein geodesic. This provides a natural way to generate additional training data while preserving the geometric structure of the underlying feature space. To control variability and avoid support concentration, we restrict interpolation to pairwise barycenters.

3.1 Computing Wasserstein barycenter in the discrete setting

[Rab+11] were the first to give a serious computational attempt. They proposed computing barycenters of point clouds in $\mathbb{R}^d$ using the sliced Wasserstein distance and optimizing via gradient descent. The key idea is to avoid solving the full optimal transport problem, to project distributions onto 1D lines, and compute 1D Wasserstein distances (easy via sorting), average over random projections. However, this approach has several limitations. It works best in low dimensions, approximate (sliced Wasserstein)

and does not generalize easily to arbitrary metric spaces.

[CD14] introduced an efficient, scalable algorithm for computing true entropically-regularized Wasserstein barycenters. Let's suppose we have N discrete probability measures $\{\nu_1, \dots, \nu_N\}$. Each ν_i is described by a list of atoms $\xi^{(i)} \in \mathbb{R}^{m_i \times d}$, and probability vector $b_i \in \Delta_{m_i}$, *i.e.* $\sum_j b_{ij} = 1$. We have $\nu_i = \sum_{j=1}^{m_i} b_{ij} \delta_{\xi^{(i)}}$. We fix the barycenter support size to k. A barycenter of the measures $\nu_1, \dots, \nu_N$ is a minimizer of f defined as follows:

$$\min_{a,X} \sum_{i=1}^{N} \lambda_i \, \mathbf{p}(a, b_i, M^{(i)})$$

where $X = \{X_1, \dots, X_k\} \in \mathbb{R}^{k \times d}$, $a \in \Delta_k$, $M^{(i)}_{jl} = \left[\|X_j - \xi^{(i)}_l\|_2^2\right]_{jl} \in \mathbb{R}^{k \times m_i}$. $M^{(i)}$ is the distance matrix between the atoms of the barycenter μ and the measure ν_i. $\mathbf{p}(a, b_i, M^{(i)}) = \mathrm{W}_2^2(\mu, \nu_i) = \min_{T \in U(a,b_i)} \langle T, M^{(i)} \rangle$, $U(a, b_i)$ is a transportation polytope, $\langle A, B \rangle = \mathrm{tr}(A^\intercal B)$ is the Frobenius dot product of matrices.

Computing Wasserstein barycenters directly is computationally demanding. Each evaluation of the objective involves solving an optimal transport problem, typically formulated as a large-scale linear program. When combined with the outer optimization over the barycenter, this leads to a nested procedure that quickly becomes intractable for empirical measures with moderate or large support sizes. As a result, naive approaches based on exact optimal transport solvers are not practical in realistic settings, which motivates the need for scalable approximations.

To overcome the computational bottleneck, [CD14] proposes an entropic regularization of the optimal transport problem, leading to efficient algorithms based on Sinkhorn iterations [Cut13]. The idea is to add an entropy penalty to the transport plan, which encourages smoother and more diffuse couplings:

$$\min_T \langle T, M \rangle - \epsilon \, \mathrm{H}(T)$$

where $\mathrm{H}(T) = -\sum_{i,j} T_{ij} \log T_{ij}$ is the entropy of the transport plan. The parameter $\epsilon > 0$ controls the trade-off. Standard OT finds a sparse transport plan (mass moves along few paths), while entropic OT encourages spread-out, smooth transport plans. When $\epsilon \to 0$, it converges to true OT.

This regularization has two important consequences. First, it makes the objective function smooth and strictly convex with respect to the transport plans, which ensures uniqueness of the optimal coupling. Second, it transforms the optimal transport problem

into a matrix scaling problem. More precisely, the optimal transport plan can be written as:

$$T = \text{diag}(u)\, K\, \text{diag}(v),$$

where $K = \exp\left(-\frac{M}{\varepsilon}\right)$, and the scaling vectors u and v are computed using Sinkhorn iterations.

From this formulation, [CD14] derived algorithm 1.

Algorithm 1 Smoothed Primal $T_\epsilon^\star$ and Dual $\alpha_\epsilon^\star$ Optima

Require: $\xi \in \mathbb{R}^{m\times d},\ b \in \Sigma_m,\ \epsilon > 0$
  initialize $X \in \mathbb{R}^{d\times k}$
  $a =$ `ones(1,k)`
  $K = \exp(-M/\epsilon)$
  $\tilde{K} = \text{diag}(a^{-1})K$
  $u =$ `ones(1,m)`
  while u changes **do**
    $u = 1./(\tilde{K}(b./(K^\intercal u)))$
  end while
  $v = b./(K^\intercal u)$
  $T_\epsilon^\star = \text{diag}(u)K\,\text{diag}(v)$
  $\alpha_\epsilon^\star = -\epsilon\log(u) + \epsilon\frac{\log(u)^\intercal \mathbf{1}_n}{n}\mathbf{1}_n$

From algorithm 1, it is possible to recover the barycenter support:

$$X = (1-\theta)X + \theta\left(\frac{1}{N}\sum_{i=1}^{N}\xi^{(i)}T_{\epsilon,i}^{\star\intercal}\right)\text{diag}(a^{-1})$$

where $\theta \in [0,1]$, is preset or set with a line-search. This corresponds to a Newton step.

3.2 The proposed method for MIL data augmentation

In this section, we propose to apply Wasserstein barycenter computation to the generation of bags for multiple instance learning.

Suppose we have a set of labeled bags $\{(\mathcal{X}_i, Y_i); 1 \leq i \leq N\}$. Each bag $\mathcal{X}_i = \{X_1^{(i)}, \dots X_{m_i}^{(i)}\}$ is represented as an empirical measure $\nu_i = \frac{1}{m_i}\sum_{j=1}^{m_i}\delta_{X_j^{(i)}}$.

A new bag is generated by computing a barycenter of a pair of measures that have the same label Y_i. This pair of measures is sampled uniformly within the corresponding

group of measures. Variability is introduced by sampling the interpolation parameter λ which is selected randomly from a uniform distribution: $\lambda \sim \mathcal{U}([0,1])$. The barycenter is computed with fixed support size k and regularization parameter $\epsilon > 0$. The generated bag corresponds to the support points $X \in \mathbb{R}^{k \times d}$ of the discrete barycenter.

Algorithm 2 Synthetic bag generation procedure

Require: label $c \in \{0,1\}$, $\epsilon > 0$, k number of barycenter support points
Select randomly ν_{l_1} and ν_{l_2} within the group of measures $\{\nu_i; Y_i = c\}$
Select $\lambda \sim \mathcal{U}([0,1])$
Given ϵ and k, compute the regularized barycenter $\mu = \text{Bar}((\nu_{l_1}, 1-\lambda), (\nu_{l_2}, \lambda))$
return support $X \in \mathbb{R}^{k \times d}$ of μ

We now discuss the geometric properties that motivate the restriction to pairwise interpolation.

In this work, we restrict interpolation to pairs of distributions. The key idea behind our choice is that the Wasserstein barycenter behaves like an average of optimally matched points across distributions [PZ19]. The barycenter is determined by transport geometry, not density averaging. This explains why mass can align and compress. It is possible to characterize the barycenter with transport maps (4.2) (as shown in [AC11]). Because of the nonlinear geometry of W_2, Wasserstein barycenters behave differently from linear averages. This geometric structure is exactly what allows support concentration phenomena.

The validity of this approach relies on the assumption that local interpolations in Wasserstein space preserve class labels. In practice, one may further enforce locality by restricting interpolation to nearby distributions (e.g., nearest neighbors with respect to the Wasserstein distance), as commonly done in feature-space augmentation methods such as SMOTE [Cha+02]. In this work, however, we opt for uniform sampling within each class in order to keep the method simple.

By restricting interpolation to pairs of measures, we also approximate local displacements in Wasserstein space, generating intermediate distributions that remain close to the original inputs while preserving variability. This pairwise construction can be interpreted as a local exploration of the Wasserstein geometry, in contrast to global averaging.

4 Numerical experiments

Experimental setup. We evaluate our Wasserstein barycenter-based data augmentation strategy on a flow cytometry dataset from a Multiple Myeloma (MM) clinical study (introduced in Chapter 1). Each biological sample corresponds to a patient represented as a bag of approximately 1500 cells, described by seven markers (FSC, SSC, CD19, CD45, CD38, CD138, H2). The dataset includes 24 patients (10 early-stage, 14 diagnosed), with all features scaled to $[0, 1]$.

The regularization parameter ϵ and the number of barycenter support points k are selected empirically to balance computational efficiency and fidelity to the true barycenter. Larger values of ϵ produce overly smooth transport plans, while smaller values increase computational cost. Similarly, small k reduces representational capacity of the support. We set $\epsilon = 1e-3$ and $k = 1500$, which provides a good trade-off (see figures 4.2 and 4.3).

We tested the trade-off between reducing the computation time, approximating the true barycenter (ϵ), losing a minimal information on the support k). Large values of ϵ lead to overly smooth transport plans and loss of geometric fidelity, while small values increase computational cost. Similarly, small support sizes k reduce the representational capacity of the barycenter. Choosing $\epsilon = 1e-3$ and $k = 1500$ seems to give a good trade-off. See figure 4.2 and 4.3 for illustration of these choices.

