



POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Advancements in Active Learning: Strategies for Imbalanced Class Settings

TESI DI LAUREA MAGISTRALE IN
MANAGEMENT ENGINEERING
INGEGNERIA GESTIONALE

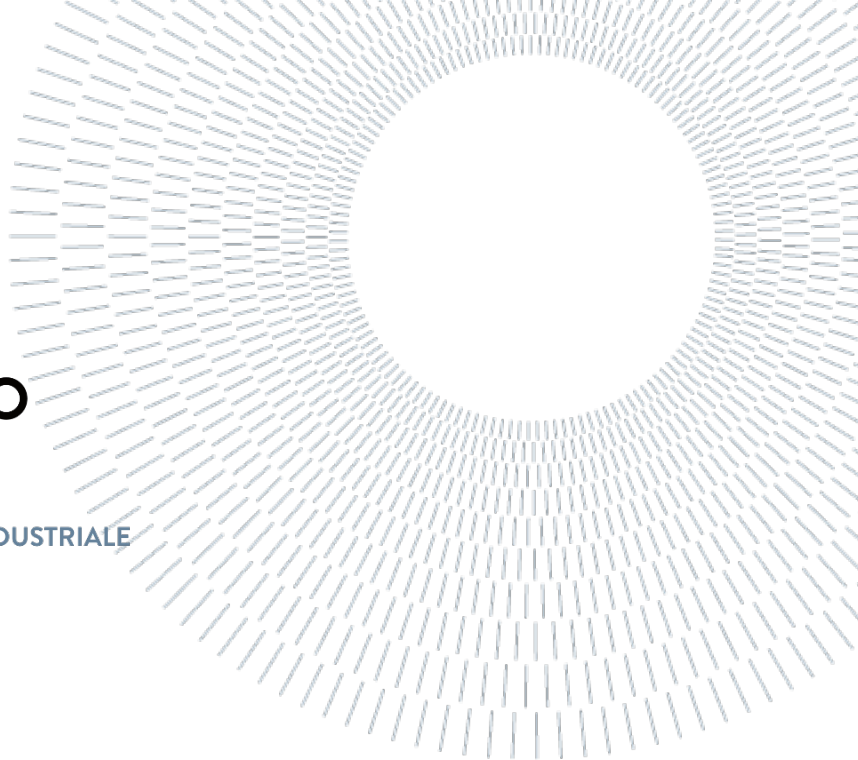
Author: **Francesco Bolognesi**

Student ID: 10661235
Advisor: Roberto Cigolini
Co-advisor: Hadis Anahideh
Academic Year: 2022-23



POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE



1 Table of Contents

Abstract	iv
Abstract in italiano	v
Introduction	1
2 State of the Art	2
2.1. Explanation of Supervised Learning and Its Reliance on Labeled Data.....	2
2.2. Introduction to the Challenges Associated with Obtaining Labeled Data	3
2.3. Introduction to Active Learning as a potential solution.....	3
2.4. Explanation of How Active Learning Works	4
2.5. Discussion on How Active Learning Can Minimize Labeling Efforts and Maximize Learning Efficiency.....	5
2.6. Central Objective: Proposing a Novel Active Learning Technique.....	6
2.7. Utilization of the Initial Percentage of Class Imbalance in the Dataset	6
2.8. Main Goal and Side Benefits of the New Technique.....	7
2.9. Brief Outline of the Methodology.....	8
2.10. Structure of the Research Work.....	9
3 Literature Review	11
3.1. Overview of Machine Learning Techniques.....	11
3.2. Understanding Active Learning.....	12
3.2.1. Introduction to Active Learning.....	12
3.2.2. Active Learning in Industry 4.0.....	13
3.2.3. Main Sampling Techniques in Active Learning	13
3.2.4. The Role of Supportive Models in Active Learning.....	14
3.2.5. Class Imbalance in Active Learning.....	15
3.2.6. Fairness in Machine Learning	15
3.2.7. Active Learning with Limited and Fixed Budget	16
3.2.8. Future Directions and Challenges in Active Learning	16
4 Methodology and Development of Novel Approaches	18
4.1. Overview.....	18
4.2. Preliminaries	18
4.2.1. Baseline Model Choice	18
4.2.2. Active Learning Framework.....	19
4.2.3. Random Sampling.....	21
4.2.4. Uncertainty Sampling Heuristic	21
4.3. The AL Rank Approach: Development and Implementation	22
4.3.1. AL Rank Methodology	22
4.3.2. AL Rank Algorithm	23

4.3.3.	Model Evaluation and Selection Unit (MESU)	24
4.4.	The AL Hybrid Approach: Development and Implementation	24
4.4.1.	Overview of AL Hybrid Approach	24
4.4.2.	AL Hybrid Model	25
4.4.3.	Hybrid Instance Selector	26
4.4.4.	Rank and Entropy Generator	27
4.4.5.	AL Hybrid Reverse Model	27
4.4.6.	Advantages of AL Hybrid Approach	28
5	Experiment Design	30
5.1.	Overview	30
5.2.	Data Collection and Description	30
5.2.1.	Source of the Data	30
5.2.2.	Data Description	31
5.3.	Data Preprocessing	32
5.3.1.	Objectives of Preprocessing	32
5.3.2.	Normalization	32
5.3.3.	Sampling Strategy	32
5.3.4.	Class Balance Adjustment	32
5.3.5.	Handling Missing Values and Outliers	32
5.4.	Experimental Setup	33
5.4.1.	Labeling Strategy	33
5.4.2.	Modeling Approach	33
5.4.3.	Data Selection via Active Learning	34
5.4.4.	Performance Metrics	35
5.4.5.	Experimental Procedures	36
5.4.6.	Results Analysis	36
6	Results and Analysis	38
6.1.	Overview	38
6.2.	Active Learning Results Across Datasets (Adult, COMPAS, ELS)	38
6.2.1.	SVM as the Auxiliary Method	39
6.2.2.	Random Forest as the Auxiliary Method	39
6.2.3.	Logistic Regression as the Auxiliary Method	40
6.3.	Overall Aggregated Insights	41
6.4.	Inclusion of Query By Committee (QBC) for Adult Dataset	42
6.5.	Fairness Metrics Analysis	43
6.5.1.	Accuracy Disparity Analysis	43
6.5.2.	Equalizing Odds Analysis	44
7	Conclusions and Future Work	46
7.1.	Conclusions	46
7.1.1.	Review of Objectives and Research Questions	46
7.1.2.	Major Findings	46
7.1.3.	Impact and Implications	47
7.2.	Limitations of the Study	47
7.3.	Managerial Implications and Future Research Paths	48
7.4.	Future Work and Potential Improvements	49

7.5. Final Thoughts	49
8 Bibliography	51
<i>List of figures</i>	<i>54</i>
<i>List of tables</i>	<i>54</i>
<i>List of equations.....</i>	<i>55</i>

Abstract

Active learning (AL) is a machine learning technique that selects the most informative samples from a large pool of unlabeled data for annotation, thus reducing the labeling cost and improving the learning performance. However, conventional AL approaches often neglect the intricate issue of class imbalance, where certain classes are either overrepresented or underrepresented in the dataset distribution. This disparity can introduce bias in sampling and compromise the overall generalization ability of the classifier. In this work, we introduce a novel threshold-based strategy for AL designed to navigate the challenges of class imbalance. This strategy dynamically adjusts to the degree of class imbalance, ensuring the selection of samples that are both informative and well-representative of minority classes. Our approach is rigorously tested on a variety of imbalanced datasets and benchmarked against state-of-the-art AL methods. Empirical results demonstrate that our proposed method significantly enhances classifier performance, especially in scenarios characterized by imbalanced class labels.

Key-words: Active learning, Informative samples, Class imbalance, Threshold-based strategy, Dynamically adjusts, Minority classes, Imbalanced datasets, Classifier performance.

Abstract in italiano

L'Active Learning (AL) è una tecnica di machine learning che seleziona i campioni più informativi da un grande insieme di dati non etichettati per l'annotazione, riducendo così il costo dell'etichettatura e migliorando le prestazioni dell'apprendimento. Tuttavia, gli approcci convenzionali all'AL spesso trascurano il delicato problema dello squilibrio di classe, dove certe classi sono sovrarappresentate o sottorappresentate nella distribuzione del *dataset*. Questa disparità può introdurre un *bias* nella selezione dei campioni e compromettere la capacità di generalizzazione del classificatore. In questo lavoro, introduciamo una nuova strategia basata su una soglia per l'AL progettata per affrontare le sfide dello squilibrio di classe. Questa strategia si adatta dinamicamente al grado di squilibrio di classe, garantendo la selezione di campioni che sono sia informativi che ben rappresentativi delle classi minoritarie. Il nostro approccio è rigorosamente testato su una varietà di set di dati sbilanciati e paragonato ai metodi AL più avanzati. I risultati empirici dimostrano che il nostro metodo proposto migliora significativamente le prestazioni del classificatore, specialmente in scenari caratterizzati da etichette di classe sbilanciate.

Parole chiave: Active learning, Campioni informativi, Squilibrio di classe, Strategia basata su soglia, Adattamento dinamico, Classi minoritarie, Dataset squilibrati, Prestazioni del classificatore

Introduction

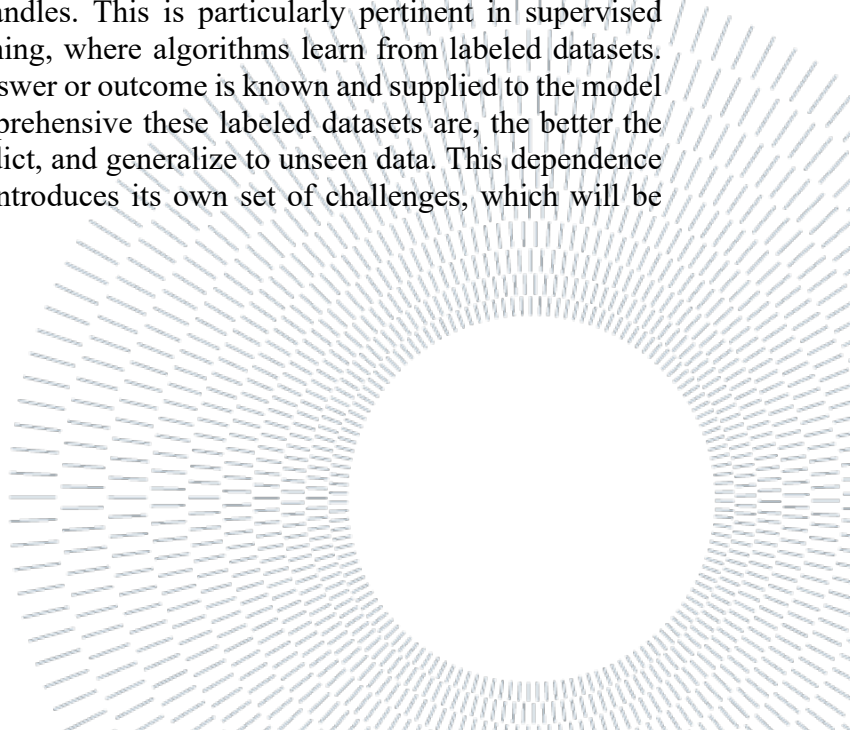
Machine learning, an interdisciplinary domain that combines facets of computer science, statistics, and mathematics, has witnessed a rapid surge in prominence and relevance in recent years. It now stands as a pivotal force behind the ongoing fourth industrial revolution, commonly termed as Industry 4.0. Machine learning's strength stems from its capability to navigate immense datasets, extract relevant trends and understandings, and streamline decision-making. This proficiency has grown notably essential within the realm of Industry 4.0, marked by its digital transformation, automation, and prolific data creation [13].

Within the framework of Industry 4.0, machine learning plays a crucial role in crafting and deploying intelligent autonomous systems, driven by data and computational intelligence [14]. It forms the backbone of advanced manufacturing technologies, including predictive maintenance and optimization of production processes. Machine learning algorithms are utilized to anticipate equipment failures, schedule timely maintenance, and thereby reduce downtime. They also enable more efficient resource allocation and supply chain management by predicting demands and identifying bottlenecks in real-time.

In addition to manufacturing, machine learning has a substantial impact on the logistics sector. It is leveraged to optimize routes, predict accurate delivery times, manage inventories, and streamline the overall supply chain management, all of which are critical aspects of Industry 4.0.

Machine learning also underpins the advancements in cyber-physical systems, another crucial component of Industry 4.0. Such systems, combining computational, networking, and physical elements, utilize machine learning for activities like sensor data interpretation, environmental awareness, and decision-making. Moreover, machine learning plays a crucial role in advancing industrial Internet of Things (IoT) solutions, aiding in the interaction of various devices and systems, and empowering the smart examination of the data they produce [15].

Yet, despite machine learning's pivotal role in Industry 4.0, its effectiveness hinges on the accessibility and caliber of the data it handles. This is particularly pertinent in supervised learning, a major branch of machine learning, where algorithms learn from labeled datasets. These are datasets in which the 'correct' answer or outcome is known and supplied to the model during training. The larger and more comprehensive these labeled datasets are, the better the machine learning algorithm can learn, predict, and generalize to unseen data. This dependence on substantial quantities of labeled data introduces its own set of challenges, which will be explored in subsequent sections.



2 State of the Art

2.1. Explanation of Supervised Learning and Its Reliance on Labeled Data

Central to machine learning is the idea of gleaning insights from data, where supervised learning stands out as a dominant method. Supervised learning encompasses a set of algorithms designed to forecast results using paired input-output data, commonly referred to as labeled data. These pairs serve as a teacher or supervisor for the model, hence the term “supervised learning”. In essence, these algorithms learn the relationship or mapping function from the input data (features) to the desired output (target or label).

The process begins with the training phase, where the supervised learning algorithm ingests a large dataset comprised of instances, each with associated labels. For example, in a spam detection task, the instances could be emails, and the labels would be binary indicators denoting whether each email is ‘spam’ or ‘not spam.’ The algorithm analyzes this dataset, learning the underlying patterns and structures that distinguish spam emails from non-spam emails.

Once the model has been trained, it can be used to predict the labels of new, unseen instances, a process known as inference. In our spam detection example, the trained model could take a new email as input and predict whether it’s spam or not based on what it has learned from the training data.

The success and performance of supervised learning heavily rely on the quality and quantity of the available labeled data [1]. The larger the training dataset, the better the model can learn the underlying patterns and generalize to unseen data. A model trained on a vast and diverse set of labeled data is less likely to overfit to the training data (a situation where the model learns the training data too well and performs poorly on unseen data) and is more likely to make accurate predictions on new data.

However, acquiring a large amount of labeled data poses a significant challenge. It often requires considerable time, effort, and resources, as it may involve manual labeling by domain experts. Moreover, for certain data categories, such as medical evaluations or confidential data, securing labels can be problematic due to concerns over privacy, regulatory limitations, or simply the limited availability of specialists capable of offering precise labels. These issues highlight the importance of adopting more effective learning strategies, like active learning, which will be explored in the subsequent section [29].

2.2. Introduction to the Challenges Associated with Obtaining Labeled Data

The creation of labeled datasets, while critical to supervised learning, is an arduous and resource-intensive process [1]. As the performance of supervised learning models hinges on the availability of large amounts of high-quality labeled data, the challenges associated with obtaining such data present a significant hurdle in machine learning [2]. One of the primary difficulties lies in the manual nature of data labeling. Each instance in a dataset must be inspected and labeled by a human annotator, often a domain expert [3]. This requirement poses multiple challenges.

First, the process can be time-consuming, particularly for large datasets. The sheer volume of data generated in the age of Industry 4.0, where every device or sensor could potentially contribute data, compounds this problem [11].

Second, there are considerable monetary costs associated with data labeling. Skilled annotators or domain experts may command high wages for their services, and the cost of labeling can quickly escalate for large datasets or datasets requiring specialized knowledge [21].

Third, obtaining labeled data in certain fields, such as healthcare or other sensitive areas, presents unique challenges. Data privacy and confidentiality regulations can restrict access to data or limit the information available for labeling [35].

Finally, there is the challenge of class imbalance, a common problem in machine learning datasets [43]. This imbalance can bias the learning process, leading the model to overfit to the majority class and perform poorly on the minority class. These challenges emphasize the need for strategies that can reduce the dependency on large volumes of labeled data. The active learning paradigm, discussed in the next section, presents one such strategy, aiming to minimize labeling effort while maximizing model performance.

