When Simpler Models Outperform Hybrid Ensemble: Re-evaluating Hybrid Ensembles for Financial Fraud Detection

MSc Research Project
Data Analytics

Sanidhya Bajaj
Student ID: X23251867

School of Computing
National College of Ireland

Supervisor: Prof. Shubham Subhnil

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Sanidhya Bajaj

Student ID: X23251867

Programme: MSc in Data Analytics **Year:** 2024 – 2025

Module: Research Practicum

Supervisor: Prof. Shubham Subhnil

Submission Due Date: 11th Aug 2025

Project Title: When Simpler Models Outperform Hybrid Ensemble: Re-evaluating Hybrid Ensembles for Financial Fraud Detection

Word Count: 8550 **Page Count**.....19.....

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Sanidhya Bajaj

Date: 11/08/2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

When Simpler Models Outperform Hybrid Ensemble: Re- Evaluating Hybrid Ensembles for Financial Fraud Detection

Sanidhya Bajaj x23251867,
Research Practicum, MSc in Data Analytics
National College of Ireland Dublin, IRELAND
URL: www.ncirl.ie

Abstract

Fraud detection in the financial services sector is important as the consequences of a mere error can result in huge losses. This research runs counter to the popular belief according to which hybrid models, like voting and stacking are always better than individual machine learning models. We assessed this with the help of a comparative assessment of simple machine learning and hybrid techniques with distinctive ensemble. The research conducted used transactional dataset that contained 50,000 records of transaction labelled to indicate that the transaction occurred was legitimate or fraudulent.

Among the hybrid ensemble methods that were evaluated were voting and stacking classifiers compared to simpler but high-accuracy individual classifiers such as Random Forest, Extra Trees and AdaBoost etc. Measurements of assessment were centered towards F1-score, recall and AUC-ROC of classification regarding the fraud class in particular because of the class-imbalanced nature of the dataset. The results showed that even with respect to detection of instances of fraud among the minorities, the more simplistic models provided at least the same and, on most occasions, better results than the hybrid ensemble models. These findings lead to the suggestion that the default use of complex ensemble procedures might not always be the best solution in investigating fraud detection but rather it is utmost to select and optimize models carefully.

1 Introduction

Financial fraud has skyrocketed due to a sharp increase in the number of internet transactions, and more effective and superior fraud detection systems are required and since it is not that easy to keep up with newer techniques and strategies of financial fraudsters a large and everlasting importance is being accelerated on the machine learning (ML) techniques. The excellent ability of ML models to reveal the unknown patterns and anomalies in the transactional data contributes greatly to quality and efficiencies of detecting fraud.

Ensemble approaches have been rapidly growing in classification applications because they have the potential to nicely integrate the strengths of multiple learning algorithms to achieve better predicted performance. Special focus has been on stacking-based hybrid models which incorporate the use of classifiers such as XGBoost, LightGBM, and CatBoost to detect frauds [1]. These hybrid methods are meant to improve upon single models and specifically on more complex and imbalanced data sets due to identifying complementary patterns.

Especially in imbalanced datasets, such fraud-related measures as recall and F1-score help to elucidate the effectiveness of the model. Recent research shows that even simple models, such as Random Forests, Decision Trees, etc., can perform competitively when properly

adjusted and applied on pre-processed data. Indicatively, in identifying fraudulent credit card transactions, Random Forest and Decision Tree algorithms generated high F1-scores and accuracies above 99 percent [3]. As much as the results achieved in the end through their suggested hybrid ensemble model were slightly better, it was also progressive. This can also assist with the argument that other less complex models still can be effective and efficient alternatives, especially where speed, transparency and regulatory filing are of primary concern.

Despite the popularity of the ensemble models in the academic research in fraud detection, there are still limitations in this area that require a more thorough investigation of their application, especially in financial systems, such as interpretability and operational limitations of the real-world deployments [2]. The necessary trade-offs, i.e., the interpretability of the systems, the computational cost and the marginal benefits of the performance of the systems, should be considered when developing fraud-detection systems. Applications within the financial sector, specifically, require consideration: applications in monetarily and reputationally high stakes contexts must avoid false alarms or suffer the loss of both their own money and their reputation.

Contrary to the usual machine learning classifiers, the thesis will determine whether hybrid ensemble models offer superior fraud detecting capabilities. To measure the trade-offs in complexities of models, efficacy of detection, and applicability of models in real life situations, comparison analysis is done using a synthetic dataset in the paper.

This study is primarily meant to critically analyze the suitability and utility of the hybrid ensemble models in detection of frauds. The following are the objectives the study is guided by:

- To investigate the extent to which the hybrid ensemble is likely to give significant enhancements on the performance of fraud detection processes compared to other techniques which can be discovered to be less complex and common in character.
- To investigate how well practices and usability have been improved in real life application with the level of added complexity behind hybrid ensemble.
- To see how hybrid ensemble deal with the issue of class imbalance which is a problem you will often encounter in fraud detection data that is found in the real world.

2 Related Work

2.1 Ensemble Learning to Identify Fraud Overview

Fraud detection systems are becoming quite significant due to a high rise in sophisticated fraudulent transactions that have been experienced alongside the rapid expansion of the digital financial services. The main contributor to the use of machine learning (ML) techniques comes in the backwardness of traditional rule-based methods to cope up with the evolving fraudster strategies. These ensemble learning methods such as bagging, boosting and stacking have gained popularity since they are able to combine many base models and therefore achieve better prediction performance. These techniques come in handy especially in high stakes applications like in fraud detection where it is necessary to identify events that do not recur with great and high precision. Also, model combination can help in reducing bias and volatility.

2.2 Ensemble Interpretability Problems

Almalki and Masud [1] propose a stacking ensemble approach where the results of three algorithms namely, XGBoost, LightGBM, and CatBoost get aggregated to achieve high accuracy of detecting financial fraud. They use such explainable AI models, as SHAP and LIME, to make their model decisions transparent, considering the problem of interpretability of sophisticated models. This highlights a very important trade-off: the ensemble models can be a great predictor, but they require often additional mechanisms to ensure explainability, particularly in highly regulated fields such as the finance sector. On the other hand, Random Forests, Decision Tree or any other model is much more transparent by default and can therefore be implemented in cases where speed matters or prompt comprehension and compliance among regulators are of foremost concern.

2.3 Hybrid Ensemble Approaches

According to the Zhang et al. [2], the hybrid ensemble model is proposed to identify the fraudulent transactions in Bitcoin; that is, their algorithm can perform a certain task. Filtering the shortcomings of classical models is done by the aggregation of good learners into an ensemble and the tuning of hyperparameters by means of the grid search. The predictive capacity is enhanced especially those that are associated with mitigation of bias and variance. The increased complexity and the necessity to use the tool like SHAP to explain the judgments produced by the model, which is why those trade-offs are evident to the reader are made clear by the authors. The additional benefits in the hybrid ensemble in terms of accuracy and AUC are so small as compared to the higher computational costs and complexity of architecture, yet the gain is such that it exceeds people in usability. That reinforces the assumption that hybrid operations to be used in real fraud detection systems must provide a context related benefit.

