TU WIEN Informatics

Meta-Learning for Blast Cell Classification in Flow Cytometry Data Using MAML and Transformer Models

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Visual Computing

eingereicht von

Christian Pratellesi, BSc.
Matrikelnummer 12118532

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Dipl.-Ing. Dr.techn. Michael Reiter

Wien, 27. November 2024

Christian Pratellesi	Michael Reiter

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

TU WIEN Informatics

Meta-Learning for Blast Cell Classification in Flow Cytometry Data Using MAML and Transformer Models

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Visual Computing

by

Christian Pratellesi, BSc.
Registration Number 12118532

to the Faculty of Informatics

at the TU Wien

Advisor: Dipl.-Ing. Dr.techn. Michael Reiter

Vienna, November 27, 2024

Christian Pratellesi	Michael Reiter

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Erklärung zur Verfassung der Arbeit

Christian Pratellesi, BSc.

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.
Ich erkläre weiters, dass ich mich generativer KI-Tools lediglich als Hilfsmittel bedient habe und in der vorliegenden Arbeit mein gestalterischer Einfluss überwiegt. Im Anhang „Übersicht verwendeter Hilfsmittel“ habe ich alle generativen KI-Tools gelistet, die verwendet wurden, und angegeben, wo und wie sie verwendet wurden. Für Textpassagen, die ohne substantielle Änderungen übernommen wurden, haben ich jeweils die von mir formulierten Eingaben (Prompts) und die verwendete IT- Anwendung mit ihrem Produktnamen und Versionsnummer/Datum angegeben.

Wien, 27. November 2024

Christian Pratellesi

Acknowledgements

As I sit down to write this, I find myself pausing to reflect on the journey that brought me to this exact moment—sitting in front of a screen, surrounded by old scribbled notes, with Kygo's "Intro" melody playing softly in the background. Like the opening notes of a song, it started quietly, hesitantly, before building into something more—something shaped by the people, places, and moments that carried me along the way.

It wasn't just me walking this path. There were challenges that felt insurmountable, moments of loss when I questioned if the sacrifices were worth it, and days when the weight of uncertainty was almost too much to bear. But through every doubt, every setback, and every fleeting victory, I leaned on the people who stood by me—even when I couldn't see the end of the road myself.

I've never been one to linger on my own accomplishments; this moment isn't just mine. It's a tribute to those who walked beside me, to their quiet strength that—whether they knew it or not—helped me keep going. They shaped the person I've become and reminded me, even in the hardest times, that there's always a rhythm worth following.

First and foremost, to my family—Mom, Dad, Alex—thank you for believing in me even when I didn't believe in myself. Your unwavering support and endless reassurance gave me the courage to take this leap, to keep moving forward, and to push through countless moments of self-doubt. You've shown me what true love and sacrifice mean, and I'll forever be grateful for everything you've done to help me get here. No matter where life takes me, you will always be with me.

To my supervisor, Michael Reiter—thank you for being both a guide and a mentor. Your expertise, patience, and thoughtful feedback shaped not only this thesis but also my growth as a researcher and as a person. You pushed me to do my best, to keep going, and reminded me that perseverance, especially in the face of challenges, is where true growth begins.

To Christian Stippel, a friend who became an essential part of this journey: I still remember those late nights at your place, debugging our real-time rendering project. Frustration mounted as we tried to bring our ideas to life, and there were moments when I was ready to quit and give it all up. But your determination and teamwork helped me push through. Without you, I wouldn't have made it as far as I did, and I'm deeply grateful to have had you by my side during those challenging times.

To my friends in Italy—despite the hundreds of kilometers between us, your unconditional support has carried me through, especially during those early days when I was alone, adjusting to a new life in Vienna. You reminded me that I was never truly alone, and your belief in me prevented me from giving up on this journey. Thank you for always being there, no matter the distance.

To dancing—thank you for being my refuge. You provided an outlet for me to express myself beyond the walls of academia and a way to stay grounded when everything felt overwhelming. You introduced me to amazing people who became wonderful friends, making Vienna feel like home and filling me with a sense of belonging I hadn't even realized I was missing. Thank you for reminding me that there's more to life than work and for being my family away from home.

And finally, to Kygo—thank you for inspiring me in ways you'll never know. You introduced me to the world of music production, a hobby that has become one of my greatest joys. Making and sharing my own songs not only kept me sane but also provided a space to channel my energy when university felt overwhelming. Through music, I found a way to express my true self, explore my imagination, and push the boundaries of my creativity, always striving to exceed what I thought was possible. Attending your concert just a week before my thesis deadline felt like the perfect end to a cycle: I arrived here with your music as my companion, began creating music because of you, and now, as I close this chapter, I finally got to experience your music live. It was a moment of peace amidst the pre-deadline chaos, reminding me of the beauty of finding rhythm and balance in all things. And as you say in "Stargazing", 'I will still be here, stargazing - I'll still look up, look up - Look up for love'. This journey may be nearing its end, but I'll keep looking up, striving for more, and remaining open to the beauty and opportunities that life has yet to offer. Your music and message have been a guiding force throughout this journey, and for that, I will always be grateful.

This thesis is the product of many hands, many hearts, and many moments of grace. To everyone who has walked this path with me, in ways big or small—you are as much a part of this as I am. While I might not always recognize my triumphs, this is one I'll carry with me forever. Thank you for everything.

Kurzfassung

Diese Arbeit untersucht die Anwendung von Model-Agnostic Meta-Learning (MAML) zur Klassifizierung von Blutzellen bei akuter lymphoblastischer Leukämie (ALL) unter Verwendung von Durchflusszytometrie-Daten. Dies erfolgt durch den Vergleich der Leistung von MAML mit einem Basismodell unter verschiedenen experimentellen Bedingungen, mit Schwerpunkt auf FewShot-Learning-Szenarien. Die Ergebnisse zeigen, dass obwohl MAML während des Trainings eine höhere Instabilität und längere Konvergenzzeit aufweist als das Basismodell, es bei größeren und vielfältigen Datensätzen sowie bei unbekannten Aufgaben eine bessere Generalisierungsfähigkeit zeigt. Dies deutet darauf hin, dass MAML geeignet ist komplexe, aufgabenspezifische Merkmale insbesondere in heterogenen medizinischen Datensätzen zu erfassen. Das Basismodell hingegen übertrifft MAML bei kleineren Datensätzen aufgrund seiner schnelleren Konvergenz und geringeren Empfindlichkeit gegenüber Hyperparametern. Die Studie trägt zur Weiterentwicklung von MetaLearning in der medizinischen Diagnostik bei, identifiziert wesentliche Schwächen von MAML und schlägt Verbesserungen wie optimiertes Hyperparameter-Tuning, Datenbalancierung und krankheitsübergreifende Generalisierung vor. Die Erkenntnisse dieser Arbeit liefern wertvolle Impulse zur Entwicklung anpassungsfähiger und effizienter Diagnoseinstrumente für die Minimal Residual Disease (MRD)-Bewertung.

Abstract

This thesis explores the application of Model-Agnostic Meta-Learning (MAML) for classifying blast cells in Acute Lymphoblastic Leukemia (ALL) and Acute Myeloid Leukemia (AML) using flow cytometry data. It evaluates MAML's performance against a baseline model across various experimental setups, focusing on few-shot learning scenarios. Results indicate that while MAML exhibits higher training instability and longer convergence times, it demonstrates superior generalization when trained on larger, diverse datasets and tested on unseen tasks. This suggests MAML's potential to capture complex, task-specific features, particularly in heterogeneous medical datasets. However, the baseline model outperforms MAML on smaller datasets due to its faster convergence and lower sensitivity to hyperparameter tuning. This study contributes to advancing meta-learning applications in medical diagnostics, identifies key limitations of MAML, and proposes improvements, including optimized hyperparameter tuning, dataset balancing, and cross-disease generalization. These findings offer insights for developing adaptable and efficient diagnostic tools for Minimal Residual Disease (MRD) assessment.

Contents

CHAPTER 1

Introduction

1.1 Motivation and Problem Definition

Over the past decade, neural networks have driven significant advancements across various fields, revolutionizing approaches in areas such as medicine, finance, and Natural Language Processing (NLP). One domain where deep learning has had a particularly profound impact is in the diagnosis and treatment of diseases, like Acute Lymphoblastic Leukemia (ALL) in children [PRL08] or Acute Myeloid Leukemia (AML) [LDB99, DWB15]. A critical aspect of ALL and AML treatment involves monitoring the number of residual blast cells in the bone marrow after therapy to evaluate treatment efficacy. This process, known as Minimal Residual Disease (MRD) assessment [Cam10], is one of the most important predictors of patient prognosis and long-term survival [CSSH$^+$00]. However, traditional MRD assessment methods rely heavily on manual interpretation by medical experts, which can be time-consuming, resource-intensive, and prone to inter-operator variability.

Advances in Machine Learning (ML) have shown promise in automating MRD assessment, potentially enhancing both speed and consistency in clinical decision-making. For example, Reiter et al. [RDS$^+$19] proposed ML-based models that utilize Gaussian Mixture Models (GMMs) to classify bone marrow cells. In contrast, Wodlinger et al. [WRW$^+$22] introduced Transformer architectures to improve the accuracy of cell classification. Figure 1.1 shows the architecture proposed by Wodlinger et al. [WRW$^+$22].

Despite these advancements, the challenge of limited labeled data in medical contexts, particularly concerning diseases like ALL and AML, remains a significant barrier to the widespread adoption of ML systems. The scarcity of labeled data arises from the diseases's rarity and the complexity of obtaining high-quality annotated samples, as manual labeling requires domain expertise and significant time investment. Traditional neural networks, which perform well in data-rich environments, often struggle in situations

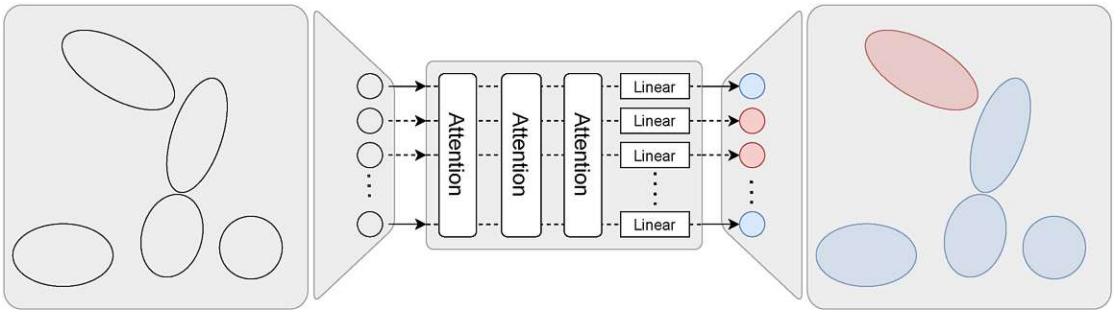

Figure 1.1: Transformer architecture proposed by Wodlinger et al. [WRW+22] to classify blast cells. The architecture consists of several attention layers followed by linear layers to output the final predictions per cell. The figure is adapted from the original paper by Wodlinger et al. [WRW+22].

with only limited training examples. This issue is particularly critical in MRD assessment, where each patient's bone marrow sample may display unique characteristics influenced by disease progression, genetic makeup, and prior treatments [BF22, RNO+21]. These variations make it difficult for conventional ML models to generalize effectively to new, unseen samples.

In response to these challenges, this thesis explores meta-learning, a machine learning paradigm that enables models to learn how to learn from limited data. Specifically, the focus will be on the application of Model-Agnostic Meta-Learning (MAML) [FAL17] to Flow Cytometry (FCM) data. MAML has emerged as a powerful technique for few-shot learning tasks, where the goal is to learn from a small number of examples and adapt quickly to new tasks. Unlike traditional training methods that optimize a model for a single task, MAML optimizes the model's initialization parameters to allow for rapid adaptation to new tasks with minimal fine-tuning.

Using MAML offers several advantages. First, unlike conventional approaches that train on combined datasets to implicitly learn generalizable features, MAML explicitly learns to adapt across tasks by focusing on meta-learning. This ensures the development of features optimized for adaptability to new, unseen cases rather than features tailored to the specific distribution of a pooled dataset. Second, this meta-learning approach can address the issue of data scarcity by enabling the model to utilize the limited labeled examples available better. Recent studies have demonstrated MAML's potential to achieve competitive results in various few-shot learning scenarios, such as on the Omniglot [LSGT11] and MiniImagenet [RL16] datasets, indicating its applicability to medical domains like MRD assessment.

This work aims to build on the existing literature by applying meta-learning to FCM data collected from pediatric ALL and AML patients. By training the model to generalize across different patient datasets, we seek to improve the accuracy and reliability of MRD assessments, even in cases where labeled data is limited. The broader goal is to contribute to developing robust and scalable machine-learning methods that can enhance clinical

decision-making in the fight against childhood leukemia.

1.2 Aim of the Thesis

The primary objective of this thesis is to apply meta-learning techniques to enhance the classification of ALL and AML blast cells. Specifically, it explores the use of MAML [FAL17], a meta-learning framework that offers flexibility in the underlying neural network architecture, allowing for adaptation to diverse contexts.

In this work, MAML is trained on datasets from Vienna (ALL and AML), Berlin (ALL), and Buenos Aires (ALL), each representing different clinical settings and patient demographics. A leave-one-out approach is employed, where the model is trained on all but one dataset for the respective disease, which is then used for evaluation. This strategy assesses the model's ability to generalize to unseen data, a critical factor for clinical applications.

This study evaluates MAML against state-of-the-art methods for MRD assessment in ALL and AML, highlighting its strengths and limitations. Key research questions include:

- What challenges arise when applying meta-learning techniques to blast cell classification and MRD assessment?
- How does MAML perform compared to existing state-of-the-art approaches for ALL and AML MRD assessment?
- How effectively can MAML adapt to new, unseen MRD assessment tasks?

This thesis contributes to research on meta-learning in medical diagnostics, analysing its potential for improving the accuracy and adaptability of MRD assessments. It also provides a foundation for future work to extend meta-learning to other medical challenges.

1.3 Contributions

This thesis contributes to the fields of meta-learning and medical data classification, focusing specifically on the automated classification of ALL and AML blast cells using FCM data and MAML. The key contributions of this work are as follows:

- **Application to multiple ALL datasets**: Three distinct datasets from Vienna, Berlin, and Buenos Aires have been utilized, each representing unique patient samples. The proposed method employs a leave-one-out strategy to evaluate MAML's generalization capacity on unseen data.

- **Application to multiple AML datasets**: A separate set of experiments was conducted on AML using data exclusively from Vienna. Tasks were defined based on distinct phenotypes (M2, M4, M5, and M7), each representing a specific subgroup of AML samples. This approach evaluates MAML's capacity to generalize across phenotypic variations within a single disease, emphasizing its adaptability to diverse cellular profiles within the same dataset.
- **Comparison with State-of-the-Art**: The performance of MAML in the context of ALL and AML classification is compared to current state-of-the-art approaches, providing insights into the strengths and weaknesses of using meta-learning for medical data analysis.
- **Evaluation of Meta-Learning in Medical Settings**: The work examines the potential and challenges of applying meta-learning techniques to medical tasks, specifically FCM data analysis, where data scarcity is a common issue.

1.4 Outline of the Thesis

This thesis is organized as follows:

- **Chapter 2: Background**
  This chapter introduces the foundational concepts necessary for understanding the thesis, including meta-learning, Transformer architectures, principles of FCM, and an overview of ALL, AML, and MRD.
- **Chapter 3: Related Work**
  This chapter reviews the relevant literature in four key areas: automated FCM data analysis, meta-learning frameworks, Transformers for classification tasks, and MRD assessment methods in the context of ALL and AML. It offers an overview of current techniques and establishes a foundation for understanding the methods employed in this thesis.
- **Chapter 4: Dataset**
  This chapter details the datasets used in the experiments. It describes the process of collecting bone marrow samples from ALL and AML patients, the structure of the datasets, and how the data is preprocessed for training.
- **Chapter 5: Methodology**
  This chapter outlines the methods employed in this thesis, with a particular emphasis on the MAML algorithm. It also details the training of Transformer-based models and describes the specific modifications made to the original MAML algorithm to adapt it for the classification task.
- **Chapter 6: Experiments**
  This chapter outlines the experimental setup, including the training and evaluation

protocols. It describes the specific configurations of the models, and the metrics employed to assess performance.

- **Chapter 7: Results**
  This chapter presents the results of the experiments and addresses all the research questions. It details the model's performance on the leave-one-out evaluation. It compares it to existing state-of-the-art approaches, discusses MAML's adaptation capabilities to new unseen tasks, and the associated challenges.

- **Chapter 8: Discussion**
  This chapter discusses the results obtained and interprets the findings in light of the challenges and potential of applying meta-learning to blast cell classification and MRD assessment tasks. It reflects on the strengths and limitations of this approach and offers insights for future work.

- **Chapter 9: Conclusion**
  This final chapter summarizes the thesis findings and highlights its contributions. It also offers suggestions for future research and potential improvements in automated medical data analysis.

CHAPTER 2

Background

The rapid advancements in ML have led to innovative approaches in medical diagnostics, particularly in automating complex tasks such as the classification of ALL or AML blast cells and the assessment of MRD. This section provides the foundational knowledge necessary to understand the methodologies and challenges addressed in this thesis. We begin by exploring the clinical significance of ALL, AML, and the current practices in MRD assessment, followed by an overview of FCM, a key technology for gathering diagnostic data. We also discuss the evolving role of deep learning in medical diagnostics and introduce meta-learning, specifically MAML, as a solution for learning from limited data. Finally, we present Transformer models, which have revolutionized data processing in various fields, and their applicability to the high-dimensional, unordered data characteristic of FCM.

2.1 ALL

ALL is an aggressive form of cancer originating in the bone marrow, where blood cells are produced. In ALL, immature lymphoblasts, precursors to lymphocytes, a type of white blood cell, begin to grow uncontrollably. This unchecked growth of abnormal cells disrupts the production of normal blood cells, resulting in impaired immune function, anemia, bleeding disorders, and an increased risk of infections. ALL progresses rapidly and can become life-threatening if left untreated. ALL predominantly affects children. In adults, it represents a smaller portion of leukemia cases but is typically more aggressive and harder to treat.

There are several subtypes of ALL, classified based on the type of lymphocyte involved, either B-cell ALL or T-cell ALL, as well as various genetic and molecular factors. These subtypes can impact both the treatment approach and the prognosis:

- **B-cell ALL** - In B-cell cancer, the cancerous cells originate from B-cell precursors, which are responsible for producing antibodies that help the immune system combat infections [Onc09, DRW+14].
- **T-cell ALL** - T-cell ALL is characterized by the proliferation of immature T-cells, which play a key role in immune responses, particularly in identifying and destroying infected or cancerous cells. This subtype is often associated with a higher white blood cell count at diagnosis and may require more intensive therapy [CSMO+09].

Despite advancements in treatment, such as chemotherapy and targeted therapies, a key factor in achieving long-term remission is the accurate monitoring of residual disease during and after treatment. Detecting even a few remaining cancer cells can provide early warning of relapse and guide subsequent therapy.

The standard treatment for ALL involves multiple stages:

- **Induction Therapy** - This initial phase aims to bring the patient into remission, meaning leukemia cells are no longer detectable in the blood or bone marrow. While most patients achieve remission after this stage, undetectable cancer cells may remain.
- **Consolidation Therapy** - This phase targets any residual leukemia cells that may have survived induction therapy. The goal is to eliminate these cells to prevent relapse.
- **Maintenance Therapy** - Over an extended period, lower doses of chemotherapy are given to prevent the recurrence of leukemia.

Each stage requires close monitoring to assess the patient's response to treatment and detect residual disease. The concept of MRD is critical in this context. While clinical remission indicates that leukemia cells are no longer detectable by standard methods, MRD techniques offer a more sensitive approach to identifying any remaining leukemia cells that could potentially lead to relapse.

2.2 AML

AML is a malignancy of the bone marrow and blood, characterized by the uncontrolled proliferation of abnormal myeloid precursor cells (blasts). These blasts accumulate in the bone marrow, inhibiting the production of normal blood cells and leading to symptoms such as anemia, infections, and bleeding. AML is the most common form of acute leukemia in adults, with incidence rates increasing significantly with age.

AML is highly heterogeneous, comprising multiple subtypes defined by genetic, cytogenetic, molecular, and immunophenotypic characteristics [LMM16, AKH+20].

The World Health Organization (WHO) classification includes categories such as AML with recurrent genetic abnormalities and AML with myelodysplasia-related changes. Each subtype presents unique clinical behaviors and prognoses.

The immunophenotypes of AML, determined through FCM, are equally diverse, reflecting the distinct stages of myeloid differentiation at which leukemic transformation occurs. Typical markers include CD33, CD34, CD45, CD99, and HLA-DR, with differences across phenotypes influencing diagnosis, monitoring, and therapeutic decisions. This diversity underscores the need for personalized medicine to address the specific biology of each patient's disease.

MRD assessment has emerged as a critical prognostic tool in AML, with MRD negativity following treatment strongly associated with better relapse-free survival and overall survival. This makes it a key metric for evaluating treatment response. Techniques such as FCM allow for sensitive MRD detection, guiding post-remission therapy decisions, including the need for consolidation chemotherapy or hematopoietic stem cell transplantation.

In summary, AML is a complex and heterogeneous disease with diverse subtypes and phenotypes that significantly influence patient outcomes. Accurate diagnostics and therapeutic monitoring, particularly through MRD assessment, are pivotal in optimizing treatment strategies and improving survival outcomes. The continued refinement of diagnostic and prognostic tools is essential for advancing AML care.

In summary, AML is a complex disease with diverse phenotypic manifestations, highlighting the need for precise diagnostics and therapeutic monitoring. The assessment of MRD is pivotal in optimizing treatment and improving patient outcomes.

2.3 MRD Assessment

MRD refers to the small number of cancerous cells that may survive treatment and remain in a patient's body, even after achieving clinical remission. Detecting MRD is crucial for guiding treatment decisions, as patients with detectable MRD face a significantly higher risk of relapse compared to those without MRD. In fact, MRD is regarded as one of the most important prognostic factors in the treatment of ALL [vDvdVBO15].

MRD can be measured using several techniques, including molecular assays such as Polymerase Chain Reaction (PCR) and FCM. Both methods are highly sensitive, enabling the detection of leukemia cells that may constitute less than one in 10,000 normal cells.

PCR - This technique amplifies specific genetic sequences found in leukemia cells, enabling their detection even in tiny quantities. PCR is especially useful for identifying genetic mutations or chromosomal rearrangements characteristic of certain types of ALL.

FCM - FCM has become one of the most widely used techniques for MRD detection. It labels cells with fluorescent markers that bind to specific surface proteins characteristic of leukemia cells. By analyzing the light emitted by these markers, FCM can classify

individual cells based on their surface proteins, offering a highly accurate and sensitive method for detecting residual leukemia cells.

The typical pipeline for MRD assessment using FCM works as follows:

- **Sample Preparation** - Bone marrow or blood samples are collected from patients and treated with antibodies labeled with fluorescent dyes. These antibodies bind to specific cell surface markers (antigens), allowing for the distinction between leukemia cells and normal cells.
- **Data Acquisition** - The sample is run through a flow cytometer, where cells pass through a laser one at a time. The laser excites the fluorescent markers, and detectors measure the emitted light.
- **Data Analysis** - The flow cytometer generates complex, high-dimensional data that is analyzed to identify leukemic cells based on their fluorescence profiles.