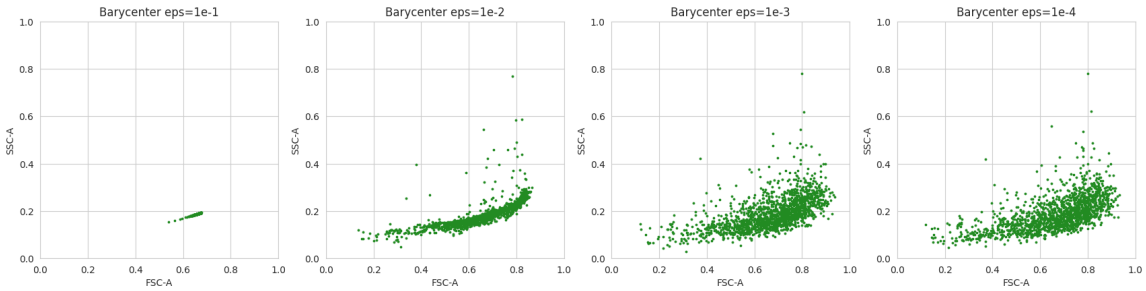


Figure 4.2: Influence of the regularization parameter ϵ on barycenter smoothness. Larger values yield smoother but less accurate barycenters, while smaller values preserve geometric structure at higher computational cost.

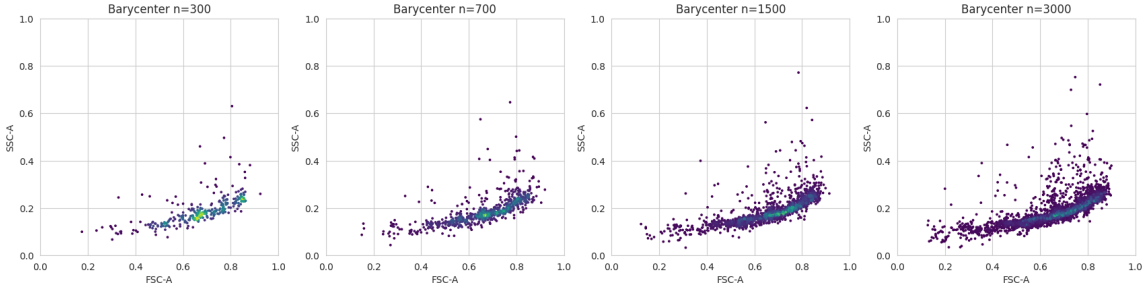


Figure 4.3: Influence of the number of support points k on Wasserstein barycenters. Increasing k improves the resolution of the estimated distribution and better captures fine-scale structures, whereas small k results in oversimplified barycenters.

Pairwise data augmentation and comparison baseline. The data augmentation method introduced in the previous section (algorithm 2), follows the pairwise interpolation procedure. Synthetic bags are generated by computing Wasserstein barycenters between pairs of empirical measures sampled uniformly within each class. In addition to this method, we consider a group interpolation strategy as a baseline for analysis purposes. In this setting, barycenters are computed using all measures within a class, with interpolation weights sampled from a Dirichlet distribution. This approach is not used for data augmentation in model training, but rather serves to illustrate the behavior of Wasserstein barycenters when averaging multiple distributions. The comparison between pairwise and group interpolation allows us to highlight the support concentration phenomenon discussed above, and provides empirical justification for restricting our method to pairwise barycenters.

- Pairwise interpolation: Barycenters are computed between pairs of patients. Pairs are sampled uniformly at random with replacement among patients of the same class.

- Group interpolation: Barycenters are computed using all patients within the same class (diagnosed MM or early stage), resulting in a single averaged distribution. Interpolation weights are sampled from a Dirichlet distribution $\mathrm{Dir}(\alpha)$, where α is a vector of ones.

Implementation details.[1] We compute barycenters using entropic regularization ($\epsilon = 10^{-3}$) with the Python library OTT-JAX [Cut+22]. The barycenter support size is fixed to $k = 1500$, matching the average number of cells per patient.

Qualitative analysis of Wasserstein interpolations. The group interpolation setting is used here solely to illustrate the concentration effect induced by averaging

[1]The source code is available at https://github.com/pamikem/distribution-learning.git

multiple measures, and is not used in the augmentation procedure. Figure 4.4 displays the distributions of the 14 patients diagnosed with Multiple Myeloma, projected onto the CD38–CD19 plane. Each point cloud corresponds to a patient distribution. Figure 4.5 illustrates the trajectory taken by the Wasserstein barycenter as the interpolation parameter varies. In figure 4.5a (pairwise interpolation), barycenters explore a wide region of the feature space. The variability (gray ellipses) remains comparable to that of the original distributions. In contrast, figure 4.5b (group interpolation) shows a strong concentration effect, where barycenters collapse toward a central region with reduced variance. Figure 4.5b shows a visibly reduced variance compared to 4.5a. These observations confirm the theoretical behavior discussed in the previous section: Wasserstein barycenters favor regions shared across input measures, possibly leading to variance shrinkage when averaging many distributions.

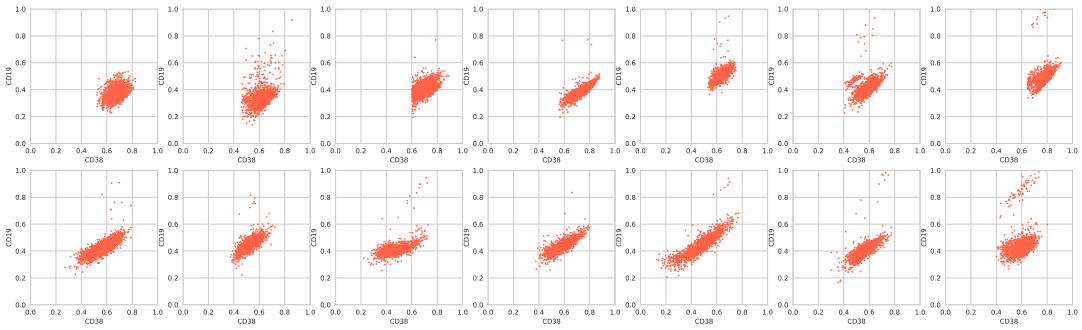
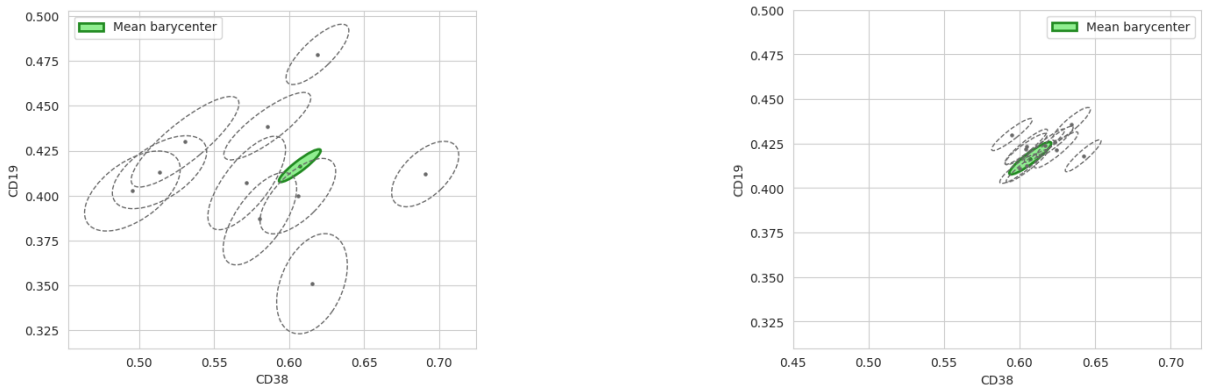


Figure 4.4: **Distribution of the 14 diagnosed Multiple Myeloma patients projected onto the CD38–CD19 feature space.** Each point cloud represents the empirical distribution of cells for a single patient, illustrating inter-patient variability in marker expression.

Impact on data augmentation. We evaluate whether Wasserstein-based augmentation improves classification performance using the AttentionMIL model [ITW18]. Results are reported over five repetitions of 2-fold cross-validation. For each class of bags, 15 synthetic bags are generated (125% increase in dataset size). We evaluate our data augmentation method in two different ways using the pairwise interpolation method. First, by doing a proper data augmentation and composing the training set of real data and synthetic data generated with Wasserstein barycenters. Second, by composing the training set with generated data only. Contrary to expectations, Wasserstein-based augmentation does not improve performance and often degrades it (see table 4.1). While recall increases to 1.0, precision drops substantially, indicating that the model over-predicts the positive class. This suggests that Wasserstein interpolation can alter class-discriminative structures rather than preserving them.



(a) (b)

Figure 4.5: **Comparison of Wasserstein barycenter interpolation strategies.** (a) Pairwise interpolation (between pair of patients in the same class) produces barycenters that explore a broad region of the feature space and preserve variability. (b) Group interpolation (all patients of a same class) leads to concentration of mass toward a central region, resulting in reduced variance. This illustrates the support collapse phenomenon when averaging multiple distributions.

Training data	Accuracy	AUC	F1	Precision	Recall
Real-only	0.68 ± 0.1	0.74 ± 0.14	0.77 ± 0.05	0.68 ± 0.11	0.93 ± 0.12
Real + Synthetic	0.58 ± 0.0	0.78 ± 0.06	0.74 ± 0.0	0.58 ± 0.0	1.0 ± 0.0
Synthetic-only	0.58 ± 0.0	0.81 ± 0.09	0.74 ± 0.0	0.58 ± 0.0	1.0 ± 0.0

Table 4.1: Classification performance of AttentionMIL [ITW18] on the Multiple Myeloma dataset under different training data configurations: (i) real-only training, (ii) training on real data augmented with Wasserstein-generated synthetic bags, and (iii) training on synthetic data only. Results are averaged over five repetitions of 2-fold cross-validation.

The experimental results highlight a limitation of standard Wasserstein interpolation in the context of data augmentation for MIL. In particular, the mass preservation constraint inherent to optimal transport may lead to overly smooth interpolations, which can dilute or distort discriminative structures within bags. This is especially problematic in settings where labels depend on sparse or localized patterns.

