Available online at www.ijournalse.org

Emerging Science Journal

(ISSN: 2610-9182)

Vol. 8, No. 6, December, 2024

SHAP-Instance Weighted and Anchor Explainable AI: Enhancing XGBoost for Financial Fraud Detection

Putthiporn Thanathamathee [1, 2], Siriporn Sawangarreerak [3*], Siripinyo Chantamunee [1, 4], Dinna Nina Mohd Nizam [5]

[1] *School of Engineering and Technology, Walailak University, Nakhon Si Thammarat 80160, Thailand.*

[2] *Research Center for Intelligent Technology and Integration, School of Engineering and Technology, Walailak University, Nakhon Si Thammarat 80160, Thailand.*

[3] *School of Accountancy and Finance, Walailak University, Nakhon Si Thammarat 80160, Thailand.*

[4] *Informatics Innovation Centre of Excellence, Walailak University, Nakhon Si Thammarat 80160, Thailand.*

[5] *User Experience Research Group, Faculty of Computing and Informatics, Universiti Malaysia Sabah, W.P.Labuan, Malaysia.*

Abstract

This research aims to enhance financial fraud detection by integrating SHAP-Instance Weighting and Anchor Explainable AI with XGBoost, addressing challenges of class imbalance and model interpretability. The study extends SHAP values beyond feature importance to instance weighting, assigning higher weights to more influential instances. This focuses model learning on critical samples. It combines this with Anchor Explainable AI to generate interpretable if-then rules explaining model decisions. The approach is applied to a dataset of financial statements from the listed companies on the Stock Exchange of Thailand. The method significantly improves fraud detection performance, achieving perfect recall for fraudulent instances and substantial gains in accuracy while maintaining high precision. It effectively differentiates between non-fraudulent, fraudulent, and grey area cases. The generated rules provide transparent insights into model decisions, offering nuanced guidance for risk management and compliance. This research introduces instance weighting based on SHAP values as a novel concept in financial fraud detection. By simultaneously addressing class imbalance and interpretability, the integrated approach outperforms traditional methods and sets a new standard in the field. It provides a robust, explainable solution that reduces false positives and increases trust in fraud detection models.

Keywords:

Fraud Detection;
SHAP-Instance Weighted;
Anchor Explainable AI;
Optuna with Hyperband.

Article History:

Received:	14	July	2024
Revised:	07	November	2024
Accepted:	12	November	2024
Published:	01	December	2024

1- Introduction

Fraud detection has become a vital concern for firms in several sectors, particularly in the financial industry, given the current digital environment. With the rise of digital transactions and the growing complexity of fraudulent operations, traditional methods of fraud detection such as rule-based systems and expert knowledge have shown little effectiveness in countering new threats [1]. Researchers and experts are exploring advanced methodologies, including machine learning algorithms, to enhance the accuracy and efficiency of fraud detection mechanisms due to the complex and ever-changing nature of fraud [2].

Many research papers in the current literature utilize statistical methods to evaluate financial fraud. Many studies have used ordinary least squares (OLS) regression and autoregressive (AR) models to thoroughly analyze fraudulent actions in the financial sector. Khaksar (2022) [3] examined the correlation between different attributes of auditors and

* **CONTACT**: siriporn.sa@wu.ac.th

DOI: http://dx.doi.org/10.28991/ESJ-2024-08-06-016

the identification of fraudulent activities in the financial statements of firms registered on the Tehran Stock Exchange during the period from 2012 to 2018. The study used a multiple regression model and identifies a strong and statistically significant correlation between various factors, including the size of the audit firm, auditor rotation, industry specialty, auditor market focus, auditor independence, audit report delay, and the identification of fraud through the restatement of financial statements. Nevertheless, it identifies a detrimental and noteworthy correlation between the duration of an auditor's tenure, the level of narcissism displayed by the auditor, the fees charged for the audit, and the specific type of auditor opinion (specifically, an unqualified opinion) with regards to the detection of fraud. The findings indicate that specific attributes of auditors, such as their knowledge in a particular industry, their impartiality, and the duration of their audits, contribute to the effectiveness of fraud detection. Conversely, factors such as prolonged tenure of auditors, narcissistic tendencies, greater fees, and positive audit opinions undermine the ability to identify fraudulent activities. Cordis (2023) [4] discovered a direct and statistically significant correlation between the political alignment index and the incidence of corporate fraud convictions. This implies that places in which lawmakers are more closely affiliated with the President's political party have a greater incidence of corporate fraud convictions.

Rahman & Jie (2024) [5] conducted a study on the correlation between the fraud triangle theory (consisting of pressure, opportunity, and rationalization) and instances of accounting fraud in Chinese firms that are publicly listed. The study utilized data from the period of 2010 to 2020. The authors employ regression analysis to examine the relationship between five factors that represent the three components of the fraud triangle: pressure (measured by leverage and liquidity ratios), opportunity (measured by auditor size), and rationalization (measured by the number of independent directors). The results indicate that having higher leverage ratios and being audited by a non-Big Four firm have a favorable effect on fraud detection. Conversely, having higher return on net equity and a larger auditor size have a negative influence on fraud detection. Maniatis (2022) [6] using the Beneish model (M-score) to ascertain if companies listed in the general index of the Athens Stock Exchange Market engaged in earnings manipulation over the period of 2017-2018. The findings indicated that only the variables of total accruals to total assets (TATA) and sales growth index (SGI) exhibited a statistically significant correlation with the M-score, which serves as an indicator of profits manipulation. Gong et al. [7] devised two variations of Benford factors, derived from Benford's Law, with the purpose of identifying instances of financial fraud in Chinese publicly traded corporations. Incorporating these Benford factors into a logistic regression model alongside conventional financial and non-financial indicators enhances the accuracy of fraud detection, thereby reducing both Type I (misclassification of fraudulent firms as non-fraudulent) and Type II (misclassification of non-fraudulent firms as fraudulent) errors when compared to using solely traditional indicators. Although statistical methods are commonly used in many studies to examine the causes and effects of financial fraud, the complex and scalable nature of fraudulent operations presents considerable obstacles that cannot be adequately addressed using statistical methods alone [1, 2].

In recent years, several researchers have applied machine learning (ML) and artificial intelligence (AI) methods for financial fraud detection due to their ability to handle large, complex datasets and identify intricate patterns. Xiuguo & Shengyong (2022) [8] constructed a multi-dimensional financial fraud indicator system and presents a Chinese textual data mining framework using word embedding and deep learning models like CNN, LSTM, and GRU. The empirical results show significant performance improvements of deep learning over traditional machine learning methods, with LSTM and GRU achieving around 95% accuracy in correctly classifying fraudulent and non-fraudulent cases. Craja et al. [9] proposed a deep learning approach for detecting financial statement fraud by combining information from financial ratios and textual data in corporate annual reports, specifically the Management Discussion & Analysis (MD&A) section. A hierarchical attention network (HAN) is employed to extract text features from the MD&A while capturing the structured hierarchy and contextual information of the documents. Pai et al. [10] introduced a support vector machine-based fraud warning (SVMFW) model to detect top management fraud in financial statements. The model integrates sequential forward selection (SFS) for feature selection, support vector machine (SVM) with particle swarm optimization (PSO) for parameter determination, and classification and regression tree (CART) to generate interpretable decision rules. The results show the SVMFW model outperforms other classifiers like logistic regression, discriminant analysis, decision trees, and neural networks in accurately detecting fraudulent financial statements.

Alfaiz & Fati (2022) [11] proposed an enhanced credit card fraud detection model using machine learning algorithms and various resampling techniques to address the class imbalance issue in the dataset. AllKNN undersampling technique with CatBoost, achieving an AUC of 97.94%, Recall of 95.91%, and F1-Score of 87.40%, outperforming previous works on the same European credit card transaction dataset. Strelcenia & Prakoonwit (2023) [12] explored various data augmentation techniques, including the introduction of a new model called K‑CGAN. After evaluating the performance of these augmentation methods using various classification techniques, the findings in-dicate that B‑SMOTE, K‑CGAN, and SMOTE outperform other methods in terms of Precision and Recall. Among these, K‑CGAN exhibits the highest F1 Score and Accuracy, making it a promising approach for enhancing credit card fraud detection in the face of imbalanced data. Chaquet-Ulldemolins et al. [13] proposed a novel methodology to apply machine learning techniques for credit fraud detection while maintaining interpretability and addressing issues like bias and lack of transparency. The key steps involve using the informative variable identifier (IVI) algorithm to select relevant features, applying recurrent

feature filters (RFF and MIFF) to further reduce dimensionality and bias, and using linear models like logistic regression, support vector machines, and gradient boosting to obtain interpretable weights for the selected features. The approach is first validated on a synthetic dataset and then applied to a real German credit dataset, achieving over 76% accuracy while identifying key interpretable features like living beyond one's means and unexpected overdrafts consistent with existing literature. The methodology allows extracting insights from powerful machine learning models while complying with regulations requiring interpretability and non-discrimination. Zhao & Bai (2022) [14] presented a method for detecting and predicting financial fraud in listed companies using machine learning algorithms and the SMOTE technique for handling imbalanced data. From a dataset of 18,060 transactions and 363 financial indicators, 13 relevant indicators were extracted using multiple feature selection models. Five single classification models (LR, RF, XGBoost, SVM, DT) and three ensemble models with voting classifiers were established. The optimal single model achieved 97-99% accuracy, while the best ensemble model combining logistic regression and XGBoost had over 99% accuracy, outperforming existing methods.

Ali et al. [15] proposed a Financial Statement Fraud (FSF) detection model using the XGBoost (eXtreme Gradient Boosting) ensemble learning technique on data from publicly available financial statements of firms in the Middle East and North Africa (MENA) region. The Synthetic Minority Oversampling Technique (SMOTE) is applied to address class imbalance in the dataset. After comparing various machine learning classifiers like Logistic Regression, Decision Tree, Support Vector Machine, Random Forest, and AdaBoost, the XGBoost algorithm is found to outperform the others. Further optimization of the XGBoost algorithm through hyperparameter tuning yields an accuracy of 96.05% in detecting FSF. Cheah et al. [16] explored techniques to address the class imbalance problem in financial fraud datasets. It evaluates the effectiveness of SMOTE, Generative Adversarial Networks (GANs), and their hybrid variants like SMOTified-GAN, SMOTE+GAN, and a proposed GANified-SMOTE method in generating synthetic fraud samples. These oversampling techniques are applied before training Feed-forward Neural Networks (FNNs), Convolutional Neural Networks (CNNs), and a hybrid FNN+CNN model. The results demonstrate that hybrid SMOTE-GAN methods outperform individual SMOTE or GAN, with the proposed GANified-SMOTE and SMOTified-GAN performing well across different amounts of generated fraud data. El Hlouli et al. [17] introduced a novel stacked autoencoder kernel extreme learning machine optimized by the dandelion algorithm (S-AEKELM-DA) for detecting fraudulent credit card transactions. Experimental results on four credit card fraud datasets demonstrate that S-AEKELM-DA achieves superior performance compared to traditional kernel ELM, stacked autoencoder ELM without optimization. Raval et al. [18] presented RaKShHA, a credit card fraud detec-tion scheme that integrates explainable artificial intelligence (XAI) with a long short-term memory (LSTM) model, called X-LSTM. XAI helps extract important features from the credit card fraud dataset, improving the LSTM model's performance. El Kafhali et al. [19] proposed an intelligent system for detecting fraudulent credit card transactions using three deep learning architectures: artificial neural networks (ANN), recurrent neural networks (RNN), and long short-term memory (LSTM). The study highlights the effectiveness of the RNN architecture combined with Bayesian hyperparameter optimization for detecting fraudulent credit card transactions.

While numerous studies have applied machine learning and artificial intelligence methods for financial fraud detection, there remain challenges in improving the performance and interpretability of these models. Previous works have focused on various techniques such as deep learning, ensemble methods, and handling imbalanced data to enhance fraud detection accuracy. However, there is a lack of research on leveraging instance weighting techniques, particularly based on feature importance, to prioritize learning from the most informative instances and improve model performance. Additionally, although some studies have addressed interpretability issues in fraud detection models, there is still a need for more effective and concise explanations that provide clear insights into the model's decision-making process. To bridge this gap, we propose the following research questions:

RQ1: How can we enhance the performance of a model for fraud detection by relevant instance weighting?

RQ2: How can we enhance the interpretability of a model for fraud detection tasks?

Based on previous research and our research questions, this study intends to provide the following contributions. First, we develop a novel approach to improve the performance of fraud detection models by incorporating relevant instance weighting techniques. Unlike traditional methods that treat all instances equally, we propose the use of SHAP (SHapley Additive exPlanations) values to quantify the importance of each instance based on its feature contributions. This innovative approach sets our work apart from previous studies. While prior research has utilized SHAP values primarily to measure feature importance and understand the impact of each input feature on the model's output, offering important insights into the model's decision-making process, our approach goes significantly further. We extend the application of SHAP values beyond feature importance to instance weighting, a novel concept in fraud detection. We design a weighting scheme that assigns higher weights to instances that are more informative or influential in the model's prediction process, a method not commonly applied in fraud detection research. This unique use of SHAP values for instance weighting represents a key innovation in our work.

The SHAP-based instance weights are integrated into the training process of the fraud detection model, such as XGBoost, to prioritize learning from the most relevant instances, enhancing the model's ability to detect fraud more

accurately. In our approach, we highlight the importance of identifying key features specific to each category of fraud, differing from earlier studies that prioritized crucial features for fraud detection as a whole. This nuanced perspective allows for more precise and targeted fraud detection, addressing a gap in existing literature. Second, we enhance the interpretability of fraud detection models through the application of explainable AI techniques. Our unique combination of SHAP-based instance weighting and Anchor Explainable AI in financial fraud detection represents a novel contribution. We implement Anchor Explainable AI to generate clear and concise explanations for the model's predictions in the form of if-then rules. Going beyond mere implementation, we rigorously evaluate the precision and coverage of the generated explanations to ensure their statistical robustness and generalizability, an aspect often overlooked in previous studies. By providing stakeholders with a transparent and understandable interpretation of the model's decision-making process, we aim to increase trust and accountability in the fraud detection system. This comprehensive approach, addressing both performance enhancement and interpretability challenges simultaneously, distinguishes our work from previous research that often focused on only one aspect. Our integrated method not only improves fraud detection accuracy but also provides valuable insights into the decision-making process, setting a new standard in the field of financial fraud detection.

The remainder of this paper is organized as follows: Section 2 provides the related works. Section 3 delves into the details of our proposed method. Section 4 presents the experimental setup and results. Section 5 engages in discussion. Finally, Section 6 summarizes the conclusions.