One challenge in MRD assessment through FCM is the variability introduced by factors such as sample handling, instrument calibration, and patient biological differences. This variability can lead to inconsistencies in results, necessitating expert knowledge for accurate interpretation. Furthermore, the manual analysis of the large datasets produced by FCM is time-consuming and prone to human error. Automating this analysis with ML techniques presents a promising approach to enhancing the speed and accuracy of MRD assessments, particularly in detecting rare subpopulations of leukemia cells.

The clinical importance of MRD monitoring in ALL and AML cannot be overstated. Numerous studies have demonstrated that patients who achieve MRD-negative status after treatment are less likely to relapse and have better long-term survival rates compared to those who remain MRD-positive [vDvdVBO15]. Consequently, MRD monitoring has been integrated into routine clinical practice to guide treatment decisions. For instance, patients with persistent MRD following induction therapy may be considered for more intensive treatment or stem cell transplants to decrease their risk of relapse.

The role of MRD goes beyond predicting relapse; it also serves as a biomarker for real-time therapy adjustments. If MRD is detected during or after consolidation therapy, clinicians may modify the treatment regimen to target the remaining leukemia cells more aggressively. Conversely, patients who are MRD-negative may be spared the toxic effects of unnecessary intensive treatments.

2.4 Flow Cytometry

FCM is a robust, high-throughput technique used to analyze the physical and chemical properties of cells or particles suspended in a fluid. This technology has become a cornerstone in both clinical diagnostics and research, particularly in fields such as immunology, oncology, and hematology, due to its capability to measure multiple

parameters on a cell-by-cell basis. FCM operates by passing cells through a laser beam, where the interaction of light with the cells yields information about their size, granularity, and the presence of specific molecular markers. The resulting data for a single cell (an event) is a collection of measurements of the concentrations of cell surface markers. Figure 2.1 shows an example of data extracted using FCM.

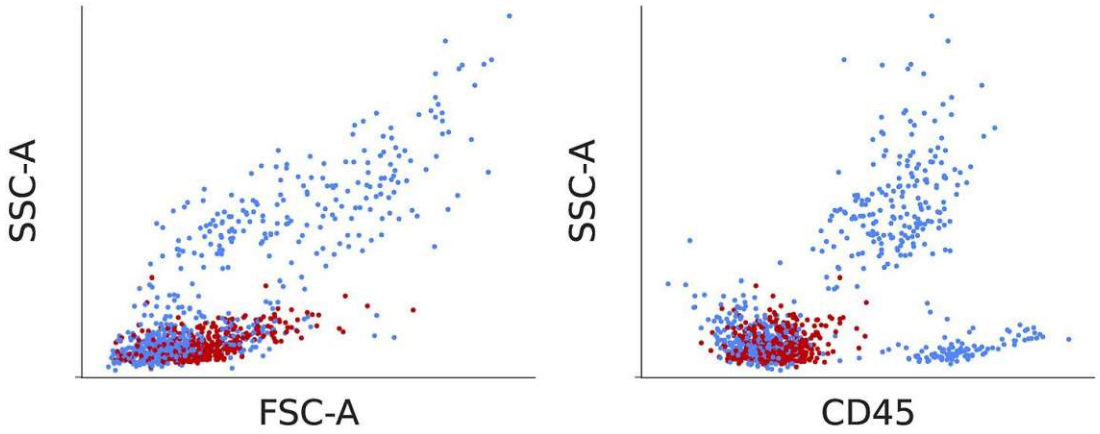

Figure 2.1: Two 2D projections of a patient's cell data extracted using FCM. Blast cells are labeled in red, while healthy cells are labeled in blue. The projections illustrate different cell populations by plotting Side Scatter (SSC) against Forward Scatter (FSC) and CD45, respectively. The figure is adapted from the original paper by Wodlinger et al. [WRW+22].

In clinical settings, such as the diagnosis and monitoring of ALL and AML, FCM is used to identify abnormal cell populations and assess MRD [DFP+02]. By analyzing surface markers on cells, FCM differentiates between healthy and malignant cells, thereby assisting in the diagnosis and prognosis of the disease.

At the core of FCM lies the principle of light scattering and fluorescence. Cells suspended in a liquid stream are directed into a narrow channel, passing through one or more laser beams. Each cell scatters the laser light, and detectors measure each cell's physical and chemical properties. Figure 2.2 illustrates the structure of a flow cytometer.

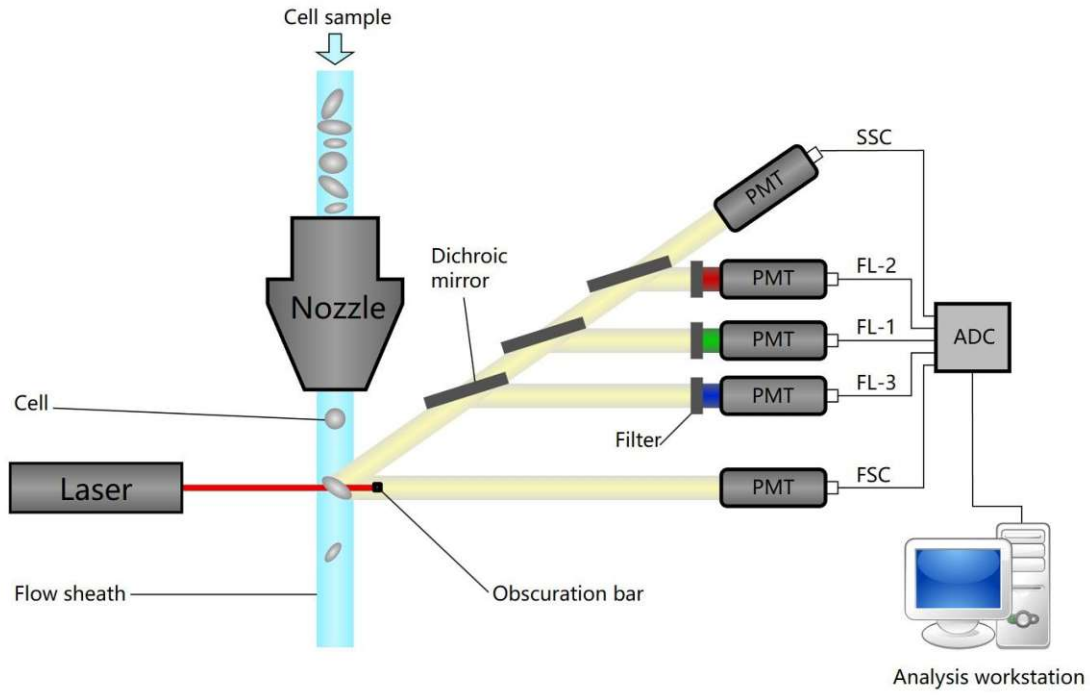

Figure 2.2: Internal structure of a flow cytometer. Cells flow individually into a laser beam, scattering forward and sideways light. The amount of scattered light is measured by sensors and converted into numerical values for each physical and chemical property of the analyzed cells. The figure is adapted from [Bio]

The physical properties measured by the flow cytometer are FSC and SSC:

- **FSC** - This measurement is proportional to the size of the cell. Larger cells scatter more light in the forward direction, making this parameter useful for evaluating cell volume.
- **SSC** - SSC measures the granularity or internal complexity of cells. Cells with more complex internal structures (e.g., granules, vesicles) scatter more light sideways.

In addition to scattering, FCM employs fluorescence to analyze specific chemical features and molecular markers on the surface or inside the cells. Cells are stained with fluorescently labeled antibodies that bind to proteins (markers) expressed on the cell surface. Detectors capture these fluorescent signals, facilitating multi-parameter analysis of each individual cell. This capability is essential for identifying distinct populations of cells, such as different immune cell subsets or leukemic blast cells.

FCM plays a critical role in diagnosing ALL and AML. The diagnosis of ALL and AML involves identifying leukemic blast cells and distinguishing them from normal blood cells.

Leukemic cells in ALL and AML often express abnormal combinations of surface markers, which can be detected by staining the cells with antibodies that specifically bind to these markers.

For example, in B-cell ALL, the leukemic cells typically express markers such as CD19, CD10, and CD34. In contrast, T-cell ALL commonly expresses markers like CD3 and CD7. FCM allows for the detection of these markers, aiding in the determination of AML or the specific subtype of ALL, which is essential for treatment planning. Table 2.1 shows the most common identifying biomarkers for ALL and AML.

Disease Type	Common Markers
B-ALL	CD10, CD19, CD20, CD22, CD 34, CD45
T-ALL	CD2, CD3, CD4, CD5, CD7, CD8
AML	CD33, CD34, CD45, CD99, CD117, HLA-DR

Table 2.1: The most common markers for identifying AML and ALL, subdividing it into type B and type T.

Several specialized techniques are used in FCM to extract the physical and chemical properties of cells. Immunophenotyping is the most common application in the diagnosis of leukemia. Cells are labeled with antibodies specific to various cell surface markers. By analyzing the combination of markers expressed by the cells, clinicians can classify them into different subsets. For example, in ALL, immunophenotyping can differentiate between B-cell and T-cell leukemias based on the expression of markers such as CD19 and CD3.

Although FCM is a robust tool, it faces challenges, particularly when applied to clinical diagnostics:

Noise and Variability - Differences in sample preparation and instrument calibration can introduce variability into the data, posing a significant challenge when comparing results across different patients or laboratories.

Rare Event Detection - In some cases, such as MRD assessment, FCM must identify extremely rare leukemic cells within a large population of normal cells. Accurately detecting these cells requires, at times, multiple rounds of staining and analysis.

Data Complexity - The multi-dimensional nature of FCM data, which involves measuring numerous parameters for each cell, can be challenging to analyze and interpret, necessitating advanced computational tools.

CHAPTER 3

Related Work

This section describes the state of the art in the main research areas related to the thesis: Automated FCM Analysis and MRD assessment, Transformers, and Meta-Learning.

Several key challenges remain despite significant advances in automated FCM analysis. One of the most critical issues is the scarcity of labeled data, especially concerning rare diseases like ALL and AML. This limitation hinders the effectiveness of traditional supervised learning methods. Additionally, many existing models are highly sensitive to noise and variability in FCM data, such as differences in sample preparation and instrument calibration, which can negatively impact model performance. The introduction of Transformer models has revolutionized various fields beyond their original application in natural language processing, extending into biomedical data analysis. Their flexibility, ability to handle high-dimensional, unordered data, and improved computational efficiency, especially with architectures like Set Transformers, make them ideal candidates for complex tasks like FCM. In the context of this thesis, Transformers offer a robust framework for classifying ALL and AML blast cells. To address the limitations of labeled data, meta-learning techniques such as MAML have shown promise. MAML enables models to adapt quickly to new tasks using only a few examples, making it particularly well-suited for clinical scenarios where data scarcity presents a significant challenge. By combining the powerful representational capabilities of Transformers with the adaptability of meta-learning frameworks, this thesis aims to advance the field of automated FCM analysis, particularly in the context of ALL and AML.

3.1 Automated FCM Analysis and MRD Assessment

FCM is a crucial tool in clinical diagnostics and biomedical research, enabling high-throughput measurement of cell properties such as size, complexity, and the expression of surface and intracellular markers. Traditionally, this technique has relied heavily on expert knowledge for manual gating and analysis, which can be time-consuming and

subjective. Automated FCM analysis provides a scalable and more objective approach, making it a vital topic in advancing modern clinical diagnostics.

The need for automation in FCM analysis has been recognized for over two decades. Early efforts primarily focused on developing rule-based systems and statistical models for identifying cell populations. For example, Pyne et al. [PHW+09] introduced an automated pipeline based on finite mixture model clustering techniques to detect cell populations based on multi-dimensional FCM data. Their approach reduced the need for manual gating but still required significant input from experts to define appropriate clustering parameters.

Following this, the work by Aghaeepour et al. [AFC+13] demonstrated that supervised ML provide a valid automated alternative to traditional manual gating. This shift towards supervised learning laid the groundwork for more data-driven and fully automated techniques to reduce human error and increase reproducibility.

With the advent of ML, more sophisticated models have been developed to automate the analysis of FCM data. One of the first breakthroughs was the application of GMMs to classify cellular subpopulations automatically. For instance, Naim et al. [NDR+14], Dundar et al. [DAYR14], and Johnsson et al. [JWF16] utilized GMMs in an unsupervised manner to identify cell populations in FCM datasets. However, GMMs can also be applied in a supervised approach, as demonstrated by Reiter et al. [RRK+16, RDS+19]. In SWIFT [NDR+14], GMMs are adapted to enhance the detection of rare cell sub-populations. BayesFlow [JWF16] utilizes a hierarchical Bayesian model that incorporates expert knowledge through informative priors. This method addresses inter-sample variation by employing a supervised approach in which GMMs are aligned with those from a labeled reference dataset [RRK+16]. Further refinement of this method is presented in [RDS+19], which introduces closed-form optimization in the fitting process. Their method learns the weights of a linear combination of multiple GMMs to predict MRD values on unseen samples. This approach demonstrated the potential of ML for MRD assessment, but there was room for improvement in predictive performance.

Recently, deep learning has been effectively utilized in the automated analysis of cellular image data [IAB+21]. However, its application to FCM data has been more limited, except for certain imaging-based FCM approaches [NDBS21, EKB+17, LCD+18]. Several studies [SLR+19, LSR+18, LSS+17] have applied neural networks with fully connected layers to individual events in FCM data. However, these methods are restricted to learning fixed decision boundaries, which limits their ability to distinguish biologically relevant sub-populations. More recently, a novel method was introduced by Zhao et al. [ZMH+20] that addresses this limitation by converting FCM data into image form and analyzing it using trained Convolutional Neural Networks (CNNs).

With the advent of Transformer models and the attention mechanism introduced by Vaswani et al. [VSP+17], significant progress has been made in automated FCM analysis and MRD assessment. Wodlinger et al. [WRW+22] further advanced the field by proposing a methodology based on Transformers that utilizes the attention mechanism to

focus on specific cell subpopulations, significantly improving upon the GMMs approach. They employed Set Transformers, first introduced by Lee et al. [LLK+19], a variation of the Transformer model designed to be invariant to input order, which is crucial when working with unordered sets of data points, such as FCM cell samples. This approach is currently regarded as the state-of-the-art method for ALL and AML MRD assessment.

3.2 Transformers

Transformers were first introduced by Vaswani et al. in 2017 in the seminal paper *Attention is All You Need* [VSP+17]. This innovation revolutionized NLP by replacing recurrent layers with an attention mechanism. The core advancement of the Transformer architecture is the self-attention mechanism, which allows the model to dynamically determine dependencies between input tokens for each input individually, enabling a more context-specific understanding compared to traditional models. By employing multiple attention heads, Transformers can simultaneously focus on different parts of the input sequence, resulting in improved feature extraction and more accurate representations.

The initial architecture of Transformers showed significant improvements in tasks such as machine translation, achieving state-of-the-art results and surpassing previously dominant models like Long Short-Term Memory (LSTM) networks. However, a major challenge of the Transformer architecture is its computational complexity. The self-attention mechanism has a time complexity of $\mathcal{O}(n^2)$, where n is the length of the input sequence. This quadratic complexity complicates the scaling of Transformers to vast datasets or extremely long input sequences without substantial computational resources. Therefore, the authors restrict input sequences to 2048 tokens.

Several variations and improvements have been proposed over the years to address the inefficiencies of the standard Transformer model. One such innovation is the Reformer [KKL20], which replaces self-attention with locality-sensitive hashing, effectively reducing the time complexity to $\mathcal{O}(n \log n)$. Other approaches include the Transformer proposed by Katharopoulos et al. [KVPF20] and the Performer introduced by Choromanski et al. [CLD+20], which reduce the complexity of the self-attention mechanism by approximating attention through kernelized methods, lowering the computational cost to $\mathcal{O}(n)$. Additionally, Wang et al. [WLK+20] introduce the Linformer, which reduces the number of required key-value pairs for efficient attention computation, making Transformers more applicable to tasks with limited computational power.

Another critical development is the Set Transformer [LLK+19], specifically designed to handle unordered input data, such as data found in FCM or point cloud analysis. Unlike traditional Transformers that are sensitive to the order of input elements, Set Transformers are permutation-invariant. This means that the order of input elements does not affect the model's predictions. This property is essential in scenarios like FCM, where each cell in the dataset is treated as an independent event, making the order of processing irrelevant.

The use of Set Transformers in biomedical data analysis, particularly in FCM, has gained attention due to their ability to process unordered sets of cells while maintaining robustness and adaptability. In the work by Wodlinger et al. [WRW+22], Set Transformers were successfully applied to FCM data for MRD assessment in ALL. By employing self-attention to capture relationships between different cells within the sample, the model could classify cells with high accuracy, leveraging the global context of the entire sample rather than treating events in isolation while remaining independent of the order in which they were presented.

Set Transformers have proven to be a powerful tool for automating the classification of biological data, offering improvements over traditional methods such as GMMs. Their ability to generalize across different datasets and tasks, along with their capacity to handle complex, high-dimensional data, make them an essential component of modern computational biology pipelines.

3.3 Meta-Learning

Meta-learning, commonly referred to as "learning to learn", represents a promising direction in ML aimed at enhancing a model's ability to adapt to new tasks with limited data. The central concept of meta-learning is to optimize a model's learning process so that it can generalize across various tasks rather than excelling at just one. This typically involves optimizing hyperparameters, such as learning rates, initialization weights, or entire learning algorithms. By learning how to learn, meta-learning models can adapt more quickly and effectively to new environments, making them particularly useful for applications where data is scarce, such as medical diagnostics or few-shot classification tasks.

One of the most influential meta-learning frameworks is MAML, introduced by Finn et al. in 2017 [FAL17]. The key innovation of MAML lies in learning the initial parameters of a model so that, with just a few gradient steps, the model can quickly adapt to a new task. Unlike other approaches, MAML is model-agnostic, meaning it makes no assumptions about the architecture of the base learner and can be applied to any model that learns through gradient descent, such as neural networks.

MAML has demonstrated its efficacy in various few-shot learning scenarios, including image classification, reinforcement learning, and language modeling. However, despite its flexibility, MAML presents certain challenges. Training MAML involves two levels of optimization: the inner loop, which updates task-specific models, and the outer loop, which refines the meta-model's parameters based on the aggregated updates from the inner loop. This bi-level optimization necessitates the computation of second-order derivatives during the outer loop, leading to significantly increased computational cost and complexity. While this reliance on second-order derivatives is crucial for accurate gradient updates, it can also introduce additional instability, particularly when dealing with noisy or highly diverse tasks. Antoniou et al. [AES18] addressed some of these

challenges by introducing techniques to enhance stability and convergence, including task-specific learning rates and additional regularization.

While MAML has been a dominant force in meta-learning, several alternative approaches have been developed to address its limitations or to target different aspects of the meta-learning problem:

Reptile - Reptile [Nic18] is a first-order approximation of MAML that simplifies the optimization process. Rather than computing second-order derivatives, Reptile updates the meta-learner by averaging gradients across tasks. This approach reduces computational costs while still preserving much of the effectiveness of MAML, though it may result in slightly lower performance in some cases.

Meta-SGD - Meta-Stochastic Gradient Descent (SGD) [LZCL17] extends MAML by allowing the model to not only learn an optimal initialization but also to learn a per-parameter learning rate, which can accelerate the adaptation process. This improvement helps the model adapt quickly to new tasks, making Meta-SGD more flexible than MAML in certain settings.

Implicit MAML - The Implicit MAML approach [RFKL19] aims to enhance the efficiency of gradient computation by utilizing implicit gradients. By solving for the stationary points of the meta-learning objective, this method eliminates the requirement for second-order derivatives, thereby increasing its scalability and computational feasibility for large datasets or models. Additionally, this approach is agnostic to the choice of inner loop optimizer and can accommodate multiple gradient steps without issues related to vanishing gradients or memory limitations.

Multimodal MAML - Vuorio et al. introduced a multimodal variant of MAML called Multimodal MAML. This approach is designed to handle multimodal task distributions involving tasks from different domains. Multimodal MAML incorporates a modulation network that helps the model identify the mode of a given task and adapt accordingly, leading to more effective task-specific learning.

Meta-Curvature - Meta-curvature, introduced by Park and Olivia [PO19], is a novel approach to meta-learning that emphasizes learning not only the model's parameters but also the curvature of the parameter space. Traditional gradient-based meta-learning methods, such as MAML, rely on a fixed update rule during adaptation. In contrast, Meta-Curvature enhances this process by enabling the learning of an optimal preconditioning matrix that adjusts the curvature of the parameter space during training. This results in more flexible and efficient updates, facilitating faster adaptation to new tasks.

CHAPTER 4

Dataset

For ALL, our evaluation uses openly accessible datasets sourced from three distinct clinical institutions. These datasets consist of bone marrow samples taken from pediatric B-ALL patients on the 15th day following induction therapy. Each sample includes manually gated data as the ground truth, identifying both blast and non-blast cell populations. These datasets are summarized in Table 4.1.

Name	City	Year	Size
vie14	Vienna	2009-2014	200
vie20	Vienna	2015-2020	319
vie	Vienna	2009-2020	519
bln	Berlin	2015	72
bue	Buenos Aires	2016-2017	65

Table 4.1: Structure of the ALL dataset used for our experiments.

For AML, we use a dataset consisting of bone marrow samples collected on the day of the diagnosis, also referred to as day zero. These samples were collected in Vienna over the span of seven years (2017-2024) and the total dataset consists of a total of 105 samples divided by phenotype. Table 4.2 summarizes the datasets.

4.1 Data Collection - ALL

Vienna - The Vienna dataset was collected at the St. Anna Children's Cancer Research Institute (CCRI) over 11 years, from 2009 to 2020, using a Becton Dickinson LSR II flow cytometer and FACSDiva software v6.2. Referred to as "vie", this dataset includes 519 samples, which we further divide into two subsets for analysis:

Phenotype	City	Year	Size
M2	Vienna	2017-2024	24
M4	Vienna	2017-2024	22
M5	Vienna	2017-2024	33
M7	Berlin	2017-2024	26

Table 4.2: Structure of the AML dataset used for our experiments.

- vie14: This subset includes 200 samples collected between 2009 and 2014, consistent with the dataset described in [RDS+19, WRW+22]. The samples were stained using a conventional seven-color drop-in panel ("B7") with the following fluorescent markers: CD20-FITC, CD10-PE, CD45-PerCP, CD34-PE-Cy7, CD19-APC, CD38-Alexa-Fluor700, and SYTO 41.

- vie20: This subset includes 319 samples collected between 2016 and 2020. The staining process utilized dried format tubes (DuraClone™ ReALB, Beckman Coulter) with the following fluorochrome-conjugated antibodies: CD58-FITC, CD34-ECD, CD10-PC5.5, CD19-PC7, CD38-APC-Alexa700, CD20-APC-Alexa750, and CD45-Krome Orange, along with the SYTO 41 dye.