Conclusion of the numerical experiments

We investigated the use of Wasserstein barycenters as a tool for data augmentation in multiple instance learning. The proposed pairwise interpolation strategy provides a principled way to generate synthetic bags by exploiting the geometry of the Wasserstein

space. Qualitative analysis shows that this approach preserves variability and avoids the support concentration phenomenon observed when averaging multiple distributions with overlapping supports, motivating then the design choice of restricting interpolation to pairs.

However, the quantitative results highlight important limitations. In our experiments, Wasserstein-based augmentation does not consistently improve classification performance and may even degrade it. This suggests that, despite its appealing geometric properties, Wasserstein interpolation does not guarantee the preservation of class-discriminative structures. In particular, the mass preservation constraint inherent to optimal transport may lead to overly smooth interpolations, diluting localized patterns that are critical for prediction in MIL settings.

This observation points to a fundamental limitation of standard optimal transport for data augmentation. A natural extension is provided by unbalanced optimal transport [Chi+18], which relaxes the marginal constraints and allows for the creation or removal of mass during transport. This additional flexibility may enable better alignment between distributions with heterogeneous supports and help preserve important structures that are not shared across all samples.

Beyond this limitation, another important challenge lies in the discrete formulation of Wasserstein barycenters used in this work. Representing distributions as finite sums of Dirac masses restricts the barycenter to a fixed support and prevents direct sampling from a continuous underlying distribution. Moreover, the computational cost of discrete barycenter estimation scales with the support size, limiting its applicability for large-scale data augmentation.

These limitations motivate the exploration of continuous formulations of optimal transport and generative barycenter models, which aim to overcome the constraints of discrete representations. We discuss these perspectives in the next section.

5 Perspectives: Beyond discrete Wasserstein barycenters

The data augmentation method proposed in this chapter relies on the computation of Wasserstein barycenters in the discrete setting, where each measure is represented as a finite sum of Dirac masses. While this formulation is computationally tractable, it

introduces two fundamental limitations. First, representing distributions as empirical measures only provides a discrete approximation of their support. As a result, the computed barycenter is itself supported on a finite set of points and does not define a continuous distribution from which new samples can be drawn. This restricts its use as a generative model for data augmentation. Moreover, the computational cost of barycenter estimation scales with both the number of input samples and the size of the barycenter support. In practice, fixing a large support size is necessary to preserve geometric fidelity, but this quickly becomes prohibitive when generating a large number of synthetic samples. These limitations motivate the need for alternative formulations that move beyond discrete representations.

Continuous and stochastic formulations of optimal transport address these limitations by avoiding explicit discretization of the measures. In these approaches, distributions are represented implicitly through samples or parameterized models, and barycenters can be defined without fixing a support size in advance.

This shift enables the construction of generative models associated with Wasserstein barycenters, allowing the sampling of arbitrarily many synthetic distributions. However, these methods still rely on access to samples from the input distributions, which remains a fundamental requirement for estimating transport maps or barycenters in practice.

5.1 From discrete to continuous optimal transport: stochastic and neural approaches

The limitations of discrete Wasserstein barycenters have motivated the development of continuous and scalable formulations of optimal transport. A first line of work focuses on removing explicit discretization by reformulating optimal transport as a stochastic optimization problem.

A major step in this direction was introduced by [Gen+16], who showed that the dual formulation of entropic optimal transport can be expressed as the maximization of an expectation. This reformulation enables the use of stochastic gradient methods, avoiding the need to manipulate large transport matrices. In the continuous setting, they propose optimizing over functions in a reproducing kernel Hilbert space (RKHS), allowing transport problems to be solved directly from samples without requiring a fixed discretization grid. This approach significantly improves scalability and opens the door to high-dimensional applications. Conceptually, this marks a shift from a combinatorial view of optimal transport toward a functional and stochastic perspective, where transport is learned through optimization in function spaces rather than computed over discrete

supports.

A second line of work goes further by directly learning the optimal transport map between distributions. Instead of approximating transport plans or dual potentials, these approaches parameterize the transport map itself, enabling a generative view of optimal transport.

[Seg+17] proposed a neural network-based framework to learn transport maps, demonstrating applications in domain adaptation and generative modeling. Building on Brenier's theorem, [Mak+20] showed that, in the quadratic case, the optimal transport map can be expressed as the gradient of a convex function. They parameterize this function using Input Convex Neural Networks (ICNNs) [AXK17], ensuring that the learned map satisfies the structural properties of optimal transport.

These neural approaches offer several advantages over discrete barycenter methods. They eliminate the need to fix a support size, avoid the smoothing bias induced by entropic regularization, and provide an explicit generative mechanism that allows sampling arbitrarily many points from the transported distribution. However, they typically involve challenging optimization problems. To address this, [Kor+19] introduced Wasserstein-2 generative networks with cycle-consistency regularization, improving training stability and making neural optimal transport more practical.

In the context of data augmentation for MIL, these continuous and neural formulations are particularly appealing. They provide a way to move beyond finite-support barycenters and enable the generation of synthetic bags through learned transport maps, while still relying on samples from the input distributions for training.

5.2 Generative Wasserstein barycenters: continuous estimation and sampling

Recent advances in continuous optimal transport naturally extend to the computation of Wasserstein barycenters, enabling formulations that overcome the limitations of discrete support representations.

A first line of work focuses on the estimation of barycenters in the continuous setting. [Li+20] introduced a stochastic formulation of the regularized Wasserstein barycenter problem, where the barycenter is represented implicitly and optimized using sampling-based methods. This approach avoids fixing a discrete support grid and instead

constructs a continuous approximation of the barycenter directly from samples. In parallel, [CAD20] proposed techniques for estimating barycenters in high-dimensional settings, addressing challenges related to scalability and sample efficiency. These methods represent an important step toward practical barycenter computation beyond the discrete regime.

A second direction builds on neural optimal transport to obtain generative barycenter models. Leveraging parameterized transport maps, [FTC20] proposed a scalable framework for computing Wasserstein barycenters using Input Convex Neural Networks. In contrast to entropic approaches, their method does not rely on regularization and instead constrains dual potentials to be convex, preserving the structure of optimal transport. A key advantage of this formulation is that the barycenter is represented as a generative model, allowing sampling of arbitrarily many points from the learned distribution. Similarly, [Kor+21] introduced a method for continuous Wasserstein-2 barycenter estimation using ICNNs with cycle-consistency regularization. Their approach avoids minimax optimization while maintaining stability and scalability, further supporting the use of neural transport maps for barycenter computation. These methods provide a practical framework for learning barycenters as generative objects rather than finite sets of points.

Despite these advances, an important limitation remains. As highlighted in [Kor+21], the evaluation of neural optimal transport methods is still an open challenge. Good performance in downstream generative tasks does not necessarily imply accurate approximation of the true optimal transport map. This raises fundamental questions about the fidelity of the resulting barycenters to the underlying Wasserstein geometry.

In the context of data augmentation for multiple instance learning, generative barycenter models offer a promising direction. By combining continuous barycenter estimation with learned transport maps, they enable the generation of arbitrarily many synthetic bags while avoiding the constraints of fixed support representations. However, these approaches still rely on access to samples from the input distributions, which remains a fundamental requirement for learning transport-based models.

While this chapter has focused on data augmentation through Wasserstein barycenters, the results highlight both the potential and the limitations of this approach. More importantly, they suggest that the role of Wasserstein geometry in multiple instance learning extends beyond interpolation between distributions. In particular, the limited ability of barycentric interpolation to preserve discriminative structures indicates that the key information may not lie in intermediate distributions, but rather in the structural relationships between distributions.

This observation motivates a broader perspective. Instead of viewing Wasserstein barycenters solely as a tool for generating synthetic data, they can be interpreted as a means of modeling shared structure and variability across distributions. In the following section, we briefly outline several perspectives and applications of Wasserstein barycenters that go beyond data augmentation and help situate our approach within a wider methodological framework.

6 Broader perspectives on Wasserstein barycenters

The data augmentation framework developed in this chapter relies on Wasserstein barycenters as a mechanism for interpolating between empirical distributions. However, this interpretation represents only one facet of their potential. More generally, Wasserstein barycenters provide a powerful tool for analyzing collections of distributions, capturing shared structure, and modeling variability in a geometrically meaningful way.

A first perspective is to interpret Wasserstein barycenters as estimators of a reference or template distribution [BLL15]. In this view, each observed distribution is modeled as a deformation of an underlying latent structure, and the barycenter captures this shared geometry. This interpretation is particularly relevant in biological applications such as single-cell analysis, where inter-patient variability and batch effects can be understood as distributional deformations around a common biological core.

Once such a reference distribution is defined, each sample can be represented through its associated optimal transport map. This enables a geometric analysis of variability directly in Wasserstein space, where differences between distributions are described as transformations rather than pointwise discrepancies. This perspective underlies statistical approaches such as Wasserstein principal component analysis [SC15; Big+17], which provide structured tools for analyzing variability across populations.

In this context, recent advances in optimal transport shift the focus from interpolation toward the learning of mappings between distributions. For instance, the CellOT framework [Bun+23] learns transport maps between unperturbed and perturbed cellular populations, enabling the prediction of treatment effects at the single-cell level. Similarly, conditional optimal transport models such as CONDOT [BKC22] incorporate contextual variables to learn families of transport maps, allowing the prediction of responses under unseen conditions.

These approaches highlight a broader paradigm: besides generating synthetic data through barycentric interpolation, one can model meaningful transformations between distributions. In applications such as biological data analysis, this shift provides a more expressive and interpretable framework for capturing variability, predicting responses to perturbations, and understanding the structure of complex populations.