2.3. Introduction to Active Learning as a potential solution

In response to the challenges associated with obtaining large volumes of labeled data, active learning emerges as a potential solution [7]. Active learning is a specialized form of machine learning where the learning algorithm actively selects the most informative samples from a pool of unlabeled data to be labeled for training [14]. In other words, the learner (or the model) participates in the data collection process by identifying the instances that it believes would be most beneficial to learn from. This strategy is in contrast to traditional supervised learning, where all training data are passively received, and every instance is considered equally valuable for learning.

Active learning is based on the hypothesis that a machine learning model can achieve comparable performance to traditional supervised learning while using significantly fewer labeled instances, as long as the right instances are selected for labeling [23]. The goal is to select those instances that reduce the model's uncertainty or increase its learning the most. These instances could be those that the model is most uncertain about or those that are most representative or diverse compared to the current training set.

This approach can be particularly beneficial in scenarios where unlabeled data are abundant, but labels are scarce or expensive to obtain [31]. By intelligently selecting the most informative instances for labeling, active learning can mitigate the need for a large labeled dataset, reducing both the time and cost associated with the data labeling process.

Additionally, active learning can address the issue of class imbalance by prioritizing the labeling of underrepresented classes, thereby improving the model's performance on these classes [38]. However, despite these advantages, active learning is not without its own challenges. This involves identifying the optimal approach for choosing informative examples, addressing the possible onset of bias during data selection, and balancing between exploration (picking varied samples) and exploitation (selecting instances with uncertainty). We will delve deeper into these challenges and potential solutions in the subsequent section.

2.4. Explanation of How Active Learning Works

At its core, active learning is a process that strategically selects the most informative instances from a pool of unlabeled data for labeling and inclusion in the training dataset. This process, known as querying, is predicated on the assumption that a learner can improve its performance more effectively by choosing the data from which it learns.

The 'oracle' in active learning refers to the source of labels for the selected instances. In practical terms, this could be a human expert who labels the data based on their domain expertise, or it could be a sophisticated system designed to generate accurate labels. The goal of the active learning system is to minimize the number of queries made to the oracle, thereby reducing the time, effort, and resources spent on labeling.

The selection of the most informative instances hinges on a variety of strategies. Here are some of the most commonly used:

1. **Uncertainty Sampling:** This strategy selects instances about which the current model is most uncertain. This could mean, for example, choosing instances that the model classifies with a confidence score near 0.5 (for a binary classification problem). The idea is that by labeling these uncertain instances, the model can learn more about the boundary between classes and improve its performance [16].
2. **Query-By-Committee (QBC):** In QBC, multiple models (the 'committee') are trained, and they vote on the labels of the unlabeled instances. The instances with the highest disagreement amongst the committee are deemed the most informative and are selected for labeling [9].
3. **Expected Model Change:** This strategy selects instances that, when labeled and added to the training set, are expected to result in the largest change to the current model. The premise here is that instances causing significant model change will yield the most learning [23].
4. **Expected Error Reduction:** This strategy selects instances that are expected to most reduce the model's generalization error when labeled and added to the training set.
5. **Variance Reduction:** This strategy aims to select instances that would most reduce the model's uncertainty (variance) about its predictions across the entire unlabeled dataset.

These strategies, while effective, are not without their challenges. For example, they may not consider the diversity of instances, leading to a potential bias towards certain types of instances.

Alternatively, some strategies may be computationally intensive, reducing the overall efficiency of the active learning process.

Addressing these challenges requires a careful balance between exploration and exploitation. Exploration involves selecting diverse instances to ensure the model learns from a broad range of data, while exploitation involves selecting uncertain instances that the model is currently struggling with. Striking this balance is crucial to the success of active learning and forms the basis of the work presented in this thesis.

2.5. Discussion on How Active Learning Can Minimize Labeling Efforts and Maximize Learning Efficiency

Active learning's unique approach to data annotation offers an efficient pathway towards high-performance machine learning models. This efficiency arises from active learning's dual ability to minimize labeling efforts while maximizing learning, thus accelerating the learning process and reducing the associated costs.

Minimizing Labeling Efforts

The traditional approach to supervised learning often involves labeling vast quantities of data, a process that can be both time-consuming and costly. Active learning addresses this challenge by intelligently selecting the most informative instances for labeling. This targeted approach reduces the number of instances that need to be annotated, thereby lowering the overall labeling effort [40].

In active learning, the model has the ability to ask for labels only for those instances where it expects to learn the most. This could be instances that the model is uncertain about, instances that are highly representative of the overall data distribution, or instances that are particularly diverse. The selected instances are then presented to the oracle for labeling.

By focusing on the most informative instances, active learning can achieve a similar level of performance to traditional supervised learning using significantly fewer labeled instances. This can be particularly beneficial in scenarios where labeled data are difficult, expensive, or time-consuming to obtain.

Maximizing Learning Efficiency

By carefully selecting the instances to label, active learning also seeks to maximize learning efficiency. Each instance that is added to the training set is chosen to provide maximum benefit to the model, whether that's reducing uncertainty, enhancing diversity, or better defining the decision boundary between classes.

This emphasis on efficient learning can lead to faster convergence of the model, meaning it can reach its optimal performance with fewer training instances. Quicker model convergence not only diminishes computational expenses but also leads to a more agile model. This agility is crucial in ever-evolving environments where rapid adaptation to shifts is essential.

It's worth noting, however, that the efficiency of active learning is heavily dependent on the strategy used to select the instances for labeling. Different strategies may be more effective in different scenarios, and part of the ongoing research in active learning involves developing and refining these strategies to ensure they provide the best possible trade-off between minimizing labeling effort and maximizing learning efficiency.

2.6. Central Objective: Proposing a Novel Active Learning Technique

The primary objective of this thesis is to propose a novel active learning technique that differentiates itself by utilizing the initial class imbalance percentage to guide the selection of instances for labelling [29].

In many traditional active learning strategies, the selection of instances for annotation is often guided by principles such as uncertainty, representativeness, or diversity. Nevertheless, these strategies do not commonly account for the class imbalance present in the dataset at the outset of the learning process. This oversight can lead to challenges as class imbalance may affect the performance of the learning algorithm, especially when it's designed with an assumption of balanced class distribution. When the class distribution is imbalanced, these models can end up being biased towards the majority class, leading to sub-optimal performance [30].

Our proposed active learning technique attempts to tackle this challenge by factoring in the initial class imbalance percentage when choosing the most informative instances for annotation.

Instead of explicitly favoring instances from the minority or majority class, our approach is to use the initial class imbalance information to guide instance selection in a more nuanced way. This novel technique aims to more effectively utilize the information available in the dataset, thereby promoting better learning efficiency and faster convergence to a desirable level of accuracy.

This technique is designed to be model-agnostic, meaning it can be paired with various machine learning algorithms, such as logistic regression, support vector machines, or random forests. This increases the potential for wide applicability across numerous domains and problem scenarios.

In summary, the central objective of this thesis is to develop and evaluate an active learning technique that integrates the initial class imbalance percentage, aims to enhance the balance between labeling efforts and learning efficiency, and seeks to accelerate the accuracy convergence process. This objective aligns with the broader goal of this thesis, which is to advance active learning methodologies and contribute to the growing body of knowledge in this field.

2.7. Utilization of the Initial Percentage of Class Imbalance in the Dataset

A distinguishing characteristic of this thesis lies in the explicit utilization of the beta value, representing the percentage of instances labeled as class 0 in the original dataset, to guide the instance selection process in active learning. The term 'beta' signifies the initial percentage of class imbalance in the dataset, capturing the unequal distribution of classes at the beginning of the learning process.

Class imbalance is a crucial concern in many machine learning scenarios, where a majority class dominates the dataset, leading to potential biases and sub-optimal model performance.

Conventional active learning methodologies often overlook the influence of class imbalance when selecting instances for labeling, focusing on criteria such as uncertainty, representativeness, or diversity.

In contrast, this thesis proposes a novel active learning approach that incorporates the beta value into the instance selection procedure. We assume that the supportive machine learning models, despite their lack of perfect accuracy in the initial iterations, can rank unlabeled instances based on the certainty of their class membership. By ordering the instances from most likely to be in class 0 to most likely to be in class 1, we can identify a threshold at the beta percentile of this ranking. This threshold serves as a proxy for the level of uncertainty, with the instance at this threshold considered the most uncertain and selected for labeling.

By leveraging the beta value, our proposed method aims to improve the efficiency and effectiveness of active learning, especially in the early stages of the learning process. This unique approach addresses the challenges posed by class imbalance and uncertainty estimation, facilitating more accurate and accelerated model convergence.

While this research offers significant insights into the relationship between the beta value and active learning, it does not aim to comprehensively investigate all possible methods for handling class imbalance. Additionally, while the proposed technique will be tested with various machine learning models, an exhaustive exploration of all possible models is beyond the scope of this thesis.

In summary, this thesis centers around the development and evaluation of an innovative active learning methodology that incorporates the beta value to guide instance selection. By integrating the initial percentage of class imbalance, we aim to enhance the efficiency and accuracy of active learning methods. This research contributes to the broader understanding of the impact of class imbalance on active learning and provides valuable insights for future investigations in this field.

2.8. Main Goal and Side Benefits of the New Technique

The primary goal of this thesis is to improve the accuracy convergence process in active learning, with the aim of achieving a desirable level of accuracy more efficiently. In addition to this overarching objective, our research also addresses the challenges associated with class imbalance and showcases the positive features of being model-agnostic.

Improving Accuracy Convergence

The central objective of this research is to develop an active learning technique that expedites the convergence to a desirable level of accuracy. Traditional active learning methods often require a significant number of labeled instances to achieve satisfactory performance, which can be time-consuming and resource-intensive. By proposing a novel active learning approach that leverages the beta value and incorporates it into the instance selection process, we aim to accelerate the accuracy convergence process.

Our technique focuses on selecting the most informative and uncertain instances for labeling, which are expected to have a substantial impact on improving the model's performance. By

strategically labeling these instances, we enhance the learning efficiency, reducing the number of labeled instances required to reach the desired level of accuracy. This approach contributes to a more efficient and effective active learning process.

Tackling Class Imbalance

In addition to improving accuracy convergence, our research also addresses the challenge of class imbalance in active learning. Class imbalance occurs when one class is significantly more dominant than the others, potentially leading to biased model predictions and inadequate representation of the minority class. Our proposed active learning technique takes into account the initial class imbalance percentage by selecting instances from both the majority and minority classes. This helps mitigate the negative effects of class imbalance, enhancing the fairness and accuracy of the model's predictions across all classes [31].

Model-Agnosticity

Another positive feature of our proposed active learning technique is its model-agnostic nature. This means that it can be applied with various machine learning models, regardless of the specific algorithm or framework used. The model-agnostic aspect increases the versatility and applicability of our technique across different domains and use cases. Researchers and practitioners can leverage our approach with popular machine learning models, such as logistic regression, support vector machines, random forests, or any other models that suit their specific needs. The model-agnostic nature allows for seamless integration of our technique into existing machine learning workflows, making it more accessible and adaptable [26].

In summary, the primary objective of this thesis is to improve the accuracy convergence process in active learning. Additionally, our research addresses the challenges of class imbalance and highlights the positive features of being model-agnostic. By leveraging the beta value and incorporating it into the instance selection process, we expedite accuracy convergence, tackle class imbalance, and contribute to the advancement of active learning methodologies. The model-agnostic nature of our technique ensures its versatility and practical applicability, benefiting various domains where accurate predictions, efficient learning, and handling of class imbalance are crucial.

2.9. Brief Outline of the Methodology

To ensure a robust evaluation and demonstration of the proposed active learning technique, we employ a systematic methodology that encompasses the utilization of diverse datasets, testing different machine learning models, and exploring a variety of beta values.

Utilization of Diverse Datasets

We use a variety of datasets to test our proposed technique. The datasets selected for this study span multiple domains and exhibit varying degrees of complexity, feature dimensions, and class imbalances. By using diverse datasets, we aim to demonstrate the applicability and robustness of our approach across a wide range of data contexts. The datasets also serve as a platform for simulating real-world scenarios, thereby providing a practical relevance to our research.

Testing Different Machine Learning Models

In our evaluation, we incorporate several popular machine learning models. By testing our active learning technique with different models, such as logistic regression, support vector machines, and random forests, we aim to establish its model-agnostic nature. This helps to ensure the broad applicability of our technique across different machine learning frameworks and use cases.

Exploring a Variety of Beta Values

Finally, we explore the impact of varying beta values on the performance of our active learning technique. Beta, in our research context, represents the initial class imbalance percentage in the dataset. By experimenting with different beta values, we investigate how the class imbalance influences the selection of instances for labeling and the overall learning efficiency. This part of the study helps to understand the interaction between class imbalance and active learning, contributing valuable insights to this field.

In summary, the methodology used in this research is designed to thoroughly evaluate our proposed active learning technique. By using diverse datasets, testing with various machine learning models, and experimenting with different beta values, we aim to demonstrate the effectiveness, robustness, and wide applicability of our approach in active learning. We believe that this methodology will allow us to make substantial contributions to the body of knowledge in active learning, particularly in tackling class imbalance, improving learning efficiency, and enhancing model performance.

2.10. Structure of the Research Work

This thesis is organized into several sections that systematically present our research work. The outline is as follows:

Chapter 2: Literature Review

In Chapter 2, we present an in-depth literature review of active learning techniques. The discussion revolves around existing methodologies, their strengths and limitations, and how they have informed the development of our proposed technique. We also delve into the issue of class imbalance in machine learning and its effects on model performance.

Chapter 3: Methodology

In Chapter 3, we introduce our proposed active learning technique, which utilizes the initial class imbalance percentage (beta value) to guide instance selection. We provide a comprehensive explanation of the technique, detailing the process of instance selection and how the beta value influences this process. Additionally, we discuss the rationale behind the model-agnostic nature of our technique.

Chapter 4: Experiment Design

Chapter 4 details the experimental design used to evaluate our proposed technique. We outline the selection of diverse datasets and the different machine learning models utilized. We also describe the process of manipulating the beta value to observe its effects on learning efficiency and model performance. The chapter concludes with a discussion on performance measures and evaluation criteria used to assess the effectiveness of our technique.

Chapter 5: Results and Discussion

In Chapter 5, we present the results of our experiments. This includes a detailed analysis of how our proposed technique performs across different datasets, machine learning models, and beta

values. We compare the performance of our technique with traditional active learning methods and discuss the implications of our findings.

Chapter 6: Conclusions and Future Work

Finally, Chapter 6 provides a summary of our research findings and their significance in the field of active learning. We reflect on the implications of our work for handling class imbalance in active learning, improving learning efficiency, and enhancing model performance. This final chapter also outlines potential avenues for future research, building upon the foundation laid by this thesis.

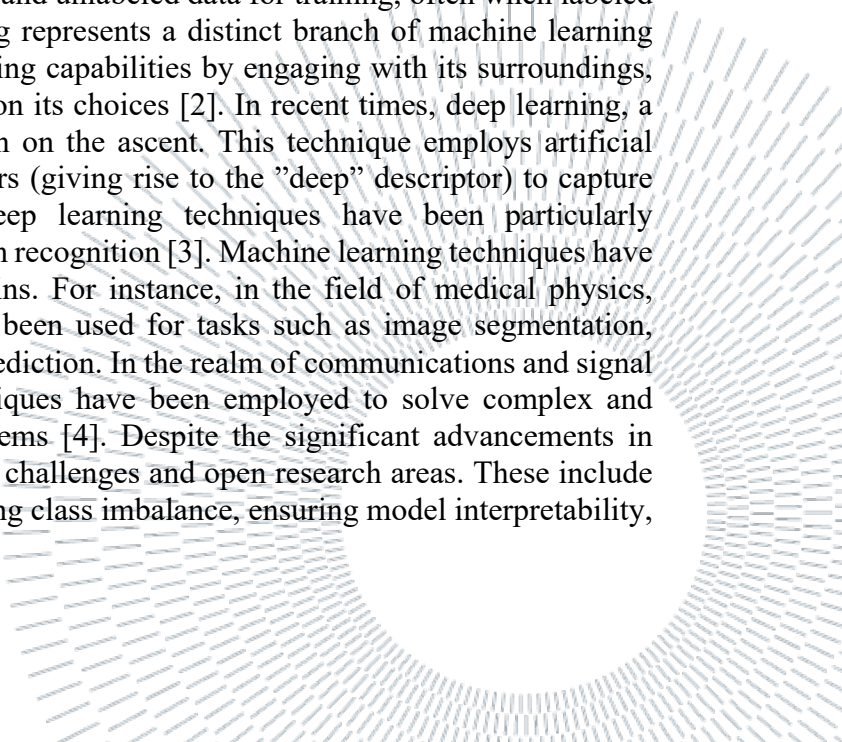
Overall, this thesis aims to present a systematic and thorough examination of our proposed active learning technique. Each chapter contributes to a holistic understanding of our research, from literature context and theoretical foundation to empirical validation and future directions.

