

REPUBLIC OF TÜRKİYE
YILDIZ TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

**WEIGHTED XGBOOST BASED ACTIVE
LEARNING FRAMEWORK FOR FRAUD
DETECTION WITH USING SMALL NUMBER OF
SAMPLES**

Mehmet Onur ERBOY

MASTER OF SCIENCE THESIS
Department of Computer Engineering
Program of Computer Engineering

Supervisor
Assoc. Prof. Dr. Ali Can KARACA

November, 2024

REPUBLIC OF TÜRKİYE
YILDIZ TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

WEIGHTED XGBOOST BASED ACTIVE LEARNING
FRAMEWORK FOR FRAUD DETECTION WITH USING
SMALL NUMBER OF SAMPLES

A thesis submitted by Mehmet Onur ERBOY in partial fulfillment of the requirements for the degree of **MASTER OF SCIENCE** is approved by the committee on 22.11.2024 in Department of Computer Engineering, Program of Computer Engineering.

Assoc. Prof. Dr. Ali Can KARACA
Yildiz Technical University
Supervisor

Approved By the Examining Committee

Assoc. Prof. Dr. Ali Can KARACA, Supervisor
Yildiz Technical University

Assoc. Prof. Hamza Osman ILHAN, Member
Yildiz Technical University

Assoc. Prof. Dr. Mehmet Zeki KONYAR, Member
Kocaeli University

I hereby declare that I have obtained the required legal permissions during data collection and exploitation procedures, that I have made the in-text citations and cited the references properly, that I haven't falsified and/or fabricated research data and results of the study and that I have abided by the principles of the scientific research and ethics during my Thesis Study under the title of Weighted XGBoost Based Active Learning Framework for Fraud Detection with using Small Number of Samples supervised by my supervisor, Assoc. Prof. Dr. Ali Can KARACA. In the case of a discovery of false statement, I am to acknowledge any legal consequence.

Mehmet Onur ERBOY

Signature

*Dedicated to my parents
and my sisters brothers
and friends
and my dearest best friend Safak Güney who passed away recently*

ACKNOWLEDGEMENTS

In my master's thesis, my family supported me in every step. So, my first thanks go to them. After them, special thanks stand up to my advisor, Assoc. Prof. Dr. Ali Can KARACA. He helped me in every step of this thesis; he gave me excellent guidance, patience, and encouragement. He helped me throughout this study, so it would not have been possible without his great support and guidance. Then, my work company, Trendyol, supported me by giving me some free time to go to lessons at university and encouraged me to work on my thesis while arranging my work duties. So here, I should thank my leaders, Onur Tařhan, Sergen Eker, and Murat Özcan, and my other colleagues. Also, I thank my close friends who supported me in this thesis research period and gave me good advice. So, I should thank Sarper Kumcu and Yasin Uygun, and special thanks to my recently passed away best friend, řafak Güney.

Mehmet Onur ERBOY

TABLE OF CONTENTS

LIST OF ABBREVIATIONS	vii
LIST OF FIGURES	viii
LIST OF TABLES	x
ABSTRACT	xii
ÖZET	xiv
1 INTRODUCTION	1
1.1 Fundamental Information	1
1.2 Literature Review	3
1.2.1 The Research about Fraud Detection	4
1.2.2 The Research about Active Learning	5
1.3 Motivation	7
1.4 Outlines of Thesis	8
2 ACTIVE LEARNING BASED FRAUD DETECTION METHOD- OLOGY	10
2.1 Fraud Detection	10
2.1.1 Isolation Forest	12
2.1.2 Support Vector Machine	13
2.1.3 Artificial Neural Network	15
2.1.4 XGBoost	16
2.1.5 Weighted XGBoost	18
2.1.6 CatBoost	20
2.2 Uncertainty Sampling Active Learning Strategies	22
2.2.1 Least Confidence	23
2.2.2 Margin Confidence	24
2.2.3 Ratio Confidence	25
2.2.4 Entropy	26
2.3 Diversity Sampling Active Learning Strategy	27
2.3.1 K-Means Cluster-Based Diversity Sampling	28

2.4	Pseudo Labeling	29
2.5	Proposed Method	30
3	DATASETS	33
3.1	Credit Card Fraud Detection Dataset	34
3.2	Credit Card Transactions Fraud Detection Dataset	36
3.3	Vehicle Loan Default Prediction Dataset	38
4	EVALUATIONS	43
4.1	Active Learning Approaches	43
4.1.1	Credit Card Fraud Detection Dataset Tests and Evaluations .	44
4.1.2	Credit Card Transactions Fraud Detection Dataset Tests and Evaluations	48
4.1.3	Vehicle Loan Default Prediction Dataset Tests and Evaluations	51
4.2	Pseudo Labeling with Active Learning Approaches	54
4.2.1	Credit Card Fraud Detection Dataset Tests and Evaluations .	55
4.2.2	Credit Card Transactions Fraud Detection Dataset Tests and Evaluations	58
4.2.3	Vehicle Loan Default Prediction Dataset Tests and Evaluations	62
4.2.4	Overall Evaluation for Pseudo Labeling with Active Learning Approaches over All Datasets	65
5	CONCLUSION	75
	REFERENCES	77
	PUBLICATIONS FROM THE THESIS	80

LIST OF ABBREVIATIONS

AE	Auto Encoder
AL	Active Learning
ANN	Artificial Neural Network
CB	Cluster-Based
DS	Diversity Sampling
EB	Entropy Based
IC3	Internet Crime Compliant Center
IF	Isolation Forest
KNN	K-Nearest Neighbours
LC	Least Confidence
LR	Logistic Regression
MC	Margin Confidence
NN	Neural Network
PCA	Principal Component Analysis
PL	Pseudo Labeling
RC	Ratio Confidence
RF	Random Forest
SVM	Support Vector Machine
US	Uncertainty Sampling
USA	United State of America

LIST OF FIGURES

Figure 2.1	Isolation Forest	13
Figure 2.2	Support Vector Machine	14
Figure 2.3	Artificial Neural Network	16
Figure 2.4	Extreme Gradient Boosting	19
Figure 2.5	The Architecture of Implemented Approach	30
Figure 3.1	t-SNE model of Credit Card Fraud Detection Dataset Sample Example	36
Figure 3.2	t-SNE model of Credit Card Transactions Fraud Detection Dataset Sample Example	37
Figure 3.3	t-SNE model of Vehicle Loan Default Prediction Dataset Sample Example	39
Figure 4.1	F1 Score Evaluation of Best Three Methods of Hybrid Active Learning used LC Uncertainty Sampling for Credit Fraud Detection Dataset	47
Figure 4.2	F1 Score Evaluation of Best Three Methods of Hybrid Active Learning used LC Uncertainty Sampling for Credit Card Transaction Fraud Detection Dataset	50
Figure 4.3	F1 Score Evaluation of Best Three Methods of Hybrid Active Learning used LC Uncertainty Sampling for Credit Fraud Detection Dataset	53
Figure 4.4	Average F1 Score Evaluation of Random, AL, AL + PL, AL + PL w/ DS Tests for IF Method over Tested Three Datasets . . .	66
Figure 4.5	Average F1 Score Evaluation of Random, AL, AL + PL, AL + PL w/ DS Tests for ANN Method over Tested Three Datasets . .	67
Figure 4.6	Average F1 Score Evaluation of Random, AL, AL + PL, AL + PL w/ DS Tests for SVM Method over Tested Three Datasets .	68
Figure 4.7	Average F1 Score Evaluation of Random, AL, AL + PL, AL + PL w/ DS Tests for CatBoost Method over Tested Three Datasets	69
Figure 4.8	Average F1 Score Evaluation of Random, AL, AL + PL, AL + PL w/ DS Tests for XGBoost Method over Tested Three Datasets	69

Figure 4.9	Average F1 Score Evaluation of Random, AL, AL + PL, AL + PL w/ DS Tests for Weighted XGBoost Method over Tested Three Datasets	70
Figure 4.10	Step-by-Step F1 Score Evaluation of Weighted XGBoost Approach Tests for Credit Card Fraud Detection Dataset	72
Figure 4.11	Step-by-Step F1 Score Evaluation of Weighted XGBoost Approach Tests for Credit Card Transactions Fraud Dataset . . .	72
Figure 4.12	Step-by-Step F1 Score Evaluation of Weighted XGBoost Approach Tests for Vehicle Loan Default Prediction Dataset . .	73



LIST OF TABLES

Table 2.1	Pseudo Code of Hybrid Active Learning with Pseudo Labeling for Fraud Detection as Algorithm	31
Table 3.1	Information of Datasets used in Thesis	33
Table 3.2	Single Sample Example for Credit Card Fraud Detection Dataset	35
Table 3.3	Single Sample Example for Credit Card Transactions Fraud Detection Dataset	39
Table 3.4	Single Sample Example for Credit Card Transactions Fraud Detection Dataset	41
Table 4.1	Accuracy, Precision, Recall, and F1 Score of Traditional Hold-out Separation as 30 % Train and 70 % Test for Credit Card Fraud Detection Dataset	44
Table 4.2	F1 Score Evaluation Metrics of Uncertainty Active Learning Using only Tests for Credit Fraud Detection Dataset	45
Table 4.3	F1 Score Evaluation of Hybrid Active Learning Tests for Credit Fraud Detection Dataset	46
Table 4.4	Accuracy, Precision, Recall, and F1 Score of Traditional Hold-out Separation as 30 % Train and 70 % Test for Credit Card Transaction Fraud Detection Dataset	48
Table 4.5	F1 Score Evaluation of Hybrid Active Learning Tests for Credit Card Transactions Fraud Detection Dataset	49
Table 4.6	Accuracy, Precision, Recall, and F1 Score of Traditional Hold-out Separation as 30 % Train and 70 % Test for Vehicle Loan Default Prediction Dataset	52
Table 4.7	F1 Score Evaluation of Hybrid Active Learning Tests for Vehicle Loan Default Prediction Dataset	53
Table 4.8	F1 Score Evaluation of Random, AL, AL + PL and AL + PL w/ DS Tests for Credit Fraud Detection Dataset	56
Table 4.9	F1 Score Evaluation of Random, AL, AL + PL and AL + PL w/ DS Tests for Credit Card Transaction Fraud Detection Dataset . .	60
Table 4.10	F1 Score Evaluation of Random, AL, AL + PL and AL + PL w/ DS Tests for Vehicle Loan Default Prediction Dataset	63

Table 4.11 Average F1 Score Evaluation of Random, AL, AL + PL and AL + PL w/ DS Conducted Tests for Tested Datasets	66
---	----



ABSTRACT

Weighted XGBoost Based Active Learning Framework for Fraud Detection with using Small Number of Samples

Mehmet Onur ERBOY

Department of Computer Engineering
Program of Computer Engineering
Master of Science Thesis

Supervisor: Assoc. Prof. Dr. Ali Can KARACA

Fraud cases are rare but costly and exhibit a continuously changing nature. However, labeled data for fraud cases is scarce and often requires expert input for labeling. Active learning addresses this by selecting the most uncertain and distinguishable samples for expert labeling, enhancing machine learning-based fraud detection. Pseudo-labeling, on the other hand, uses machine learning predictions as actual labels for selected samples. Scope of the thesis, we proposed a method that uses a hybrid active learning approach of Least Confidence uncertainty sampling and K-means cluster-based diversity sampling combined with a hybrid approach with pseudo labeling and diversity sampling using weighted XGBoost as a machine learning method. Evaluations show that our approach achieves nearly identical F1 scores using less than 4% of samples in training, compared to traditional methods requiring 70% of samples. We evaluated combinations of Isolation Forest, Support Vector Machine, Neural Network, XGBoost, and CatBoost with uncertainty-based active learning strategies as Margin Confidence, Ratio Confidence, Entropy and hybrid approaches using cluster-based diversity sampling. We also examined the effect of pseudo-labeling for each combination across three fraud detection datasets. Our proposed approach outperformed all tested methods. Least Confidence uncertainty sampling, coupled with pseudo-labeling and K-means cluster-based diversity sampling, proved to be the most effective strategy for labeling fraud cases within the scope of this thesis.

Keywords: Active learning, fraud detection, weighted XGBoost, least confidence, pseudo labeling



Az Sayıda Örnek Kullanarak Dolandırıcılık Tespiti Yapan Ağırlıklı XGBoost Tabanlı Aktif Öğrenme Methodu

Mehmet Onur ERBOY

Bilgisayar Mühendisliği Anabilim Dalı
Bilgisayar Mühendisliği Programı
Yüksek Lisans Tezi

Danışman: Doç. Dr. Ali Can KARACA

Dolandırıcılık vakaları nadir görülür ve karşılaşıldığında büyük maddi kayıplara yol açar. Ayrıca, sürekli değişen bir yapıya sahiptir. Bu nedenle, dolandırıcılığın tespiti, makine öğrenimi teknikleriyle çözülmesi gereken önemli bir problemdir. Ancak dolandırıcılık vakaları için yeterli etiketli veri bulunmadığından genellikle uzman görüşüne ihtiyaç duyulur. Aktif öğrenme, uzmanların etiketlemesi için en belirsiz ve ayırt edici örnekleri seçerek doğru öğelerin etiketlenmesini sağlar ve makine öğrenimi algoritmalarıyla dolandırıcılığın tespitinde kullanılır. Sözde etiketleme ise makine öğreniminin etiket tahminlerini, belirli örnekler için gerçek etiket olarak kullanır. Bu tezde, en düşük güven belirsizlik örnekleme ve K-means küme tabanlı çeşitlilik örnekleme yönteminin hibrit bir aktif öğrenme yaklaşımını, sözde etiketleme ve çeşitlilik örnekleme ile birleştirerek ağırlıklı XGBoost yöntemini kullandık. Nadir görülen dolandırıcılık vakaları için az veriyle çalışabilecek bir model geliştirdik. Değerlendirmelerimiz, geleneksel % 70 veriyle yapılan eğitim yöntemine kıyasla % 4'ten az veriyle benzer F1 skorlarına ulaşılabilirdiğini gösterdi. İzolasyon Ormanı, Destek Vektör Makinesi, Sinir Ağı, XGBoost ve Catboost yöntemlerini, belirsizlik tabanlı ve küme tabanlı çeşitlilik örnekleme stratejileriyle karşılaştırdık. Üç farklı dolandırıcılık veri kümesi üzerinde yaptığımız testlerde, önerilen yöntem diğer yaklaşımlardan daha iyi performans gösterdi. En düşük güven belirsizlik örnekleme, sözde etiketleme ve K-means çeşitlilik örnekleme yönteminin birleşimi en etkili yöntem olarak belirlendi.

Anahtar Kelimeler: Aktif öğrenme, hile tespiti, ağırlıklı XGBoost, az güven, sözde etiketleme



1

INTRODUCTION

The introductory chapter of this thesis establishes the foundation necessary to understand the effectiveness of integrating active learning (AL) and pseudo labeling(PL) with diversity sampling (DS) in fraud detection. This chapter begins with a Fundamental Information section, which introduces the core concepts underpinning fraud detection systems and the challenges associated with imbalanced datasets. Following this, a Literature Review provides a comprehensive examination of existing research. Within this review, two sub-sections are dedicated to addressing prior work in Fraud Detection and Active Learning. These reviews are essential for contextualizing the current study within the broader field, identifying limitations in existing approaches, and highlighting the potential of combining methodologies to enhance fraud detection.

The Motivation section then articulates the specific challenges that inspire this research, particularly the low prevalence of fraudulent data and the need for more efficient sampling strategies. Finally, the Outlines of Thesis section provides a road map for the document, outlining the structure of each subsequent chapter and its role in addressing the research questions.

1.1 Fundamental Information

In our increasingly interconnected world, there are some everyday things like hot sunny weather on a summer day in northern countries, crowded people after work time, or high money transfer of a bank account when the salary is paid. However, some situations do not adhere to these typical patterns, like traffic jams on Sunday mornings due to unexpected heavy traffic accidents or sudden heavy rain and a big storm appearing just after sunny weather was observed one hour earlier. These kinds of situations are not typical; they are instead referred to as anomalies, events that rarely happen but stand out sharply against normal patterns. Some anomalies may even become malicious, affecting organizations or innocent individuals. Examples

include unexpectedly high volumes of money transfers between bank accounts in the banking sector, massive endpoint calls targeting a website to destabilize it via cyber-attacks, or the posting of fake reviews on e-commerce platforms to manipulate product rankings. Anomalies affecting organizations, such as monetary losses or reputational harm, and anomalies affecting individuals, such as financial losses, legal issues, or prestige damage, are generally referred to as fraud.

In the realm of fraud, a diverse range of cases can be observed. One example is unexpected credit card transactions, which should be labeled as fraud to facilitate further examination and contribute to datasets for research, as noted in studies like [1]. Some money transfers may violate regulatory standards, particularly in anti-money laundering efforts, as discussed in research such as [2]. In e-commerce platforms, sellers may try to gain an unfair advantage by using fake traffic to boost their product listings and attract a larger customer base, a trend analyzed in studies on e-commerce fraud like [3]. The prevalence of fraud and its adverse financial and reputational impacts have spurred extensive research into detection and prevention methodologies. For instance, the Internet Crime Complaint Center (IC3) in the USA reported a total financial loss of approximately 27.6 billion USD over a five-year period from 2018 to 2022, corresponding to more than 3.25 million complaints. Of these, over 10 billion USD in losses were recorded in 2022 alone. Detecting and preventing true fraud cases is therefore essential for both protecting assets and maintaining organizational integrity.

Despite the severity of fraud cases recorded, malicious fraud still represents only a tiny fraction of all transaction events, making fraud detection highly challenging. As noted in research such as [4], fraud detection datasets are often highly imbalanced, with fraudulent transactions constituting a small minority compared to normal ones. In cases where labeled data is limited, expert opinion becomes essential for determining the labels of unlabeled samples. However, despite the value of expert judgment, examining every sample for fraud verification is impractical. Therefore, rather than presenting all samples to experts, effective sampling and visualization techniques are needed, as demonstrated in studies such as [5]. This calls for a well-designed human-computer interaction framework that efficiently involves experts in labeling critical samples.

To support this labeling process efficiently, active learning and pseudo labeling can be particularly valuable. Active learning focuses on selecting the most relevant samples for experts to label, thereby optimizing the use of limited expert resources, as suggested by research in [6]. On the other hand, pseudo labeling, which involves using machine predictions as labels for unlabeled data, can further enhance the labeled dataset by allowing the model to make initial classifications on low-confidence samples. These pseudo labeled samples are then reviewed in conjunction with active learning strategies to improve model accuracy progressively. Together, active learning and pseudo labeling facilitate an effective labeling process, allowing for more balanced and representative datasets in fraud detection.

1.2 Literature Review

The persistent challenge of fraud detection within the landscape of computer science has generated a substantial body of research, especially as digital transactions continue to proliferate globally. Fraud detection, a field traditionally characterized by imbalanced datasets and a high error cost, has necessitated the development of innovative machine learning and data-driven techniques. These techniques enhance detection accuracy while minimizing false positives, which can have considerable financial and reputational consequences. Over time, various approaches, including supervised learning, semi-supervised learning, and unsupervised learning, have been explored extensively within fraud detection research, each seeking to address the inherent challenges posed by this domain.

In recent years, two methodologies—active learning and pseudo-labeling—have garnered particular attention due to their potential for optimizing performance in scenarios with limited labeled data. Active learning, a strategy where the model actively selects the most informative samples for labeling, has shown promise in efficiently dealing with data scarcity, a prevalent issue in fraud detection. This approach aligns well with the resource-intensive nature of labeling fraud cases, where expert input is often required. Similarly, pseudo-labeling, which involves assigning labels to unlabeled data based on model predictions, has expanded training datasets in a semi-supervised learning framework. Pseudo labeling, mainly when applied iteratively, can reinforce model learning, potentially improving detection rates even in low-prevalence scenarios such as fraud detection.

This literature review is structured to examine relevant studies and continued sub-sections thoroughly organized into two primary sections. The first subsection named by "The Research about Fraud Detection" outlines the evolution of machine learning approaches within fraud detection, emphasizing the unique constraints and challenges specific to this field. The second subsection named "The Research about Active Learning" delves into the advancements in active learning and its applications in data-scarce domains, focusing on the techniques and adaptations suited for fraud detection. Together, these sections provide a comprehensive overview of the existing research, setting a foundation for the proposed methodologies in this study.

1.2.1 The Research about Fraud Detection

There are many different sectors and working areas keen on fraud cases. According to a survey conducted about fraud detection [7], general research on fraud detection problems could be categorized into credit card fraud, telecommunication fraud, healthcare fraud, automobile insurance fraud, and online auction fraud over sectors. The essential challenges could also be concept drift, skewed class distribution, and the reduction of large amounts of data. Generally, fraud detection and prevention research suffers from a lack of balanced data. Fraud cases have rapidly changed in nature. We should detect new patterns, and our detected patterns and cases should be counted as a high minority in our datasets. As an example given in the research [3], sellers behave unexpectedly differently about selling products to increase their traffic and listing upper places in e-commerce platforms. To achieve this aim, they behave differently daily because platforms detect their inappropriate behavior patterns, and they change their strategies to reach their desires. Therefore, fraud cases are counted as highly limited by how it has rapidly changed nature day by day despite having no bounded pattern with historic fraud cases. Causes possible money loss probability kept on the future mentioned by [8] they should investigated and detected correctly.

In the e-commerce world, there is research conducted by Ge Zhang and his colleagues [3] shows that a graph neural-network-based approach named eFraudCom could help detect e-commerce platform fraud cases. They compared their methods with convenient machine learning algorithms such as Isolation Forest (IF) and Auto-Encoder (AE). They presented two different approaches to detect fraud cases. One of them tried to detect fraud cases that overlapped previous fraud cases with known patterns. Other approaches tried to detect new fraud cases using a time series-based approach. Some research like [1] studied fraud detection framework which has computational power on the client side with federated

learning approach over fraud detection problem. They investigated the effectiveness of methods like IF, Support Vector Machine (SVM), K-Nearest Neighbour (KNN), and Artificial Neural Network (ANN) in terms of the scope of the research.

In fraud cases, the decision boundary between normal and fraud samples is generally not noticeable. The research conducted by Schrijver and his colleagues [9] shows that fraudulent and known claims generally could not be separated well. They compared the Random Forest (RF), Logistic Regression (LR), and KNN methods in fraud case detection problems in their research with supervised approaches. According to another research, [10], researchers researched credit fraud dataset, and they used an ensemble learning approaches containing a voting structure using SVM, KNN, RF, Bagging, and Boosting classifiers approaches. They pointed out that a voting framework using given approaches helped detect fraud samples over their used dataset. In another research conducted by Hajek and his colleagues [11], many methods compared over the XGBoost algorithm. According to their evaluations, XGBoost outperformed better results compared to the used other methods scope of their research to find fraud cases.