Chapter 5

Bayesian classification in Multiple Instance Learning

Individual cells can be viewed as samples drawn from an underlying distribution specific to each patient. As shown in Chapter 1, while these distributions vary across patients, the dependency structure between markers can remain stable. The differences between patients might then be captured at the level of distributions rather than individual observations (cells). In this chapter, we propose a Bayesian approach to MIL that explicitly models distributions of cells. To address the variability across patients, we estimate these distributions by pooling cells from patients belonging to the same group (e.g., healthy or diseased). The resulting classifier evaluates the likelihood of the observed cells under each group. Unlike standard Naive Bayes approaches, our method incorporates a structured dependency model between features, based on partitions of variables learned in Chapter 1, allowing us to capture relevant interactions. The proposed framework combines non-parametric density estimation (via kernel density estimation) with a Bayesian decision rule adapted to the MIL setting. This leads to a flexible model capable of capturing complex, multimodal distributions. In addition to strong classification performance, the model provides interpretable instance-level contributions, enabling the identification of discriminative regions of the feature space and potentially revealing biologically meaningful cell populations associated with the disease.

1 Bayes classifier

The Bayes classifier provides a probabilistic framework for classification. Given an observation x and a class label c, the optimal decision rule is to assign x to the class with

highest posterior probability:

$$\hat{y} = \arg\max_{c} P(Y = c \mid X = x).$$

Using Bayes' theorem, this posterior can be expressed as:

$$P(Y = c \mid X = x) \propto P(X = x \mid Y = c)\, P(Y = c),$$

where $P(Y = c)$ denotes the class prior and $P(X = x \mid Y = c)$ the class-conditional likelihood.

In practice, estimating the full joint distribution $P(X \mid Y)$ is often difficult, especially in high-dimensional settings. The *Naive Bayes classifier* addresses this issue by introducing a simplifying assumption: the features are conditionally independent given the class label. If $X = (X_1, \ldots, X_d)$, this leads to:

$$P(X = x \mid Y = c) = \prod_{j=1}^{d} P(X_j = x_j \mid Y = c).$$

Despite the fact that this assumption is rarely satisfied by real-world data, Naive Bayes often performs remarkably well in practice.

A classical application of Naive Bayes is document classification, where texts are represented using a bag-of-words model [Har54; ZJZ10]. In this representation, a document is viewed as an unordered collection of words, and each word occurrence is assumed independent given the document class. This approach has been successfully applied to tasks such as spam detection and topic classification, illustrating that strong independence assumptions can still yield effective classifiers when combined with appropriate statistical estimation.

In our setting, we have a collection of instances (cells), and labels are defined at the bag level (patient). As discussed in Chapter 1, patients differ through the distributions from which their cells are drawn, while the dependency structure between markers remains stable across individuals. This variability across patients introduces an additional challenge: even within the same group (e.g., healthy or diseased), individual patients may exhibit different empirical distributions due to biological and sampling variability. To address this issue, rather than modeling each patient-specific distribution independently, we propose to estimate group-level distributions of patients by aggregating information across patients belonging to the same group. In other words, we learn a joint distribution of cells for each group, which captures the common statistical structure while smoothing

out patient-specific fluctuations.

Importantly, unlike the standard Naive Bayes classifier, we do not assume mutual independence between features. As shown in Chapter 1, markers can exhibit structured dependencies that carry relevant information and remain stable across patients. To account for this, we introduce a structured factorization of the distribution based on a partition of variables: features within the same group may be dependent, while different groups are treated as independent. This provides a flexible alternative to the Naive Bayes assumption while preserving tractability. Classification can then be performed by evaluating how likely the set of cells observed for a given patient is under each group-level distribution.

2 A Bayes Classifier for MIL

2.1 Intuition

Our goal is to distinguish patients based on the distributions from which their cells are drawn. Concretely, we compare how cell densities differ across disease conditions. Figure 5.1 illustrates this point using 2D histograms from FlowCAP-AML data (introduced in 1.1). Although the overall support of the distributions is similar, certain regions of the space show clear differences in concentration. Some areas contain a higher density of cells from AML subjects (sick patients), while others are more populated for healthy patients. These local differences provide discriminative information. Importantly, we do not assume prior knowledge about which regions are relevant; identifying such regions is precisely one of the objectives of the model. In particular, regions where the cells of sick patients are significantly more concentrated than the cells of healthy patients, may correspond to abnormal cell populations associated with the disease.

We adopt the following perspective: cells from patients belonging to the same group (healthy or diseased) are viewed as independent samples from a common distribution. This assumption allows us to pool cells across patients within each group, thereby mitigating patient-level variability and enabling robust estimation of group-level distributions. Under this model, a patient is represented as a finite sample drawn from the corresponding group distribution, and classification amounts to evaluating how likely this sample is from each group.

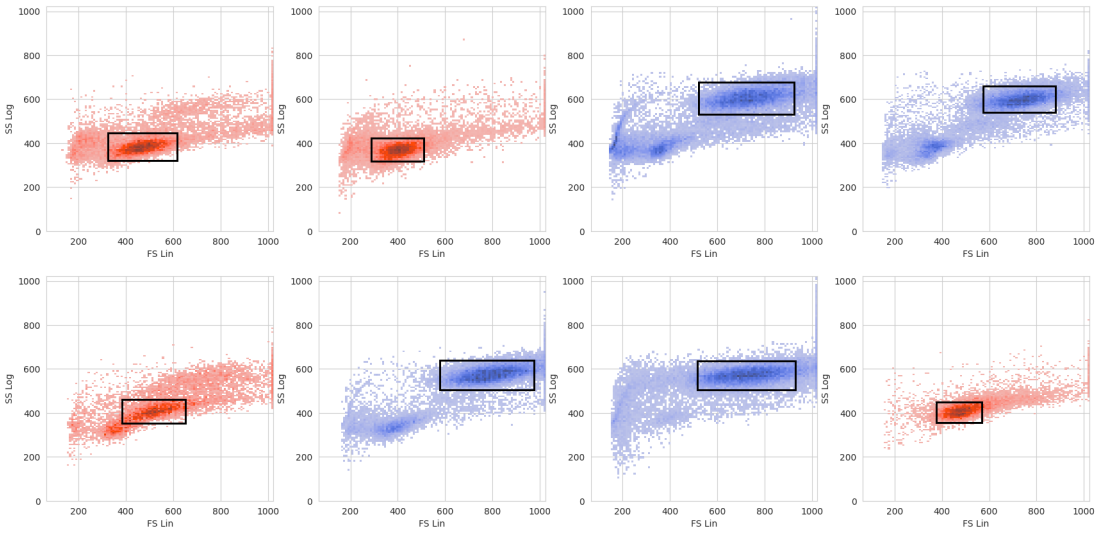


Figure 5.1: Comparison of cell density distributions between AML (red) and non-AML (blue) patients in a two-dimensional feature space. Each panel shows a 2D Histogram (FSC vs SSC) on cells pooled from patients of the same class. Although the global support of the distributions is similar, clear local differences appear: certain regions exhibit higher cell concentration in AML patients, while others are more populated in non-AML patients. These localized density shifts illustrate that discrimination between classes can be achieved by comparing where cells concentrate rather than relying on global separation.

2.2 The proposed method

We now formalize the probabilistic model. Let $\{(\mathcal{X}_i, Y_i); 1 \leq i \leq n\}$ be a set of labeled bags ($Y_i \in \{0, 1\}$). Each bag $\mathcal{X}_i$ has m_i instances $X_j^{(i)} \in \mathbb{R}^d$. We would like to model $\mathbf{P}(Y_i|\mathcal{X}_i)$. Following the intuition developed previously, we assume that for each class $c \in \{0, 1\}$, there exists a patients group-specific distribution $\mathbf{P}(\mathcal{X}|Y = c)$ and the instances within a bag are independent and identically distributed conditional on the class label. Under these assumptions and by using Bayes formula, the likelihood of a bag can be factorized as follows:

$$\begin{aligned}
\mathbf{P}(Y_i|\mathcal{X}_i) &= \frac{\mathbf{P}(\mathcal{X}_i|Y_i)\mathbf{P}(Y_i)}{\mathbf{P}(\mathcal{X}_i)} \\
\mathbf{P}(Y_i|\mathcal{X}_i) &\propto \mathbf{P}(\mathcal{X}_i|Y_i)\mathbf{P}(Y_i) \\
&\propto \mathbf{P}(Y_i) \prod_{j=1}^{m_i} \mathbf{P}(X_j^{(i)}|Y_i)
\end{aligned}$$

To model the distribution of individual instances, we incorporate the dependency structure as studied in Chapter 1. Let $\mathcal{P}$ be a partition of the feature set $\{1, \cdots, d\}$, where each subset $S \in \mathcal{P}$ groups variables with strong dependencies. We assume that

features are independent across groups. This leads to the factorization:

$$\mathbf{P}(X_j^{(i)}|Y_i) = \prod_{S\in\mathcal{P}} \mathbf{P}(X_{jS}^{(i)}|Y_i)$$

To estimate the distribution of each patients group, we adopt a non-parametric approach based on kernel density estimation (KDE). This choice is motivated by both empirical and practical considerations. The cell distributions exhibit complex, potentially multimodal structures that are not well captured by simple parametric models. In particular, experiments conducted in Chapter 2 with Gaussian mixture models showed that Gaussian assumptions can be too restrictive to accurately represent the data. Similar parametric approaches have been explored in the multiple-instance learning literature, for example through a Naive Bayes classifier with Gaussian distributions at the instance level [Xu03], but such assumptions may fail to capture subtle patterns in the data. KDE provides a flexible alternative that can capture localized variations in density without imposing strong structural assumptions. Let $G_1, \ldots, G_N$ denote the set of observed cells from class c. Then, for a given group S of features, the density is estimated as:

$$\mathbf{P}(x_S|Y_i = c) = \frac{1}{(2\pi)^{d/2} N h_S^d} \sum_{l=1}^{N} \mathrm{K}\left(\frac{G_{l,S} - x_S}{h_S}\right)$$

where $\mathrm{K}(u) = \exp\left(-\frac{1}{2}\|u\|^2\right)$, and h_S is the bandwidth for the group S.