2- Related Works

2-1- Altman Z-Score Model

The Altman Z-score model, created by Altman (1968) [20], is a financial model utilized to forecast the likelihood of a company's insolvency. The model identifies and quantifies particular financial parameters that can function as indicators or predictors of a company's probability of facing legal action for bankruptcy. According to the study, Equation 1 has a high level of accuracy, up to 94%, in predicting non-bankrupt enterprises one year ahead.

$$Z = 0.012X_1 + 0.014X_2 + 0.033X_3 + 0.006X_4 + 0.999X_5 \quad (1)$$

where: X_1represents Working Capital/Total Assets, X_2 represents Retained Earnings/Total Assets, X_3 represents EBIT/Total Assets, X_4 represents Market Value Equity/Book Value of Total Debt, X_5 represents Sales/Total Assets, and Z represents the Overall Index.

The criteria for consideration are as follows: if the value of Z is greater than 2.99, it predicts that the business will not experience bankruptcy. However, if the value of Z is less than 1.81, it predicts that the business will experience bankruptcy. If the value of Z falls between 1.81 and 2.99, it is classified as being in the "Grey area," indicating that the company may or may not go bankrupt.

2-2- EM Score Model

Altman (1983) [21] introduced the Emerging Market Scoring Model (EM Score Model), which is an additional bankruptcy prediction model. This approach eliminates sales-related factors to specifically target the problem of variations in company size. The EM Score Model is applicable to non-manufacturing organizations and has been empirically validated on small and medium-sized businesses in the United States. Equation 2 represents the model.

$$Z = 3.25 + 6.56X_1 + 3.26X_2 + 6.72X_3 + 1.05X_4 \quad (2)$$

where: X_1 is Working Capital/Total Assets, X_2 is Retained Earnings/Total Assets, X_3 is EBIT/Total Assets, X_4 is Book Value Equity/Book Value of Total Debt, and Z is Overall Index.

The criteria for consideration are as follows: if the value of Z is greater than 2.6, it predicts that the business will not experience bankruptcy. However, if the value of Z is less than 1.11, it predicts that the business will experience bankruptcy. If the value of Z falls between 1.11 and 2.6, it is classified as being in the "Grey area," indicating that the company may or may not go bankrupt.

2-3- Explainable Artificial Intelligence (XAI)

Explainable Artificial Intelligence (XAI) is a discipline that specifically aims to improve the comprehensibility and clarity of Machine Learning (ML) algorithms and the results they produce. With the increasing prominence of ML algorithms, particularly opaque models, there is a growing demand to comprehend and elucidate these techniques in numerous fields, notably in finance-related contexts [22]. XAI seeks to fulfill this requirement by establishing methodologies and strategies that offer understanding in the decision-making procedure of these intricate models, hence enhancing their transparency and reliability. Medianovskyi et al. [23] introduced an explainable AI (XAI) method for

predicting financial crises in small and medium-sized firms (SMEs). They utilized different regression models, such as Gradient Boosting (XGBoost and Catboost), Random Forests, Logistic Regression, and Artificial Neural Networks. The researchers utilized the Shapley's Additive Explanations (SHAP) framework to analyze and confirm the accuracy of the predictions made by these models. The study's findings demonstrated that the Catboost model exhibited superior performance compared to the other models in accurately anticipating financial crises within small and medium-sized enterprises (SMEs). Torky et al. [24] introduced an Explainable Artificial Intelligence (XAI) model that utilizes the Pigeon Inspired Optimization (PIO) algorithm and the Gradient Boosting classifier to identify the underlying causes of financial crises. The feature selection was optimized using the PIO algorithm, and the classification and recognition of financial crisis roots were compared using Gradient Boosting and Random Forest classifiers.

The XAI model demonstrated superior performance compared to the Random Forest classifier, achieving training and testing accuracies of 99% and 96.7%, respectively. Tran et al. [25] utilized machine learning methods to predict the financial instability of publicly listed firms in Vietnam from 2010 to 2021. They employed SHAP (Shapley Additive Explanations) values to understand the results of the model. The study discovered that extreme gradient boosting and random forest models outperformed other models in terms of recall, F1 scores, and AUC (Area Under the Curve). The researchers found that certain factors, including the long-term debts to equity ratio, enterprise value to revenues ratio, account payable to equity ratio, and diluted earnings per share (EPS), had a significant influence on the model outputs. This suggests that these factors are important in predicting financial distress among Vietnamese listed companies, as determined by the Shapley values. Nallakaruppan et al. [26] proposed an explainable AI (XAI) framework that integrates machine learning models with techniques like Local Interpretable Model-Agnostic Explanations (LIME), Shapley Additive Explanations (SHAP), and partial dependence plots to evaluate credit risk. The researchers performed trials on a loan prediction dataset and found that the random forest model beat other models, reaching the greatest accuracy of 0.998, sensitivity of 0.998, and specificity of 0.997.

2-4- SHapley Additive exPlanations (SHAP)

Lundberg & Lee [27] developed SHAP, a comprehensive framework that has proven to be an effective method for enhancing the interpretability of machine learning models. SHAP values measure the impact of each input feature on the model's output, offering important insights into the model's decision-making process. By determining the contribution of individual features, SHAP enables users to understand the factors influencing the model's predictions, thereby increasing model transparency and explainability. Shapley values, denoted as $\emptyset_i$, for a specific feature i are calculated by taking the average of the incremental contributions of feature i over all potential combinations S of the set of features N, excluding feature i itself, as shown in Equation 3.

$$\emptyset_i = \frac{1}{n!}\sum_{S\in N-\{i\}}|S|!\,(n-1-|S|)!\,[f(S\cup\{i\})-f(s)] \quad (3)$$

where S represents a subset of features, f refers to the predictive model, N is the complete set of features, and n is the number of features. For a given observation x, the prediction model's output is elucidated using a linear function g, which is derived from Equation 4.

$$f(x) = g(x^*) = \emptyset_0 + \sum_{i=1}^{M} \emptyset_i x_i \quad (4)$$

Here, x represents the specific instance under examination, x_i denotes the simplified version of the input, M signifies the total number of these simplified input features, and $\emptyset_0$ is the baseline value used when no input features are present.

2-5- Anchor (Anchors: High-Precision Model-Agnostic Explanations)

The ANCHOR techniques provide explanations for predictions made by black-box classification models by defining decision limits that serve as effective "anchors" for the prediction. An anchor explanation is a principle that links a forecast to a particular local environment within the instance being analyzed [28]. This rule guarantees that modifications to the instance's other feature values do not have a substantial impact on its capacity to clarify the forecast. The Anchors technique employs reinforcement learning methods and a graph search algorithm to reduce the amount of model calls required during runtime and effectively overcome local optima. The Anchor method utilizes a perturbation-based approach to provide local explanations in the form of straightforward IF-THEN rules. In contrast to LIME's surrogate models, this approach generates rules that can be scored and effectively explain occurrences that have not been encountered before. The concept of coverage in Anchor ensures that these rules can accurately elucidate any possible unseen situations. The Anchor approach utilizes reinforcement learning techniques to tackle the exploration or multi-armed bandit issue in order to discover anchors. Anchor analyzes the neighboring instances or perturbations of each given instance, enabling the algorithm to function without taking into account the internal structure and parameters of the black-box model. As a result, the structure and characteristics of the black box stay concealed and unchanged

throughout the procedure. Anchor's model-agnostic technique is universally applicable to any type of models, as it does not rely on the exact internal mechanisms of the black box being discussed. Through the examination of the disruptions in each case, Anchor has the ability to produce specific explanations in the shape of straightforward and reusable IF-THEN rules. This allows for a lucid and comprehensible comprehension of the model's forecasts [29]. Equation 5 provides a formal definition of an anchor.

$$\mathbb{E}_{\mathcal{D}x(z|A)}\left[1_{\hat{f}(x)=\hat{f}(z)}\right] \geq \tau, A(x) = 1 \tag{5}$$

where x signifies the data point under investigation, such as a record in a dataset. A represents a set of conditions or the generated rule/anchor, where $A(x) = 1$ is true if the features specified by A match the attributes of x. f denotes the model being explained, like an artificial neural network, used to predict outcomes for x and its perturbations. $\mathcal{D}_x(\cdot|A)$ is the distribution of instances near x that satisfy A. The parameter $0 \leq \tau \leq 1$ sets a precision threshold, with rules needing to achieve a local fidelity of at least τ to be considered reliable explanations.

2-6- Optuna for Hyperparameter Optimization

Optuna is a popular open-source package that simplifies the process of tweaking hyperparameters in machine learning applications [30]. It provides numerous notable advantages that make it a powerful tool for optimizing hyperparameters. An important benefit of this is its define-by-run style API, which is influenced by deep learning frameworks. This API allows users to programmatically specify the range of values to search for hyperparameters, providing the ability to easily adjust and customize the optimization process. This feature is especially beneficial when working with intricate or developing machine learning models, as it enables effortless modification of hyperparameters as the project advances. Optuna offers a noteworthy benefit in terms of its effective pruning and sampling process, which encompasses two primary strategies: efficient searching and efficient performance estimation. These strategies synergistically combine to establish a cost-efficient optimization approach that intelligently searches the hyperparameter space while limiting the computational resources needed. Optuna can expedite the process of hyperparameter tuning by effectively searching for potential hyperparameter combinations and properly assessing their performance. This allows for the rapid identification of ideal settings, resulting in a reduction in the overall time and resources required for hyperparameter tuning. Optuna is widely acclaimed for its straightforward setup, which makes it accessible to both experienced and inexperienced machine learning practitioners. The user-friendly interface and detailed documentation of Optuna enable users to seamlessly incorporate it into their current projects, simplifying the process of adjusting hyperparameters and expediting the creation of high-performing models. This level of accessibility allows a broader spectrum of users to utilize the potential of hyperparameter optimization, irrespective of their proficiency in machine learning.

2-7- Hyperband Bandit-Based Approach

Hyperband is a novel hyperparameter optimization approach introduced by Li et al. [31]. It effectively allocates computational resources to different alternative configurations. This method enhances the process of finding the best hyperparameters in machine learning models by efficiently devoting additional resources to promising configurations and swiftly discarding poor ones. The main benefit of Hyperband is its flexible allocation of resources mechanism. First, the algorithm allocates a limited number of resources to each setup. During the optimization process, Hyperband incrementally raises the allocation for configurations that demonstrate potential. This strategy enables the algorithm to efficiently recognize and give priority to the most promising configurations at an early stage, thereby conserving computational resources and time. In addition, Hyperband efficiently manages the trade-off between exploration and exploitation, which are two crucial components of efficient optimization. Hyperband guarantees a comprehensive exploration of the hyperparameter space by trying out various hyperparameter configurations and prioritizing the most promising ones, which are likely to yield optimal performance. The results of our investigation show that the combination of Hyperband and Optuna improves the process of adjusting hyperparameters. Hyperband is a pruning technique that dynamically ranks hyperparameter sets based on their performance and eliminates those that are less promising. By integrating Hyperband with Optuna, the process of searching for ideal hyperparameters becomes more streamlined and impactful, resulting in enhanced optimization results. This integration expedites the process of finding the optimal hyperparameter settings, ultimately leading to the creation of machine learning models that perform better.

3- Proposed Methodology

Figure 1 illustrates the proposed technique for detecting financial fraud. It consists of three main processes: 1) Advanced optimization of XGBoost with SHAP-Instance Weighting preceded by combining RENN and SMOTE balancing., 2) Generating Anchor explanations for XGBoost model predictions., and 3) Feature importance analysis using SHAP.

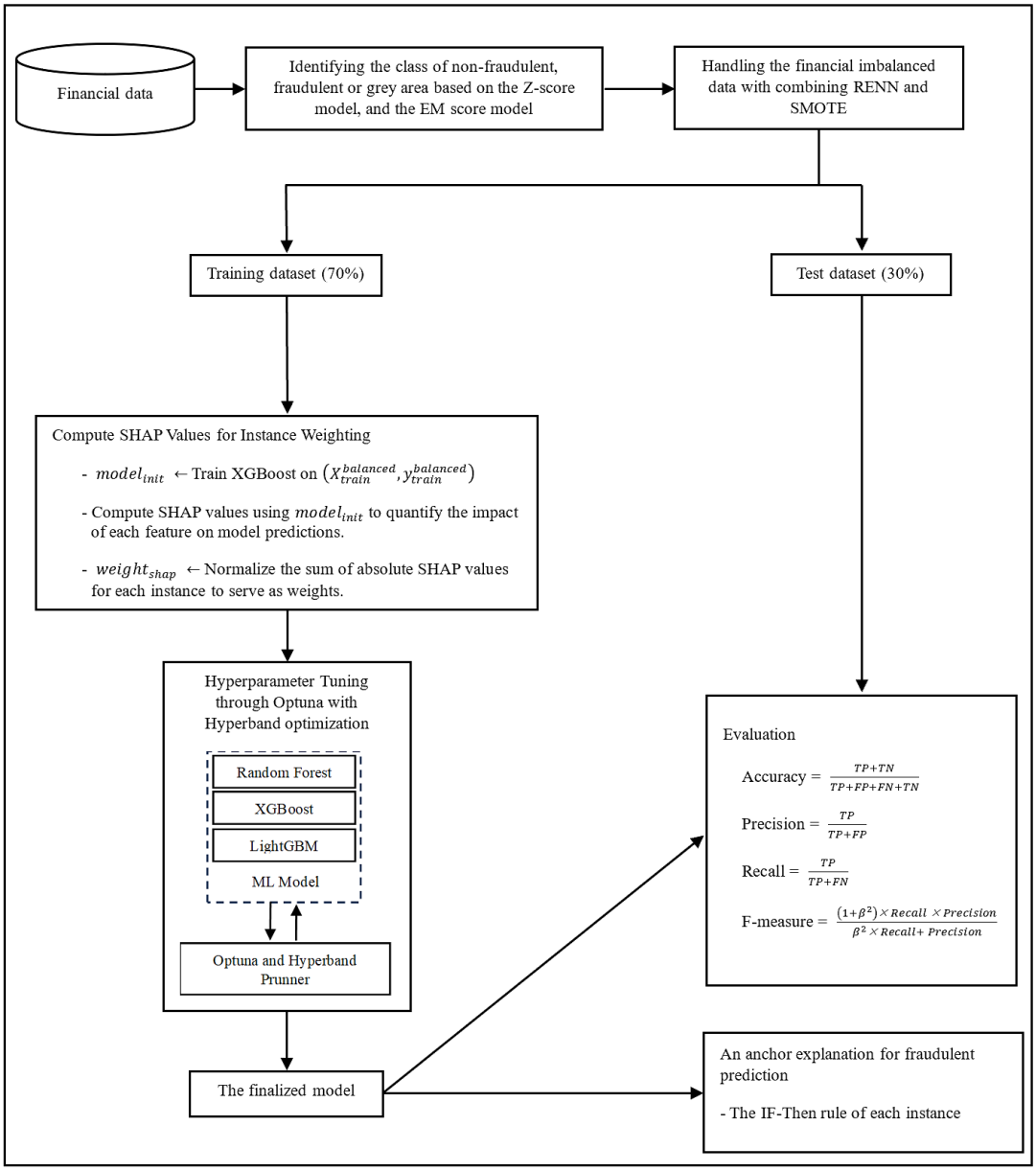

Figure 1. Proposed method for financial fraud detection

Algorithm 1 outlines a method for detecting fraudulent financial statements. Here's a summary of the algorithm:

- Load the training and testing datasets and perform necessary preprocessing steps, including scaling, and handling missing values.
- Apply the RENN (Repeated Edited Nearest Neighbors) method to undersample the data. RENN iteratively removes instances that are misclassified by their nearest neighbors, refining the training set by eliminating noisy or misleading samples. This step produces a more balanced dataset.
- Apply SMOTE (Synthetic Minority Over-sampling Technique) to the dataset obtained after RENN. SMOTE generates synthetic samples for the minority class, resulting in a more balanced dataset.
- Train an initial XGBoost model on the balanced data to serve as a basis for computing SHAP (SHapley Additive exPlanations) values. The criteria determining instance importance are based on the magnitude of SHAP values. Instances with larger absolute SHAP values across features are considered more influential in shaping the model's decision boundaries. The SHAP-based instance weighting process involves:
  - Initial model training: An initial XGBoost model is trained on the balanced dataset.
  - SHAP value computation: SHAP values are calculated for each feature across all instances in the training set using this initial model. These values quantify the impact of each feature on the model's predictions.
  - Instance weight derivation: For each instance, the absolute SHAP values across all features are summed. This sum represents the total impact of that instance on the model's predictions.