3 Literature Review

In the evolving landscape of machine learning, various methodologies and techniques have been developed and refined to address a wide array of challenges. As we delve into this literature review, our aim is to provide a comprehensive overview of these techniques, emphasizing the significance and intricacies of Active Learning (AL). Starting with a broad overview of machine learning techniques, we gradually narrow our focus to the heart of AL, exploring its foundational concepts, its applications in contemporary scenarios like Industry 4.0, and the different sampling techniques that play a pivotal role in its operation. This review also sheds light on the synergy between supportive models and active learning, the significance of fairness in machine learning, and the challenges of applying AL under constrained resources. As we conclude, we will discuss the future prospects and challenges, offering a lens into what the next frontier for AL might be. Join us on this enlightening journey as we uncover the intricacies of active learning and its place in the broader world of machine learning.

3.1. Overview of Machine Learning Techniques

Machine learning, a subset of artificial intelligence, has seen significant advancements over the past few decades. It involves the development of algorithms that enable computers to learn from and make decisions or predictions based on data [1]. Machine learning methodologies can generally be divided into four main categories: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. Supervised learning, the most prevalent form, entails educating a model using a dataset with known labels to forecast outcomes for data it hasn't encountered. In contrast, unsupervised learning works with data lacking labels, aiming to uncover underlying patterns or inherent structures within. Semi-supervised learning merges elements of both, utilizing a mix of labeled and unlabeled data for training, often when labeled datasets are scarce. Reinforcement learning represents a distinct branch of machine learning wherein an agent evolves its decision-making capabilities by engaging with its surroundings, earning rewards or facing penalties based on its choices [2]. In recent times, deep learning, a subcategory of machine learning, has been on the ascent. This technique employs artificial neural networks that possess multiple layers (giving rise to the "deep" descriptor) to capture intricate representations within data. Deep learning techniques have been particularly successful in areas such as image and speech recognition [3]. Machine learning techniques have found wide applicability in various domains. For instance, in the field of medical physics, machine learning and deep learning have been used for tasks such as image segmentation, radiation therapy planning, and outcome prediction. In the realm of communications and signal processing, machine learning-based techniques have been employed to solve complex and analytically intractable optimization problems [4]. Despite the significant advancements in machine learning techniques, there are still challenges and open research areas. These include dealing with high-dimensional data, handling class imbalance, ensuring model interpretability,



and addressing privacy and security concerns. The following sections will delve deeper into some of these issues, particularly focusing on active learning strategies and their role in addressing these challenges.

3.2. Understanding Active Learning

3.2.1. Introduction to Active Learning

Active learning, a notable subset of machine learning, has attracted considerable interest because of its capacity to reduce the reliance on labeled data. In conventional machine learning approaches, models are usually trained on an extensive collection of labeled data. However, the process of labeling, especially for voluminous datasets, can be both expensive and time-intensive. Active learning tackles this challenge by enabling the model to cherry-pick the most insightful data points for labeling, potentially cutting down the amount of labeled data needed to reach a specified performance benchmark [5]. Though the idea behind active learning isn't novel, with investigations tracing back to the 1970s [6], the actual term "active learning" only became prevalent in the 1990s. Over the past ten years, the domain has witnessed a pronounced uptick in enthusiasm, spurred by the rise in access to extensive unlabeled datasets and a mounting demand for models adept at harnessing such data. In the sphere of Industry 4.0, where industrial procedures churn out vast data quantities, active learning becomes particularly pertinent. Manually labeling this avalanche of data often proves to be an unfeasible task. For example, in predictive maintenance applications, active learning can be used to select the most informative data points for labeling, such as those that are most likely to represent a machine failure [7]. There are several strategies for selecting data points in active learning, including uncertainty sampling, query-by-committee, and expected model change. Uncertainty sampling, where the model selects the data points for which it is most uncertain, is one of the most widely used strategies [8]. Query-by-committee involves maintaining a committee of models and selecting data points where the committee members disagree [9]. Expected model change strategies select data points that are expected to result in the greatest change to the current model. A key aspect of active learning is the use of a supportive model to guide the selection of data points. This model, which is trained on the currently labeled data, is used to evaluate the informativeness of the unlabeled data points. The choice of supportive model can have a significant impact on the performance of the active learning process [10]. However, active learning also presents several challenges. One of these is the issue of class imbalance, where one class has many more examples than the other. This can bias the active learning process towards the majority class, resulting in poor performance on the minority class [11]. Another challenge is ensuring fairness in the active learning process, as the model's decisions about which data points to label can potentially introduce bias [12]. Ultimately, active learning usually functions within a set labeling budget, implying there's a restriction on how many data points can be labeled. This scenario creates a balance between exploration, where the model chooses data points to enhance its understanding of the data distribution, and exploitation, in which it picks data points to bolster its present performance. In essence, while active learning emerges as a potent tool for gleaning insights from vast unlabeled datasets, applicable across numerous domains, it also presents certain hurdles. These include handling class imbalance, guaranteeing fairness, and navigating within a predetermined labeling budget.

3.2.2. Active Learning in Industry 4.0

Industry 4.0, also known as the fourth industrial revolution, is characterized by the integration of cyber-physical systems, the Internet of Things, and cloud computing. It has brought about a significant transformation in the manufacturing industry, with artificial intelligence playing a pivotal role in this revolution [13]. Active learning, as a subset of machine learning, has found its applications in various aspects of Industry 4.0. It has been used to optimize production processes, improve product quality, and enhance decision-making. For instance, active learning algorithms have been used in the context of financial big data, where they help in the prediction of financial trends and risk management [14]. Moreover, active learning has been employed in the health sector, which is an integral part of Industry 4.0. It has been used to predict health outcomes and assist in decision-making processes [15]. However, despite these advancements, there are still challenges that need to be addressed. These include dealing with the high dimensionality of data, managing the trade-off between exploration and exploitation, and handling class imbalance, among others. The following sub-sections will delve deeper into these challenges and discuss the various active learning strategies that have been proposed to tackle them.

3.2.3. Main Sampling Techniques in Active Learning

3.2.3.1. Uncertainty Sampling

Uncertainty sampling is a popular strategy in active learning where the learner queries the instances about which it is most uncertain. The goal is to select the most informative instances to label, thereby improving the learning efficiency. In the context of binary classification, uncertainty sampling often involves selecting instances closest to the decision boundary, where the classifier's confidence is the lowest. This is typically measured by the margin sampling strategy, which selects instances for which the difference between the class probabilities is smallest [16]. Another common approach in uncertainty sampling is entropy sampling, where instances are selected based on the entropy of their predicted class probability distribution. The higher the entropy, the greater the uncertainty, and thus the instance is considered more informative [17]. In multi-class problems, a common strategy is to select instances that have the smallest difference between the first and second most probable labels [18]. This strategy is often referred to as margin sampling. Uncertainty sampling has been widely used in various applications due to its simplicity and effectiveness. However, it also has limitations. For example, it tends to query outliers or noisy instances, which can lead to a decrease in performance [19]. Moreover, it assumes that the learner's uncertainty is a good indicator of the instance's informativeness, which may not always be the case. Recent studies have proposed various modifications and enhancements to uncertainty sampling. For instance, [20] introduced a proactive learning framework that combines uncertainty sampling with representativeness to select not only uncertain but also representative instances. Despite its limitations, uncertainty sampling remains a fundamental strategy in active learning, and understanding its principles and variations is crucial for developing more advanced active learning methods.

3.2.3.2. Query by Committee

Query by Committee (QBC) is a popular active learning strategy that involves maintaining a committee of models trained on the current labeled set. The committee then votes on the labels

of the unlabeled instances. Instances about which the committee disagrees the most are considered the most informative and are selected for labeling [9].

The QBC strategy is based on the principle of maximizing disagreement among the committee members. The disagreement is often quantified using measures such as vote entropy or KL-divergence [21]. The committee members are typically trained on bootstrapped subsets of the labeled data, and their diversity is crucial for the effectiveness of the QBC strategy [22].

In the context of Industry 4.0, QBC has been used for tasks such as fault detection in industrial processes. The QBC strategy has also been combined with other active learning strategies, such as uncertainty sampling, to create hybrid methods that leverage the strengths of both approaches [21].

However, QBC has some limitations. The need to maintain and train multiple models can make QBC computationally expensive. Moreover, the effectiveness of QBC depends on the diversity of the committee members, which can be challenging to ensure in practice [22].

3.2.3.3. Expected Model Change

Expected Model Change (EMC) is an active learning strategy that focuses on selecting instances that, when labeled and added to the training set, are expected to result in the most significant change in the current model. This strategy is based on the idea that instances that cause a substantial change in the model are likely to be the most informative.

The EMC strategy often involves calculating a measure of the expected change in the model's parameters or predictions if an instance were to be labeled and added to the training set [23]. This measure can be used to rank the unlabeled instances and select the one that is expected to cause the most significant change.

In the context of Industry 4.0, EMC has been used for tasks such as predictive maintenance and quality control, where it is crucial to quickly learn and adapt to new patterns in the data [24]. The EMC strategy has also been combined with other active learning strategies to create hybrid methods that leverage the strengths of multiple approaches [25].

However, EMC has some limitations. The need to estimate the expected change in the model for each unlabeled instance can make EMC computationally expensive. Moreover, the effectiveness of EMC depends on the accuracy of the model's change estimates, which can be challenging to ensure in practice.

3.2.4. The Role of Supportive Models in Active Learning

Active Learning (AL) is a powerful machine learning paradigm where the learning algorithm selects the instances from which it learns. The goal is to choose instances that are expected to increase the model's performance the most, given a limited labeling budget. Supportive models play a crucial role in this process, guiding the selection of instances for labeling. Supportive models in AL are typically probabilistic models that provide not only class predictions but also uncertainty estimates, often in the form of predicted probabilities. These probabilities are used to identify instances that are informative, i.e., instances for which the model is uncertain about the correct label. The most uncertain instances are then selected for labeling.

Different types of models can be used as supportive models in AL, including Support Vector Machines (SVM), logistic regression, and random forests. Each of these models has its own strengths and weaknesses, and the choice of model can significantly impact the performance of the AL process.

SVMs, for instance, are known for their robustness and ability to handle high-dimensional data. However, they do not naturally provide probability estimates, which are crucial for instance selection in AL. Techniques such as Platt scaling can be used to obtain probabilities from

SVMs, but these are known to be less reliable than probabilities obtained from models that naturally provide them, such as logistic regression or random forests [26].

Logistic regression models, on the other hand, are simple and interpretable models that provide reliable probability estimates. They have been widely used in AL, especially in the initial iterations where the supportive models are not stable yet. A study by Menon et al. [27] found that logistic regression models had good calibration, but machine learning algorithms had poorer calibration despite having comparable Brier scores. This highlights the inherent limitation of interpreting Brier score, which is a composite measure of both discrimination and calibration, in terms of calibration alone.

Random forests are another popular choice for supportive models in AL. They are known for their robustness and ability to handle high-dimensional and noisy data. Moreover, they naturally provide probability estimates by averaging the predictions of the individual trees in the forest. However, these probabilities can be overly confident, especially when the number of trees in the forest is large, which can lead to suboptimal instance selection in AL [28]. In conclusion, the choice of supportive model in AL is a crucial decision that can significantly impact the performance of the AL process. Different models have their own strengths and weaknesses, and the best choice of model depends on the specific characteristics of the data and the task at hand.

3.2.5. Class Imbalance in Active Learning

Class imbalance is a prevalent issue in real-world datasets, where one class significantly outnumbers the other(s). This imbalance can lead to biased learning, with the model favoring the majority class, resulting in poor performance on the minority class. In the context of active learning, class imbalance can further exacerbate the problem, as the model may be less likely to select instances from the minority class for labeling, thereby missing out on valuable information that could improve its performance [29]. Several strategies have been proposed to address class imbalance in active learning. One approach is to adjust the selection criterion to favor instances from the minority class, thereby giving priority to the underrepresented class during instance selection [30]. Another strategy involves balancing the classes in the dataset through oversampling of the minority class or undersampling of the majority class [31]. Unequal probability sampling has also been proposed, where instances from the minority class are given a higher probability of being selected for labeling [32]. Additionally, sample weighting has been used to balance the classes, with instances from the minority class assigned higher weights [33]. There are also model-specific solutions that have been developed to address class imbalance in active learning. For instance, a method has been proposed for handling class imbalance in Support Vector Machines (SVMs) by adjusting the SVM decision boundary towards the minority class [34]. Another approach involves dealing with class imbalance in Random Forests by modifying the tree-building process to favor the minority class [35]. Despite these efforts, there remains a need for an active learning framework that can effectively handle class imbalance across different models and datasets. The proposed AL-Rank and AL-Hybrid models aim to address this need by incorporating class imbalance in their instance selection process, thereby ensuring that the minority class is adequately represented.

3.2.6. Fairness in Machine Learning

Fairness in machine learning is a topic of extensive interest, with a focus on mitigating discrimination and bias in algorithmic decision-making. At a high level, fairness is partitioned into individual fairness, which deals with discrimination against individuals, and group fairness, which considers parity over different demographic groups [36].

One popular notion of fairness is based on model independence or demographic parity, also referred to as statistical parity or disparate impact. This concept requires the sensitive characteristic to be statistically independent of the score [37].

In addition to independence, fairness can be defined using the notions of separation and sufficiency. The separation model allows correlation between the score and a sensitive attribute to the extent that it is justified by the target variable. The sufficiency model requires independence of a target variable and a sensitive attribute conditional to the scores.

The goal of improving fairness in learning problems can be achieved by intervention at pre-processing, in-processing (algorithms), or post-processing strategies. Pre-processing strategies involve the fairness measure in the data preparation step to mitigate the potential bias in the input data [38]. In-process approaches incorporate fairness in the design of the algorithm to generate a fair outcome [39]. Post-process methods manipulate the outcome of the algorithm to mitigate the unfairness of the outcome for the decision-making process.

3.2.7. Active Learning with Limited and Fixed Budget

Active learning strategies are often employed in scenarios where labeling data is costly or time-consuming. In such cases, it is crucial to make the most out of a limited labeling budget. Several recent studies have proposed innovative approaches to tackle this challenge.

A novel active learning method that works efficiently with deep networks by predicting target losses of unlabeled inputs has been proposed [40]. This method is task-agnostic and has been validated through image classification, object detection, and human pose estimation, demonstrating its potential to maximize the utility of a limited labeling budget.

In the context of semantic segmentation under a domain shift, an active learning approach that uses region impurity and prediction uncertainty to guide instance selection has been introduced [41]. This region-based approach has been shown to make more efficient use of a limited budget than image-based or point-based counterparts.

An Online Adaptive Asymmetric Active learning algorithm has been proposed, which is based on a new asymmetric strategy and second-order optimization [42]. This algorithm has been theoretically analyzed for its mistake bound and cost-sensitive metric bounds, providing a robust framework for active learning under a limited budget.

In summary, these studies highlight the importance of innovative active learning strategies in making the most out of a limited and fixed labeling budget. They demonstrate the potential of these strategies in a range of applications, from image classification and object detection to semantic segmentation and industrial applications.

3.2.8. Future Directions and Challenges in Active Learning

Active learning, as a field of machine learning, has seen significant advancements in recent years. However, there are still numerous challenges and potential directions for future research. One of the primary challenges in active learning is scalability. As datasets grow larger and more complex, the computational cost of active learning algorithms can become prohibitive. This is particularly true for methods that require retraining the model after each query, such as uncertainty sampling [16]. Recent research has proposed solutions such as deep active learning [43], which leverages the representational power of deep learning models to handle high-dimensional data.