1.2.2 The Research about Active Learning

In the research about fraud detection, there could be many different approaches that we could use to solve the fraud detection problem. Despite many approaches, there are generally constraints such as limited fraud samples, the need for expert opinion to detect fraud cases, and a vast data payload with no label information. Therefore, we needed a wise approach to handle these typical constraints. Active learning could be a helpful approach to handling a limited expert count to detect fraud cases by representing effective samples to label.

A prepared book on human-machine learning interaction using active learning approaches written by Robert Munro [6] describes usable active learning approaches on the active learning side. According to the book, we have highly limited experts to gather data labels and a vast data payload to label. To this end, we should present wisely which sample should be labeled by the expert. By motivating from this aim, several approaches are listed in the book, but the active learning approaches can be classified into two main classes: uncertainty sampling and diversity sampling. Uncertainty sampling (US) focused on which samples we needed clarification on their labels and how I could gather the most informative information when I put it into an expert label. With uncertainty sampling, we tried to find the most unsure samples from the coming cluster. Diversity sampling focused on which data sample cluster has the most distinguished samples and how

I can present the most different contents. This research tried to find the most distinguishable samples because some samples have nearly the same valued content and differ from only 1 or 2 features. With minimal expertise as an expert source, we did not want to show duplicate samples to be labeled as normal or fraudulent while wasting our valuable source. Using active learning approaches from uncertainty and diversity sampling helps us organize our train and test cluster wisely, and this optimization could enrich our gathered success after our training process.

Active learning approaches are highly effective approaches to many different competitive tasks. One of its used environments is a robot arm training process of research conducted by Chen and her colleagues [12]. They prepared a machine learning approach, and they tried to train their model by selecting training data parts with active learning. They pointed out that active learning helped their model succeed, and their robot arm could more successfully avoid possible target crashes. Another research [13] use active learning approach over text classification. They observed that active learning approaches enriched their solution and evaluation metrics. According to the results, active learning entropy-based strategies under uncertainty methods outperformed best. Despite our mentioned uncertainty and diversity sampling, some hybrid approaches stand up to enrich solutions by combining them. Research by Ash and colleagues [14] shows that using hybrid approaches over active learning could be best overall using a single uncertainty approach or diversity approaches. Therefore, a hybrid approach that contains an uncertainty approach in the first place and diversity sampling in the second place could be helpful to enrich solutions.

Active learning can be used to solve many different kinds of problems. One of them is cryptocurrency anomaly detection. Cunha and colleagues [15] conducted an active learning approach to finding anomalies in cryptocurrency transactions. They used RF, LR, and XGboost approaches as machine learning methods. Another research [16] concerns credit card fraud cases detected by some machine learning algorithms helped by active learning approaches. They studied some supervised and unsupervised algorithms. In conclusion, they pointed out that there is an opportunity to increase the accuracy of fraud detection by 5% by combining active learning methods with stochastic semi-supervised learning. So, active learning was an excellent catalyst for finding fraud cases with tremendous success. In another research area of finding anti-money laundering transactions conducted by Labanca and his colleagues [2], active learning enhanced the results using a combination of supervised and unsupervised methods. They measured that active learning improved the detection rate up to 50% and reduced the overall cost by 20%. In the research on [17], they said that they said that active learning helps detect

some labels with expert and model communication with the ensemble learning approach preferred. Another conducted research [18] shows that active learning is a good choice when we have no labeling data content or limited labeled data. They conducted a cold start training procedure in which there are only 2% to 10% labels of data, and it has good evaluation metrics for that circumstance. Another research [19] shows that hybrid approaches using together uncertainty and diversity sampling over fraud detection could overcome traditional machine learning approaches.

Active learning enhanced the results of machine learning algorithms. We learned that the hybrid approach resulted in the best metrics according to research [19]. We see that XGBoost could perform good results over fraud detection problems according to another research [11]. With these aspects, we proposed a method that weighted XGBoost using hybrid active learning approaches over fraud detection datasets. We also observed that IF, SVM, NN, and XGboost are commonly used for fraud detection problems. We aimed to challenge our used methods and approaches over possible combinations. Further part, named Methodology, deep dive into knowledge about our proposed methods and opponents.

1.3 Motivation

According to the research outlined in the Literature Review on fraud detection, active learning, and pseudo labeling, both active learning and pseudo labeling approaches demonstrate significant potential for enhancing fraud detection performance. Given this potential, this study aims to investigate the combined effectiveness of active learning and pseudo labeling in addressing the unique challenges of fraud detection.

One of the primary challenges in fraud detection is the highly imbalanced nature of data, where only a limited subset is labeled as fraudulent compared to the vast amount of normal data. Labeling samples efficiently requires substantial expertise and a significant time commitment, making the process resource-intensive. Active learning provides a valuable solution by facilitating the selection of the most informative samples for expert labeling, thereby optimizing expert time and effort. By strategically choosing the most relevant samples for labeling, active learning reduces uncertainty and allows the model to leverage clusters of similar samples, improving labeling efficiency and minimizing redundant expert effort.

However, effective fraud detection models generally require a substantial number of labeled samples to achieve high accuracy. Active learning allows the selection of training samples judiciously, enabling a training process to commence with relatively few labeled samples. This approach is particularly valuable in scenarios with limited or no labeled data, as active learning can guide experts in labeling only essential samples, enabling the model to achieve acceptable accuracy with a minimized dataset. In scenarios akin to a "cold start," where labeled data is scarce, active learning can deliver strong predictive accuracy by focusing expert attention on the most informative samples.

While active learning effectively reduces the number of samples requiring expert labeling, it still requires expert input for a subset of critical samples to finalize fraud classifications. However, some samples reveal their classification only after initial model training, often through predictable patterns identified by the machine learning model itself. Here, pseudo labeling complements active learning by enabling the model to label samples based on its predictions, thus expanding the labeled dataset without additional expert intervention. By combining pseudo labeling with active learning, this approach leverages the machine learning model's knowledge to label less ambiguous samples autonomously, further reducing the demand on expert resources.

In this research, we aim to assess the effectiveness of combining active learning and pseudo labeling for fraud detection under conditions of limited training data. Specifically, we examine how the careful selection of samples for expert input, combined with model-generated pseudo labels, can enhance fraud detection accuracy and reduce reliance on large labeled datasets. This study explores various configurations of active learning and pseudo labeling to identify optimal solutions to the challenges presented by fraud detection in imbalanced datasets.

1.4 Outlines of Thesis

This thesis is organized into five chapters. Chapter One provides a general introduction to the thesis, including basic concepts, a review of relevant literature, and the motivations behind this study. Chapter Two presents a detailed examination of the approaches used for fraud detection, active learning, and pseudo labeling, covering the computational methods and potential applications of each. The chapter concludes with a description of our proposed methodology for addressing the fraud detection problem, highlighting the most effective combination identified through

testing. Chapter Three offers an overview of the three datasets used within the scope of this thesis. Chapter Four delivers an evaluation of four uncertainty sampling methods and one diversity sampling method, applied across six machine learning models: SVM, IF, ANN, CatBoost, XGBoost, and weighted XGBoost. The first subsection focuses on these sampling techniques' performance across the models, while the second subsection evaluates the combined impact of pseudo labeling, uncertainty sampling, and diversity sampling within active learning frameworks for each machine learning model. Finally, Chapter Five summarizes the findings of this research, highlighting the performance improvements achieved in fraud detection through the application of pseudo labeling and active learning approaches.



ACTIVE LEARNING BASED FRAUD DETECTION METHODOLOGY

Fraud detection is a complex topic with numerous approaches available, depending on the dataset structure and the nature of the fraud cases. Active learning strategies are known to assist in addressing fraud detection challenges by leveraging machine learning methods. Active learning methods are generally categorized into two main types: uncertainty sampling and diversity sampling. These approaches typically rely on expert opinions to label selected samples based on specific rule sets.

In addition to active learning, pseudo labeling is an effective technique for leveraging patterns identified during the model training process. By using insights gathered through training, pseudo labeling allows us to improve model evaluation metrics without requiring external expert decisions.

In the next section, we defined and compared our fraud detection and uncertainty active learning strategies methods. The following section is about the diversity sampling active learning strategy, in which we use K-Means Cluster-Based (CB) diversity sampling. The following section is about pseudo labeling, which is a practical usage area for sure trained samples to decrease expert opinion on unsupervised training processes. The final section is about our proposed methodology. The figure of the proposed method's stages and detailed information can be seen in the last section of the scope of this chapter.

2.1 Fraud Detection

Fraud detection has become an essential field of study within data science and machine learning, driven by the financial and reputational impacts of fraud across industries. Fraud, by its nature, is typically rare yet costly, often existing within highly imbalanced datasets where fraudulent transactions constitute only a small fraction of all instances. This rarity, combined with the evolving strategies of

fraudsters, makes fraud detection a complex and dynamic challenge that requires adaptable, efficient, and accurate methods. Machine learning has emerged as a powerful tool in this domain, offering approaches that can identify subtle patterns and anomalies indicative of fraudulent activity.

Traditional rule-based methods, while effective in detecting known fraud patterns, often fall short when faced with novel or subtle fraud schemes. Machine learning approaches, by contrast, have demonstrated the ability to detect complex, hidden relationships within data that would be difficult to identify using conventional methods. A range of supervised and unsupervised algorithms are commonly employed in fraud detection to enhance accuracy and mitigate the high cost of false positives.

Scope of thesis, various machine learning techniques are investigated for their effectiveness in identifying fraud, each offering unique strengths suited to different aspects of the problem. IF, an unsupervised anomaly detection algorithm, isolates observations by randomly partitioning the feature space, making it particularly effective at detecting outliers. SVM is another approach explored for fraud detection. SVM classifies instances by finding an optimal hyperplane that maximizes the margin between classes, useful in distinguishing fraud from legitimate transactions when sufficient labeled data is available.

Gradient boosting techniques such as CatBoost and XGBoost have also been applied to the task of fraud detection. CatBoost, a gradient boosting method designed to handle categorical variables efficiently, and XGBoost, known for its computational efficiency and regularization features, both provide robust models that can capture complex, non-linear patterns in imbalanced datasets. An adaptation of XGBoost, referred to as Weighted XGBoost, is used in this study to further enhance sensitivity to fraudulent cases by weighting minority class instances proportionally higher, allowing the model to prioritize fraud detection even in highly skewed datasets.

Additionally, artificial neural networks (ANN) are utilized for their ability to model intricate patterns through multiple layers of abstraction, making them particularly powerful in scenarios with large datasets and complex fraud patterns. With their capacity to learn deep representations of data, artificial neural networks offer flexibility in identifying evolving fraud patterns that may elude simpler models.

By investigating these methods this study aims to compare their performance in fraud detection and identify which approaches offer the most effective solutions in various scenarios. Each algorithm contributes distinct advantages, from anomaly detection and classification margin optimization to boosting techniques and deep learning, providing a comprehensive toolkit for addressing the challenges of fraud detection in real-world applications. Detailed information about these methods is given in further subsections.

2.1.1 Isolation Forest

The Isolation Forest (IF) algorithm is a robust and efficient method for anomaly detection, particularly suitable for high-dimensional datasets. It is founded on the principle that anomalies, or fraud cases, are fewer in number and structurally distinct from normal samples. Unlike traditional methods that model normal behavior, IF isolates anomalies directly, identifying them based on differences from the typical distribution of normal samples.

IF operates by constructing an ensemble of isolation trees, a unique approach that enhances its effectiveness in detecting anomalies. Each tree is created by recursively partitioning a random subset of the data. During each split, a feature is selected at random, along with a corresponding split value. The partitioning process continues until all points are isolated, or a specified tree depth is reached. The length of the path from the root node to a terminal node serves as a measure of how well-separated a sample is; samples with shorter paths are more likely to be classified as anomalies. Once the forest is complete, the average path length across all isolation trees is calculated for each sample, with significantly shorter path lengths typically indicating potential fraud cases.

This isolation-based approach allows IF to efficiently identify fraud samples within large, high-dimensional datasets. The algorithm is less prone to over-fitting compared to some other methods, due to its reliance on random splits rather than complex models. However, because it partitions data randomly, its performance may vary if fraud samples are not easily distinguishable from normal samples and are interspersed among them. Under these conditions, IF may struggle to accurately isolate fraud cases.

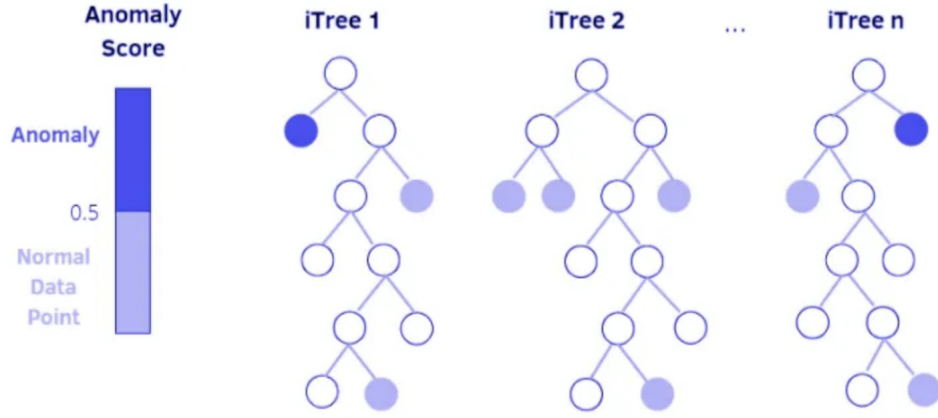


Figure 2.1 Isolation Forest

Figure 2.1[20] provides an illustrative example of the IF structure. In this figure, trees generated by the method are depicted with hierarchical leaf levels. Each leaf represents a sample from the dataset, where samples with leaves closer to the root are more likely to be identified as fraud cases. Conversely, samples with leaves positioned farther from the root are generally considered normal.

2.1.2 Support Vector Machine

Support Vector Machine (SVM) is a robust supervised machine learning method used for classification and regression tasks. It is generally used for the fraud detection problem. The main idea behind SVM is finding the appropriate hyperplane in N-dimensional space that distinctly and successfully classifies the samples.

Support vectors, crucial elements in the SVM model, are generated for samples that are closest to the hyperplane. They play a significant role in influencing the position and orientation of the hyperplane. The distance between the hyperplane and the nearest samples on each side is referred to as the margin. The primary goal of the SVM model is to maximize this margin, thereby enhancing the model's robustness.

The shape of the hyperplane in SVM is adaptable to the nature of the dataset. If the data is linearly separable, SVM finds a straight line in two dimensions. In three dimensions, a plane serves as the hyperplane. In higher dimensions, the hyperplane acts as a separation boundary. For non-linear data structures, SVM utilizes a kernel function to transform the data into a higher dimension where a hyperplane can be used to separate classes. In our approach, we employed the Radial Basis Function as the kernel in our SVM model. For our fraud detection problem, SVM aimed

to generate an appropriate hyperplane to represent the separation between normal and fraud samples. In this approach, the selection of the right features, such as transaction amount, time, and frequency, plays a crucial role in distinguishing fraud cases.

SVM has an effective calculation method over high-dimensional spaces. It has a versatile structure because it allows using different kernel functions. These changeable functions affect the finding of the correct hyperplane based on the dataset structure. SVM has a memory-efficient calculation structure. However, it is crucial to select the correct kernel function. It is sensitive to the choice of the regularization parameter and kernel parameters. It can be prone to over-fitting, especially in very high-dimensional spaces. So, it is vital to care about these features of the algorithm.

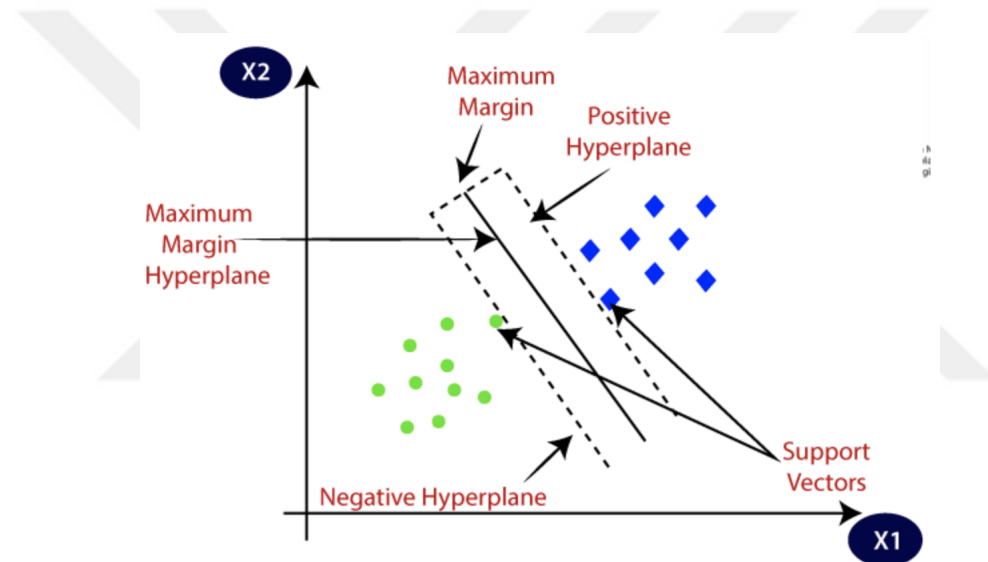


Figure 2.2 Support Vector Machine

There is a Figure placed in Figure 2.2[21] to explain the SVM methodology. The SVM method tried to generate a hyperplane to distinguish different classes of samples. Here, a successful hyperplane was found for the given samples. With this hyperplane, we could differentiate all different class-labeled samples. SVM tried to generate this hyperplane that was placed as far away for each different labeled class sample cluster. We called this distance a margin. So, an SVM model is thriving as the greatness of margin distance.

2.1.3 Artificial Neural Network

Artificial neural networks (ANN) are a foundational element of deep learning. It could be accepted that it is a subset of machine learning inspired by the architecture of the human brain. Nowadays, ANN is particularly used in complex problems in which relationships between inputs and outputs are non-linear.

ANN is basically generated from its essential building stone named neurons, which receive inputs, process them, and generate outputs. Placing neurons with different orders or structures changes the ANN. Placing different orders and sequentially gives us layers of ANNs. These layers are classified according to the places they are used. In the input taker, neurons are accepted as the input layer. After several calculations, calculated served into outside. We call the neurons in that place the output layer. There could be several different neurons as parallel or sequential; we call these neurons hidden layers. The hidden layer could contain more than one layer step. In these layers, neurons process the coming data with activation functions like sigmoid, ReLU, or tanh. According to this function, neural gathers coming data and forwards it to the next neuron. So, arranging neurons' places, layer structures, and activation functions influenced the output.

According to the structure of neurons, layer arrangement, and activation function, the artificial neural network is initialized with default value first. Input fed into an artificial neural network in the first place. According to the coming values, all neurons process the values and generate an output value until reaching the output layer neurons. This process is named forward propagation. In the final place, the ANN model evaluates the final values of output layers and starts to change the weights of the hidden layer neurons' effect into the model until it reaches the input neurons. This process is called back-propagation. With sequential processes named forward propagation and back-propagation, we arrange the weight of connected neurons effect over the model, and this process continues until there is no effective neuron weight change.

This approach is capable of learning non-linear relationships and is highly adaptable to various types of data. However, finding the right weights of neurons is a challenging problem. It's a complex task that requires large amounts of data to train effectively. Additionally, we generally don't know the hidden layer structure, which means there's no proper regularization to prevent over-fitting.

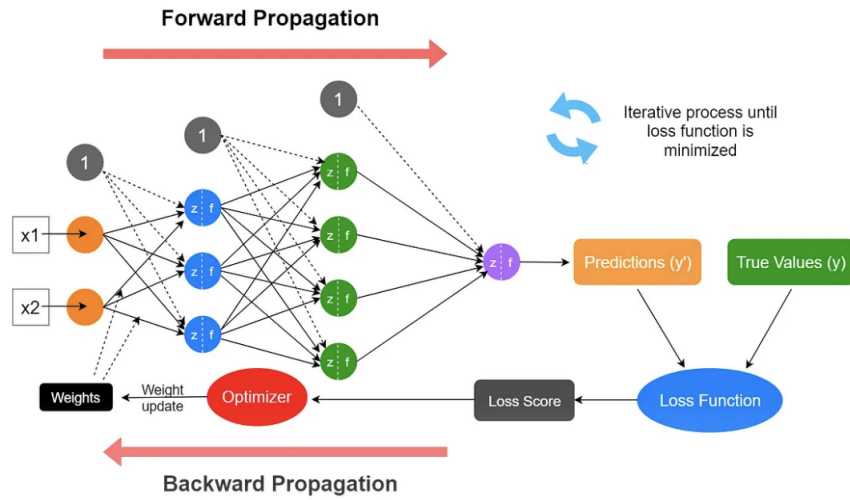


Figure 2.3 Artificial Neural Network

Figure 2.3[22] explains the ANN method which takes the features from the input layer. Then, many neurons in the hidden layer are generated, and many hidden layers contain numerous and different valued neurons. In the first place, all neurons and their connection have the same weight. According to the first data feed, all neurons calculate coming data and calculate an output value. They served this output value into their connected subsequent neurons. This process is named forward propagation. In the output layer, we obtained a value set. ANN compares these values with its fed data. According to the difference between its calculated and final data, ANN started to update the weight of neuron connections to obtain more valid results. This process continued until the input layer was reached. This process is named back-propagation. All these calculations are named epochs. These steps are repeated until the weights of connections of neurons nearly do not change or are given a strict maximum number of epochs if we reach them.

The Multi-Layer Perceptron classifier was used in the scope of this research. It has 64 hidden layers and allows a maximum of 10 iterations in each epoch.

2.1.4 XGBoost

Extreme Gradient Boosting (XGBoost), is an advanced ensemble learning technique based on gradient boosting that has gained widespread popularity due to its effectiveness in classification and regression tasks, particularly in high-stakes fields like fraud detection. Developed by Tianqi Chen and Carlos Guestrin in 2016 [23], XGBoost distinguishes itself from traditional gradient boosting by introducing several optimizations that enhance both its computational efficiency and predictive accuracy. These optimizations include regularization, tree pruning,

parallel processing, and a novel tree structure for handling missing values, which together contribute to its robust performance, even with highly imbalanced datasets.

XGBoost is an ensemble method that builds a series of decision trees sequentially, with each tree constructed to minimize the residual errors of its predecessors. The primary goal of the algorithm is to minimize an objective function that combines a differentiable loss function with a regularization term that penalizes model complexity. By iteratively adjusting the weights of training instances based on previous errors, XGBoost can capture complex, non-linear relationships in the data.