Berlin - The bln Dura dataset, also referred to as "bln", consists of 72 samples collected in 2016 at Charité Berlin, as documented in [RDS+19, WRW+22]. These samples were analyzed using a Navios flow cytometer (Beckman Coulter, Brea, CA) with an 8-color multiparameter FCM panel ("B8") and processed using a specialized dried format tube (DuraClone™, Beckman Coulter). The panel comprises seven fluorochrome-conjugated antibodies: CD58-FITC, CD10-PE, CD34-PerCPCy5.5, CD19-PC7, CD38-APC, CD20-APC-Alexa750, and CD45-Krome Orange, along with SYTO 41.

Buenos Aires - The bue Dura dataset (referred to as "bue") consists of 65 samples collected from 2016 to 2017 at Garrahan Hospital in Buenos Aires, as reported in [RDS+19, WRW+22]. This dataset was processed using a FACSCanto II flow cytometer (Becton Dickinson, San Jose, CA) with FACSDiva software v8.0.1. The staining panel is identical to the one of the bln Dura dataset, utilizing the DuraClone™ cocktail tube ("B8," Beckman Coulter, Brea, CA).

The gating hierarchy to track the blast events differs from clinic to clinic, and different instruments were used to acquire data for each dataset. Therefore, there are differences in dynamics range and settings reflected in the data, justifying the use of each dataset as a separate task for ALL. Figure 4.1 summarizes the data acquisition procedure for each clinic.

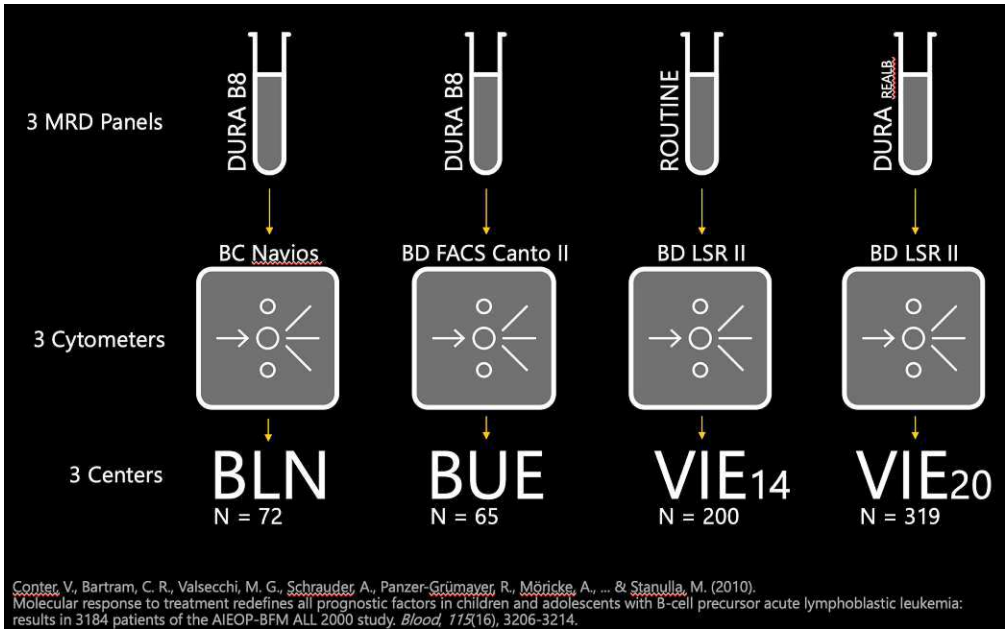

Figure 4.1: Illustration of the different data acquisition procedures for each clinic in ALL.

4.2 Data Collection - AML

The AML samples used in this study were acquired using two advanced, pre-designed antibody cocktail tubes: the Colony Forming Unit (CFU) tube and the Leukemia Associated Immunophenotype (LAIP) tube. These tubes combine well-established and innovative markers in knowledge-based configurations, forming a sophisticated multicolor staining panel for FCM. Initially developed for detecting and quantifying MRD in AML, the samples were collected during routine clinical practice at the St. Anna Hospital of Vienna between 2017 and 2024. Each patient contributed a pair of samples: one processed with the LAIP tube and the other with the CFU tube.

4.3 Dataset Structure

In our FCM analysis, each sample is represented by a matrix known as the event matrix. This matrix captures quantitative, cell-specific measurements, with each row representing a unique cell and each column corresponding to a measured marker or feature.

Let N represent the total number of cells in a sample, typically ranging from 10^5 to 10^6. The exact number varies between samples due to differences in sample preparation or biological context. The number of markers, denoted by m, are the measured molecular characteristics for each cell, generally ranging from 10 to 20, depending on the specific

antibodies used in the assay. Each marker corresponds to a targeted antibody that binds to a specific antigen on the cell's surface or intracellular components, providing crucial information about the cell's phenotypic profile.

While the numbers N and m can vary, a subset of markers, referred to as the *base panel*, is consistently measured across all samples. This base panel ensures a degree of uniformity in the dataset, allowing for comparative analysis among samples and enabling robust model training. Measurements from markers outside this base panel are excluded from our analysis to maintain a standardized set of features, effectively fixing m to include only the base panel markers.

Each row vector in the matrix, denoted by x_n for $n \in \{1, \ldots, N\}$, provides a quantitative profile of the markers for an individual cell. The values in these vectors, which indicate the fluorescence intensity for each marker, provide insights into the presence and density of specific cellular markers. This allows for the identification and classification of various cell types based on their expression profiles, which is especially valuable in distinguishing leukemic blast cells in ALL and AML from normal lymphoid cells.

Flow cytometers analyze cells individually as they pass through the machine in a fluid stream, creating an ordered sequence based on acquisition timing. However, for analysis, the exact order of cells is often irrelevant. Thus, rather than treating the dataset as a sequential structure, we can view it as an unordered set of feature measurement vectors. In this representation, each sample is treated as a collection of vectors (each representing marker measurements for a cell) without regard to their acquisition order. This bag of vectors approach aligns well with ML methods designed for unordered data, such as those leveraging Set Transformers and permutation-invariant architectures.

Figure 4.2 shows an example of the structure of our dataset used for ALL.

4.4 Challenges of Medical Data

Medical data exhibits several distinctive properties that pose unique challenges for its application in ML and artificial intelligence. These challenges arise from the data's specific characteristics, including heterogeneity, regulatory constraints, and quality issues, often less pronounced in other domains. Properly addressing these challenges is essential for developing reliable diagnostic and prognostic tools in healthcare, particularly for diseases like ALL and AML.

Class Imbalance and Rare Event Detection - Medical datasets often exhibit extreme class imbalance, particularly in conditions like ALL and AML, where leukemic cells (positive cases) may represent only a small fraction of the total cells analyzed. The clinical significance of rare events further complicates this imbalance: detecting a few MRD-positive cells among thousands of normal cells is critical for determining treatment efficacy. This challenge is unique to medical data due to the disparity in class importance. Missing rare positive cases can have severe clinical consequences.

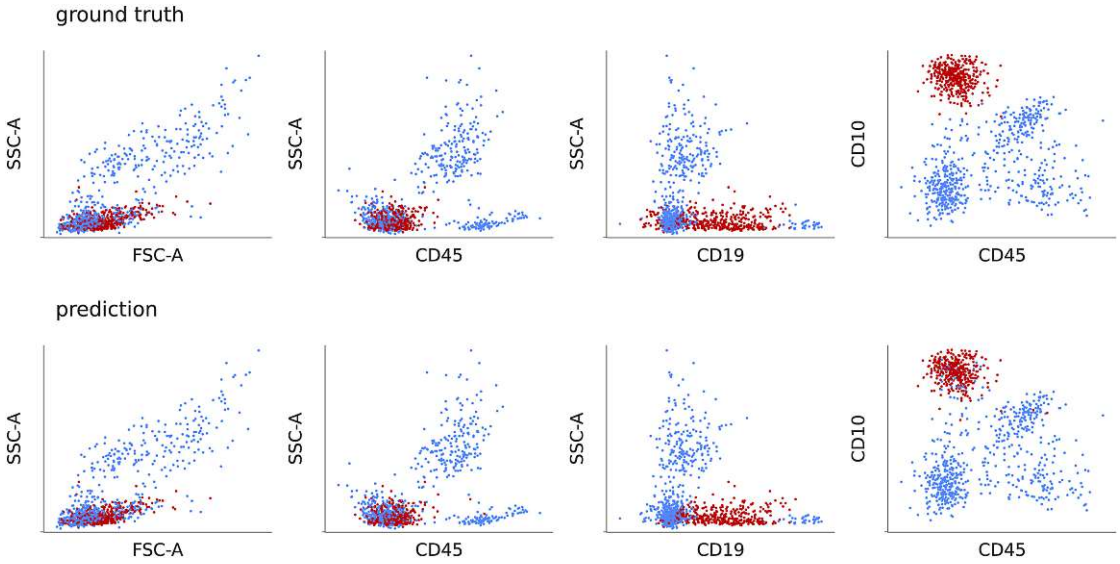

Figure 4.2: Example of the dataset structure we used for our experiments for ALL. On top is the ground truth, while on the bottom are the predicted results of a baseline model architecture. Red dots represent blast cells while blue dots represent healthy cells. The figure is adapted from the original paper by Wodlinger et al. [WRW+22].

Heterogeneity Across Institutions - Medical data is inherently heterogeneous due to variations in demographics, diagnostic practices, and equipment across institutions. For example, FCM data from Vienna, Berlin, and Buenos Aires differ not only in patient populations but also in the instrumentation and protocols used, which affects the data distribution. This necessitates that ML models generalize well while remaining robust to site-specific artifacts, often requiring domain adaptation or meta-learning techniques to harmonize the differences.

Data Privacy and Ethical Constraints - Medical data is highly sensitive and subject to stringent regulations, such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States and the General Data Protection Regulation (GDPR) in Europe. These regulations limit how data can be collected, shared, and analyzed. For ML applications, the challenge lies in ensuring data privacy while preserving the utility of the data. Approaches like differential privacy and federated learning have emerged as promising solutions for training models on sensitive data without compromising privacy [RHL+20].

Data Quality and Annotation Challenges - Medical data often suffers from quality issues, including missing values, noise, and inconsistent annotations. Unlike other fields where missing data can be handled statistically, in medicine, information gaps can lead to misdiagnoses or treatment delays. For example, variations in FCM protocols or marker intensity measurements can obscure critical patterns, necessitating robust preprocessing pipelines and noise-resilient models.

The challenges of annotation are particularly acute. Unlike crowd-sourced annotations in other domains, medical annotations require expert input, which is costly and time-intensive. In diseases like ALL and AML, annotations must identify cell types and reflect subtle morphological or biomarker-based differences.

Interpretability and Explainability - In the healthcare field, model interpretability is crucial. Unlike other domains where "black-box" models may be acceptable, clinicians in medicine must understand how a model arrives at its predictions to trust its outputs and incorporate them into decision-making processes. Many complex models, particularly deep learning methods, often lack this interpretability, raising concerns about their adoption in clinical settings.

Models must provide explainability to identify which features or inputs are most influential in driving a prediction [KWK+23]. Techniques like Shapley Additive Explanations (SHAP) [Lun17] and Local Interpretable Model-Agnostic Explanations (LIME) [RSG16] have been developed to interpret predictions from complex models, as well as Transformers models that mimic the decision process or human experts [KWK+23]. However, balancing model accuracy and interpretability remains an ongoing challenge.

CHAPTER 5

Methodology

The proposed method to enhance the model's ability to generalize to new, unseen diseases, referred to as tasks, using only a few training examples, involves a neural network structure with two main components: a Transformer network for classifying cells as blast or non-blast cells, and a meta-learning framework designed to train the Transformer in a manner that facilitates learning task-general features and quick adaptation to unseen diseases. In this section, we will first go through the details of the MAML framework. Then, we will explain the Transformer architecture used alongside MAML and, lastly, we will describe our training strategy.

5.1 Problem Formulation

The primary purpose of this work is to develop a model capable of classifying blast cells in ALL and AML using FCM data while effectively generalizing to unseen diseases (tasks) with limited training data. With samples consisting of cells as input, the goal is to classify every cell in the sample. Traditional ML approaches face challenges in this domain due to small dataset sizes and high variability in cell features. To address this, we propose a solution that leverages **Meta-Learning** with a **Transformer-based architecture**, designed to adapt to new, unseen tasks with minimal data quickly.

In our problem formulation, each dataset (as presented in Tables 4.1 and 4.2) is considered a separate task. Every dataset consists of a collection of samples and the model must classify all individual cells within a sample as either blast cells (cancerous cells) or non-blast cells (healthy cells). The challenges arise because the model must generalize across different tasks despite having limited data from each specific task. The aim is to train a model that can learn general features applicable across multiple tasks (diseases) and then, most importantly, quickly adapt to unseen datasets such that the overall performance on unseen data increases compared to models trained with standard training frameworks on multiple datasets.

The classification task is thus framed as a few-shot learning problem. In this scenario, the model is trained using a meta-learning approach: learning generalizable features from a set of training tasks and rapidly adapting to a new task (e.g., data from a different dataset or disease) with only a small number of examples.

The crucial components of the problem are:

- **Tasks** - Each dataset or phenotype (as presented in Chapter 4) is considered a task, with the goal of classifying each cell of a sample as blast or non-blast using sample wide information via self-attention.
- **Objective** - To develop a model capable of generalizing to new, unseen tasks using limited training data.
- **Challenge** - The model must be trained using small sample sizes and extensive variability in data across tasks.

We apply MAML to a Transformer architecture in this context. MAML allows the model to learn effective initialization parameters, enabling it to adapt to new tasks with minimal fine-tuning. The Transformer architecture, known for its self-attention mechanism, is used to capture the relationships between individual cells in a FCM sample, providing robust features for classification.

In summary, the problem is framed as a task-general classification challenge in a few-shot learning context. The meta-learning approach aims to optimize the model's ability to generalize to new tasks and adapt quickly with limited data.

5.2 Meta-Learning Framework

A key aspect of this thesis is the use of meta-learning, often referred to as "learning to learn". Meta-learning optimizes the learning process, enabling models to adapt more effectively to new tasks with minimal data by introducing an inductive bias that guides learning toward solutions suitable for a wide range of tasks. This inductive bias allows the model to effectively generalize across tasks by encoding prior knowledge that influences the search space of potential solutions. In particular, we apply MAML, a meta-learning framework first introduced by Finn et al. [FAL17]. The core idea behind MAML is to learn an optimal initialization for the model's parameters, such that it can quickly adapt to new tasks with only a few training examples, making it particularly suited for few-shot learning. MAML is expected to learn general, higher-level features rather than task-specific ones. These general features are useful when the model encounters new, unseen tasks, as it can quickly adapt using the knowledge learned during the meta-training phase.

The following part of the section demonstrates the MAML algorithm 5.1 specific to the classification task that we are addressing using pseudocode.

Algorithm 5.1: MAML for Few-Shot Supervised Learning

Input: $p(\tau)$: distribution over tasks
Input: α, β: step size hyperparameters
randomly initialize θ
while *not done* **do**
  Sample a batch of tasks $\tau_i \sim p(\tau)$
  forall τ_i **do**
    Sample K datapoints $D = x^{(i)}, y^{(j)}$ from τ_i
    Evaluate $\nabla_\theta L_{\tau_i}(f_\theta)$ using D and L_{τ_i} in 5.1
    Compute adapted parameters with gradient descent: $\theta_i^{'} = \theta - \alpha \nabla_\theta L_{\tau_i}(f_\theta)$
    Sample datapoints $D_i^{'} = x^{(i)}, y^{(j)}$ from τ_i for the meta update
  end
  Update $\theta \leftarrow \theta - \beta \nabla_\theta \sum_{\tau_i \sim p(\tau)} L_{\tau_i}(f_{\theta_i^{'}})$ using each $D_i^{'}$ and L_{τ_i} in 5.1
end

The algorithm requires two input parameters: $p(\tau)$, which is a distribution over the tasks used to train MAML, in our case, the different datasets collected from patients treated for ALL or the different AML phenotype datasets, and two learning rates, α, and β. The learning rate α controls the task-specific updates (inner loop), while β governs the updates to the meta-model (outer loop).

The first step is initializing the meta-model with a random set of weights θ. We then sample a batch of tasks from the task distribution $p(\tau)$. For each sampled task, the meta-model is fine-tuned on a **support set** and evaluated on a **query set** specific to that task. The task-specific models are adapted using α, while the performance on the query set affects the meta-model's update. In the final step, the meta-model's parameters θ are updated by computing the gradients of the summed loss from all the tasks' query sets using the learning rate β. In this case the loss function is a Binary Cross-Entropy (BCE) loss described by Equation 5.1

$$BCE = -(y \log(p) + (1 - y) \log(1 - p)) \tag{5.1}$$

where log is the natural logarithm, y is a binary indicator (0 or 1) for the correct class to be predicted, and p is the output probability of the model.

A central aspect of MAML is its bi-level optimization structure, which involves two loops: the **inner loop** and the **outer loop**. These loops allow the model to learn both task-specific adaptations and generalized knowledge across tasks. In addition, task sampling is crucial in ensuring the meta-model's ability to generalize to new, unseen tasks. Figure 5.1 shows the workflow of MAML.

The **inner loop** focuses on task-specific adaptation. For each task sampled from the distribution $p(\tau)$, the model uses the small labeled support set to fine-tune its parameters.

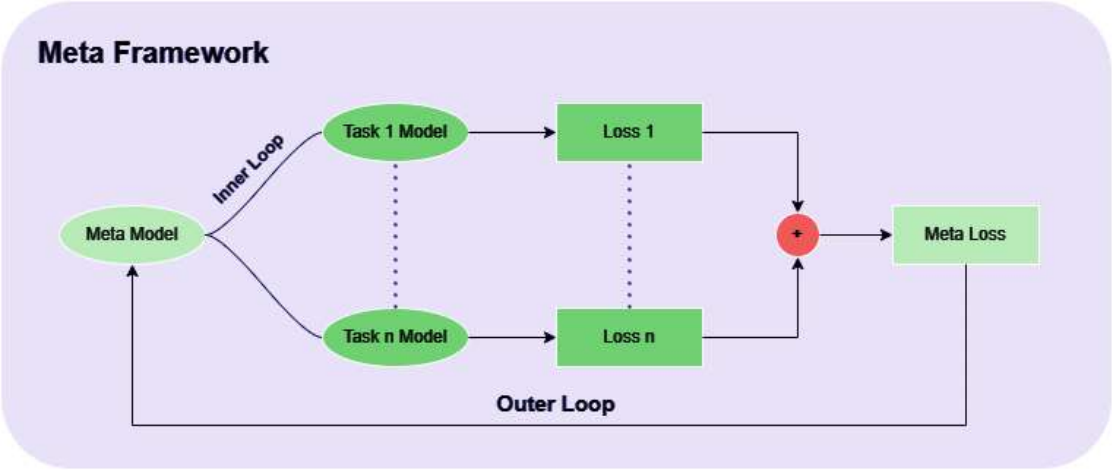

Figure 5.1: Illustration of the general MAML workflow, showing how the meta-model adapts to each task by fine-tuning task-specific models and combining their losses to update the meta-model's parameters.

The model learns task-specific features during this phase by performing a few gradient descent updates on the support set. However, these updates only apply to the task-specific model and do not immediately affect the meta-model. For instance, in the case of ALL, the task might be to classify blast cells for a specific patient's dataset. The support set would contain a small number of labeled cells (blast and non-blast), and the inner loop updates the task-specific model based on these few examples. The fine-tuned model is then evaluated on the query set (which acts as the test data for the task) to compute the task-specific loss. This inner loop allows the model to specialize in each task. But, more importantly, the meta-learning framework uses the results from these tasks in the next phase to improve the model's general ability to adapt.

The **outer loop** is where the generalization across tasks happens. After the inner loop has fine-tuned the task-specific models, these models are evaluated on their respective query sets to obtain a task-specific loss. The meta-model's parameters are then updated based on the sum of these losses across all tasks through a gradient descent step. This update process helps the meta-model learn good initial weights, enabling it to quickly adapt to new tasks with only a few examples in future inner loop phases. By aggregating the information across multiple tasks, the outer loop ensures that the meta-model captures more generalized, task-agnostic features rather than features specific to one task. In the context of ALL and AML, this means that the meta-model learns broad, cross-patient features of blast cell classification so that when it encounters a new patient from a different clinic or belonging to a different phenotype, it can adapt quickly. As stated by Finn et al. [FAL17], to encourage the emergence of such general-purpose features, the aim will be to learn a model in such a way that the gradient-based learning rule can make rapid progress on new tasks drawn from the distribution of tasks $p(\tau)$. So, as shown in Figure 5.2 adapted from the paper by Finn et al. [FAL17], "the aim will

be to find model parameters that are sensitive to changes in the task such that small changes in the parameters will produce large improvements on the loss function of any task drawn from $p(\tau)$ when altered in the direction of the gradient of that loss". [FAL17]

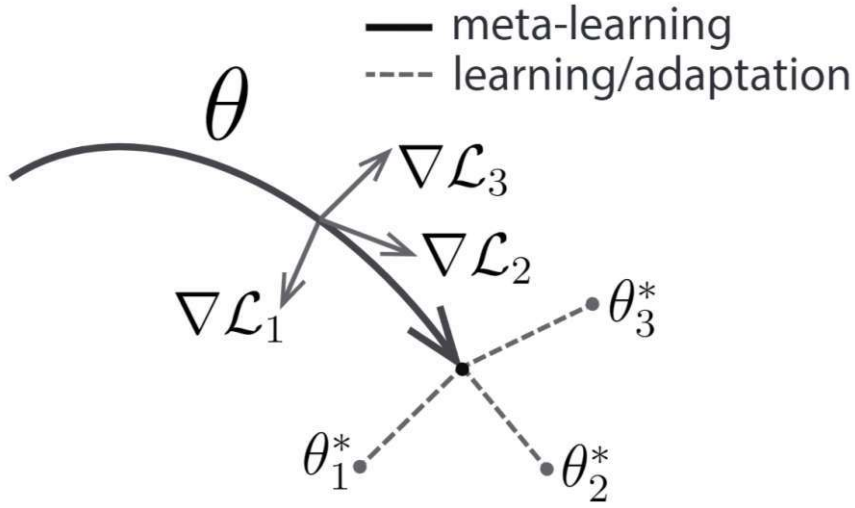

Figure 5.2: Illustration of the idea of the MAML meta-update, showing how the combined gradients of the task-specific models are used to perform the higher-level meta-update to encourage quick adaptation to new tasks. The figure is adapted from the original paper by Finn et al. [FAL17].

While the original MAML algorithm 5.1 has been proven effective in a variety of domains, practical considerations required us to make certain modifications. Below, we present the pseudocode of our adapted version of the MAML algorithm. Each modification was made to address specific limitations or optimize certain aspects of the training process. Following the pseudocode, we will discuss these modifications in detail, explaining the reasons behind each one and their potential impact on model performance. Algorithm 5.2 shows the exact steps implemented in the context of this thesis.

The adapted version of MAML presented here differs from the original algorithm in several ways to better suit the practical challenges encountered during the development process. These changes were motivated by computational efficiency, framework expansion, and the inherent instability of MAML. Below, we will walk through each modification, detailing its expected impact on performance.