Another challenge is dealing with high-dimensional data. Traditional active learning methods often struggle in high-dimensional spaces due to the curse of dimensionality. Recent work has proposed using dimensionality reduction techniques in conjunction with active learning to

address this issue [44]. However, more research is needed to develop methods that can effectively handle high-dimensional data without sacrificing the benefits of active learning. Integrating active learning with other machine learning techniques is also a promising direction for future research. For example, combining active learning with reinforcement learning can potentially lead to more efficient learning algorithms [45]. Similarly, integrating active learning with transfer learning can help to leverage knowledge from related tasks, potentially reducing the amount of labeling required.

Furthermore, the development of active learning methods for complex tasks, such as sequence learning and structured prediction, is an important area for future research. While some work has been done in this area [5], there is still much room for improvement.

Finally, the theoretical understanding of active learning is still limited. While some theoretical guarantees exist for certain active learning methods [46], a comprehensive theoretical framework for active learning is still lacking. Developing such a framework is a challenging but important direction for future research.

In conclusion, while active learning has made significant strides, there are still many challenges to overcome and directions to explore. The continued development of this field promises to yield more efficient and effective machine learning algorithms.

4 Methodology and Development of Novel Approaches

4.1. Overview

The methodology chapter serves as the backbone of this thesis, detailing the comprehensive set of techniques, approaches, and innovations employed in our study. It not only sets the foundation for our experiments but also introduces novel techniques designed to improve and expand the existing paradigms of active learning.

Section 3.2 delves into the core building blocks of our methodology. Starting with our choice of baseline models, it subsequently paints the landscape of the active learning framework that serves as the foundation for subsequent developments. Two common strategies, Random Sampling and the Uncertainty Sampling Heuristic, are detailed, providing context for their traditional roles in active learning.

Moving forward to Section 3.3, the spotlight shifts to our original contribution: the AL Rank approach. Here, we unfold the intricacies of the AL Rank methodology, describing its core algorithm and the innovative Model Evaluation and Selection Unit (MESU). This section underscores our efforts to rethink and redefine the conventions of active learning.

Section 3.4 introduces the AL Hybrid approach, another significant contribution of this thesis. A broad overview lays the groundwork, followed by a deep dive into the AL Hybrid model itself. The chapter proceeds to elucidate on the mechanisms behind the Hybrid Instance Selector and the Rank and Entropy Generator. A special emphasis is placed on the AL Hybrid Reverse model, highlighting its differences and nuances compared to its counterpart. To culminate, we encapsulate the myriad advantages of the AL Hybrid approach, making a case for its efficacy and potential in the field.

In essence, this chapter is not just a recount of techniques but an exploration of new horizons in active learning. It symbolizes a journey from understanding traditional models to forging new paths with the AL Rank, AL Hybrid, and AL Hybrid Reverse approaches.

4.2. Preliminaries

4.2.1. Baseline Model Choice

In this research, we have chosen to use uncertainty sampling active learning using Shannon entropy as our baseline model. This decision is motivated by the fact that uncertainty sampling is one of the most common and widely used approaches in active learning. It is a well-established technique that selects the data point for which the current model is least certain about its label, thus making it a suitable baseline for our novel solution.

Uncertainty sampling operates by maximizing the Shannon entropy over the probabilities of the labels at each iteration, which is a measure of the uncertainty or randomness of the data. The effectiveness of this strategy heavily relies on the precision and stability of the model, especially in the initial iterations where the model is still in its early stages of learning. Therefore, using the predicted probabilities directly can potentially lead to suboptimal selection of instances. However, the model's predictions are still useful for ranking observations, and instead of depending on the absolute predicted probabilities, we consider the relative ranks of these probabilities.

In addition to uncertainty sampling, we also used random sampling as a second baseline. Random sampling is a simple yet effective method that selects instances randomly from the pool of unlabeled data. Despite its simplicity, random sampling can sometimes be hard to outperform, especially in situations where the data is highly diverse or the model's predictions are not reliable.

The choice of these two baselines allows us to compare and contrast the performance of our proposed AL-Rank and AL-Hybrid models against both a sophisticated, uncertainty-based approach and a simple, random approach. This comparison will provide valuable insights into the effectiveness of our novel active learning strategies.

4.2.2. Active Learning Framework

The active learning framework is a dynamic process that aims to construct an accurate model in the most efficient manner. It operates in a scenario where we have a classifier and an unlabeled data pool, denoted as U . The data pool U is assumed to be an independent and identically distributed (i.i.d.) sample set derived from an unknown distribution. Each data instance $x_i \in U$ is associated with a class label y_i , which can take on one of K possible values $0, \dots, K - 1$.

The ultimate goal of the active learning process is to learn a classifier function $M : \mathbb{R}^d \rightarrow [0, K - 1]$ that maps the feature space X to the labels y . Given an input x , the predicted label is denoted as $\hat{y} = M(x)$. This classifier function is the core of the active learning framework, as it is responsible for making predictions on the unlabeled data instances.

The active learning process operates by sequentially selecting instances from the data pool U to be labeled by an expert oracle. This results in a labeled set L , which is then used for training the classifier. This process is illustrated in Algorithm 1, which represents the conventional active learning algorithm (see Table 1).

Algorithm 1 Active Learning

- 1: **for** $t = 1$ to B **do**
 - 2: $\mathbf{x}^* = \arg \max_{\mathbf{x} \in \mathcal{U}} \mathcal{H}(y|\mathbf{x}, \mathcal{L})$
 - 3: $y =$ label of \mathbf{x}^* obtained from the labeling oracle
 - 4: Add $\langle \mathbf{x}^*, y \rangle$ to \mathcal{L}
 - 5: Train the classifier M_t using \mathcal{L}
 - 6: **end for**
 - 7: **return** trained model M_t
-

Table 1: Active Learning Algorithm

At each iteration t , the algorithm selects the data point \mathbf{x}^* that maximizes the Shannon entropy (H) over the probabilities of the labels. The label of \mathbf{x}^* is then obtained from the labeling oracle, and the point is added to the labeled dataset \mathcal{L} . The classifier M_t is then trained using \mathcal{L} , and this process continues until the labeling budget B is exhausted.

The process of uncertainty sampling in active learning is visually represented in Figure 1. In each cell of the figure, the value at the top represents the auxiliary model’s predicted probability for a given point. Conversely, the value at the bottom of each cell indicates the corresponding percentile rank of the predicted probability.

However, the labeling process is often costly and constrained by a limited budget of B . Therefore, the challenge lies in devising an effective sampling strategy that optimally utilizes the budget to construct the most accurate model. This is where the concept of active learning comes into play, as it seeks to strategically select the most informative instances for labeling, thereby maximizing the utility of the labeling budget.

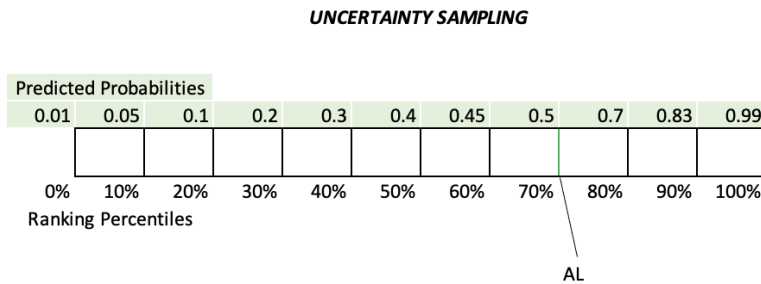


Figure 1: Uncertainty Sampling

The active learning framework is a powerful tool for constructing accurate models, especially in scenarios where labeled data is scarce or expensive to obtain. However, its effectiveness heavily relies on the sampling strategy used to select instances for labeling. In this research, we propose two novel active learning strategies, AL-Rank and AL-Hybrid, which aim to improve the efficiency of the active learning process by leveraging rank-based and entropy-based criteria for instance selection.

4.2.3. Random Sampling

Random sampling stands as a straightforward but potent technique employed across various research domains, including active learning. Differing from intricate sampling approaches, random sampling eschews specific criteria or algorithmic guidelines when choosing samples. It functions by haphazardly picking instances from the reservoir of unlabeled data. This element of unpredictability ensures every instance possesses an equal likelihood of selection, negating potential biases.

Within active learning, random sampling acts as a foundational strategy, offering a reference point for contrasting with more advanced methods. Though basic, it can occasionally pose a challenge to surpass, particularly when dealing with highly varied data or when a model's predictions lack consistency. In such scenarios, the inherent unpredictability of the method might serendipitously choose valuable instances that alternative strategies might miss.

However, the efficacy of random sampling hinges on the dataset's volume and variety. In expansive and diverse datasets, it can likely encompass a broad spectrum of instances, enhancing the model's robustness. Conversely, with smaller or uniform datasets, its efficacy might wane, given the reduced odds of picking significant samples.

Regardless of these nuances, random sampling retains its importance in the active learning sphere. It paves the way for crafting superior sampling techniques and establishes a standard to gauge their effectiveness. In our study, we enlist random sampling as a foundational model, juxtaposed with uncertainty sampling, aiming for an exhaustive analysis against our introduced AL-Rank and AL-Hybrid models.

4.2.4. Uncertainty Sampling Heuristic

Uncertainty sampling is a widely adopted strategy in active learning, which operates on the principle of selecting the data point $\mathbf{x}^* \in D$ for which the current model is least certain about its label. This strategy is based on the concept of Shannon entropy, a measure of the uncertainty or randomness of the data. At each iteration t , the classifier M_{t-1} selects the data point that maximizes the Shannon entropy (H) over the probabilities of the labels.

The Shannon entropy is calculated using the following equation:

Equation 1: Shannon Entropy - Data Point to Be Selected

$$\mathbf{x}^* = \operatorname{argmax}_{\mathbf{x} \in U} \mathcal{H}(y|\mathbf{x}, \mathcal{L})$$

where \mathcal{H} is defined as:

Equation 2: Shannon Entropy – Entropy Value

$$\mathcal{H}(\mathbf{x}, \mathcal{L}) = - \sum_{i=1}^n P(y_i|\mathbf{x}, \mathcal{L}) \log_2 P(y_i|\mathbf{x}, \mathcal{L})$$

In this context, \mathcal{H} represents the Shannon entropy, $P(y_i|\mathbf{x}, \mathcal{L})$ is the conditional probability of each outcome y_i given the unlabeled data point \mathbf{x} and the labeled set \mathcal{L} . The logarithm is base 2 as entropy is often measured in bits. The summation is taken over all outcomes of the random variable Y .

The active learning algorithm, as illustrated in Algorithm 1, employs this principle of uncertainty sampling. It iteratively selects the next data point from \mathcal{U} to be labeled, using the classifier trained in the previous step, M_t , to obtain probabilities of the labels. The algorithm obtains the label of the point with the maximum entropy from the labeling oracle, adds the point to the labeled dataset \mathcal{L} , and uses it to train the classifier M_t . This process continues until the labeling budget is exhausted.

The functioning of the traditional Max Entropy active learning technique is visually represented in Figure 1. In each cell of the figure, the value at the top represents the auxiliary model's predicted probability for a given point. Conversely, the value at the bottom of each cell indicates the corresponding percentile rank of the predicted probability. In the context of Max Entropy, the active learning model seeks to select the point with a predicted probability value closest to 0.5. This is because points with probabilities close to 0.5 are considered the most uncertain, thereby having the highest entropy, and are deemed to be the most informative for the learning model.

The effectiveness of the uncertainty sampling strategy heavily relies on the precision and stability of the auxiliary model. In the initial iterations, where the model is still in its early stages of learning, these predictions might not be reliable. In these situations, using the predicted probabilities directly can potentially lead to suboptimal selection of instances. However, the model's predictions are still useful for ranking observations. Instead of depending on the absolute predicted probabilities, we consider the relative ranks of these probabilities. This approach allows us to leverage the uncertainty inherent in the data to guide the selection of instances for labeling, thereby enhancing the efficiency and accuracy of the active learning process.

4.3. The AL Rank Approach: Development and Implementation

The AL-Rank approach is a novel active learning strategy that optimizes the labeling efficiency by leveraging a single auxiliary model, such as an SVM, logistic regression, or a random forest model, to rank the unlabeled instances according to their predicted class probabilities. This approach is based on the hypothesis that the auxiliary model has a higher fidelity in sorting the samples from the most likely to belong to class 0 to the most likely to belong to class 1, rather than estimating the precise numerical probabilities. Therefore, by systematically selecting instances for labeling based on these rankings, the AL Rank approach achieves faster accuracy convergence.

4.3.1. AL Rank Methodology

The AL-Rank methodology is visually represented in Figure 2, which contrasts the difference in approach between Max Entropy active learning and our proposed AL Rank method. In each cell of the figure, the value at the top represents the predicted probability of a point by the auxiliary model, while the bottom value denotes the respective percentile rank of this probability.

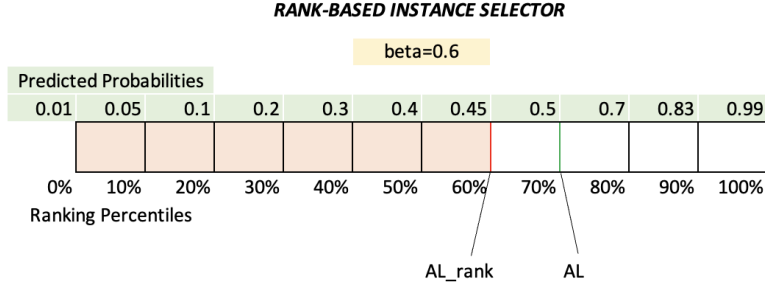


Figure 2: Rank-Based Instance Selector ($\beta=0.6$)

The green line in the figure is drawn at the point with a predicted probability closest to 0.5. This represents the selection made by the Max Entropy active learning technique, indicating the point deemed most informative based on the entropy criterion. On the other hand, the red line is positioned at the β percentile point, representing the point chosen by the AL Rank method. This selection operates on our hypothesis that the rank order of points is more reliable than their predicted probability values, especially in the initial iterations. By aligning with the β percentile, the AL Rank method aims to select a point that is closer to the true decision boundary, thus being potentially more informative for model learning.

4.3.2. AL Rank Algorithm

The AL-Rank framework follows an iterative process similar to standard active learning methods, as illustrated in Algorithm 2 (see Table 2). At the core of the framework is the Model Evaluation and Selection Unit (MESU). The MESU identifies an unlabeled instance $\langle X(i) \rangle$ from the pool U and obtains its label from an oracle. Once labeled, the instance $\langle X(i), y(i) \rangle$ is added to L , the labeled set, and used to train the supportive model M_t . In the subsequent iteration $t + 1$, the MESU employs M_t to select the next instance to be labeled based on the generated ranking. This iterative process continues until the pre-allocated labeling budget is exhausted.

Algorithm 2 Active Learning Ranking (AL Rank)

Require: $U, L, M_{t-1}, B, \beta, N$

- 1: Initialize L by randomly selecting N instances from each class in U , and remove selected instances from U
 - 2: **while** $|L| < (B + 2N)$ **do**
 - 3: Generate rank list r for U using M_{t-1}
 - 4: $\mathbf{x}^* = \arg \max_{\mathbf{x} \in U}$ point at the β percentile of r
 - 5: Label \mathbf{x}^* with oracle, resulting in label y^*
 - 6: $L = L \cup \{(\mathbf{x}^*, y^*)\}$, and remove \mathbf{x}^* from U
 - 7: Retrain M_t using L
 - 8: **end while**
 - 9: **return** trained model M_t
-

Table 2: AL Rank Algorithm

4.3.3. Model Evaluation and Selection Unit (MESU)

The MESU consists of two essential computational components that contribute to the AL- Rank approach. The upper component, known as the Rank-Based Instance Selector, is responsible for choosing the next instance from U to be labeled. Instead of selecting based on the highest predicted probability or highest uncertainty, the Rank-Based Instance Selector utilizes the instance's rank in the ordered list generated from the auxiliary model's predicted probabilities. The instances are ranked from the most probable to be in class 0 to the most probable to be in class 1.

The Rank-Based Instance Selector introduces a unique selection criterion that sets the AL- Rank approach apart from conventional active learning methods. It selects the instance located at the β percentile of the ranked list. This β value used for selecting the percentile is derived from prior knowledge or assumptions about the percentage of instances belonging to class 0 in the original dataset. By focusing on this specific percentile, the AL-Rank approach aims to circumvent potential mispredictions that are more likely to occur in the early iterations of the active learning process.