To classify a bag $\mathcal{X}_i$, we compare posterior probabilities:

$$\begin{aligned}
&\mathbf{P}(Y_i = 1|\mathcal{X}_i) > \mathbf{P}(Y_i = 0|\mathcal{X}_i) \\
\Leftrightarrow\ &\mathbf{P}(Y_i = 1)\mathbf{P}(\mathcal{X}_i|Y_i = 1) > \mathbf{P}(Y_i = 0)\mathbf{P}(\mathcal{X}_i|Y_i = 0) \\
\Leftrightarrow\ &\frac{\mathbf{P}(Y_i = 1)}{\mathbf{P}(Y_i = 0)} > \prod_j \frac{\mathbf{P}(X_j^{(i)}|Y_i = 0)}{\mathbf{P}(X_j^{(i)}|Y_i = 1)} \\
\Leftrightarrow\ &\sum_j \log \mathbf{P}(X_j^{(i)}|Y_i = 0) - \log \mathbf{P}(X_j^{(i)}|Y_i = 1) < \log \mathbf{P}(Y_i = 1) - \log \mathbf{P}(Y_i = 0)
\end{aligned}$$

Let $\alpha = \mathbf{P}(Y = 1)$ denote the class prior. We estimate α as $\hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{Y_i=1}$ the ratio of positive bags in the training set. Note that there are no label on individual instances. We define the classification criterion for a bag $\mathcal{X}_i$ as:

$$\mathcal{L}(\mathcal{X}_i) = \log\left(\frac{\hat{\alpha}}{1-\hat{\alpha}}\right) - \sum_j \log \frac{\mathbf{P}(X_j^{(i)}|Y_i = 0)}{\mathbf{P}(X_j^{(i)}|Y_i = 1)}$$

We predict:

$$\hat{Y}_i = \begin{cases} 0 & \text{if } \mathcal{L}(\mathcal{X}_i) > 0 \\ 1 & \text{otherwise} \end{cases} \tag{5.1}$$

Note that the threshold on $\mathcal{L}(\mathcal{X}_i)$ used in the above decision rule, can also be chosen according to the ROC curve (false positive rate vs true positive rate).

Finally, this formulation naturally provides a measure of the contribution of each instance to the classification of a bag:

$$\ell(X_j^{(i)}) = \log \frac{\mathbf{P}(X_j^{(i)}|Y_i = 1)}{\mathbf{P}(X_j^{(i)}|Y_i = 0)}$$

This quantity reflects how strongly a given cell contributes to the classification of a patient and will be used for interpretation in the next section.

3 Numerical experiments

We evaluate the MIL Bayes classifier on the FlowCAP-AML dataset. The dataset consists of 359 patients (316 healthy vs 43 diagnosed with AML). We split the data at the patient level into 70% for training and 30% for testing. Patients group conditional densities are estimated using kernel density estimation (KDE). The bandwidth parameters $(h_S)_{S\in\mathcal{P}}$ is selected via cross-validation. We use a fixed partition of variables obtained from the ISDE algorithm [Puj22], as identified in section 3.2 of Chapter 1. Densities are evaluated on a grid of 3000 points. We use standard classification metrics at the patient level: Accuracy, F1-score, and Area Under the ROC Curve (AUC).

Classification performance. The proposed method[1] achieves near-perfect performance on this dataset. On the training set, we obtain high accuracy (0.99) and F1-score (0.97), along with strong AUC values (0.98). On the test set, performance remains equally strong, with all metrics reaching values close to 1. These results indicate that the model is able to effectively discriminate between healthy and diseased patients based only on the distributions of their cells. The absence of a noticeable gap between training and test performance suggests that the model generalizes well and does not suffer from overfitting, despite the class imbalance (0.12 sick and 0.88 healthy patients).

[1]The code used to reproduce the experiments, is publicly available at https://github.com/pamikem/not-naive-bayes-MIL.git

Interpretation at the cell level. A key advantage of the proposed Bayesian formulation is its interpretability at the instance level. Each cell contributes additively to the classification decision through its log-likelihood ratio $\ell(x)$ defined in the previous section. Using this score, we can rank the cells of a patient according to their individual contribution. Cells with the largest positive values of $\ell(x)$ provide the strongest evidence in favor of the group of sick patients, while negative values support the group of healthy patients. To identify the most contributing cells, we rank instances by their score. A subset of top-contributing cells is then selected using a threshold based on the decision rule (5.1). Figure 5.2 illustrates the regions of the feature space associated with high contributions. Cells (in red) located in these regions play a central role in the classification decision. Removing them significantly alters the predicted status of the patient. This confirms that the model captures differences between the two groups of patients. Importantly, these regions can be interpreted as candidates for abnormal cell populations. Rather than imposing prior knowledge, the model identifies them automatically based on statistical differences between healthy and diseased distributions.

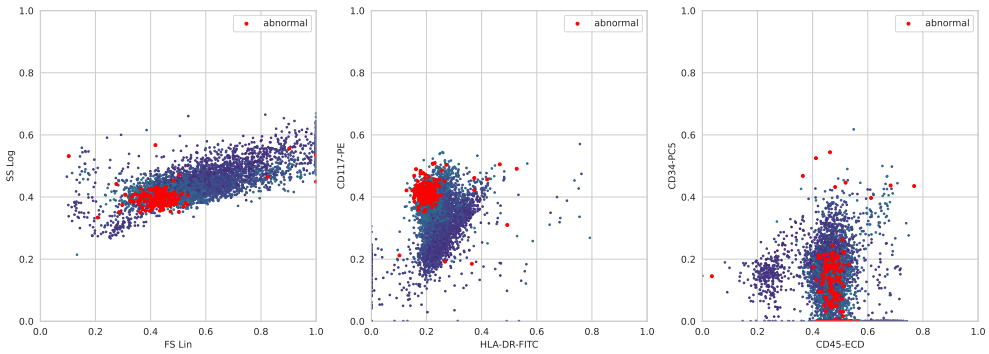


Figure 5.2: **Visualization of cell-level contributions to the classification of a sick patient.** Each cell is assigned a score based on its likelihood ratio between AML (sick) and non-AML (healthy) density models. Regions labeled as "abnormal" correspond to areas where cells have a high contribution to the classification decision. Cells are ranked according to their individual contribution, and a subset of top-contributing cells is selected using a threshold derived from the F1-score criterion. These cells represent the most influential instances: removing them significantly alters the predicted label of the patient, highlighting their central role in the decision process.

Impact of the dependency structure of variables. To assess the importance of modeling dependencies between variables, we compare the performance of the classifier for different partitions of variables. These partitions correspond to alternative assumptions about the dependency structure, ranging from structured groupings (as obtained via ISDE) to the fully independent case (Naive Bayes). The results (see table 5.1)

show that the partition derived from the ISDE procedure consistently yields the best performance. Alternative partitions lead to slightly lower, but still competitive, results. Interestingly, even the fully independent model performs well on this dataset, indicating that the classification task is relatively robust to the choice of dependency structure of variables. However, the best results are obtained when incorporating the empirically learned partition.

	Accuracy		**F1 score**		**AUC**	
	Train	Test	Train	Test	Train	Test
Partition 1	**0.99**	**1.0**	**0.97**	**1.0**	**0.98**	**1.0**
Partition 2	0.99	1.0	0.95	1.0	0.98	1.0
Partition 3	0.98	0.99	0.91	0.96	0.97	0.99
All independent (Naive Bayes)	0.97	1.0	0.89	1.0	0.97	1.0

Table 5.1: Comparison of the performance of the proposed classifier under different assumptions on the dependency structure between variables. The results show that all configurations achieve strong performance, including the fully independent model corresponding to Naive Bayes. However, the best partition (partition 1) obtained via ISDE [Puj22] consistently yields the best results. Partitions 2 and 3 are less optimal (log-likelihood-wise) but still yield strong classification performance too.

Beyond cell-level analysis, the proposed framework also provides insight at the patient level. Figure 5.3 presents radar plots summarizing the average expression of abnormal cells identified on each biological marker. The belief score corresponds to the log-likelihood ratio over all cells of a patient. These visualizations highlight different patterns related to the abnormal cells and reinforce the fact that these cells can behave differently from one patient to another. These findings are consistent with known cytometric patterns for AML diagnosis reported in the literature [WB01; BKA+19].



Figure 5.3: **Radar plots showing the profile (average expression on each biological marker) of the abnormal cells identified in AML diagnosed patients.** Each subfigure corresponds to a single diseased patient and displays a radar plot summarizing the average expression of the selected cells across different markers. Each axis corresponds to a biological marker. The belief score represents the likelihood ratio aggregated over all cells of a patient, quantifying how strongly the data supports one class (sick) over the other (healthy). In addition, "top instances" correspond to the subset of cells with the highest individual contributions. Comparing the top instances to the less contributing instances reveals whether the classification is driven by global distributional differences or by a small number of highly informative cells.

Conclusion

The analysis of flow cytometry data illustrates a fundamental challenge in modern machine learning. Observations are collected at a fine-grained level (individual cells) while clinically, decisions are formulated at a higher level (patients). This mismatch, combined with the heterogeneity and variability of biological systems, makes classical supervised approaches difficult to apply directly. In this context, this thesis has explored the Multiple Instance Learning (MIL) framework as a principled way to bridge these two levels of information, with the dual objective of improving patient-level classification and uncovering meaningful cellular patterns associated with disease.