The objective function L in XGBoost at each iteration t is given by:

$$L^{(t)} = \sum_{i=1}^n \ell \left(y_i, \hat{y}_i^{(t)} \right) + \sum_{k=1}^t \Omega(f_k) \quad (2.1)$$

where n is the number of training samples, $\ell \left(y_i, \hat{y}_i^{(t)} \right)$ is the loss function, typically the mean squared error for regression or log loss for classification, which measures the difference between the true value y_i and the predicted value $\hat{y}_i^{(t)}$ at iteration t , f_k represents each decision tree in the ensemble, and $\Omega(f_k)$ is a regularization term that penalizes the complexity of each tree f_k helping to prevent over-fitting.

Regularization is a critical component in XGBoost, as it addresses the tendency of boosted trees to overfit, particularly in cases of noisy or imbalanced data. The regularization term $\Omega(f_k)$ for a tree-based model is defined as:

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \quad (2.2)$$

where T is the number of leaves in the tree, γ is a parameter that controls the penalty for adding a new leaf node, and λ is a parameter that controls the L_2 regularization on the weights w_j of the leaves.

By controlling γ and λ , XGBoost mitigates over-fitting by discouraging the model from creating excessively complex trees. This balance between bias and variance allows XGBoost to generalize effectively to new, unseen data.

XGBoost greedily constructs trees, selecting splits that optimize the reduction in loss at each step. The optimal split for a node is determined by the gain G , which is defined as the improvement in the objective function when a node is split. For a feature j , the gain from a split at threshold s is calculated as:

$$G = \frac{1}{2} \left(\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right) - \gamma \quad (2.3)$$

where G_L and G_R represent the sum of the first-order gradients for the left and right nodes after the split, H_L and H_R represent the sum of the second-order gradients for the left and right nodes, λ is the regularization parameter controlling the penalty for large weights, and γ penalizes additional leaf nodes to control tree complexity.

The gain function captures both the predictive power of each split and the regularization effects, thus selecting splits that not only reduce loss but also maintain the model's simplicity.

2.1.5 Weighted XGBoost

Weighted Extreme Gradient Boosting (Weighted XGBoost) is an adaptation of the standard XGBoost method tailored specifically for handling highly imbalanced datasets, such as those commonly encountered in fraud detection, where fraud cases often represent only a small fraction of the total instances. In this variant, a parameter known as **scale_pos_weight** is introduced to account for the imbalance between the positive (fraud) and negative (normal) classes. By setting **scale_pos_weight** to the ratio of fraud labeled samples to normal samples, the algorithm emphasizes the minority class, increasing its impact on the model's learning process and helping to improve detection rates for rare events.

In traditional version of XGBoost, all instances are weighted equally, which can lead to suboptimal performance when the data is highly imbalanced, as the model tends to prioritize accuracy on the majority class, often neglecting the minority class. Weighted XGBoost, by contrast, assigns higher importance to fraud cases through the **scale_pos_weight** parameter, effectively increasing the contribution of fraud instances in the calculation of the objective function. As a result, the model is incentivized to minimize classification errors in the minority class, leading to a more balanced and accurate performance.

The algorithm proceeds by iteratively constructing decision trees that minimize a regularized objective function, with each tree trained to reduce the residual errors of its predecessors. By adjusting the model to assign a greater penalty for misclassifying minority class instances, weighted XGBoost achieves improved sensitivity to fraud cases, thus addressing the critical need for accuracy in imbalanced classification tasks.

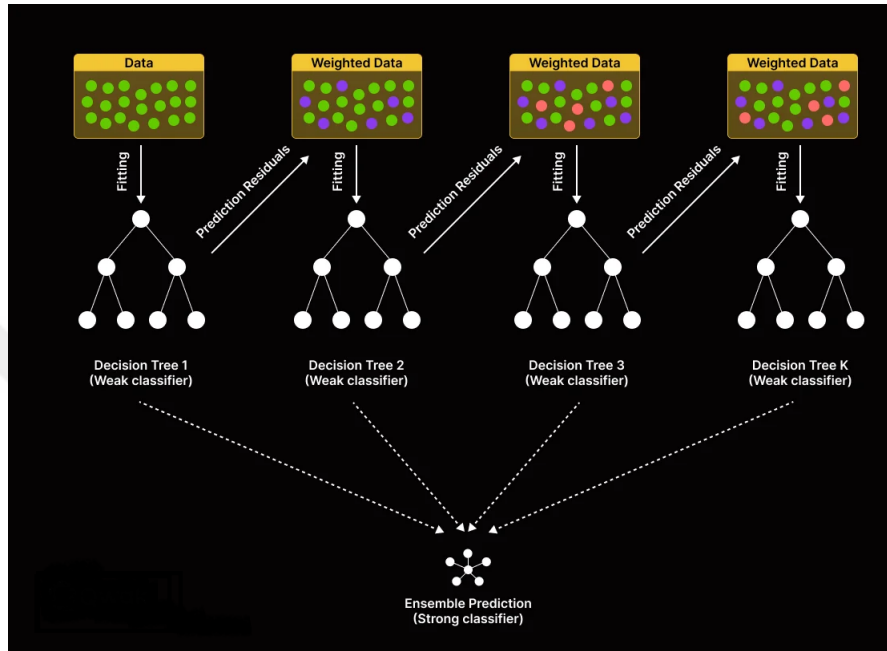


Figure 2.4 Extreme Gradient Boosting

There is a figure in Figure 2.4[24] that explains the Gradient Boosting method. XGBoost method is based on Gradient Boosting method but it can be handle calculation part faster and more convenient way. The Weighted XGBoost algorithm combines gradient boosting with multiple computational and algorithmic optimizations that enhance both speed and accuracy, making it a powerful tool for fraud detection. By leveraging regularization, parallel processing, and efficient split selection, Weighted XGBoost can model complex relationships while avoiding over-fitting. Its adaptability to highly imbalanced datasets, through configurable hyper-parameters like **scale_pos_weight**, further underscores its suitability for real-world fraud detection applications, where accurate identification of fraudulent activities is paramount.

2.1.6 CatBoost

CatBoost, short for Categorical Boosting, is a gradient boosting method designed to handle categorical data effectively, making it well-suited for datasets with a significant proportion of categorical features. Developed by Yandex, CatBoost builds on the principles of gradient boosting but incorporates innovative techniques to mitigate issues such as over-fitting and target leakage, which are common challenges in gradient boosting frameworks.

CatBoost, like other gradient-boosting methods, builds an ensemble of decision trees by iteratively adding trees that correct the errors of the previous ones. Formally, given a dataset (X, y) where X presents the features and y the target variable, gradient boosting aims to learn a function $F(X)$ that minimizes the loss function $L(y, F(X))$. Starting with an initial prediction $F_0(x)$, each subsequent iteration m adds a decision tree $h_m(X)$ to approximate the residuals of the previous predictions:

$$F_m(X) = F_{m-1}(X) + \alpha h_m(X) \quad (2.4)$$

where α is a learning rate parameter that controls the contribution of each new tree to the overall model. The goal is to iteratively minimize the loss function by adjusting the predictions $F_m(X)$ until they closely match the target values y .

One of CatBoost's primary advantages is its native support for categorical data. Most gradient boosting methods require categorical features to be converted into numerical values using methods like one-hot encoding or target encoding. These approaches can lead to high dimensionality or target leakage. CatBoost, however, uses a unique encoding scheme that assigns a value to each category based on the average target value of the feature while taking care to prevent target leakage. This encoding uses a randomized permutation of the dataset to calculate the mean target value for each category, excluding the current sample. Formally, for a categorical feature x_i with categories $C_1, C_2, \dots, C_k$, Catboost calculates the target encoding for a category C_j at instance i as :

$$Encoding(C_j) = \frac{\sum_{t=1}^{i-1} y_t * I(x_t = C_j) + \alpha}{\sum_{t=1}^{i-1} I(x_t = C_j) + \alpha} \quad (2.5)$$

where I is the indicator function, y_t is the target value for sample t , and α is a smoothing parameter to prevent over-fitting. This encoding process ensures that each sample's encoding depends only on the data points that appeared before it in the permutation, thus reducing target leakage.

Catboost introduces a unique ordered boosting approach to address the gradient bias problem. Traditional gradient boosting methods calculate the gradient based on the predictions from previous trees, which can lead to target leakage and over-fitting. In contrast, Catboost uses ordered boosting, where it splits the dataset into multiple folds, and each fold is used to calculate predictions and gradients independently. This prevents the model from using the current target data to improve predictions, reducing the risk of over-fitting.

One of the primary advantages of Catboost is its innovative approach to handling categorical features. Unlike other gradient boosting methods that require categorical features to be transformed into numerical representations using techniques such as one-hot encoding or target encoding, Catboost has a unique mechanism that reduces the risk of target leakage. By encoding categorical features using a permutation-based scheme, Catboost ensures that each sample's encoding is computed based on prior samples only, thereby preventing information from future samples from influencing the encoding. This approach not only mitigates target leakage but also avoids the high-dimensionality issues associated with one-hot encoding, making Catboost highly efficient on datasets with many categorical features.

CatBoost is also known for its ability to reduce over-fitting, thanks to its ordered boosting technique. In traditional gradient boosting, the model calculates gradients based on predictions from previous trees, which can inadvertently introduce target leakage and lead to over-fitting. Catboost, however, implements an ordered boosting approach where predictions for a sample are made using a subset of trees built without access to the target value of that sample. This method significantly reduces over-fitting, particularly in small datasets, where the risk of memorizing the target values is higher. By preventing the model from “seeing” the actual target values during training, ordered boosting enhances the model's generalization ability.

In terms of computational efficiency, CatBoost uses symmetric decision trees, where splits occur along the same feature across all branches in each tree. This symmetry allows Catboost to perform both training and inference more quickly than traditional boosting methods, which build asymmetrical trees. Faster training

and prediction times make Catboost particularly advantageous when working with large datasets or when rapid model deployment is required. The use of symmetric trees also helps reduce the computational cost associated with tree building, making Catboost an appealing choice for high-dimensional data.

Additionally, CatBoost includes mechanisms for automatic hyperparameter tuning, which can simplify the training process for users. In most machine learning methods, tuning hyperparameters like the learning rate, tree depth, and the number of iterations is essential for achieving optimal model performance. Catboost incorporates several optimizations that enable it to automatically select near-optimal values for many of these parameters, reducing the need for extensive manual tuning. This feature makes Catboost accessible and user-friendly, particularly for practitioners who may not have extensive experience with machine learning.

In summary, CatBoost's advantages lie in its ability to handle categorical features effectively, reduce over-fitting, improve computational efficiency, and simplify the model-tuning process. These characteristics make CatBoost a powerful choice for a wide range of machine-learning tasks, particularly those involving datasets with numerous categorical variables and high-dimensional data. Its robustness against target leakage and over-fitting, combined with efficient training and prediction times, allows it to excel in scenarios where other gradient boosting methods might struggle.

2.2 Uncertainty Sampling Active Learning Strategies

In machine learning, active learning has become a valuable strategy for constructing accurate models in situations where labeled data is limited or expensive to obtain. Active learning strategies aim to optimize the learning process by selectively querying data points that will most improve model performance. uncertainty sampling is a key approach within this framework, particularly useful in domains with high-stakes decision-making, such as fraud detection, where data imbalance and rarity of critical cases are common challenges. By focusing on instances where the model has the least confidence, uncertainty sampling targets data points that are likely to provide the most information for refining model boundaries and reducing classification errors.

Uncertainty Sampling works by selecting samples the model finds most ambiguous, typically those near the decision boundary where it has the lowest confidence. Unlike diversity-based methods, which prioritize broad coverage of the feature space, uncertainty sampling zeroes in on instances with high predictive uncertainty,

which is likely to yield the greatest reduction in model error upon labeling. This approach is particularly advantageous in scenarios like fraud detection, where subtle, ambiguous patterns may differentiate fraudulent from non-fraudulent cases. The method ensures the model is repeatedly challenged with difficult-to-classify examples, fostering improved sensitivity to rare and nuanced patterns.

In fraud detection, uncertainty sampling can help the model better recognize subtle cues and edge cases that may indicate fraudulent behavior, thereby reducing the likelihood of overlooking atypical fraud instances. Its targeted approach facilitates a more focused and resource-efficient active learning process, ultimately leading to a model that is more attuned to real-world complexities and anomalies.

There are several techniques commonly used within uncertainty sampling. Some of them could be examined as Least Confidence, Margin Confidence, Ratio Confidence, and Entropy-Based Sampling. These methods quantify uncertainty in different ways, such as by focusing on the confidence level of the top prediction, the difference between the highest and second-highest class probabilities, or the entropy of the prediction distribution. By selecting samples based on these criteria, uncertainty Sampling actively addresses the model's areas of weakness, ensuring that each queried instance offers maximum potential for enhancing accuracy and robustness. The following subsection details our selected uncertainty sampling on active learning approaches.

2.2.1 Least Confidence

Least Confidence (LC) uncertainty sampling is a widely used approach in active learning, especially effective in scenarios where the model's confidence in its predictions needs to be quantified. This method selects samples for which the model has the lowest confidence in its prediction, thereby focusing on the most ambiguous cases for labeling. By prioritizing highly uncertain samples, LC helps improve model performance efficiently by concentrating on data points that provide the most learning value.

In LC, uncertainty is calculated based on the highest predicted probability among the possible classes for each sample. For a given sample, let p represent the vector of probability values output by the model, where each element p_i corresponds to the model's predicted probability of that sample belonging to $class_i$. In LC sampling, only the highest probability value, denoted p_{max} , is considered, which represents the model's most confident prediction. LC is then calculated as:

$$LC = 1 - p_{max} \quad (2.6)$$

where $p_{max} = \max(p)$. This formula subtracts the highest probability from 1, producing a measure of uncertainty; the closer the value of LC is to 1, the lower the model's confidence in its prediction, making it an ideal candidate for expert labeling.

In practice, the LC metric is calculated for all samples in a given batch, and the samples with the highest LC scores are selected as the most uncertain. This targeted selection allows the model to learn progressively from its most challenging cases, the data points it is least confident about. This method is beneficial for imbalanced datasets, as it helps refine the model's decision boundary in regions where fraud cases (or minority class cases) are more complex to identify.

2.2.2 Margin Confidence

Margin Confidence (MC) uncertainty sampling is an active learning strategy that targets samples based on margin between the two highest predicted class probabilities. Unlike LC, it focuses solely on the most probable class, MC considers the difference between the top two class probabilities, making it effective for identifying samples that lie near the decision boundary of the model. This approach is advantageous in cases where the model's uncertainty stems from low confidence in a single class and close competition between two potential classes.

For a given sample, let p denote the vector of predicted class probabilities output by the model, where each element p_i represents the probability of the sample belonging to $class_i$. To compute the MC score, the probabilities are sorted in descending order, and the two highest probabilities, p_{max1} and p_{max2} , are selected. The MC uncertainty score is then defined as the difference between these two highest probabilities:

$$MC = p_{max1} - p_{max2} \quad (2.7)$$

where p_{max1} and p_{max2} represent the highest and second-highest probabilities, respectively. In MC sampling, a smaller margin indicates greater uncertainty, as the model is nearly equally likely to assign the sample to either of these top classes.

The MC metric is calculated for all samples in a batch, and samples with the smallest margins are prioritized as the most uncertain. By focusing on these closely competing samples, MC sampling helps the model refine its decision boundary, especially in the most ambiguous areas. This strategy is highly effective for imbalanced datasets, where distinguishing between similar classes is crucial, as it allows the model to focus on cases that could easily be misclassified.

2.2.3 Ratio Confidence

Ratio Confidence (RC) uncertainty sampling is an active learning technique designed to identify samples where the model’s classification confidence is ambiguous. RC considers the ratio between the two highest class probabilities, unlike other uncertainty methods. This ratio-based approach is beneficial in cases where the model’s uncertainty comes from closely competing probabilities, and it enables more nuanced detection of samples near the decision boundary.

For a given sample, let p represent the vector of predicted class probabilities, where each element p_i corresponds to the probability of the sample belonging to $class_i$. To compute the RC score, the probabilities are sorted in descending order to identify the two highest probabilities p_{max1} and p_{max2} . The RC uncertainty score is then defined as the ratio of these two probabilities:

$$RC = p_0/p_1 \quad (2.8)$$

where p_{max1} is the highest probability, and p_{max2} is the second-highest probability. In RC sampling, a higher ratio signifies greater uncertainty, as it indicates that the model assigns nearly equal probabilities to the top two classes.

The RC score is calculated for each sample in a batch, and those with the highest RC values are prioritized as the most uncertain. By selecting samples with a high RC score, the model focuses on cases where the distinction between the top two classes is least clear, refining its decision boundary where it is most ambiguous. This approach is convenient for imbalanced datasets, as it enables the model to concentrate on learning from complex samples to classify, improving its ability to handle subtle differences between classes, such as fraud and non-fraud instances.

2.2.4 Entropy

Entropy-Based (EB) uncertainty sampling is an active learning approach that leverages the concept of entropy to assess the model’s uncertainty about class predictions. Entropy, quantifies the uncertainty or disorder within a probability distribution. In classification, entropy measures the uncertainty in the model’s predicted class probabilities, with higher entropy indicating more significant uncertainty about the correct class label.

For a given sample, let p represent the vector of class probabilities predicted by the model, where each p_i corresponds to the probability that the sample belongs to $class_i$. The entropy H of this probability distribution is calculated as:

$$H = - \sum_{i=1}^C p_i \log(p_i) \quad (2.9)$$

where C is the total number of classes, and p_i is the probability of the sample being classified as $class_i$. Entropy values range from 0 to $\log(C)$ with higher values indicating greater uncertainty. When all probability mass is concentrated, close to 0 entropy, signaling high confidence. Conversely, entropy approaches its maximum when probability values are spread evenly across multiple classes, reflecting high uncertainty.

In EB uncertainty sampling, entropy serves as an uncertainty metric to pinpoint samples where the model is least certain. For each sample, entropy is computed, and the samples with highest entropy values are chosen for labeling. This approach enables the model to prioritize ambiguous samples, where it assigns similar probabilities to multiple classes. By concentrating on high-entropy samples, EB sampling ensures that the model focuses on data points where its predictions are most uncertain.

The use of EB sampling is particularly advantageous in imbalanced datasets, such as those commonly encountered in fraud detection, where certain classes are much rarer than others. By selecting samples with high entropy, the model is guided to focus on instances that will likely enhance its ability to distinguish between fraud and normal classes. This entropy-based approach not only improves the model’s decision boundary but also enables the model to gain crucial information from samples where class probabilities are nearly evenly distributed. This, in turn, refines the model’s understanding of difficult-to-classify cases, providing a robust foundation for active learning.

2.3 Diversity Sampling Active Learning Strategy

In machine learning, active learning has emerged as a critical strategy for efficiently building accurate models, especially when labeled data is scarce or costly. Active learning strategies seek to optimize model training by selectively querying data points that enhance the model's performance. Diversity sampling is a prominent approach within this framework, which is precious in scenarios involving imbalanced datasets or high-dimensional data, such as fraud detection. By focusing on querying samples that best represent the overall distribution of the dataset, diversity sampling addresses critical challenges in learning from imbalanced data, where fraudulent instances may be rare but varied. It ensures that the model learns from a broad range of data points.

Diversity Sampling operates by selecting a set of informative data points that are both uncertain and diverse. Unlike uncertainty-based methods that prioritize samples where the model has low confidence, diversity sampling leverages distance-based or clustering approaches to comprehensively cover as many aspects of the feature space as possible. This approach is advantageous in domains like fraud detection, where fraud cases are often anomalous and do not conform.

Several techniques are commonly employed within diversity sampling, including K-means clustering and maximum-minimum distance sampling. These techniques group data points into clusters based on feature similarities and then select representatives from each cluster. By doing so, diversity sampling ensures that the active learning process does not solely focus on a single or narrow class of data, which is especially valuable in highly imbalanced datasets. This leads to a more comprehensive and balanced model better suited for detecting rare but impactful cases like fraud.

In fraud detection, diversity Sampling can help mitigate the bias towards non-fraudulent data and improve the model's sensitivity to rare yet varied fraudulent patterns. Diversity Sampling's robustness ensures an efficient and robust active learning process, ultimately enhancing the model's generalization capacity and real-world applicability. The following subsection describes diversity sampling active learning approach, the K-Means cluster-based diversity sampling method.

2.3.1 K-Means Cluster-Based Diversity Sampling

The K-Means cluster-based diversity sampling approach enhances the diversity of selected samples in active learning. It is particularly useful in scenarios involving high-dimensional or imbalanced data, such as fraud detection. This method leverages the K-means clustering method to identify representative samples that can enrich the learning process by ensuring that the queried data points are informative and varied.

The K-means clustering algorithm is an iterative unsupervised learning method designed to partition a dataset into k clusters where each cluster is represented by its center, also known as the centroid. Given a set of data points $x_1, x_2, \dots, x_n$, K-means aims to minimize the within-cluster variance by finding centroids $c_1, c_2, \dots, c_k$ that best represent each cluster. Formally, the objective function is defined as:

$$J = \sum_{i=1}^k \sum_{x_j \in S_i} \|x_j - c_i\|^2 \quad (2.10)$$

where S_i denotes the set of data points assigned to cluster i , c_i represents the centroid of $cluster_i$ and $\|x_j - c_i\|^2$ is the Euclidean distance between data point x_j and c_i .

The algorithm started with the initialization process. In this process, Select k initial centroids randomly from the dataset or based on a specific heuristic. After that part, the assignment continued. For each data point, x_j , assign it to the nearest centroid c_j based on the Euclidean distance. The next operation is updating centroids. The algorithm recomputes each centroid c_j as the mean of all data points assigned to it. Repeat the assignment and update steps until convergence occurs when the centroids stabilize and do not change significantly between iterations.

This iterative process clusters the dataset into k distinct groups, with each centroid representing the central point of a cluster. In the context of active learning, the centroids act as diverse, representative samples of the entire data distribution.

In diversity sampling for active learning, the K-Means algorithm is applied to the pool of unlabeled data. By clustering this data, the method identifies the most informative samples across the feature space, ensuring a broad representation of the dataset's structure. The centroids, as outputs of the clustering process, are then selected as the queried samples because they represent each cluster's most central, typical points. This approach is particularly beneficial in datasets with rare or

imbalanced classes, as it enables the model to learn from various data distributions rather than being overly influenced by the dominant class.

When used as part of a hybrid active learning strategy, K-Means cluster-based diversity sampling can be combined with uncertainty sampling to refine the sample selection process further. In this hybrid approach, uncertain samples are prioritized while ensuring diversity by choosing samples that are also the centroids of clusters. This combination improves the model's exposure to uncertain cases and prevents redundancy by avoiding the over-selection of similar data points. Consequently, this hybrid strategy leads to a more balanced and comprehensive training set, enhancing the model's capacity to generalize across varied data points, including those representing less frequent patterns, such as fraudulent transactions.