The rationale behind selecting the β percentile is to prioritize instances that are expected to provide the most valuable information for refining the decision boundary. By considering the class distribution in the dataset, the approach identifies instances near the decision boundary, which are more likely to be challenging and prone to misclassification. By focusing on these instances early on, the AL-Rank approach aims to accelerate the learning process and improve the model's performance more rapidly.

By incorporating prior knowledge or assumptions about the class distribution through the β value, AL-Rank takes a proactive approach to mitigate the potential impact of mispredictions in the early stages of active learning. This strategy helps to optimize the allocation of labeling resources by prioritizing instances that are likely to contribute the most significant information for improving the model's accuracy.

4.4. The AL Hybrid Approach: Development and Implementation

4.4.1. Overview of AL Hybrid Approach

The AL Hybrid approach is a novel active learning strategy that combines the strengths of classical active learning techniques with the rank-based methodology of the AL Rank approach. This model aims to expedite the convergence of accuracy by adopting a more adaptive and informed approach to instance selection.

The underlying rationale for the Hybrid approach stems from the observation that as the number of iterations increases, the auxiliary model becomes more reliable, thus enhancing the precision of traditional active learning methods. As such, the Hybrid approach has been designed to optimize performance by dynamically switching between AL Rank and traditional active learning methods as the model evolves. In the development of the Hybrid approach, we utilized the Maximum Entropy active learning technique.

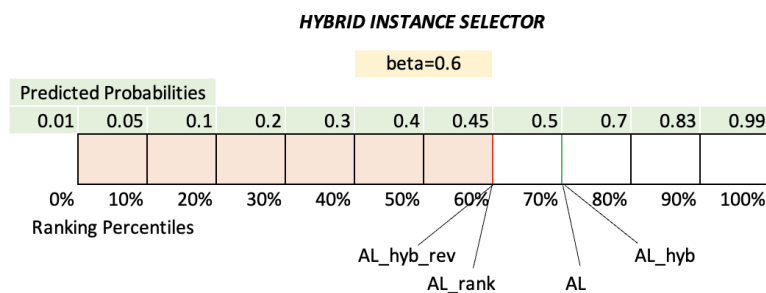


Figure 3: Hybrid Instance Selector ($\beta=0.6$)

The functioning of the AL Hybrid approach is visually represented in Figure 3. In this figure, the top value in each cell corresponds to the predicted probability of a point by the auxiliary model, while the bottom value denotes the respective percentile rank of this probability. The green line in the figure is drawn at the point with a predicted probability closest to 0.5, representing the selection made by the Max Entropy active learning technique. On the other hand, the red line is positioned at the β percentile point, representing the point chosen by the AL Hybrid method. This selection operates on our hypothesis that the rank order of points is more reliable than their predicted probability values, especially in the initial iterations. By aligning with the β percentile, the AL Hybrid method aims to select a point that is closer to the true decision boundary, thus being potentially more informative for model learning.

4.4.2. AL Hybrid Model

In the initial stages of the active learning process, the AL Hybrid model employs the classical active learning approach, where instances with high uncertainty are selected for labeling. This strategy ensures the exploration of uncertain regions in the feature space, which are likely to be the most informative for the learning model. By focusing on these high-uncertainty instances, the model reduces potential biases introduced by a single model and enhances the diversity of the labeled dataset.

As the active learning process progresses and the supportive model becomes more reliable, the AL Hybrid model integrates the rank-based methodology of AL Rank. This methodology leverages the instance rankings generated by the supportive model to prioritize the labeling process.

Algorithm 3 Active Learning Hybrid (AL Hybrid)

Require: $U, L, M_{t-1}, B, \beta, N$

- 1: Initialize L by randomly selecting N instances from each class in U , and remove selected instances from U
 - 2: **while** $|L| < (B + 2N)$ **do**
 - 3: Generate entropy values for U using M_{t-1}
 - 4: Pre-select $\mathbf{x}^{\text{pre}} = \arg \max \text{entropy of } \mathbf{x}$
 - 5: Generate rank list r for U using M_{t-1}
 - 6: Evaluate rank of \mathbf{x}^{pre} in r
 - 7: **if** $(\beta \geq 0.5 \text{ and rank of } \mathbf{x}^{\text{pre}} \geq \beta)$ or $(\beta < 0.5 \text{ and rank of } \mathbf{x}^{\text{pre}} < \beta)$ **then**
 - 8: $\mathbf{x}^* = \mathbf{x}^{\text{pre}}$
 - 9: **else**
 - 10: $\mathbf{x}^* = \text{point at the } \beta \text{ percentile of } r$
 - 11: **end if**
 - 12: Label \mathbf{x}^* with oracle, resulting in label y^*
 - 13: $L = L \cup \{(\mathbf{x}^*, y^*)\}$, and remove \mathbf{x}^* from U
 - 14: Retrain M_t using L
 - 15: **end while**
 - 16: **return trained model** M_t
-

Table 3: AL Hybrid Algorithm

Instead of solely relying on the absolute values of the predicted probabilities, the rank-based methodology exploits the relative ordering of these probabilities. This approach is based on the assumption that the supportive model has a higher fidelity in sorting the samples from the most likely to belong to class 0 to the most likely to belong to class 1, rather than estimating the precise numerical probabilities.

The AL Hybrid model follows an iterative process, as outlined in Algorithm 3 (see Table 3). At the core of this process is the Model Evaluation and Selection Unit (MESU), which is responsible for identifying the most informative unlabeled instance from the pool U and obtaining its label from an oracle. Once labeled, the instance is added to L , the labeled set, and used to train the supportive model M_t . In the subsequent iteration $t + 1$, the MESU employs M_t to select the

next instance to be labeled based on the generated ranking. This iterative process continues until the pre-allocated labeling budget is exhausted.

By combining the strengths of classical active learning and the AL Rank approach, the AL Hybrid model aims to adapt its instance selection mechanism to the reliability of the supportive model. This adaptive approach allows the model to leverage the benefits of both methodologies, resulting in a more efficient and effective labeling process that expedites accuracy convergence.

4.4.3. Hybrid Instance Selector

The Hybrid Instance Selector is a key component of the AL Hybrid model that is responsible for identifying the most informative instance from the unlabeled data pool U to be labeled next. This component operates by employing a two-stage approach that combines the principles of classical active learning and the rank-based methodology of AL Rank.

In the first stage, the Hybrid Instance Selector pre-selects an instance based on its entropy value, which is a measure of the uncertainty or randomness of the data. Higher entropy values indicate a greater level of uncertainty, suggesting that the model is less confident in its prediction.

Consequently, instances with higher entropy are considered more informative and are prioritized for labeling (Algorithm 3, Line 5).

Following the entropy-based pre-selection, the second stage focuses on evaluating the rank of the pre-selected instance relative to the β threshold in the ranked list generated from the supportive model's predicted probabilities (Algorithm 3, Lines 7-9). The Hybrid Instance Selector operates on the principle of choosing instances that are closer to the minority class, determined based on the original β ratio. If the β ratio is above 0.5, indicating that the majority class is class 0 (the negative class), the minority class is class 1 (the positive class). Conversely, if the β value is below 0.5, the minority class is class 0.

The Hybrid Instance Selector's decision-making process combines the uncertainty-based selection of instances with high entropy and the rank-based selection that focuses on instances closer to the minority class. This integration enables the Hybrid Instance Selector to adaptively prioritize instances for labeling, taking into account both the informative nature of high entropy and the potential for refinement near the decision boundary. By leveraging these two methodologies, the Hybrid Instance Selector aims to enhance the efficiency and effectiveness of the instance selection process in active learning, thereby expediting the convergence of accuracy.

4.4.4. Rank and Entropy Generator

The Rank and Entropy Generator is another crucial component of the AL Hybrid model. It is responsible for generating the entropy values and the rank list that are used by the Hybrid Instance Selector for instance selection (Algorithm 3, Lines 4 and 6).

The entropy values are generated based on the current model's predictions. Entropy is a measure of the uncertainty or randomness of the data. Higher entropy values indicate a greater level of uncertainty, suggesting that the model is less confident in its prediction. Consequently, instances with higher entropy are considered more informative and are prioritized for labeling.

The rank list, on the other hand, is generated based on the predicted class probabilities of the instances in the unlabeled data pool U . The instances are ranked from the most probable to belong to class 0 to the most probable to belong to class 1. This rank list is then used by the Hybrid Instance Selector to prioritize instances for labeling based on their rank relative to the β threshold.

By generating these entropy values and rank list, the Rank and Entropy Generator provides the necessary information for the Hybrid Instance Selector to make informed decisions about which instances to label. This, in turn, enhances the efficiency and effectiveness of the instance selection process in active learning, thereby expediting the convergence of accuracy.

4.4.5. AL Hybrid Reverse Model

The AL Hybrid Reverse model is a variant of the AL Hybrid model that operates in a reverse order. While the AL Hybrid model selects the pre-selected instance for labeling if it is closer to the minority class than the β threshold, the AL Hybrid Reverse model selects the pre-selected instance if it is closer to the majority class.

This decision-making process is illustrated in Figure 4. For instance, consider a scenario where the pre-selected instance with the highest entropy is found at the 70th percentile in the ranking (as shown in Figure 4). If the original dataset's β value is 0.4, indicating that 40% of the instances belong to class 0, then the rank of the pre-selected instance is closer to the majority class (class 1). In this case, the AL Hybrid Reverse model, following the classical active learning strategy, would choose the pre-selected sample (i.e., the 70th percentile sample) for labeling.

By operating in a reverse order, the AL Hybrid Reverse model offers an alternative approach to instance selection in active learning. It provides a different perspective on the trade-off between exploration and exploitation in the learning process, potentially leading to different learning dynamics and performance outcomes.

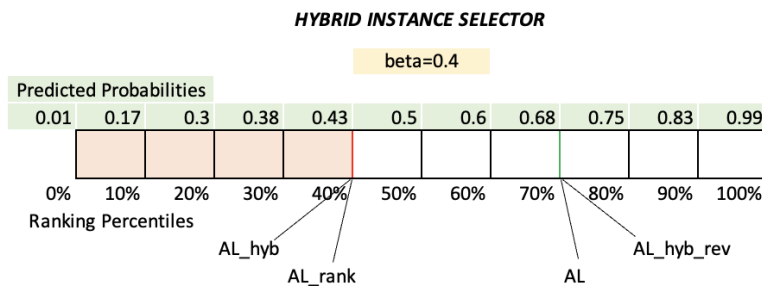


Figure 4: Hybrid Instance Selector (beta=0.4)

This model can be particularly useful in scenarios where the majority class is more diverse or harder to learn, thus requiring more focus on the instances closer to the majority class.

4.4.6. Advantages of AL Hybrid Approach

The AL Hybrid approach offers several advantages that make it a compelling choice for active learning.

1. **Adaptive Learning Strategy:** The AL Hybrid models dynamically switch between the classical active learning and the AL Rank approach based on the reliability of the model's predictions. This adaptive strategy allows the models to leverage the strengths of both methodologies, leading to a more efficient and effective labeling process.
2. **Efficient Use of Labeling Resources:** By incorporating the rank-based methodology, the AL Hybrid approach prioritizes instances that are likely to contribute the most significant information for improving the model's accuracy. This strategy optimizes the allocation of labeling resources, thereby enhancing the efficiency of the learning process.
3. **Accelerated Accuracy Convergence:** The AL Hybrid approach aims to expedite the convergence of accuracy by adopting a more informed approach to instance selection. By focusing on instances that are closer to the decision boundary, the models can refine the decision boundary more effectively, leading to faster improvement in model accuracy.
4. **Robustness to Initial Model Inaccuracy:** The AL Hybrid approach is designed to mitigate the potential impact of mispredictions in the early stages of active learning. By

considering the rank order of instances rather than their absolute predicted probabilities, the models can be more robust to initial model inaccuracies.

5. **Flexibility:** The AL Hybrid approach is flexible and can be easily adapted to different learning scenarios. The β parameter can be adjusted based on prior knowledge or assumptions about the class distribution in the dataset, allowing the models to be tailored to the specific characteristics of the learning task.

In summary, the AL Hybrid approach provides a versatile and efficient solution for active learning. By combining the strengths of classical active learning and the AL Rank approach, it offers a powerful tool for constructing accurate models, especially in scenarios where labeled data is scarce or expensive to obtain.

5 Experiment Design

5.1. Overview

Chapter 4 delves into the meticulous design and details of the experimental framework. This comprehensive exploration encompasses every facet of the study, ensuring a profound understanding of the process flow and methodologies incorporated.

- **Data Collection and Description:** This segment elucidates the origin of the data, its inherent characteristics, and the reasons for its selection. It serves as the foundational stone, shedding light on the very start of our research pipeline.
- **Data Preprocessing:** Recognizing the importance of pristine data, this section discusses the measures undertaken to refine the dataset. From normalization to addressing class imbalances and eliminating anomalies, each step is explained with its underlying rationale.
- **Experimental Setup:** Serving as the heart of this chapter, this extensive section unravels the entire blueprint of our experimentation. The section starts with the strategies employed for labeling data, progressing to elucidate the modeling approach. A dedicated focus is given to the active learning methodologies, underscoring their importance in our study. The modeling phase is complemented with a rigorous set of metrics designed to evaluate both the performance and fairness of our strategies. Finally, the section concludes by laying out the procedural steps that guide the execution of our experiments.

With this chapter, we aim to offer a clear and detailed map of our experimental journey, paving the way for the results and discussions in the succeeding chapters.

5.2. Data Collection and Description

The datasets chosen for this research play a pivotal role in addressing the thesis's objectives. They were selected based on their relevance to the research questions, comprehensiveness, and data integrity. This section delves into a meticulous description of each dataset, elaborating on its source, characteristics, and significance concerning the study's goals.

5.2.1. Source of the Data

Adult Dataset: Sourced from the 1994 United States Census Bureau data, the Adult dataset, colloquially known as the "Census Income" dataset, was curated by Ronny Kohavi and Barry Becker. With an intent to predict an individual's income bracket using personal details, it offers

a blend of demographic and socio-economic variables. Its relevance to this study stems from its comprehensive representation of a broad population, ensuring varied data patterns.

COMPAS Dataset: The COMPAS dataset is derived from the Broward County Sheriff's Office in Florida, encompassing scores from 2013 to 2014. Acquired through a public records request, it was historically utilized for informed pretrial decisions in Broward County. This dataset is crucial for our research due to its emphasis on risk scores, which resonate with our analytical objectives.

ELS:2002 Dataset: Designed for a holistic evaluation of the US education policy, the ELS:2002 dataset is monumental in understanding educational trajectories. Capturing dynamics ranging from school attributes, parental involvement, to post-high school transitions, its richness in features makes it indispensable for this study.

5.2.2. Data Description

Adult Dataset:

- **Size:** Comprises 48,842 entries.
- **Time Span:** Data from the year 1994.
- **Missing Values:** 3,620 entries have incomplete data, yielding 45,222 complete rows.
- **Features:** Includes 14 variables such as Age, Workclass, Education, Occupation, Race, and more.
- **Class Distribution:** Oriented for binary classification, it segments data as 'above 50K'

COMPAS Dataset:

- **Size:** Started with 18,610 individuals, narrowed down to 11,757 after focusing on pretrial scores.
- **Time Span:** Data from 2013-2014.
- **Features:** Evaluates defendants on three COMPAS scores with a range from 1 to 10.
- **Class Distribution:** Scores are labeled as "Low" (1-4), "Medium" (5-7), and "High" (8-10).
- **Data Integrity:** A scrutiny revealed a 3.75

ELS:2002 Dataset

- **Objective:** Assesses multiple educational policy aspects, capturing a student's journey from high school to employment or further studies.
- **Features:** Encompasses demographics, academic performance, employment details, and more.
- **Time Span:** Focuses on data trends and transitions over multiple years.

5.3. Data Preprocessing

Effective preprocessing of data is an integral step in machine learning experimentation. The aim of this preprocessing phase was to transform the datasets to be suitable for our experimental framework and to ensure robustness and consistency in model training.

5.3.1. Objectives of Preprocessing

Our primary preprocessing objectives included:

1. **Normalization:** Standardizing the data to ensure all features are within a similar scale, improving convergence and performance of certain algorithms.
2. **Sampling:** Creating balanced test sets for consistent evaluation and understanding the model's capacity to generalize.
3. **Class Balance Adjustment:** Adapting the datasets according to the input beta value, ensuring even representation and eliminating potential biases.