1 Conclusions on the research work

A first contribution of this work lies in the characterization of the data itself. Through an exploratory analysis of cellular distributions at the patient level, we highlighted both the intra-patient variability and the dependencies between biological variables. These observations emphasize that flow cytometry data cannot be adequately modeled using simplistic assumptions, and that understanding their statistical structure is a necessary step toward designing relevant learning methods.

Building on these insights, we developed several modeling strategies within the MIL framework. A probabilistic approach based on Gaussian mixtures was first introduced to represent patients as combinations of latent cellular populations, enabling the identification of substructures associated with disease while maintaining interpretability. In parallel, we proposed a local MIL framework based on neural networks, which departs from standard global (bag-level) aggregation strategies by learning instance-level representations and structuring them into meaningful regions of a latent space. Together, these approaches illustrate two complementary directions: one grounded in probabilistic modeling, the other in representation learning, both aiming to reconcile predictive performance with interpretability at the cellular level.

Beyond these models, this work also explored extensions of the MIL framework by adopting a distributional perspective on the data. By interpreting each patient as an empirical distribution of cells, we introduced a data augmentation strategy based on Wasserstein barycenters, enabling the generation of synthetic samples directly in the space of distributions. Finally, we proposed a Bayesian classifier for MIL, in which cells distributions are modeled by group of patients, allowing for flexible modeling of complex dependencies and providing interpretable contributions at the instance level.

Taken together, these contributions support a common perspective for analyzing flow cytometry data for clinical purposes. Rather than viewing the data as collections of invidual observations (cells), it is interesting to consider them as sets of observations coming from different patients. From this standpoint, Multiple Instance Learning appears as an appropriate framework for linking local variability to global decisions. More broadly, this work suggests that learning from distributions provides a unifying paradigm for addressing problems where data are organized as sets of observations and annotated at a more global level than traditionally at the level of observations.

2 Perspectives

The results of this thesis highlight the potential of learning from set-structured data. While the proposed approaches demonstrate that Multiple Instance Learning can effectively bridge cell-level observations and patient-level decisions, they also reveal several open challenges related to modeling heterogeneity, integrating distributional representations, and evaluating model reliability. These challenges naturally suggest several directions for future research.

Modeling inter-patient variability. A first key direction involves modeling the biological variability between patients. Although the approaches developed in this thesis capture cellular subpopulations through structured mechanisms such as mixture models or latent space partitioning, they remain limited in their ability to represent complex, patient-specific variability. In particular, the relationship between cellular patterns and patient-level outcomes is assumed to be shared across individuals, which may not fully reflect the diversity of biological processes. Recent work, such as MixMIL [Eng+23], proposes to address this limitation by introducing generalized linear mixed-effects models [BC93; Bat+15] within the MIL framework. In this setting, the contribution of cellular representations to the final prediction is modeled as a random, patient-dependent effect, allowing for a more flexible description of heterogeneity and uncertainty. This perspective

suggests a promising direction: combining representation learning (neural networks), aggregation mechanisms (MIL with attention), and probabilistic modeling (generalized linear mixed models) within a unified framework.

Towards a unified framework: learning from distributional and set-structured data. A second, more fundamental direction is to situate Multiple Instance Learning within the broader framework of learning from data structured as probability measures. Throughout this thesis, several contributions implicitly adopt this perspective by modeling patients as empirical distributions of cells, rather than as finite sets of independent observations. This viewpoint is particularly natural in a context where variability between patients is more meaningfully captured at the distributional level than at the level of individual cells. This perspective connects MIL to a rich body of work on distribution regression, initiated in [Póc+13], and further developed through kernel-based approaches and mean embeddings [Mua+12; Mua+17], scalable projection methods [Oli+14], and theoretical analyses with consistency guarantees [Sza+15; Sza+16]. In this framework, each observation is treated as a probability measure, and the learning task consists in mapping distributions to labels. The authors of [DR16; DLR16] show that MIL can be interpreted as a special case of this setting. This unification offers a powerful conceptual advantage: it allows different approaches like kernel methods, deep learning architectures for sets, and optimal transport techniques to be interpreted within a common mathematical framework. From this point of view, classical supervised learning appears as a particular case in which each distribution is reduced to a Dirac measure. More broadly, this suggests that learning from distributional data provides a natural and general paradigm for handling data that are annotated at a distribution level, and may serve as a foundation for future developments in both theory and applications.

Evaluation metrics, robustness, and uncertainty. A third important direction concerns model evaluation, robustness, and uncertainty quantification. The MIL framework is inherently hierarchical, with data generated through a two-level process (instances $\rightarrow$ bag $\rightarrow$ label). However, evaluation is most often performed using global metrics at the bag level, such as accuracy or AUC score, which do not fully reflect the quality of the internal modeling of the data. In particular, these metrics fail to capture several important aspects: the relevance and stability of instance-level contributions, the consistency of the aggregation mechanism, and the faithfulness of the learned representations to the underlying data structure. Recent works have emphasized the need for evaluation strategies that account for instance-level explanations and internal model behavior [Jav+22; EER22]. Moreover, the distributional nature of the data introduces an additional source of variability: predictions may depend not only on the model parameters, but also on the specific sample of instances observed for each

bag. This phenomenon, related to the two-stage sampled problem, has been studied in [Sza+15; Sza+16]. In this context, methods such as conformal prediction [VGS05; AB23] can provide a promising framework for quantifying uncertainty, by producing prediction sets with statistical guarantees rather than point estimates. More broadly, improving the reliability of MIL models requires developing tools that jointly assess predictive performance, interpretability, and robustness to sampling variability. This is particularly important in biomedical applications, where decisions must be both accurate and trustworthy.

These directions reflect a broader evolution in machine learning: a shift from learning with fully annotated, independent observations toward frameworks capable of handling structured, heterogeneous, and weakly annotated data. In many real-world settings, particularly in biology, fine-grained annotations are costly, not always reliable, or simply unavailable. In this context, learning from distributions or structured data is an opportunity to discover new patterns that remain inaccessible to traditional supervised approaches.

Flow cytometry References

[Agh+13] Nima Aghaeepour et al. "Critical assessment of automated flow cytometry data analysis techniques". In: *Nature methods* 10.3 (2013), pp. 228–238.

[Ami+13] El-ad David Amir et al. "viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia". In: *Nature biotechnology* 31.6 (2013), pp. 545–552.

[Bio25a] Creative Biolabs. *Immunofluorescence*. https://www.antibody-creativebiolabs.com/immunofluorescence-if.htm. Accessed: 2025-11-27. 2025.

[Bio25b] AAT Bioquest. *Fundamentals of Flow Cytometry*. https://www.aatbio.com/resources/assaywise/2019-8-1/fundamentals-of-flow-cytometry. Accessed: 2025-11-27. 2025.

[BKA+19] Maria Basharat, Saleem Ahmed Khan, Dawood Ahmed, et al. "Immunophenotypic characterisation of morphologically diagnosed cases of Acute Myeloid Leukaemia (AML)". In: *Pakistan journal of medical sciences* 35.2 (2019), p. 470.

[Eck+20] Jan-Niklas Eckardt et al. "Application of machine learning in the management of acute myeloid leukemia: current practice and future prospects". In: *Blood Advances* 4.23 (2020), pp. 6077–6085.

[Ng+24] David P Ng et al. "Recommendations for using artificial intelligence in clinical flow cytometry". In: *Cytometry Part B: Clinical Cytometry* 106.4 (2024), pp. 228–238.

[Now+19] Malgorzata Nowicka et al. "CyTOF workflow: differential discovery in high-throughput high-dimensional cytometry datasets". In: *F1000Research* 6 (2019), p. 748.

[Rus+19] Marc A Russo et al. "Expansion and activation of distinct central memory T lymphocyte subsets in complex regional pain syndrome". In: *Journal of neuroinflammation* 16.1 (2019), p. 63.

[Sal+19] Haneen T Salah et al. "Machine learning applications in the diagnosis of leukemia: Current trends and future directions". In: *International journal of laboratory hematology* 41.6 (2019), pp. 717–725.

[Sam+16] Nikolay Samusik et al. "Automated mapping of phenotype space with single-cell data". In: *Nature methods* 13.6 (2016), pp. 493–496.

[Sha05] Howard M Shapiro. *Practical flow cytometry*. John Wiley & Sons, 2005.

[Van+15] Sofie Van Gassen et al. “FlowSOM: using self-organizing maps for visualization and interpretation of cytometry data”. In: *Cytometry Part A* 87.7 (2015), pp. 636–645.

[WB01] Edward G Weir and Michael J Borowitz. “Flow cytometry in the diagnosis of acute leukemia”. In: *Seminars in hematology*. Vol. 38. 2. Elsevier. 2001, pp. 124–138.

Multiple instance learning References

[Alp+15] Ethem Alpaydın et al. “Single-vs. multiple-instance classification”. In: *Pattern recognition* 48.9 (2015), pp. 2831–2838.

[Amo13] Jaume Amores. “Multiple instance classification: Review, taxonomy and comparative study”. In: *Artificial intelligence* 201 (2013), pp. 81–105.

[ATH02] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. “Support vector machines for multiple-instance learning”. In: *Advances in neural information processing systems* 15 (2002).

[Bab+08] Boris Babenko et al. “Simultaneous learning and alignment: Multi-instance and multi-pose learning”. In: *Workshop on Faces in'Real-Life'Images: Detection, Alignment, and Recognition*. 2008.

[Bab08] Boris Babenko. “Multiple instance learning: algorithms and applications”. In: *View Article PubMed/NCBI Google Scholar* 19 (2008).

[BK98] Avrim Blum and Adam Kalai. “A note on learning from multiple-instance examples”. In: *Machine learning* 30.1 (1998), pp. 23–29.

[Car+18] Marc-André Carbonneau et al. “Multiple instance learning: A survey of problem characteristics and applications”. In: *Pattern Recognition* 77 (2018), pp. 329–353.