Integrating K-Means cluster-based diversity sampling within the active learning framework ensures that queried data points are diverse and representative. This leads to a more robust and effective learning process in highly imbalanced datasets.

2.4 Pseudo Labeling

Pseudo labeling (PL) is a semi-supervised learning technique that leverages labeled and unlabeled data to improve model performance. In traditional supervised learning, only labeled data are used, which can be a limiting factor when dealing with highly imbalanced datasets, such as those commonly encountered in fraud detection. Pseudo labeling addresses this by allowing the model to assign pseudo labels to unlabeled data based on its current predictions, effectively transforming specific high-confidence predictions into additional labeled data. These pseudo labeled instances are then incorporated into the training process, gradually refining the model's ability to identify fraudulent patterns.

In the context of fraud detection, pseudo labeling offers distinct advantages. Given the scarcity of fraud cases, acquiring a fully labeled dataset can be costly and time-intensive. By employing pseudo labeling, it is possible to augment the training set with high-confidence predictions of fraud and non-fraud cases, allowing the model to learn from a more diverse and representative dataset. This is particularly useful for capturing subtle patterns and improving detection performance on rare fraudulent events.

When combined with active learning, pseudo labeling can further enhance the efficiency of fraud detection models. Active learning identifies the most uncertain samples in the dataset and queries experts for labels, improving the model

with minimal labeling effort. pseudo labeling can complement this by pseudo labeling additional high-confidence samples, enabling the model to learn from expert-confirmed instances and reliably predicted ones.

The hybrid approach of active learning and pseudo labeling is a cost-effective solution for fraud detection, where balancing accuracy and efficiency is crucial. Active learning can focus on low-confidence or highly ambiguous transactions for expert labeling, while pseudo labeling can make use of high-confidence non-labeled samples. This not only expands the training data pool but also reduces the manual labeling workload, providing a robust and cost-effective solution for fraud detection.

The adaptive framework of pseudo labeling with active learning has the potential to significantly enhance the model's exposure to diverse fraud patterns. By dynamically incorporating both expert-labeled and pseudo labeled samples, the model's generalization capabilities and sensitivity to rare fraud cases can be greatly improved. This promising approach effectively tackles the challenges posed by highly imbalanced fraud datasets.

2.5 Proposed Method

In this study, we propose a hybrid active learning approach that combines Least Confidence uncertainty sampling with K-means cluster-based diversity sampling and pseudo labeling. This methodology aims to enhance fraud detection performance, particularly in scenarios with highly imbalanced datasets, by leveraging minimal expert intervention and augmenting the training set through pseudo labeling. By employing these techniques in tandem, our approach improves the model's F1 score while significantly reducing the reliance on labeled data.

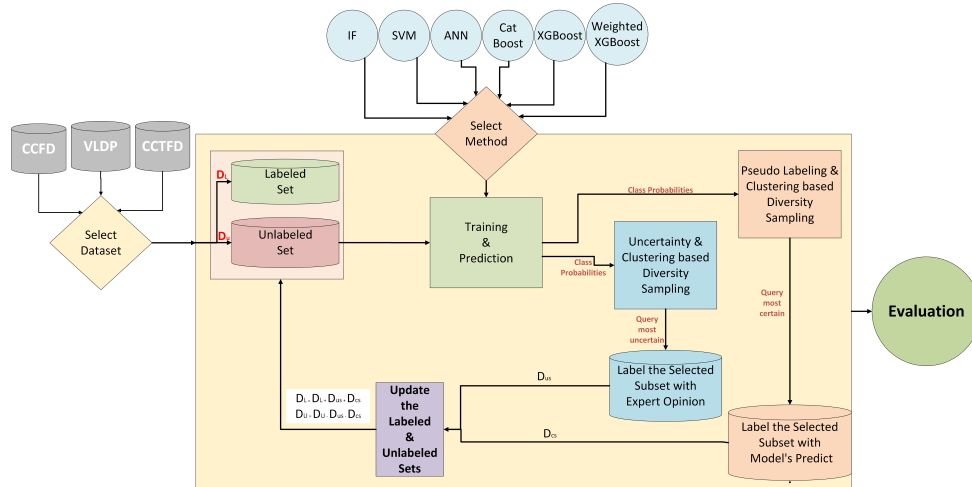


Figure 2.5 The Architecture of Implemented Approach

Figure 2.5 presents the working way of our proposed method. It started with selecting an initial training set of less than 4% of the total dataset, with the remaining data designated for testing. At each iteration, the model undergoes an active learning and pseudo labeling process. For active learning, we apply LC uncertainty sampling to select the samples about which the model is least certain. These samples are then subjected to K-means clustering, where cluster centroids represent the most uncertain instances, which are then labeled by an expert and added to the training set. This step ensures that the model receives highly informative samples to improve its performance incrementally.

For pseudo labeling, we simultaneously select the samples about which the model is most certain, clustering these samples using K-means clustering to select cluster centroids as representatives of confident predictions. These samples are then pseudo labeled with the model's predictions and added to the training set to enhance the model's exposure to diverse, confidently predicted examples, ultimately reducing the need for extensive expert labeling. Detailed steps as pseudo codes can be seen below as Algorithm 1 under Table 2.1.

Table 2.1 Pseudo Code of Hybrid Active Learning with Pseudo Labeling for Fraud Detection as Algorithm

Algorithm 1 Hybrid Active Learning with Pseudo Labeling for Fraud Detection Require: Dataset $\mathcal{D}$, initial training set $\mathcal{D}_{\text{train}}$, test set $\mathcal{D}_{\text{test}}$, number of samples N, number of clusters k, stopping criterion L Ensure: Trained model with an improved F1 score 1: while not reached stopping criterion L do 2: Active Learning Part 3: Select $5 * n$ samples from the most uncertain set (Least Confidence sampling) 4: Apply K-means clustering to the selected samples with k 5: Choose the cluster centroids as the most uncertain k samples 6: Add these samples to $\mathcal{D}_{\text{train}}$ with true labels (expert labeling) 7: Pseudo Labeling Part 8: Select $5 * N$ samples from the most certain set 9: Apply K-means clustering to the selected samples with k 10: Choose the cluster centroids as the most certain k samples 11: Add these samples to $\mathcal{D}_{\text{train}}$ with predicted labels (pseudo labeling) 12: Update Model 13: Train the selected model on updated $\mathcal{D}_{\text{train}}$ 14: Test the model on $\mathcal{D}_{\text{test}}$ and update performance metrics 15: Remove found samples from $\mathcal{D}_{\text{test}}$ 16: end while 17: Evaluate the final model on held-out test data to assess F1 score
--

The process iterates until a specified stopping criterion is met, typically based on a predetermined loop count, at which point the final model is evaluated to assess the effectiveness of the proposed hybrid sampling strategy. The hybrid approach, which combines Least Confidence uncertainty sampling with K-means cluster-based diversity sampling and pseudo labeling, demonstrates notable improvements in fraud detection performance over traditional methods. The main strength of this approach lies in its ability to selectively target the most informative samples at each iteration, using a carefully balanced combination of expert-labeled data and model-predicted labels to optimize training efficiency and accuracy.

Other methods explored have distinct advantages but encounter limitations in the context of fraud detection. For example, IF operates unsupervised and thus lacks targeted learning from labeled fraud data, making it less effective at capturing subtle fraud patterns. SVM, while powerful for classification, can struggle with highly imbalanced datasets, often over-fitting to the majority class. ANNs require large labeled datasets to avoid over-fitting, a significant challenge given the scarcity of labeled fraud cases. XGBoost, though robust in handling imbalanced data, depends heavily on the initial training set distribution and may overlook rare fraud cases in limited-label scenarios. Weighted XGBoost improves sensitivity to minority classes but can still be inefficient due to redundant sampling of similar cases. Lastly, while CatBoost effectively handles categorical features, it does not prioritize the most informative samples, which limits its efficiency in fraud detection tasks with limited labeled data. Additionally, alternative uncertainty sampling methods like Margin Confidence, Ratio Confidence, and Entropy-based sampling can identify uncertain instances but may not target the most valuable samples with the same level of precision as our approach.

Our method's advantage over these approaches lies in its hybrid design, which iteratively incorporates the most uncertain and diverse samples alongside confident pseudo labeled instances. This strategy allows the model to learn from a small but highly informative dataset, reducing the dependency on extensive expert labeling and enabling it to capture nuanced fraud patterns effectively. By combining uncertainty sampling, diversity sampling, and pseudo labeling, our method achieves a balanced and adaptive training process, leading to significantly improved F1 scores and enhanced robustness against fraud cases. This targeted methodology addresses real-world challenges in fraud detection, such as data imbalance and evolving fraud patterns, providing a comprehensive solution for accurate and efficient fraud identification.

3

DATASETS

Fraud detection is a broad research topic with numerous datasets available for analysis. Various datasets can be found on open platforms such as Kaggle, academic publications, cloud storage, proprietary platforms, and within company data warehouses. Given the vast range of potential datasets across these sources, carefully select datasets suitable for fraud detection research. To this end, a comprehensive benchmark study was conducted by Prince Grover and his colleagues [4]. Their research examined numerous datasets across various criteria, identifying those that could be effectively applied to fraud detection.

Grover and colleagues highlighted several key challenges of fraud detection datasets, including the high imbalance between fraud and normal cases, high feature cardinality, low sample independence, and non-identical distribution. Their findings identified nine datasets potentially suitable for fraud detection research. In this thesis, three of these recommended datasets were selected to evaluate the proposed methodology. Summary information on these datasets is provided in the table below, with detailed information available in the subsequent dataset subsections.

Table 3.1 Information of Datasets used in Thesis

Dataset Name	Feature Count	Fraud Count	Fraud Case Rate	Sample Count
Credit Card Fraud Detection Dataset	30	492	0.172%	284.807
Vehicle Loan Default Prediction Dataset	33	50.611	14,6%	345.546
Credit Card Transactions Fraud Detection Dataset	9	9651	0.52%	1.852.394

3.1 Credit Card Fraud Detection Dataset

The first dataset selected for this study is the Credit Card Fraud Detection dataset from Kaggle [25], which comprises credit card transaction data collected from the banking sector. This dataset is highly imbalanced, with only 0.172% of samples labeled as fraud, and includes nearly 30 features.

Originating from real-world banking data, the dataset was published as a result of Principal Component Analysis (PCA) to address confidentiality concerns. As a result, the actual transaction values were transformed, producing anonymized numerical features without directly interpretable meanings. The PCA-processed dataset, which consists exclusively of numerical features, was made available as an open-source dataset on Kaggle, and we chose to use it in this thesis due to its accessible and standardized format. In a systematic review survey [26] conducted by Suraya Nurain Kalid and his colleagues shows that Credit Card Fraud Detection Dataset used 60 different works scope in reviewed 87 works on detecting credit card frauds and payment defaults problem. So its commonly used dataset on fraud detection works especially on credit card topic.

The dataset contains 284,807 samples, of which only 492 are labeled as fraud. This 0.172% fraud rate underscores the challenge of working with highly imbalanced data to achieve reliable fraud detection predictions.

To visualize the dataset, we applied t-SNE to evaluate the distribution of fraud and normal data in a two-dimensional space. The t-SNE output in Figure 3.1 shows fraud-labeled samples as red points and normal samples as blue points. This figure provides a visual representation of the dataset's distribution for analysis.

Our t-SNE model reveals that fraud-labeled samples constitute a significant minority within the dataset. While some fraud samples are clustered, many fraud instances are widely dispersed across the dataset. This distribution presents a challenge, as it complicates distinguishing between fraud and normal samples for accurate labeling.

The following Table 3.2 represents a sample record in a structured format, showing details such as transaction date, customer information, merchant details, and fraud indicator. This data format is essential for processing and analysis in fraud detection tasks.

Table 3.2 Single Sample Example for Credit Card Fraud Detection Dataset

Column Name	Column Value
Time	0
V1	-1.3598071336738
V2	-0.0727811733098497
V3	2.53634673796914
V4	1.37815522427443
V5	-0.338320769942518
V6	0.462387777762292
V7	0.239598554061257
V8	0.0986979012610507
V9	0.363786969611213
V10	0.0907941719789316
V11	-0.551599533260813
V12	-0.617800855762348
V13	-0.991389847235408
V14	-0.311169353699879
V15	1.46817697209427
V16	-0.470400525259478
V17	0.207971241929242
V18	0.0257905801985591
V19	0.403992960255733
V20	0.251412098239705
V21	-0.018306777944153
V22	0.277837575558899
V23	-0.110473910188767
V24	0.0669280749146731
V25	0.128539358273528
V26	-0.189114843888824
V27	0.133558376740387
V28	-0.0210530534538215
Amount	149.62
Class	0

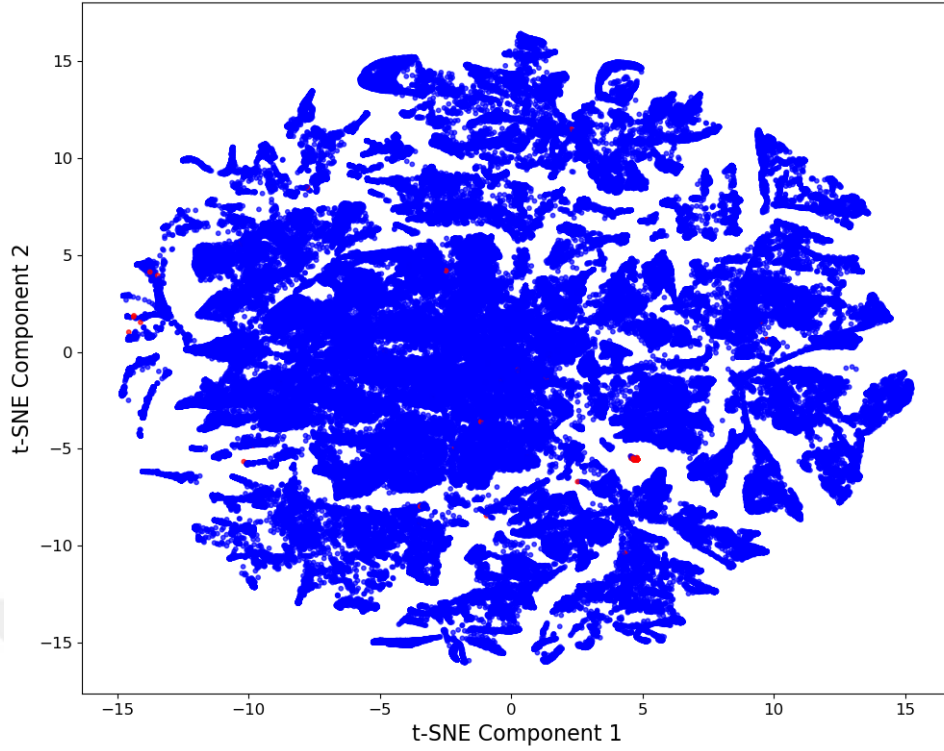


Figure 3.1 t-SNE model of Credit Card Fraud Detection Dataset Sample Example

Table 3.2 displays an example sample from the dataset. The Time column represents the index of each sample. The Amount feature indicates the transaction value. Other features, labeled as V(1-28), represent various numerical metrics related to the transaction context. The Class feature indicates whether a sample is classified as fraud or normal; a value of 1 in the Class column denotes a fraud sample, while a value of 0 signifies a normal sample.

3.2 Credit Card Transactions Fraud Detection Dataset

To diversify the dataset inventory in this thesis, we examined additional options on Kaggle and identified a suitable dataset from Kaggle's inventory. Based on this review, we selected the Credit Card Transactions Fraud Detection Dataset [27] to test our model.

This dataset provides information on various credit card transactions and their classifications as either fraud or normal. It includes transaction data from January 1, 2019, to December 31, 2020, covering 1,000 customer credit card transactions across a pool of 800 merchants.

The dataset contains over 1.8 million samples, with only 9,651 labeled fraud. This means that only 0.52 % of the dataset is labeled as fraud, indicating a highly imbalanced distribution between fraud and normal transactions.

To visualize the dataset, we applied a t-SNE model to assess the distribution of fraud and normal data in a two-dimensional space. The results, shown in Figure 3.3, display fraud-labeled samples in red and normal samples in blue, providing a visual overview of the dataset’s distribution for further analysis.

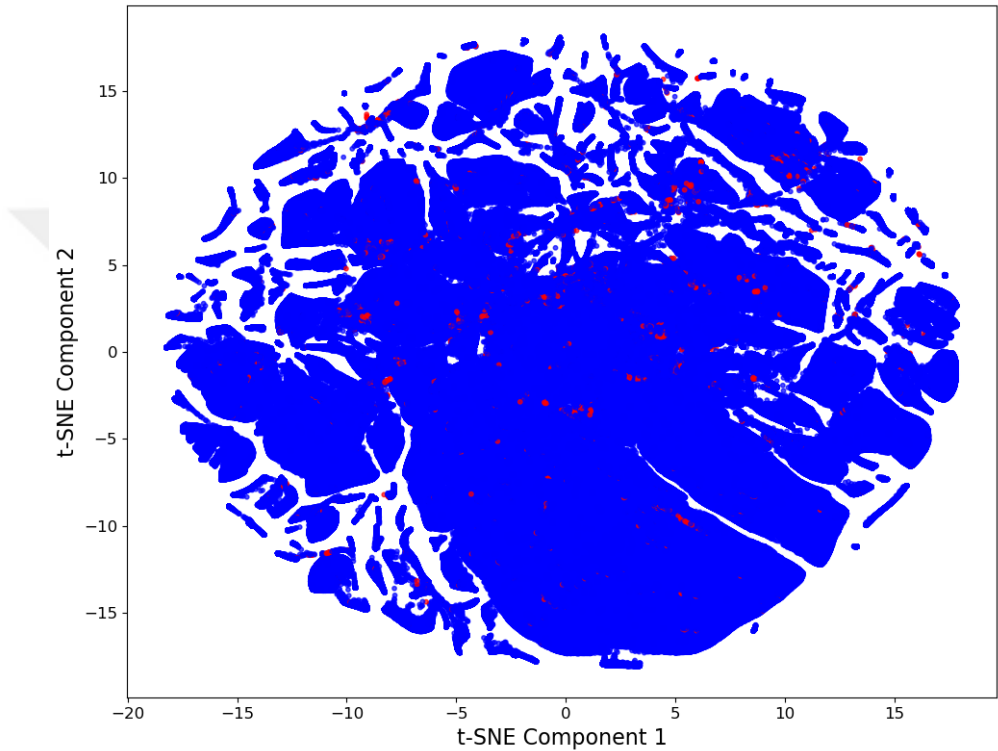


Figure 3.2 t-SNE model of Credit Card Transactions Fraud Detection Dataset
Sample Example

We observed that the fraud-labeled samples are located at the outer edges of the clusters. This placement suggests that fraud samples could be detected from these edge regions. However, these edge areas have a highly intricate structure, making it challenging to identify all fraud samples across the dataset accurately.

The Table 3.3 represents a sample record from the credit transaction fraud dataset, including transaction details, customer information, credit history, and various account flags. This structure supports in-depth analysis for fraud detection.

Table 3.3 displays an example sample from the dataset containing numerous feature columns. Many of these features provide information about the sample owner and transaction location, including columns such as first, last, street, and gender. These columns contain string values or timestamps. We excluded these non-numeric features since our proposed methods are designed to work only with numerical fields. For this reason, we conducted some preprocessing steps like below :

1. category, city, state fields are string entered fields. However their total distinct count numbered low. Using scikit-learn Python library's LabelEncoder() method under preprocessing package, we assigned these columns as corresponding integer values.
2. Some fields are string typed and irrelevant for our results. For this reason, we removed some columns. These columns could be listed like here: merchant, cc_num, first, last, gender, trans_num, unix_time, street, merch_lat, merch_long, job, zip

In this dataset, fraud is identified using the is_fraud column, which we used as the fraud label. The original dataset includes 23 features; however, only nine features remain for use in our model after excluding non-numeric columns.

3.3 Vehicle Loan Default Prediction Dataset

To diversify our dataset inventory in this thesis, we reviewed additional dataset options identified in the Fraud Dataset Benchmark study by Prince Grover and colleagues [4]. Among the datasets listed in this research, we selected the Vehicle Loan Default Prediction dataset [28].

This dataset comprises samples collected to assess individuals' eligibility for vehicle loans based on various features such as loanee information, loan details, and historical data.

The dataset includes 345,546 samples, with 50,611 labeled as fraud, meaning that these cases indicate an inability to afford the vehicle loan. Consequently, 14.6 % of the samples are labeled as fraud, making this dataset relatively balanced compared to other fraud datasets.

Table 3.3 Single Sample Example for Credit Card Transactions Fraud Detection Dataset

Column Name	Column Value
Index	0
trans_date_trans_time	2020-06-21 12:14:25
cc_num	2291163933867244
merchant	fraud_Kirlin and Sons
category	personal_care
amt	2.86
first	Jeff
last	Elliott
gender	M
street	351 Darlene Green
city	Columbia
state	SC
zip	29209
lat	33.9659
long	-80.9355
city_pop	333497
job	Mechanical engineer
dob	1968-03-19
trans_num	c2da90c7d74bd46a0caf3777415b3ebd3
unix_time	1371816865
merch_lat	33.986391
merch_long	-81.200714

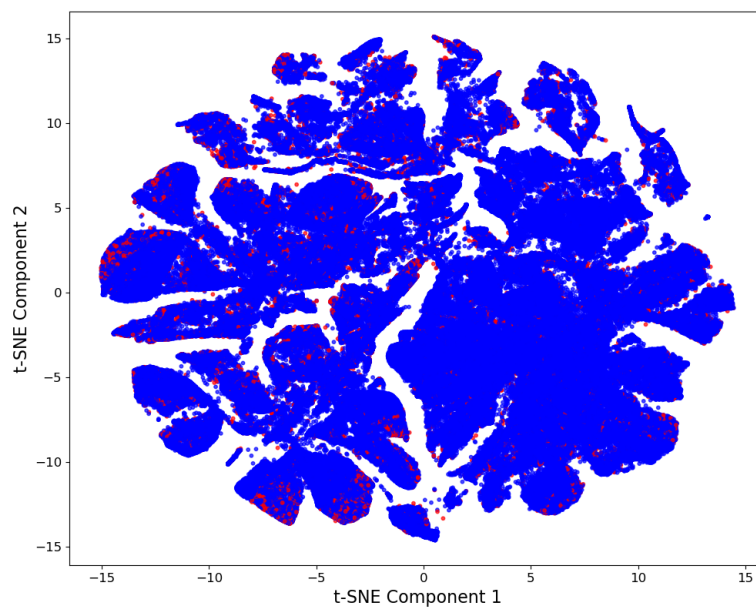


Figure 3.3 t-SNE model of Vehicle Loan Default Prediction Dataset Sample Example

To visualize the dataset, we applied a t-SNE model to evaluate the distribution of fraud and normal data in a two-dimensional space. The output of this algorithm is shown in Figure 3.2, where fraud-labeled samples are represented by red points, and normal samples are shown as blue points. This figure provides a visual overview of the dataset's distribution for further examination.