5.3.2. Normalization

To make sure that all features contributed equally to the model performance, and to aid the convergence of algorithms like SVM, the MinMax scaler from the sklearn library was applied. This transformation rescaled every feature to lie between 0 and 1.

5.3.3. Sampling Strategy

For evaluation consistency, test sets were created with a particular focus on class balance:

- ELS and Adult datasets: 4000 observations each, ensuring a 50-50 split between the two classes.
- COMPAS dataset: 2400 observations, maintaining the same balanced class distribution.

5.3.4. Class Balance Adjustment

Post-normalization, an additional step was introduced to align the datasets with the desired beta value. This involved oversampling with replacement, ensuring the representation matched the target beta value. After this modification:

- ELS and Adult datasets: 6000 samples were randomly chosen for training.
- COMPAS dataset: 3400 samples were used.

5.3.5. Handling Missing Values and Outliers

Given the high-quality nature of the datasets, minimal missing values were detected. In the few instances where they occurred, a mean-imputation strategy was adopted. Outliers, identified using the IQR method, were suitably capped to maintain data integrity without introducing bias.

5.4. Experimental Setup

The integrity and effectiveness of the research are anchored to a meticulously constructed experimental setup. This setup, underpinned by structured procedures, has been designed to guarantee both the repeatability of the results and to serve the overarching research goals. In this section, we will delve deep into the core aspects of the setup.

5.4.1. Labeling Strategy

To ensure robust model training and foster unbiased evaluation, a two-tier labeling strategy was implemented.

5.4.1.1. Initial Labeling

A foundational step in our experimental design was the initialization of our labeled dataset. Given the binary nature of our classes (0 and 1), we aimed for a balanced representation to preclude any inherent bias. We carefully selected three instances randomly from each class, culminating in an initial labeled set with six instances. This even-handed representation augments the model's ability to discern patterns with impartiality.

5.4.1.2. Iterative Labeling

Post initialization, an iterative strategy was adopted to further amplify the labeled dataset. In every iteration, an instance was judiciously moved from the unlabeled dataset to the labeled one. This strategic transfer ensured a gradual enrichment of the labeled dataset, enabling the models to learn progressively. The iterative process was capped once our labeling budget, set at 200 labels, was completely utilized. This iterative approach is particularly effective as it facilitates model improvement with each successive instance addition.

5.4.2. Modeling Approach

For this research, the choice of models is pivotal. Multiple models were considered to ensure comprehensive evaluation and to cater to the diverse nature of the data.

5.4.2.1. Supportive Models

Three models were chosen as the linchpins of our experiments:

1. Support Vector Machine (SVM): Recognized for its prowess in high-dimensional spaces, the SVM used comes with a linear kernel, ideal for our dataset structure.
2. Logistic Regression: This model, equipped with a liblinear solver, is particularly adept at binary classification tasks, aligning with our class structure.
3. Random Forest: A ensemble-based model, our random forest implementation underwent meticulous tuning. Specifically, cross-validation was employed to tune the maximum features, yielding four for both ELS and COMPAS datasets, and six for the Adult dataset. Beyond this, we adhered to the default hyperparameters provided by the sklearn library, given their empirically proven efficiency for a myriad of tasks.

This triad of models was chosen not just for their individual strengths but also for their collective ability to provide diverse insights, ensuring our findings are both comprehensive and robust.

5.4.3. Data Selection via Active Learning

In our experiments, active learning plays a pivotal role in improving the efficiency of our model's learning process. Two primary active learning strategies have been used in our research: uncertainty sampling and Query By Committee (QBC). Both techniques have been chosen to serve as baselines against which the performance of our proposed methods can be compared.

5.4.3.1. Uncertainty Sampling

Uncertainty sampling is a fundamental active learning strategy that queries labels for the data points about which the model's prediction is most uncertain.

1. Principle: At its essence, uncertainty sampling aims to refine the model's learning by focusing on instances where it exhibits the least confidence. The strategy posits that concentrating on these uncertain instances will allow the model to achieve higher accuracy with fewer labeled examples, thereby optimizing the learning curve.
2. Implementation: The measure of uncertainty varies with the nature of the problem (be it binary classification, multi-class classification, regression, etc.). For classification problems, one typical approach is to select data points where the predicted class probabilities are close to 0.5 in binary classification or where the entropy of the predicted class distribution is highest in multi-class scenarios.
3. Benefits:
 - a. Reduces label cost: By prioritizing the most uncertain instances, the need for numerous labels diminishes to achieve a particular model performance level.
 - b. Accelerates learning: Given that models are trained based on instances they find challenging, their performance improves at a quicker pace.

5.4.3.2. Query By Committee (QBC)

In the realm of active learning, Query By Committee (QBC) stands as a prominent strategy. The method is rooted in maintaining a committee of models and leveraging their disagreements to identify the data instances for which labels should be queried. The steps and reasoning for adopting QBC in our experiments are detailed below:

1. Rationale: The main impetus behind employing QBC is to introduce another baseline, contrasting with uncertainty sampling. This differentiation aids in ensuring a comprehensive evaluation of our newly proposed methods.
2. Methodology: The crux of QBC lies in the use of a committee of models. In our study, we utilized a committee composed of two identical model architectures but trained in distinct ways. Specifically, our committee comprised models like SVM, LR, and RF. The training involved techniques like bagging to induce diversity among the models. This diversity is crucial for creating disagreements, which QBC exploits.

3. Disagreement Measure: To gauge the level of disagreement among committee members, we employed the vote entropy. This measure effectively captures the uncertainty across models about which class a particular instance belongs to.
4. Iteration Mechanism: At each active learning iteration, the data point with the highest disagreement (as per the vote entropy) is selected. Its label is then queried and appended to the training set for the next iteration.
5. Results Aggregation: After training our auxiliary models (SVM, LR, and RF) on the adult dataset, we calculated the mean performance for all these models. This aggregated result acts as the baseline against which the performance of our new models is juxtaposed.

5.4.4. Performance Metrics

In the active learning paradigm, the choice of evaluation metrics is pivotal in assessing the model's performance and understanding the effectiveness of different strategies. This study focuses predominantly on accuracy and the Convergence Area Under The Curve (CAUC). Moreover, with the increasing emphasis on fairness in machine learning, we've integrated essential fairness metrics.

5.4.4.1. Accuracy

Accuracy, as a widespread metric, illustrates the model's capability to predict labels correctly for the given dataset. Within our study, accuracy serves a critical role, offering insights from the continuous active learning iterations. This metric is assessed at three significant budget points: $b=50$, $b=100$, and $b=200$, representing the accuracy after 50, 100, and 200 labeled samples, respectively, hence detailing the model's refinement over iterations.

5.4.4.2. Convergence Area Under The Curve (CAUC)

As active learning models leverage incrementally labeled data, gauging their convergence rate becomes essential. CAUC serves this purpose, computing the integral of the accuracy curve up to a set budget, encapsulating the model's efficiency across iterations. A pronounced CAUC indicates swift model enhancements, marking a proficient active learning approach.

5.4.4.3. Fairness Metrics

Given the urgency of fairness in machine learning, we've incorporated fairness metrics to ensure our models neither perpetuate nor magnify inherent biases in the data.

1. Accuracy Disparity: This metric measures the difference in accuracy between the protected (P) and non-protected (NP) groups. Mathematically, it is represented as:

Equation 3: Accuracy Disparity

$$\text{Accuracy Disparity} = |\text{Accuracy}_P - \text{Accuracy}_{NP}|$$

A smaller disparity implies that the model is treating both groups more uniformly.

2. Equalized Odds: Equalized odds ensures that both true positive rates (TPR) and false positive rates (FPR) are consistent across the protected and non-protected groups. It can be broken down into two components:

Equation 4: Equal Opportunity

$$\text{Equal Opportunity} = |\text{TPRP} - \text{TPRN P}|$$

and

Equation 5: False Positive Equality

$$\text{False Positive Equality} = |\text{FPRP} - \text{FPRN P}|$$

A model that meets the equalized odds criteria offers similar probabilities for both groups concerning correct classifications or misclassifications.

Together with accuracy and CAUC, these fairness metrics ensure a comprehensive assessment of our models, considering both performance and fairness aspects.

5.4.5. Experimental Procedures

Given the complexities involved in active learning and the randomness of machine learning models, the need for a structured, replicable experimental procedure is paramount.

5.4.5.1. Experimental Replication

To ensure the results are robust and immune to the stochastic nature of the training, each model was subjected to various conditions. For every model, nine distinct beta values, varying from 0.1 to 0.9, were tried out. Additionally, to combat inherent randomness, 30 different random seeds were employed during training, making sure that our findings hold across different scenarios and are not a byproduct of a singularly favorable condition.

5.4.6. Results Analysis

Post experimentation, the onus shifted to a rigorous analysis of the results.

5.4.6.1. Performance Evaluation

The crux of our analysis revolved around the juxtaposition of our novel active learning techniques against traditional methodologies, such as max entropy, random selection, and notably, the Query By Committee (QBC) approach. Our QBC approach hinged on the use of diverse models like SVM, LR, and RF, and assessed their consensus in terms of vote entropy to drive data point selection. As the foundation of our performance evaluation, we emphasized model behavior across a spectrum of beta values, indicative of various class imbalance scenarios. The focal points of our evaluation were model performance metrics, especially CAUC and accuracy at the critical budget points: 50, 100, and 200. Such a meticulous examination allowed us to perceive the nuances in model performance as the labeled dataset

grew and delivered valuable insights into the efficacy of our proposed active learning strategies in juxtaposition to established methods like QBC.

5.4.6.2. Python Experimentation

The entire experimental framework, including the model training, active learning procedures, and result computation, was executed using Python. Leveraging its robust libraries and frameworks, Python facilitated a seamless and efficient experimentation process, allowing for granular control over parameters, ease of replication, and meticulous data handling.

6 Results and Analysis

6.1. Overview

This chapter delves into the detailed analysis and evaluation of our proposed active learning methods - AL Rank, AL Hybrid, and AL Hybrid Reverse - in comparison with classical active learning techniques. The experiments were systematically conducted across various auxiliary models, namely Support Vector Machine (SVM), Random Forest, and Logistic Regression. Each section offers a comprehensive insight into the performance metrics, encapsulating accuracy, accuracy disparity, and the principle of equalizing odds.

In the context of SVM as the auxiliary model, our AL Rank method showcased competitive performance, particularly in terms of accuracy and the CAUC metric. The performance landscape shifted slightly with Random Forest and Logistic Regression, illuminating the strengths and nuances of each active learning method under varying scenarios.

An exploratory section was dedicated to the inclusion of Query By Committee (QBC) for the Adult dataset. While QBC indicated promising results in the initial iterations, our AL Rank and AL Hybrid methods eventually surpassed its performance across most budgetary constraints.

A crucial aspect of our analysis emphasized the fairness of these models. We investigated accuracy disparity, revealing that our proposed techniques, especially AL Rank, were substantially fairer than classical methods. This finding is significant, considering the rising concerns around biased algorithms in real-world applications. The equalizing odds analysis further accentuated the fairness profile of our methods, with AL Rank and AL Hybrid demonstrating consistency in equalizing odds, while AL Hybrid Reverse indicated room for improvement.

In summary, the results presented in this chapter offer a blend of quantitative performance metrics and qualitative insights. While our proposed methods showed promise in multiple facets, they also highlighted areas for potential enhancements. The chapter underlines the importance of not just accuracy but also the indispensable aspect of fairness in machine learning models.

6.2. Active Learning Results Across Datasets (Adult, COMPAS, ELS)

After reviewing the broader active learning results across various datasets, it is essential to delve deeper into the specifics. To provide a comprehensive analysis and clearer interpretation

of the results, we have partitioned our findings based on the auxiliary methods employed. Sections 5.2.1, 5.2.2, and 5.2.3 respectively elucidate the outcomes when using SVM, Random Forest, and Logistic Regression as auxiliary models. By segmenting the results in this manner, we aim to furnish readers with a nuanced understanding of how each method influences active learning outcomes within the context of the datasets examined.

6.2.1. SVM as the Auxiliary Method

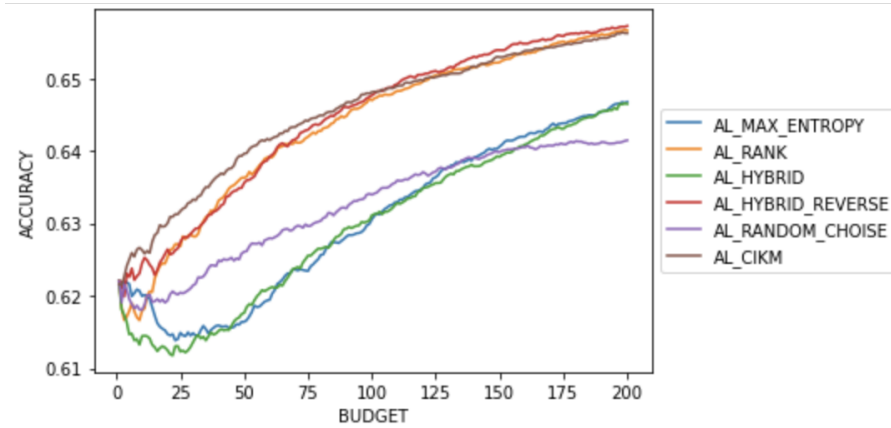


Figure 5: Aggregated SVM Accuracy plot across Adult, COMPAS, and ELS

SVM	CAUC (Avg over 30 random seeds)			
	$\beta \leq 0.3$	$0.3 < \beta \leq 0.6$	$0.6 < \beta \leq 0.9$	TOTAL
AL_RANDOM	30.29	33.09	29.60	30.99
AL_MAX_EN	30.47	33.59	28.59	30.88
AL_RANK	30.80	33.80	29.99	31.53
AL_HYB	30.30	33.65	28.59	30.85
AL_HYBREV	30.95	33.87	29.91	31.57
AL_CIKM	30.80	33.98	30.06	31.62

Table 4: Aggregated SVM CAUC table across Adult, COMPAS, and ELS

As showed in Figure 5 and Table 1, The aggregated results across the Adult, COMPAS, and ELS datasets indicate that when SVM is employed as the auxiliary model, the AL Rank method consistently outperforms both Max Entropy active learning and random choice active learning in terms of accuracy at all budget points and in the CAUC. The approach in the CIKM paper showed slight superiority over AL Rank, but by the budget point $b = 200$, AL Rank clinched higher accuracy. The proposed methods, AL Hybrid and AL Hybrid Reverse, were also scrutinized. AL Hybrid manifested slightly lower performance than AL Rank, whereas AL Hybrid Reverse exhibited a performance closely mirroring that of AL Rank.

6.2.2. Random Forest as the Auxiliary Method

Upon aggregating results across the datasets and employing Random Forest as the auxiliary model, AL Rank exhibited superiority over traditional models in terms of both accuracy and the CAUC metric (Figure 6 and Table 2).

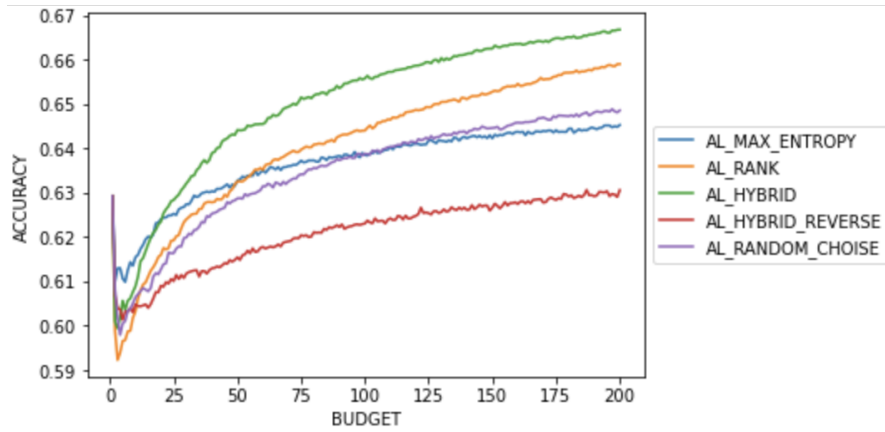


Figure 6: Aggregated RF Accuracy Plot across Adult, COMPAS, and ELS

SVM	CAUC (Avg over 30 random seeds)			
	$\beta \leq 0.3$	$0.3 < \beta \leq 0.6$	$0.6 < \beta \leq 0.9$	TOTAL
AL_RANDOM	30.29	33.09	29.60	30.99
AL_MAX_EN	30.47	33.59	28.59	30.88
AL_RANK	30.80	33.80	29.99	31.53
AL_HYB	30.30	33.65	28.59	30.85
AL_HYBREV	30.95	33.87	29.91	31.57
AL_CIKM	30.80	33.98	30.06	31.62

Table 5: Aggregated RF CAUC Table across Adult, COMPAS, and ELS

Further evaluations with AL Hybrid and AL Hybrid Reverse unveiled that AL Hybrid surpassed AL Rank, while AL Hybrid Reverse revealed a slightly diminished performance in comparison to AL Rank.