[CBW06] Yixin Chen, Jinbo Bi, and James Ze Wang. “MILES: Multiple-instance learning via embedded instance selection”. In: *IEEE transactions on pattern analysis and machine intelligence* 28.12 (2006), pp. 1931–1947.

[CDP19] Veronika Cheplygina, Marleen De Bruijne, and Josien PW Pluim. “Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis”. In: *Medical image analysis* 54 (2019), pp. 280–296.

[Che+14] Veronika Cheplygina et al. “Classification of COPD with multiple instance learning”. In: *2014 22nd International Conference on pattern recognition*. IEEE. 2014, pp. 1508–1513.

[Che+15] Veronika Cheplygina et al. “Label stability in multiple instance learning”. In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2015, pp. 539–546.

[CTL15] Veronika Cheplygina, David MJ Tax, and Marco Loog. “Multiple instance learning with bag dissimilarities”. In: *Pattern recognition* 48.1 (2015), pp. 264–275.

[DLL97] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. “Solving the multiple instance problem with axis-parallel rectangles”. In: *Artificial intelligence* 89.1-2 (1997), pp. 31–71.

[DLR16] Gary Doran, Andrew Latham, and Soumya Ray. “A Unifying Framework for Learning Bag Labels from Generalized Multiple-Instance Data.” In: *IJCAI*. 2016, pp. 1425–1431.

[Doo+02] Daniel R Dooly et al. “Multiple-instance learning of real-valued data”. In: *Journal of Machine Learning Research* 3.Dec (2002), pp. 651–678.

[DR14] Gary Doran and Soumya Ray. “A theoretical and empirical analysis of support vector machine methods for multiple-instance classification”. In: *Machine learning* 97.1 (2014), pp. 79–102.

[DR16] Gary Doran and Soumya Ray. “Multiple-instance learning from distributions”. In: *Journal of Machine Learning Research* 17.128 (2016), pp. 1–50.

[EER22] Joseph Early, Christine Evers, and Sarvapali Ramchurn. “Model agnostic interpretability for multiple instance learning”. In: *arXiv preprint arXiv:2201.11701* (2022).

[Eng+23] Jan P Engelmann et al. “Mixed models with multiple instance learning”. In: *arXiv preprint arXiv:2311.02455* (2023).

[FF10] James Foulds and Eibe Frank. “A review of multi-instance learning assumptions”. In: *The knowledge engineering review* 25.1 (2010), pp. 1–25.

[Fou08] James Richard Foulds. “Learning instance weights in multi-instance learning”. PhD thesis. The University of Waikato, 2008.

[Gär+02] Thomas Gärtner et al. “Multi-instance kernels”. In: *ICML*. Vol. 2. 3. 2002, p. 7.

[ITW18] Maximilian Ilse, Jakub Tomczak, and Max Welling. “Attention-based deep multiple instance learning”. In: *International conference on machine learning*. PMLR. 2018, pp. 2127–2136.

[Jav+22] Syed Ashar Javed et al. “Additive mil: Intrinsically interpretable multiple instance learning for pathology”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 20689–20702.

[JK24] Jaeseok Jang and Hyuk-Yoon Kwon. "Are multiple instance learning algorithms learnable for instances?" In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 10575–10612.

[KH15] Melih Kandemir and Fred A Hamprecht. "Computer-aided diagnosis from weak supervision: A benchmarking study". In: *Computerized medical imaging and graphics* 42 (2015), pp. 44–50.

[Lee+19] Juho Lee et al. "Set transformer: A framework for attention-based permutation-invariant neural networks". In: *International conference on machine learning*. PMLR. 2019, pp. 3744–3753.

[Li+13] Yan Li et al. "Multiple-instance learning as a classifier combining problem". In: *Pattern recognition* 46.3 (2013), pp. 865–874.

[Mai21] Paulo Maia. *An Introduction to Multiple Instance Learning*. https://nilg.ai/202105/an-introduction-to-multiple-instance-learning. Accessed: 2026-03-16. 2021.

[Mel+14] Jaime Melendez et al. "A novel multiple-instance learning-based approach to computer-aided detection of tuberculosis on chest X-rays". In: *IEEE transactions on medical imaging* 34.1 (2014), pp. 179–192.

[ML97] Oded Maron and Tomás Lozano-Pérez. "A framework for multiple-instance learning". In: *Advances in neural information processing systems* 10 (1997).

[Mua+12] Krikamol Muandet et al. "Learning from distributions via support measure machines". In: *Advances in neural information processing systems* 25 (2012).

[Oli+14] Junier Oliva et al. "Fast distribution to real regression". In: *Artificial Intelligence and Statistics*. PMLR. 2014, pp. 706–714.

[Póc+13] Barnabás Póczos et al. "Distribution-free distribution regression". In: *artificial intelligence and statistics*. PMLR. 2013, pp. 507–515.

[Que+12] Gwénolé Quellec et al. "A multiple-instance learning framework for diabetic retinopathy screening". In: *Medical image analysis* 16.6 (2012), pp. 1228–1240.

[Que+17] Gwenolé Quellec et al. "Multiple-instance learning for medical image and video analysis". In: *IEEE reviews in biomedical engineering* 10 (2017), pp. 213–234.

[RC05] Soumya Ray and Mark Craven. "Supervised versus multiple instance learning: An empirical comparison". In: *Proceedings of the 22nd international conference on Machine learning*. 2005, pp. 697–704.

[RDK00] Jan Ramon, Luc De Raedt, and Stefan Kramer. "Multi instance neural networks". In: *Proceedings of the ICML-2000 workshop on attribute-value and relational learning*. 2000, pp. 53–60.

[ST12] Sivan Sabato and Naftali Tishby. "Multi-instance learning with any hypothesis class". In: *The Journal of Machine Learning Research* 13.1 (2012), pp. 2999–3039.

[Sza+15] Zoltán Szabó et al. "Two-stage sampled learning theory on distributions". In: *Artificial Intelligence and Statistics*. PMLR. 2015, pp. 948–957.

[Sza+16] Zoltán Szabó et al. "Learning theory for distribution regression". In: *Journal of Machine Learning Research* 17.152 (2016), pp. 1–40.

[Vur+06] Volkan Vural et al. "Batch classification with applications in computer aided diagnosis". In: *European Conference on Machine Learning*. Springer. 2006, pp. 449–460.

[Wan+18] Xinggang Wang et al. "Revisiting multiple instance neural networks". In: *Pattern recognition* 74 (2018), pp. 15–24.

[WYZ08] Hua-Yan Wang, Qiang Yang, and Hongbin Zha. "Adaptive p-posterior mixture-model kernels for multiple instance learning". In: *Proceedings of the 25th international conference on Machine learning*. 2008, pp. 1136–1143.

[WZ00] Jun Wang and Jean-Daniel Zucker. "Solving multiple-instance problem: A lazy learning approach". In: (2000).

[XF04] Xin Xu and Eibe Frank. "Logistic regression and boosting for labeled bags of instances". In: *Pacific-Asia conference on knowledge discovery and data mining*. Springer. 2004, pp. 272–281.

[Xu+14] Yan Xu et al. "Weakly supervised histopathology cancer image segmentation and classification". In: *Medical image analysis* 18.3 (2014), pp. 591–604.

[Xu03] Xin Xu. "Statistical learning in multiple instance problems". PhD thesis. The University of Waikato, 2003.

[Zah+17] Manzil Zaheer et al. "Deep sets". In: *Advances in neural information processing systems* 30 (2017).

[ZG01] Qi Zhang and Sally Goldman. "EM-DD: An improved multiple-instance learning technique". In: *Advances in neural information processing systems* 14 (2001).

Autoencoders References

[Ber+24] Kamal Berahmand et al. "Autoencoders and their applications in machine learning: a survey". In: *Artificial Intelligence Review* 57.2 (2024), p. 28.

[BKG23] Dor Bank, Noam Koenigstein, and Raja Giryes. "Autoencoders". In: *Machine learning for data science handbook: data mining and knowledge discovery handbook* (2023), pp. 353–374.

[GF19] Isobel Claire Gormley and Sylvia Frühwirth-Schnatter. "Mixture of experts models". In: *Handbook of mixture analysis*. Chapman and Hall/CRC, 2019, pp. 271–307.

[KW+19] Diederik P Kingma, Max Welling, et al. "An introduction to variational autoencoders". In: *Foundations and Trends® in Machine Learning* 12.4 (2019), pp. 307–392.

[LLY18] Jianran Liu, Chan Li, and Wenyuan Yang. "Supervised learning via unsupervised sparse autoencoder". In: *IEEE Access* 6 (2018), pp. 73802–73814.

[MSC18] Meng Ma, Chuang Sun, and Xuefeng Chen. "Deep coupling autoencoder for fault diagnosis with multimodal sensory data". In: *IEEE Transactions on Industrial Informatics* 14.3 (2018), pp. 1137–1145.

[Pra+21] David Pratella et al. "A survey of autoencoder algorithms to pave the diagnosis of rare diseases". In: *International journal of molecular sciences* 22.19 (2021), p. 10891.

[Wan+19] Wei Wang et al. "Clustering with orthogonal autoencoder". In: *IEEE Access* 7 (2019), pp. 62421–62432.

[Yan+17] Bo Yang et al. "Towards k-means-friendly spaces: Simultaneous deep learning and clustering". In: *international conference on machine learning*. PMLR. 2017, pp. 3861–3870.

Wasserstein Barycenters References

[AC11] Martial Agueh and Guillaume Carlier. "Barycenters in the Wasserstein space". In: *SIAM Journal on Mathematical Analysis* 43.2 (2011), pp. 904–924.