- Weight normalization: The sums are normalized to create instance weights, ensuring they add up to 1 across all instances.
- Weight integration: These normalized weights are incorporated into the XGBoost training process, giving higher importance to instances with larger SHAP value sums.

- Optimize the model parameters using Optuna, an optimization framework:
  - Configure the objective function to integrate SHAP-influenced instance weights in the training process and dynamically adjust XGBoost parameters based on trial suggestions from Optuna.
  - Use cross-validation on the balanced dataset to evaluate the model's performance and minimize log loss.
  - Execute an Optuna study with a HyperbandPruner to efficiently search through the hyperparameter space and find the optimal settings that minimize the defined objective function.
- Train the final model using the optimized parameters and instance weights:
  - Configure the final XGBoost model with the best parameters identified from the Optuna optimization.
  - Train the final model on the entire balanced training dataset, using the SHAP-influenced instance weights to prioritize learning from instances deemed more important based on the SHAP analysis.

Algorithm 1. Advanced Optimization of XGBoost with SHAP-Instance Weighting Preceded by combining RENN and SMOTE Balancing

Inputs:
- X_{train}: Training feature matrix, where $X_{train} \in \mathbb{R}^{n \times m}$, with n instances and m features.
- y_{train}: Training labels vector, where $y_{train} \in \mathbb{R}^{n}$.
- X_{test}: Test feature matrix, analogous to X_{train}.
- y_{test}: Test labels vector, analogous to y_{train}.

Outputs:
- $model_{final}$: Optimally trained XGBoost classifier.
- $accuracy_{final}$: Accuracy of $model_{final}$ on X_{test}.
- Additional metrics: Confusion matrix, precision, recall, and F1-score.

Step 1: Initialization
1.1 Data Loading and Preprocessing:
- datasets X_{train}, y_{train}, X_{test}, and y_{test} are correctly loaded and preprocessed.

Step 2: Handle Imbalance Data with RENN
2.1 Initialize Repeated Edited Nearest Neighbours (RENN):
- Configure RENN to iteratively edit the training set by removing samples that are misclassified by their k nearest neighbors.

2.2 Apply RENN to Training Data:
- $(X_{train}^{balanced}, y_{train}^{balanced})$ ← Apply RENN to X_{train}, y_{train} to achieve a more balanced dataset.

2.3 Apply SMOTE:
- $(X_{train}^{balanced}, y_{train}^{balanced})$ ← Apply SMOTE to the dataset obtained after RENN to further balance the classes by generating synthetic samples for the minority class.

Step 3: Compute SHAP Values for Instance Weighting
3.1 Train Initial XGBoost Model:
- $model_{init}$ ← Train XGBoost on $(X_{train}^{balanced}, y_{train}^{balanced})$

3.2 Calculate SHAP Values:
- Compute SHAP values using $model_{init}$ to quantify the impact of each feature on model predictions.

3.3 Generate Instance Weights:
- $weight_{shap}$ ← Normalize the sum of absolute SHAP values for each instance to serve as weights.

Step 4: Hyperparameter Tuning via Optuna
4.1 Define Objective Function f
- Configure f to optimize XGBoost parameters dynamically, incorporating $weight_{shap}$ into the model training.
- Use cross-validation on $X_{train}^{balanced}$ with $weight_{shap}$ to evaluate model performance.

4.2 Execute Optuna:
- Configure and execute an Optuna study to minimize the log loss and identify the best model parameters $param_{best}$.

Step 5: Train Optimized Model
5.1 Final Model Configuration:
- $model_{final}$ ← Configure the final XGBoost model with $param_{best}$.

5.2 Train Model with Instance Weights:
- Train $model_{final}$ on $X_{train}^{balanced}$ using $weight_{shap}$.

Step 6: Performance Evaluation
6.1 Make Predictions:
- $predictions$ ← Use $model_{final}$ to generate predictions on X_{test}.

6.2 Compute Metrics:
- Evaluate $accuracy_{final}$, confusion matrix, and other performance metrics.

The process in Algorithm 2 for generating Anchor Explanations for XGBoost model predictions can be summarized as follows:

- Provide the explainer with the training data, which allows it to understand the data distribution and relationships between features.
- The explainer uses this information to simulate perturbations around the instance being explained. This helps identify the crucial features and their values that significantly influence the model's prediction.
- Apply the Anchor algorithm to the selected test instance using the explainer. The algorithm generates perturbed versions of the instance and observes the consistency of the model's output.
- The Anchor algorithm aims to find the smallest set of features and their conditions that, when held constant, maintain the same prediction across various perturbations.
- Set a high confidence threshold = 0.95 to ensure the anchor is statistically robust and reliable. This threshold determines the minimum required precision before considering an anchor valid.
- Evaluate the anchor's precision by calculating how frequently the model's prediction remains consistent when the anchor's specified conditions are met, compared to a wider range of data instances.
- Assess the anchor's coverage by determining the fraction of the dataset where the anchor applies. High coverage indicates a generalizable rule, while low coverage may suggest overfitting to the specific instance's context.
- Present the anchor as a set of easy-to-understand if-then rules, making it accessible to stakeholders.

Algorithm 2. Generating Anchor Explanations for XGBoost Model Predictions

Objective: To provide interpretable, rule-based explanations for predictions made by an XGBoost model using the Anchor Explainer methodology, enhancing transparency and aiding in compliance with regulatory requirements.

Inputs:
- X_{train}: Training feature matrix, where $X_{train} \in \mathbb{R}^{n \times m}$, with n instances and m features.
- y_{train}: Training labels vector, where $y_{train} \in \mathbb{R}^{n}$.
- X_{test}: Test feature matrix, analogous to X_{train}.
- y_{test}: Test labels vector, analogous to y_{train}.
- $model_{final}$: A trained XGBoost classifier.
- $class\_names$: A list of class names in the target variable.
- $feature\_names$: A list of feature names in the dataset.
- idx: Index specifying the test instance to be explained.

Outputs:
- Anchor: A set of rules that explain the conditions under which the model makes a specific prediction for the test instance.
- Precision: The accuracy of the anchor in predicting the same outcome as the model for other similar instances.
- Coverage: The proportion of instances that the anchor applies to within the dataset.

Steps:
1. **Initialize the Anchor Explainer:**
    - Construct an AnchorTabularExplainer using the training dataset X_{train}, class names $class\_names$, feature names $feature\_names$.
2. **Prepare the Instance for Explanation:**
    - Select the instance at index idx from X_{test} for which the explanation is to be generated.
3. **Generate the Anchor Explanation:**
    - Use the Anchor Explainer to compute an explanation for the selected instance.
    - Set a confidence threshold (0.95) to determine the required precision of the anchor.
4. **Evaluate the Explanation:**
    - Determine the precision of the anchor, which measures how often the model prediction remains the same under the conditions specified by the anchor across all data points.
    - Calculate the coverage, which quantifies the proportion of dataset instances that satisfy the anchor's conditions.
5. **Report the Explanation:**
    - Present the anchor as a set of if-then rules.

The steps outlined in Algorithm 3 for conducting Feature Importance Analysis with SHAP can be concisely described as follows:

- Load the trained model and the feature matrix along with the feature names.
- Create the SHAP explainer object based on the trained model. Compute SHAP values for each instance in the feature matrix. These values represent the contribution of each feature to the model's prediction for each instance, providing a local explanation for the model's behavior.
- By averaging the SHAP values for each feature across all instances, obtain a global measure of feature importance. The mean SHAP values indicate the overall impact of each feature on the model's predictions for each class.

- Create DataFrames for feature importance for each class. Populate these DataFrames with features and their corresponding mean SHAP values.
- Sort the DataFrames in descending order of importance values to highlight the most influential features. Display these sorted DataFrames to help understand which features are most critical for the model's predictions for each class.

Algorithm 3. Feature Importance Analysis Using SHAP

Inputs:
- X_{train}: Training feature matrix, where $X_{train} \in \mathbb{R}^{n \times m}$, with n instances and m features.
- $model_{final}$: A trained XGBoost classifier.
- $feature\_names$: A list of feature names in the dataset.

Outputs:
- $importance\_class_i$: feature importance for class i.

Step1: Initialize
1.1 Load the trained $model_{final}$.
1.2 Load the feature matrix X_{train} and the list of feature names $feature\_names$.

2. Compute SHAP Values:
2.1 Initialize the SHAP explainer using the trained model $model_{final}$.
- $explainer \leftarrow$ shap.Explainer($model_{final}$)

2.2 Compute SHAP values for the feature matrix X_{train} using the explainer.
- $shap\_values \leftarrow explainer$.shap_values($X_{train}$)

3. Calculate Mean SHAP Values:
3.1 Compute the mean SHAP value for each feature across all instances.

4. Create DataFrames for Feature Importance:
4.1 Initialize $importance\_dfs$ as an empty.
4.2 Set $n_{classes}$ to the number of classes in $mean\_shap\_values$.
For i from 0 to $n_{classes} - 1$:
- 4.2.1 create $importance\_df$ a DataFrame with columns:
  - 'feature': $feature\_names$
  - '$importance\_class_i$': $mean\_shap\_values[:, i]$
- 4.2.2 Sort $importance\_df$ by $importance\_class_i$ in descending order.
- 4.2.3 Append $importance\_df$ to $importance\_dfs$.

5. Sort and Display Feature Importances:
For i from 0 to $n_{classes} - 1$:
- 5.1 Print "Feature importance for class i:"
- 5.2 Print $importance\_dfs[i]$.

4- Experimental Setup and Results

4-1- Data Description

In a comprehensive analysis, we examined the financial statements of 959 companies listed on the Stock Exchange of Thailand over a ten-year period from 2013 to 2022. The study encompassed a total of 7,530 financial statements, each of which was evaluated using a set of 35 carefully selected accounting descriptor variables [32], along with variables from the Altman Z-score and EM score models, resulting in a total of 40 variables used in this research. Based on these variables, we classified the statements into three categories: non-fraudulent (class 0), fraudulent (class 1), and a grey area (class 2) for statements that exhibited ambiguous characteristics that could not be definitively classified as either fraudulent or non-fraudulent.

To illustrate the findings, Table 1 presents a representative sample of fraud findings and descriptor variables for three of the companies included in the study. Out of the total 7,530 financial statements analyzed, 410 were identified as non-fraudulent (class 0), 312 were determined to be fraudulent (class 1), and a substantial 6,808 statements were classified in the grey area (class 2), highlighting the complexity and uncertainty involved in detecting financial statement fraud.

To identify the class of non-fraudulent, fraudulent or grey area, this research defines three classes: 0 for non-fraudulent, 1 for fraudulent, and 2 for grey area. These groups are derived from the values of the Z-score model and the EM Score model, which is consistent with the research by Elewa [33], who stated that applying Altman Z-Score models has a significant impact on the quality of financial distress predictability. This study has set the following conditions:

- If the Z-score is below 1.81 and the EM Score is below 1.11, the company is classified as non-fraudulent.
- If the Z-score exceeds 2.99 and the EM Score exceeds 2.6, the company is classified as fraudulent.
- Otherwise, the company is assumed to be in the grey area, highlighting the complexity and uncertainty involved in detecting financial statement fraud.