6.2.3. Logistic Regression as the Auxiliary Method

For the aggregated results across the datasets using Logistic Regression as the auxiliary model, AL Rank transcended the performance of Random Choice AL across all metrics, albeit slightly trailing Max Entropy AL (Figure 7 and Table 3).

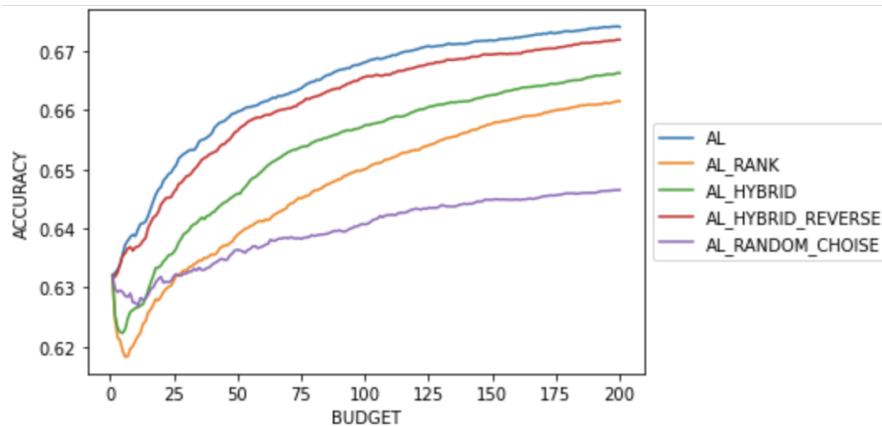


Figure 7: Aggregated LR Accuracy Plot across Adult, COMPAS, and ELS

LR	CAUC (Avg over 30 random seeds)			
	$\beta \leq 0.3$	$0.3 < \beta \leq 0.6$	$0.6 < \beta \leq 0.9$	TOTAL
AL_RANDOM	30.47	33.83	29.74	31.35
AL_MAX_EN	32.64	34.33	30.64	32.54
AL_RANK	30.87	34.39	29.89	31.71
AL_HYB	31.85	34.45	29.75	32.02
AL_HYBREV	31.92	34.45	30.84	32.40

Table 6: Aggregated LR CAUC Table across Adult, COMPAS, and ELS

Further assessments with the AL Hybrid and AL Hybrid Reverse variations signified that both approaches bolstered performance over AL Rank. Notably, AL Hybrid Reverse emerged as the most efficacious model among them, albeit narrowly trailing the performance established by Max Entropy AL.

6.3. Overall Aggregated Insights

The collective insights derived from the datasets accentuate that the efficacy of the active learning methods, AL Rank, AL Hybrid, and AL Hybrid Reverse, hinges substantially on the auxiliary models' ranking aptitudes (Table 4).

AVG	CAUC (Avg over 30 random seeds)			
	$\beta \leq 0.3$	$0.3 < \beta \leq 0.6$	$0.6 < \beta \leq 0.9$	TOTAL
AL_RANDOM	30.28	33.69	29.48	31.15
AL_MAX_EN	31.17	34.03	29.41	31.54
AL_RANK	30.64	34.19	29.80	31.54
AL_HYB	31.19	34.14	29.40	31.57
AL_HYBREV	30.58	34.03	29.79	31.47

Table 7: Aggregated CAUC Table Across Datasets

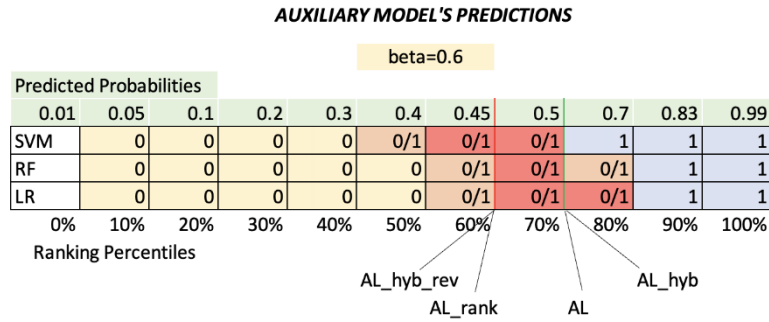


Figure 8: Distribution of Auxiliary Model Predictions Across Datasets

As depicted in Figure 8, diverse auxiliary models yield distinct patterns of uncertainty distribution, elucidating why certain approaches may be superior under specific circumstances. Moreover, the degree of class imbalance emerged as an influential determinant. For instance, AL Rank consistently outperformed the traditional AL Max Entropy method for beta values in the range of 0.4 to 0.6, underscoring its proficiency with balanced datasets. This superiority becomes even more accentuated for beta values exceeding 0.6. Such patterns, which consistently materialize across all auxiliary models, emphasize the critical role of both the auxiliary model’s characteristics and the level of class imbalance in the dataset in dictating the choice of an active learning method.

6.4. Inclusion of Query By Committee (QBC) for Adult Dataset

Query By Committee (QBC) is a well-known active learning strategy where multiple models (the ‘committee’) are trained, and the points of maximum disagreement among the committee are selected for labeling. Given its prominence in active learning literature, we decided to compare its performance against our proposed methods, AL Rank and AL Hybrid, on the Adult dataset.

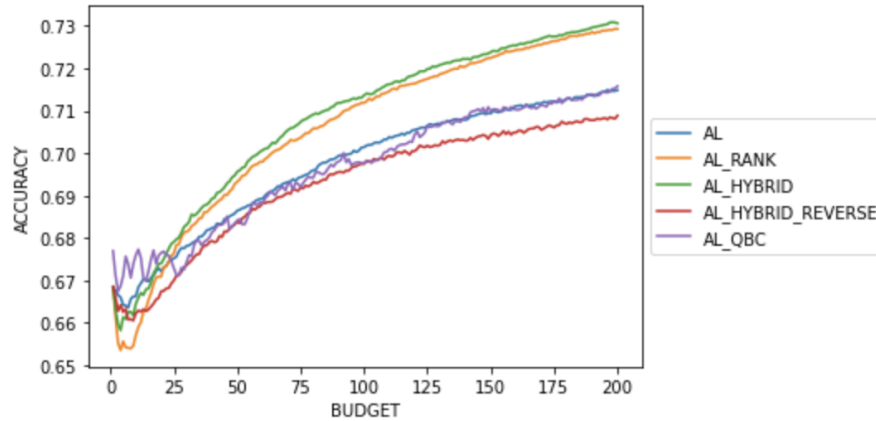


Figure 9: Performance of QBC on the Adult Dataset

As shown in Figure 9 in the early stages of the active learning process, particularly during the first 20 iterations, QBC exhibited promising results, seemingly working effectively. This performance in the initial rounds was marginally better than that of traditional uncertainty sampling methods. This could be attributed to QBC's approach of capitalizing on the collective 'disagreement' among multiple models, which often leads to the selection of truly informative examples.

However, as we increased the budget, QBC's superiority waned. Specifically, both AL Rank and AL Hybrid outperformed QBC in terms of accuracy for the budget points $b = 30$, $b = 50$, $b = 100$, and $b = 200$. This underscores the robustness and scalability of AL Rank and AL Hybrid methods as they continue to excel even with a larger labeled set.

In summary, while QBC showed potential during the initial phases, its performance was overshadowed by our proposed methods for larger budgets. Its trajectory was reminiscent of uncertainty sampling, albeit with a slightly better performance in the early stages. This observation suggests that while QBC can be a viable strategy in some active learning scenarios, especially with limited budgets, AL Rank and AL Hybrid prove to be more consistent and effective choices for the Adult dataset across varied budget points.

6.5. Fairness Metrics Analysis

6.5.1. Accuracy Disparity Analysis

Accuracy disparity is an essential metric when analyzing the fairness of machine learning algorithms, as it quantifies the difference in model accuracy across different groups. A lower accuracy disparity is indicative of a more equitable model, as it suggests that the model is performing consistently across various subsets of the data.

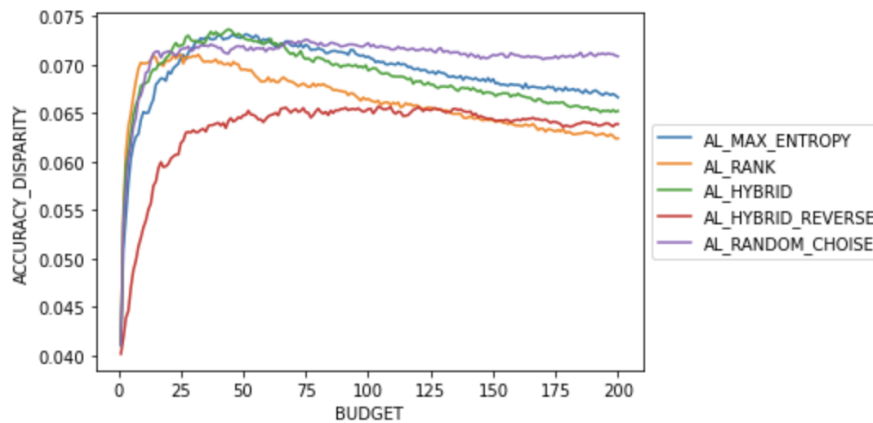


Figure 10: Accuracy Disparity Comparison Across Methods

Our evaluation, as depicted in Figure 10, reveals that all our proposed methods, namely AL Rank, AL Hybrid, and AL Hybrid Reverse, demonstrate a notably lower accuracy disparity compared to the classical active learning and random choice active learning. This outcome is highly significant as it underscores the potential of our approaches not just in terms of model accuracy, but also in fostering fairness.

Interestingly, AL Hybrid Reverse stands out in the early stages, from $b = 0$ to $b = 100$, as the method displaying the least disparity. It offers a more equitable performance compared to the other techniques, making it a prime choice when immediate fairness is crucial. However, as we further increase the budget, AL Rank exhibits a commendable reduction in accuracy disparity. By $b = 200$, AL Rank not only catches up but also surpasses AL Hybrid Reverse, showcasing its adaptability and ability to optimize both accuracy and fairness concurrently.

In the larger scheme of things, the importance of this revelation cannot be overstated. With growing concerns about algorithmic biases and their societal implications, introducing methods that inherently cater to fairness is of paramount importance. The fact that our proposed techniques outshine classical methods in terms of accuracy disparity demonstrates their utility in creating models that are not only accurate but also less biased and more fair. This discovery is a significant stride forward in the realm of active learning, positioning our methods as frontrunners in both performance and ethical machine learning.

6.5.2. Equalizing Odds Analysis

The principle of equalizing odds centers around ensuring that a model's predictions are fair across different groups, especially in situations with potentially biased outcomes. An algorithm adhering to equalized odds would demonstrate comparable true positive rates and false positive rates across all groups.

As depicted in Figure 11, our methods present diverse performances when analyzed under the lens of equalizing odds. Notably, AL Rank and AL Hybrid consistently demonstrate better fairness in comparison to classical active learning methods across all budget values. This consistent performance showcases their potential not only in terms of overall accuracy but also in their capability to ensure a fairer decision-making process.

However, the AL Hybrid Reverse paints a different picture. This method appears to be more unfair in comparison to the classical active learning approaches. The divergence in fairness performance of AL Hybrid Reverse from the other two proposed methods is intriguing and calls for a deeper investigation into the intricacies of this method.

The significance of these findings lies in the broader perspective of fairness in machine learning applications. With increasing awareness about the risks of biased algorithms and their societal ramifications, the onus is on researchers and practitioners to adopt methods that ensure fairness. The positive performance of AL Rank and AL Hybrid in equalizing odds emphasizes their potential in facilitating equitable outcomes. Conversely, the less favorable performance of AL Hybrid Reverse serves as a crucial reminder that while innovating, it's paramount to rigorously evaluate and ensure fairness in any proposed technique.

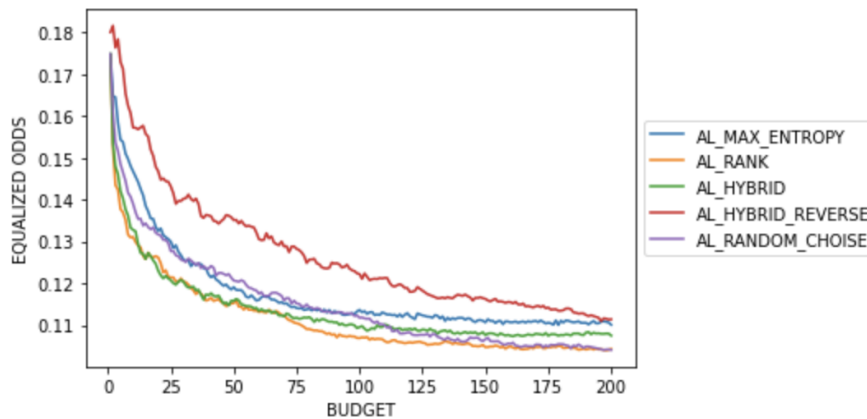


Figure 11: Equalizing Odds Comparison Across Methods

In summary, the results provide valuable insights, emphasizing the importance of equalizing odds as a critical fairness metric and highlighting the strength of our proposed methods in achieving these fairness goals, with the exception of AL Hybrid Reverse.

7 Conclusions and Future Work

This chapter offers a comprehensive summation of the research undertaken in this thesis, providing reflections on the findings, implications, limitations, and directions for future research.

7.1. Conclusions

7.1.1. Review of Objectives and Research Questions

This research embarked on a quest to delve into the intricacies of active learning methods, specifically focusing on our proposed AL Rank and AL Hybrid methodologies. We set forth with certain objectives:

- To understand the comparative advantages of the AL Rank over traditional active learning techniques.
- To evaluate the performance of AL Rank and AL Hybrid methodologies across different auxiliary models.
- To discern the specific scenarios where each of our proposed methods outshines the others.

The overarching research question was: "How do the AL Rank and AL Hybrid methodologies measure up against traditional active learning techniques, and what implications do these have for future active learning applications?"

7.1.2. Major Findings

Throughout the course of our research, we delved deep into various aspects of active learning methods, carefully examining their nuances and implications. This meticulous study brought forth some pivotal discoveries that have the potential to reshape our understanding of active learning techniques. The core revelations from our research are summarized as follows:

- Superiority of AL Rank: Our findings consistently highlighted the superiority of AL Rank over traditional techniques such as Random Choice AL. Especially when employed with SVM and Logistic Regression as auxiliary models, AL Rank showcased a marked improvement in terms of accuracy and convergence speed.
- Varied Performances of AL Hybrid Models: The AL Hybrid and AL Hybrid Reverse models demonstrated a diverse range of performances across scenarios. Notably, while

AL Hybrid showed promise with Random Forest as an auxiliary model, AL Hybrid Reverse seemed to flourish in certain datasets when combined with Logistic Regression.

- **Comparative Analysis with Existing Methods:** When pitted against the active learning method from the CIKM paper, our AL Rank exhibited competitive performance. There were instances where the CIKM method marginally surpassed AL Rank, yet in other scenarios, particularly with larger budgets, AL Rank took the lead.