Computational Efficiency - To enhance the computational efficiency of the MAML framework, we used the First-Order MAML (FOMAML) variation, which is specifically designed to reduce the computational load that makes standard MAML challenging to deploy in certain environments. The original MAML algorithm is computationally intensive due to its reliance on second-order gradients, which require backpropagating through both the task-specific updates and the meta-model. Essentially, MAML optimizes for a good initialization by training over multiple tasks, necessitating that each task's

Algorithm 5.2: Adapted MAML for Few-Shot Supervised Learning

Input: $p(\tau)$: distribution over tasks
Input: α, β: step size hyperparameters
randomly initialize θ
for *n outer loops* **do**
  Initialize the meta loss $L_{tot} = 0$
  forall *training tasks* **do**
    Divide the tasks training set into support and query sets
    Clone the meta-model to generate the task-specific model
    for *n inner loops* **do**
      Evaluate $\nabla_{\theta_i} L_\tau(f_{\theta_i})$ using the support set and L_τ in 5.1
      Compute adapted parameters with gradient descent: $\theta_i' = \theta_i - \alpha \nabla_{\theta_i} L_\tau(f_{\theta_i})$
      Compute a weighted loss $L_\tau = w_i \cdot L_\tau(f_{\theta_i'})$ using the query set and MSL weights
      Add L_τ to the meta loss as $L_{tot} = L_{tot} + L_\tau$
    end
  end
  Update $\theta \leftarrow \theta - \beta \nabla_\theta L_{tot}$ using BCE in 5.1
end

gradients backpropagate to the meta-level, a process that significantly increases the computational overhead.

In the original MAML, the process is designed to make the model parameters highly adaptable, allowing for fine-tuning with only a few steps of gradient descent on new tasks. However, the necessity of computing second-order derivatives makes training slow and resource-intensive, especially as the number of tasks or model complexity increases. This inefficiency not only lengthens training time but also limits scalability and real-world applicability, particularly in resource-constrained settings.

FOMAML, on the other hand, simplifies this process by approximating the update steps without calculating the full second-order gradients. Instead of backpropagating through the task-specific gradient updates to the original meta-model, FOMAML only applies the gradients calculated within each adapted task-specific model, treating the task update as a first-order approximation. This significantly reduces the number of operations needed since the meta-gradient update does not require differentiation through the inner loop adaptation process. While this approach may sacrifice some theoretical precision, it has been shown to perform comparably to full MAML in many cases, especially when computational resources are limited. Thus, by using FOMAML, we achieved a more computationally feasible training process while retaining the adaptability benefits of MAML for new, unseen tasks, as shown in Table 5.1.

Framework	Average Val F1 Score	Training Time
MAML	0.73	3 days
FOMAML	0.71	5 hours

Table 5.1: Comparison in validation f1 scores and training times between the MAML and FOMAML frameworks.

Framework Expansion - To expand the original MAML framework and enhance its adaptability across various tasks, we incorporated several hyperparameters that were not originally included in MAML. These include distinct inner and outer loop hyperparameters, a support-to-query ratio, and learning rate schedulers. While these additions make the framework more versatile, they also introduce complexity due to the increased number of parameters that require tuning:

- **Inner and Outer Loop Hyperparameters** - In standard MAML, there are relatively few hyperparameters, primarily focused on the step size for task-specific updates and the meta-optimization step. We introduced separate hyperparameters for the inner and outer loops to enhance the framework's adaptability. The inner loop, or task-specific loop, updates the model's parameters within each task using a small number of gradient steps. In contrast, the outer loop optimizes the meta-model across tasks. These additional settings for the number of gradient steps in each loop allow for fine-tuning of the training depth, influencing both the speed of adaptation and model generalizability. However, the trade-off is that fine-tuning these parameters becomes more complex, increasing the risk of suboptimal hyperparameter configurations.

- **Support-to-Query Ratio** - Another important hyperparameter is the ratio between each task's support and query sets. The support set, used for task-specific adaptation, and the query set, which evaluates the model's performance post-adaptation, play critical roles. Adjusting this ratio allows for control over the size of each subset, impacting how well the model generalizes from the support to the query data within a task. In tasks with limited data, fine-tuning this ratio is crucial, as an imbalanced support-query setup could lead to overfitting on the support set or insufficient generalization on the query samples. However, optimizing this ratio further increases the number of configurations to test, complicating the training process.

- **Learning Rate Schedulers** - We implemented two types of learning rate schedulers for both the meta and task-specific updates. For the meta-model, we applied a cosine annealing scheduler [LH16] that gradually decreases the learning rate with each outer loop step. Cosine annealing has proven effective in various domains by allowing smoother convergence and preventing abrupt parameter changes, which

is particularly useful for high-dimensional data like FCM. Instead, for the task-specific models, we applied a scheduler that uses a "Reduce on Plateau" approach, which decreases the learning rate within each inner loop when improvements in the meta-loss plateau. This strategy helps prevent overshooting the optimal point and encourages convergence in the later stages of meta-training.

- **Number of Examples per Outer Loop** - This hyperparameter controls the number of examples sampled from each task's dataset during each outer loop iteration. By adjusting the number of examples processed in each meta-update, this setting directly impacts computational efficiency, as it determines the batch size at the task level. Smaller sample sizes allow for quicker updates but may lead to less stable training, particularly when data is sparse or noisy. Conversely, using larger sample sizes per outer loop enhances the robustness of gradient updates but can slow down training. Fine-tuning this parameter is essential for balancing efficiency with model stability, especially for larger datasets with diverse cell types.
- **Task Balance Hyperparameter** - To manage the variability in dataset size between different sources (e.g., Vienna with 519 samples versus Buenos Aires with only 65), we included a task dataset balance hyperparameter. This parameter sets the proportion of samples taken from each dataset during training. It prevents over-represented datasets from dominating the training, which could cause the model to overfit larger datasets and underperform on smaller ones. Controlling the balance ensures that each dataset contributes fairly, facilitating generalization across diverse data sources. This is particularly important in medical datasets, where heterogeneity across sources reflects different population demographics and equipment setups. However, by restricting the number of examples from the larger datasets, we also decrease the information the meta-model can derive from the larger tasks. Hence, we need to consider this trade-off.

However, these additional hyperparameters introduce another layer of complexity, as the initial learning rate and scheduler settings directly impact performance. Balancing these hyperparameters to avoid excessive tuning and potential overfitting remains challenging, requiring a balance between adaptability and computational efficiency.

In summary, expanding the MAML framework with these hyperparameters and schedulers creates a more flexible architecture capable of handling the demands of varied tasks, particularly when data is limited. However, managing the added complexity without overcomplicating the tuning process presents its own challenges, highlighting the importance of carefully selecting and evaluating each hyperparameter.

Instability - To address the inherent instability of MAML, we incorporated a technique known as Multi-Step Loss (MSL) [AES18], which stabilizes the meta-learning process by refining the task-specific updates in the inner loop. While effective for learning with limited data, MAML often suffers from instability during training, particularly in scenarios with noisy data, diverse task distributions, or high learning rates. This

instability arises because MAML's original optimization strategy involves backpropagating through multiple task-specific updates, which can produce large gradients that destabilize the parameters of the meta-model, leading to oscillatory or diverging behaviors.

MSL addresses these challenges by evaluating the loss at several points during the task-specific update process rather than only at the final step. In traditional MAML, the meta-objective considers only the loss after a predefined number of gradient updates on each task. In contrast, MSL assesses the intermediate loss at various stages throughout these updates and aggregates them to create a smoother, MSL function that stabilizes the gradients. This approach prevents the optimizer from overreacting to abrupt changes in task-specific gradients and promotes a consistent improvement path for each update.

In our modified MAML implementation, MSL operates by:

1. Evaluating the intermediate steps within each task's inner loop. As the model takes gradient steps to adapt to a specific task, we calculate the loss at each intermediate step.

2. Aggregating the losses across these steps forms the final meta-objective. This aggregation dampens the influence of any single noisy gradient, resulting in a more stable overall meta-update.

3. Applying the aggregated MSL in the outer loop smooths the meta-objective, allowing the model to learn in a more stable and controlled manner. This improvement enhances convergence speed and reduces oscillatory behavior during training.

The MSL is shown in Equation 5.2 and is calculated as:

$$MSL = \sum_{i=0}^{I} w_i \nabla L(f_{\theta_i}) \tag{5.2}$$

where I is the total number of inner loops, f_{θ_i} is the task-specific model at the i-th inner loop, and w_i is the weight assigned to the gradients of the i-th iteration.

Our framework has as many weight parameters as we have inner loops, and the weights are learned during training. We use an Adam optimizer [Kin14] with a learning rate of $5e^{-4}$ for the weights.

By distributing the influence of each task's gradients over multiple steps, MSL reduces the likelihood of extreme gradient shifts. This is particularly beneficial for high-variance tasks or noisy data, which are common in clinical applications like ALL and AML diagnosis. This approach results in a more consistent optimization trajectory, helping the meta-model converge more reliably, especially when learning from a diverse dataset with varying feature distributions. Our experiments found that MSL improved training

stability and model performance, ultimately enhancing the model's ability to generalize across new tasks.

Integrating MSL into our MAML framework mitigated the instability issues of the standard MAML algorithm, as we can see from Figures 5.3, 5.4, 5.5, and5.6.

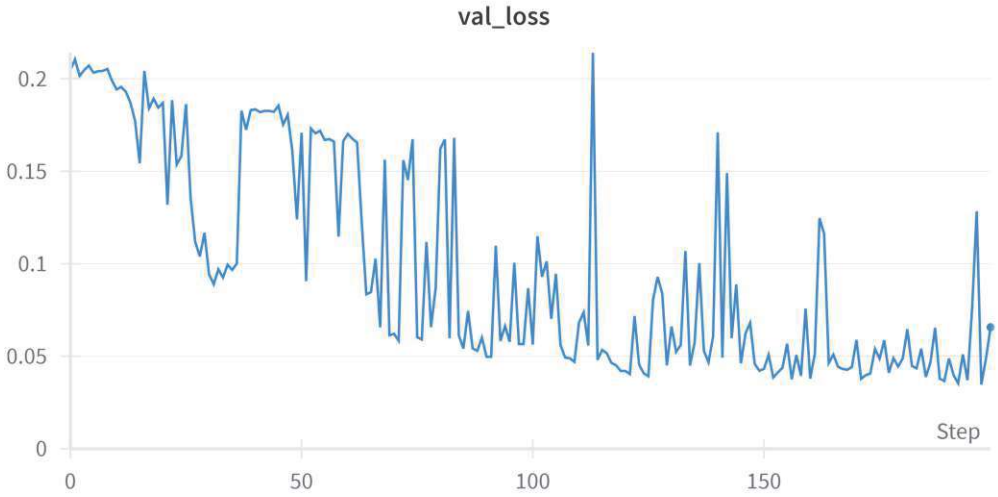

Figure 5.3: Illustration of the error on the validation set over time plotted against the epochs for a MAML model trained without using the MSL optimization for stability.

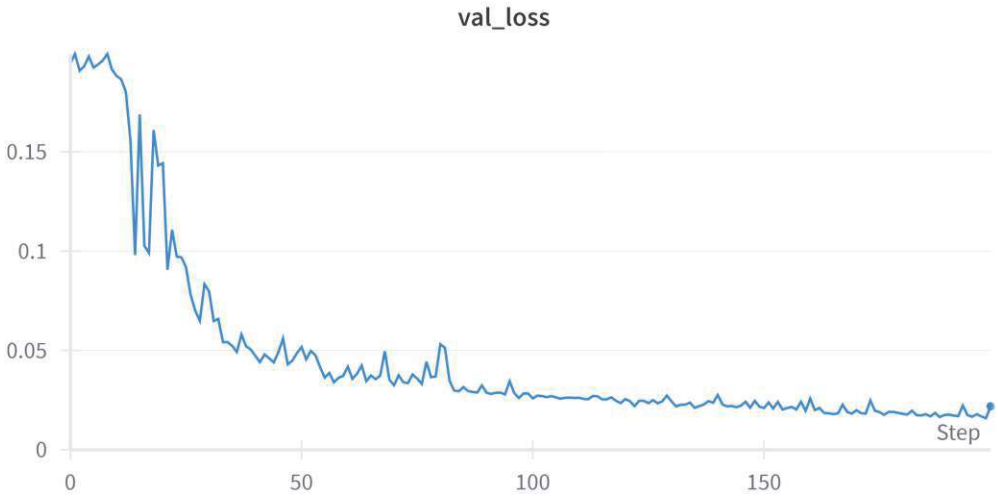

Figure 5.4: Illustration of the error on the validation set over time plotted against the epochs for a MAML model trained by using the MSL optimization for stability.

Figures 5.3 and 5.4 illustrate the behavior of the loss function over the epochs. In

Figure 5.3, we can immediately observe that the loss is highly unstable, characterized by significant jumps and spikes without smooth convergence. This behavior is mitigated by using MSL loss, as shown in Figure 5.4, where the loss converges more smoothly and stably.

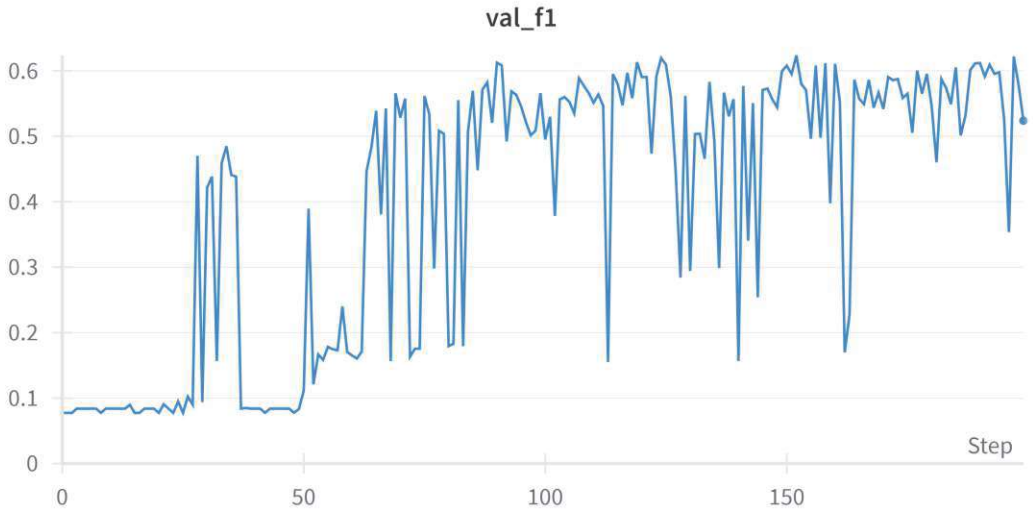

Figure 5.5: Illustration of the f1 score plotted against the epochs without using MSL.

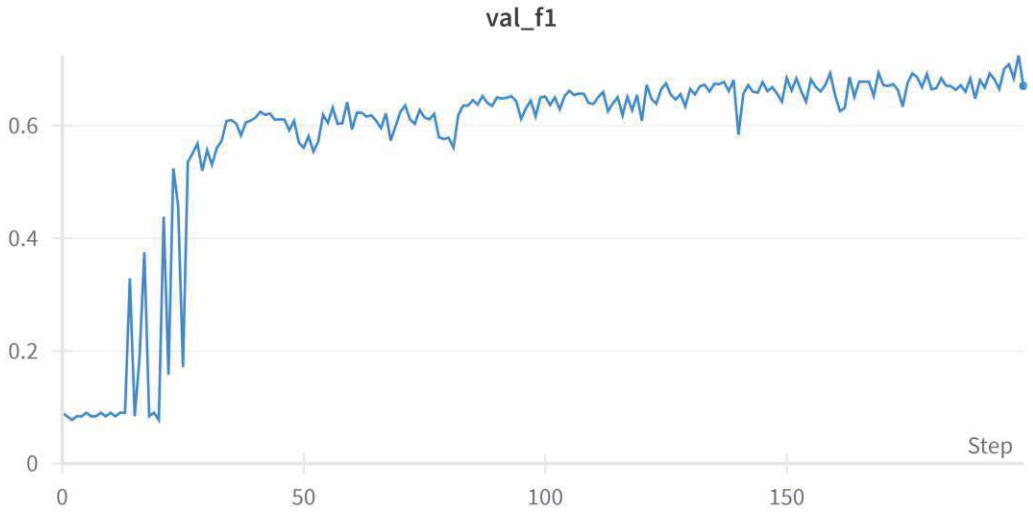

Figure 5.6: Illustration of the f1 score plotted against the epochs using MSL.

Figures 5.5 and 5.6 depict the validation F1 score across the epochs. In Figure 5.5, it is evident that the validation F1 score exhibits high instability, marked by significant

fluctuations and spikes, lacking smooth convergence. In contrast, Figure 5.6 demonstrates that employing MSL loss mitigates this instability, leading to a smoother and more stable convergence of the validation F1 score.

So, MSL has proven to be an efficient solution to the problem of MAML instability, making it a reliable approach for our application.

5.3 Transformer Model Architecture

For our experiments, we utilize the same Transformer architecture employed by Woedlinger et al. [WRW+22], around which we build our MAML-based framework. In this section, we describe the integration of the Transformer model within our MAML-based architecture, specifically detailing the adaptations required to handle FCM data, which lacks the typical grid structure found in image and sequential text data. Unlike text or images that can be organized along one- or two-dimensional grids, FCM data is unstructured, making conventional neural network approaches less effective. Instead, Transformers are a suitable choice, as they use self-attention to model relationships within sets of vectors, allowing them to adapt effectively to non-sequential data like ours.

Standard Transformers utilize a self-attention mechanism, described in Equation 5.3, which enables models to capture interactions between elements in an input set. The key components of self-attention are the query, key, and value matrices (Q, K, V) derived from the input embeddings. By computing the attention score as a scaled dot product between Q and K, Transformers model relationships across the entire input set. However, this self-attention mechanism is computationally demanding, with memory requirements that grow quadratically with input size [KVPF20].

$$Attention(Q, K, V) = Softmax(Q^T K)V \tag{5.3}$$

For FCM data, the quadratic complexity of self-attention can result in prohibitive memory consumption, as each sample comprises high-dimensional measurements for potentially thousands of cells. Therefore, directly applying traditional Transformers would exceed memory limits, especially with larger batch sizes or higher feature dimensions.

Woedlinger et al. [WRW+22] employed an Induced Attention Block (IndAttnBlock) as a memory-efficient alternative to standard self-attention to address this. This approach reduces memory demands from quadratic to linear complexity by incorporating a two-stage attention mechanism. Inspired by Lee et al. [LLK+19], the IndAttnBlock introduces a small set of learnable parameters called inducing points, which serve as intermediary features. These induced points enable the attention operation to be decomposed into two sequential steps:

- The inducing points $I \in \mathbb{R}^{k \times h}$ with $k \in \mathbb{N}$ and $h \in \mathbb{R}^{k \times h}$ serve as a condensed representation, attending to the input set $X \in \mathbb{R}^{N \times m}$ to capture global information.

- These attended features are then used as queries to attend back to the input set, consolidating relevant information from the input data.

The IndAttnBlock is shown in Equation 5.4

$$IndAttnBlock(X) = TransfBlock(X, TransfBlock(I, X)) \tag{5.4}$$

where the TransfBlock is described in Equation 5.5

$$TransfBlock(X, Y) = LayerNorm(AttnBlock(X, Y) + FF(AttnBlock(X, Y))) \tag{5.5}$$

LayerNorm is the normalization layer proposed by [Ba16], and AttnBlock is described by Equation 5.6.

$$AttnBlock(X, Y) = LayerNorm(X + Attn(X, Y, Y)) \tag{5.6}$$

By breaking down the $Q^T K$ operation into two smaller attention steps, IndAttnBlock maintains the representational power of self-attention while minimizing memory consumption. This approach has been shown to perform well on unordered, non-grid data, preserving the permutation invariance that is critical for processing FCM data. Figure 5.7, adapted from the paper by Woedlinger et al. [WRW+22], shows the whole transformer block described so far.

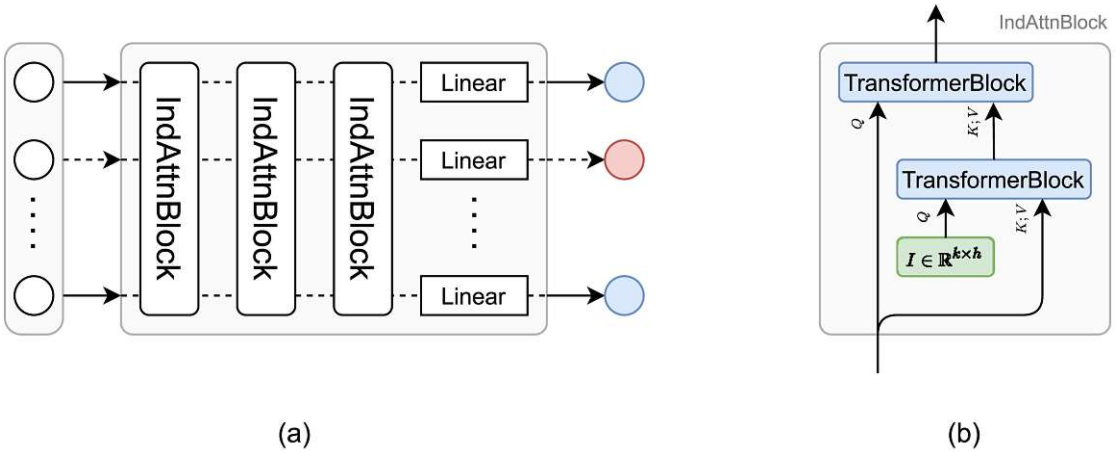

Figure 5.7: Illustration of the Transformer block used in the Transformer architecture employed in our experiments with MAML. The figure is adapted from the original paper by Woedlinger et al. [WRW+22].

The modified Transformer model proposed by Woedlinger et al. [WRW+22] consists of a sequence of three IndAttnBlocks followed by a linear output layer. The absence of positional encoding further simplifies the model since the data does not follow a sequential order. The model is equivariant, which means that for every input token, there is a corresponding output token, and the embedding of one event in the sample uses information from potentially all other events of the sample (global information). This is motivated by the fact that clinicians find subpopulations by using their spatial relation to different clusters. We utilize a lightweight configuration with 16 inducing points, a latent embedding size of 32, and 4 attention heads per layer. This configuration balances computational efficiency and model capacity, making it feasible to process FCM samples without extensive computational resources.

In our architecture, we apply BCE loss (Equation 5.1) during training to accurately classify each cell as either a blast or non-blast cell.

This Transformer configuration, optimized for computational efficiency, adaptability, and performance, serves as an effective backbone for MAML, enabling meta-learning across diverse tasks with FCM data.

5.4 Training Strategy

This section outlines the training strategy used to achieve optimal performance and computational efficiency in adapting the MAML-Transformer model to FCM data. The following hyperparameters were selected based on preliminary experiments that compared various configurations, resulting in the choices detailed below.

Learning Rates and Schedulers - We set the initial learning rate to 0.0005 for both task-specific and meta-updates. The meta learning rate (β) governs how quickly the meta-model updates in response to feedback from task-specific models. In contrast, the task learning rate (α) determines the rate of gradient updates within each task during inner loop adaptation. Given the sensitivity of the meta-learning framework to overfitting and convergence stability, we implemented learning rate schedulers to modulate the learning rates dynamically:

- **Meta LR Scheduler** - For meta-optimization, we applied a cosine annealing learning rate scheduler with parameters $T_{\max} = 10$ and $\eta_{\min} = 0.0002$. This strategy gradually decreases the learning rate, allowing for rapid learning at the start, followed by progressively slower updates to reach convergence, effectively balancing exploration and stability.