Table 1. Example of descriptor variables and fraud results from three companies

Accounting descriptors and fraudulent results	Company A	Company B	Company C
Total liabilities	169,244.41	929,984.31	5,524,225.62
Total assets	1,194,886.85	934,723.9	11,507,070.22
Gross margin	313,414.43	-168,394.35	387,359.79
Net Profit	262,708.26	-214,395.69	313,071.37
Primary business income	1,365,663.97	2,614,806.62	14,993,711.63
Cash and deposit	145,315.63	76,886.89	338,199.11
Accounts receivable	354,774.37	353,327.71	1,911,728.06
Inventory/Primary business income	0.240747956	0.117177208	0.073995055
Inventory/total assets	0.275156438	0.327792774	0.096415551
Gross profit/Total asset	0.262296325	-0.180154108	0.033662764
net profit/Total asset	0.219860366	-0.229367934	0.027206871
Currents assets/Total asset	0.704869921	0.788051242	0.32013723
net profit/Primary business income	0.192366692	-0.081992943	0.020880178
Accounts receivable/Primary business income	0.259781599	0.135125752	0.127501989
Primary business income/Total asset	4.35737426	-15.5278762	38.70745497
Current assets/Current liabilities	5.671755427	0.814454876	0.992302173
Primary business income/Total asset	4.35737426	-15.5278762	38.70745497
Cash/Total asset	0.121614553	0.082256258	0.029390549
Inventory/Current liabilities	2.214053935	0.338775462	0.298851091
Total debt/total equity	0.166090406	942.0901687	0.952207061
Long term debt/Total asset	0.017363326	0.027348204	0.157451588
Net profit/Gross profit	0.838213671	1.273176267	0.808218556
Total debt/Total asset	0.141640533	0.994929422	0.480072296
Total asset/Capital and reserve	1.184270352	946.8914552	1.982155913
Long term debt/Capital and reserve	0.020562872	25.89578078	0.312093596
Fixed assets/Total asset	0.261297335	0.176481312	0.520956392
Deposit and Cash/Current assets	0.172534746	0.104379326	0.091806095
Capital and reserve/Total debt	5.961582956	0.00106147	1.050885847
Accounts receivable/Total asset	0.296910431	0.378002221	0.166135082
Gross profit/Primary business income	0.229496008	-0.064400307	0.025834817
Undistributed profit/net profit	1	1	1
Primary business profit/Last year's Primary business profit	1.069390559	-3.648553661	0.853148258
Primary business income/Last year's Primary business income	1.115159338	0.547763025	0.98665878
Accounts receivable/Last year's Accounts receivable	1.325021435	1.257276133	1.00609027
Total asset/Last year's Total asset	1.180041223	0.750837455	0.952386286
Working capital/total assets	0.580592715	-0.179529977	-0.002483478
Retained Earnings / Total Assets	0.442730992	-0.576654614	0.288378561
Earnings Before Interest and Taxes / Total Assets	0.262296325	-0.180154108	0.046508434
Sales / Total Assets	1.134496844	2.770217248	1.263091101
Book Value of Equity / Book Value of Total Debt	6.020817349	0.00106147	1.05019175
Z-score	4.14	0.40	2.21
EM score	21.32	-2.39	6.27
Fraudulent results	**0**	**1**	**2**

4-2-Assessment Matrices

To evaluate the efficacy of the prediction model, the researchers utilized multiple performance indicators, such as accuracy, precision, and recall. The results of the categorization process were arranged and presented in a confusion matrix, as depicted in Table 2. This matrix displays the allocation of accurate and inaccurate predictions for each category, with the counts indicating the percentage of instances falling into each cell. The confusion matrix uses the labels {P, N} to represent the positive and negative testing data, respectively. The labels {Y, N} indicate the classifier's predictions for the positive and negative classes, as explained in reference [34].

Table 2. Confusion matrix

	Actual Positive (P)	**Actual Negative (N)**
Predicted Positive (Y)	TP (true positives)	FP (false positives)
Predicted Negative (N)	FN (false negatives)	TN (true negatives)

The confusion matrix distinguishes between the number of accurate predictions for positive occurrences, known as "true positive" (TP), and the number of accurate predictions for negative cases, known as "true negative" (TN). Conversely, the number of inaccurate predictions for positive occurrences is referred to as "false positive" (FP), while the number of inaccurate predictions for negative instances is called "false negative" (FN). In order to assess the effectiveness of the prediction model using the datasets, the researchers employed a set of assessment metrics, as specified in Equations 6 to 9.

Accuracy:

$$\frac{TP + TN}{TP + FP + FN + TN} \tag{6}$$

Precision:

$$\frac{TP}{TP + FP} \tag{7}$$

Recall:

$$\frac{TP}{TP + FN} \tag{8}$$

F-measure:

$$\frac{(1 + \beta^2) \times Recall \times Precision}{\beta^2 \times Recall + Precision} \tag{9}$$

The variable β in the equation acts as a coefficient that allows for fine-tuning the trade-off between precision and recall. The F-measure is often provided a value of 1, which ensures that both precision and recall are given equal importance in its calculation. A high F-measure value indicates that the learning method performs well for the target class, as it means that both recall, and precision values are high. Put simply, a high F-measure indicates that the algorithm is successfully recognizing a significant fraction of the true positive cases (high recall) while also keeping the number of false positive predictions low (high precision).

4-3-Experimental Configuration

The studies were performed on a Windows 11 computer system equipped with an NVIDIA GeForce RTX 4070 Ti graphics card boasting 8GB of memory and 32GB of system RAM. Python was the main programming language utilized for development, and the project made use of several machine learning tools. The available models for classification are RandomForestClassifier for random forest models, XGBClassifier for gradient boosting, and lightgbm for efficient gradient boosting. The optuna library can be used for hyperparameter optimization, while the HyperbandPruner is used for pruning less promising hyperparameter configurations. To handle imbalanced datasets, the imbalanced-ensemble library can be utilized. Lastly, the anchor-exp library can be used for generating rule-based explanations of model predictions. The study utilized these libraries to train and optimize the machine learning models.

4-4- Experimental Results

4-4-1- Result of Handling Imbalanced Financial Data Through Optuna with Hyperband Optimization

Equation 10 illustrates the procedure of normalizing input data through the utilization of a standardization approach. Standardization is the process of converting the characteristics of a dataset to conform to a normal distribution with a mean of zero and a standard deviation of one. Z-score normalization, commonly referred to as this method, effectively shifts the data's center to zero and adjusts the scale to produce a variance of one. Standardization is a process that rescales the features so that their values are stated in terms of the number of standard deviations they deviate from the mean of the feature distribution.

$$Z(x) = \frac{x - \bar{x}}{\sigma} \tag{10}$$

In the given context, x represents a specific example of the feature vector, which has n dimensions and belongs to the n-dimensional integer space, represented as Z^n. The average values of the characteristics are denoted by $\bar{x}$, which is a vector of n dimensions and belongs to Z^n. Similarly, the variability of the characteristics is represented by σ, which is a vector in n-dimensional space, belonging to the set Z^n. The notations are employed to depict the process of standardization, wherein the feature values are adjusted to be centered on their respective means ($\bar{x}$) and scaled by their corresponding standard deviations (σ) in order to attain a standardized representation of the data.

To maintain data integrity and completeness, the study sourced financial statements from The Stock Exchange of Thailand's official filings, with additional verification against companies' annual reports. Gaps in the dataset were filled using mean imputation, where null values in each column were replaced with that column's rounded average. This technique ensured a complete dataset, though it's worth noting that such an approach may introduce bias, particularly if the missing data is not randomly distributed across the dataset. This method was chosen for its simplicity and effectiveness in providing a complete dataset for analysis, while acknowledging its potential limitations in preserving the statistical properties of the original data.

This study utilized tree-based machine learning algorithms, along with Optuna and Hyperband, to identify and forecast instances of fraudulent behavior. The initial tree-based model employed was the Random Forest (RF), initially introduced by Breiman [35]. Random Forest (RF) algorithm creates a collection of decision trees. It randomly chooses features for each tree and combines the results of all trees to generate the ultimate forecast of the model. The second machine learning approach used in this research is XGBoost [36], which is based on tree-based algorithms. XGBoost utilizes two fundamental principles in ensemble learning: bagging and boosting. Bagging involves training models simultaneously and creating trees by separate sampling, which enhances the stability and accuracy of the models. On the other hand, boosting constructs trees in a sequential manner, where each tree specifically targets the shortcomings of the preceding one, enabling incremental improvements in the learning process. Ultimately, we utilized LightGBM to forecast instances of fraudulent behavior. LightGBM, developed by Ke et al. [37], is a highly efficient method that utilizes the gradient-boosting architecture. The method utilizes an innovative approach that relies on gradients and samples data instances from one side to filter and produce segmentation values. In addition, LightGBM utilizes distinctive feature bundling to decrease the number of dimensions in the feature space, leading to a very efficient training process.

In order to tackle the problem of class imbalance in the dataset, we assessed and contrasted four distinct sampling techniques: SMOTE, ADASYN, Tomek Links, and Repeated Edited Nearest Neighbors (RENN). Upon applying these strategies to balance the dataset, we proceeded to partition it into a training set, which accounted for 70% of the data, and a test set, which accounted for 30% of the data. The training set was employed for hyperparameter optimization using Optuna and Hyperband, whereas the test set was exclusively used to assess the performance of the ultimate model. Table 3 displays the performance outcomes of various classification algorithms on the test set, utilizing a variety of sampling approaches to tackle class imbalance. When evaluating SMOTE and ADASYN, XGBoost (XGB) demonstrated a minor superiority with SMOTE in terms of overall metrics, specifically in accuracy and precision for class 1. These findings indicate that SMOTE is superior to ADASYN in achieving dataset balance and enhancing the efficacy of XGBoost in identifying fraudulent instances. When comparing Tomek Links and RENN, we find that RENN consistently performs better than Tomek Links in all three algorithms (XGB, RF, and Light GBM). RENN outperforms Tomek Links in terms of accuracy scores for all three algorithms. Furthermore, RENN exhibits superior equilibrium between precision and recall for all three categories, especially for XGBoost. The investigation indicates that utilizing XGBoost with SMOTE for oversampling and RENN for undersampling yields the most optimal overall performance, specifically for class 1 (fraudulent) and class 2 (grey area).

Table 3. The performance of XGBoost, Random Forest, and LightGBM after applying balancing methods

Dataset	#Test data	Accuracy			Precision			Recall			F-measure		
		XGB	RF	Light GBM	XGB	RF	Light GBM	XGB	RF	Light GBM	XGB	RF	Light GBM
Original data													
class 0	115				0.9059	0.9318	0.8935	0.8496	0.6930	0.8061	0.8769	0.7879	0.8476
class 1	89	0.9627	0.9543	0.9610	0.8901	0.9287	0.9110	0.8901	0.8315	0.8901	0.8901	0.8662	0.8905
class 2	2055				0.9688	0.9571	0.9665	0.9722	0.9751	0.9727	0.9705	0.9860	0.9696
SMOTE													
class 0	2051				**0.9651**	0.9076	0.5006	**0.9785**	0.9737	0.9740	**0.9718**	0.9395	0.6470
class 1	2051	**0.9728**	0.9449	0.6461	**0.9737**	0.9608	0.9790	**0.9800**	0.9776	0.9712	**0.9768**	0.9691	0.9751
class 2	2026				**0.9891**	0.9713	0.0000	**0.9598**	0.8828	0.0000	**0.9698**	0.9250	0.0000
ADASYN													
class 0	2052				0.9651	0.9081	0.5006	0.9766	0.9737	0.9790	0.9708	0.9398	0.6470
class 1	2074	0.9717	0.9444	0.6466	0.9719	0.9560	0.9761	0.9838	0.9817	0.9708	0.9759	0.9679	0.9735
class 2	2032				0.9785	0.9746	0.0000	0.9783	0.8786	0.0000	0.9683	0.9242	0.0000
Tomek Links													
class 0	94				0.9455	0.9567	0.9677	0.8736	0.4268	0.8311	0.9082	0.5931	0.8943
class 1	92	0.9683	0.9413	0.9661	0.9335	0.9483	0.9219	0.8713	0.6430	0.8604	0.9013	0.7671	0.8901
class 2	2039				0.9707	0.9408	0.9679	0.9771	0.9785	0.9671	0.9739	0.9593	0.9724
RENN													
class 0	53				**0.9600**	0.9542	0.9565	**0.8957**	0.6026	0.8268	**0.9120**	0.7474	0.9000
class 1	105	**0.9689**	0.9582	0.9678	**0.9690**	0.9635	0.9300	0.8471	0.7419	0.8843	**0.9084**	0.8449	0.8966
class 2	1910				**0.9696**	0.9570	0.9682	**0.9784**	0.9721	0.9679	**0.9740**	0.9684	0.9725

After a thorough evaluation, we determined that integrating the RENN (Repeated Edited Nearest Neighbors) undersampling method with the SMOTE (Synthetic Minority Over-sampling Technique) oversampling approach, along with Optuna and Hyperband, yielded the most promising results. RENN helps to remove noisy and ambiguous instances from the majority class, while SMOTE generates synthetic examples for the minority class, effectively balancing the class distribution. The results in Table 4 highlight the effectiveness of the approach, combining RENN and SMOTE balancing methods with optimally tuned XGBoost hyperparameters, in creating a highly accurate and efficient model for fraud detection across all classes of instances.

Table 4. The Performance of XGBoost with combining RENN and SMOTE balancing methods and the Optuna Hyperband optimal values of dataset

Dataset	#Test data	Accuracy	Precision	Recall	F-measure	Optimal hyperparameter values	Time (minutes)
class 0	1952		0.9974	0.9914	0.9987	'max_depth': 5 'gamma': 0.0266 'colsample_bytree': 0.5773 'subsample': 0.7557 'min_child_weight': 1	2.27
class 1	1918	0.9869	0.9834	0.9978	0.9967		
class 2	1886		0.9985	0.9805	0.9952		

4-4-2- Result of using SHAP-Instance Weighting to Improve the Performance of Fraud Detection

After finding that the combination of RENN and SMOTE balancing methods with Optuna Hyperband optimization creates a highly precise and effective model for detecting fraud across all instance classes, we strive to further improve the model's performance in identifying fraudulent cases based on Algorithm 1. To achieve this, we incorporate relevant instance weighting techniques into the process. We suggest using SHAP values to measure the importance of each instance based on its feature contributions. By developing a weighting scheme that gives higher weights to more informative or influential instances in the model's decision-making process, we can prioritize learning from the most relevant examples. These SHAP-based instance weights are integrated into the training process of the XGBoost fraud detection model, allowing it to focus on the most informative instances. The integration of SHAP-based instance weighting into the training process, along with the emphasis on key features specific to each fraud category, presents a novel approach to enhancing fraud detection performance. By utilizing the insights provided by SHAP values and prioritizing learning from the most informative instances, we aim to develop an even more accurate and efficient model for detecting fraudulent activities across different categories.

Table 5 demonstrates the performance of the XGBoost model with SHAP-based instance weights for fraud detection. The results showcase the effectiveness of this approach in improving the model's ability to identify and classify fraudulent instances accurately.

Table 5. The performance of the XGBoost model with SHAP-based instance weights for fraud detection

Dataset	#Test data	Accuracy	Precision	Recall	F-measure	Optimal hyperparameter values	Time (minutes)
class 0	1952		0.9984	1.0000	0.9992	'max_depth': 8 'gamma': 0.0494 'colsample_bytree': 0.3718 'subsample': 0.9212 'min_child_weight': 2 'learning_rate': 0.0953	
class 1	1918	0.9977	0.9949	1.0000	0.9974		2.30
class 2	1886		1.0000	0.9931	0.9965		

Figure 2 displays the ROC curve of each class. The curves for all classes are positioned extremely close to the top-left corner of the plot, indicating near-perfect classification ability. The Area Under the Curve (AUC) scores further quantify this outstanding performance, with class 0 achieving a perfect score of 1.000000, class 1 reaching 0.999994, and class 2 attaining 0.999992. These remarkably high and consistent AUC values across all classes suggest that the model maintains an excellent balance between sensitivity and specificity, regardless of the classification threshold. This implies that the model can effectively distinguish between non-fraudulent, fraudulent, and grey zone cases with an extremely low rate of misclassification.