Talukder et al. [3] offered a quite well optimized hybrid ensemble model via data balancing with the help of Instance Hardness Threshold (IHT) and hyperparameters tuning with the help of grid search through a use of several classifiers (Decision Tree, Random Forest, K-Nearest Neighbors, and Multilayer Perceptron) on credit card fraud detection. They just warn that individuals should not believe that the ideal points in their model (100 percent pinpoint and AUC) apply in real life activities. The authors directly address the dangers of overfitting and performance inflation, and the way they impact either a balanced or an imbalanced dataset within a controlled environment of a proper experiment. Within such a critical approach, one should remember about the necessity to assess the ensemble approaches in relation to the generalizability, interpretation, and applicability in conducting fraud environments and comparative performance ratings.

2.4 Ensembles Deep Learning

A review about deep learning ensembles by kora et al. [5] can be examined, in detail, concerning the usage of deep neural networks (CNNs, and LSTMs among others) in the context of fraud detection tasks using ensemble set-up. The feature learning power of these models is strong, however the review hints at low interpretability, slow performance of training, and demands many computations. The disadvantages in this case are especially evident with such a distinction of line of work as the fraud detection, in which the

transparency of the model is the essential factor, and the instant response is paramount. In addition, the review observes that even though the deep ensemble techniques are naturally more complex, they tend to work quite consistently better than the common machine learning ensemble techniques. This will help with the purpose of the current study to prove that the simpler models can provide comparable or even better results than the complex approaches without being as expensive.

2.5 Simpler Models vs. Complex Ensembles

Though sometimes complex ensemble models and deep learning have shown some marginal increase in recall and F1-score, simple models, including Random Forests, Decision Trees, and Logistic Regression, have often been nearly as good, Ali et al. [7] examined 93 machine learning strategies applied to the financial fraud detection. In addition, they can be more easily understood, require less complicated processing requirements and are also easier to apply in, especially, high-stakes sectors like financial services. The results imply that the trade-offs between transparency and resources costs are likely to be not as important as minimal advantages associated with complex models.

2.6 Supervised vs. Unsupervised Learning

Supervised methods and Unsupervised methods, in general, and methods based on recognizing anomalous transaction patterns without being labelled, specifically, are also currently explored with the aim of fraud detection [8]. These techniques may also be computationally effective and often succeed where rare fraud patterns are to be found. The fact remains that despite any correctness, they are often not transparent to deal with and subject to interpretability and temporal stability concerns, especially in dynamic real-world versions of fraud. Thus, despite the knowledge that can be conveyed through these techniques being massive, the limitations shine more light on the importance of adopting supervised and interpretable models.

2.7 Handling Oversampling and Class Imbalance

The effect of a class imbalance and an application of the oversampling strategies have become a subject of a thorough study in recent fraud-detection research works. A 2024 study had the authors specifically investigate the possibility of building a hybrid ensemble model by inserting gradient boosting models, XGBoost and LightGBM, into the Synthetic Minority Oversampling Technique (SMOTE). Using them on transaction data, they realized that the pairing with SMOTE could lead to drastic improvements of key performance measures, including F1- score, recall and precision, by about 56% compared to more conventional classifiers, e.g., Random Forests and Logistic Regression [11]. It was also found that although the oversampling strategies such as SMOTE could effectively increase recall and overall accuracy, the model could suffer instabilities, induced by the wrong tuning of SMOTE parameters, which are reflected by the increased variation of performance results [11]. Also, during the last year another study made the same observation, which is that newer and more sophisticated models, where an ensemble approach with boosted algorithms, as in Gradient Boosting Machines (GBM) and the improved Random Forest (RF) are resoundingly superior over generally simpler models. Nonetheless, the study stresses the feasibility of the trade-offs associated with these complex models and especially in terms of interpretability and cost of

reliance [6].

2.8 Voting Ensemble Methods

The soft voting classifier is one of the recent voting-based ensemble methods that have received a lot of attention in studies. Compared to hard voting methods, soft voting classifications generally surpass class-imbalanced scenarios [10]. When techniques that aggregate predictions via probabilities such as soft voting methods were used, scientists discovered that they resulted in superior aggregate predictions across all classes and allowed decisions to be made in a fairer way. Aware of the fact, however, that the performance of the soft voting ensembles is highly dependent on the quality of individual base classifiers and their variability, the study aims at doing exactly that. Enlarging the number of classifiers by itself does not guarantee an improvement in the ensemble performance and can lead to an enforcement of the ensemble, more succinctly, the accuracy of the aggregate ensemble prediction is less than accurate as that of the base model that has best performance. Thus, the calibration and reserved selection of component classifiers in ensemble-based fraud detection tasks should also be enforced.

The challenges brought by the class imbalance and overlapping data distributions have been captured in the past works on credit card fraud detection. These issues are hard to programmable in conventional machine learning programs. It was seen that ensemble methods can be used to increase the accuracy of detection especially with boosting and bagging. Research states that feature selection process and data pretreatments play an important role in improving the performance of models. The literature research suggests that the hybrid models are getting increasingly popular because such models combine numerous algorithms to better capture fraudulent trends in unbalanced datasets [4].

2.9 Hybrid Ensemble Model Study

Saraf and Phakatkar [9] present that a hybrid ensemble model has been proposed which operates a combination of both techniques namely AdaBoost and Random Forest, also known as bagging, to enhance the detection of credit card fraud. They deal with overclassification problems using oversampling, and their method shows superior AUC scores (they solve the problem of class imbalance since the AUC scores go to 98.27 on the European credit card data and 99.3 on the simulated one). The success of blending both the strong and the weak classifiers (use of K-nearest neighbors and use of Decision Trees) is also brought into focus in the publication of hybrid ensemble. Their results however indicate that its combinations are in several instances below various types of hybrid model which are built on stronger learners, boosting procedures and the RF. This also adds to the hypothesis that model selection and building of ensembles despite to a strong extent influences predictive achievement on fraud honing cases.

2.10 Comparison of Public Datasets

Sundaravadivel et al. employed Random Forest with SMOTE in a comprehensive study of credit card fraud of public data. They discovered that the Random Forest model worked exceedingly well and demonstrated good fraud detection ability in terms of accuracy of detecting fraudulent transactions up to 99.5 percent and a high read rate [12]. Conversely, Niu, Wang, and Yang resorted to the same Kaggle dataset comprising 492 of 284,807 transactions

involving fraud cases to compare a supervised and unsupervised approach, including XGBoost, Random Forest, One-Class SVM, Auto-Encoders, RBMs, and GANs. Their result was that the AUROC values of unsupervised methods (such as RBM and GAN) were 0.961 and 0.954, respectively, whereas the supervised methods (such as XGBoost and Random Forest) performed better with an AUROC value of 0.989 and 0.988, respectively [15], which led them to the conclusion that finely tuned classical classifiers are successful in the field.

2.11 Trade-offs between performance and complexity

Recently in a set of papers on the performance of strong classifier hybrid models, small, but statistically significant, improvements in AUC have been found where Random Forest is joined with Gradient Boosting [6]. However, such advantages come together with the extra computational complexity and longer training needs. The authors therefore conclude that the marginal benefits linked with faster and more accurate predictions in cases where this is necessary -e.g. fraud detection may not justify the cost incurred by complexity of the model. Taken together, their findings support the overall rule to be considered in literature so far: hybrid ensembles must provide provable and problem-specific benefits, prior to justification.