- **Task LR Scheduler** - To adjust the learning rate for task-specific updates, we used a "Reduce on Plateau" scheduler. This scheduler reduces the learning rate by a factor of 0.9 whenever the validation loss stagnates, with a patience setting of 5. This strategy facilitates further refinement when the model reaches performance plateaus,

preventing premature convergence and promoting fine-grained improvements during task adaptation.

Outer Loop and Inner Loop Parameters - In ALL experiments, each training epoch consists of four outer loop iterations, with each outer loop comprising four inner loop updates to adapt to the task-specific objective. As for AML, we set the number of outer loops to three and the number of inner loops to five. This balance between the number of inner and outer loops enables the model to maintain generalization capabilities while effectively adapting to individual tasks within the meta-learning framework.

- **Examples per Outer Loop** - To enhance training efficiency, we selected a subset of 10 examples for each outer loop iteration regarding ALL experiments. This approach ensures task adaptation remains computationally feasible while maintaining sufficient variability among the examples. As for AML experiments, given the scarcity of data, we used 5 examples for each outer loop.
- **Support-Query Ratio** - In ALL, we used a support-to-query ratio of 0.5, meaning that for each task, half of the data was allocated to the support set (for adaptation) and the other half to the query set (for evaluation). This ratio was optimal for balancing adaptation strength and evaluation stability across tasks. On the other hand, for AML, we chose a support-to-query ratio of 0.6, taking 3 examples for training and 2 for testing, giving the model enough examples to adapt to a specific task while also having enough examples allocated for the meta-update.
- **Dataset Balancing Ratio** - In ALL, we balanced the contributions from each dataset (vie, bln, and bue) by setting a sampling ratio of 3:1:1. This allocation emphasizes the vie data, as it is the largest subset within our full dataset, while still allowing for contributions from the smaller bln and bue datasets. This ratio addresses the imbalance between the datasets, ensuring that the model does not disproportionately learn from the smaller datasets or overfit the larger ones. In AML, since all the phenotype datasets (M2, M4, M5, and M7) have similar sizes, we set the sampling ratio to 1:1:1:1, which sets a perfect balance between all tasks.

These hyperparameter choices and training strategies were specifically selected to address the complexities of meta-learning with FCM data, achieving a balance between task adaptation efficiency and computational practicality. By applying carefully chosen learning rate schedules and balancing the contributions of different datasets, this training strategy aims to maximize model performance across the various domains represented in the meta-training set.

CHAPTER 6

Experiments

6.1 Experimental Setup

Our experiments were conducted in a controlled environment to ensure consistency and reproducibility. We carefully selected the computational resources, software versions, and framework configurations to maximize performance while minimizing potential sources of variability.

Hardware and Software Environment - The model was trained on an NVIDIA RTX A5000 Graphics Processing Unit (GPU) with 24 GB of memory, utilizing Compute Unified Device Architecture (CUDA) 12.1 to accelerate deep learning operations and optimize training speed. The GPU's memory capacity was critical for handling the computational demands of transformer-based architectures within the MAML framework, particularly given the large and complex datasets used in the experiments.

We employed Python 3.10.12 as the programming language, along with PyTorch version 2.2.2 [PGM+19], which is compatible with CUDA 12.1. PyTorch was chosen for its flexibility, extensive library support, and efficient integration with CUDA, allowing for parallelized operations on the GPU. Training and code management were conducted in Visual Studio Code, selected for its versatility, debugging capabilities, and seamless integration with Python and version control systems.

Frameworks and Libraries - The training setup included essential libraries and dependencies to support the experiments. PyTorch's built-in modules were utilized to construct and optimize the transformer model. At the same time, additional libraries were integrated to handle meta-learning functionalities within the MAML framework besides our implementation, such as the learn2learn library [AMD+20] which made working with meta-learning straightforward. With a CUDA-enabled configuration, the experimental framework leveraged GPU acceleration for all forward and backward passes, enabling efficient, extensive training and fine-tuning.

Configuration - We fixed all initial random seeds to ensure consistency across experiments, and data loading was configured to use deterministic options where available. Model weights, learning rates, and dataset loading were logged throughout training, enabling systematic performance evaluations. Additionally, output logs and model checkpoints were saved periodically to monitor convergence rates and identify potential overfitting during each experimental run.

6.2 Experiments Description

To address our core research questions, we designed a series of experiments centered around meta-training using the MAML framework. We assessed the model's adaptability and performance relative to conventional non-meta-learning approaches. Here, we describe each set of experiments in detail, including the methodologies used to assess comparative performance and evaluate the adaptability of MAML.

To assess the performance of MAML relative to baseline methods, we applied a "leave-one-out" training approach. For ALL, using three datasets from Vienna, Berlin, and Buenos Aires (abbreviated as vie, bln, and bue), we sampled two of them for meta-training and reserved the third for testing. For AML, having four datasets each representing a different phenotype (M2, M4, M5, and M7), we sampled three of them for meta-training and used the fourth one for testing. Specifically:

- **Meta Training** - Datasets were selected for meta-training for each experiment, with a 75-25 split between the training and validation sets. The training set is further divided at runtime into support and query sets. To encourage robustness, we reshuffled these support and query sets at each epoch, enabling each example to switch between roles across epochs and further diversify the training process.

- **Evaluation on the Unseen Task** - After meta-training for a maximum of 400 epochs in ALL, with early stopping set to 100 epochs, and, following the configurations from Woedlinger et al. [WRW+22], for a maximum of 100 epochs in AML, with early stopping set to 50 epochs, we evaluated the generalization performance by testing the meta-trained model on the full dataset of the remaining task. The non-fine-tuned meta-model was directly applied to this dataset, allowing us to examine its initial cross-task generalization capabilities. This process was repeated for each combination of tasks, both in ALL (i.e., vie-bln → bue, vie-bue → bln, bln-bue → vie) and AML (i.e., M2-M4-M5 → M7, ...).

- **Baseline Comparison** - For comparison, we implemented the baseline approach by training a standard model using the same training examples in a conventional, non-meta-learning setup. By directly contrasting the performance of the meta-learning model with that of the baseline model, we could better understand the advantages conferred by MAML for MRD assessment.

To evaluate the adaptability of MAML on new, unseen tasks, we fine-tuned the meta-trained model using a small subset of samples (1, 5, and 10 examples, respectively) from the left-out task and assessed its performance on the remaining examples in that dataset. This fine-tuning allowed us to quantify the model's rapid adaptation to new contexts, a key advantage of meta-learning.

We also fine-tuned the baseline models with 1, 5, and 10 examples from the unseen task to compare the adaptability between MAML and non-meta-learning approaches. By conducting these adaptation experiments across various train-validation splits, we aimed to verify the robustness and consistency of the results.

This setup allowed us to systematically investigate the effectiveness and adaptability of MAML for blast cell classification and MRD assessment while also ensuring a reliable baseline for comparison with traditional training approaches.

6.3 Performance Metrics

The following metrics were used to evaluate the effectiveness and adaptability of our MAML-based model in comparison to baseline approaches. These metrics include classification-focused measures such as accuracy, precision, recall, and F1-scores, as well as performance metrics related to training efficiency and adaptation speed.

Precision - Precision (Equation 6.1) assesses a model's ability to correctly identify positive instances among all predicted positives, which is crucial in scenarios with imbalanced classes.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{6.1}$$

Recall - Recall (Equation 6.2) measures the model's ability to identify all relevant instances (i.e., true positives) out of the total actual positive instances.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{6.2}$$

Average F1-Score - The F1-score (Equation 6.3) is the harmonic mean of precision and recall, providing a balanced perspective on a model's performance. The average F1-score (Equation 6.4) refers to the mean F1-score calculated across all samples (S is the number of samples).

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6.3}$$

$$\text{Average F1-Score} = \frac{1}{S} \sum_{i=1}^{S} \text{F1-Score}_i \tag{6.4}$$

Median F1-Score - The median F1-score offers an additional measure of robustness by capturing the central tendency of F1-scores across classes, thereby reducing the impact of outliers in the distribution.

Training Time - Training time (Equation 6.5) measures the total duration required to train the model within the meta-learning framework. This metric is essential for evaluating the computational efficiency of the MAML approach, particularly when compared to conventional training methods.

$$\text{Training Time} = \sum_{i=1}^{\text{epochs}} \text{Time per Epoch}_i \tag{6.5}$$

Adaptation Speed - Adaptation speed (Equation 6.6) quantifies how quickly the MAML-trained model can adjust to a new, unseen task using a limited number of support examples. This is typically measured by evaluating the model's performance over a specified number of fine-tuning steps on the unseen task.

$$\text{Adaptation Speed} = \text{Performance after fine-tuning} - \text{Initial Performance} \tag{6.6}$$

CHAPTER 7

Results

This chapter presents a detailed analysis of the outcomes from our experiments, evaluating both the effectiveness and adaptability of our meta-learning approach, specifically MAML, in the context of blast cell classification and MRD assessment. It aims to answer the research questions by comparing quantitative metrics across baseline and meta-learning-based methods, examining model performance under various configurations, and exploring adaptability to new, unseen tasks. The chapter synthesizes these results to inform the potential practical implications of our approach, guiding future research and applications in clinical settings.

7.1 ALL - Quantitative Results

In this section, we present the quantitative results of our experiments for ALL across various metrics. We evaluated average and median F1-scores to assess classification quality and additional metrics such as adaptation speed to evaluate efficiency and adaptability. The results are organized by primary research questions, focusing on the model's performance compared to baseline methods and its ability to adapt to new, unseen tasks.

7.1.1 How does the performance of MAML compare to existing state-of-the-art approaches for ALL MRD assessment?

The first set of experiments involved a single training run for each model (MAML and the baseline) across three different task combinations, following a "leave-one-out" structure. In this setup, MAML outperformed the baseline approach on the Berlin and Buenos Aires datasets, while it underperformed compared to the baseline model on the Vienna dataset. This initial observation indicated that MAML's meta-learning approach can leverage in a better way the information enclosed in a more extensive dataset (Vienna)

and use it to generalize on different smaller tasks (Berlin and Buenos Aires). However, when the models were trained on the smaller tasks (Berlin and Buenos Aires) and then tested on the larger task (Vienna), MAML did not produce the anticipated improvements for blast cell classification, and MRD assessment.

To ensure that the observed performance discrepancy was not due to random variability or specific data splits, we repeated the MAML training and evaluation process on four additional random splits of the data. This provided a more robust evaluation of MAML's consistency across different data partitions. The averaged results from these five runs were consistent with the initial findings, demonstrating that MAML's performance on Vienna remained consistently lower than that of the baseline model but higher than the baseline on Berlin and Buenos Aires. This systematic evaluation reinforced the conclusion that, under the current setup, MAML struggles to achieve comparable performance to a baseline method when trained with very small datasets but can leverage and apply the knowledge of a larger dataset to smaller datasets better. Table 7.1 reports the comparison results.

Run	Avg F1	Median F1
Baseline Vienna	0.67	0.80
MAML Vienna	0.63	0.75
Baseline Berlin	0.58	0.72
MAML Berlin	0.64	0.75
Baseline Buenos Aires	0.70	0.88
MAML Buenos Aires	0.76	0.90

Table 7.1: Performance of MAML compared to the baseline model in terms of average and median F1 scores averaged on 5 runs without any adaptation on ALL. The dataset names in the rows specify the dataset that was used for testing, hence the remaining two were use for training the corresponding model.

7.1.2 How efficiently can a model trained using MAML adapt to new, unseen tasks in ALL?

To investigate how well both the MAML and baseline models adapt to new, unseen tasks, we conducted a series of adaptation experiments. Specifically, each model was fine-tuned on one of the left-out third tasks (Vienna, Berlin, or Buenos Aires) with a learning rate of 0.0001 after training on the other two tasks. For each adaptation, we used sets of 1, 5, and 10 examples to determine whether increasing the number of adaptation examples improved the model's performance. These experiments were repeated five times for each adaptation setting, resulting in a total of 15 adaptation runs per model (15 for MAML and 15 for the baseline).

The results, displayed in the tables within the respective subsections for each task, reveal

distinct patterns of adaptability based on the task and model. For the Vienna dataset, the baseline model demonstrated slightly better adaptation results, confirming that it remains the best-performing model even after adaptation. In contrast, for the Berlin dataset, both models achieved similar performance metrics after adaptation, indicating that the baseline model adapted more effectively given that it started from a worse initial point. Lastly, for the Buenos Aires dataset, MAML exhibited superior adaptation performance, reinforcing its status as the best model for this dataset. The sections below discuss the dataset-specific adaptation results in greater detail.

Vienna - For Vienna, the baseline model consistently demonstrated a stronger adaptability profile than MAML, as evidenced by both the average and median F1 scores. With 10 adaptation examples and 10 training epochs, the baseline model achieved a 4% increase in average F1 score and a 3% improvement in median F1 score over the initial performance, indicating better adaptation to the Vienna task. This suggests that, for the Vienna dataset, the baseline model was better equipped to generalize from a few examples, despite lacking the meta-learning structures designed to enhance adaptability in MAML. Table 7.2 shows the gains/losses in percentage concerning the respective initial performance reported in Table 7.1.

Run	1 Epoch		5 Epochs		10 Epochs	
	Avg F1	Median F1	Avg F1	Median F1	Avg F1	Median F1
Baseline 1-shot	+0%	+0%	+0%	+0%	+1%	+1%
MAML 1-shot	+0%	+1%	+0%	+1%	+0%	+1%
Baseline 5-shot	+0%	+0%	+0%	-1%	+1%	+0%
MAML 5-shot	-1%	+0%	+0%	+0%	+2%	+1%
Baseline 10-shot	+0%	+1%	+3%	+3%	+4%	+3%
MAML 10-shot	+1%	+0%	+1%	+1%	+2%	+2%

Table 7.2: Performance of MAML compared to the baseline model in terms of adaptation capability to the Vienna dataset. The comparison is made by computing the average percentage gains/losses on the F1 scores of the models after 1, 5, and 10 epochs of training.

Berlin - Berlin provided interesting evidence of the baseline's adaptability compared to MAML. With 10 adaptation examples and 10 epochs, MAML only achieved a 2% increase in average F1 score and a 4% improvement in median F1 score over its initial performance before adaptation. On the other hand, the baseline showed increases of 6% and 9%, respectively. However, despite the baseline model's superior adaptability performance, likely due to its low initial performance, MAML's performance remained unbeaten regarding the average F1 score. In contrast, the baseline surpassed MAML in the median F1 score. This suggests that, while MAML demonstrated lower relative improvements on Berlin with a limited number of adaptation examples, its final performance aligned closely with the one of the adapted baseline. Table 7.3 shows the gains and losses in percentage relative to the initial performance reported in Table 7.1.

Run	1 Epoch		5 Epochs		10 Epochs	
	Avg F1	Median F1	Avg F1	Median F1	Avg F1	Median F1
Baseline 1-shot	+1%	+1%	+1%	+1%	+0%	+1%
MAML 1-shot	+0%	+1%	-1%	+1%	+0%	+1%
Baseline 5-shot	+1%	+0%	+2%	+2%	+4%	+6%
MAML 5-shot	+1%	+2%	+0%	+1%	+0%	+1%
Baseline 10-shot	+1%	+3%	+4%	+5%	+6%	+9%
MAML 10-shot	+0%	+3%	+2%	+3%	+2%	+4%

Table 7.3: Performance of MAML compared to the baseline model in terms of adaptation capability to the Berlin dataset. The comparison is made by computing the average percentage gains/losses on the F1 scores of the models after 1, 5, and 10 epochs of training.

Buenos Aires - In the case of Buenos Aires, MAML and the baseline model exhibited very similar adaptation performance. In this scenario, MAML achieved a lower average F1 score by 1% and a 1% improvement in median F1 score compared to its initial performance. The baseline model's median F1 score remained unchanged, while its average F1 score also improved by 1%. Despite these marginal improvements in performance metrics, MAML's absolute performance in Buenos Aires remained significantly higher than that of the baseline model. Table 7.4 presents the percentage gains and losses relative to the respective initial performance reported in Table 7.1.

Run	1 Epoch		5 Epochs		10 Epochs	
	Avg F1	Median F1	Avg F1	Median F1	Avg F1	Median F1
Baseline 1-shot	+0%	+0%	+0%	-1%	+1%	+0%
MAML 1-shot	-2%	+1%	-1%	+1%	-2%	+0%
Baseline 5-shot	-1%	-1%	+0%	-1%	+1%	+0%
MAML 5-shot	-2%	-1%	-3%	-1%	-2%	+1%
Baseline 10-shot	+0%	+0%	-1%	-3%	-4%	-5%
MAML 10-shot	-2%	-1%	-4%	-2%	-2%	-1%

Table 7.4: Performance of MAML compared to the baseline model in terms of adaptation capability to the Buenos Aires dataset. The comparison is made by computing the average percentage gains/losses on the F1 scores of the models after 1, 5, and 10 epochs of training.

In summary, these results indicate that while MAML can adapt better to certain tasks, the baseline model also achieves strong adaptability performance on others. Table 7.5 displays the test metrics of both models after adaptation. In the Vienna dataset, the baseline model outperforms MAML in both average and median F1 scores, suggesting that the baseline is more effective at leveraging knowledge from smaller datasets. In

contrast, for the Buenos Aires dataset, MAML achieved the highest scores, significantly surpassing the baseline. This indicates that MAML can better utilize the knowledge contained in larger datasets and adapt that information to smaller, similar tasks. Finally, for the Berlin dataset, both models exhibit comparable performance after adaptation. The baseline model has a lower average F1 score but a higher median F1 score, suggesting more consistent performance across all samples. Conversely, MAML has a higher average F1 score but a lower median F1 score, indicating it may have slightly more outliers, primarily positive ones, that elevate the average F1 score but may lead to less consistency in predictions across all samples. Overall, the performances of both models differ only slightly, resulting in similar outcomes.

Run	Avg F1	Median F1
Baseline Vienna	0.71	0.83
MAML Vienna	0.65	0.77
Baseline Berlin	0.64	0.81
MAML Berlin	0.66	0.79
Baseline Buenos Aires	0.71	0.88
MAML Buenos Aires	0.75	0.91

Table 7.5: Performance of MAML compared to the baseline model in terms of average and median F1 scores averaged across samples after adaptation for a few epochs on ALL. The dataset names in the rows specify the dataset that was used for testing, hence the remaining two were use for training the corresponding model.

7.2 ALL - Qualitative Analysis

The qualitative analysis in this section examines the behavior and learning characteristics of the MAML model compared to the baseline model on ALL. It utilizes several visual representations of validation loss, validation F1 scores, and predicted MRD values per sample. Through these analyses, we observe how MAML and the baseline model differ in terms of convergence speed, stability, and overall performance consistency on validation data.

Validation Loss Comparison - The validation loss plots (Figures 7.1 and 7.2) highlight a significant difference in stability and convergence between MAML and the baseline model. MAML starts with a higher initial loss and exhibits a more volatile trajectory throughout the training process. This volatility indicates that MAML, while having a downward trend, it struggles with stability issues. In contrast, the baseline model converges more quickly and exhibits a smoother loss trajectory, reflecting a more consistent learning pattern.

Additionally, the final validation loss value for MAML remains higher than that of the baseline model despite the longer training time. This result suggests that MAML is less

capable of generalizing well to unseen validation data, potentially due to meta-optimization challenges or the limited effectiveness of MAML in scenarios with fewer training samples. This behavior is consistent with previous findings regarding the instability of MAML in complex classification tasks, where the model's iterative fine-tuning approach may lead to erratic loss patterns over epochs.

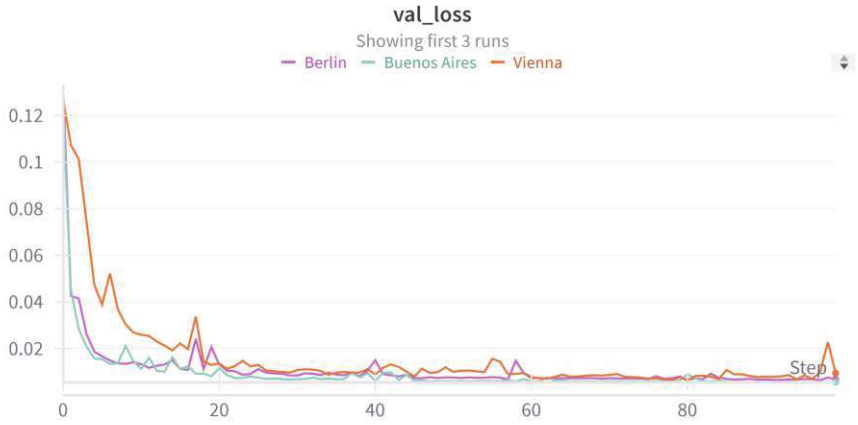

Figure 7.1: Illustration of the validation loss during training of the baseline models on ALL. In the legend, the written datasets are the ones on which the models were tested, training has been done on the remaining datasets.

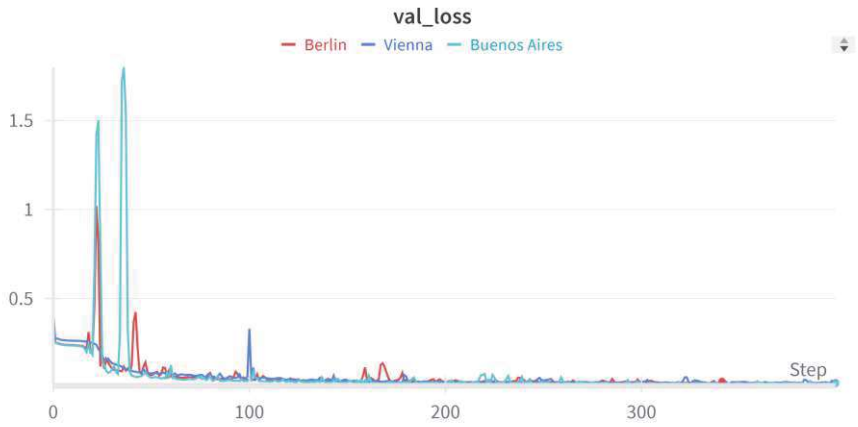

Figure 7.2: Illustration of the validation loss during training of the MAML models on ALL. In the legend, the written datasets are the ones on which the models were tested, training has been done on the remaining datasets.

Validation F1 Score Comparison - In the validation F1 score plots (Figures 7.3 and 7.4), the baseline model demonstrates faster convergence, achieving relatively high validation scores earlier in the training process. In contrast, the MAML model shows

greater fluctuation in F1 scores, indicating an inconsistent ability to classify samples accurately over time. This inconsistency may highlight MAML's sensitivity to the inner loop fine-tuning process, where the model's learned representations can vary significantly depending on the sampled task data for each iteration.

That said, the baseline model converges faster but reaches similar validation F1 scores to MAML, which starts converging about 30 to 80 epochs later but is able to catch up with the baseline.

