International Journal of Engineering Applied Sciences and Technology, 2024
Vol. 9, Issue 04, ISSN No. 2455-2143, Pages 187-199
Published Online August 2024 in IJEAST (http://www.ijeast.com)

AUTOMOBILE INSURANCE FRAUD DETECTION USING ENSEMBLE LEARNING MODELS

Navin Duwadi
Faculty of Science, Health and Technology
Nepal Open University, Lalitpur, Nepal

Dr. Bhoj Raj Ghimire
Asst. Prof. Faculty of Science, Health and Technology
Nepal Open University, Lalitpur, Nepal

Abstract— **Automobile insurance fraud is a universal problem that has negative effects on both insurance companies and policyholders. This research proposes a novel ensemble learning model to accurately detect potential fraudulent vehicle insurance claims. By leveraging advanced machine learning techniques and addressing the challenge of imbalanced data, our model aims to enhance fraud detection efficiency and reduce financial losses. Our approach combines a stacking ensemble learner with carefully selected base classifiers, meta-classifier and data pre-processing techniques. We evaluate the model's performance on a real-world dataset of insurance claims, demonstrating superior results compared to existing methods. Notably, our model achieves high accuracy, recall, precision and area under curves, ensuring comprehensive detection and minimizing false positives.**

Keywords: **ensemble learning, machine learning, fraud detection, over sampling, under sampling, SMOTE**

I. INTRODUCTION

Insurance fraud is an enduring challenge within the insurance industry, exerting substantial financial pressure and undermining trust in the system. The landscape of insurance fraud has evolved significantly over time, driven by the ingenuity of fraudsters who continually adapt and refine their techniques [1]. As these fraudulent methods have grown in complexity, so too has the frequency and severity of insurance fraud incidents. One particularly concerning facet of insurance fraud is vehicle insurance fraud, where individuals or groups conspire to fabricate or inflate claims related to property damage or personal injuries resulting from accidents. These fraudulent activities, often executed with careful planning and deception, impose a heavy burden on insurance providers, policyholders, and, ultimately, the broader economy.

Fraud poses a significant challenge, leading to substantial financial losses for many insurance companies. Leveraging data mining techniques can mitigate some of these losses by tapping into extensive customer data repositories. In addition to address concerns related to scalability and efficacy, the task of fraud detection encounters technical hurdles, particularly the imbalanced nature of datasets. This issue, however, has not received widespread attention within the insurance fraud detection community. The data employed for insurance fraud detection, and fraud detection in general, tends to exhibit an imbalance, with fraudulent cases forming the minority class and legitimate cases representing the majority. Utilizing the data in its raw form yields high accuracy in predicting legitimate cases but falls short in identifying any instances of fraudulent activity [2].

Various supervised and unsupervised learning techniques have been suggested to enhance the detection of fraud in automobile industry. Several supervised machine learning models, such as logistic regression, support vector machine, decision tree, naïve bays, random forests, and neural networks, can be employed for this purpose [3]. These models can be trained on extensive insurance data to identify risk factors for fraud activity and create predictive models. A key advantage of ML models is their capacity to manage large and complex datasets, including unstructured data. Moreover, ML-based models have the ability to learn and adapt over time, allowing them to be retrained as new data becomes available, thereby enhancing their accuracy and performance [4]. However, traditional ML models are often challenged by issues related to data quality, bias, and interpretability. Due to the sensitive nature of this task, it is crucial to identify an accurate model. These days' ensemble models have demonstrated notable improvements over traditional single classifiers.

The primary hurdle in insurance fraud detection lies in the highly imbalanced distribution of regular and fraudulent transactions. This imbalance challenges the reliability of any machine learning algorithm in discerning patterns within

fraudulent transactions. To address this issue, one potential approach is through sampling. The goal of sampling is to modify the distribution of either the minority or majority class to achieve an approximately uniform distribution [5]. However, this adjustment carries the risk of the classifier either overfitting or under fitting normal transactions, potentially leading to misclassifying a fraudulent transaction as normal or vice versa. Various sampling techniques exist to tackle the class imbalance problem, including oversampling the minority class, undersampling the majority class, a combination of both, and the application of SMOTE [6].