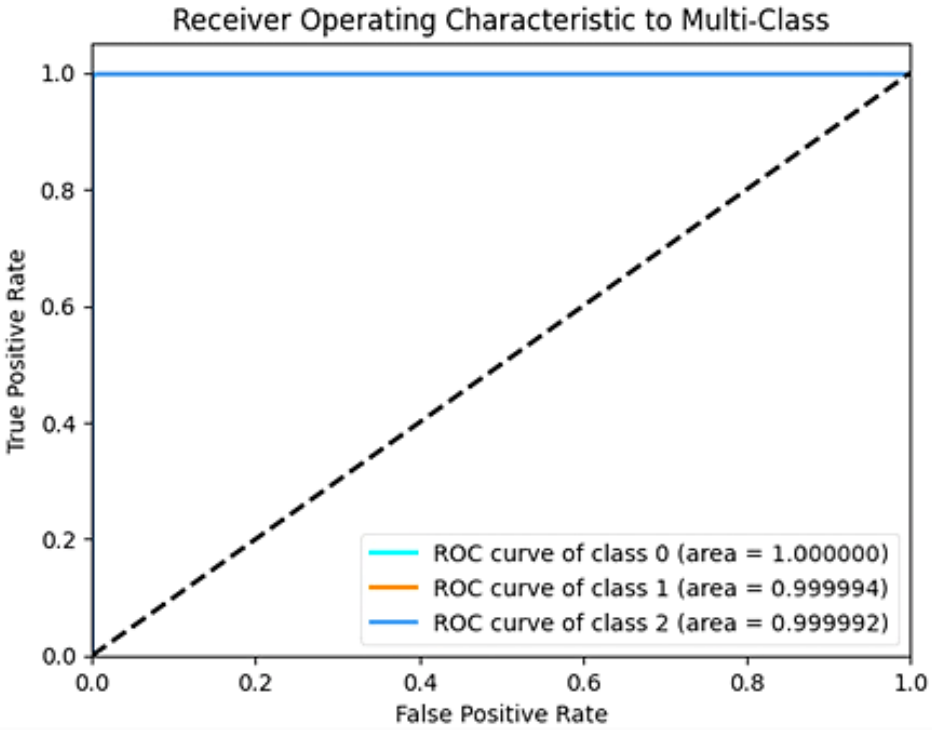

Figure 2. The ROC curve and AUC score of each class

4-4-3- Result of an Anchor Explanation for Fraud Detection and Feature Important of Each Class

For fraudulent prediction, it is essential to comprehend the reasons behind a specific prediction. To address this, our study employs Anchor XAI (Algorithm 2), a technique that generates human-interpretable IF-Then rules to explain individual predictions. By establishing a threshold of explanation certainty at 0.95 in the context of fraud prediction, we can utilize precision and coverage metrics to evaluate the quality and relevance of the identified rules in distinguishing between non-fraudulent, fraudulent, and grey area instances. Table 6 showcases the top 10 example IF-Then rule results derived from the Anchor XAI approach. These rules provide clear and concise explanations for the model's predictions, enabling users to grasp the key factors that contribute to the classification of an instance as non-fraudulent, fraudulent, or grey area. Moreover, the application of SHAP values, as outlined in Algorithm 3, identifies the key features influencing each prediction by computing average SHAP values. This approach reveals how different factors impact the

classification for each class. Such insights are crucial for identifying the financial and operational indicators that suggest a company's likelihood of engaging in fraudulent activities. The top 5 significant features for each fraud category are displayed in Table 7, providing a clear view of the specific attributes the model uses to form its predictions.

Table 6. Anchor's top 10 IF-Then rules for each fraudulent class

Class	If-Then rules	Precision
class 0	If Currents assets/Total asset > 0.46 and Gross profit/Total asset > 0.11 and Net Profit > 565252.55 and Sales / Total Assets > 0.37 and Total assets ≤ 2881404.19 Then **Non-fraudulent**.	1.0000
	If Accounts receivable/Last year's Accounts receivable ≤ 1.51 and Gross profit/Total asset > 0.11 and Net Profit > 565252.55 and Total assets ≤ 2881404.19 and Working capital/total assets > 0.35 Then **Non-fraudulent**.	1.0000
	If Gross profit/Total asset > 0.11 and Inventory/Primary business income ≤ 0.18 and Inventory/total assets > 0.27 and Sales / Total Assets > 0.88 Then **Non-fraudulent**.	0.9992
	If Currents assets/Total asset > 0.70 and Gross profit/Total asset > 0.11 and Net Profit > 565252.55 and Primary business profit/Last year's Primary business profit ≤ -0.12 and Total debt/Total asset ≤ 0.30 Then **Non-fraudulent**.	0.9965
	If Gross profit/Total asset > 0.11 and Inventory/Primary business income ≤ 0.10 and Inventory/total assets > 0.12 and Sales / Total Assets > 0.88 Then **Non-fraudulent**.	0.9944
	If Accounts receivable > 931559.04 and Currents assets/Total asset > 0.70 and Gross profit/Total asset > 0.11 and Sales / Total Assets > 0.37 and Total debt/Total asset ≤ 0.30 Then **Non-fraudulent**.	0.9936
	If Net Profit > 565252.55 and Sales / Total Assets > 1.78 and Working capital/total assets > 0.35 Then **Non-fraudulent**.	0.9919
	If Gross profit/Total asset > 0.11 and Inventory/Current liabilities > 0.86 and Sales / Total Assets > 0.88 and net profit/Primary business income ≤ 0.08 Then **Non-fraudulent**.	0.9911
	If Accounts receivable > 931559.04 and Currents assets/Total asset > 0.70 and Retained Earnings / Total Assets > 0.30 and Sales / Total Assets > 0.88 Then **Non-fraudulent**.	0.9911
	If Gross profit/Total asset > 0.11 and Inventory/Current liabilities > 0.86 and Inventory/Primary business income ≤ 0.10 and Sales / Total Assets > 0.88 Then **Non-fraudulent**.	0.9904
class 1	If Retained Earnings / Total Assets ≤ 0.11 and Total assets ≤ 1133373.03 and Total debt/Total asset > 0.74 Then **Fraudulent**.	0.9966
	If Primary business income/Total asset ≤ -1.88 and Retained Earnings / Total Assets ≤ -0.59 and Total debt/total equity > 1.99 Then **Fraudulent**.	0.9964
	If Current assets/Current liabilities ≤ 0.66 and Retained Earnings / Total Assets ≤ 0.11 and Total asset/Capital and reserve ≤ 1.28 Then **Fraudulent**.	0.9946
	If Retained Earnings / Total Assets ≤ -0.59 and Working capital/total assets ≤ -0.18 Then **Fraudulent**.	0.9932
	If Current assets/Current liabilities ≤ 1.46 and Inventory/Current liabilities ≤ 0.10 and Retained Earnings / Total Assets ≤ -0.59 and Total assets ≤ 1133373.03 Then **Fraudulent**.	0.9931
	If Gross profit/Primary business income ≤ -0.10 and Retained Earnings / Total Assets ≤ -0.59 and Total asset/Capital and reserve > 2.85 Then **Fraudulent**.	0.9929
	If Retained Earnings / Total Assets ≤ 0.11 and Total debt/Total asset > 0.74 and Total liabilities ≤ 1307195.55 Then **Fraudulent**.	0.9927
	If Primary business income/Total asset.1 ≤ -1.88 and Retained Earnings / Total Assets ≤ -0.59 and Total debt/total equity > 1.99 Then **Fraudulent**.	0.9923
	If Current assets/Current liabilities ≤ 0.66 and Gross margin ≤ -63281.56 and Working capital/total assets ≤ -0.18 Then **Fraudulent**.	0.9911
	If Current assets/Current liabilities ≤ 1.46 and Retained Earnings / Total Assets ≤ -0.59 and Total asset/Last year's Total asset ≤ 0.94 and Total assets ≤ 1133373.03 Then **Fraudulent**.	0.9911
class 2	If Retained Earnings / Total Assets > 0.11 and Sales / Total Assets ≤ 0.37 and Total debt/Total asset > 0.51 Then **Grey area**.	1.0000
	If Retained Earnings / Total Assets > 0.11 and Sales / Total Assets ≤ 0.88 and Total asset/Capital and reserve > 1.76 Then **Grey area**.	1.0000
	If Earnings Before Interest and Taxes / Total Assets ≤ 0.05 and Retained Earnings / Total Assets > 0.11 and Sales / Total Assets ≤ 0.88 Then **Grey area**.	1.0000
	If Current assets/Current liabilities ≤ 1.46 and Retained Earnings / Total Assets > 0.11 and Sales / Total Assets ≤ 0.88 Then **Grey area**.	1.0000
	If Primary business income/Total asset > 17.74 and Retained Earnings / Total Assets > 0.11 and Sales / Total Assets ≤ 0.88 Then **Grey area**.	1.0000
	If Retained Earnings / Total Assets > 0.11 and Sales / Total Assets ≤ 0.88 and Working capital/total assets ≤ 0.09 Then **Grey area**.	1.0000
	If Primary business income/Total asset ≤ -1.88 and Retained Earnings / Total Assets > 0.11 and Sales / Total Assets ≤ 0.88 Then **Grey area**.	1.0000
	If Retained Earnings / Total Assets > 0.11 and Sales / Total Assets ≤ 0.37 and Working capital/total assets ≤ 0.09 Then **Grey area**.	1.0000
	If Primary business income/Total asset.1 ≤ -1.88 and Retained Earnings / Total Assets > 0.11 and Sales / Total Assets ≤ 0.88 Then **Grey area**.	1.0000
	If Long term debt/Total asset > 0.18 and Primary business income/Total asset > 17.74 and Retained Earnings / Total Assets > 0.11 and Sales / Total Assets ≤ 1.78 Then **Grey area**.	1.0000

Table 7. Top 5 important features of each fraudulent class by SHAP values

Class	Important features
class 0	1. Sales / Total Assets 2. Retained Earnings / Total Assets 3. net profit/Total asset 4. Gross profit/Total asset 5. Working capital/total assets
class 1	1. Retained Earnings / Total Assets 2. Working capital/total assets 3. Gross profit/Primary business income 4. Book Value of Equity / Book Value of Total Debt 5. Total debt/Total asset
class 2	1. Sales / Total Assets 2. Retained Earnings / Total Assets 3. Working capital/total assets 4. net profit/Total asset 5. Gross profit/Total asset

5- Discussion

The following section presents our proposed approaches, addressing the research questions and offering an analysis of fraud prediction methods:

5-1- RQ1: How Can We Enhance the Performance of a Model for Fraud Detection by Relevant Instance Weighting?

The proposed SHAP-based instance weighting method in Algorithm 1 and the result in Table 5 demonstrate superior performance across all metrics, particularly in achieving perfect recall (1.0000) for fraudulent instances. This is crucial in fraud detection scenarios where minimizing false negatives (undetected frauds) is often prioritized. The success of the SHAP-based instance weighting can be attributed to its ability to leverage the importance of individual instances in the training process. By assigning weights based on SHAP values, the model effectively focuses on the most influential samples for distinguishing between fraudulent and non-fraudulent cases. This approach goes beyond traditional sampling techniques by not just addressing class imbalance, but also emphasizing the most informative instances within each class. The improvement in precision (from 0.9834 to 0.9949 for class 1) and accuracy (from 0.9869 to 0.9977) from Table 4 to Table 5 suggests that the SHAP-based weighting not only maintains the model's ability to correctly identify fraudulent instances but also reduces false positives. This balance is crucial for practical applications where both missed frauds and false alarms can have significant consequences. Moreover, the perfect recall score achieved by the SHAP-based method indicates its robustness in capturing all fraudulent instances in the test set. This is a significant advancement, especially considering the challenges typically associated with detecting financial fraud due to its evolving nature and the sophisticated techniques employed by fraudsters.

The proposed SHAP-based instance weighting method, as presented in Table 5, offers a nuanced approach to fraud detection by distinguishing between three classes: non-fraudulent (class 0), fraudulent (class 1), and grey area (class 2). The implications and benefits of this approach for each class are detailed as follows:

1) Non-fraudulent (class 0): The model achieves a high precision of 0.9984 for this class, indicating a strong ability to correctly identify legitimate transactions or companies. This is crucial for maintaining trust in the financial system and avoiding unnecessary investigations or restrictions on legitimate businesses. By accurately distinguishing normal financial activities, the model reduces the number of false positives, which are instances incorrectly flagged as suspicious. This efficiency is vital for resource allocation in financial institutions, ensuring that investigative efforts are focused on truly suspicious cases.

2) Fraudulent (class 1): The recall score of 1.000 for fraudulent instances highlights the model's capability to identify all fraudulent cases in the test set. This is critical for fraud detection as it ensures that no fraudulent activity goes unnoticed. With a precision of 0.9949, the model confirms that when it identifies an instance as fraudulent, it is almost always correct. This high precision minimizes the risk of false alarms, which are costly and disruptive. The balance between recall and precision ensures that the model is not only comprehensive in detecting all frauds but also accurate in its predictions, reducing both false negatives (missed frauds) and false positives (innocent entities being flagged).

3) Grey area (class 2): The introduction of a grey area class acknowledges the complexity of financial transactions that may exhibit some, but not all, characteristics of fraud. This nuanced classification allows for a more refined analysis

of potentially risky activities. By categorizing these borderline cases, the model can better inform risk management and compliance strategies, ensuring that entities or transactions that fall into this category receive appropriate scrutiny without being prematurely classified as fraudulent. The recall (0.9931) and F-measure (0.9965) for this class in Table 5 suggest it effectively identifies and manages these ambiguous cases, thereby enhancing the robustness of the fraud detection system.

The study presents a comparison of existing fraudulent detection methods in terms of their accuracy. Table 8 provides a comprehensive comparison, including information on the datasets used, the number of testing samples in each dataset, and the fraudulent detection techniques employed. Each method utilizes a different approach to accurately predict fraud from financial data.