In the work by Ibanga, the introduction of oversampling methods such as SMOTE and ADASYN to enhance model sensitivity to detect fraud was explored [13]. They addressed class imbalance, which is often a problem in the datasets of fraud detection, by generating synthetic examples of the minority class. Bagging, boosting and stacking ensemble methods were also used in the search of resilient prediction. The findings indicated that in most of the datasets, a panel like Random Forest and SMOTE or Bagging Classifier and ADASYN were more accurate in most cases compared to others. The above signifies the utility of the ensemble models combined with oversampling approach to detect fraud more precisely in practice.

Moreover, according to the new studies, the ensemble solutions and personalized oversampling mechanisms significantly improve fraud detection when used in tandem. Ranjan et al. [16] established a stacking ensemble with respect to Random Forest and hybrid resampling to achieve decent classification levels of the European dataset. Mim et al. [17] have reached an AUC of 0.9936 applying a soft-voting ensemble on top of XGBoost, KNN, MLP with SMOTE and ADASYN. The efficiency of clustering-based oversampling in the context of classical ensembles was indicated by Wang [18], through the introduction of SMOTE-KMEANS, combined with RF and SVM that yielded an AUC of around 0.96.

3 Dataset Description

The study has been conducted based on the Fraud Detection Transactions Dataset, which is available [14] on the Kaggle platform. There is no doubt that the dataset is a simulated type of data, but it is programed to be as near to reality in real life financial transactions of the banking. It records actual fraud indicators, in the combination of the device features, risk scores and behavioral patterns.
Characteristics of Dataset:

- 50,000 unique transactions
- Label: Fraud_Label (1 = Fraud ,0 = Legitimate)

- Characteristics: 21 explanatory features
- imbalanced: The minority class Fraud is present.

Our dataset imitates the behavior of the users throughout the time frame as it contains both stationary (Transaction_Type, Device_Type) and temporal (Daily_Transaction_Count, Risk_Score) characteristics.

Category	**Features**
Transaction-level	Transaction_Amount, Transaction_Type, Timestamp, Location, Merchant_Category
User behavior	Daily_Transaction_Count, Avg_Transaction_Amount_7d, Failed_Transaction_Count_7d
Account info	Account_Balance, Card_Type, Card_Age, Authentication_Method
System attributes	Device_Type, IP_Address_Flag, Is_Weekend, Risk_Score

4 Exploratory Data Analysis

Exploratory data analysis played a key role while understanding the dataset and blocking a major data leak that would have resulted in misleading accuracy of the model, had it not been checked.

4.1 Data leakage in Failed_Transaction_Count_7d

Early model training activities resulted in measures of performance that were far too high. In a further analysis, we wanted to plot the fraud rate versus Failed_Transaction_Count_7d to plot an artificial spike because it is not practical:

- The rate of fraud remained approximately 15%, up to 3.
- At the level of value = 4 fraud rate rose to 100%.

It was obvious that the feature encoding fraudulent activity directly could be either data building artifacts or label annotate leakage. This aspect was removed prior to the training process to maintain the generalization ability of the model by preventing overfitting.

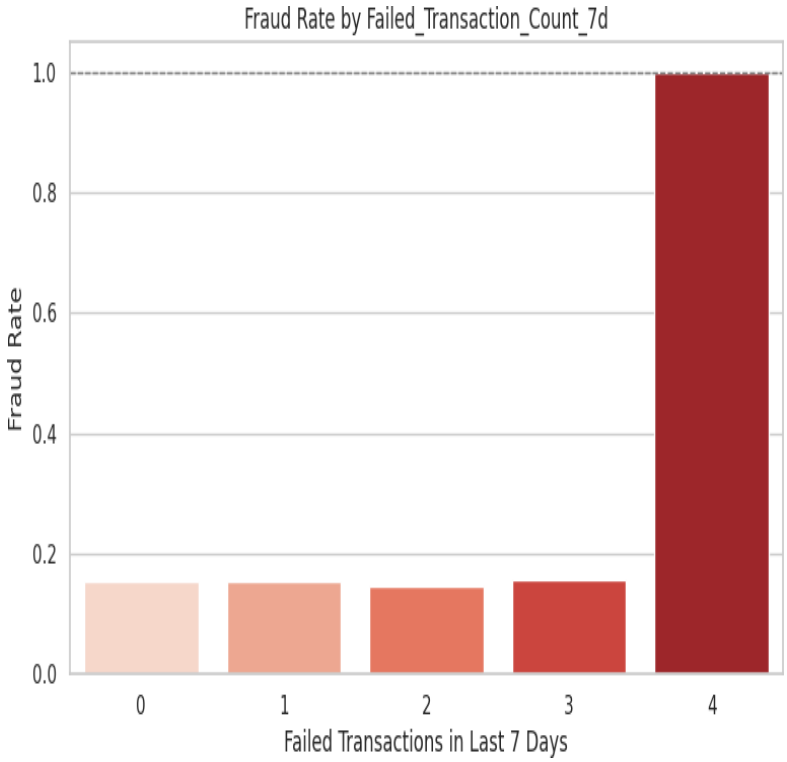

Figure 1: Fraud Rate vs. Failed_Transaction_Count_7d

4.2 Class Imbalance Overview

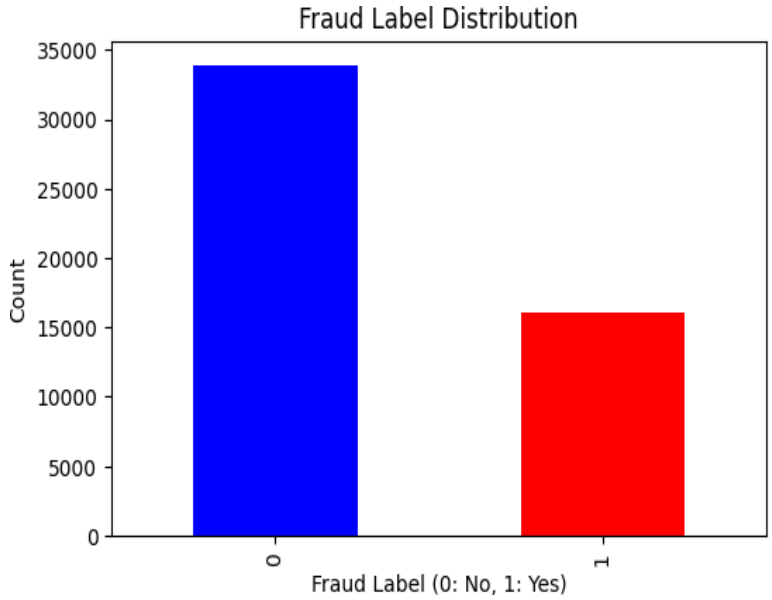

Figure 2: Fraud vs Legitimate Transactions count

Observing the fraud label distribution plot, it is evident that a class imbalance is present in the dataset with transactions labelled as 0 (non-fraudulent) being considerably more than transactions labelled as 1 (fraudulent). This imbalance indicates the low incidences of fraud cases as it is a reality when it comes to fraud detection in the real world. As such, in such skewed distributions, performance metrics such as recall, F1-score, and AUC- ROC are more suitable to be used to assess and are more instructive than plain accuracy which may be misleading.