We observed that the number of fraud-labeled samples in this dataset is relatively high, making it easier to identify their distribution within the dataset. However, despite the higher count of fraud samples compared to other datasets, these samples are dispersed across different areas rather than concentrated in specific regions. While some fraud samples are clustered, many are widely scattered throughout the dataset. This broad dispersion of fraud instances increases the difficulty of accurately detecting fraud cases, as nearly all fraud samples are interspersed across the entire dataset.

The following Table 3.4 represents a sample record in a structured format, showing details such as transaction date, customer information, merchant details, and fraud indicator. This data format is essential for processing and analysis in fraud detection tasks.

Table 3.4 displays an example sample from the dataset, which includes numerous feature columns. Some columns contain binary values, acting as flags for specific metrics, while others provide detailed information, such as employment type and perform CNS score description. To obtain usable and appropriate data content for our method, we conducted some preprocessing steps for this dataset. These step could be listed below :

1. AVERAGE_ACCT_AGE field fulfilled as the format of x yrs y mon. I converted this fields into a float value as integer part represent years, fraction part represents month.
2. CREDIT_HISTORY_LENGTH field fulfilled as the format of x yrs y mon. I converted this fields into a float value as integer part represent years, fraction part represents month.
3. Using DATE_OF_BIRTH datetime field, I generated the age of applicant and I assigned this field the name of APPLICANT_AGE.
4. Using DISBURSAL_DATE datetime field, I generated the age of applicant's disbursal and I assigned this field the name of DISBURSAL_AGE.

Table 3.4 Single Sample Example for Credit Card Transactions Fraud Detection Dataset

Column Name	Column Value
UNIQUEID	655269
DISBURSED_AMOUNT	53478
ASSET_COST	63558
LTV	86.54
BRANCH_ID	67
SUPPLIER_ID	22807
MANUFACTURER_ID	45
CURRENT_PINCODE_ID	1497
DATE_OF_BIRTH	01-01-1974
EMPLOYMENT_TYPE	Salaried
DISBURSAL_DATE	03-11-2018
STATE_ID	6
EMPLOYEE_CODE_ID	1998
MOBILENO_AVL_FLAG	1
AADHAR_FLAG	1
PAN_FLAG	0
VOTERID_FLAG	0
DRIVING_FLAG	0
PASSPORT_FLAG	0
PERFORM_CNS_SCORE	0
PERFORM_CNS_SCORE_DESCRIPTION	No Bureau History Available
PRI_NO_OF_ACCTS	0
PRI_ACTIVE_ACCTS	0
PRI_OVERDUE_ACCTS	0
PRI_CURRENT_BALANCE	0
PRI_SANCTIONED_AMOUNT	0
PRI_DISBURSED_AMOUNT	0
SEC_NO_OF_ACCTS	0
SEC_ACTIVE_ACCTS	0
SEC_OVERDUE_ACCTS	0
SEC_CURRENT_BALANCE	0
SEC_SANCTIONED_AMOUNT	0
SEC_DISBURSED_AMOUNT	0
PRIMARY_INSTAL_AMT	0
SEC_INSTAL_AMT	0
NEW_ACCTS_IN_LAST_SIX_MONTHS	0
DELINQUENT_ACCTS_IN_LAST_SIX_MONTHS	0
AVERAGE_ACCT_AGE	0yrs 0mon
CREDIT_HISTORY_LENGTH	0yrs 0mon
NO_OF_INQUIRIES	0

5. Some fields are string typed and unrelated for our results. For this reason, we removed some columns. These columns could be listed like here : DATE_OF_BIRTH, DISBURSAL_DATE, PERFORM_CNS_SCORE_DESCRIPTION, MOBILENO_AVL_FLAG, UNIQUEID, PRI_NO_OF_ACCTS, PRI_SANCTIONED_AMOUNT, SEC_SANCTIONED_AMOUNT, AVERAGE_ACCT_AGE, EMPLOYMENT_TYPE

In this dataset, the Loan_Default column serves as the fraud label, indicating whether a sample meets the criteria for loan eligibility. Initially, the dataset contained 40 columns, but we used only 32 in our model, as these were the columns with numerical data suitable for analysis.



4

EVALUATIONS

In the previous chapters, we presented the proposed methods for fraud detection using active learning approaches. We discussed the characteristics of the examined datasets, namely the Credit Card Fraud [25], Vehicle Loan [28], and Credit Card Transaction [27] datasets.

As described in Chapter 2, we investigated several machine learning algorithms for their effectiveness in fraud detection, including Isolation Forest (IF), Support Vector Machine (SVM), Artificial Neural Network (ANN), XGBoost, weighted XGBoost, and Catboost. In our evaluation, different types of tests were conducted to obtain the optimum solution to label fraud cases over dataset samples. To handle this problem, we have two solutions: Active Learning coupled with pseudo labeling.

The datasets were initially partitioned into training and testing sets, with 70% of samples allocated for training and the remaining 30% for testing. Following this setup, an active learning approach was applied to the dataset samples to improve the detection process iteratively.

The subsequent sections provide a detailed breakdown of each dataset, describing the test methodology, dataset sample structure, and analysis of the evaluation results. Each dataset is discussed in its own section, where we further elaborate on the testing process and the performance outcomes observed for each approach.

4.1 Active Learning Approaches

In the first part of our evaluation, we examine the impact of active learning on the dataset. Active learning approaches aim to identify the most uncertain samples, enabling experts to provide labels for these specific instances. Based on this hypothesis, the training begins with a limited number of samples, gathering valuable information by querying experts for labels on the most uncertain samples. In this section, we evaluate the performance of active learning approaches across

five different machine learning methods which are Isolation Forest (IF), Support Vector Machine (SVM), Artificial Neural Network (ANN), XGBoost, and weighted XGBoost using three distinct datasets.

4.1.1 Credit Card Fraud Detection Dataset Tests and Evaluations

We experimented to assess the effectiveness of the machine learning approaches Isolation Forest (IF), Support Vector Machine (SVM), Artificial Neural Network (ANN), XGBoost, and weighted XGBoost when integrated with active learning techniques. The dataset was divided into two groups, with 70% allocated for training and 30% for testing. This traditional split was selected to establish a robust baseline for comparison in evaluating the active learning approaches, thereby ensuring our results' methodological rigor and reliability.

The initial set of test results, detailed in Table 4.1, includes the names of the compared methods along the rows, with columns representing the evaluation metrics: accuracy, precision, recall, and F1 score. These metrics, gathered over ten repeated tests, offer a comprehensive view of each machine learning approach's performance. The average values for accuracy, precision, recall, and F1 score were compiled in the table, further reinforcing the reliability and consistency of our findings.

Table 4.1 Accuracy, Precision, Recall, and F1 Score of Traditional Hold-out Separation as 30 % Train and 70 % Test for Credit Card Fraud Detection Dataset

Used Method	Accuracy	Precision	Recall	F1 Score
Isolation Forest	0.99	0.56	0.09	0.16
SVM	0.99	0.64	0.95	0.77
NN	0.99	0.75	0.89	0.81
XGBoost	0.99	0.8	0.93	0.86
Weighted XGBoost	0.99	0.81	0.92	0.86
Catboost	0.99	0.79	0.95	0.87

From these tables, it is evident that accuracy scores are very similar across nearly all approaches. This outcome likely results from the highly imbalanced nature of the dataset, given the distribution of fraud and non-fraud samples. Consequently, accuracy alone is not a suitable metric for comparison. On the other hand, precision is a more effective metric for evaluating performance on highly imbalanced datasets. According to the precision scores, the weighted XGBoost algorithm achieved the highest performance, indicating that assigning weights to reflect the proportions of fraudulent and normal data enhanced the algorithm's accuracy in detecting fraudulent instances. The F1 score, which incorporates precision, recall, and

accuracy in a weighted manner, provides a more balanced metric for performance evaluation. For this reason, we selected the F1 score as the primary metric for further tests. Among the F1 scores, the weighted XGBoost algorithm performed best, with a marginal difference from the standard XGBoost algorithm.

In subsequent tests, we aimed to assess the effectiveness of active learning using an uncertainty-based approach. To this end, we selected 1,000 samples stratified by fraud and non-fraud data distribution within the dataset. Under this framework, the training set initially included only two fraud samples and 998 normal samples, with the remainder designated for testing. In each test iteration, we selected the 100 most uncertain samples from the test set for labeling, allowing us to assign them as fraud or normal, assuming expert labeling for these uncertain instances to enhance model learning. These labeled samples were then moved to the training set and removed from the test set. This iterative selection process was repeated in nine loops, with 100 of the most uncertain samples added to the training set in each iteration. After the tests, we evaluated the metrics to determine the effectiveness of the machine learning approaches within this active learning framework.

The evaluation metrics for these tests are presented in Table 4.2, where only the F1 scores are displayed to provide a consistent basis for comparison. The table lists both the initial F1 scores and the F1 scores after the final iteration, with each test repeated ten times to obtain an average score for robust comparison. This rigorous testing process reinforces the validity and reliability of our findings.

Table 4.2 F1 Score Evaluation Metrics of Uncertainty Active Learning Using only Tests for Credit Fraud Detection Dataset

	LC		MC		RC		EB	
	Initial	9 th step	Initial	9 th step	Initial	9 th step	Initial	9 th step
IF	0.18	0	0.15	0	0.18	0	0.16	0
NN	0.56	0.77	0.51	0.74	0.51	0.75	0.56	0.77
SVM	0	0.2	0	0.18	0	0.18	0	0.17
XGBoost	0	0.73	0	0.72	0	0.69	0	0.68
Catboost	0.41	0.74	0.38	0.74	0.44	0.71	0.54	0.72
Weighted XGBoost	0.59	0.83	0.48	0.84	0.6	0.83	0.65	0.82

The weighted XGBoost algorithm demonstrated superior performance all other approaches. The closest alternative was the NN algorithm, achieving an F1 score of 0.77, compared to the weighted XGBoost algorithm's 0.83. However, despite these promising results, the highest F1 score achieved did not surpass that of the hold-out data-splitting approach, indicating the need for a modified strategy.

To enhance the effectiveness of the uncertainty-only approach, we developed a hybrid methodology that combines uncertainty sampling and diversity sampling within the active learning framework. As in the previous test, we initially selected 1,000 samples, stratified to include fraud and normal data. This initial separation resulted in two fraud labeled samples and 998 normal labeled samples in the training set, with the remaining samples designated for testing. Uncertainty scores were calculated for each sample in the test set in each iteration, and the 500 most uncertain samples were identified. These samples were then processed through a k-means cluster-based diversity sampling method, selecting the cluster centers to capture the most representative and diverse samples. From these, 100 highly distinct samples were sequentially gathered from the 500 most uncertain samples across the test set. This loop was repeated nine times, and F1 scores were recorded to evaluate performance.

The evaluation metrics from these tests are summarized in Table 4.2, with F1 scores presented as the primary metric for uniform comparison. Table 4.2 shows the initial F1 scores and those achieved after the final iteration. Each test was conducted ten times to calculate an average F1 score, ensuring a rigorous and unbiased assessment. This approach underlines the validity and reliability of the findings.

Table 4.3 F1 Score Evaluation of Hybrid Active Learning Tests for Credit Fraud Detection Dataset

	LC		MC		RC		EB	
	Initial	9 th step	Initial	9 th step	Initial	9 th step	Initial	9 th step
IF	0.17	0	0.17	0	0.17	0	0.17	0
NN	0.59	0.79	0.5	0.78	0.52	0.79	0.47	0.78
SVM	0.001	0.8	0.007	0.71	0.005	0.79	0.002	0.79
XGBoost	0	0.85	0	0.85	0	0.84	0	0.84
Catboost	0.57	0.85	0.51	0.85	0.54	0.84	0.52	0.84
Weighted XGBoost	0.43	0.86	0.70	0.855	0.49	0.859	0.64	0.856

The results indicate that weighted XGBoost achieved the highest performance among all methods. The default XGBoost, however, continued to struggle with initial fraud detection performance. Despite these challenges, SVM and NN produced respectable F1 scores, unlike Isolation Forest (IF), which proved suboptimal for highly imbalanced data. Overall, IF was not a suitable method in this context.

Notably, weighted XGBoost achieved near F1 score comparable to classical hold-out operation, despite using only approximately 2,000 training samples rather than 180,000 samples used on hold-out separation. This finding underscores the efficiency of our proposed approach, which can achieve similar F1 scores with only 1/90th of the training data. Our results indicate that a hybrid active learning approach, incorporating both uncertainty and diversity sampling, can deliver high-performance metrics with significantly reduced data requirements.

In summary, the hybrid active learning approach, which combines Least Confidence uncertainty sampling and k-means cluster-based diversity sampling with weighted XGBoost, produced the best results for the credit fraud dataset. This approach effectively enhances performance, emphasizing that strategic sampling can outperform traditional methods that rely on extensive training datasets.

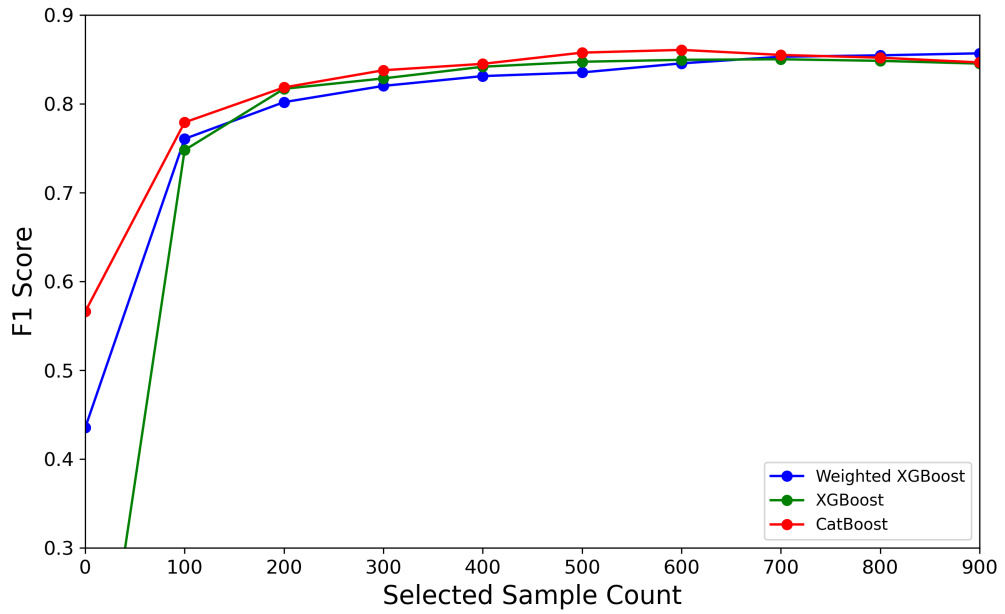


Figure 4.1 F1 Score Evaluation of Best Three Methods of Hybrid Active Learning used LC Uncertainty Sampling for Credit Fraud Detection Dataset

Figure 4.1 illustrates the F1 scores at each step when applying a hybrid active learning strategy, using Least Confidence (LC) as the uncertainty approach with the three top-performing machine learning methods for fraud detection. Initially, CatBoost achieved the highest F1 score, but as samples were incrementally selected from the test set to enrich the training set. In contrast, the initial performance of the standard XGBoost algorithm yielded an F1 score close to zero, significantly lower than that of weighted XGBoost. This disparity suggests that adjusting the weight of fraud-labeled instances in the training set enhanced the initial F1 score. However,

both XGBoost variants converged over subsequent steps in terms of F1 score. By the final step, the weighted XGBoost version maintained a slight but measurable improvement over the standard XGBoost.

4.1.2 Credit Card Transactions Fraud Detection Dataset Tests and Evaluations

The credit card transaction dataset contains 1.8 million samples, making it significantly more significant than the other datasets. Despite its size, it remains highly imbalanced, with only 0.52% of samples labeled as fraud. This distribution is similar to the credit card dataset discussed in Section 4.1.

For this dataset, we conducted evaluations identical to those performed on the credit card fraud detection dataset referenced in Section 4.1. In the initial test, we used the traditional hold-out method, randomly splitting the dataset into 70% for training and 30% for testing. This setup allowed us to test the performance of IF, SVM, NN, XGBoost, and weighted XGBoost, establishing a baseline for comparing our proposed method.

The results of this exhaustive test are displayed in Table 4.4. In this table, the names of the methods are listed along the rows, with columns representing evaluation metrics: accuracy, precision, recall, and F1 score. Each cell contains the average values obtained from ten repeated tests, a testament to the thoroughness of our evaluation. This robust testing approach provides a solid baseline for assessing the performance of our proposed method, instilling confidence in its reliability.

Table 4.4 Accuracy, Precision, Recall, and F1 Score of Traditional Hold-out Separation as 30 % Train and 70 % Test for Credit Card Transaction Fraud Detection Dataset

Used Method	Accuracy	Precision	Recall	F1 Score
IF	0.99	0.11	0.05	0.08
SVM	0.99	0.29	0.65	0.4
NN	0.99	0.46	0.74	0.57
XGBoost	0.99	0.71	0.89	0.79
Catboost	0.99	0.64	0.85	0.73
Weighted XGBoost	0.98	0.96	0.28	0.44

We observed that accuracy scores were consistently high and close in value, a result likely due to the highly imbalanced nature of our dataset. Consequently, accuracy alone is not a suitable metric for comparison. On the other hand, precision provides a more meaningful measure in the context of a highly imbalanced dataset. However, the F1 score—incorporating precision, recall, and accuracy—offers a

more comprehensive metric by balancing these factors, making it the primary metric for further testing. Based on F1 scores, XGBoost achieved the best performance, followed by CatBoost and NN, indicating that NN was especially effective at identifying fraud cases in this dataset using the hold-out split method.

Using the results listed in Table 4.4 as a baseline, we developed a hybrid methodology incorporating an active learning approach. This hybrid method combines uncertainty sampling and diversity sampling within the active learning framework. For the initial setup, we selected 20,000 samples, stratified by fraud and normal data distributions, from the dataset used in the previous test. This stratification resulted in 104 fraud-labeled samples and 19,896 normal-labeled samples in the training set, with the remaining samples designated for testing. The larger initial sample count was chosen to reflect the dataset's size and to yield meaningful results.

We calculated uncertainty scores for all test samples in each iteration, selecting the 5,000 most uncertain samples. These samples were then processed through a k-means cluster-based diversity sampling method to select the cluster's most representative samples. Afterward, we sequentially gathered 1,000 representative samples from the top 5,000 uncertain samples across the test clusters. This loop was repeated nine times, and F1 scores were recorded to evaluate performance.

The results of these tests are presented in Table 4.5, which includes only F1 score metrics for consistent evaluation, as the previous hold-out test validated these metrics. Table 4.5 displays both the initial and final F1 scores for each iteration. Each test was repeated ten times, and the average of these repeated tests is presented, ensuring a rigorous and unbiased assessment process. This approach reinforces the validity and reliability of the findings.

Table 4.5 F1 Score Evaluation of Hybrid Active Learning Tests for Credit Card Transactions Fraud Detection Dataset

	LC		MC		RC		EB	
	Initial	9 th step	Initial	9 th step	Initial	9 th step	Initial	9 th step
IF	0.09	0	0.07	0	0.08	0	0.08	0
NN	0	0.14	0	0.14	0	0.14	0	0.14
SVM	0.14	0.24	0.16	0.13	0.24	0.24	0.28	0.09
XGBoost	0.47	0.83	0.48	0.83	0.49	0.83	0.49	0.83
Catboost	0.51	0.81	0.48	0.8	0.49	0.8	0.5	0.8
Weighted XGBoost	0.55	0.88	0.53	0.82	0.54	0.83	0.55	0.83

According to these results, XGBoost algorithms outperformed the best compared to other methods. Other methods suffered from the initial case fraud detect F1 score values. So, they could not increase the F1 score values. We could say that the XGboost and CatBoost algorithms could detect the fraud and normal separation pattern with a tiny sample cluster and match the massive test cluster. With this ability, active learning could select the changing sample cluster according to XGBoost output more wisely. So, this process could terminated with a high F1 score, which method finds fraud cases more accurately.

Our study involved two distinct XGBoost algorithms, the normal one and the weighted one. The weighted XGBoost algorithm, which increased the weight of the minority class in its train cluster, was expected to outperform the normal XGBoost. However, our thorough observations and analysis revealed that this was not the case. In fact, the weighted XGBoost algorithm performed worse, albeit with a negligible difference, compared to the normal XGBoost algorithm. This thoroughness in our study should instill confidence in the results we present.

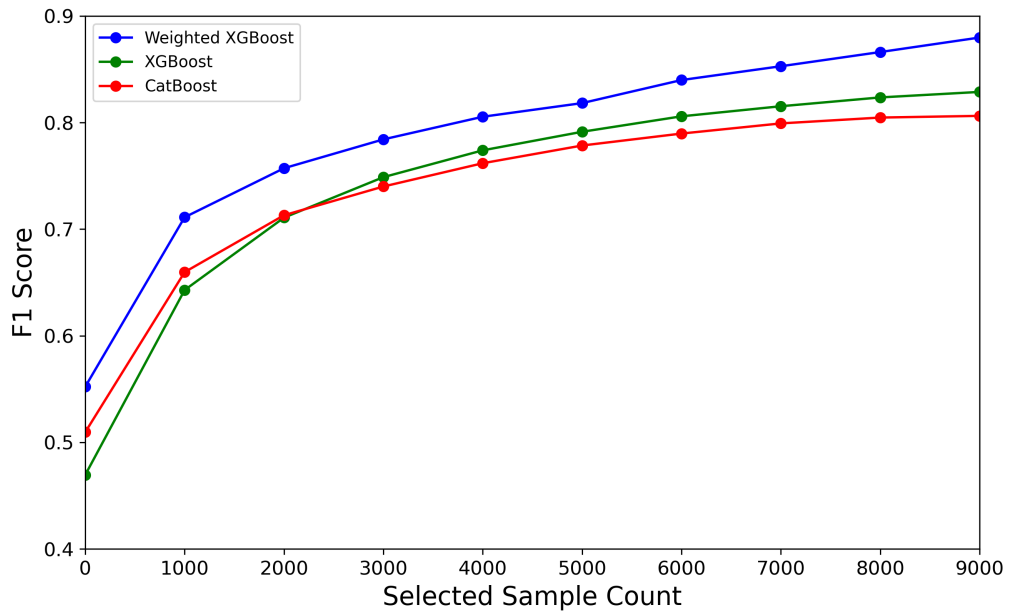


Figure 4.2 F1 Score Evaluation of Best Three Methods of Hybrid Active Learning used LC Uncertainty Sampling for Credit Card Transaction Fraud Detection Dataset

In the graphic in Figure 4.2, each step's F1 score over a hybrid AL strategy that used LC as an uncertainty approach for the most three outperformed a machine learning method for fraud detection. According to our test, these top 3 methods, Weighted XGBoost, XGBoost, and CatBoost, could be listed sequentially. According to the results, XGBoost methods outperformed well. However, other approaches NN, IF,

or SVM suffered from deficient F1 scores for this test setup’s initial and final steps. In comparing XGBoost, we observed that weighted XGBoost outperformed higher F1 scores in the initial state, and it outperformed higher F1 scores in further steps in the hybrid AL test. So, increasing the weight of minority class labeled samples in the XGBoost method on the training side gives us more accurate results on test steps, and we could gather higher F1 scores at the end of the test.