This paper presents a fraud detection approach utilizing a dataset that was re-sampled with both oversampling and undersampling technique. The research involved designing insurance fraud detection models, referred to as base-classifier models, using random forest, support vector machine, logistic regression, AdaBoost, and XGBoost classifiers. The innovative application of stacking and voting to process the fraud detection models holds the potential for improved results. A common challenge with imbalanced data is accurately measuring classifier performance. Recent studies on imbalanced datasets have employed more effective performance metrics such as recall, precision, and area under the curve (AUC) [7]. In this paper, the models' performance was evaluated using recall, precision, and AUC curves.

This research seeks to address the critical issue by examining the potential of ensemble learning models, complemented by advanced techniques for handling imbalanced data, and enhanced through feature engineering and hyperparameters tuning. By leveraging the power of ensemble models, this paper aims to improve the accuracy and efficiency of detecting fraudulent automobile insurance claims, thereby mitigating financial losses and promoting the integrity of the insurance ecosystem. This paper outlines the comprehensive strategy for conducting this research, with the ultimate goal of advancing fraud detection capabilities within the automobile insurance sector.

II. LITERATURE SURVEY

Ensemble learning, which uses multiple base estimators, has been proven to be a powerful tool in detecting insurance fraud, achieving better predictive accuracy than a single base learner. In the healthcare, financial and automobile sector, ensemble machine learning techniques have been used for fraud detection. Various classifiers were employed, including K-Nearest Neighbours (KNN), Artificial Neural Network (ANN), Linear Discriminant Analysis (LDA), Gradient Boosting Machine (GBM), Bagging classifier, and stacking meta-estimator. The best accuracy was obtained when feature selection was applied to the Stacking classifier. Other research papers applied machine learning techniques such as Decision Trees, Bagging, Random Forests, and Boosting for fraud detection in health insurance, with the best results achieved using the ensemble technique. In the auto insurance sector, an ensemble learning method, Auto Insurance Multi-modal Learning (AIML), was proposed, which includes feature extraction from multi-modal data, feature engineering, and tree-based classification [8], [9], [10], [11], [1]. These studies demonstrate the effectiveness of ensemble learning in detecting insurance fraud across different sectors, but continuous research and development of these models are necessary to maintain their effectiveness as fraud patterns change over time.

Addressing class imbalance in machine learning is crucial for achieving accurate predictions, especially in scenarios where one class has significantly fewer samples than the other. Techniques such as data sampling methods (e.g., random under-sampling and over-sampling), weighted loss functions (e.g., focal loss), ensemble methods, data augmentation using generative adversarial networks (GANs), and hybrid approaches that combine multiple techniques have been proposed to handle data imbalance effectively [12], [13], [14], [15].

Undersampling and oversampling is a prevalent technique in insurance fraud detection, addressing the issue of imbalanced data distribution where fraudulent cases often form the minority class. Multiple studies has been conducted with different sampling methods, SMOTE and ROSE, to remove class imbalance in automobile insurance fraud detection, revealing that models built using the feature selection perform slightly better. Similarly, an innovative method based on building insurance fraud detection models using Decision tree (DT), Support vector machine (SVM) and Artificial Neural Network (ANN), on data partitions derived from under-sampling of the majority class. A novel hybrid Automobile Insurance Fraud Detection System was proposed where undersampling of the majority class was performed by using a fuzzy clustering algorithm, eliminating the outliers from the majority class samples s. ([16], [17], [18], [19]). In conclusion, some cases undersampling improves the performance of fraud detection models by dealing with the imbalanced data distribution problem and in some cases oversampling improves the performance of the fraud detection model.