4.3 Fraudulent Activity Patterns throughout the Years

To be able to monitor the temporal changes and fluctuations in frauds, a line plot of the number of fraudulent transactions per day within the 2023 calendar year was developed. As depicted by the graphic, there are consistent fluctuations over the entire calendar year, which means that fraud is not evidently seasonal but is scattered unevenly over time.

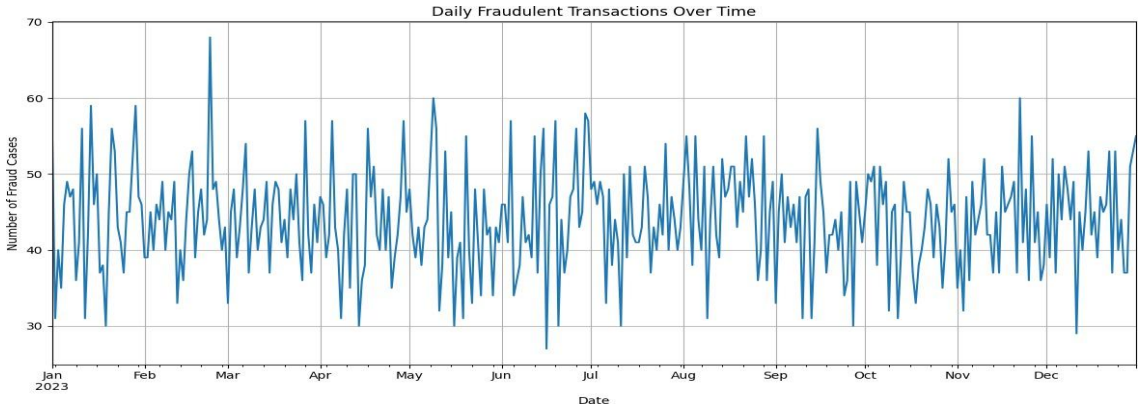

Figure 3: Fraudulent Activity Patterns throughout the Years

Key findings:

- The frequency of fraudulent transactions remains steadfast all through the year with the daily figures normally turning out to be 35 to 55.
- It is observed that the data fluctuates a lot, it appears that a batch or group of fraud occurrences happened on days, although there seems no apparent seasonality in the data.
- The frequent escalations of over 60 fraud cases are indicators of days that are outliers, and this can be due to organized attacks or bugs on the system or active ended-month transaction periods.
- It shows no recognizable upsurge or downward trend over the period of the year, although the fluctuations exist, thus indicating that fraudulent traffic is still actively present, and relatively consistent in the terms of its volume.

Based on this time information, we can predict when the levels of fraud may be higher or when concept drift may happen on a time basis. Two of the time series modes of models that can be used to improve the system to detect abnormal patterns used in fraudulent practice and dynamically change with new patterns of transactions are the recurrent neural networks (RNNs) and the time-based anomaly detection.

4.4 Transaction Amount Distribution by Fraud_Label

To understand better the role of this marker in the classification of fraud, a boxen plot of the distribution of the Transaction_Amount variable between frauds (Fraud_Label=1) and non-fraudulent (Fraud_Label=0) transaction was developed.

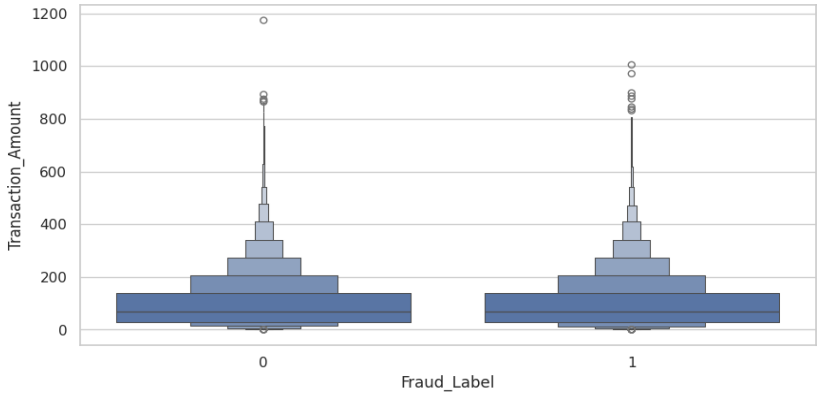

Fig 4: Transaction_AmountDistribution by Fraud_Label

Key findings:

- Fraudulent and non-fraudulent transactions in both cases have positive effects with transactions being clustered in the low values of less than 200 units.
- The general distribution forms and interquartile ranges (IQRs) are also similar in both classes and thus Transaction_Amount may not be a good predictor of fraud in both the classes.

Although the number of transactions alone is rather low, this description confirms the notion that it remains an assisting variable, which does make a difference when being analyzed alongside the other situational or behavioral attributes.

4.5 Risk Score Density Plot by Fraud Label

A density plot of the score distribution between fraudulent and valid transactions to detect the predictive relation between the Risk_Score and the fraudulent activity was developed.

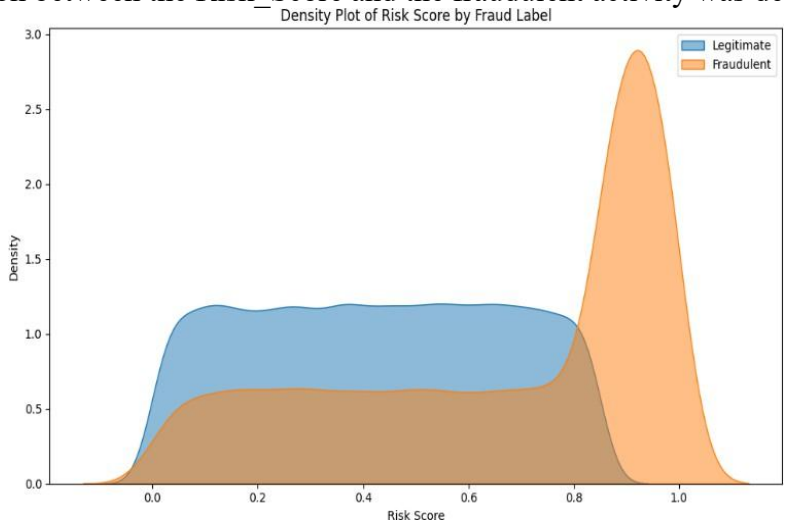

Figure 4 - Risk Score Density Plot by Fraud Label

The two classes can be easily contrasted with each other due to this

visualization:

- High concentrations of risk ratings near the value of 1.0 are experienced with fraudulent transactions which imply that the attribute is highly indicative of fraudulent conduct.
- Legitimate transactions, however, spread out over lower the risk scores and a prominent peak is at the 0.2-0.6 range.

The distance between the two distributions is evidence that Risk_Score is an efficient indicator of classifying fraud. The fact that it has been put in the model contributes significantly to the distinction of high-risk trades.

Such a level of separability, however, also raises the problem of feature dependency: in case the actual Risk_Score is a composite or derived score, as it should be in the real-life systems, its presence at the time of training may cause either bias or over-fitting.