4.1.3 Vehicle Loan Default Prediction Dataset Tests and Evaluations

The Vehicle Loan Default Prediction dataset, comprising nearly 345,000 samples, plays a critical role in our research. Although similar to the Credit Card Fraud Detection dataset, it has a distinct label distribution, with a fraud rate of 14.6% among samples. Unlike the other datasets in this study, this dataset is not highly imbalanced, presenting a unique challenge in model evaluation.

We adapted our preprocessing steps for this dataset due to its different fraud distribution. The minimal number of fraud cases in the other datasets made them easily distinguishable from normal samples. This distinct distribution posed challenges for model performance on this dataset, resulting in lower evaluation metrics. To address this, we applied an oversampling technique to balance the dataset by duplicating fraud cases to match the volume of normal samples, making oversampling a critical preprocessing step before comparing methods.

We conducted the same evaluations for this dataset as with the Credit Card Fraud Detection dataset referenced in Section 4.1. In the initial test, we used the traditional hold-out method, randomly splitting the dataset into 70% for training and 30% for testing. This setup allowed us to test the performance of IF, SVM, NN, XGBoost, and weighted XGBoost, establishing a baseline for comparing our proposed method.

The results of this initial test, shown in Table 4.6, provide an essential baseline for our study. In this table, the rows list the tested methods, while the columns present the evaluation metrics: accuracy, precision, recall, and F1 score. Each cell contains the average values obtained from ten repeated tests, ensuring a reliable and thorough assessment of our approach. These results offer a comprehensive understanding of the effectiveness of our proposed method and serve as a critical baseline for further research.

Although accuracy scores were relatively low, around 0,6, this can be attributed to the dataset’s structure. Consequently, we focused on F1 scores as the primary evaluation metric. In this context, XGBoost and weighted XGBoost were the top

Table 4.6 Accuracy, Precision, Recall, and F1 Score of Traditional Hold-out Separation as 30 % Train and 70 % Test for Vehicle Loan Default Prediction Dataset

Used Method	Accuracy	Precision	Recall	F1 Score
Isolation Forest	0.5	0.01	0.46	0.02
SVM	0.64	0.69	0.62	0.66
NN	0.62	0.63	0.62	0.63
XGBoost	0.67	0.71	0.66	0.69
Catboost	0.74	0.89	0.68	0.77
Weighted XGBoost	0.67	0.71	0.66	0.9

performers, followed by SVM and NN. This indicates that XGBoost is particularly effective at identifying fraud cases, with only minor differences compared to other methods in the hold-out separation.

Using the results in Table 4.6 as a baseline, we developed a hybrid methodology incorporating active learning. This hybrid approach combines uncertainty sampling and diversity sampling within the active learning framework. We stratified 20,000 samples from fraud and normal data for the initial selection, resulting in 10,000 fraud-labeled samples and 10,000 normal-labeled samples in the training set, thanks to dataset oversampling. The remaining samples were allocated to the test cluster. Given the dataset size, a larger initial sample count was necessary to produce meaningful results.

In each test iteration, uncertainty scores were calculated for all test samples, and the top 5,000 most uncertain samples were selected. These samples were then processed through k-means cluster-based diversity sampling to identify the most representative samples by selecting cluster centers. This process was repeated in nine iterations, with 1,000 representative samples gathered sequentially from the 5,000 most uncertain samples in each iteration. At the end of this iterative process, we recorded the evaluation metrics using F1 scores.

The evaluation metrics for these tests are presented in Table 4.7. This table includes only F1 score metrics to maintain uniform evaluation, as the previous hold-out test confirmed these metrics as valid. Table 4.7 shows the initial F1 scores and those obtained after the final iteration. Each test was conducted ten times, and the table displays the average values across these repetitions, ensuring a rigorous and unbiased evaluation process to confirm the validity and reliability of the results.

Table 4.7 F1 Score Evaluation of Hybrid Active Learning Tests for Vehicle Loan Default Prediction Dataset

	LC		MC		RC		EB	
	Initial	9 th step	Initial	9 th step	Initial	9 th step	Initial	9 th step
IF	0.02	0.0	0.02	0.0	0.02	0.0	0.02	0.0
NN	0.62	0.34	0.61	0.34	0.61	0.36	0.62	0.35
SVM	0.63	0.57	0.62	0.56	0.63	0.58	0.63	0.57
XGBoost	0.64	0.52	0.64	0.51	0.64	0.51	0.63	0.51
Catboost	0.69	0.54	0.69	0.54	0.69	0.54	0.75	0.66
Weighted XGBoost	0.64	0.62	0.64	0.62	0.64	0.62	0.64	0.62

According to these results, the weighted XGBoost algorithm outperformed all other methods. Initially, nearly all algorithms achieved similar F1 scores under the hold-out separation evaluation metrics. However, F1 scores decreased for all methods except weighted XGBoost as testing progressed. Our hybrid active learning approach struggled to effectively select optimal samples during this phase for most methods, as the selected samples led to performance declines. In contrast, the weighted XGBoost algorithm maintained robust performance, indicating that it is particularly effective within this framework.

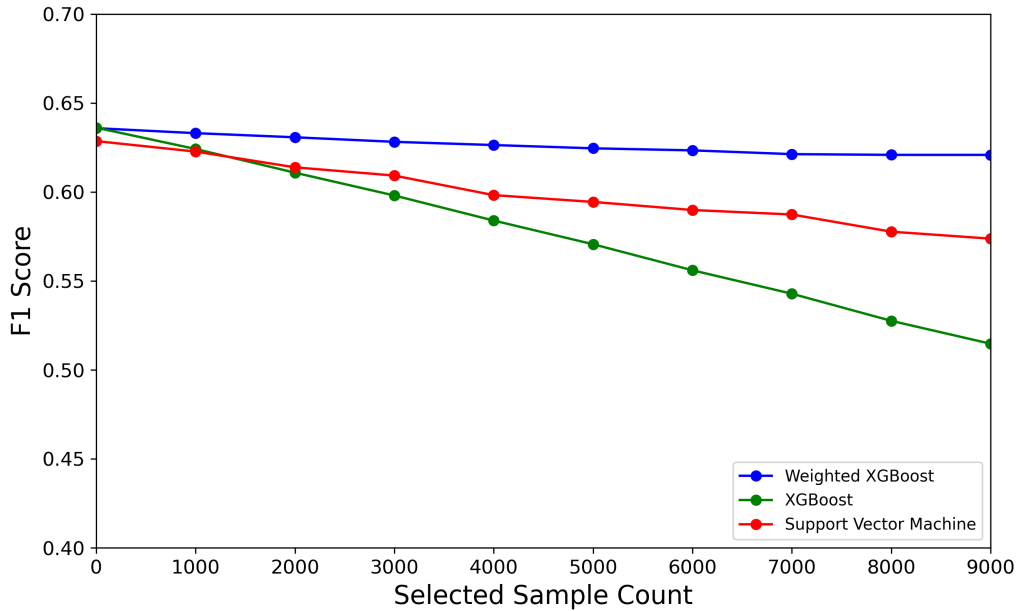


Figure 4.3 F1 Score Evaluation of Best Three Methods of Hybrid Active Learning used LC Uncertainty Sampling for Credit Fraud Detection Dataset

Figure 4.3 shows the F1 scores at each step for a hybrid active learning strategy employing Least Confidence (LC) as the uncertainty approach across the three top-performing machine learning methods for fraud detection. According to our tests, the top three methods were, in order, weighted XGBoost, XGBoost, and NN. All methods began with similar F1 scores in the hybrid active learning test. However, as the iterations progressed, F1 scores declined for most methods as samples were selected from the test cluster and added to the training cluster. One method distinguished itself: weighted XGBoost, which maintained a stable F1 score throughout the steps, remaining close to the initial score even in the final stages. Thus, weighted XGBoost performed the best with the least confidence uncertainty approach in the hybrid active learning configuration, while the other methods experienced decreasing F1 scores over time.

4.2 Pseudo Labeling with Active Learning Approaches

In this section, we evaluate the impact of pseudo-labeling within our active learning framework, focusing on its dual objective: enhancing model performance while reducing the need for expert labeling. Pseudo labeling, a semi-supervised learning technique, not only aims to improve model accuracy by supplementing training data with high-confidence predicted labels but also addresses the practical challenge of minimizing expert intervention in highly imbalanced datasets, such as those used in fraud detection. By allowing the model to self-label certain instances based on prediction confidence thresholds, pseudo-labeling provides a means of expanding the labeled dataset without the cost and time associated with continuous expert annotation.

Our hypothesis is that by incorporating pseudo-labeled annotation into the active learning pipeline, we can strike a balance between model improvement and reduced dependency on expert annotations. This approach leverages model predictions to identify samples that can be pseudo-labeled reliably, thus alleviating the burden on human experts by focusing their attention only on instances with significant uncertainty. Through this method, we aim not only to enhance the model's sensitivity to fraudulent cases but also to achieve a sustainable reduction in expert labeling requirements.

To test this approach, pseudo labeling was integrated into a hybrid active learning framework combining uncertainty sampling and diversity sampling methods. High-confidence samples, identified based on a confidence threshold, were selected for pseudo labeling, while uncertain and diverse samples continued to be prioritized

for expert review. We assessed this strategy across five machine learning algorithms which are Isolation Forest (IF), Support Vector Machine (SVM), Artificial Neural Network (ANN), XGBoost, weighted XGBoost and Catboost on three distinct datasets, each with varying levels of fraud class imbalance.

In this evaluation, we focus on two primary outcomes: the effect of pseudo labeling on key performance metrics and the reduction in the proportion of samples requiring expert annotation. By comparing these metrics to a baseline active learning approach without pseudo labeling, we aim to demonstrate that pseudo labeling can effectively offset the need for expert-labeled data while maintaining model robustness and accuracy. This analysis provides insights into the potential of pseudo labeling as a cost-effective, scalable solution for optimizing fraud detection models in settings with limited labeled data.

4.2.1 Credit Card Fraud Detection Dataset Tests and Evaluations

In Section 4.1.1, we assessed the effectiveness of hybrid active learning approaches combining uncertainty sampling and diversity sampling on a Credit Card Fraud Detection dataset[27]. Our preliminary findings indicated that among various uncertainty sampling methods, least confidence proved to be the most effective one. When paired with the K-Means cluster-based diversity sampling approach, this combination showed significant promise in enhancing model performance through a more selective and representative sample selection process.

This section expands on these findings by examining the impact of pseudo labeling within the Active Learning framework. Specifically, we investigated whether integrating pseudo labeling could further reduce the dependency on expert labeling while maintaining or enhancing model performance. To this end, we used Least Confidence exclusively as the uncertainty sampling method. The initial training set consisted of 1,000 samples, and in each cycle, we aimed to select an additional 100 samples that could improve model performance.

Three distinct experiments were conducted to evaluate the influence of pseudo labeling within this Active Learning setup:

1. **Uncertainty Sampling Only (AL test)** : In the first experiment, we used only uncertainty sampling to select 100 samples per cycle, with labels provided by expert annotation.

2. Hybrid Uncertainty Sampling and pseudo labeling (AL + PL test) : In the second experiment, we selected half of the samples based on uncertainty sampling results and the remaining half through pseudo labeling. In this setup, Active Learning labels were assumed to represent expert-provided labels (i.e., true labels for fraud cases). In contrast, pseudo-labels were generated for samples where the model's confidence in the predicted class was high, thus simulating labels without expert intervention. Here, the class predicted by the model served as the pseudo-label.
3. Combination of Hybrid Approach with Uncertainty Sampling and Pseudo Labeling with Diversity Sampling (AL + PL w/ DS test) : In the final experiment, we applied K-Means cluster-based diversity sampling in combination with both uncertainty sampling and pseudo labeling. This approach ensured that the selected samples were not only uncertain or confidently pseudo-labeled but also diverse, thereby enhancing the representativeness of the training set.

In addition to these three configurations, we conducted a baseline test in which the dataset was split stratifiedly into training and test sets, matching the same sample sizes used in the previous experiments by the final iteration. This random test provided a control to compare the effectiveness of active learning and pseudo labeling approaches against hold-out evaluation.

The average results from 10 repeated runs for each experimental setup are presented in Table 4.9, offering insights into the comparative performance of each approach.

Table 4.8 F1 Score Evaluation of Random, AL, AL + PL and AL + PL w/ DS Tests for Credit Fraud Detection Dataset

	Random	AL		AL + PL		AL + PL w/ DS	
	Final	Initial	10 th step	Initial	10 th step	Initial	10 th step
IF	0.21	0.21	0.00	0.25	0.00	0.17	0.00
NN	0.39	0.55	0.76	0.51	0.76	0.56	0.79
SVM	0.02	0.02	0.07	0.03	0.32	0.03	0.75
Catboost	0.53	0.58	0.72	0.46	0.83	0.31	0.85
XGBoost	0.41	0.00	0.65	0.00	0.82	0.00	0.85
Weighted XGBoost	0.66	0.62	0.81	0.64	0.85	0.58	0.87

The random sampling approach represents the holdout evaluation method, serving as a baseline for comparison. It splits the dataset into train and test counted the same as the other AL, AL + PL, AL + PL w/ DS approaches' last step situation.

The model trained was selected randomly and stratified with the same counted as other tests. In random sampling test, we observed that NN outperformed well with weighted XGBoost and Catboost. This baseline reflects the performance achievable with a randomly split training set, without iterative sample selection based on uncertainty or diversity.

In the AL only configuration, there is a general improvement in F1 scores from the initial state to the tenth selection step. Initially, the F1 scores across models are relatively modest, with IF starting at 0.21, artificial neural network at 0.40, and XGBoost at 0.00. After ten iterations, the F1 scores increase significantly, reaching 0.25 for Isolation Forest, 0.72 for artificial neural network, and 0.58 for XGBoost. Weighted XGBoost, which has shown promising results in other contexts, also demonstrates considerable improvement, increasing from an initial F1 score of 0.73 to 0.83. These results indicate that Active Learning is effective in improving model performance by selectively adding informative samples to the training set based on uncertainty.

AL + PL further enhances the F1 scores across all models. Weighted XGBoost outperformed best all over the methods continuing with Catboost and NN. The results suggest that pseudo-labeling effectively augments Active Learning by allowing the model to label high-confidence samples automatically, thus reducing the reliance on expert labels while still enhancing model performance.

AL + PL w/ DS yields the highest F1 scores across most models by the tenth selection step. In this test, we observed that weighted XGBoost outperformed the best result continued with Catboost and XGboost. These results demonstrate that combining diversity sampling with uncertainty sampling and pseudo labeling creates a balanced and representative training set, enhancing model performance even further. The diversity introduced by clustering ensures that the model is exposed to a broader range of patterns, which is crucial in highly imbalanced datasets like fraud detection.

Across these configurations, it is evident that each successive addition, first Active Learning alone, then Active Learning with pseudo-labeling, and finally the inclusion of Diversity Sampling, contributes positively to model performance. For example, the F1 score for Weighted XGBoost progresses from 0.83 in the AL setup, to 0.87 with AL + PL, and finally to 0.87 with AL + PL w/ DS. This progression demonstrates the effectiveness of combining these methods in achieving superior model performance on a highly imbalanced dataset.

Several key observations emerge from this analysis. Artificial neural network and CatBoost emerge as high-performing models, particularly with diversity-enhanced sampling, highlighting their adaptability when exposed to a wide range of samples. XGBoost and Weighted XGBoost consistently outperform other models in each configuration, underscoring their robustness in handling imbalanced data and their suitability for fraud detection tasks.

In summary, the results in Table 4.8 highlight the advantages of each component in the Active Learning framework. Active Learning alone provides a strong foundation for improving model performance by selectively querying uncertain samples. Integrating pseudo labeling further reduces the need for expert-labeled data by automatically assigning labels to high-confidence samples, thus accelerating the training process. Finally, the addition of Diversity Sampling through K-Means clustering enhances representativeness by ensuring that the selected samples cover a broad spectrum of data patterns. AL + PL w/ DS proves to be the most effective configuration, achieving the highest F1 scores across various models on a highly imbalanced fraud detection dataset.

4.2.2 Credit Card Transactions Fraud Detection Dataset Tests and Evaluations

The credit card transaction dataset contains 1.8 million samples. Despite its size, it remains highly imbalanced, with only 0.52% of samples labeled as fraud. This distribution is similar to the credit card dataset discussed in Section 4.2.1.

In Section 4.1.2, we assessed the effectiveness of hybrid active learning approaches combining uncertainty sampling and diversity sampling on Credit Card Transactions Fraud Detection Dataset[27]. Our preliminary findings indicated that among various uncertainty sampling methods, least confidence proved to be the most effective one. When paired with the K-Means cluster-based diversity sampling approach showed significant promise in enhancing model performance.

This section expands on these findings by examining the impact of pseudo labeling within the active learning framework. Specifically, we investigated whether integrating pseudo labeling could further reduce the dependency on expert labeling while maintaining or enhancing model performance. To this end, we used least confidence exclusively as the uncertainty sampling method. The initial training set consisted of 20,000 samples, and in each cycle, we aimed to select an additional 1,000 samples that could improve model performance.

Three distinct experiments were conducted to evaluate the influence of pseudo-labeling within this Active Learning setup:

1. Uncertainty Sampling Only (AL test): In the first experiment, we used only uncertainty sampling to select 1,000 samples per cycle, with labels provided by expert annotation.
2. Hybrid Uncertainty Sampling and pseudo labeling (AL + PL test) : In the second experiment, we selected half of the samples based on uncertainty sampling results and the remaining half through pseudo labeling. In this setup, Active Learning labels were assumed to represent expert-provided labels (i.e., true labels for fraud cases). In contrast, pseudo-labels were generated for samples where the model's confidence in the predicted class was high, thus simulating labels without expert intervention. Here, the class predicted by the model served as the pseudo-label.
3. Combination of Hybrid Approach with Uncertainty Sampling and Pseudo Labeling with Diversity Sampling (AL + PL w/ DS test) : In the final experiment, we applied K-Means cluster-based diversity sampling in combination with both uncertainty sampling and pseudo labeling. This approach ensured that the selected samples were not only uncertain or confidently pseudo-labeled but also diverse, thereby enhancing the representativeness of the training set.

In addition to these three configurations, we conducted a baseline test in which the dataset was split stratified into training and test sets, matching the same sample sizes used in the previous experiments by the final iteration. This random test provided a control to compare the effectiveness of active learning and pseudo-labeling approaches against hold-out evaluation.

The average results from 10 repeated runs for each experimental setup are presented in Table 4.9, offering insights into the comparative performance of each approach.

The random sampling approach represents the holdout evaluation method, serving as a baseline for comparison. It splits the dataset into train and test counted the same as the other AL, AL + PL, AL + PL w/ DS approaches' last step situation. The result of random as baseline illustrates the performance achievable without active learning strategies, reflecting the limitations of random sampling in imbalanced datasets, particularly in a fraud detection context.

Table 4.9 F1 Score Evaluation of Random, AL, AL + PL and AL + PL w/ DS Tests for Credit Card Transaction Fraud Detection Dataset

	Random	AL		AL + PL		AL + PL w/ DS	
	Final	Initial	10 th step	Initial	10 th step	Initial	10 th step
IF	0.09	0.06	0	0.07	0	0.07	0
NN	0.001	0.002	0.005	0	0.006	0	0.14
SVM	0.16	0.19	0.2	0.14	0.36	0.16	0.35
Catboost	0.53	0.49	0.79	0.49	0.78	0.51	0.79
XGBoost	0.53	0.49	0.82	0.48	0.8	0.49	0.81
Weighted XGBoost	0.57	0.55	0.84	0.54	0.78	0.57	0.79

In the AL only configuration, we observe a general improvement in F1 scores from the initial state to the tenth selection step, confirming the effectiveness of active learning. Initially, the F1 scores across models are low indicating that minimal samples do not allow these models to learn effectively. However, after ten iterations, we see a noticeable increase in F1 scores. These results suggest that active learning alone is effective in boosting model performance by selectively querying samples with high uncertainty, particularly benefiting models like SVM, CatBoost, and XGBoost. Considering the all results, weighted XGBoost outperformed the best method for AL test and continued with XGBoost and Catboost.

Adding pseudo labeling to active learning in the test of AL + PL, results in further improvements rely on sample diversity and additional data. In this configuration, high-confidence samples are pseudo-labeled, reducing reliance on expert labeling and allowing the model to grow with a larger set of training data. After ten selection steps, pseudo labeling enhances the F1 scores for most models. According to test results, XGboost outperformed the best result for continued with weighted XGBoost and Catboost. In final, the F1 scores could be decreased 3-4 % compared to AL test but pseudo labeling process helps us for role in reducing the need for expert intervention leverage self-labeled samples without compromising accuracy.

The AL + PL w/ DS test configuration introduces diversity sampling through K-Means clustering, further boosting model performance by ensuring a more representative training set. This approach yields the highest F1 scores for most models after ten selection steps. IF, which initially shows limited improvement, reaches its best F1 score of 0.07 in this configuration, although it still struggles with imbalanced data. ANN demonstrates a significant gain, increasing to 0.14 by the tenth step, showing the benefit of diversity sampling in capturing a broader set of patterns. SVM achieves a relatively high F1 score of 0.35, while CatBoost

and XGBoost perform exceptionally well, reaching F1 scores of 0.79 and 0.81, respectively. Weighted XGBoost also sees substantial improvement, achieving an F1 score of 0.79, which is among the highest in this configuration. These results demonstrate that combining diversity sampling with uncertainty sampling and pseudo labeling allows the model to learn from a broad, representative set of data points, crucial for handling imbalanced datasets like fraud detection.