Machine learning is very useful for combating insurance fraud, but its effectiveness is important on optimal hyperparameter tuning. This process involves systematically evaluating different configurations for settings like learning rates or tree depths within the model. Hyperparameter tuning refines a model's ability to generalize beyond the training data, preventing overfitting and enhancing its performance on unseen fraudulent claims. Additionally, tuning can identify configurations that optimize training speed and resource usage. Furthermore, it allows us to strike a balance between accurately detecting fraud (recall) and minimizing false positives (precision) by tailoring the model to the insurance company's cost structure. Techniques like grid search and random search can be employed for this exploration, while robust evaluation metrics like AUC and cost-sensitive scores guide the selection of the best hyperparameter configuration [20], [21], [22], [23], [24]. Ultimately, hyperparameter tuning

empowers machine learning models to become more efficient, accurate, and adaptable fraud detection tools within the insurance domain.

Ensemble stacking, combined with hyperparameter tuning, offers a powerful approach to enhancing model performance in insurance fraud detection. Stacking leverages the strengths of multiple base models, each potentially trained with different hyperparameter configurations. Hyperparameter tuning ensures each base model operates at its peak efficiency by optimizing settings like learning rates or tree depths. This optimization not only improves the individual models' performance but also allows the stacking ensemble to learn more robust patterns from their diverse outputs. The final model in the stacking framework, often called the meta-learner, then combines the predictions from the tuned base models, leading to a more accurate and generalizable fraud detection system [25], [20], [26], [27]. This approach can be particularly beneficial when dealing with complex fraud patterns that might be challenging for a single model to capture effectively.

In the reference [28], the author conducted an extensive analysis of the existing research on Medicare fraud detection, specifically focusing on the performance of two popular gradient boosting techniques: CatBoost and XGBoost. Through their review, the authors observed that both CatBoost and XGBoost have been increasingly applied in this context, showcasing superior performance in terms of accuracy, precision, and recall when compared to traditional methods. The literature suggested that these algorithms excel in handling large and complex healthcare datasets, effectively capturing subtle patterns and anomalies associated with fraudulent claims. Furthermore, the review indicated that model interpretability remains a challenge with these techniques, emphasizing the need for further research in this area to ensure transparency in fraud detection systems. Hancock and Khoshgoftaar's review offers valuable insights into the applicability and potential advantages of CatBoost and XGBoost in the critical domain of Medicare fraud detection.

In the study [29], the author investigated the effectiveness of ensemble learning methods for fraud detection in the context of credit card. Their findings indicated that ensemble methods, which combine multiple base classifiers, offer a promising approach for enhancing the accuracy and reliability of health insurance fraud detection systems. By using technique like boosting, the ensemble models demonstrated superior performance in identifying potentially fraudulent claims compared to individual classifiers. The research also highlighted the importance of addressing the issue of class imbalance, a common challenge in fraud detection, through data pre-processing techniques to ensure a more balanced representation of legitimate and fraudulent cases. This study provides valuable insights into the practical application of ensemble learning methods for improving the detection of fraudulent activities within the finance sector, thereby potentially saving substantial costs and safeguarding the integrity of the insurance system [30].

In the work [31], the author focused on the challenging task of fraud detection using large-scale imbalanced datasets. Their findings indicated that dealing with imbalanced data in fraud detection is a crucial concern, and it requires specialized techniques for effective model training. The authors explored various methodologies, including resampling methods, ensemble learning, and cost-sensitive learning, to address the class imbalance issue. They discovered that a combination of these strategies, particularly a tailored ensemble approach, yielded improved results in terms of fraud detection performance. The study emphasized the significance of understanding and managing imbalanced datasets to enhance the accuracy and reliability of fraud detection systems, particularly in scenarios where the fraudulent class is significantly underrepresented, making these findings valuable for practitioners and researchers working in fraud detection with large-scale imbalanced datasets.

In the article [32], the author have investigated the use of machine learning techniques for predicting insurance fraud. Hybrid learning methods, which involve combining different algorithms seems to have greater flexibility, and have demonstrated superior performance compared to traditional approaches. Ensemble learning has gained importance recently due to its reliability and adaptability across various approaches. Recent studies indicate that ensembles not only enhance prediction accuracy but also address the machine learning challenges such as overfitting, class imbalance, and concept drift [33]. The appeal of ensemble models and their applications lies in their ability to generalize well. While building ensembles is resource-intensive in terms of time and effort, it can be viewed as a one-time investment, as once assembled, ensembles consistently yield highly efficient results [2].