Consequently, the comparison of model performance with and without these criteria might

also benefit future research to recreate the real deployment scenario, although it is retained to obtain the baseline comparison.

5 Research Methodology

The methodology employed in the thesis was set on a comparative experimental design that was aimed at explaining the overall effectiveness of the hybrid ensemble methods and the basic models in detecting financial fraud. In this section, procedures and explanations of choosing the model, assessment criteria, hybridization strategies, and preprocessing of the dataset are outlined.

5.1 Data Exploration

The study will be based on a dataset with 50, 000 transactions and attributes of 21 each. It is synthetic information; it is published to the general population, and it is supposed to emulate the behaviors of both actual and artificial financial transactions. Each transaction is classified as either a 0 or 1 in which 0 represents a non- fraudulent example whereas 1 indicates a fraudulent one. Prior to developing a model, the exploratory data analysis (EDA) has been performed to have an overview of the distribution of the classes, the type of features and the condition of the missing values. This initial analysis was a necessity because it the contribution it made towards assuring the quality of the data that would be obtained and provided some guidance as far as preprocessing activities were concerned.

5.2 Data Preprocessing

- **Missing Values Processing**: The columns that were no longer necessary and which had to be dropped were Transaction_ID, User ID, Timestamp and Failed_Transaction Count 7d. No imputation or backfilling was required as the understanding of the data was that there are no missing responses in the dataset.
- **Categorical Encoding:** Categorical variables were converted to prevent occurrence of the problem of multicollinearity by undergoing one-hot encoding (drop='first'). Machine learning models can use such encoded features with the help of this format.
- **Standardization:** The numerical characteristics were standardized (z-score normalization), and then the data was centered in 0 and standard deviation 1. This will make sure that there is an equitable participation of any feature during training.
- **Class Imbalance**: In this work, methods such as SMOTE for resampling were not used. Even though such methods can assist in the class distribution balancing, such methods can lead to overfitting or misevaluation of the performance rates because they include artificial tendencies found not in real fraud cases. The original distribution was not altered since the aim was to determine how effectively models could handle the inherently unbalanced fraud detection, which takes place in a real-life bank and financial organization. This enabled the further assessment of the models in sensitivity to be more veracious, namely, examining the extent to which they were able to identify the minority (fraud) cases without puffing the recall or the precision rate.
- **Train-Test Splitting:** Stratified sampling was used to respect the original proportion of fraud and non- fraud among the testing and the training data to segment the data into the testing and the training data.

5.3 Baseline Machine Learning Models

Some of the traditional machine learning models were trained and tested independently to have a benchmark to compare and make hybrids.

- Decision Tree
- Random Forest
- Gradient Boosting
- AdaBoost
- Support Vector Machine
- K-Nearest Neighbours (KNN)
- Naive Bayes
- Extra Trees
- XGBoost
- Bagging Classifier
- CatBoost
- LightGBM

The models were run using the default settings or without modifying the parameters to the pre-processed data to obtain the base results. Various combinations of learning strategies have been experimented on with the trials e.g. distance-based classifiers, probabilistic models, boosting algorithms, and tree-based and so on. These baseline tests were employed as the standard comparison on basis of which subsequent comparison is to be carried out with the ensemble and hybrid methodologies.

5.4 System Architecture Design

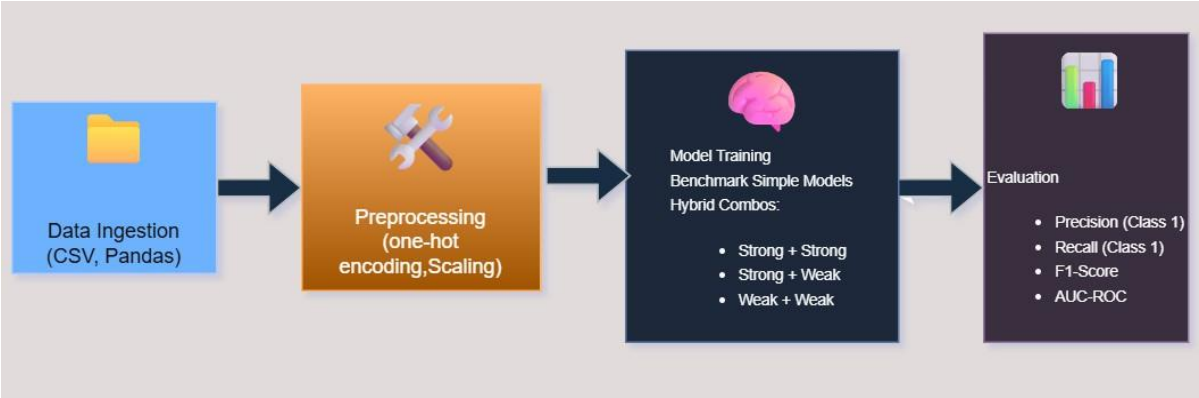

Figure 5: End-to-End Architecture Design

5.5 Evaluation Metrics

a) Accuracy

- **Definition**: The percentage of all the correct predictions out of the total ones made.
- **Formula**: Accuracy = (TP + TN) / (TP + TN + FP + FN)
- **Interpretation in your project**: The accuracy of the model implies how often the model was used to determine whether a transaction was fraudulent or not. It can be misleading to apply it in imbalanced dataset, such as the legal transactions in dataset but a much lower instance of fraudulent transactions.

b) Precision (Class 1)

- **Definition**: What was the number of transactions that were predicted to be fraudulent (Class 1), but turned out to be fraudulent?
- **Formula**: Precision = TP / (TP + FP)
- **In your wildfire case**: High precision denotes a good model as it is usually accurate in classifying a transaction as fraudulent. It is essential because there is a risk of false positives that may end up in unnecessary banning of transactions, restrictions on accounts, or dissatisfied customers.

c) *Recall (Class 1)*

- **Definition**: What was the number of correct identifications in the number of all the real Class 1 instances found by the model?
- **Formula**:

  Recall = TP / (TP + FN)
- **In your context**: Although it may incorrectly find some legitimate transactions, it is very good that the high recall shows that the model can identify fraud effectively. This is necessary in minimizing false negatives that may cause loss of money.

d) *F1-Score (Class 1)*

- **Definition**: Harmonic means of Precision and Recall. Balances both metrics.
- **Formula**:

  F1-Score = 2 * (Precision * Recall) / (Precision + Recall)
- **Use in project**: F1-Score is very useful in the fraction of the imbalance in the classes to identify fraud. The larger the F1-score the better because it means that there are not many false positives or false negatives and that the model is catching fraudulent transactions.

1. *AUC-ROC (Area Under Curve - Receiver Operating Characteristic)*

- **Definition**: measures the model capability to discriminate among classes over all levels.
- **Scale**:
  - 0.5 = Random guessing
  - 1.0 = Perfect separation
- **Use in project**: A large AUC means that the model is characteristically able to draw the line between genuine and false transactions which is crucial when it comes to changing the threshold and real-time fraud detection pipeline.