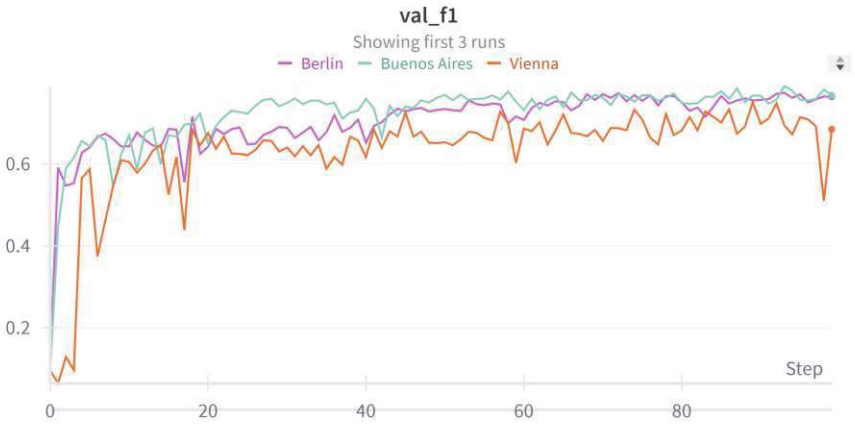

Figure 7.3: Illustration of the validation f1 score during training of the baseline models on ALL. In the legend, the written datasets are the ones on which the models were tested, training has been done on the remaining datasets.

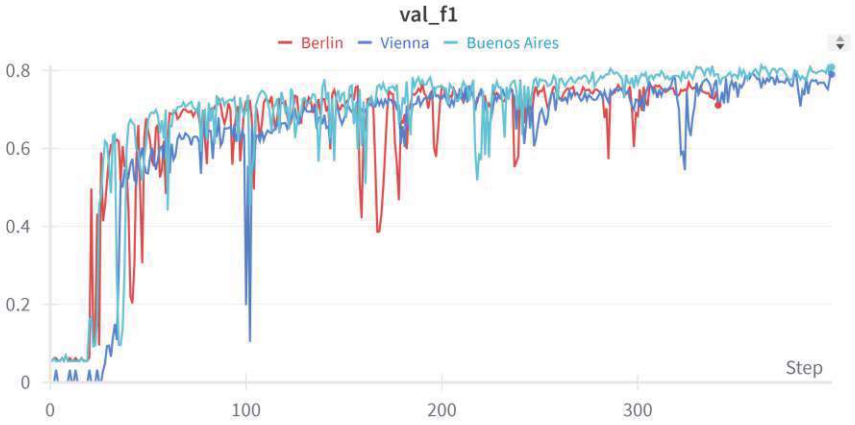

Figure 7.4: Illustration of the validation f1 score during training of the MAML models on ALL. In the legend, the written datasets are the ones on which the models were tested, training has been done on the remaining datasets.

Predicted MRD Values - To analyze and compare the predicted MRD values for the baseline and MAML models, we take the adapted models on each dataset and visualize the predictions vs the ground truth. The MRD plots that follow contain colored dots, each representing the F1 score of a single sample and ranging from red (F1 = 0.0) to green (F1 = 1.0). The dashed lines correspond to MRD values of $5e-4$ which is, as mentioned by Woedlinger et al. [WRW+22], "the lower necessary resolution for patient stratification according to the current international therapy trials of the allied study groups of the iBFM consortium. Predictions that are within the range of either less than 3 times or more than 1/3 of the true MRD (tolerance window) are considered acceptable (correct) predictions." [DGR+08]

- **Vienna** - Figures 7.5 and 7.6 display the output MRD values for the Vienna dataset. The plots confirm the results of the previous quantitative evaluation, suggesting that the baseline model produces better MRD predictions than MAML on Vienna. Both models generate many predictions outside of the tolerance window, resulting in both underestimations and overestimations of the MRD. However, while the amount of MRD overestimation is similar across both models, the baseline model shows significantly fewer MRD underestimations and exhibits an overall tighter distribution around the ground truth values. This observation confirms the baseline model's superior capability in making predictions after being trained on smaller datasets, indicating that it is the most suitable model for this task.

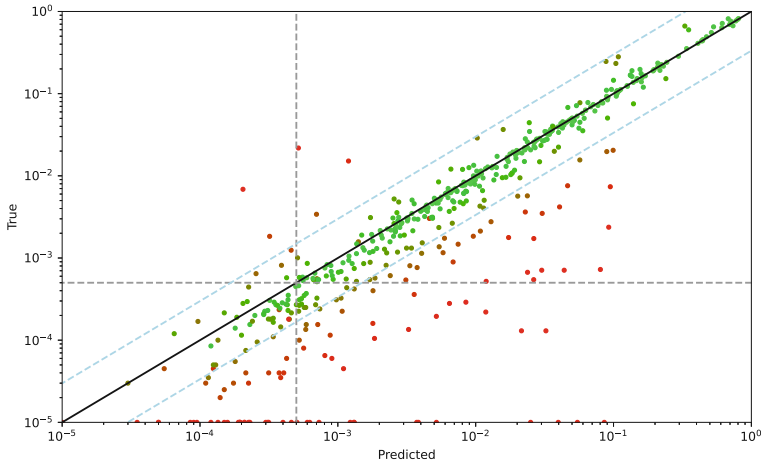

Figure 7.5: Illustration of the output MRD plot generated by the baseline model after adaptation on Vienna.

- **Berlin** - In the figures displaying predicted MRD values per sample (Figures 7.7 and 7.8) for Berlin, we observe that, despite the very similar performances, the MAML model produces predictions that align more closely with the expected MRD

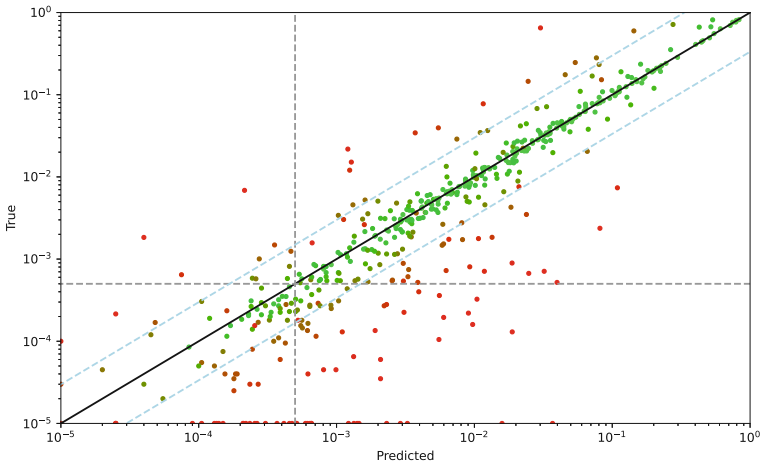

Figure 7.6: Illustration of the output MRD plot generated by MAML after adaptation on Vienna.

distribution. It has fewer outliers and a higher correlation between predictions and true MRD values. Both models exhibit erroneous outputs, all of which tend to overestimate the actual MRD. However, while both models generate false outputs, the predictions made by MAML deviate less from the expected values and fall more consistently within the tolerance window. This result suggests that the baseline model's predictions are slightly more prone to extremes, making it a less reliable choice for practical MRD assessment applications.

- **Buenos Aires** - Figures 7.9 and 7.10 display the output MRD values for the Buenos Aires dataset. The MAML model's outputs exhibit a higher concordance between predictions and ground truth values, which is reflected in its higher measured performance metrics. The baseline predictions include four samples that fall outside the expected tolerance values, featuring both underestimations and overestimations of the MRD. In contrast, the MAML model predicts five samples outside the tolerance window. However, even though MAML has more samples that exceed the tolerance limits, the outliers are less pronounced compared to those predicted by the baseline model. Overall, the distribution predicted by MAML is more consistent and has a higher correlation between predictions and ground truth values. This result confirms that the MAML model produces safer and more reliable MRD scores for this task.

Hyperparameter Sensitivity - The plot examining MAML's sensitivity to hyperparameters (Figure 7.11) reveals that minor adjustments can significantly impact the model's performance and convergence behavior. Specifically, increasing the number of outer loops from 3 to 4 enables the model to begin converging, with validation F1 scores

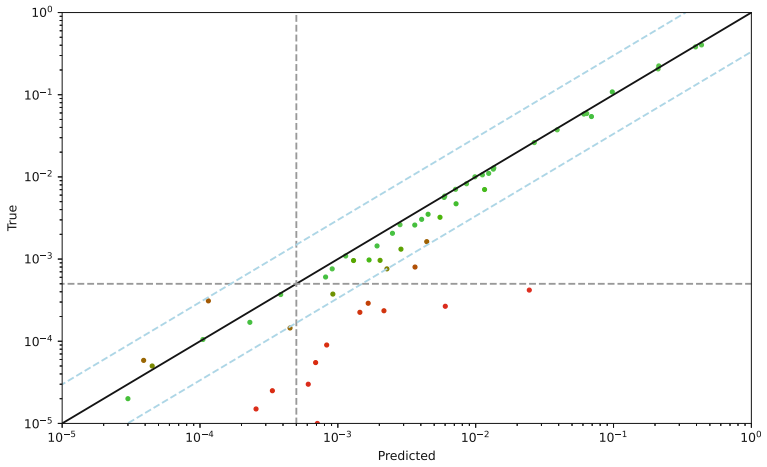

Figure 7.7: Illustration of the output MRD plot generated by the baseline model after adaptation on Berlin.

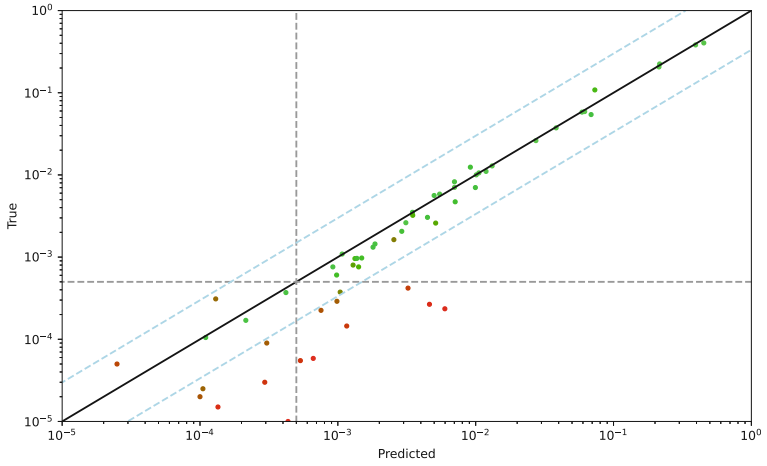

Figure 7.8: Illustration of the output MRD plot generated by MAML after adaptation on Berlin.

improving consistently over epochs. In contrast, when the outer loop is set to 3, the model fails to converge, with validation F1 scores remaining flat at around 10-20%.

This observation highlights MAML's high sensitivity to meta-optimization settings, particularly in scenarios where task complexity requires a delicate balance of inner- and outer-loop iterations. The strong difference in performance resulting from this single hyperparameter change suggests that MAML's success in adapting to unseen tasks

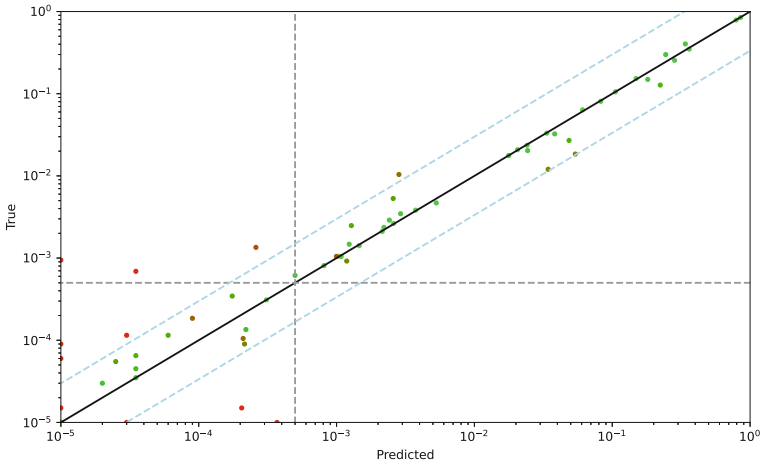

Figure 7.9: Illustration of the output MRD plot generated by the baseline model after adaptation on Buenos Aires.

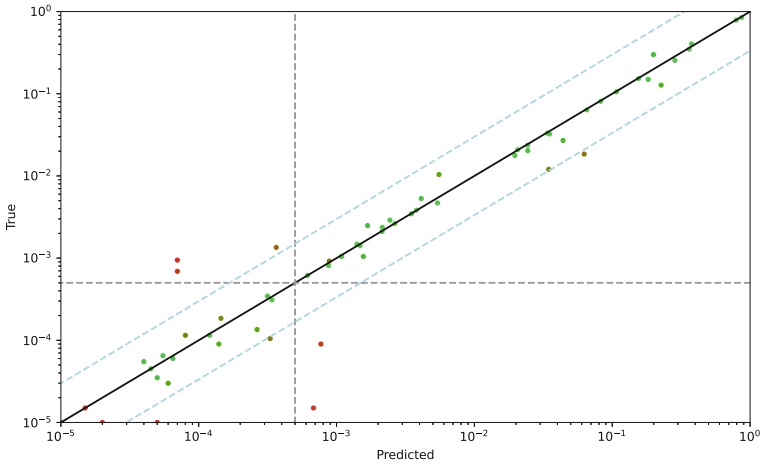

Figure 7.10: Illustration of the output MRD plot generated by MAML after adaptation on Buenos Aires.

relies on finely tuned meta-training configurations. Such sensitivity implies that small deviations from optimal settings can lead to suboptimal or entirely stalled learning, impacting the model's robustness and reliability in practical applications.

Furthermore, this hyperparameter sensitivity may contribute to MAML's volatility in performance metrics, as observed in previous plots of validation loss and F1 scores. Since MAML relies on task-specific fine-tuning, even minor shifts in parameters during the

meta-learning phase can disproportionately affect its capacity to generalize and adapt, especially in complex classification scenarios such as MRD assessment.

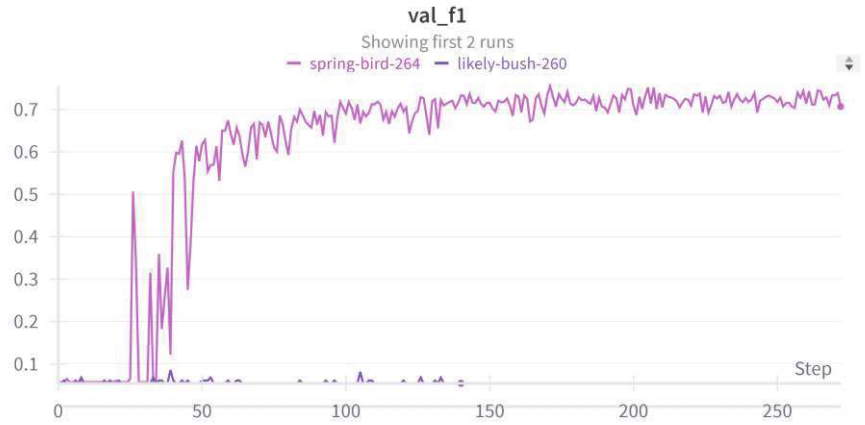

Figure 7.11: Illustration of the hyperparameter sensitivity of the MAML framework. Both the displayed runs have the same hyperparameters besides the number of outer loops. The run that converged has 4 outer loops, while the one that didn't has 3 outer loops. Changing one hyperparameter changed completely the outcome of the training procedure.

Overall, the qualitative analysis underscores the baseline model's advantages in convergence speed and stability while emphasizing MAML's better alignment with expected MRD outcomes. The analysis also highlights the limitations of the MAML model in terms of volatility and instability. These insights align with the quantitative results, suggesting that while MAML holds potential for adaptability and its performance surpasses that of the baseline in certain tasks, its application in blast cell classification and MRD assessment may require additional stabilization techniques, alternative meta-learning strategies, or a broader set of tasks with more data to enhance performance.

7.3 AML - Quantitative Results

In this section, we present the quantitative results of our experiments for AML across multiple metrics. We analyze the efficacy of the MAML architecture, emphasizing the model's capacity to address the unique challenges of AML phenotypic diversity and its implications for MRD assessment and discuss how these findings align with the previous results for ALL.

7.3.1 How does the performance of MAML compare to existing state-of-the-art approaches for AML MRD assessment?

The first set of experiments for AML involved training both models (MAML and the baseline) on datasets structured by phenotypic subtypes of AML. Tasks were divided based on phenotypic characteristics, enabling the evaluation of the models' ability to generalize across distinct phenotypes. In this scenario, MAML demonstrated improved performance compared to the baseline when trained on phenotypes with a larger number of samples (M5, M7, M2) and tested on the phenotype with fewest samples (M4). This finding aligns with MAML's meta-learning approach, which leverages patterns from richer datasets to enhance performance on smaller, distinct tasks.

Conversely, when trained on smaller phenotypes (M2, M4, M7) and tested on the phenotype with the largest sample size, MAML underperformed relative to the baseline model. This outcome suggests that the baseline's conventional training approach is more effective when adapting to phenotypes with higher complexity and exploits better the variability in small datasets.

To validate these results, the MAML training and evaluation process was repeated across four additional random splits of the data. The findings remained consistent across these trials, confirming that MAML excels in leveraging larger phenotypic datasets to generalize effectively to smaller, related tasks but struggles in scenarios where task-specific data diversity or scale increases. These observations highlight the strengths and the current limitations of MAML in addressing the phenotypic heterogeneity of AML. Table 7.6 provides a detailed summary of all the comparative results.

Run	Avg F1	Median F1
Baseline M2	0.93	0.96
MAML M2	0.90	0.95
Baseline M4	0.78	0.90
MAML M4	0.88	0.90
Baseline M5	0.74	0.83
MAML M5	0.60	0.61
Baseline M7	0.71	0.77
MAML M7	0.71	0.84

Table 7.6: Performance of MAML compared to the baseline model in terms of average and median F1 scores on different runs without any adaptation on AML. The dataset names in the rows specify the dataset that was used for testing, hence the remaining ones were use for training the corresponding model.

7.3.2 How efficiently can a model trained using MAML adapt to new, unseen phenotypes in AML?

To investigate how well the MAML and baseline models adapt to new, unseen AML phenotypes, a series of adaptation experiments were conducted. Both models were initially trained on all but one phenotype and then fine-tuned on the left-out phenotype using a learning rate of 0.0005. Adaptation was performed using sets of 1, 5, and 10 examples to assess whether increasing the number of adaptation samples improved model performance. Each experiment was repeated five times for each adaptation setting.

The results, presented in the following subsections, reveal distinct trends in adaptability based on the phenotype and model used. Interestingly, the baseline model exhibited stronger adaptation capabilities, either maintaining its initial advantage or surpassing the MAML. These findings provide a nuanced perspective on the models' strengths and limitations in adapting to the phenotypic heterogeneity of AML. The sections below offer a detailed analysis of the adaptation results for each phenotype.

M2 - For the M2 phenotype, the baseline model was the one performing best without any adaptation, although MAML is very close in performance. After adaptation, the performance of the baseline didn't improve at all, it only got worse in each run that we executed, most likely because of the very high initial performance. On the other hand, after adaptation, MAML's best run achieved an increase in average F1 score of 2% and an increase of 1% in median F1 score, almost reaching the baseline model in terms of performance metrics. So, MAML showed a better adaptation pattern on the M2 phenotype but the baseline model remained the one performing best. Table 7.7 shows the gains/losses in percentage of the two models after adaptation.

Run	1 Epoch		5 Epochs		10 Epochs	
	Avg F1	Median F1	Avg F1	Median F1	Avg F1	Median F1
Baseline 1-shot	+0%	-1%	-5%	-2%	-9%	-9%
MAML 1-shot	+0%	+0%	+0%	+0%	-1%	+0%
Baseline 5-shot	+0%	+0%	-2%	-1%	-2%	-1%
MAML 5-shot	+1%	+1%	-1%	+0%	+2%	+1%
Baseline 10-shot	-1%	+0%	-5%	-3%	-3%	+0%
MAML 10-shot	+1%	+1%	+0%	-2%	+1%	+0%

Table 7.7: Performance of MAML compared to the baseline model in terms of adaptation capability to the M2 phenotype. The comparison is made by computing the average percentage gains/losses on the F1 scores of the models after 1, 5, and 10 epochs of training.

M4 - For the M4 phenotype, MAML outperformed the baseline model prior to adaptation, particularly in terms of the average F1 score, where it showed a significant advantage. However, after adaptation, the baseline model demonstrated stronger improvements, surpassing MAML in both average and median F1 scores, albeit by a narrow margin.

This suggests that while MAML excelled in its initial performance on the M4 phenotype, the baseline model exhibited better adaptability during fine-tuning. Table 7.8 summarizes the gains and losses in performance metrics for both models following adaptation.

Run	1 Epoch		5 Epochs		10 Epochs	
	Avg F1	Median F1	Avg F1	Median F1	Avg F1	Median F1
Baseline 1-shot	-1%	-2%	-2%	-3%	+8%	+1%
MAML 1-shot	+0%	-1%	+1%	+1%	+1%	+1%
Baseline 5-shot	+3%	+1%	+12%	+1%	+9%	+1%
MAML 5-shot	+1%	+0%	+0%	-1%	+2%	+1%
Baseline 10-shot	-3%	-3%	+14%	+3%	+14%	+3%
MAML 10-shot	+0%	-2%	+2%	+0%	+3%	+1%

Table 7.8: Performance of MAML compared to the baseline model in terms of adaptation capability to the M4 phenotype. The comparison is made by computing the average percentage gains/losses on the F1 scores of the models after 1, 5, and 10 epochs of training.

M5 - For the M5 phenotype, MAML initially underperformed significantly compared to the baseline model, particularly in terms of average and median F1 scores. After adaptation, MAML showed improvements in performance; however, it remained consistently inferior to the baseline across all metrics. The baseline model not only maintained its advantage but also demonstrated slight gains in both average and median F1 scores following adaptation. This indicates that, for the M5 phenotype, MAML struggled to adapt effectively, while the baseline model continued to perform robustly. Table 7.9 highlights the percentage gains and losses for both models post-adaptation.

Run	1 Epoch		5 Epochs		10 Epochs	
	Avg F1	Median F1	Avg F1	Median F1	Avg F1	Median F1
Baseline 1-shot	+4%	+7%	+3%	+6%	+0%	+4%
MAML 1-shot	+3%	+12%	+10%	+20%	+4%	+6%
Baseline 5-shot	+1%	+0%	+9%	+10%	+6%	+7%
MAML 5-shot	+4%	+3%	+9%	+9%	+6%	+4%
Baseline 10-shot	+5%	+6%	+4%	+9%	+1%	-7%
MAML 10-shot	+6%	+8%	+9%	+17%	+15%	+22%

Table 7.9: Performance of MAML compared to the baseline model in terms of adaptation capability to the M5 phenotype. The comparison is made by computing the average percentage gains/losses on the F1 scores of the models after 1, 5, and 10 epochs of training.

M7 - For the M7 phenotype, MAML and the baseline model performed similarly regarding the average F1 score before adaptation, yielding nearly identical results. However, MAML

had a notable advantage in the median F1 score, outperforming the baseline by 7%. After adaptation, the baseline model showed significant improvement, surpassing MAML in both average and median F1 scores. This suggests that while MAML initially leveraged its ability to generalize well across tasks, the baseline model's adaptability allowed it to exceed MAML after fine-tuning. Table 7.10 presents the percentage gains and losses in performance metrics for both models following adaptation.