Table 8. A comparison of current methods based on the accuracy of fraudulent detection

Method	Technique	Accuracy	Dataset	Number of testing data (Non-Fraud: Fraud: Grey zone)	Training time (Minutes)
Zhao and Bai (2022) [14]	LR+XGBOOST	0.98532	The listed companies from the stock market in Shanghai and Shenzhen and the Hong Kong stock market.	5354:5354	1.54
Ali et al. (2023) [15]	Optimized XGBoost	0.9605	Osiris database	The original data had a ratio of 1798:102, but this could not be identified after using SMOTE.	Not specified
Cheah et al. (2023) [16]	GANified-SMOTE	0.8900 (F1-score)	Credit card fraud dataset from Kaggle	The original data had a ratio of 56864:98, but this could not be identified after using GANified-SMOTE.	Not specified
Proposed method	SHAP-Instance Weighted with RENN and SMOTE through XGBoost	**0.9977**	The listed companies on the Stock Exchange of Thailand	1952:1918:1886	2.30

Table 8 provides a comparative analysis of various methods for detecting fraudulent activities, focusing on the accuracy of each approach, the datasets used, the number of testing data instances, and the training times. The proposed method, which utilizes SHAP-based instance weighting combined with RENN and SMOTE through XGBoost, achieves a remarkable accuracy of 0.9977. This accuracy surpasses that of Zhao & Bai (2022) [14] who used LR+XGBOOST with an accuracy of 0.98532, Ali et al. (2023) [15] with Optimized XGBoost at 0.9605, and Cheah et al. (2023) [16] who reported an F1-score of 0.8900 using GANified-SMOTE. The dataset for the proposed method comprises listed companies on the Stock Exchange of Thailand, with a balanced distribution of 1952 non-fraud, 1918 fraud, and 1886 grey zone instances. This diverse dataset ensures a robust evaluation of the model's performance. In contrast, the datasets used by other methods vary, including the stock markets in Shanghai and Shenzhen, the Osiris database, and a credit card fraud dataset from Kaggle. The number of testing data instances for the proposed method also provides a balanced evaluation, whereas other methods either did not specify their testing data ratios post-SMOTE application or had less balanced datasets. The training time for the proposed method is 2.30 minutes, which is reasonable considering the high accuracy achieved. Zhao and Bai reported a slightly faster training time of 1.54 minutes, but with lower accuracy.

The synergistic combination of XGBoost, renowned for its proficiency in processing structured data [38], with Optuna Hyperband for fine-tuning parameters [30, 39], and SHAP-based instance weighting creates a powerful approach. This integrated method effectively tackles both model optimization and the challenge of determining instance significance concurrently. In the realm of fraud detection, where datasets are often characterized by significant class imbalances, this approach proves particularly valuable. Although XGBoost already demonstrates a strong capability in managing imbalanced data [38], the incorporation of SHAP-based instance weighting further enhances this ability. This strategy is in line with research by Antwarg et al. (2021) [40], which demonstrated that weighting instances could lead to substantial improvements in classification accuracy when dealing with imbalanced datasets.

5-2- RQ2: How Can We Enhance the Interpretability of a Model for Fraud Detection Tasks?

This study utilizes Anchor XAI principles to provide insight into the rationale behind prediction results, clarifying the factors that lead to classifications of non-fraudulent, fraudulent, or grey area cases through the use of if-then rules. As a result, users of financial statements can leverage both the prediction outcomes and the underlying reasons for each classification when making decisions about investments, loans, or other business transactions with a company. This approach enhances the practical application of the model's findings in real-world financial contexts.

Table 6 provides an illustration of an if-then rule used to classify instances as non-fraudulent, fraudulent, or grey area. The following is a representative example from this table:

1) A non-fraudulent (class 0)

If the result is classified as class 0, financial statement users can proceed with standard due diligence when considering transactions with that company. This classification suggests that the company's financial activities appear legitimate and align with normal business operations. An example of an if-then rule for this class might be as follows:

- Rule: "If Currents assets/Total asset > 0.46 and Gross profit/Total asset > 0.11 and Net Profit > 565252.55 and Sales / Total Assets > 0.37 and Total assets ≤ 2881404.19 then non-fraudulent." indicates a combination of financial ratios that suggest a healthy, efficiently operating company. The high current assets to total assets ratio (>0.46) implies strong liquidity, indicating the company can easily meet its short-term obligations. A gross profit to total assets ratio exceeding 0.11 suggests efficient use of assets to generate profit. The substantial net profit (>565252.55) indicates the company is not only generating sales but also managing its expenses effectively to maintain profitability. The sales to total assets ratio (>0.37) demonstrates good asset utilization in generating revenue, a sign of operational efficiency. The cap on total assets (≤2881404.19) focusing on smaller enterprises is significant, as it aligns with findings by Ge & McVay [41] that while smaller firms may have less sophisticated internal controls, they are also subject to different market pressures and scrutiny compared to larger firms.
- Rule: "If Accounts receivable/Last year's Accounts receivable ≤ 1.51 and Gross profit/Total asset > 0.11 and Net Profit > 565252.55 and Total assets ≤ 2881404.19 and Working capital/total assets > 0.35 then non-fraudulent." suggests a combination of financial indicators that point to a healthy, well-managed company. The accounts receivable ratio (≤1.51) indicates stable or improving collection practices, as it shows the company isn't experiencing an unusual increase in uncollected sales. A gross profit to total assets ratio exceeding 0.11 demonstrates efficient use of assets in generating profit. The substantial net profit (>565252.55) suggests the company is not only generating sales but also managing expenses effectively to maintain profitability. The cap on total assets (≤2881404.19) focusing on smaller enterprises is significant, as it relates to findings by Ge and McVay [41] that while smaller firms may have less sophisticated internal controls, they are also subject to different market pressures and scrutiny compared to larger firms. The high working capital to total assets ratio (>0.35) is consistent with Beaver's (1966) [42] research on the importance of liquidity ratios in predicting firm success, Collectively, these ratios paint a picture of a company with consistent sales practices, strong profitability, good liquidity, and efficient operations – all hallmarks of a legitimate, well-managed business.
- Rule: "If Gross profit/Total asset > 0.11 and Inventory/Primary business income ≤ 0.18 and Inventory/total assets > 0.27 and Sales / Total Assets > 0.88 then non-fraudulent." suggests a combination of financial indicators that point to a healthy, efficiently operating company. The gross profit to total assets ratio exceeding 0.11 indicates effective use of assets in generating profit, suggesting the company has a strong profit margin relative to its asset base. The inventory ratios (Inventory/Primary business income ≤ 0.18 and Inventory/total assets > 0.27) are consistent with Beneish's (1999) M-score model [43], which identifies unusual changes in inventory as potential red flags for fraudulent activity. The balance displayed here suggests efficient inventory management without signs of manipulation. The high sales to total assets ratio (>0.88) demonstrates excellent asset utilization in generating revenue, indicating operational efficiency. This ratio also aligns with Ou & Penman's (1989) [44] findings on financial statement analysis for equity valuation, suggesting a well-managed, non-fraudulent entity. Collectively, these ratios paint a picture of a company with strong profitability, efficient inventory management, and high operational efficiency – all characteristics of a legitimate, well-managed business. The combination of these positive financial indicators, particularly the balance between inventory levels and sales performance, makes it unlikely that the company is engaging in fraudulent activities, as fraudulent companies often struggle to maintain consistent, positive performance across multiple financial metrics simultaneously.

2) Fraudulent (class 1)

If the result is classified as class 1 (fraudulent), financial statement users should exercise extreme caution and conduct extensive due diligence before considering any transactions with that company. This classification suggests that the company's financial activities exhibit characteristics associated with fraudulent practices and do not align with normal, legitimate business operations. An example of an if-then rule for this fraudulent class might be as follows:

- Rule: "If Retained Earnings / Total Assets ≤ 0.11 and Total assets ≤ 1133373.03 and Total debt/Total asset > 0.74 then fraudulent." highlights potential indicators of fraudulent activity in a company's financial structure. A retained earnings to total assets ratio of 0.11 or less is alarming as it suggests the company has amassed minimal profits in relation to its assets. Additionally, a debt-to-asset ratio exceeding 0.74 indicates an unusually high level of leverage, with debt financing over 74% of the company's assets. This elevated leverage may incentivize financial statement manipulation to satisfy debt covenants or preserve credit access. In cases of fraud, firms might inflate assets or underreport liabilities to artificially improve this ratio. These financial characteristics align with Skousen et al. (2009) [45] and Cressey's (1953) [46] Fraud Triangle theory and its later expansion into the Fraud Diamond by Wolfe & Hermanson (2004) [47], which propose that fraud emerges from the convergence of pressure, opportunity, rationalization, and capability. The combination of low retained earnings and high leverage signifies financial distress, creating substantial pressure – a key component of these fraud theories. Furthermore, when viewed through the lens of Kahneman & Tversky's (1979) [48] prospect theory, as applied to corporate fraud by Abdel-Khalik (2014) [49], the potential losses suggested by low retained earnings and high leverage could lead to increased risk-taking behavior, potentially including fraudulent activities.
- Rule: "If Primary business income/Total asset ≤ -1.88 and Retained Earnings / Total Assets ≤ -0.59 and Total debt/total equity > 1.99 then fraudulent." outlines a set of financial indicators that strongly suggest fraudulent

activity. A primary business income to total asset ratio of -1.88 or lower indicates the company is incurring substantial losses relative to its assets, a situation that is untenable in normal business operations. Similarly, retained earnings to total assets ratio of -0.59 or lower shows that accumulated losses surpass half of the company's total assets. The third criterion, a total debt to total equity ratio exceeding 1.99, reveals an alarmingly high reliance on debt financing, particularly troubling when considered alongside the other negative financial indicators. This combination of ratios aligns with research by Dechow et al. [50], who found that aggressive earnings management can progress into fraudulent financial reporting. The extremely negative income and retained earnings ratios observed in this rule suggest a level of earnings manipulation that likely crosses the threshold into fraudulent territory.

- Rule: "If Current assets/Current liabilities ≤ 0.66 and Retained Earnings / Total Assets ≤ 0.11 and Total asset/Capital and reserve ≤ 1.28 then fraudulent." presents a combination of financial ratios that together strongly indicate potential fraudulent activity within a company. A current ratio of 0.66 or less points to severe liquidity issues, suggesting the company may struggle with short-term obligations, possibly due to manipulation of current assets or liabilities to hide financial distress. Persons (1995) [51] found that companies with lower liquidity are more prone to fraudulent financial reporting to conceal their true financial state. The low Retained Earnings to Total Assets ratio (≤0.11) implies minimal accumulated profit relative to company size, potentially indicating persistent unprofitability or inappropriate dividend payouts despite poor performance, which could signal earnings manipulation or resource misappropriation. Paradoxically, the Total asset to Capital and reserve ratio (≤1.28) suggests low leverage and a strong equity position, contradicting the other ratios. This inconsistency might indicate asset overvaluation, liability underreporting, or other balance sheet manipulations intended to present a falsely robust financial position. These discrepancies align with the fraud triangle theory proposed by Cressey (1953) [46] and later expanded into the fraud diamond by Wolfe & Hermanson (2004) [47], with poor liquidity and profitability ratios demonstrating pressure, and the apparent equity structure manipulation suggesting both opportunity and capability. This unusual combination of financial characteristics, presenting an inconsistent picture of poor liquidity and profitability alongside a seemingly strong equity structure, suggests potential fraudulent financial engineering or complex schemes to mislead stakeholders about the company's true financial health, warranting further investigation in any fraud detection system.

3) Grey area (class 2)

If the result is classified as class 2 (grey area), financial statement users should exercise increased caution and conduct thorough due diligence before considering any transactions with that company. This classification suggests that the company's financial activities exhibit characteristics that are neither clearly fraudulent nor definitively legitimate, falling into an area of uncertainty. The grey zone classification indicates potential risks or anomalies that warrant closer examination and may not fully align with normal business operations. An example of an if-then rule for this grey zone class might be as follows:

- Rule: "If Retained Earnings / Total Assets > 0.11 and Sales / Total Assets ≤ 0.37 and Total debt/Total asset > 0.51 then grey area." indicates a combination of financial ratios that suggest a company's situation is neither clearly fraudulent nor definitively legitimate, hence falling into a grey zone. The Retained Earnings to Total Assets ratio being greater than 0.11 suggests the company has accumulated some profits over time, which is generally a positive sign. However, the Sales to Total Assets ratio being less than or equal to 0.37 indicates relatively low asset turnover, potentially signaling inefficient use of assets or declining sales. The Total debt to Total asset ratio exceeding 0.51 shows that more than half of the company's assets are financed by debt, which could be concerning depending on the industry and company's stage. This high leverage, combined with low asset turnover, creates a mixed picture. While the company has retained earnings, its operational efficiency appears questionable, and its high debt levels could pose risks. The high debt ratio, in particular, aligns with Kanapickienė & Grundienė's (2015) [52] findings on the importance of leverage in fraud risk assessment, but doesn't conclusively indicate fraudulent activity. This combination of factors doesn't clearly indicate fraud, but it does raise questions about the company's financial health and management, warranting further investigation and caution from financial statement users.

- Rule: "If Retained Earnings / Total Assets > 0.11 and Sales / Total Assets ≤ 0.88 and Total asset/Capital and reserve > 1.76 then grey area." indicates a combination of financial ratios that place a company in an ambiguous financial position, hence the grey zone classification. The Retained Earnings to Total Assets ratio being greater than 0.11 suggests the company has accumulated some profits over time, which is generally a positive sign of financial health. However, the Sales to Total Assets ratio being less than or equal to 0.88, while not critically low, indicates that the company's asset turnover could be improved, potentially signaling some inefficiency in utilizing assets to generate sales. The Total asset to Capital and reserve ratio exceeding 1.76 implies that the company is significantly leveraged, with assets substantially exceeding the shareholders' equity. This high leverage, combined with the moderate asset turnover, creates an ambiguous financial position. While the company shows some profitability through retained earnings, its operational efficiency and capital structure raise questions. This combination of factors doesn't clearly indicate fraud or definitive financial health, but rather suggests a complex

financial situation that warrants closer examination. The grey zone classification acknowledges this ambiguity, indicating that financial statement users should approach this company's financials with caution and conduct thorough due diligence before making any decisions [21].