[AXK17] Brandon Amos, Lei Xu, and J Zico Kolter. "Input convex neural networks". In: *International conference on machine learning*. PMLR. 2017, pp. 146–155.

[Big+17] Jérémie Bigot et al. "Geodesic PCA in the Wasserstein space by convex PCA". In: (2017).

[BKC22] Charlotte Bunne, Andreas Krause, and Marco Cuturi. "Supervised training of conditional monge maps". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 6859–6872.

[BLL15] Emmanuel Boissard, Thibaut Le Gouic, and Jean-Michel Loubes. "Distribution's template estimate with Wasserstein metrics". In: *Bernoulli* (2015), pp. 740–759.

[Bun+23] Charlotte Bunne et al. "Learning single-cell perturbation responses using neural optimal transport". In: *Nature methods* 20.11 (2023), pp. 1759–1768.

[CAD20] Samuel Cohen, Michael Arbel, and Marc Peter Deisenroth. "Estimating barycenters of measures in high dimensions". In: *arXiv preprint arXiv:2007.07105* (2020).

[CD14] Marco Cuturi and Arnaud Doucet. "Fast computation of Wasserstein barycenters". In: *International conference on machine learning*. PMLR. 2014, pp. 685–693.

[Cha+02] Nitesh V Chawla et al. "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* 16 (2002), pp. 321–357.

[Chi+18] Lenaic Chizat et al. "Unbalanced optimal transport: Dynamic and Kantorovich formulations". In: *Journal of Functional Analysis* 274.11 (2018), pp. 3090–3123.

[Cut+22] Marco Cuturi et al. "Optimal transport tools (ott): A jax toolbox for all things wasserstein". In: *arXiv preprint arXiv:2201.12324* (2022).

[Cut13] Marco Cuturi. "Sinkhorn distances: Lightspeed computation of optimal transport". In: *Advances in neural information processing systems* 26 (2013).

[Fla+21] Rémi Flamary et al. "Pot: Python optimal transport". In: *Journal of Machine Learning Research* 22.78 (2021), pp. 1–8.

[FTC20] Jiaojiao Fan, Amirhossein Taghvaei, and Yongxin Chen. "Scalable computations of wasserstein barycenter via input convex neural networks". In: *arXiv preprint arXiv:2007.04462* (2020).

[Gen+16] Aude Genevay et al. "Stochastic optimization for large-scale optimal transport". In: *Advances in neural information processing systems* 29 (2016).

[GMP23] Nicolas Guigui, Nina Miolane, and Xavier Pennec. "Introduction to riemannian geometry and geometric statistics: from basic theory to implementation with geomstats". In: *Foundations and Trends in Machine Learning* 16.3 (2023), pp. 329–493.

[Gul+17] Ishaan Gulrajani et al. "Improved training of wasserstein gans". In: *Advances in neural information processing systems* 30 (2017).

[Kor+19] Alexander Korotin et al. “Wasserstein-2 generative networks”. In: *arXiv preprint arXiv:1909.13082* (2019).

[Kor+21] Alexander Korotin et al. “Do neural optimal transport solvers work? a continuous wasserstein-2 benchmark”. In: *Advances in neural information processing systems* 34 (2021), pp. 14593–14605.

[Li+20] Lingxiao Li et al. “Continuous regularized wasserstein barycenters”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 17755–17765.

[Mak+20] Ashok Makkuva et al. “Optimal transport mapping via input convex neural networks”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 6672–6681.

[McC97] Robert J McCann. “A convexity principle for interacting gases”. In: *Advances in mathematics* 128.1 (1997), pp. 153–179.

[NN07] Frank Nielsen and Richard Nock. “On the centroids of symmetrized bregman divergences”. In: *arXiv preprint arXiv:0711.3242* (2007).

[PC19] Gabriel Peyré and Marco Cuturi. *Computational optimal transport: With applications to data science*. Now Foundations and Trends, 2019.

[PZ19] Victor M Panaretos and Yoav Zemel. “Statistical aspects of Wasserstein distances”. In: *Annual review of statistics and its application* 6.1 (2019), pp. 405–431.

[Rab+11] Julien Rabin et al. “Wasserstein barycenter and its application to texture mixing”. In: *International conference on scale space and variational methods in computer vision*. Springer. 2011, pp. 435–446.

[SC15] Vivien Seguy and Marco Cuturi. “Principal geodesic analysis for probability measures under the optimal transport metric”. In: *Advances in Neural Information Processing Systems* 28 (2015).

[Seg+17] Vivien Seguy et al. “Large-scale optimal transport and mapping estimation”. In: *arXiv preprint arXiv:1711.02283* (2017).

[Vil+09] Cédric Villani et al. *Optimal transport: old and new*. Vol. 338. Springer, 2009.

Other References

[AB23] Anastasios N Angelopoulos and Stephen Bates. “Conformal prediction: A gentle introduction”. In: *Foundations and Trends in Machine Learning* 16.4 (2023), pp. 494–591.

[Aug24] Midhun T Augustine. "A survey on universal approximation theorems". In: *arXiv preprint arXiv:2407.12895* (2024).

[Bat+15] Douglas Bates et al. "Fitting linear mixed-effects models using lme4". In: *Journal of statistical software* 67 (2015), pp. 1–48.

[BC93] Norman E Breslow and David G Clayton. "Approximate inference in generalized linear mixed models". In: *Journal of the American statistical Association* 88.421 (1993), pp. 9–25.

[Cha+13] Frédéric Chazal et al. "Persistence-based clustering in Riemannian manifolds". In: *Journal of the ACM (JACM)* 60.6 (2013), pp. 1–38.

[DLR77] Arthur P Dempster, Nan M Laird, and Donald B Rubin. "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the royal statistical society: series B (methodological)* 39.1 (1977), pp. 1–22.

[GB10] Xavier Glorot and Yoshua Bengio. "Understanding the difficulty of training deep feedforward neural networks". In: *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings. 2010, pp. 249–256.

[Gir21] Christophe Giraud. *Introduction to high-dimensional statistics*. Chapman and Hall/CRC, 2021.

[Goo+14] Ian J Goodfellow et al. "Generative adversarial nets". In: *Advances in neural information processing systems* 27 (2014).

[Har54] Zellig S Harris. "Distributional structure". In: *Word* 10.2-3 (1954), pp. 146–162.

[HBL23] Olympio Hacquard, Gilles Blanchard, and Clément Levrard. "Statistical learning on measures: an application to persistence diagrams". In: *arXiv preprint arXiv:2303.08456* (2023).

[HJ20] Wenchong He and Zhe Jiang. "Semi-supervised learning with the em algorithm: A comparative study between unstructured and structured prediction". In: *IEEE Transactions on Knowledge and Data Engineering* 34.6 (2020), pp. 2912–2920.

[Kit+02] Josef Kittler et al. "On combining classifiers". In: *IEEE transactions on pattern analysis and machine intelligence* 20.3 (2002), pp. 226–239.

[KL20] Patrick Kidger and Terry Lyons. "Universal approximation with deep narrow networks". In: *Conference on learning theory*. PMLR. 2020, pp. 2306–2327.

[Lau07] Steffen Lauritzen. "Exchangeability and de Finetti's Theorem". In: *Lecture Notes, University of Oxford, http://www. stats. ox. ac. uk/ steffen/teaching/grad/definetti. pdf* (2007).

[Laz+23] Walid Laziri et al. “Méthodes de réduction de dimension basées sur l’algorithme FINE pour le clustering de patients à partir de données de cytométrie en flux”. In: *54es Journées de Statistiques de la SFDS*. 2023.

[LL16] Maria V Liberti and Jason W Locasale. “The Warburg effect: how does it benefit cancer cells?” In: *Trends in biochemical sciences* 41.3 (2016), pp. 211–218.

[Mak98] Yuly Makovoz. “Uniform approximation by neural networks”. In: *Journal of Approximation Theory* 95.2 (1998), pp. 215–228.

[Mik+25] Pierre-André Mikem et al. *Multiple Instance Learning and data augmentation via Wasserstein barycenters: Application to flow cytometry data*. ML at Aussois 2025 - Autumn School on Recent Advances in Machine Learning. Poster. Nov. 2025. URL: https://inria.hal.science/hal-05499580.

[MKM25] Pierre-André Mikem, Christine Keribin, and Pascal Massart. “Méthodes d’apprentissage faiblement supervisées pour la classification de patients à partir de données de cytométrie en flux”. In: *JDS 2025 - 56ème édition des Journées de Statistiques*. SFdS. Marseille, France, June 2025. URL: https://hal.science/hal-05407929.

[Mua+17] Krikamol Muandet et al. “Kernel mean embedding of distributions: A review and beyond”. In: *Foundations and TrendsÂ® in Machine Learning* 10.1-2 (2017), pp. 1–141.

[Puj22] Louis Pujol. “Independence Structure Density Estimation”. In: *arXiv* (2022).

[Sch25] Carsten Schelp. *An alternative way to plot the Covariance ellipse*. https://carstenschelp.github.io/2018/09/14/Plot_Confidence_Ellipse_001.html. Accessed: 2025-12-30. 2025.

[Vas+17] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).

[VGS05] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

[Wan+21] Yingfan Wang et al. “Understanding how dimension reduction tools work: an empirical approach to deciphering t-SNE, UMAP, TriMAP, and PaCMAP for data visualization”. In: *Journal of Machine Learning Research* 22.201 (2021), pp. 1–73.

[ZJZ10] Yin Zhang, Rong Jin, and Zhi-Hua Zhou. “Understanding bag-of-words model: a statistical framework”. In: *International journal of machine learning and cybernetics* 1.1 (2010), pp. 43–52.

[ZMP16] Kuang Zhou, Arnaud Martin, and Quan Pan. "The belief noisy-or model applied to network reliability analysis". In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 24.06 (2016), pp. 937–960.