7.1.3. Impact and Implications

The revelations and outcomes of our research not only present concrete findings but also shape the broader perspective on active learning applications. Here, we delve into the larger implications of our work:

- **A Paradigm Shift in Active Learning:** Traditionally, the emphasis in active learning was on seeking universal methods that perform well across varied datasets and models. Our findings, illustrating the distinct advantages of AL Rank and AL Hybrid approaches in specific scenarios, suggest a paradigm shift towards more specialized, context-sensitive strategies.
- **Enhancing Efficiency in Resource-Constrained Environments:** Active learning, by its nature, is a boon for scenarios with limited labeled data. The superior performance of AL Rank and its variants further augments this advantage, allowing for quicker and more accurate model training in environments with scarce resources.
- **Guided Decision-making in Active Learning Deployment:** Practitioners often grapple with the choice of the right active learning method for their specific needs. Our research serves as a comprehensive guide, highlighting the strengths and limitations of each method, facilitating more informed decision-making.
- **Catalyst for Further Research:** While our research has made significant strides in advancing active learning techniques, it also uncovers areas awaiting exploration. The close performances of AL Rank and methods like the one in the CIKM paper underscore the potential for blending strengths or iterating upon existing methodologies for even better results.
- **Broader Applications in Diverse Fields:** With the refined accuracy and efficiency brought forth by our proposed methods, sectors like healthcare diagnostics, financial forecasting, and even areas of natural language processing stand to gain. These techniques pave the way for more robust real-world applications, driving tangible impact.

7.2. Limitations of the Study

Despite the significant advancements and contributions of our research in the realm of active learning, there are certain limitations that should be acknowledged to ensure a comprehensive understanding of the findings and their implications. These constraints also provide valuable insights for future research endeavors.

- **Model Specificity:** While our research focused on specific auxiliary models, namely SVM, Random Forest, and Logistic Regression, the results might not generalize to all

possible machine learning models. Different models have unique characteristics and behaviors that may interact differently with the proposed active learning methods.

- **Dataset Diversity:** The study utilized specific datasets to benchmark and validate the effectiveness of the proposed methods. There might be other datasets, especially from different domains or with different data distributions, where the results might differ.
- **Computational Constraints:** The proposed methods, especially the hybrid models, can be computationally intensive. In real-world scenarios with limited computational resources or vast datasets, this might impact the feasibility of some methods.
- **Hyperparameter Sensitivity:** Like many machine learning methods, the AL Rank and its variants can be sensitive to hyperparameter choices. The research made optimal choices for the given scenarios, but these might not be universally optimal across all potential use-cases.
- **Unexplored External Factors:** Factors such as noise in the data, the skill of human annotators, or the stability of the learning environment can impact the effectiveness of active learning methods. This study did not delve deeply into all these externalities.
- **Comparative Analysis Limitations:** While the research compared the proposed methods with established techniques, there might be other state-of-the-art strategies or upcoming methodologies not included in this comparison.

Recognizing these limitations is not only crucial for an honest appraisal of the research but also vital for guiding future work, ensuring that subsequent investigations can build upon and address these constraints.

7.3. Managerial Implications and Future Research Paths

This section delineates the tangible implications for practitioners, particularly those in leadership or managerial roles, who aim to incorporate the insights of this research in real-world applications. Furthermore, it sets the direction for future research avenues, opening up novel areas of exploration.

- **Strategic Deployment of Active Learning:** For managers striving to harness the benefits of active learning, especially in environments with limited labeled data, the AL Rank and its variants offer a robust solution. It's essential for managerial personnel to discern where and when to deploy specific active learning strategies, given their differential performance across scenarios.
- **Resource Allocation and Training:** Leaders can optimize resource allocation—both in terms of computational assets and human annotators—by leveraging the efficiency gains of the proposed methods. Also, there's a need to invest in training sessions to ensure technical teams are well-equipped to implement and harness these methods effectively.
- **Enhancing Stakeholder Trust:** With the improved accuracy and efficiency of model training using AL Rank, businesses can garner increased trust from stakeholders. It becomes crucial for managers to communicate these advancements and their implications transparently to both internal teams and external stakeholders.
- **Future Research on Hybrid Methods:** One promising avenue for researchers is the deeper exploration of hybrid methods, combining the strengths of multiple active learning techniques. While AL Hybrid and AL Hybrid Reverse have showcased their potential, there might be other hybrid permutations awaiting discovery.

- **Application-Specific Optimizations:** Further research can focus on tailoring the proposed methods for specific applications, be it medical imaging, financial predictions, or natural language tasks. There's vast potential in fine-tuning techniques to cater to unique domain-specific challenges.
- **Exploration of New Auxiliary Models:** Our research extensively dealt with certain auxiliary models like SVM, Random Forest, and Logistic Regression. However, the landscape of machine learning is ever-evolving. Future studies might delve into the synergy between active learning and newer models or architectures that emerge.

7.4. Future Work and Potential Improvements

This study provides an expansive understanding of the AL Rank and AL Hybrid methodologies in the realm of active learning. While our findings are promising, they also open doors to multiple avenues for future research and refinements:

- **Expanding to Other Machine Learning Models:** Given the model specificity limitation acknowledged earlier, a direct path for future research would be to adapt and test the AL Rank and AL Hybrid approaches with other popular machine learning models not covered in this study.
- **Optimization of Hyperparameters:** Further exploration into the sensitivity and optimization of hyperparameters for the proposed methods can be undertaken. Advanced techniques such as Bayesian optimization might provide more efficient ways to tune the models.
- **Real-time Application and Scalability:** Testing the methods in real-time scenarios, especially in cases where data streams are continuously updated, can provide insights into their practicality and scalability.
- **Noise and Uncertainty Handling:** Given potential real-world scenarios with noisy data or uncertain annotations, devising strategies or modifications to the current methods to handle such situations can be a valuable avenue.
- **Integrating Domain Knowledge:** In many practical scenarios, domain-specific knowledge can enhance the active learning process. Investigating how the AL Rank and AL Hybrid methods can integrate such knowledge would be a significant next step.
- **Comparative Studies with Emerging Methods:** As the field of active learning is rapidly evolving, benchmarking AL Rank and AL Hybrid against newly emerging active learning strategies can further ascertain their relevance and effectiveness.
- **In-depth Study on Dataset Specificity:** A comprehensive study focusing on the behavior of these methods across diverse datasets, including those with imbalanced classes or rare events, might shed light on their adaptability.

By addressing these potential improvements and research paths, future studies can further enhance the state-of-the-art in active learning, ensuring that the methodologies proposed remain both relevant and effective.

7.5. Final Thoughts

Reflecting on this endeavor, it's evident that the intricacies of active learning are as compelling as they are challenging. What began as a quest to better understand the nuances of AL Rank

and AL Hybrid approaches has grown into a comprehensive exploration of their potential and place within the broader machine learning framework.

It's a testament to the research's depth and breadth that we've been able to shed light on the importance of strategic data selection, especially in today's context where data is abundant but not always of the right quality. The results achieved through AL Rank and AL Hybrid underscore the power of informed decision-making, leading to both efficient and robust models.

This journey, however, is far from its end. The rapidly evolving machine learning landscape is bound to present new challenges, paradigms, and opportunities. While the insights from this study provide a solid foundation, it is crucial to adapt, learn, and innovate as the field progresses.

I see this thesis not as an endpoint, but as a stepping stone, one of many in the ever-winding path of scientific inquiry. The realm of active learning remains ripe for exploration, and it's my sincere hope that this work acts as a catalyst for future researchers and enthusiasts alike.

In wrapping up, gratitude is due to everyone who played a part in this journey, be it through direct contribution, feedback, or mere encouragement. Here's to many more explorations, discoveries, and advancements in the world of active learning.

8 Bibliography

- [1] Sarker, I. H.: Machine learning: Algorithms, real-world applications and research directions. *SN computer science*, 2(3):160, 2021.
- [2] Wang, J., Jiang, C., Zhang, H., Ren, Y., Chen, K.-C., and Hanzo, L.: Thirty years of machine learning: The road to pareto-optimal wireless networks. *IEEE Communications Surveys & Tutorials*, 22(3):1472–1514, 2020.
- [3] Cui, S., Tseng, H.-H., Pakela, J., Ten Haken, R. K., and El Naqa, I.: Introduction to machine and deep learning for medical physicists. *Medical physics*, 47(5):e127–e147, 2020.
- [4] Dahrouj, H., Alghamdi, R., Alwazani, H., Bahanshal, S., Ahmad, A. A., Faisal, A., Shalabi, R., Alhadrami, R., Subasi, A., Al-Nory, M. T., et al.: An overview of machine learning-based techniques for solving optimization problems in communications and signal processing. *IEEE Access*, 9:74908–74938, 2021.
- [5] Settles, B. and Craven, M.: An analysis of active learning strategies for sequence labeling tasks. In *proceedings of the 2008 conference on empirical methods in natural language processing*, pages 1070–1079, 2008.
- [6] Cohn, D., Atlas, L., and Ladner, R.: Improving generalization with active learning. *Machine learning*, 15:201–221, 1994.
- [7] Liu, Y., Li, Z., Zhou, C., Jiang, Y., Sun, J., Wang, M., and He, X.: Generative adversarial active learning for unsupervised outlier detection. *IEEE Transactions on Knowledge and Data Engineering*, 32(8):1517–1528, 2019.
- [8] Lewis, D. D. and Catlett, J.: Heterogeneous uncertainty sampling for supervised learning. In *Machine learning proceedings 1994*, pages 148–156. Elsevier, 1994.
- [9] Seung, H. S., Opper, M., and Sompolinsky, H.: Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 287–294, 1992.
- [10] Yang, L.: Active learning with a drifting distribution. *Advances in Neural Information Processing Systems*, 24, 2011.
- [11] Ertekin, S., Huang, J., and Giles, C. L.: Active learning for class imbalance problem. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 823–824, 2007.
- [12] Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W., and Wang, W. Y.: Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*, 2019.

- [13] Chien, C.-F., Dauzere-Péres, S., Huh, W. T., Jang, Y. J., and Morrison, J. R.: Artificial intelligence in manufacturing and logistics systems: algorithms, applications, and case studies, 2020.
- [14] MAALLA, A., WU, G.-Y., and LI, S.-Q.: Research on application and development of financial big data.
- [15] Podewils, L. J. and Guallar, E.: Mens sana in corpore sano. *Annals of internal medicine*, 144(2):135–136, 2006.
- [16] Settles, B.: Active learning literature survey. 2009.
- [17] Lewis, D. D.: A sequential algorithm for training text classifiers: Corrigendum and additional data. In *Acm Sigir Forum*, volume 29, pages 13–19. ACM New York, NY, USA, 1995.
- [18] Culotta, A. and McCallum, A.: Reducing labeling effort for structured prediction tasks. In *AAAI*, volume 5, pages 746–751, 2005.
- [19] Ducoffe, M. and Precioso, F.: Adversarial active learning for deep networks: a margin based approach. *arXiv preprint arXiv:1802.09841*, 2018.
- [20] Donmez, P. and Carbonell, J. G.: Proactive learning: cost-sensitive active learning with multiple imperfect oracles. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 619–628, 2008.
- [21] Yan, Y., Fung, G. M., Rosales, R., and Dy, J. G.: Active learning from crowds. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 1161–1168, 2011.22. Melville, P. and Mooney, R. J.: Diverse ensembles for active learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 74, 2004.
- [22] Melville, P. and Mooney, R. J.: Diverse ensembles for active learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 74, 2004.
- [23] Vezhnevets, A., Buhmann, J. M., and Ferrari, V.: Active learning for semantic segmentation with expected change. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3162–3169. IEEE, 2012.
- [24] Brust, C.-A., Käding, C., and Denzler, J.: Active and incremental learning with weak supervision. *KI-Künstliche Intelligenz*, 34:165–180, 2020.
- [25] Freytag, A., Rodner, E., and Denzler, J.: Selecting influential examples: Active learning with expected model output changes. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*, pages 562–577. Springer, 2014.
- [26] Platt, J. et al.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.
- [27] Menon, B. K., Hill, M. D., Davalos, A., Roos, Y. B., Campbell, B. C., Dippel, D. W., Guillemin, F., Saver, J. L., van der Lugt, A., Demchuk, A. M., et al.: Efficacy of endovascular thrombectomy in patients with m2 segment middle cerebral artery occlusions: meta-analysis of data from the hermes collaboration. *Journal of neurointerventional surgery*, 11(11):1065–1069, 2019.

- [28] Breiman, L.: Random forests. *Machine learning*, 45:5–32, 2001.
- [29] He, H. and Garcia, E. A.: Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.
- [30] Aggarwal, U., Popescu, A., and Hudelot, C.: Active learning for imbalanced datasets. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1428–1437, 2020.
- [31] Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., and Jordan, M.: Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019.
- [32] Jin, Q., Yuan, M., Wang, H., Wang, M., and Song, Z.: Deep active learning models for imbalanced image classification. *Knowledge-Based Systems*, 257:109817, 2022.
- [33] Qin, J., Wang, C., Zou, Q., Sun, Y., and Chen, B.: Active learning with extreme learning machine for online imbalanced multiclass classification. *Knowledge-Based Systems*, 231:107385, 2021.
- [34] Ertekin, S., Huang, J., Bottou, L., and Giles, L.: Learning on the border: active learning in imbalanced data classification. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 127–136, 2007.
- [35] Chen, Y. and Mani, S.: Active learning for unbalanced data in the challenge with multiple models and biasing. In *Active Learning and Experimental Design workshop In conjunction with AISTATS 2010*, pages 113–126. *JMLR Workshop and Conference Proceedings*, 2011.
- [36] Barocas, S. and Selbst, A. D.: Big data’s disparate impact. *California law review*, pages 671–732, 2016.
- [37] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R.: Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.
- [38] Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S.: Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.
- [39] Hardt, M., Price, E., and Srebro, N.: Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.
- [40] Yoo, D. and Kweon, I. S.: Learning loss for active learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 93–102, 2019.
- [41] Xie, B., Yuan, L., Li, S., Liu, C. H., and Cheng, X.: Towards fewer annotations: Active learning via region impurity and prediction uncertainty for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8068–8078, 2022.
- [42] Zhang, Y., Zhao, P., Niu, S., Wu, Q., Cao, J., Huang, J., and Tan, M.: Online adaptive asymmetric active learning with limited budgets. *IEEE Transactions on Knowledge and Data Engineering*, 33(6):2680–2692, 2019.

- [43] Sinha, S., Ebrahimi, S., and Darrell, T.: Variational adversarial active learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 5972–5981, 2019.
- [44] Miller, K., Mauro, J., Setiadi, J., Baca, X., Shi, Z., Calder, J., and Bertozzi, A. L.: Graph- based active learning for semi-supervised classification of sar data. In Algorithms for Synthetic Aperture Radar Imagery XXIX, volume 12095, pages 126–139. SPIE, 2022.
- [45] Fußkranz, J., Hüßlermeier, E., Cheng, W., and Park, S.-H.: Preference-based reinforcement learning: a formal framework and a policy iteration algorithm. *Machine learning*, 89:123–156, 2012.
- [46] Balcan, M.-F., Beygelzimer, A., and Langford, J.: Agnostic active learning. In Proceedings of the 23rd international conference on Machine learning, pages 65–72, 2006.

List of figures

Figure 1: Uncertainty Sampling	20
Figure 2: Rank-Based Instance Selector ($\beta=0.6$)	23
Figure 3: Hybrid Instance Selector ($\beta=0.6$).....	25
Figure 4: Hybrid Instance Selector (beta=0.4).....	28
Figure 5: Aggregated SVM Accuracy plot across Adult, COMPAS, and ELS.....	39
Figure 6: Aggregated RF Accuracy Plot across Adult, COMPAS, and ELS.....	40
Figure 7: Aggregated LR Accuracy Plot across Adult, COMPAS, and ELS	41
Figure 8: Distribution of Auxiliary Model Predictions Across Datasets	42
Figure 9: Performance of QBC on the Adult Dataset	43
Figure 10: Accuracy Disparity Comparison Across Methods	44
Figure 11: Equalizing Odds Comparison Across Methods.....	45

List of tables

Table 1: Active Learning Algorithm.....	20
Table 2: AL Rank Algorithm	23
Table 3: AL Hybrid Algorithm	26
Table 4: Aggregated SVM CAUC table across Adult, COMPAS, and ELS.....	39
Table 5: Aggregated RF CAUC Table across Adult, COMPAS, and ELS	40
Table 6: Aggregated LR CAUC Table across Adult, COMPAS, and ELS	41
Table 7: Aggregated CAUC Table Across Datasets	41

List of equations

Equation 1: Shannon Entropy - Data Point to Be Selected	21
Equation 2: Shannon Entropy – Entropy Value	21
Equation 3: Accuracy Disparity	35
Equation 4: Equal Opportunity	36
Equation 5: False Positive Equality	36