Paper [34] conducted a study focusing on the application of machine learning techniques in the detection of fraud in motor insurance. The primary objective of the research was to create a model for identifying fraudulent motor insurance claims using classification algorithms. The study proposed an optimal model through the utilization of specific evaluation criteria. The investigation encompassed motor insurance claims data sourced from Sri Lanka Insurance, with a dataset comprising 30,098 motor claims. The study employed Artificial Neural Network, Random Forest, and XGBoost algorithms as classifiers to determine the fraudulent nature of claims. The dataset was partitioned into training, validation, and testing sets for a comprehensive evaluation of these algorithms. However, the study acknowledged that when feeding data into a machine learning model with an imbalanced class variable, a bias toward the majority class might lead to misclassifying fraudulent claims as normal claims. To address this issue, the Synthetic Minority Oversampling Technique (SMOTE) was applied in conjunction with ensemble models.

International Journal of Engineering Applied Sciences and Technology, 2024
Vol. 9, Issue 04, ISSN No. 2455-2143, Pages 187-199
Published Online August 2024 in IJEAST (http://www.ijeast.com)

The model's performance was assessed using various criteria, including recall, precision, f1-score, precision-recall (PR) curve, and receiver operating characteristics (ROC) curve. Random Forest and XGBoost classifiers involve parameters requiring the researcher's decision, hyperparameter tuning was implemented and assessed. The findings indicated that Random Forest and XGBoost models outperformed neural network models. Although there was minimal disparity between Random Forest and XGBoost models, the Random Forest model with tuned hyperparameters demonstrated slightly superior performance compared to other models. The study found that ensemble models, such as the Random Forest and XGBoost models, exhibit superior performance in predicting motor insurance fraud claims. This underscores the significance of leveraging ensemble techniques to transform weak learners into strong learners.

III. METHODOLOGY

A. Data Collection and Description –

A diverse and representative dataset of automobile insurance claims is collected, containing both legitimate and fraudulent claims. The dataset includes features such as claim details, policyholder information, accident information, and historical claim patterns. Vehicle Insurance Claim Fraud Detection is a dataset sourced from a real American insurance company, provided by Oracle for educational purposes. This dataset focuses on the identification of fraudulent activities related to vehicle insurance claims. Vehicle insurance fraud involves colluding to submit false or exaggerated claims regarding property damage or personal injuries resulting from an accident. Dataset description is as follows,

Table1: Data Description

Variables	**Description**
Month	Month of accident
Week Of Month	week of month of accident
Day Of Week	day of the week of accident
Make	manufacturing company of the vehicle
Accident Area	Location of accident
Day Of Week Claimed	day of the week of claimed
Month Claimed	month of claimed
Week Of Month Claimed	week of the month of claimed
Sex	sex of the person of claimed
Marital Status	marital status of the person claimed
Age	age of the person claimed
Fault	owner of the insurance policy either policy holder or third party
Policy Type	type of insurance policy
Vehicle Category	category of the vehicle
Vehicle Price	price of the vehicle
Fraud Found_P	label of the data for fraud or non-fraud
Base Policy	insurance 's base policy
Variables	Description
Policy Number	policy number of the insurance
Rep Number	repair number
Deductible	amount that is deductible
Driver Rating	rating of the driver
Days Policy Accident	days policy accident
Days Policy Claim	days policy claimed
Past Number Of Claims	total number of the claims in the past
Age Of Vehicle	age of the vehicle
Age Of Policy Holder	age of the policy holder

International Journal of Engineering Applied Sciences and Technology, 2024
Vol. 9, Issue 04, ISSN No. 2455-2143, Pages 187-199
Published Online August 2024 in IJEAST (http://www.ijeast.com)