- Rule: "If Earnings Before Interest and Taxes / Total Assets ≤ 0.05 and Retained Earnings / Total Assets > 0.11 and Sales / Total Assets ≤ 0.88 then grey area." highlights a combination of financial ratios that places a company in an ambiguous financial position, hence the "grey zone" classification. The first ratio, EBIT to Total Assets, measures a company's operating profitability relative to its total assets, and a ratio less than or equal to 0.05 suggests that the company is generating a low level of operating profits compared to its asset base, potentially indicating inefficiencies in operations, high operating costs, or a combination of both. The second ratio, Retained Earnings to Total Assets, provides insight into a company's cumulative profitability over time, and a ratio greater than 0.11 implies that the company has been profitable in the past and has chosen to reinvest a portion of those profits back into the business, which is generally a positive sign. However, the third ratio, Sales to Total Assets, also known as Asset Turnover, measures how efficiently a company uses its assets to generate sales, and a ratio less than or equal to 0.88 indicates that the company is generating a relatively low level of sales compared to its total assets, suggesting that the company may not be utilizing its assets effectively to generate revenue. When these three ratios are considered together, they paint a mixed picture of the company's financial health, placing it in a "grey area" that suggests the company may be experiencing operational challenges or inefficiencies despite its historical profitability, and further investigation would be warranted to understand the underlying reasons for the company's current financial performance. The "grey area" classification serves as a warning sign to financial statement users, indicating that they should approach the company's financials with caution and conduct thorough due diligence before making any investment or lending decisions [43].

By examining the if-then rules for individual instances, we've identified the key features for each class by SHAP values in Algorithm 3, as presented in Table 7. The significant characteristics for each category are outlined below:

1) A non-fraudulent (class 0)

The top 5 important features for identifying non-fraudulent companies (class 0) are Sales/Total Assets, Retained Earnings/Total Assets, net profit/Total asset, Gross profit/Total asset, and Working capital/total assets. These ratios are crucial indicators of a company's financial health, efficiency, and stability, which are typically strong in legitimate, well-managed businesses.

Sales/Total Assets, also known as the asset turnover ratio, measures how efficiently a company uses its assets to generate sales. A high ratio indicates effective asset utilization, which is characteristic of well-run, non-fraudulent companies. Retained Earnings/Total Assets reflects a company's ability to accumulate profits over time, suggesting sustainable operations. The net profit/Total asset and Gross profit/Total asset ratios are key profitability metrics, indicating a company's ability to generate earnings from its assets. Strong, consistent profitability is often associated with legitimate business operations. Lastly, working capital/total assets is a measure of liquidity and operational efficiency, with a healthy ratio suggesting the company can meet its short-term obligations and operate smoothly.

These ratios align with several established theories and empirical findings in accounting and finance literature. The importance of Sales/Total Assets is supported by Ou and Penman (1989) [44], who found it to be a significant predictor of future earnings. Altman (1968) [20] included Retained Earnings/Total Assets in his Z-score model for predicting bankruptcy, highlighting its importance in assessing financial health. The profitability ratios (net profit/Total asset and Gross profit/Total asset) are consistent with Dechow et al.'s (2011) [53] predictors of financial statement reliability. Beneish (1999) [43] also emphasized the importance of gross margin in his M-score model for detecting earnings manipulation. The Working capital/total assets ratio aligns with Beaver's (1966) [42] findings on the importance of liquidity ratios in predicting firm success. Furthermore, Kaminski et al. (2004) [54] found that these types of ratios, when considered collectively, can be effective in distinguishing between fraudulent and non-fraudulent companies. The combination of these ratios also addresses multiple aspects of the fraud triangle theory proposed by Cressey (1953) [46], particularly by indicating an absence of financial pressure that often motivates fraudulent behavior.

2) Fraudulent (class 1)

The top 5 important features for identifying fraudulent companies (class 1) are Retained Earnings/Total Assets, Working capital/total assets, Gross profit/Primary business income, Book Value of Equity/Book Value of Total Debt, and Total debt/Total asset. These ratios are crucial indicators that, when showing unusual or extreme values, can signal potential fraudulent activities or financial statement manipulation.

Retained Earnings/Total Assets reflects a company's cumulative profitability over time. In fraudulent cases, this ratio might be unusually low or negative, indicating persistent losses or aggressive dividend policies that deplete retained earnings. Working capital/total assets is a measure of liquidity; extreme values could suggest manipulation of current assets or liabilities. Gross profit/Primary business income can reveal unusual profit margins that might indicate revenue or cost manipulation. The Book Value of Equity/Book Value of Total Debt ratio provides insights into a company's

capital structure; significant deviations could suggest hidden liabilities or overstated equity. Lastly, Total debt/Total asset ratio indicates the extent of leverage; extremely high leverage could create pressures that motivate fraudulent behavior.

The importance of Retained Earnings/Total Assets is supported by Beneish (1999) [43], who included a similar ratio in his M-score model for detecting earnings manipulation. Dechow et al. (2011) [53] found that extreme values in working capital accruals, related to the Working capital/total assets ratio, were associated with a higher likelihood of material misstatements. The Gross profit/Primary business income ratio aligns with findings by Summers and Sweeney (1998) [55], who noted that gross margin index is a significant predictor of fraud. The capital structure ratios (Book Value of Equity/Book Value of Total Debt and Total debt/Total asset) are consistent with the findings of Persons (1995) [51], who identified leverage as a key factor associated with fraudulent financial reporting. These ratios also relate to the pressure component of the fraud triangle theory proposed by Cressey (1953) [46] and expanded by Wolfe and Hermanson (2004) [47] into the fraud diamond.

3) Grey area (class 2)

The top 5 important features for identifying companies in the grey zone (class 2) are Sales/Total Assets, Retained Earnings/Total Assets, Working capital/total assets, net profit/Total asset, and Gross profit/Total asset. These ratios are crucial indicators that, when showing ambiguous or moderate values, can signal a company's financial position that is neither clearly fraudulent nor definitively healthy, hence falling into a grey area.

The Sales/Total Assets and Retained Earnings/Total Assets ratios, prominent in our grey zone features, are key components of Altman's Z-score model (1968) [20]. However, Grice & Ingram (2001) [56] found that the effectiveness of these ratios in predicting financial distress varies over time and across industries, supporting the need for a grey zone classification that acknowledges this variability. The Working capital/total assets ratio's importance in the grey zone is supported by Sharma & Iselin (2003) [57], who found that liquidity ratios are significant in predicting financial distress, but their predictive power is not absolute. This aligns with the grey area concept where liquidity might be concerning but not clearly indicative of fraud or imminent failure. Profitability ratios (net profit/Total asset and Gross profit/Total asset) in the grey zone context are particularly interesting. Burgstahler & Dichev (1997) [58] provided evidence of earnings management to avoid earnings decreases and losses, suggesting that companies in the grey zone might engage in such practices without crossing into clear fraudulent territory.

5-3- Implementable Framework for Detecting Financial Fraud

In this study, we present a deployment model for detecting fraudulent activities, focusing on a company recently scrutinized by the Securities and Exchange Commission of Thailand for alleged misconduct in its 2022 financial reporting [59]. Our model leverages financial indicator data spanning a ten-year period (2013-2022) from the mentioned company. The outcomes of our predictive analysis are presented in Table 9.

Table 9. The anchor if-then rule fraud detection results from deployment model

Year	If-Then rules	Precision
2013	If Sales/Total Assets ≤ 0.88 AND Primary business income/Total asset > 5.87 AND Deposit and Cash/Current assets ≤ 0.06 AND Capital and reserve/Total debt > 0.21 Then **Grey area**.	0.9949
2014	If Sales/Total Assets ≤ 0.88 AND Primary business income/Total asset > 17.74 AND Retained Earnings / Total Assets > -0.59 Then **Grey area**.	0.9899
2015	If Sales/Total Assets ≤ 0.88 AND Working capital/total assets > 0.09 AND Cash/Total asset ≤ 0.02 Then **Grey area**.	0.9903
2016	If Sales/Total Assets ≤ 0.37 AND Working capital/total assets > 0.09 AND Cash/Total asset ≤ 0.02 Then **Grey area**.	0.9825
2017	If Sales / Total Assets ≤ 0.88 AND Working capital/total assets > 0.09 AND Cash/Total asset ≤ 0.02 Then **Grey area**.	0.9832
2018	If Retained Earnings / Total Assets ≤ 0.11 AND net profit/Total asset ≤ -0.07 Then **Fraudulent**.	0.9526
2019	If Sales/Total Assets ≤ 1.78 AND Net profit/Gross profit ≤ 0.76 AND Accounts receivable > 931559.04 Then **Grey area**.	0.9512
2020	If Sales/Total Assets ≤ 0.88 AND Primary business income/Total asset.1 > 5.87 AND Total asset/Capital and reserve > 1.76 AND Retained Earnings / Total Assets > 0.11 Then **Grey area**.	1.0000
2021	If Sales/Total Assets ≤ 0.88 AND Primary business income/Total asset > 5.87 AND Primary business profit/Last year's Primary business profit > 1.27 AND Accounts receivable/Total asset > 0.19 Then **Grey area**.	1.0000
2022	If Retained Earnings/Total Assets ≤ 0.11 AND Working capital/total assets ≤ -0.18 Then **Fraudulent**.	0.9609

Incorporating the knowledge of these important features for fraudulent (class 1) and grey area (class 2) classifications provides additional depth to our year-by-year analysis of Table 9.

For the years classified as "grey area" (2013-2017, 2019-2021), we observe a consistent presence of Sales/Total Assets and Working capital/total assets ratios in the anchor if-then rules. This aligns with our identified top features for grey area companies. The recurring appearance of these ratios supports Altman's (1968) Z-score model [20], which

emphasizes the importance of these metrics in assessing financial health. The presence of these indicators in the grey zone classification years suggests that the company's financial position was ambiguous, neither clearly healthy nor fraudulent, as noted by Dechow et al. (2011) [53] in their study on predicting material accounting misstatements.

Our model's fraudulent classification in 2018 is noteworthy, particularly when considering our identified top features for fraudulent companies. The rule for this year emphasizes Retained Earnings/Total Assets and net profit/Total asset ratios, which are among our key indicators for fraudulent activities. This aligns with Beneish's (1999) [43] research on detecting earnings manipulation. Similarly, the 2022 fraudulent classification incorporates Retained Earnings/Total Assets and Working capital/total assets, supporting Persons' (1995) [51] findings on identifying fraudulent financial reporting. The model's ability to detect potential fraud in 2018, four years before the 2022 regulatory action, highlights its predictive strength and sensitivity to subtle financial irregularities. This early identification, focusing on specific ratios, demonstrates our model's alignment with established financial theory, particularly Beneish's work. Such early detection capabilities underscore our model's effectiveness in identifying potential fraudulent activities before they become evident to regulators.

The grey area classifications in later years (2019-2021) incorporate additional ratios such as Primary business income/Total asset and Accounts receivable/Total asset. While these are not in our top 5 for grey zone, their inclusion supports Kaminski et al.'s (2004) [54] assertion that a wide range of financial ratios can be relevant in fraud detection, depending on the specific context.

The recent scrutiny of this company by the Securities and Exchange Commission of Thailand for alleged misconduct in its 2022 financial reporting provides a compelling real-world context for our analysis. This regulatory action aligns remarkably with our model's classification of the company as "fraudulent" in 2022, offering a form of external validation for the model's effectiveness. The progression from grey area classifications in previous years (2019-2021) to a fraudulent classification in 2022 supports Rezzae's (2005) [60] concept of fraud existing on a continuum, providing valuable insights into the development of financial misconduct over time. This alignment between our model's predictions and actual regulatory action underscores the potential application of such analytical approaches in regulatory oversight, possibly serving as an early warning system for identifying high-risk companies. The ten-year span of our analysis (2013-2022) proves particularly valuable in tracking the company's financial health trajectory leading up to the alleged misconduct, reinforcing the importance of longitudinal studies in fraud detection, as emphasized by Kaminski et al. (2004) [54]. The prominence of specific indicators like Retained Earnings/Total Assets and Working capital/total assets in the 2022 fraudulent classification rule aligns with our identified top features for fraudulent companies, further validating the significance of these metrics in fraud detection. Ultimately, this real-world regulatory action not only adds significant weight to our model's findings but also presents an opportunity for further refinement and validation of our approach, potentially enhancing its accuracy and applicability in similar cases. This case study demonstrates the practical relevance of sophisticated analytical approaches in identifying potential financial misconduct, while also highlighting the complex and evolving nature of fraudulent activities in corporate finance.

6- Conclusions

The research demonstrates the efficacy of using SHAP-based instance weighting in conjunction with XGBoost and Anchor Explainable AI to enhance fraud detection in financial datasets. By leveraging SHAP values, we assign importance weights to individual instances, ensuring that the model prioritizes the most informative samples during training. This method significantly improves the model's precision and recall, particularly in detecting fraudulent activities, as evidenced by the perfect recall score for fraudulent instances and the substantial increase in accuracy and precision metrics. The study also highlights the interpretability benefits provided by Anchor Explainable AI, which generates easy-to-understand if-then rules for the model's predictions. These rules offer clear insights into the factors influencing each classification, thus enhancing transparency and trust in the model's decision-making process. The approach effectively distinguishes between non-fraudulent, fraudulent, and grey area cases, providing nuanced insights that can inform more effective risk management and regulatory compliance strategies. Overall, the integration of SHAP-based instance weighting and Anchor Explainable AI into the fraud detection framework presents a robust and interpretable solution for identifying and understanding fraudulent activities in financial data. This approach not only addresses the challenges of class imbalance and model interpretability but also sets a foundation for future advancements in the field of fraud detection.

The implications of this research are significant. Enhanced fraud detection is achieved through SHAP-based instance weighting, which improves the model's ability to detect fraudulent activities by focusing on the most informative instances, leading to higher accuracy and recall. Improved model interpretability is achieved by integrating Anchor Explainable AI, which provides clear if-then rules that help stakeholders understand the reasoning behind each classification, crucial for regulatory compliance and building user trust. Resource optimization is facilitated by accurately distinguishing between non-fraudulent, fraudulent, and grey area cases, allowing financial institutions to allocate investigative resources more efficiently. The methods proposed are applicable across various sectors within the financial industry, providing a robust framework for enhancing fraud detection systems in real-world scenarios.

However, there are some limitations to this approach. The model's performance heavily relies on the quality and comprehensiveness of the financial data used, with any gaps or inaccuracies potentially affecting accuracy. The integration of SHAP values and Anchor explanations adds to the computational load, requiring significant processing power and time, especially for large datasets. While the proposed methods are effective for the datasets used in this study, their applicability to other datasets and types of fraud might vary, necessitating additional validation across diverse datasets.

Future work should focus on broader validation by testing the proposed methods on a wider range of datasets and fraud types to validate their generalizability and robustness across different financial contexts. Developing methods to optimize the computational efficiency of SHAP and Anchor explanations can facilitate real-time fraud detection applications. Investigating additional feature engineering techniques to identify new financial indicators could further enhance the model's performance. Exploring the integration of SHAP-based instance weighting with other advanced machine learning models and techniques, such as deep learning, can further improve detection accuracy. Finally, investigating ways to integrate the model outputs with regulatory frameworks can streamline compliance and reporting processes for financial institutions.