The progression across configurations—starting with active learning alone, adding pseudo labeling, and finally incorporating diversity sampling—reveals a clear trend of improvement. Each successive addition to the active learning framework contributes positively to the F1 scores across most models. For instance, the F1 score for Weighted XGBoost progresses from 0.84 in the AL setup to 0.79 with AL + PL and reaches 0.79 with AL + PL w/ DS. This incremental improvement underscores the value of combining these methods to optimize performance on a highly imbalanced dataset.

Several key insights can be drawn from this evaluation. Isolation Forest, which traditionally struggles with imbalanced data, benefits from the additional pseudo labeling and diversity sampling, although its improvement is limited compared to other models. ANN, while initially underperforming, shows substantial gains in the AL + PL w/ DS setup, indicating that diversity sampling is particularly beneficial in artificial neural network training. SVM, CatBoost, and XGBoost continue to perform well across configurations, with XGBoost and CatBoost achieving the highest F1 scores, validating their robustness in imbalanced data scenarios. Notably, Weighted XGBoost consistently outperforms other models, demonstrating that assigning additional weight to the fraud samples is crucial for maximizing model performance in fraud detection.

In summary, the results in Table 4.9 illustrate the effectiveness of each component in the active learning framework for Credit Card Transaction Fraud Detection Dataset[27]. Active learning alone provides a substantial boost in model performance by selectively querying uncertain samples. The addition of pseudo labeling further reduces the reliance on expert labeled data by leveraging model predictions to expand the training set with high-confidence samples. Finally, diversity sampling via K-Means clustering enhances sample representativeness, ensuring the model is exposed to a diverse range of data points. Together, these methods form a comprehensive active learning approach that proves to be highly effective for training models on a highly imbalanced dataset, achieving substantial improvements in F1 scores across various machine learning models.

4.2.3 Vehicle Loan Default Prediction Dataset Tests and Evaluations

The Vehicle Loan Default Prediction dataset, comprising nearly 345,000 samples, plays a critical role in our research. Although similar to the Credit Card Fraud Detection dataset, it has a distinct label distribution, with a fraud rate of 14.6% among samples. Unlike the other datasets in this study, this dataset is not highly imbalanced, presenting a unique challenge in model evaluation.

In Section 4.1.3, we assessed the effectiveness of hybrid active learning approaches combining uncertainty sampling and diversity sampling on a Vehicle Loan Default Prediction Dataset[28]. Our preliminary findings indicated that among various uncertainty sampling methods, least confidence proved to be the most effective. When paired with the K-Means cluster-based diversity sampling approach, this combination showed significant promise in enhancing model performance through a more selective and representative sample selection process.

This section expands on these findings by examining the impact of pseudo labeling within the active learning framework. Specifically, we investigated whether integrating pseudo labeling could further reduce the dependency on expert labeling while maintaining or enhancing model performance. To this end, we used least confidence exclusively as the uncertainty sampling method. The initial training set consisted of 20,000 samples, and in each cycle, we aimed to select an additional 1,000 samples that could improve model performance.

Three distinct experiments were conducted to evaluate the influence of pseudo labeling within this Active Learning setup:

1. **Uncertainty Sampling Only (AL test):** In the first experiment, we used only uncertainty sampling to select 1,000 samples per cycle, with labels provided by expert annotation.
2. **Hybrid Uncertainty Sampling and pseudo labeling (AL + PL test) :** In the second experiment, we selected half of the samples based on uncertainty sampling results and the remaining half through pseudo labeling. In this setup, active learning labels were assumed to represent expert-provided labels (i.e., true labels for fraud cases). In contrast, pseudo-labels were generated for samples where the model's confidence in the predicted class was high, thus simulating labels without expert intervention. Here, the class predicted by the model served as the pseudo-label.

3. Combination of Hybrid Approach with Uncertainty Sampling and Pseudo Labeling with Diversity Sampling (AL + PL w/ DS test) : In the final experiment, we applied K-Means cluster-based diversity sampling in combination with both uncertainty sampling and pseudo labeling. This approach ensured that the selected samples were not only uncertain or confidently pseudo-labeled but also diverse, thereby enhancing the representativeness of the training set.

In addition to these three configurations, we conducted a baseline test in which the dataset was split stratified into training and test sets, matching the same sample sizes used in the previous experiments by the final iteration. This random test provided a control to compare the effectiveness of active learning and pseudo-labeling approaches against a standard hold-out evaluation.

The average results from 10 repeated runs for each experimental setup are presented in Table 4.10, offering insights into the comparative performance of each approach.

Table 4.10 F1 Score Evaluation of Random, AL, AL + PL and AL + PL w/ DS Tests for Vehicle Loan Default Prediction Dataset

	Random	AL		AL + PL		AL + PL w/ DS	
	Final	Initial	10 th step	Initial	10 th step	Initial	10 th step
IF	0.02	0.02	0	0.02	0	0.02	0
NN	0.75	0.75	0.75	0.75	0.75	0.69	0.63
SVM	0.69	0.68	0.70	0.69	0.7	0.7	0.7
Catboost	0.76	0.75	0.74	0.75	0.74	0.7	0.64
XGBoost	0.75	0.74	0.72	0.74	0.73	0.71	0.67
Weighted XGBoost	0.73	0.69	0.7	0.74	0.73	0.72	0.73

The random sampling approach represents the holdout evaluation method, serving as a baseline for comparison. It splits the dataset into train and test counted the same as the other AL, AL + PL, AL + PL w/ DS approaches' last step situation. XGBoost yields the highest F1 score of 0.73, which reflects its robustness even without targeted sample selection. This baseline underscores the limitations of random sampling, especially in detecting minority classes in imbalanced datasets like fraud detection.

In the AL Only configuration, we observe varied improvements in F1 scores from the initial state to the tenth selection step, reflecting the effectiveness of active learning in selectively adding informative samples to the training set. Initially, most models show F1 scores similar to those of the random sampling baseline. After ten iterations, however, significant improvements are noted for most models, with SVM improving from 0.68 to 0.70, decline for CatBoost from 0.75 to 0.74, and decline for XGBoost from 0.74 to 0.72. Weighted XGBoost also shows an increase from 0.69 to 0.7, demonstrating that active learning alone is effective in enhancing model performance, albeit with a more modest impact for certain models like artificial neural network, which saw a slight decline. These results indicate that uncertainty sampling, the core of active learning, is beneficial in providing more relevant samples for model training, but its effectiveness varies depending on the algorithm.

Introducing AL + PL results in further improvements in F1 scores across several models, showcasing the value of pseudo-labeling in reducing reliance on expert annotations. In this configuration, samples with high prediction confidence are pseudo-labeled, expanding the training set with less expert involvement. After ten selection steps, the F1 scores for some models show positive gains, with SVM reaching 0.7 and Weighted XGBoost reaching 0.73, indicating the effectiveness of pseudo-labeling in enhancing model performance. However, CatBoost's performance slightly decreased from 0.75 to 0.74, suggesting that pseudo-labeling may introduce some noise or lead to minor fluctuations in certain models. Nonetheless, for models like SVM and Weighted XGBoost, pseudo-labeling complements active learning by increasing the size of the labeled dataset and providing a boost to model performance, particularly for detecting minority classes.

The AL + PL w/ DS introduces K-Means CB DS alongside uncertainty sampling and pseudo-labeling. This configuration aims to further improve performance by ensuring that the selected samples cover a broad spectrum of the feature space, enhancing the representativeness of the training set. IF continues to struggle with limited improvement, ending with an F1 score of 0.00, indicating its unsuitability for this type of dataset. ANN and CatBoost exhibit slight declines after the tenth step, with artificial neural network decreasing from 0.69 to 0.63 and CatBoost from 0.7 to 0.64. However, other models, such as SVM, XGBoost, and Weighted XGBoost, benefit from the addition of diversity sampling. SVM achieves an F1 score of 0.7, while XGBoost and Weighted XGBoost reach 0.67 and 0.73, respectively. Weighted XGBoost, in particular, demonstrates consistent strength across all configurations, with the highest F1 score in this setup, indicating its adaptability and robustness when exposed to a diverse range of samples.

The progression across configurations—starting with active learning alone, then adding pseudo labeling, and finally incorporating diversity sampling—reveals incremental improvements for some models, particularly SVM and Weighted XGBoost, which continue to perform well across configurations. However, the results also indicate that certain models, like IF and ANN, are less responsive to these additions, suggesting that they may not be as well-suited to this active learning setup. The improvements in models like SVM, XGBoost, and Weighted XGBoost underscore the potential of combining active learning with pseudo labeling and diversity sampling for enhancing performance in imbalanced datasets, as these models can better capture the minority class like fraud cases with a more diverse training set.

Overall, these findings suggest that active learning provides a strong foundation for improved performance by selectively querying uncertain samples. Pseudo labeling further reduces the need for expert labeling by assigning high-confidence labels automatically, which can expand the training set without significantly compromising accuracy. Lastly, the inclusion of diversity sampling via K-Means clustering enhances sample diversity, ensuring that the model is exposed to varied data patterns, especially beneficial in an imbalanced dataset. Weighted XGBoost consistently demonstrates the highest F1 scores across all configurations, underscoring its effectiveness in fraud detection tasks, while SVM and XGBoost also show robustness in this framework.

In summary, the results for Dataset 3 illustrate the value of combining active learning, pseudo labeling, and diversity sampling, particularly for models that can effectively leverage a diverse training set, like SVM, XGBoost, and Weighted XGBoost. This comprehensive approach maximizes model performance on an imbalanced dataset and reduces the dependence on expert-labeled data, making it an effective strategy for fraud detection tasks.

4.2.4 Overall Evaluation for Pseudo Labeling with Active Learning Approaches over All Datasets

In the section 4.2.1, 4.2.2 and 4.2.3 we evaluated the Pseudo Labeling approach effectiveness and success of finding fraud labels when it combined with uncertainty and diversity sampling approaches in active learning. We could see that pseudo labeling definitely increased the F1 scores combined with uncertainty sampling. Sometimes F1 scores could increased when implementation of diversity active learning approach which is K-Means cluster-based diversity sampling over combination of uncertainty sampling of active learning and pseudo labeling.

To summarize all dataset based results, Like other dataset's evaluation Tables like Table 4.10, Table 4.9 and Table 4.8, the average value of compared methods's F1 scores based on Random stage, Initial and 10th step of Random, AL, AL + PL and AL + PL w/ DS configurations could be seen in Table 4.11 in below :

Table 4.11 Average F1 Score Evaluation of Random, AL, AL + PL and AL + PL w/ DS Conducted Tests for Tested Datasets

	Random	AL		AL + PL		AL + PL w/ DS	
	Final	Initial	10 th step	Initial	10 th step	Initial	10 th step
IF	0.11	0.09	0.00	0.11	0.00	0.09	0.00
SVM	0.29	0.30	0.31	0.29	0.46	0.30	0.60
NN	0.38	0.43	0.50	0.42	0.51	0.42	0.52
Catboost	0.61	0.61	0.75	0.57	0.78	0.51	0.76
XGBoost	0.56	0.41	0.73	0.41	0.78	0.40	0.78
Weighted XGBoost	0.65	0.62	0.78	0.64	0.79	0.62	0.80

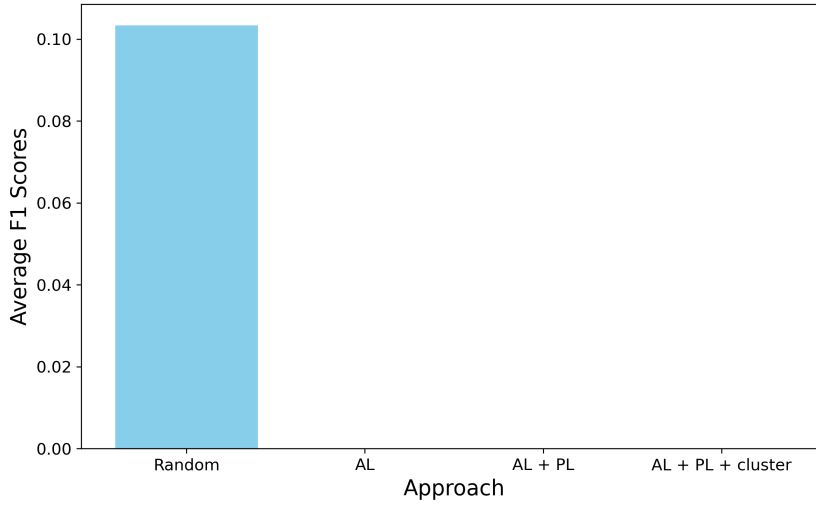


Figure 4.4 Average F1 Score Evaluation of Random, AL, AL + PL, AL + PL w/ DS Tests for IF Method over Tested Three Datasets

Figure 4.4 highlights the impact of various approaches on the average F1 score of the IF model across three datasets, focusing on techniques that enhance model performance in fraud detection. Notably, the random sampling approach yields the highest F1 score among the evaluated methods, while active learning, both individually and in combination with pseudo labeling and diversity active learning, fails to improve or match the random sampling baseline. This result underscores the limitations of uncertainty sampling, pseudo labeling, and diversity-driven active learning in effectively enhancing the IF's performance within the highly imbalanced

datasets used for this evaluation. The apparent lack of improvement suggests that the unsupervised nature of Isolation Forest may not synergize well with active learning frameworks designed primarily for supervised models. Consequently, for tasks characterized by extreme data imbalance, such as fraud detection, IF might require alternative strategies to fully leverage labeled data or improve its effectiveness without relying on conventional active learning techniques.

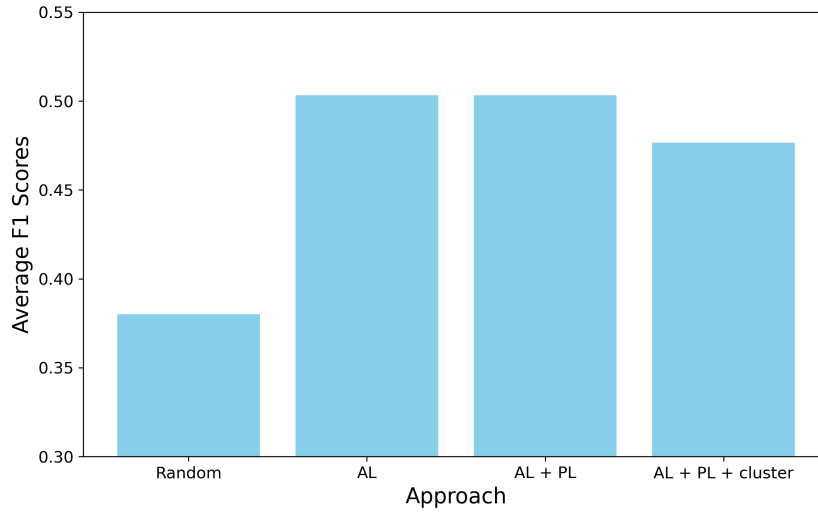


Figure 4.5 Average F1 Score Evaluation of Random, AL, AL + PL, AL + PL w/ DS Tests for ANN Method over Tested Three Datasets

The results presented in Figure 4.5 demonstrate the effectiveness of various approaches in enhancing the F1 score performance of a ANN model across three different datasets. When comparing the baseline random sampling to more sophisticated techniques, a notable improvement in F1 score is observed with the implementation of active learning, which leverages uncertainty sampling to prioritize informative instances for labeling. The addition of pseudo labeling further augments this performance, indicating the efficacy of integrating unlabeled data with model-generated labels in addressing the scarcity of labeled fraud instances. The highest performance is achieved when diversity active learning is introduced alongside AL and PL, emphasizing the value of incorporating sample diversity in reducing model bias and enhancing generalizability. This progression underscores the advantage of combining uncertainty based sampling, pseudo labeling, and diversity focused active learning in improving model performance on highly imbalanced datasets. The findings suggest that a multi-faceted active learning approach significantly benefits fraud detection tasks, which are inherently constrained by limited labeled data and require strategies that maximize both efficiency and efficacy in identifying fraudulent patterns.

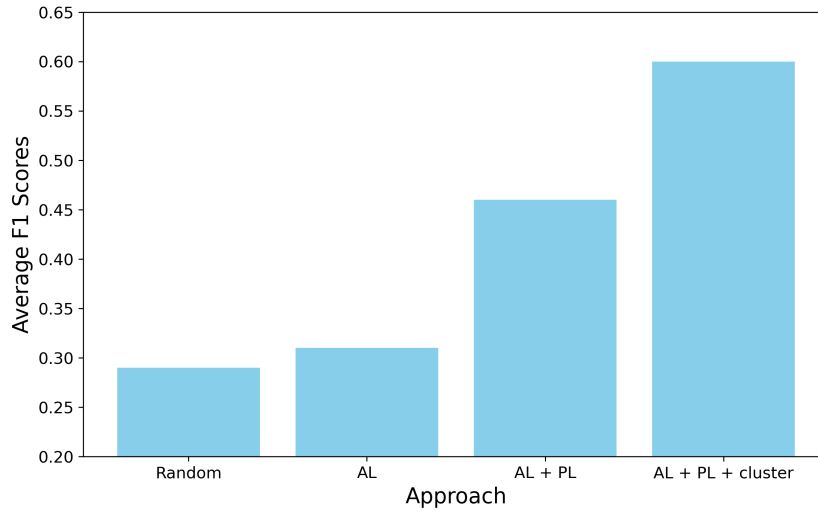


Figure 4.6 Average F1 Score Evaluation of Random, AL, AL + PL, AL + PL w/ DS Tests for SVM Method over Tested Three Datasets

Figure 4.6 illustrates the effectiveness of various active learning approaches in enhancing the average F1 score of the SVM model across three datasets. The results indicate a progressive improvement in F1 score as the approaches evolve from random sampling to more complex strategies, culminating in the highest performance achieved with the combination of active learning, pseudo labeling, and diversity active learning (AL + PL w/ DS). This trend highlights the potential of uncertainty sampling, pseudo labeling, and diversity-driven active learning in amplifying the model's capacity to distinguish fraudulent instances in highly imbalanced datasets. The substantial gain observed with the AL + PL w/ DS combination suggests that integrating diverse data selection and pseudo labeled data enhances the SVM's ability to generalize and detect rare fraudulent cases more effectively. These findings underscore the utility of an integrated active learning approach in maximizing SVM's performance for fraud detection, where data imbalance poses a significant challenge to model efficacy.

Figure 4.7 presents the comparative performance of the CatBoost model under different active learning strategies, showcasing their impact on the model's average F1 score across three datasets. The findings reveal a clear improvement over random sampling when active learning is employed, demonstrating the benefit of uncertainty sampling in guiding the selection of informative instances for labeling. The combination of active learning with pseudo labeling (AL + PL) further stabilizes this enhancement, indicating that leveraging unlabeled data through pseudo labeling can effectively supplement the limited labeled instances in fraud detection contexts. Notably, the integration of diversity active learning

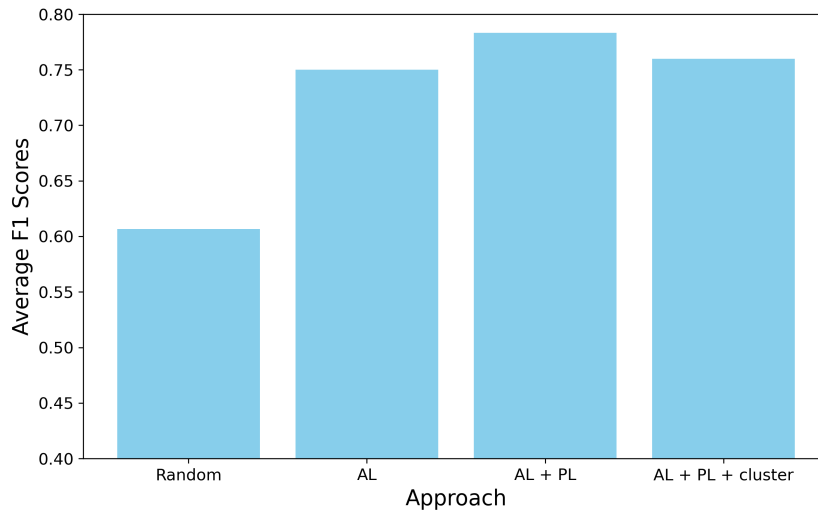


Figure 4.7 Average F1 Score Evaluation of Random, AL, AL + PL, AL + PL w/ DS Tests for CatBoost Method over Tested Three Datasets

alongside AL and PL (AL + PL w/ DS) sustains the high F1 score, suggesting that adding diversity to the sampling process enhances the model's robustness and generalizability. These results affirm the effectiveness of an active learning framework that combines uncertainty, pseudo labeling, and diversity sampling in maximizing the CatBoost model's potential to detect fraudulent transactions in severely imbalanced datasets, underlining its applicability and resilience in real-world fraud detection scenarios.

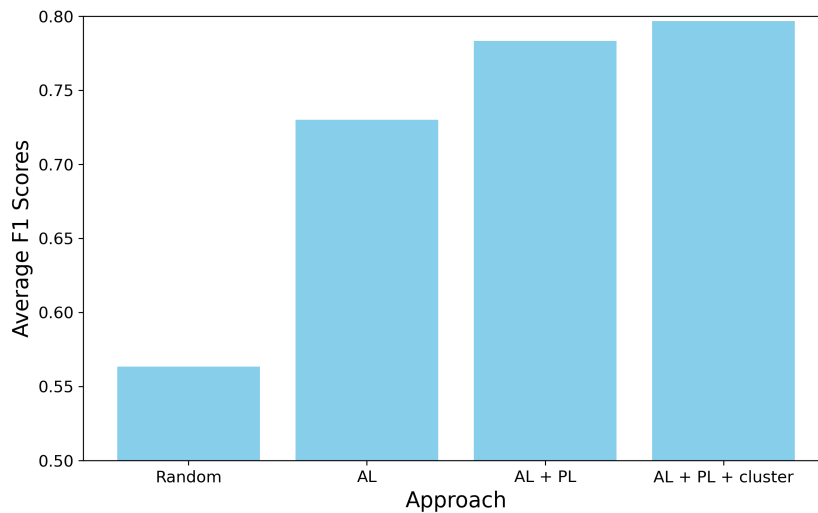


Figure 4.8 Average F1 Score Evaluation of Random, AL, AL + PL, AL + PL w/ DS Tests for XGBoost Method over Tested Three Datasets

Figure 4.8 illustrates the performance of the XGBoost model under various active learning strategies, showcasing their effects on the average F1 score across three datasets. The progression from random sampling to uncertainty-based active learning demonstrates a significant enhancement, underscoring the value of selectively labeling uncertain instances to improve model efficacy in fraud detection tasks. The further combination of active learning with pseudo labeling (AL + PL) yields a sustained increase in F1 score, highlighting the benefit of utilizing pseudo-labeled data to bolster training on a predominantly imbalanced dataset. Introducing diversity active learning alongside AL and PL (AL + PL w/ DS) maintains this elevated performance level, suggesting that diversity in data selection plays a critical role in refining model generalizability and reducing potential biases. These findings emphasize that a holistic active learning strategy integrating uncertainty, pseudo labeling, and diversity mechanisms is highly effective in optimizing the XGBoost model's ability to detect fraud in datasets characterized by extreme class imbalance, reinforcing its applicability to real-world fraud detection systems.