5.6 Hybrid Ensemble Design

To find out whether hybrid models indeed tend to perform better than the techniques of which they are a combination, an ensemble of classifiers was constructed by combining the existing techniques of classification. The combinations are:

- **Strong + Strong:** Random Forest + Extra Trees
- **Weak + Weak:** Decision Tree + KNN
- **Weak + Strong**: Decision Tree + Random Forest

5.7 Ranking and Comparison

A model comparison table (section 6.5) was constructed to order the models based on the criteria of interest in fraud-detection (F1-Score of the fraud class). Hybrid models were compared to determine did hybridization led to the significant advances

6 Results and Analysis

At the experimental stage, both hybrid ensemble models and simple machine learning classifiers were tested on the dataset and trained to determine their performance by evaluating the accuracy of their predictions. Performance evaluation concentrated more on the indicators sensitive to skewed distribution because only a few fraudulent transactions are fractions of the records, because of the overall class imbalance. Specifically, it was the use of AUC-ROC (Area Under the ROC Curve) Precision, Recall, and F1-Score (reported per class). The measures concerning the minority (fraudulent) class were paid more attention since more accurate fraud detention is crucial to performance of the system in practice.

6.1 Performance of Individual Classifiers

Though results of all classifiers were always high (with the accuracy value over 0.75 in many cases), initial results of individual models indicated that, due to the class imbalance, that score was not by itself informative. Instead, more insights were provided by metrics such as F1-Score and Recall of the fraud class.

- Random Forest, Gradient Boosting, AdaBoost and Extra Trees showed very high precision (near or equal to 1.0) of the fraud class. Notwithstanding their rather small F1-scores in terms of fraud (F1_1) because they had very low recall, this model outperformed other models in this indicator, which points at their much-improved ability (but clearly not perfect) to identify fraudulent transactions.
- There was a worse f1 and recall of detecting fraud when using Decision Tree and K-Nearest Neighbors (KNN), which were faster and less challenging to use.
- Naive Bayes and Support Vector Machines exhibited competitive scores, especially when it comes to Precision and AUC though they were not consistent over the recall.
- Having high precision and AUC scores that reflect proper modelling, calibration and generalization, CatBoost and LightGBM worked well overall, as their balance where fraud detection is concerned can be deemed good. It could also be improved in terms of recognizing the minority class with them portraying low memory in the identification of fraud cases like most of the other ensemble models.

6.2 Performance of Hybrid Ensembles Interpretation

The hybrid models were built by stacking weak and strong learners by means of voting (hard/soft) techniques. The ideal was to find out whether combining classifiers would lead to overfitting, redundancy, or unnecessary complexity, or increase the performance in detection.

Strong + Strong Combination

- Random Forest + Extra Trees (Stacking) edged out all other individual models by a small margin overall, achieving a large AUC-ROC (0.744) and overall score of 0.644

on the fraud F1-Score.

- The Voting Classifier (Hard) implementation of RF and Extra Trees achieved a lower value of fraud recall, which suggests that stacking could be an effective approach of capturing complexity compared to simple averaging.

Strong + Weak Combinations

- On the second line were Random Forest + Decision Tree (Stacking), which almost equal RF alone, with high AUC-ROC (0.745), as well as an impressive test-level fraud F1-Score of 0.644.
- What this means is that hybridizing a weaker learner with a stronger learner need not necessarily lead to a noticeable gain compared with the stronger learner alone.

Weak + Weak Combinations

- Even though the Decision Tree + KNN combination worked better than each of the base models individually, it was still lot worse than the best.
- As for example, stacking (DT+KNN) did better than KNN and DT, but still underperformed compared to RF and GBM hybrids, reaching an F1-Score of 0.619 for the indicator fraud.

6.3 Analysis of Hybrid Performance

- Generally, the measure of F1 score and area under the ROC curve were higher with stacking as compared to voting.
- The uncertainty under probability averaging is better captured compared to hard majority voting which is reflected in the fact that soft voting was usually successful compared to hard voting.
- Curiously, hybrids, i.e., combinations of two already strong models, e.g. RF + ET did not perform a lot better when compared to their individual performance. This supports the findings that we should use hybrids sparingly, and not every combination yields to meaningful gains.

6.4 Summary of Findings

- **Best Model Overall**: The best of all models and the one with the maximum AUC-ROC is stacking (RF
  + ET) and very closely tied with the maximum F1_1 scores. Both Random Forest and Extra Trees achieved the same performance, especially with F1_1, so they are all superior at the detection of fraud.
- **Best Individual Model**: Random Forest (individually matched most hybrid performances)
- **Best Hybrid**: Stacking (RF + ET) and Stacking (DT + RF): The two stacking strategies that performed best among all the hybrid models that were evaluated were Stacking (RF + ET) and Stacking (DT + RF). The latter showed the highest overall power of distinguishing between fraud and non-fraud in general with an AUC-ROC of 0.745. Nevertheless, regarding the identification of fraud, the stack Random Forest + Extra Trees provided the highest F1-score in the case of fraud class (F1_1 = 0.6442), meaning that it has a slightly better balance between recall and precision.

Although the two models perform almost identically on most of the measures, the relative strengths of two models differ: RF + Extra Trees stacking was the best in terms of balanced fraud detection and DT + RF stacking was the best in terms of overall ranking performance. We have found the results that well designed hybrid ensembles may provide some performance boosts even when compared to even the best individual stand-alone classifiers.

- **Worst Individual Model**: KNN and Decision Tree
- **Worst Hybrid Model**: Hard Voting (DT + KNN) despite improvement from its individual classifiers, it underperformed all top models

6.5 Model Comparison Table (Sorted by F1-Score for Fraud)

Rank	Model Name	F1_1	Accuracy	AUC-ROC
1	Extra Trees	0.644544	0.831467	0.7431
2	Random Forest	0.644544	0.831467	0.741
3	Bagging Classifier	0.644544	0.831467	0.7432
4	AdaBoost	0.644544	0.831467	0.7378
5	StackingClassifier (RF + ET)	0.644211	0.831	0.7445
6	Gradient Boosting	0.644182	0.8312	0.739
7	Stacking (RF + DT)	0.64402	0.830933	0.7452
8	XGBoost	0.644001	0.831067	0.7401
9	Voting Classifier (Soft RF + ET)	0.643997	0.8304	0.7413
10	CatBoost	0.64383	0.830867	0.7404
11	LightGBM	0.643569	0.8306	0.7407
12	Stacking Classifier (DT + KNN)	0.619487	0.7964	0.7435
13	Naive Bayes	0.618052	0.792933	0.7389
14	SVM	0.627301	0.811067	0.7386
15	Voting Classifier (Soft DT + KNN)	0.594074	0.749733	0.7282

16	Voting Classifier (Soft DT + RF)	0.58599	0.7238	0.7502
17	Decision Tree	0.580918	0.7224	0.6898
18	Voting Classifier (Hard RF + ET)	0.637648	0.829067	—
19	Stacking (DT + KNN)	0.619487	0.7964	0.7435
20	Voting Classifier (Hard DT + KNN)	0.470659	0.763667	—
21	K-Nearest Neighbour	0.454072	0.713533	0.6656

7 Key Findings

- **Simpler models work as effectively as or even better than hybrids**: Random Forest and Extra Trees were trees-based models that produced one of the highest F1-Scores and AUC-ROC values individually. These models would go on to demonstrate that simpler classifiers could yield results that are adequate and in certain cases better than fancier ensemble structures. This confirms the outcomes of other studies that there is no confirmation of complexity in the actions of detecting fraud leading to better performance [7].