Run	1 Epoch		5 Epochs		10 Epochs	
	Avg F1	Median F1	Avg F1	Median F1	Avg F1	Median F1
Baseline 1-shot	+1%	+5%	-4%	-12%	+11%	+12%
MAML 1-shot	-2%	+0%	-2%	-1%	-8%	-4%
Baseline 5-shot	+1%	+8%	+5%	+7%	+16%	+14%
MAML 5-shot	-8%	-3%	+2%	-2%	+8%	+2%
Baseline 10-shot	+15%	+9%	+15%	+13%	+16%	+15%
MAML 10-shot	+10%	+3%	-8%	-4%	+0%	-3%

Table 7.10: Performance of MAML compared to the baseline in terms of adaptation capability to the M7 phenotype. The comparison is made by computing the average percentage gains/losses on the F1 scores of the models after 1, 5, and 10 epochs of training.

In summary, the adaptation results for the M2, M4, M5, and M7 phenotypes highlight varying performance patterns between MAML and the baseline model. For M2, the baseline model maintained its lead with minimal improvement after adaptation, while MAML demonstrated a better adaptation trend but could not surpass the baseline. In contrast, for M4, MAML initially outperformed the baseline, particularly in the average F1 score, but the baseline model's superior adaptability allowed it to surpass MAML after adaptation. For M5, MAML lagged behind the baseline both before and after adaptation, indicating limited adaptability for this phenotype. Lastly, for M7, MAML showed an initial advantage in median F1 score, but the baseline model overtook it in both average and median F1 scores after adaptation. These results emphasize the nuanced differences in how the two models handle specific phenotypes and adapt to new tasks. Table 7.11 shows the test metrics of the best-performing models for each phenotype after adaptation.

7.4 AML - Qualitative Analysis

The qualitative analysis in this section explores the behavior and learning dynamics of the MAML model compared to the baseline model on AML. This analysis leverages visual representations of validation loss, validation F1 scores, and predicted MRD values for individual samples across various phenotypes. By examining these figures, we assess key aspects such as convergence speed, stability, and consistency in performance across validation data. These qualitative observations provide deeper insights into the

Run	Avg F1	Median F1
Baseline M2	0.93	0.96
MAML M2	0.92	0.96
Baseline M4	0.92	0.93
MAML M4	0.91	0.91
Baseline M5	0.83	0.93
MAML M5	0.75	0.83
Baseline M7	0.87	0.92
MAML M7	0.81	0.87

Table 7.11: Performance of MAML compared to the baseline model in terms of average and median F1 scores after adaptation for a few epochs on AML. The dataset names in the rows specify the dataset that was used for testing, hence the remaining ones were use for training the corresponding model.

models' comparative strengths, weaknesses, and MAML's adaptability in addressing the phenotypic heterogeneity of AML.

Validation Loss Comparison - The validation loss plots (Figures 7.12 and 7.13) reveal notable differences in stability and convergence between MAML and the baseline model for AML. MAML exhibits significant instability, characterized by a highly volatile trajectory throughout training, failing to demonstrate clear patterns of convergence. This instability may be attributed to the inherent challenges of meta-learning, particularly in handling the phenotypic diversity and complexity of AML datasets.

In comparison, the baseline model shows a smoother and more consistent loss trajectory, converging earlier and resulting in a lower validation loss. This suggests that the baseline model is more effective at optimizing its parameters and generalizing to unseen validation data under our experimental setup. Moreover, MAML's final validation loss remains not only higher than that of the baseline but also fluctuates even in later epochs, indicating that it struggles to stabilize and fully adapt to the task.

These results underscore the challenges of applying MAML to AML phenotypes, where the complexity and variance within the data may exacerbate the difficulties of meta-optimization and hinder the model's ability to achieve reliable convergence. This behavior highlights a key limitation of MAML in handling tasks that require high stability and consistent learning patterns.

Validation F1 Score Comparison - In the validation F1 score plots (Figures 7.14 and 7.15), the baseline model demonstrates rapid convergence, achieving relatively high validation scores in the early stages of training before flattening out. This behavior reflects the baseline model's stability and consistent learning pattern, allowing it to maintain reliable performance with minimal fluctuations throughout the training process.

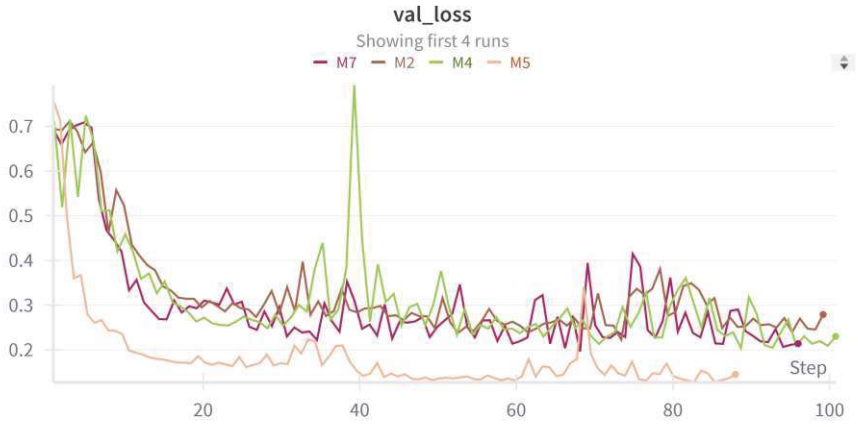

Figure 7.12: Illustration of the validation loss during training of the baseline models on AML. In the legend, the written datasets are the ones on which the models were tested, training has been done on the remaining datasets.

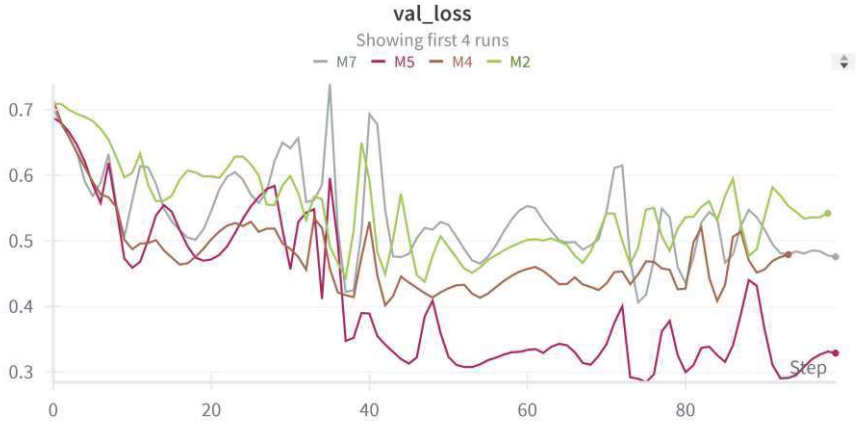

Figure 7.13: Illustration of the validation loss during training of the MAML models on AML. In the legend, the written datasets are the ones on which the models were tested, training has been done on the remaining datasets.

In contrast, the MAML model, while also reaching high validation F1 scores early in training, exhibits significantly greater fluctuations over time. This volatility suggests that MAML struggles with stability, likely due to its sensitivity to the inner loop fine-tuning process, where the sampled task data in each iteration can lead to substantial variation in the model's learned representations. Although MAML achieves scores comparable to the baseline, its fluctuating trajectory indicates less consistency in its ability to generalize effectively to unseen validation data.

These results highlight the baseline model's advantage in achieving stable and predictable

performance, while MAML shows potential but requires more robust optimization to address its inherent instability in AML classification tasks.

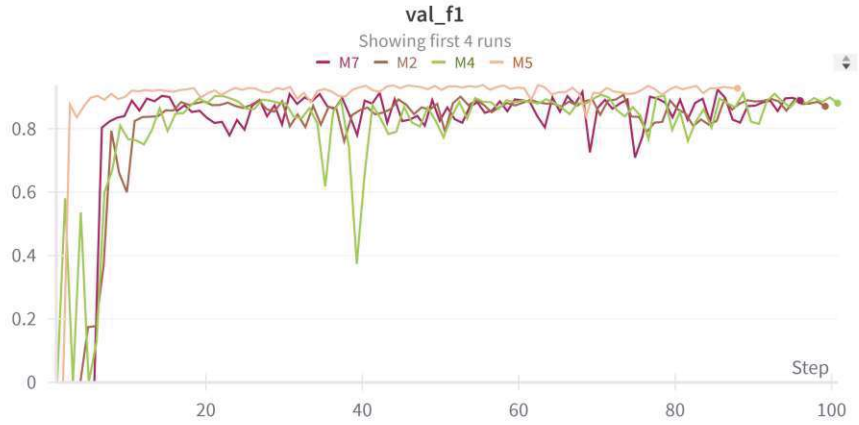

Figure 7.14: Illustration of the validation f1 score during training of the baseline models on AML. In the legend, the written datasets are the ones on which the models were tested, training has been done on the remaining datasets.

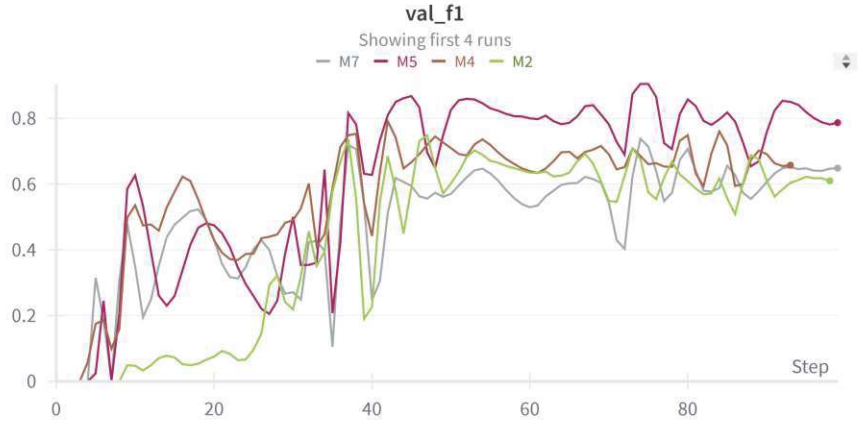

Figure 7.15: Illustration of the validation f1 score during training of the MAML models on AML. In the legend, the written datasets are the ones on which the models were tested, training has been done on the remaining datasets.

Predicted MRD Values - In this section, we present the predicted MRD output plots for the M4 and M5 phenotypes as representative examples. While similar analyses were conducted for all phenotypes, we focus on these two to highlight their contrasting behaviors: M4 represents a phenotype with smaller training datasets, while M5 represents a phenotype trained on larger datasets. These examples offer key insights into the models' performance patterns under varying training conditions.

- **M4** - Figures 7.16 and 7.17 display the predicted MRD values for the M4 phenotype. The results indicate that both models perform reasonably well in predicting MRD values, with most of their predictions falling within the tolerance window. Notably, MAML has a complete absence of outliers, as all its predictions lie within the tolerance range. However, its predictions are slightly less concordant with the expected ground truth compared to the baseline. In contrast, the baseline model produces three outliers outside the tolerance window, all of which are underestimations of the MRD. Despite this, the predictions of the baseline model correlate better with the ground truth MRD values for the M4 phenotype. These results suggest that MAML demonstrates superior reliability in avoiding extreme deviations, while the baseline model retains a slight advantage in overall prediction precision.

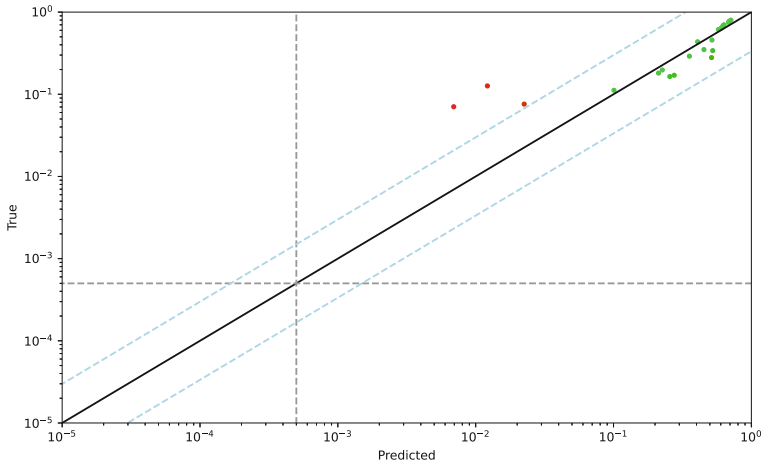

Figure 7.16: Illustration of the output MRD plot generated by the baseline model before adaptation on M4.

- **M5** - Figures 7.18 and 7.19 illustrate the predicted MRD values for the M5 phenotype. In this case, the baseline model demonstrates predictions that correlate better with the ground truth, with only two outliers outside the tolerance window, both of which are overestimations. Most of the baseline model's predictions correlate closely with the ground truth, reflecting its ability to make precise predictions when trained on smaller datasets. In contrast, MAML produces slightly less concordant predictions within the tolerance window and shows a higher number of outliers compared to the baseline. Specifically, the MAML outliers represent underestimations of the MRD, which could significantly impact downstream clinical decisions. These findings highlight the baseline model's superior performance on the M5 phenotype, particularly in its ability to generate consistent predictions and limit extreme deviations.

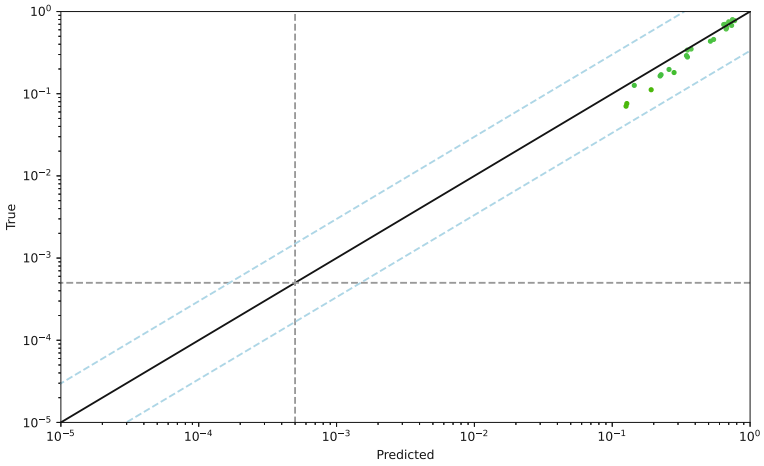

Figure 7.17: Illustration of the output MRD plot generated by MAML before adaptation on M4.

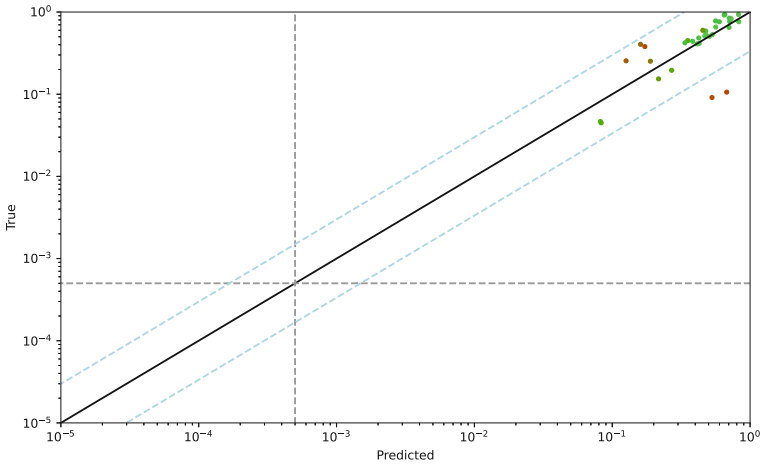

Figure 7.18: Illustration of the output MRD plot generated by the baseline model before adaptation on M5.

The qualitative analysis of AML results highlights important aspects of both MAML and baseline model performance. The validation loss and F1 score plots reveal that while MAML shows potential for capturing complex relationships and generalizing across phenotypic diversity, it struggles with stability and consistency during training, particularly regarding loss fluctuations and convergence patterns. In contrast, the baseline model exhibits more stable and predictable behavior, often converging faster

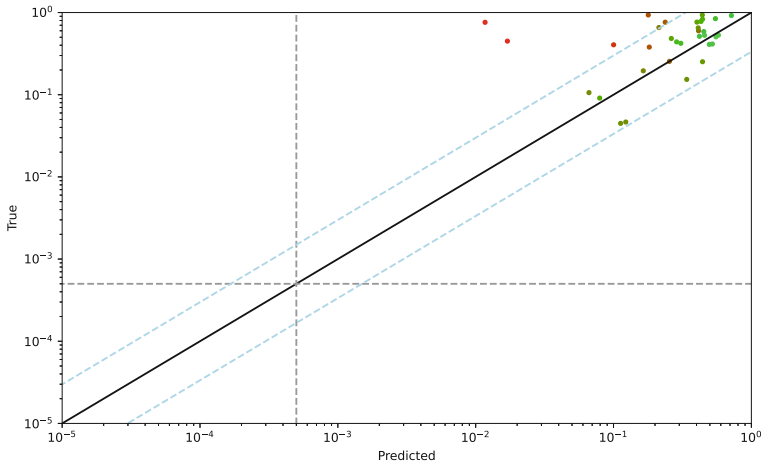

Figure 7.19: Illustration of the output MRD plot generated by MAML before adaptation on M5.

and maintaining tighter validation loss and F1 score trajectories.

The MRD output plots further underscore these findings. For phenotypes trained on larger datasets, MAML demonstrates greater reliability in avoiding extreme outliers, whereas, for phenotypes with smaller datasets, the baseline model produces predictions that correlate better with the ground truth MRD values. This suggests that MAML excels in scenarios requiring adaptability to diverse phenotypic data but faces challenges in achieving precision and stability, especially when trained on smaller, less diverse tasks.

Overall, these observations provide a nuanced understanding of the strengths and limitations of MAML in the context of AML. While it offers promising adaptability for phenotypically diverse tasks, further refinements are necessary to address its instability and enhance its practical applicability in clinical settings.

CHAPTER 8

Discussion

8.1 Interpretation of Results

The experimental outcomes from both the baseline and MAML models reveal several critical insights regarding the effectiveness of meta-learning in blast cell classification and MRD assessment. Key observations include MAML's slower convergence and increased volatility in validation metrics, such as F1 scores and validation loss, compared to the baseline. It requires more epochs to converge, with its validation loss consistently remaining slightly higher than that of the baseline. This indicates that MAML's optimization dynamics introduce additional complexity, leading to slower convergence and less stable training.

In ALL experiments where the tasks are datasets from three different clinics, despite its instability, MAML outperforms the baseline in cross-dataset generalization when trained on combinations that include Vienna, the largest dataset in ALL. Specifically, MAML performs better when trained on Vienna-Buenos Aires and tested on Berlin, as well as when trained on Vienna-Berlin and tested on Buenos Aires. Similarly, when trained on AML, MAML performs better than the baseline when it is trained on the phenotypes with the largest amount of samples. This suggests that MAML excels at leveraging the diversity and richness of larger datasets during meta-training to adapt effectively to new, unseen tasks. This aligns with MAML's theoretical strength in capturing meta-knowledge that facilitates rapid adaptation.

Conversely, the baseline model performs better in scenarios involving smaller datasets. In ALL, the baseline outperforms MAML when trained on Berlin-Buenos Aires and tested on Vienna. As for AML, MAML performs worse than the baseline when the largest M5 and M7 phenotypes are left out. This indicates that the baseline model's conventional training approach is more effective at learning task-specific features in limited-data scenarios, as it focuses directly on minimizing the loss for the given training data rather than optimizing for task adaptability.

The analysis of adaptation performance emphasizes that MAML can improve its performance on specific tasks by only seeing a few examples. However, the baseline model shows similar adaptability patterns. This suggests that while MAML's meta-learning structure is designed for fast adaptation, it may not provide the expected benefits when trained on smaller datasets for ALL and AML MRD assessment.

8.2 Challenges and Limitation

Several challenges emerged during the research, most of which relate to the inherent complexity of MAML and its high resource requirements:

- **High Hyperparameter Sensitivity** - MAML's sensitivity to hyperparameters posed significant challenges, as minor adjustments could lead to dramatic changes in performance. For instance, increasing the number of outer loops from 3 to 4 significantly improved the model's convergence. However, when set to 3, the validation F1 score remained flat, barely exceeding 10-20%. This instability in hyperparameters underscores the difficulty of achieving consistent results without meticulous tuning, a process that can be both time-consuming and resource-intensive. A solution to finding optimal hyperparameters can be a more extensive search using tuning algorithms, such as grid or random search, to train several runs of MAML and select the configuration that performs best.
- **Extended Training Times** - MAML's higher-order backpropagation requires considerably longer training times. In this study, training the model with higher-order MAML took approximately three days, while the FOMAML approach required only 4 to 5 hours. This significant difference in training duration underscores the high computational demands of MAML, which may limit its scalability and applicability in real-world clinical settings where rapid model updates are crucial.
- **Training Instability** - The training process exhibited signs of instability, particularly when MAML was not used with MSL regularization. Comparisons of validation loss between MSL and non-MSL runs show a significant advantage for using MSL regularization. Regularized models demonstrate higher stability and often achieve higher validation F1 scores compared to the non-regularized MAML runs.
- **Increased Complexity** - MAML's framework introduces a plethora of hyperparameters that require tuning, increasing the complexity of implementation and optimization. Key hyperparameters include the meta-learning rate, task-specific learning rate, parameters for learning rate schedulers for both the task-specific and meta-models, the number of inner and outer loops, support/query split parameters, and the number of examples per loop. Additionally, given the unbalanced dataset in ALL, where Vienna has 519 examples while Berlin and Buenos Aires each have between 60 and 70, decisions had to be made on whether to balance all tasks or

leverage Vienna's additional data. This decision involves further tuning, such as determining whether to take twice, thrice, or six times as many examples from Vienna as from Berlin and Buenos Aires.

- **Slower Convergence** - Compared to the baseline model, which exhibited rapid convergence, MAML required an average of 100 epochs to achieve stable validation metrics. This need for extended training underscores MAML's inefficiency in quickly reaching optimal parameters, a significant limitation in scenarios where training time is critical.

8.3 Potential Improvements and Future Work

Considering the challenges identified, several potential improvements could enhance MAML's performance and reliability in this application domain:

- **Optimized Hyperparameter Search** - To address MAML's sensitivity to hyperparameters, a more robust approach can be using automated tuning frameworks, such as Optuna or Ray Tune, to systematically explore combinations of inner and outer loop parameters, support/query splits, and learning rates.
- **Experiment with FOMAML** - The long training times associated with higher-order MAML could potentially be mitigated by experimenting more extensively with FOMAML, which simplifies the gradient computation in the meta-optimization step. Although FOMAML's performance may differ slightly from that of higher-order MAML, its substantially lower computational cost makes it a viable alternative in resource-limited scenarios.
- **Data Balancing Techniques** - To address the imbalance between Vienna and the other tasks, advanced data augmentation or synthetic data generation methods might be employed. Alternatively, resampling techniques could help balance the training data without reducing the available examples from Vienna, thus preserving valuable information while minimizing bias.
- **Alternative Meta-Learning Algorithms** - Given the limitations observed with MAML, exploring alternative meta-learning methods could provide insights into whether different approaches might achieve better performance. These methods may offer a more stable training experience with fewer hyperparameter dependencies, potentially improving adaptability and convergence times.
- **Experiments with Additional Diseases** - Another promising avenue is to test MAML's performance across distinct diseases, adding to AML and ALL and treating each disease type as a unique task. This could be achieved by using the Feature-Agnostic Transformer-based Encoder (FATE) architecture proposed by Weijler et al. [WKR+24] to transform features from different diseases into a common feature space. The current study's use of three ALL datasets as separate

tasks may not provide MAML with sufficient task diversity to generalize effectively across diseases. By incorporating data from multiple diseases, MAML may better learn generalizable patterns and leverage variations in cell morphology and other relevant features, thereby improving its adaptability to novel, unseen medical tasks.