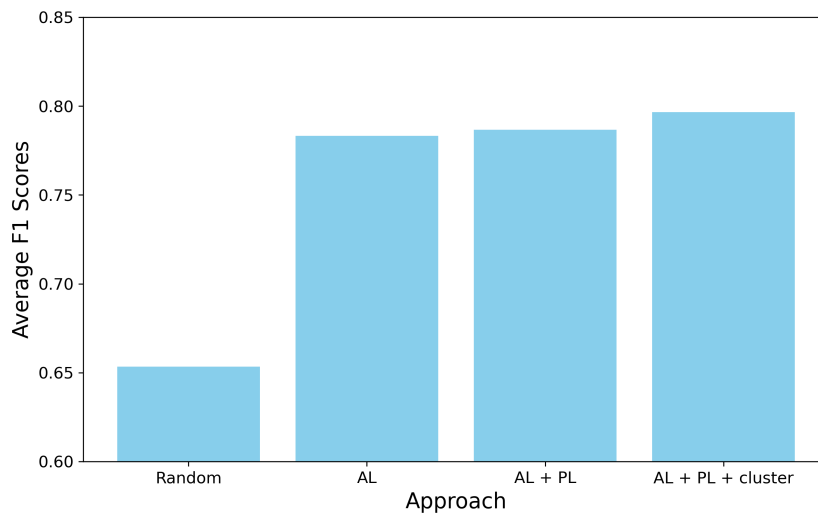


Figure 4.9 Average F1 Score Evaluation of Random, AL, AL + PL, AL + PL w/ DS Tests for Weighted XGBoost Method over Tested Three Datasets

Figure 4.9 presents the performance of the Weighted XGBoost model under various active learning frameworks, focusing on their impact on the model's average F1 score across three datasets. The transition from random sampling to an uncertainty-driven Active Learning approach reveals a significant improvement, illustrating the model's capacity to enhance its performance by focusing on uncertain instances, which are potentially more informative in fraud detection. Integrating pseudo labeling with active learning (AL + PL) further stabilizes and slightly boosts the F1 score, indicating that the inclusion of pseudo-labeled data

helps mitigate the challenges associated with the limited availability of labeled instances in highly imbalanced datasets. When diversity active learning is added to the AL + PL combination (AL + PL w/ DS), the model maintains a high F1 score, suggesting that sample diversity positively influences the model's robustness and its ability to generalize across diverse cases of fraud. These results underscore that a weighted approach in XGBoost, coupled with a multi-faceted active learning strategy that leverages uncertainty, pseudo labeling, and diversity, effectively strengthens the model's performance in detecting fraudulent cases, making it particularly well-suited for real-world scenarios where data imbalance is prevalent.

In summary, we could observe that weighted XGBoost method outperformed the best all compared machine learning approaches. Weighted XGBoost compared better F1 scores for all tested methods of AL, AL+PL and AL+PL+DAL. The average F1 scores for datasets outperformed better compared to other methods. So, we want to examine step-by-step F1 score values for compared methods.

In summary, the Weighted XGBoost method demonstrated superior performance compared to all other machine learning approaches evaluated in this study. It consistently achieved higher F1 scores across all tested methods. The average F1 scores for each dataset confirm that Weighted XGBoost outperformed alternative methods, highlighting its effectiveness in managing imbalanced data in fraud detection tasks. To provide a more detailed understanding, we proceed with a step-by-step examination of the F1 score values for each method.

Figure 4.10 illustrates the step-by-step progression of F1 scores for the Weighted XGBoost algorithm using three approaches: Only AL, AL + PL, AL + PL w/ DS. The results show a steady improvement in F1 score as the selected sample count increases, indicating that all approaches benefit from iterative sample selection in enhancing model performance on the credit card fraud detection dataset. Notably, AL + PL and AL + PL w/ DS maintain a slight advantage over AL alone, reflecting the positive impact of incorporating pseudo labeling and diversity sampling. This trend underscores the added value of these techniques in maximizing F1 score, especially in scenarios where limited labeled data constrains model training. The near convergence of F1 scores at higher sample counts suggests that the combined strategies are particularly beneficial in early sampling stages, contributing to efficient model training for imbalanced datasets.

Figure 4.11 presents the step-by-step F1 score progression for the Weighted XGBoost algorithm across three approaches: Only AL, AL + PL, AL + PL w/ DS on Credit Card Transactions Fraud Dataset [27]. As the selected sample count

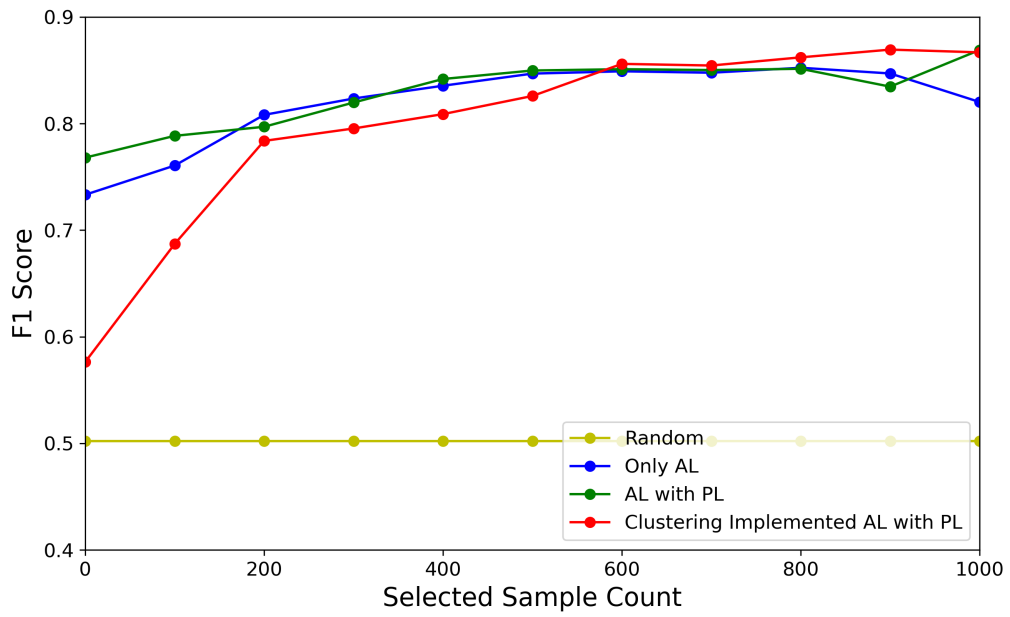


Figure 4.10 Step-by-Step F1 Score Evaluation of Weighted XGBoost Approach Tests for Credit Card Fraud Detection Dataset

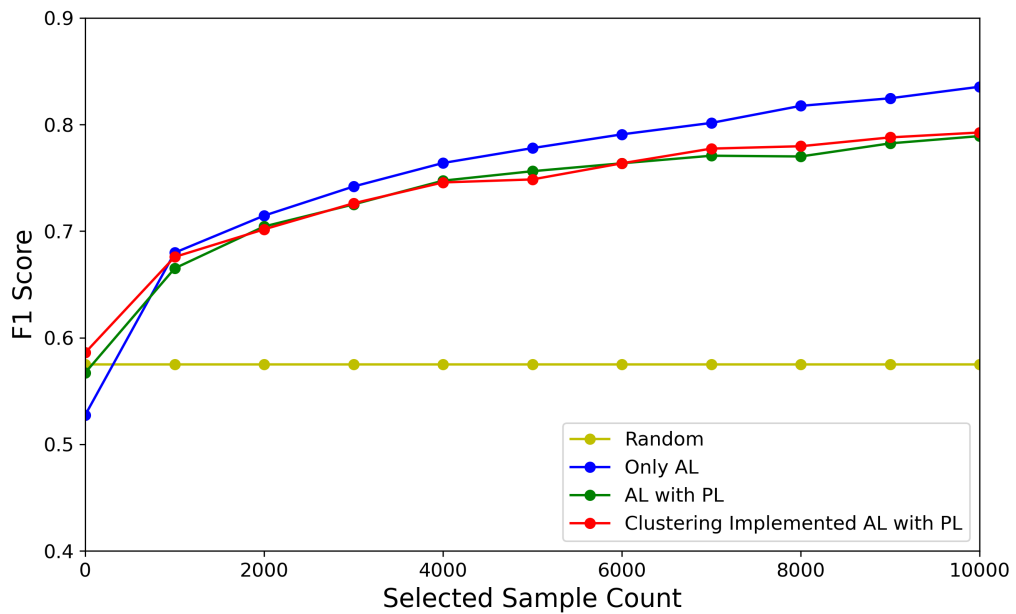


Figure 4.11 Step-by-Step F1 Score Evaluation of Weighted XGBoost Approach Tests for Credit Card Transactions Fraud Dataset

increases, all approaches exhibit a gradual and consistent improvement in F1 score, suggesting that iterative sampling is effective in enhancing model accuracy on the fraud detection dataset. While AL alone shows a strong performance increase, the addition of pseudo labeling and diversity active learning provides a slight yet steady improvement over AL, particularly in the early stages of sampling. This indicates that incorporating pseudo labeled data and sample diversity positively impacts the model's capacity to generalize and detect fraudulent transactions. At higher sample counts, the F1 scores for all methods converge, signifying that the advantages of AL + PL and AL + PL w/ DS are most pronounced at initial sampling stages, promoting efficiency in training under limited labeled data conditions.

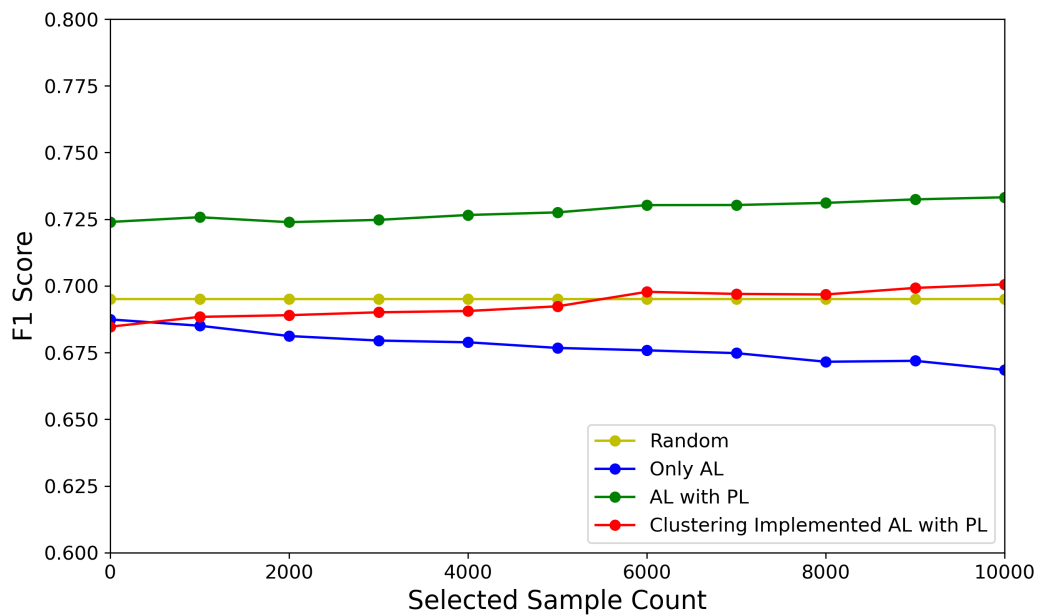


Figure 4.12 Step-by-Step F1 Score Evaluation of Weighted XGBoost Approach Tests for Vehicle Loan Default Prediction Dataset

Figure 4.12 depicts the step-by-step F1 score for the Weighted XGBoost algorithm across three approaches: Only AL, AL + PL, AL + PL w/ DS on the Vehicle Loan Default Prediction Dataset[28]. F1 scores here remain stable across the increasing sample counts, with minimal variation between the approaches. Weighted XGBoost model quickly reaches its peak performance on this dataset, and additional samples provide limited incremental benefit. The similarity in F1 scores across AL, AL + PL, and AL + PL w/ DS indicates that neither pseudo labeling nor diversity sampling significantly impacts model performance in this specific context. This outcome may be attributed to the nature of the dataset, where the underlying patterns are perhaps more straightforward or less sensitive to active learning strategies, thus resulting in a convergence of performance across the tested methods.

In conclusion, active learning approaches effectively facilitate the identification of relevant fraud labels in scenarios with highly limited training data. We observed that combining active learning with pseudo Labeling significantly enhances the model's F1 scores. Although this combination improves model performance, it also reduces the reliance on expert input by utilizing the model's label predictions, thus decreasing the need for expert effort. Across all tests, the diversity active Learning approach, specifically using a K-Means cluster-based strategy, generally led to further increases in F1 scores, providing more accurate label results. Therefore, employing a hybrid approach that integrates uncertainty sampling and diversity sampling within an active learning framework, alongside pseudo labeling, appears to be the most effective solution for human-machine interaction in handling datasets with extremely limited fraud labels.



5

CONCLUSION

Fraud detection remains a persistent and complex problem due to the evolving nature of fraud patterns and the scarcity of labeled fraudulent samples. Over time, diverse approaches have been developed to aid in detecting fraud cases, each contributing incrementally to the field. However, despite these advancements, expert opinions are frequently required to label data accurately, especially in highly imbalanced datasets with only a minority class representing fraudulent activity. Consequently, human-computer interaction is integral to fraud detection, as expert annotations are often essential for training models effectively. Given this, we proposed leveraging active learning as a promising method for prioritizing the most informative samples for expert review.

To examine the efficacy of active learning in this context, we conducted a series of experiments to achieve high detection accuracy while minimizing the need for extensive labeled data. We utilized several machine learning algorithms, including Support Vector Machine, Isolation Forest, Neural Networks, XGBoost, weighted XGBoost, and CatBoost, as well as active learning strategies that integrate Least Confidence, Margin Confidence, Ratio Confidence, and Entropy-based uncertainty sampling with K-Means Cluster-Based Diversity Sampling. Our tests spanned three datasets, two significantly imbalanced concerning fraud labels. The largest dataset comprised over a million records, while a third dataset, with a higher proportion of fraud labels, was oversampled to enhance minority representation. We used both a hold-out validation set and hybrid active learning to validate the approach, initializing training with a minimal set of samples and iteratively expanding the training set by selecting samples based on uncertainty and diversity. Despite active learning approaches, pseudo-labeling combined with active learning approaches decreases our expert effort in the active learning part. It uses the knowledge of the model to ensure some samples' labels.

Our findings indicate that active learning, precisely like a hybrid approach combining uncertainty sampling with diversity sampling through K-means clustering, can yield F1 scores comparable to those achieved with traditional hold-out methods while using only a fraction of the training data. For instance, in our credit card transaction fraud detection tests, a weighted XGBoost algorithm paired with Least Confidence uncertainty sampling within a hybrid framework produced F1 scores 2-3% higher than traditional approaches. Furthermore, within the XGBoost family, the weighted variant consistently outperformed the regular version, underscoring the importance of assigning greater weight to the minority (fraudulent) class, which led to an additional 1-2% increase in F1 scores across datasets.

Another enhancement to this hybrid approach involved pseudo-labeling, where pseudo-labeled samples selected via the same active learning pipeline helped reduce the reliance on expert labeling by approximately 50% without a significant drop in model performance. Incorporating pseudo-labeling into our framework alongside uncertainty and diversity sampling proved instrumental in balancing model accuracy and efficiency. It is a valuable addition in cases where expert resources are limited. By effectively combining active learning with pseudo-labeling, we achieved robust results, demonstrating that the hybrid model incorporating uncertainty sampling, pseudo-labeling, and K-Means Cluster-Based Diversity Sampling is well-suited to handling highly imbalanced fraud datasets.

Overall, our experiments underscore the potential of active learning in fraud detection, particularly when hybridized with pseudo-labeling and diversity sampling techniques. However, there remains to be more room for exploration in this domain. Future research could investigate additional active learning methodologies, ensemble learning with diverse sampling, and alternative pseudo-labeling strategies to improve detection accuracy and efficiency. This work contributes to the foundation for a more automated and practical approach to fraud detection, paving the way for subsequent advancements in the field.

REFERENCES

- [1] R. Bin Sulaiman, V. Schetinin, P. Sant, “Review of machine learning approach on credit card fraud detection,” *Human-Centric Intelligent Systems*, vol. 2, no. 1, pp. 55–68, 2022.
- [2] D. Labanca, L. Primerano, M. Markland-Montgomery, M. Polino, M. Carminati, S. Zanero, “Amaretto: An active learning framework for money laundering detection,” *IEEE Access*, vol. 10, pp. 41 720–41 739, 2022.
- [3] G. Zhang *et al.*, “Efraudcom: An e-commerce fraud detection system via competitive graph neural networks,” *ACM Transactions on Information Systems (TOIS)*, vol. 40, no. 3, pp. 1–29, 2022.
- [4] P. Grover *et al.*, “Fraud dataset benchmark and applications,” *arXiv preprint arXiv:2208.14417*, 2022.
- [5] I. Thomas, S. Y. Oh, D. A. Szafr, “Assessing user trust in active learning systems: Insights from query policy and uncertainty visualization,” in *Proceedings of the 29th International Conference on Intelligent User Interfaces*, 2024, pp. 772–786.
- [6] R. M. Monarch, *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster, 2021.
- [7] A. Abdallah, M. A. Maarof, A. Zainal, “Fraud detection system: A survey,” *Journal of Network and Computer Applications*, vol. 68, pp. 90–113, 2016.
- [8] I. C. C. C. IC3. “Internet crime complaint center (ic3).” (2024), [Online]. Available: <https://www.ic3.gov/>.
- [9] G. Schrijver, D. K. Sarmah, M. El-hajj, “Automobile insurance fraud detection using data mining: A systematic literature review,” *Intelligent Systems with Applications*, p. 200 340, 2024.
- [10] A. R. Khalid, N. Owoh, O. Uthmani, M. Ashawa, J. Osamor, J. Adejoh, “Enhancing credit card fraud detection: An ensemble machine learning approach,” *Big Data and Cognitive Computing*, vol. 8, no. 1, p. 6, 2024.
- [11] P. Hajek, M. Z. Abedin, U. Sivarajah, “Fraud detection in mobile payment systems using an xgboost-based framework,” *Information Systems Frontiers*, vol. 25, no. 5, pp. 1985–2003, 2023.
- [12] N. Chen, A. Klushyn, A. Paraschos, D. Benbouzid, P. Van der Smagt, “Active learning based on data uncertainty and model sensitivity,” in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2018, pp. 1547–1554.

- [13] C. Schröder, A. Niekler, M. Potthast, “Revisiting uncertainty-based query strategies for active learning with transformers,” *arXiv preprint arXiv:2107.05687*, 2021.
- [14] J. T. Ash, C. Zhang, A. Krishnamurthy, J. Langford, A. Agarwal, “Deep batch active learning by diverse, uncertain gradient lower bounds,” *arXiv preprint arXiv:1906.03671*, 2019.
- [15] L. L. Cunha, M. A. Brito, D. F. Oliveira, A. P. Martins, “Active learning in the detection of anomalies in cryptocurrency transactions,” *Machine Learning and Knowledge Extraction*, vol. 5, no. 4, pp. 1717–1745, 2023.
- [16] F. Carcillo, Y.-A. Le Borgne, O. Caelen, G. Bontempi, “Streaming active learning strategies for real-life credit card fraud detection: Assessment and visualization,” *International Journal of Data Science and Analytics*, vol. 5, pp. 285–300, 2018.
- [17] X. Tang, Y. S. Astle, C. Freeman, “Deep anomaly detection with ensemble-based active learning,” in *2020 IEEE International Conference on Big Data (Big Data)*, IEEE, 2020, pp. 1663–1670.
- [18] R. Barata, M. Leite, R. Pacheco, M. O. Sampaio, J. T. Ascensão, P. Bizarro, “Active learning for imbalanced data under cold start,” in *Proceedings of the Second ACM International Conference on AI in Finance*, 2021, pp. 1–9.
- [19] D. Kemper, “Keeping up with fraud: An active learning approach for imbalanced non-stationary data streams,” 2017.
- [20] L. Priya. “Isolation forest.” (2024), [Online]. Available: <https://lekha-bhan88.medium.com/anomaly-detection-using-isolation-forest-and-local-outlier-factor-158985d6a45>.
- [21] Javatpoint. “Support vector machine.” (2024), [Online]. Available: <https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm>.
- [22] R. Pramoditha. “Neural network.” (2024), [Online]. Available: <https://medium.com/data-science-365/overview-of-a-neural-networks-learning-process-61690a502fa>.
- [23] T. Chen, “Xgboost: Extreme gradient boosting,” *R package version 0.4-2*, vol. 1, no. 4, 2015.
- [24] H. Deng, Y. Zhou, L. Wang, C. Zhang, “Ensemble learning for the early prediction of neonatal jaundice with genetic features,” *BMC Medical Informatics and Decision Making*, vol. 21, Dec. 2021. doi: 10.1186/s12911-021-01701-9.
- [25] A. Machine Learning Group ULB. “Credit card fraud detection.” (2024), [Online]. Available: <https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud>.
- [26] S. N. Kalid, K.-C. Khor, K.-H. Ng, G.-K. Tong, “Detecting frauds and payment defaults on credit card data inherited with imbalanced class distribution and overlapping class problems: A systematic review,” *IEEE Access*, 2024.

- [27] K. Shenoy. “Credit card transactions fraud detection dataset.” (2024), [Online]. Available: <https://www.kaggle.com/datasets/kartik2112/fraud-detection>.
- [28] A. Paul. “Vehicle loan default prediction.” (2024), [Online]. Available: <https://www.kaggle.com/datasets/avikpaul4u/vehicle-loan-default-prediction>.



PUBLICATIONS FROM THE THESIS

Conference Papers

1. M. Onur Erboy, A. C. Karaca, “Weighted XGBoost Based Active Learning Framework for Fraud Detection with using Small Number of Samples from Imbalanced Dataset”, in 3rd International Conference on Advanced Engineering, Technology, and Applications (ICAETA2024), Italy, 2024. https://doi.org/10.1007/978-3-031-70924-1_51