- **Hybrid models must be well designed to work:** Hybrid models need to be well constructed to yield substantial returns. As the findings show, not every hybrid can be better than its separate parts despite not much improvement being gained by combining models with different applications, such as Decision Tree and Random Forest, and similar models, such as Random Forest and Extra Trees. This is more so in case they lack sufficient complementary diversity in the models.

- **Due to poor recall rate of fraud, some hybrid models failed significantly**: The recall of the fraud class (class 1) was quite low in most of the hybrid and ensemble models, even though the overall performance indicators appeared to be high, including the accuracy and F1-scores lesser in magnitude. As an example, even the most successful stacking models on RF and ET combinations had less than 0.48 recall. This would create the illusion of efficacy since a lot of requests for fraud would remain unchecked under real- world application.

- **In unbalanced data sets the F1-Score is more important than accuracy:** As the valid transactions count was much more the accuracy remained illusorily high in nearly all the models. Conversely, F1- Score of the class of fraud (class 1) was accurately showing the ability of the model to predict the occurrence of minority classes. This statistic played an important role when determining the feasibility of a model in terms of detecting fraud.

- **Stacking outdoes soft and hard voting**: This finding has been true in the F1 as well as AUC-ROC performance of stacking models, which beat the voting-based ensembles in almost all the types of hybrids. This means that the base models have more benefits that can be exploited by meta-learners as compared to static models of voting because they do not learn by watching the activity of the base model.

- **Hard voting was often unnecessarily challenging and lacked finesse:** In several of these setups the F1-scores and recall were worse with hard instead of soft-voting. Hard voting is less flexible whereby models differ as they do not provide chance of confidence. As a result, the base model contributions were missing, containing a few encouraging contributions.

- **Some Strong individual models outperformed by weak + weak hybrids (DT+KNN):** Although the performances of two weak learners outperformed each of them individually, they were still not capable of beating the well-optimized single classifiers. This once again shows that hybridization cannot always lead to better performance unless the balance of complementary strengths is given due consideration.
- The financial implication is that missing fraudulent transactions (false negative result) is much costlier compared to detecting actual ones, and therefore, recall is the critical parameter when detecting frauds. Models that are highly precise or have low recall come with serious implications for operations. High recall can ensure that cases of fraud are detected even if it implies going across a couple of false positives.

- **Hybrid models have a major drawback potential over interpretability:** Despite the high performance, ensemble and hybrid models often fail in being explainable. It has become a problem in regulated industries such as the financial sector where companies will be required to give a reason as to why a transaction has been flagged. Researchers ought to consider adding XAI tools, such as SHAP, in the future.

- **The difference in performance of the best models was insignificant:** whereas some hybrids received higher ranking in AUC-ROC or F1-Score than others, the actual improvements were often in the range of 0.5-1% gains. These minor increases may not be justified by the increased deployment cost or training difficulty on real world uses.

- **Redundant hybrids may exist** where two base learners are the same and they give predictions which overlap each other, then there is no extra use by adding the ensemble which is waste of computing resources and unnecessary complexity to the pipeline besides not being efficient in such cases.

- **Training and inference time matter in the practical deployment:** The Hybrid stacks take relatively longer in the training and inference process as compared to Random Forest and other types of easier models. It is possible for real-time fraud detection scenarios (though at the expense of the slight accuracy trade-offs) to adapt to low-latency and low-complexity models.

8 Discussion

The results of the research show the importance of context-based choice of the model to find financial crime. Due to the prohibitive price of false negatives and the value of openness that has been driven by the regulations, interpretability takes a higher precedence over accuracy or memory in fraud detection. It is crucial given that, when institutions do not detect fraudulent transactions, they can lose a lot of money and reputation.

Less complex models, such as Random Forest, Extra Trees, and Naive Bayes, consistently made good results in this environment, in terms of recall and F1-score of the minority (fraud) class. Their clarity and the easiness of use also endorse their effectiveness in case of high-stake situations. The models help to gain clear knowledge of the decision-making process, especially in regulatory situations where explainability is an absolute requirement.

On the other hand, hybrid ensemble models increased the complexity of models with regards to computers and building. Even though there were cases where certain of the hybrid models showed minor increases in accuracy or AUC-ROC, these advantages often were at the cost of a more involved deployment and reduced interpretability. Moreover, hybrids were often of no significant additional value to simply replicate the best base model performance.

The main conclusion associated with this study is the fact that not every ensemble strategy is beneficial. Namely, ensemble of models with similar decision boundaries, such as Random Forest as well as Extra Trees, did not present substantially superior results when compared to a single model. On the negative side, however, weak and strong learner combinations resulted in modest, non-systematic improvement, in support of the fact that model diversity is central to ensemble performance.

Finally, this paper stresses that the focus on careful, precise design and selection should not be compromised on the idea of greater model complexity. In a fraud detection situation, especially in real life, simplicity is often encountered with reliability and the understanding of the fractions therefore, practitioners must evaluate critically whether benefits of performance improvement procured by using complex models are greater when compared with computational and operations costs involved.

9 Conclusion

This research aimed at evaluating the efficiency of hybrid ensemble models in carrying out the task of financial fraud detection against simpler machine learning paradigms. The research showed certain interesting reflections

concerning the behavior, performance and practical usefulness of models in an experiment-based approach that entailed training, testing, and comparing a few classifiers and hybrid combinations over a synthetic fraud data set.

Conclusively, it can be said that, in comparison with the more complex hybrid ensemble strategies, which sometimes revealed minor improvements, the simple tree models such as Random Forest and Extra Trees showed similar performance in all instances. These base models performed well when it came to unbalanced datasets however, they had a significant performance with high AUC-ROC and F1-scores of the fraud class, which is one of the most significant criteria in fraud detection. They also offered operational advantages such as lower computing cost, faster training, and greater interpretability, which also made them extraordinarily beneficial alternatives in real world, high-stake financial applications.

Conversely, hybrid ensembles, especially the ones that are created using the stacking or voting of strong learner improved performance by mostly 0.2 to 0.5 percent based on measures such as AUC-ROC. These hybrids only tended to emulate the performance of their most efficient base model. Moreover, they also introduced an extra ontology to the explainability, training process and model architecture. This begs the question of whether they can be used in production environments, more so in such areas as banking where decision making must be transparent and traceable thereon.

Another pertinent revelation is the importance of aligning model evaluation with corporate goals. Even though measures such as accuracy are universally reported, they can easily mislead in terms of finding a fraud due to the class imbalance. They should focus, however, rather on the metrics that are better indicators of the ability of a particular model to detect the scarce variety of fraudulent events (e.g., recall, precision, F1-Score, and AUC- ROC). Also, the cost of false negatives (missed fraud) is significantly higher than false positives, thus the recall is more important than accuracy.

This analysis ends with the recommendation that practitioners should not make an arbitrary resort to complex hybrid models. Results justify an ethical approach to the model selection that bears on business meaningfulness, interpretability, memory, and practice feasibility.