Police Report Filed	police report filed date
Witness Present	Witness present or not?
Agent Type	type of the agent
Number Of Supplements	number of the supplements done
Address Change Claim	address change claim
Number Of Cars	number of cars owned by policy holder
Year	insurance year

B. Data Pre-Processing

Data pre-processing stands as a crucial initial step in modelling any dataset. The strength of a machine learning algorithm is greatly influenced by the cleanliness of the data. In our dataset, numerous categorical variables are present, prompting the use of a prominent technique known as one-hot encoding. This method transforms these categorical variables into binary ones, enhancing the modelling process and mitigating bias in the model. Fortunately, our data doesn't contain missing values, so we do not need to do anything about it. It's imperative to convert the target variable from categorical to an integer before modelling, ensuring the model comprehends the necessity to predict a binary outcome. Given the relatively straightforward nature of our data, our pre-processing steps remain minimal. However, before inputting the data into any algorithm, addressing the imbalance in the data becomes essential. To balance the data before modelling multiple resampling technique such as synthetic minority oversampling techniques (SMOTE) and random under sampling have been used.

C. Fraud Detection Models

Traditionally, machine learning uses a single model to solve a problem. Ensemble methods take a different approach. They combine multiple models, like a team of experts, to get a more accurate and reliable solution. There are different ways to create ensembles, but this research focuses on combining individual classifiers. There are two ways to combine classifiers: base classifier and meta-classifier combining methods. Simple methods like averaging and voting work well when all the models perform similarly, but they struggle with outliers and uneven performance. Meta-combining methods like stacking and grading are theoretically more powerful, but they can be more complex to train and prone to overfitting.

To measure the effectiveness of an ensemble classifier, accuracy is typically used, which represents the percentage of correct predictions. However, this can be misleading for problems with uneven class sizes. In such cases, even a simple rule can appear very accurate. Instead, this research uses recall to find positive examples. Recall is less affected by the majority class, making it a better measure for imbalanced problems. To visualize how well the model can distinguish between positive and negative classes AUC ROC is used. This is a graphical plot that illustrates the trade-off between two key performance metrics for binary classification true positive rate and false positive rate.

D. MODEL DEVELOPMENT

An ensemble model was developed using the selected technique. The base learners was trained on the pre-processed dataset, and their predictions was combined using appropriate aggregation methods to make the final fraud detection decision.

1) BASE LEARNERS

In this study, an ensemble approach was employed to develop an effective insurance fraud detection model. The dataset was used to evaluate the performance of different base learner algorithms. The base learners were trained using 10-fold cross-validation.

a) Random Forest

The Random Forest algorithm is widely recognized as one of the most prominent machine learning techniques. Notably, it demands minimal information planning, modelling, or demonstration yet consistently yields accurate results. Random Forests build upon the concept of decision trees. To be more specific, Random Forests consist of collections of decision trees, collectively enhancing prediction accuracy. The term forest is employed because it essentially constitutes a collection of decision trees [35].

The fundamental concept involves constructing numerous decision trees based on independent subsets of the dataset. At each node of these trees, a random selection of n features from the set of features is made, and the optimal split based on these features is determined. This ensemble approach contributes to the robustness and effectiveness of the Random Forest algorithm in making accurate predictions [6].

b) Logistic regression

Logistic regression is a technique used to model the likelihood of a discrete outcome based on input variables. Typically, logistic regression is applied to model binary outcomes, where the result can assume two values such as true/false or yes/no. However, it can be extended to handle scenarios with more than two possible discrete outcomes, a variation known as multinomial logistic regression. This statistical technique is particularly valuable in the analysis of classification problems. In scenarios where the objective is to ascertain whether a new

sample is best categorized into a specific group, logistic regression proves to be a useful analytical tool. In the realm of machine learning, where many problems involve classification tasks such as fraud detection, logistic regression serves as a valuable and applicable analytic technique [36].