7- Declarations

7-1- Author Contributions

Conceptualization, P.T., S.S., and S.C.; methodology, P.T. and S.C.; software, P.T. and S.C.; validation, P.T., S.S., S.C., and D.N.M.N.; formal analysis, P.T., S.S., and S.C.; investigation, P.T., S.S., and S.C.; resources, S.S.; data curation, P.T., S.S., S.C., and D.N.M.N.; writing—original draft preparation, P.T., S.S., S.C., and D.N.M.N.; writing—review and editing, P.T., S.S., S.C., and D.N.M.N.; visualization, P.T., S.S., S.C., and D.N.M.N.; supervision, P.T. and S.S.; project administration, S.S. All authors have read and agreed to the published version of the manuscript.

7-2- Data Availability Statement

The data presented in this study are available in the article.

7-3- Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

7-4- Ethical Approval

Permission for the study was obtained from the ethics committee of Walailak University, Thailand (protocol no. WUEC-24-207-01).

7-5- Institutional Review Board Statement

Not applicable.

7-6- Informed Consent Statement

Not applicable.

7-7- Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancies have been completely observed by the authors.

8- References

[1] West, J., & Bhattacharya, M. (2016). Intelligent financial fraud detection: A comprehensive review. Computers & Security, 57, 47-66. doi:10.1016/j.cose.2015.09.005.

[2] Abdallah, A., Maarof, M. A., & Zainal, A. (2016). Fraud detection system: A survey. Journal of Network and Computer Applications, 68, 90-113. doi:10.1016/j.jnca.2016.04.007.

[3] Khaksar, J., Salehi, M., & DashtBayaz, M. L. (2022). The relationship between auditor characteristics and fraud detection. Journal of Network and Computer Applications, 20(1), 79-101. doi:10.1108/JFM-02-2021-0024.

[4] Cordis, A. (2023). Political alignment and corporate fraud: Evidence from the United States of America. Journal of Applied Accounting Research. doi:10.1108/JAAR-06-2022-0159.

[5] Rahman, M. J., & Jie, X. (2024). Fraud detection using fraud triangle theory: Evidence from China. Journal of Financial Crime, 31(1), 101–118. doi:10.1108/JFC-09-2022-0219.

[6] Maniatis, A. (2022). Detecting the probability of financial fraud due to earnings manipulation in companies listed in Athens stock exchange market. Journal of Financial Crime, 29(2), 603–619. doi:10.1108/JFC-04-2021-0083.

[7] Gong, Y., Li, J., Xu, Z., & Li, G. (2022). Detecting financial fraud using two types of Benford factors: Evidence from China. Procedia Computer Science, 214, 656–663. doi:10.1016/j.procs.2022.11.225.

[8] Xiuguo, W., & Shengyong, D. (2022). An Analysis on Financial Statement Fraud Detection for Chinese Listed Companies Using Deep Learning. IEEE Access, 10, 22516-22532. doi:10.1109/ACCESS.2022.3153478.

[9] Craja, P., Kim, A., & Lessmann, S. (2020). Deep learning for detecting financial statement fraud. Decision Support Systems, 139, 113421. doi:10.1016/j.dss.2020.113421.

[10] Pai, P. F., Hsu, M. F., & Wang, M. C. (2011). A support vector machine-based model for detecting top management fraud. Knowledge-Based Systems, 24, 314–321. doi:10.1016/j.knosys.2010.10.003.

[11] Alfaiz, N. S., & Fati, S. M. (2022). Enhanced Credit Card Fraud Detection Model Using Machine Learning. Electronics, 11(4), 662. doi:10.3390/electronics11040662.

[12] Strelcenia, E., & Prakoonwit, S. (2023). Improving Classification Performance in Credit Card Fraud Detection by Using New Data Augmentation. AI, 4(1), 172–198. doi:10.3390/ai4010008.

[13] Chaquet-Ulldemolins, J., Gimeno-Blanes, F.-J., Moral-Rubio, S., Muñoz-Romero, S., & Rojo-Álvarez, J.-L. (2022). On the Black-Box Challenge for Fraud Detection Using Machine Learning (I): Linear Models and Informative Feature Selection. Applied Sciences, 12(7), 3328. doi:10.3390/app12073328.

[14] Zhao, Z., & Bai, T. (2022). Financial Fraud Detection and Prediction in Listed Companies Using SMOTE and Machine Learning Algorithms. Entropy, 24(8), 1157. doi:10.3390/e24081157.

[15] Ali, A. A., Khedr, A. M., El-Bannany, M., & Kanakkayil, S. (2023). A Powerful Predicting Model for Financial Statement Fraud Based on Optimized XGBoost Ensemble Learning Technique. Applied Sciences, 13(4), 2272. doi:10.3390/app13042272.

[16] Cheah, P. C. Y., Yang, Y., & Lee, B. G. (2023). Enhancing Financial Fraud Detection through Addressing Class Imbalance Using Hybrid SMOTE-GAN Techniques. International Journal of Financial Studies, 11(3), 110. doi:10.3390/ijfs11030110.

[17] El Hlouli, F. Z., Riffi, J., Sayyouri, M., Mahraz, M. A., Yahyaouy, A., El Fazazy, K., & Tairi, H. (2023). Detecting Fraudulent Transactions Using Stacked Autoencoder Kernel ELM Optimized by the Dandelion Algorithm. Journal of Theoretical and Applied Electronic Commerce Research, 18(4), 2057-2076. doi:10.3390/jtaer18040103.

[18] Raval, J., Bhattacharya, P., Jadav, N. K., Tanwar, S., Sharma, G., Bokoro, P. N., Elmorsy, M., Tolba, A., & Raboaca, M. S. (2023). RaKShA: A Trusted Explainable LSTM Model to Classify Fraud Patterns on Credit Card Transactions. Mathematics, 11(8), 1901. doi:10.3390/math11081901.

[19] El Kafhali, S., Tayebi, M., & Sulimani, H. (2024). An Optimized Deep Learning Approach for Detecting Fraudulent Transactions. Information, 15(4), 227. doi:10.3390/info15040227.

[20] Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The Journal of Finance, 23(4), 589-609. doi:10.2307/2978933.

[21] Altman, E. I. (1983). Corporate financial distress: A complete guide to predicting, avoiding, and dealing with bankruptcy. John Wiley & Sons. doi:10.1002/9781118267806.

[22] Martins, T., de Almeida, A. M., Cardoso, E., & Nunes, L. (2024). Explainable Artificial Intelligence (XAI): A Systematic Literature Review on Taxonomies and Applications in Finance. IEEE Access, 12, 618-629. doi:10.1109/access.2023.3347028.

[23] Medianovskyi, K., Malakauskas, A., Lakstutiene, A., & Yahia, S. B. (2023). Interpretable machine learning for SME financial distress prediction. 12th International Conference on Information Systems and Advanced Technologies "ICISAT 2022". ICISAT 2022. Lecture Notes in Networks and Systems, Istanbul, Turkey, doi.org/10.1007/978-3-031-25344-7_42.

[24] Torky, M., Gad, I., & Hassanien, A. E. (2023). Explainable AI Model for Recognizing Financial Crisis Roots Based on Pigeon Optimization and Gradient Boosting Model. International Journal of Computational Intelligence Systems, 16, 50. doi:10.1007/s44196-023-00222-9.

[25] Tran, K. L., Le, H. A., Nguyen, T. H., & Nguyen, D. T. (2022). Explainable Machine Learning for Financial Distress Prediction: Evidence from Vietnam. Data, 7(11), 160. doi:10.3390/data7110160.

[26] Nallakaruppan, M. K., Balusamy, B., Shri, M. L., Malathi, V., & Bhattacharyya, S. (2024). An Explainable AI framework for credit evaluation and analysis. Applied Soft Computing, 153, 111307. doi:10.1016/j.asoc.2024.111307.

[27] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. 31st International Conference on Neural Information Processing Systems (NIPS). Curran Associates, New York, United States. doi:10.48550/arXiv.1705.07874.

[28] Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: high-precision model-agnostic explanations. AAAI Conference on Artificial Intelligence, 2-7 February 2018, Louisiana, United States. doi:10.1609/aaai.v32i1.11491.

[29] Band, S. S., Yarahmadi, A., Hsu, C. C., Biyari, M., Sookhak, M., Ameri, R., Dehzangi, I., Chronopoulos, A. T., & Liang, H. W. (2023). Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods. Informatics in Medicine Unlocked, 40, 101286. doi:10.1016/j.imu.2023.101286.

[30] Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, United States. doi:10.48550/arXiv.1907.10902.

[31] Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., & Talwalkar, A. (2017). Hyperband: A novel bandit-based approach to hyperparameter optimization, Journal of Machine Learning Research, 18, 6765–6816. doi:10.48550/arXiv.1603.06560.

[32] Sawangarreerak, S., & Thanathamathee, P. (2021). Detecting and Analyzing Fraudulent Patterns of Financial Statement for Open In-novation Using Discretization and Association Rule Mining. Journal of Open Innovation: Technology, Market, and Complexity, 7(2), 128. doi.org/10.3390/joitmc7020128.

[33] Elewa, M. M. (2022). Using Altman Z-Score Models for Predicting Financial Distress for Companies - The Case of Egypt panel data analysis. Alexandria Journal of Accounting Research, 6(1), 1-28. doi:10.21608/aljalexu.2022.225155.

[34] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2006). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. doi.org/10.1613/jair.953.

[35] Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32. doi:10.1023/A:1010933404324.

[36] Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13-17 August 2016, New York, NY, USA. doi:10.1145/2939672.2939785.

[37] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, California, United States.

[38] Anbananthen, K. S. M., Busst, M. B. M. A., Kannan, R., & Kannan, S. (2023). A Comparative Performance Analysis of Hybrid and Classical Machine Learning Method in Predicting Diabetes. Emerging Science Journal, 7(1), 102-115. doi:10.28991/ESJ-2023-07-01-08.

[39] Abraham, A., Mohideen, H. S., & Kayalvizhi, R. (2023). A Tabular Variational Auto Encoder-Based Hybrid Model for Imbalanced Data Classification with Feature Selection. IEEE Access, 11, 122760-122771. doi:10.1109/ACCESS.2023.3329139.

[40] Antwarg, L., Miller, R. M., Shapira, B., & Rokach, L. (2021). Explaining anomalies detected by autoencoders using Shapley Additive Explanations. Expert Systems with Applications, 186, 115736. doi:10.1016/j.eswa.2021.115736.

[41] Ge, W., & McVay, S. (2005). The disclosure of material weaknesses in internal control after the Sarbanes-Oxley Act. Accounting Horizons, 19(3), 137-158. doi:10.2308/acch.2005.19.3.137.

[42] Beaver, W. H. (1966). Financial ratios as predictors of failure. Journal of Accounting Research, 4, 71-111. doi:10.2307/2490171.

[43] Beneish, M. D. (1999). The detection of earnings manipulation. Financial Analysts Journal, 55(5), 24-36.

[44] Ou, J. A., & Penman, S. H. (1989). Financial statement analysis and the prediction of stock returns. Journal of Accounting and Economics, 11(4), 295-329. doi:10.1016/0165-4101(89)90017-7.

[45] Skousen, C. J., Smith, K. R., & Wright, C. J. (2009). Detecting and predicting financial statement fraud: The effectiveness of the fraud triangle and SAS No. 99. Advances in Financial Economics, 13, 53-81. doi:10.1108/S1569-3732(2009)0000013005.

[46] Cressey, D. R. (1953). Other people's money: A study in the social psychology of embezzlement. American Journal of Sociology, 59(6). doi:10.15388/Teise.2021.120.10.

[47] Wolfe, D. T., & Hermanson, D. R. (2004). The fraud diamond: Considering the four elements of fraud. The CPA Journal, 74, 38.

[48] Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. Econometrica, 47, 263-291. doi:10.2307/1914185.

[49] Abdel-Khalik, A. R. (2014). Prospect theory predictions in the field: Risk seekers in settings of weak accounting controls. Journal of Accounting Literature, 33(1-2), 58-84. doi:10.1016/j.acclit.2014.10.001.

[50] Dechow, P. M., Sloan, R. G., & Sweeney, A. P. (1996). Causes and consequences of earnings manipulation: An analysis of firms subject to enforcement actions by the SEC. John Wiley & Sons., 13(1), 1-36. doi:10.1111/j.1911-3846.1996.tb00489.x.

[51] Persons, O. S. (1995). Using financial statement data to identify factors associated with fraudulent financial reporting. Journal of Applied Business Research, 11(3), 38. doi:10.19030/jabr.v11i3.5858.

[52] Kanapickienė, R., & Grundienė, Ž. (2015). The model of fraud detection in financial statements by means of financial ratios. Procedia-Social and Behavioral Sciences, 213, 321-327. doi:10.1016/j.sbspro.2015.11.545.

[53] Dechow, P. M., Ge, W., Larson, C. R., & Sloan, R. G. (2011). Predicting material accounting misstatements. Contemporary Accounting Research, 28(1), 17-82. doi:10.1111/j.1911-3846.2010.01041.x.

[54] Kaminski, K. A., Wetzel, T. S., & Guan, L. (2004). Can financial ratios detect fraudulent financial reporting?. Managerial Auditing Journal, 19(1), 15-28. doi:10.1108/02686900410509802.

[55] Summers, S. L., & Sweeney, J. T. (1998). Fraudulently misstated financial statements and insider trading: An empirical analysis. The Accounting Review, 73(1), 131-146.

[56] Grice, J. S, & Ingram, R. W. (2001). Tests of the generalizability of Altman's bankruptcy prediction model. Journal of Business Research, 54(1), 53-61. doi:10.1016/S0148-2963(00)00126-0.

[57] Sharma, D. S., & Iselin, E. R. (2003). The relative relevance of cash flow and accrual information for solvency assessments: A multi-method approach. Journal of Business Finance & Accounting, 30(7-8), 1115-1140. doi:10.1111/1468-5957.05421.

[58] Burgstahler, D., & Dichev, I. (1997). Earnings management to avoid earnings decreases and losses. Journal of Accounting and Economics 1997, 24(1), 99-126. doi:10.1016/S0165-4101(97)00017-7.

[59] SEC. (2024). The Securities and Exchange Commission (SEC), Bangkok, Thailand. Available online: https://www.sec.or.th/EN/Pages/Home.aspx, (accessed on November 2024).

[60] Rezaee, Z. (2005). Causes, consequences, and deterrence of financial statement fraud. Critical Perspectives on Accounting, 16(3), 277-298. doi:10.1016/S1045-2354(03)00072-8.