10 Future Work

Although this thesis has presented valuable details on the performance of hybrid ensemble models and basic classifiers in terms of their ability to detect financial fraud, a lot of gaps will have to be filled with additional studies. The next part indicates the key means towards enhancing the breadth, interpretability, and relevance of findings:

10.1 Model optimization and Threshold Tunning

The models under this study determine the existence of transactions as either legal or as fraudulent based on default probability criteria that tend to be 0.5. Nonetheless, deterministic rules may not yield an optimal performance in detecting fraud where the fraud vs non-fraud ratio is severely unbalanced and false negative has a high cost. In future studies, threshold tuning ought to be implemented depending on the cost-matrices of each firm following the establishment of probabilities of the models using such techniques as Platt Scaling or Isotonic Regression. This would allow practitioners to find a better trade-off between precision (false alarms) and memory (fraud detection) depending upon their willingness to take risk of making false alarms.

10.2 Interpretability by applying Explainable AI (XAI) Approaches

Hybrid and ensemble models are often characterized with Black-box algorithms, especially those that are based on stacking and boosting. Such XAI tools as Partial Dependence Plots (PDPs), LIME (Local Interpretable Model- Agnostic Explanations), and SHAP (SHapley Additive exPlanations) will help to deconstruct their inferences. Interpretability (as well as user trust) can be enhanced not only with regulated monetary scenarios where decisions about fraud detection must be explained, as has been shown by Almalki & Masud [1]).

10.3 Evaluation using Real Data

Artificial data that is applied to this analysis and is beneficial to benchmarking may not be very suitable in portraying the noise, drift and complexity of real financial data. It could be a future direction to test the models on real-world data (e.g. anonymized transactions data coming out of banks or open financial APIs). This would facilitate the capacity to check the robustness of the model, identify the errors that arise in edge cases and understand how the real behavior of the users affects the outcome of the detection.

10.4 Hands on Temporal Validation and Concept Drift

Fraud tendencies are also changing as long as the behavior of attackers and the conditions of the market. In future research, the retraining of sliding windows and the use of temporal cross validation should be applied to result in a similar behavior of the models when they are dealing with new data. Periodic retraining of the online learning or batch models will better model the performance in the real world, especially with systems that need retraining to best

handle emergent fraud methods. As time goes by, one may also track the metrics of concept drift like KL divergence or changes in data distribution and commit them to model resilience.

REFERENCES

[1] Almalki, F. and Masud, M., 2025. *Financial fraud detection using explainable AI and stacking ensemble methods*. arXiv. Available at: https://arxiv.org/abs/2505.10050.

[2] Zhang, L., 2024. *A hybrid ensemble model to detect Bitcoin fraudulent transactions*. *Expert Systems with Applications*. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0952197624019699.

[3] Talukder, A., et al., 2024. *A hybrid dependable ensemble ML model using IHT-LR*. *Cybersecurity*, 7(1). Available at: https://cybersecurity.springeropen.com/articles/10.1186/s42400-024-00221-z

[4] Islam, M.A., Uddin, M.A., Aryal, S. & Stea, G., 2023. *An ensemble learning approach for anomaly detection in credit card data with imbalanced and overlapped classes. Journal of Information Security and Applications*, 73, 103538. Available at: https://www.sciencedirect.com/science/article/pii/S2214212623002028.

[5] *Mohammed, A. & Kora, R., 2023. A comprehensive review on ensemble deep learning: Opportunities and challenges. Journal of King Saud University – Computer and Information Sciences, 35(2), pp.757–774. Available at: https://www.sciencedirect.com/science/article/pii/S1319157823000228.*

[6] Hu, T., 2025. Financial fraud detection system based on improved Random Forest and Gradient Boosting Machine (GBM). *arXiv*, arXiv:2502.15822. Available at: https://arxiv.org/abs/2502.15822

[7] Ali, A., Razak, S.A., Othman, S.H., Eisa, T.A.E., Al-Dhaqm, A., Nasser, M., Elhassan, T., Elshafie, H. & Saif, A., 2021. *Financial fraud detection based on machine learning: A systematic literature review. Applied Sciences*, 11(14), 6317. Available at: https://www.mdpi.com/2076-3417/12/19/9637

[8] Zadafiya, N., Karasariya, J. & Parthku, P., 2022. *Detecting credit card frauds using Isolation Forest and Local Outlier Factor – Analytical insights. Proceedings of the IEEE International Conference on Computing, Communication and Green Engineering (ICCCGE)*. IEEE. Available at: https://ieeexplore.ieee.org/document/9716541.

[9] Saraf, S. & Phakatkar, A., 2022. *Detection of Credit Card Fraud using a Hybrid Ensemble Model*. *International Journal of Advanced Computer Science and Applications*, 13(9). Available at: https://thesai.org/Publications/ViewPaper?Code=ijacsa&Issue=9&SerialNo=53&Volume=13

[10] Mim, M.A., Majadi, N. & Mazumder, P., 2024. A soft voting ensemble learning approach for credit card fraud detection. *Heliyon*, 10, e1497X. Available at: https://www.sciencedirect.com/science/article/pii/S240584402401497X.

[11] Zheng, Q., Yu, C., Cao, J., Xu, Y., Xing, Q. & Jin, Y., 2024. Advanced payment security system: XGBoost, LightGBM and SMOTE integrated. *arXiv* [preprint], arXiv:2406.04658. Available at: https://arxiv.org/abs/2406.04658 [Accessed 31 July 2025].

[12] P. Sundaravadivel, R. A. Isaac, D. Elangovan, D. KrishnaRaj, V. V. L. Rahul, and R. Raja, "Optimizing credit card fraud detection with random forests and SMOTE," *Scientific Reports*, vol. 15, 2025. [Online]. Available: https://doi.org/10.1038/s41598-025-00873-y

[13] Ibanga, J. (2024). *Resolving data imbalance in financial fraud detection by combining machine learning models and ensemble learning strategies*. ResearchGate. https://doi.org/10.13140/RG.2.2.35330.08644

[14] Ashar, S. (Samayashar), 2025. *Fraud Detection Transactions Dataset*. Kaggle [dataset]. Available at: https://www.kaggle.com/datasets/samayashar/fraud-detection-transactions-dataset

[15] X. Niu, L. Wang, and X. Yang, "A comparison study of credit card fraud detection: Supervised versus unsupervised," *arXiv preprint arXiv:1904.10604*, 2019. [Online]. Available: https://arxiv.org/abs/1904.10604

[16] Ranjan, S., Agrawal, N., Sharma, V., & Pal, S. (2024). *Credit card fraud detection by using ensemble*

method of machine learning. ResearchGate. Retrieved August 1, 2025, from https://www.researchgate.net/publication/378380428_Credit_Card_Fraud_Detection_by_Using_Ensemble_Method_of_Machine_Learning

[17] Mim, M. A., Majadi, N., & Mazumder, P. (2024). *A soft voting ensemble learning approach for credit card fraud detection*. Heliyon, 10(6), e1497X. https://doi.org/10.1016/j.heliyon.2024.e1497X

[18] Wang, X. (2025). *A robust fraud detection system using SMOTE, K-means clustering, and hybrid ensemble learning*. arXiv preprint arXiv:2503.21160. https://arxiv.org/abs/2503.21160