- **Explainability in Medical Diagnostics** - Enhancing the interpretability of MAML's outputs could increase its utility in clinical settings. Techniques such as attention visualization or SHAP values could provide insights into the decision-making process, fostering trust among medical practitioners.
- **Integration into Clinical Workflows** - Future efforts should focus on developing user-friendly systems that integrate the proposed model into clinical workflows for MRD assessment. Testing these systems in real-world environments could provide valuable feedback for further refinement.

Future work could significantly enhance the adaptability and reliability of MAML and other meta-learning approaches for medical diagnostics. More exhaustive hyperparameter tuning could mitigate the sensitivity and instability observed during training, while computationally efficient variants like FOMAML may address the long training times, enabling faster experimentation and deployment. Addressing dataset imbalance through advanced augmentation or resampling can ensure more equitable learning across tasks, which is particularly critical in MRD assessment involving diverse patient populations. Exploring alternative meta-learning algorithms and increasing task diversity through cross-disease generalization could reveal more robust approaches to learning transferable representations, thereby improving MAML's ability to adapt to unseen tasks and medical contexts. Together, these advancements would not only enhance the technical performance of MRD assessments but also increase their clinical relevance, ultimately contributing to more accurate and reliable diagnostic tools for healthcare applications.

CHAPTER 9

Conclusion

This thesis explored the application of MAML, a prominent meta-learning framework, to enhance blast cell classification in ALL and AML using a few-shot learning approach. The study combined the advantages of MAML with Transformer architectures to address the challenges of learning from limited data in ALL and AML FCM datasets.

The thesis began with a review of the state of the art in automated FCM analysis, meta-learning, and transformer-based models, emphasizing their relevance to medical diagnostics and MRD assessment. The background section provided foundational knowledge on MAML, Transformers, and the intricacies of FCM data. A detailed account of dataset preparation and structure preceded the presentation of the methodology, which highlighted the design and implementation of the MAML-based and baseline models. Finally, the experiments and results sections analyzed the comparative performance of the models.

This work made the following key contributions:

- **Application of MAML to Blast Cell Classification** - This study represents an attempt to integrate MAML with Transformer-based architectures in the context of FCM data.
- **Performance Analysis Across Dataset Scenarios** - Conducted a comprehensive evaluation of MAML and baseline models under varying training and testing configurations, emphasizing the relationship between dataset size and model performance.
- **Insights into Meta-Learning** - Provided evidence of MAML's adaptability advantages, particularly in cross-dataset generalization, while highlighting its computational trade-offs.

This thesis demonstrated that MAML, despite its instability and slower convergence, can offer significant advantages in learning from diverse datasets. Its ability to generalize across tasks makes it a promising candidate for real-world applications, particularly in resource-constrained medical settings. However, its trade-offs, including higher validation loss and computational complexity, must be carefully considered and subjected to future research.

By comparing MAML with a baseline model, this work highlights specific strengths, such as MAML's ability to leverage larger datasets for improved generalization to smaller, unseen tasks, as well as limitations, including instability during training and sensitivity to hyperparameters. These findings lay the groundwork for future research aimed at addressing these challenges, such as developing more robust meta-optimization techniques or exploring alternative meta-learning algorithms. Furthermore, the study advances the application of ML in medical diagnostics by providing insights into how adaptable models can manage phenotypic diversity in AML and cross-clinic diversity in ALL, ultimately supporting clinicians in improving patient outcomes through more precise and efficient MRD assessments.

Overview of Generative AI Tools Used

In the process of writing this thesis, generative AI tools were employed to enhance workflow and improve the quality of the text. ChatGPT 4o was utilized to streamline the text, guide the overall structure, and clarify technical concepts, ensuring a logical progression throughout the chapters. Additionally, EditGPT was used to refine the text by addressing grammar, clarity, and style issues. These tools significantly contributed to maintaining consistency, reducing redundancy, and saving time during the drafting and editing phases. While the intellectual content and research remain the author's own, the AI tools provided valuable support in organizing and articulating ideas effectively. Here is a list of example prompts that were used for this thesis:

- How can I structure my thesis on meta-learning for blast cell classification using transformers?
- Can you provide a meaningful list of sections and subsections for my thesis?
- What should I include in the background section of my thesis?
- How would you describe the dataset's internal structure?
- Can you improve this text on AML and MRD assessment?
- What kind of figures can I include in the results section to explain MAML's instability?

List of Figures

List of Tables

Glossary

anemia A medical condition characterized by a deficiency of red blood cells or hemoglobin in the blood, resulting in reduced oxygen transport to tissues and organs. 7, 8

antibody A protein the immune system produces in response to foreign substances, such as bacteria, viruses, or toxins. Antibodies specifically bind to these antigens to neutralize or mark them for destruction by other immune cells. They play a crucial role in the body's defense against infections and diseases. 8, 10, 12, 13, 22–24

antigen A substance, typically a protein or polysaccharide, recognized by the immune system as foreign or harmful. Antigens trigger an immune response, prompting the production of antibodies that bind to them, helping the body defend against infections or diseases. 10, 24

blast cell Immature precursor cells found in the bone marrow. In the context of ALL, these cells multiply uncontrollably, leading to disease progression. 1–3, 5, 7, 12, 15, 24, 25, 27, 30, 45, 47, 48, 58, 69, 73

bone marrow The spongy tissue inside bones that produces blood cells, including red blood cells, white blood cells, and platelets. Bone marrow is a key focus in leukemia diagnosis and treatment. 1, 2, 4, 7, 8, 21

fine-tuning The process of updating a pre-trained model's weights using a small dataset from a new domain to adapt it to a specific task. 2, 28, 30, 32, 33, 43, 45, 46, 52, 53, 57, 61, 64

flow cytometer An instrument used in FCM that analyzes cells individually as they pass through a laser beam. It measures the scattered light and fluorescence emitted from cells to determine their properties. 10–12, 21, 22

hematology The branch of medicine that studies blood, blood-forming organs, and blood diseases. Hematology focuses on conditions such as anemia, clotting disorders, leukemia, lymphoma, and other blood-related disorders, as well as the management of blood transfusions and bone marrow transplants. 10

immunology The branch of biology and medicine that focuses on the immune system, its structure, function, and disorders. Immunology studies how the body defends itself against pathogens (such as bacteria, viruses, and fungi) and foreign substances, as well as the mechanisms behind autoimmune diseases, allergies, and immunodeficiencies. 10

inductive bias Prior knowledge or assumptions consistent with the process that generated the training data incorporated into a learning model to guide its predictions and generalizations. 28

inner loop The process within meta-learning where the model adapts to a specific task using its current parameters and task-specific data. 29, 30, 32–35, 40, 41, 53, 64

meta-learning A machine learning paradigm where models learn to learn. It focuses on enabling algorithms to adapt quickly to new tasks with limited data by leveraging knowledge from previously seen tasks. 2–5, 7, 15, 18, 19, 25, 27, 28, 30, 34, 40, 41, 43–47, 49, 58, 63, 69–74

meta-model A generalized model trained using meta-learning techniques, which serves as the foundation for creating task-specific models. 29–35, 40, 44, 70

meta-objective The objective function in meta-learning that measures how well the meta-model generalizes across tasks after task-specific adaptations. 35

oncology The branch of medicine that focuses on diagnosing, treating, and preventing cancer. Oncologists study the causes, progression, and treatment options for various types of cancer, including surgery, chemotherapy, radiation, and emerging therapies like immunotherapy. 10

outer loop In meta-learning, the process that updates the meta-model's parameters using feedback from multiple tasks. 29, 30, 33–35, 41, 55, 56, 58, 70

phenotype A phenotype refers to the observable characteristics or traits of an organism, such as its physical appearance, biochemical properties, or behavior. 9, 21, 28–30, 41, 44, 59–63, 65, 66, 68, 69

query set Set of examples used to evaluate the task-specific model after it has been fine-tuned on the support set. 29, 30, 33, 41

stem cell A type of undifferentiated cell with the unique ability to develop into different types of specialized cells in the body. Stem cells can divide and renew themselves and are essential for growth, tissue repair, and regeneration. 9, 10

support set Small set of labeled examples used to fine-tune a task-specific model during the inner-loop optimization of the meta-learning process. 29, 30, 33, 41

Acronyms

ALL Acute Lymphoblastic Leukemia. 1–4, 7–13, 15, 17, 18, 21–24, 26, 27, 29, 30, 35, 41, 44, 47, 48, 51, 58, 69–71, 73, 74

AML Acute Myeloid Leukemia. 1–4, 7–13, 15, 17, 21, 23, 24, 26, 27, 29, 30, 35, 41, 44, 58–60, 62, 63, 65, 67–71, 73, 74

BCE Binary Cross-Entropy. 29, 40

CCRI Children's Cancer Research Institute. 21

CFU Colony Forming Unit. 23

CNNs Convolutional Neural Networks. 16

CUDA Compute Unified Device Architecture. 43

FATE Feature-Agnostic Transformer-based Encoder. 71

FCM Flow Cytometry. 2–4, 7, 9–13, 15–18, 22, 23, 25, 27, 28, 34, 38–41, 73

FOMAML First-Order MAML. 31–33, 70–72

FSC Forward Scatter. 11, 12

GDPR General Data Protection Regulation. 25

GMMs Gaussian Mixture Models. 1, 16–18

GPU Graphics Processing Unit. 43

HIPAA Health Insurance Portability and Accountability Act. 25

LAIP Leukemia Associated Immunophenotype. 23

LIME Local Interpretable Model-Agnostic Explanations. 26

LSTM Long Short-Term Memory. 17

MAML Model-Agnostic Meta-Learning. 2–5, 7, 15, 18, 19, 27–36, 38–40, 43–74

ML Machine Learning. 1, 2, 7, 10, 16, 18, 24, 25, 27, 74

MRD Minimal Residual Disease. 1–5, 7–11, 13, 15–18, 23, 24, 44, 45, 47, 48, 51, 54–58, 62, 65–70, 72–74

MSL Multi-Step Loss. 34–38, 70

NLP Natural Language Processing. 1, 17

PCR Polymerase Chain Reaction. 9

SGD Stochastic Gradient Descent. 19

SHAP Shapley Additive Explanations. 26, 72

SSC Side Scatter. 11, 12

WHO World Health Organization. 9

Bibliography

[AES18] Antreas Antoniou, Harrison Edwards, and Amos Storkey. How to train your maml. In *International conference on learning representations*, 2018.

[AFC+13] Nima Aghaeepour, Greg Finak, FlowCAP Consortium, Dream Consortium, Holger Hoos, Tim R Mosmann, Ryan Brinkman, Raphael Gottardo, and Richard H Scheuermann. Critical assessment of automated flow cytometry data analysis techniques. *Nature methods*, 10(3):228–238, 2013.

[AKH+20] Marlon Arnone, Martina Konantz, Pauline Hanns, Anna M Paczulla Stanger, Sarah Bertels, Parimala Sonika Godavarthy, Maximilian Christopeit, and Claudia Lengerke. Acute myeloid leukemia stem cells: the challenges of phenotypic heterogeneity. *Cancers*, 12(12):3742, 2020.

[AMD+20] Sébastien M R Arnold, Praateek Mahajan, Debajyoti Datta, Ian Bunner, and Konstantinos Saitas Zarkias. learn2learn: A library for Meta-Learning research. August 2020.

[Ba16] Jimmy Lei Ba. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[BF22] Jonathan R Brestoff and John L Frater. Contemporary challenges in clinical flow cytometry: small samples, big data, little time. *The journal of applied laboratory medicine*, 7(4):931–944, 2022.

[Bio] Creative Biolabs. Flow cytometry.

[Cam10] Dario Campana. Minimal residual disease in acute lymphoblastic leukemia. *Hematology 2010, the American Society of Hematology Education Program Book*, 2010(1):7–12, 2010.

[CLD+20] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020.

[CSMO+09] Elaine Coustan-Smith, Charles G Mullighan, Mihaela Onciu, Frederick G Behm, Susana C Raimondi, Deqing Pei, Cheng Cheng, Xiaoping Su, Jeffrey E Rubnitz, Giuseppe Basso, et al. Early t-cell precursor leukaemia: a subtype of very high-risk acute lymphoblastic leukaemia. *The lancet oncology*, 10(2):147–156, 2009.

[CSSH+00] Elaine Coustan-Smith, Jose Sancho, Michael L Hancock, James M Boyett, Frederick G Behm, Susana C Raimondi, John T Sandlund, Gaston K Rivera, Jeffrey E Rubnitz, Raul C Ribeiro, et al. Clinical importance of minimal residual disease in childhood acute lymphoblastic leukemia. *Blood, The Journal of the American Society of Hematology*, 96(8):2691–2696, 2000.

[DAYR14] Murat Dundar, Ferit Akova, Halid Z Yerebakan, and Bartek Rajwa. A non-parametric bayesian model for joint cell clustering and cluster matching: identification of anomalous sample phenotypes with random effects. *BMC bioinformatics*, 15:1–15, 2014.

[DFP+02] Michael N Dworzak, Gertraud Froschl, Dieter Printz, Georg Mann, Ulrike Potschger, Nora Muhlegger, Gerhard Fritsch, and Helmut Gadner. Prognostic significance and modalities of flow cytometric minimal residual disease detection in childhood acute lymphoblastic leukemia. *Blood, The Journal of the American Society of Hematology*, 99(6):1952–1958, 2002.

[DGR+08] Michael Norbert Dworzak, Giuseppe Gaipa, Richard Ratei, Marinella Veltroni, Angela Schumich, Oscar Maglia, Leonid Karawajew, Allessandra Benetello, Ulrike Pötschger, Zvenyslava Husak, et al. Standardization of flow cytometric minimal residual disease evaluation in acute lymphoblastic leukemia: Multicentric assessment is feasible. *Cytometry Part B: Clinical Cytometry: The Journal of the International Society for Analytical Cytology*, 74(6):331–340, 2008.

[DRW+14] Marco L Davila, Isabelle Riviere, Xiuyan Wang, Shirley Bartido, Jae Park, Kevin Curran, Stephen S Chung, Jolanta Stefanski, Oriana Borquez-Ojeda, Malgorzata Olszewska, et al. Efficacy and toxicity management of 19-28z car t cell therapy in b cell acute lymphoblastic leukemia. *Science translational medicine*, 6(224):224ra25–224ra25, 2014.

[DWB15] Hartmut Döhner, Daniel J Weisdorf, and Clara D Bloomfield. Acute myeloid leukemia. *New England Journal of Medicine*, 373(12):1136–1152, 2015.

[EKB+17] Philipp Eulenberg, Niklas Köhler, Thomas Blasi, Andrew Filby, Anne E Carpenter, Paul Rees, Fabian J Theis, and F Alexander Wolf. Reconstructing cell cycle and disease progression using deep learning. *Nature communications*, 8(1):463, 2017.

[FAL17] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[IAB+21] Muhammad Shahid Iqbal, Iftikhar Ahmad, Luo Bin, Suleman Khan, and Joel JPC Rodrigues. Deep learning recognition of diseased and normal cell representation. *Transactions on Emerging Telecommunications Technologies*, 32(7):e4017, 2021.

[JWF16] Kerstin Johnsson, Jonas Wallin, and Magnus Fontes. Bayesflow: latent modeling of flow cytometry cell populations. *BMC bioinformatics*, 17:1–16, 2016.

[Kin14] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[KKL20] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*, 2020.

[KVPF20] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pages 5156–5165. PMLR, 2020.

[KWK+23] Florian Kowarsch, Lisa Weijler, FLorian Kleber, Matthias Wödlinger, Michael Reiter, Margarita Maurer-Granofszky, and Michael Dworzak. Explainable techniques for analyzing flow cytometry cell transformers. *arXiv preprint arXiv:2307.14581*, 2023.

[LCD+18] Yuqian Li, Bruno Cornelis, Alexandra Dusa, Geert Vanmeerbeeck, Dries Vercruysse, Erik Sohn, Kamil Blaszkiewicz, Dimiter Prodanov, Peter Schelkens, and Liesbet Lagae. Accurate label-free 3-part leukocyte recognition with single cell lens-free imaging flow cytometry. *Computers in biology and medicine*, 96:147–156, 2018.

[LDB99] Bob Lowenberg, James R Downing, and Alan Burnett. Acute myeloid leukemia. *New England Journal of Medicine*, 341(14):1051–1062, 1999.

[LH16] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[LLK+19] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*, pages 3744–3753. PMLR, 2019.

[LMM16] Sheng Li, Christopher E Mason, and Ari Melnick. Genetic and epigenetic heterogeneity in acute myeloid leukemia. *Current opinion in genetics & development*, 36:100–106, 2016.

[LSGT11] Brenden Lake, Ruslan Salakhutdinov, Jason Gross, and Joshua Tenenbaum. One shot learning of simple visual concepts. In *Proceedings of the annual meeting of the cognitive science society*, volume 33, 2011.

[LSR+18] Roxane Licandro, Thomas Schlegl, Michael Reiter, Markus Diem, Michael Dworzak, Angela Schumich, Georg Langs, and Martin Kampel. Wgan latent space embeddings for blast identification in childhood acute myeloid leukaemia. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 3868–3873. IEEE, 2018.

[LSS+17] Huamin Li, Uri Shaham, Kelly P Stanton, Yi Yao, Ruth R Montgomery, and Yuval Kluger. Gating mass cytometry data by deep learning. *Bioinformatics*, 33(21):3423–3430, 2017.

[Lun17] Scott Lundberg. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.

[LZCL17] Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. Meta-sgd: Learning to learn quickly for few-shot learning. *arXiv preprint arXiv:1707.09835*, 2017.

[NDBS21] Noga Nissim, Matan Dudaie, Itay Barnea, and Natan T Shaked. Real-time stain-free classification of cancer cells and blood cells using interferometric phase microscopy and machine learning. *Cytometry Part A*, 99(5):511–523, 2021.

[NDR+14] Iftekhar Naim, Suprakash Datta, Jonathan Rebhahn, James S Cavenaugh, Tim R Mosmann, and Gaurav Sharma. Swift—scalable clustering for automated identification of rare cell populations in large, high-dimensional flow cytometry datasets, part 1: Algorithm design. *Cytometry Part A*, 85(5):408–421, 2014.

[Nic18] A Nichol. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018.

[Onc09] Mihaela Onciu. Acute lymphoblastic leukemia. *Hematology/oncology clinics of North America*, 23(4):655–674, 2009.

[PGM+19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative

style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[PHW+09] Saumyadipta Pyne, Xinli Hu, Kui Wang, Elizabeth Rossin, Tsung-I Lin, Lisa M Maier, Clare Baecher-Allan, Geoffrey J McLachlan, Pablo Tamayo, David A Hafler, et al. Automated high-dimensional flow cytometric data analysis. *Proceedings of the National Academy of Sciences*, 106(21):8519–8524, 2009.

[PO19] Eunbyung Park and Junier B Oliva. Meta-curvature. *Advances in neural information processing systems*, 32, 2019.

[PRL08] Ching-Hon Pui, Leslie L Robison, and A Thomas Look. Acute lymphoblastic leukaemia. *The Lancet*, 371(9617):1030–1043, 2008.

[RDS+19] Michael Reiter, Markus Diem, Angela Schumich, Margarita Maurer-Granofszky, Leonid Karawajew, Jorge G Rossi, Richard Ratei, Stefanie Groeneveld-Krentz, Elisa O Sajaroff, Susanne Suhendra, et al. Automated flow cytometric mrd assessment in childhood acute b-lymphoblastic leukemia using supervised machine learning. *Cytometry Part A*, 95(9):966–975, 2019.

[RFKL19] Aravind Rajeswaran, Chelsea Finn, Sham M Kakade, and Sergey Levine. Meta-learning with implicit gradients. *Advances in neural information processing systems*, 32, 2019.

[RHL+20] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, et al. The future of digital health with federated learning. *NPJ digital medicine*, 3(1):1–7, 2020.

[RL16] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *International conference on learning representations*, 2016.

[RNO+21] Giovanni Riva, Vincenzo Nasillo, Anna Maria Ottomano, Giuliano Bergonzini, Ambra Paolini, Fabio Forghieri, Beatrice Lusenti, Patrizia Barozzi, Ivana Lagreca, Stefania Fiorcari, et al. Multiparametric flow cytometry for mrd monitoring in hematologic malignancies: clinical applications and new challenges. *Cancers*, 13(18):4582, 2021.

[RRK+16] Michael Reiter, Paolo Rota, Florian Kleber, Markus Diem, Stefanie Groeneveld-Krentz, and Michael Dworzak. Clustering of cell populations in flow cytometry data using a combination of gaussian mixtures. *Pattern Recognition*, 60:1029–1040, 2016.

[RSG16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[SLR+19] Jakob Scheithe, Roxane Licandro, Paolo Rota, Michael Reiter, Markus Diem, and Martin Kampel. Monitoring acute lymphoblastic leukemia therapy with stacked denoising autoencoders. In *Computer Aided Intervention and Diagnostics in Clinical and Medical Images*, pages 189–197. Springer, 2019.

[vDvdVBO15] Jacques JM van Dongen, Vincent HJ van der Velden, Monika Brüggemann, and Alberto Orfao. Minimal residual disease diagnostics in acute lymphoblastic leukemia: need for sensitive, fast, and standardized technologies. *Blood, The Journal of the American Society of Hematology*, 125(26):3996–4009, 2015.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[WKR+24] Lisa Weijler, Florian Kowarsch, Michael Reiter, Pedro Hermosilla, Margarita Maurer-Granofszky, and Michael Dworzak. Fate: Feature-agnostic transformer-based encoder for learning generalized embedding spaces in flow cytometry data. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7956–7964, 2024.

[WLK+20] Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020.

[WRW+22] Matthias Wödlinger, Michael Reiter, Lisa Weijler, Margarita Maurer-Granofszky, Angela Schumich, Elisa O Sajaroff, Stefanie Groeneveld-Krentz, Jorge G Rossi, Leonid Karawajew, Richard Ratei, et al. Automated identification of cell populations in flow cytometry data with transformers. *Computers in Biology and Medicine*, 144:105314, 2022.

[ZMH+20] Max Zhao, Nanditha Mallesh, Alexander Höllein, Richard Schabath, Claudia Haferlach, Torsten Haferlach, Franz Elsner, Hannes Lüling, Peter Krawitz, and Wolfgang Kern. Hematologist-level classification of mature b-cell neoplasm using deep learning on multiparameter flow cytometry data. *Cytometry Part A*, 97(10):1073–1080, 2020.