c) Support vector machine

Support Vector Machine (SVM) is a powerful supervised machine learning algorithm used for classification and regression tasks. It works by finding the optimal hyperplane that best separates the data points of different classes in a high-dimensional space. The main objective of SVM is to maximize the margin between the closest data points of different classes, known as support vectors, and the hyperplane [37]. By doing so, SVM achieves a robust separation of classes, making it effective in handling linear as well as non-linear classification problems through the use of kernel functions. These kernel functions enable SVM to project data into a higher-dimensional space where a linear separator can be found. SVM is widely used in various applications, including image recognition, text categorization, and bioinformatics, due to its effectiveness in high-dimensional spaces and its ability to handle complex datasets [38].

d) AdaBoost

Adaptive boosting, also known as AdaBoost, is an ensemble machine learning technique for classification tasks. It combines multiple weak learners, typically decision trees with a single split, into a strong learner. AdaBoost works by iteratively training these weak learners, placing higher emphasis on instances that were misclassified by previous learners. This approach progressively improves the overall model's performance by focusing on the harder-to-learn examples. AdaBoost has become a popular technique due to its effectiveness in boosting the performance of weak learners and its relative resistance to overfitting [39].

e) XGBoost:

XGBoost, the earliest among the three Gradient Boosting Decision Trees (GBDTs) utilized in our approach, was introduced by [40] in 2016. Previous research has demonstrated that XGBoost is a particularly effective choice for classifying imbalanced Big Data sets. Notably, XGBoost introduces several enhancements to the standard GBDT technique. One significant improvement is the incorporation of an enhanced loss function during the training phase, featuring an additional regularization term designed to address overfitting. Moreover, XGBoost improves the process of calculating splits within the ensemble of Decision Trees it employs. Chen & Guestrin [40] introduced an "approximate algorithm" to estimate optimal split values, particularly beneficial when dealing with datasets too large to fit into main memory or in distributed environments.

Another notable advancement is XGBoost's ability to effectively handle sparse data, which often exhibits near-constant values with occasional deviations. XGBoost's sparsity-aware split finding feature allows it to efficiently leverage sparse data, facilitating the construction of Decision Trees with improved effectiveness. Overall, XGBoost incorporates these features to enhance its performance, making it a robust choice for classifying imbalanced and large-scale datasets [41].

f) Ensemble Voting

A Voting Classifier is an ensemble machine learning technique that combines the predictions of multiple different models to improve the overall classification accuracy. This method operates under the principle that aggregating the predictions of several diverse models can yield a more accurate and robust prediction than relying on a single model. In a voting classifier, each base model (such as logistic regression, random forest, and support vector machines) makes a prediction, and the final output is determined by a majority vote (hard voting) or the average of the predicted probabilities (soft voting) [42]. Hard voting considers the class that receives the most votes, while soft voting sums the probabilities of each class and selects the class with the highest total. Voting classifiers are particularly effective when the base models have complementary strengths and weaknesses, as the ensemble can balance out the individual models' biases and reduce the risk of overfitting, leading to enhanced performance on a wide range of datasets.

g) Ensemble Stacking

Stacking, also known as stacked generalization, is a powerful ensemble learning technique that utilizes a two-level learning approach. In the first level, multiple base learners (logistic regression, support vector machines) are trained on the original dataset. The predictions from these base learners become the input for a final meta-learner [43]. This meta-learner is trained to combine the strengths of the individual base learners and create a more robust ensemble model. Stacking can achieve superior performance compared to individual base learners by leveraging their complementary strengths and potentially mitigating their weaknesses. A recent paper by Iqbal et al., [44] explores the application of stacking ensembles for anomaly detection in time series data. Their work demonstrates the effectiveness of stacking in identifying complex anomalies that might be missed by individual models, showcasing the ongoing development of stacking techniques for various classification tasks.

2) META CLASSIFIER

To generate the final predictions, a meta-learner was trained using the outputs from the base learners obtained during cross-validation. Logistic Regression was chosen as the meta-learner algorithm because of its ability to learn linear combinations of predictions from different base learners. The predictions from

the Random Forest, Logistic Regression, Support Vector Machine, XGBoost, and Adaboost models for each validation fold were combined to form a new training dataset at the meta-level. This dataset had the same number of rows as the original samples, with features corresponding to the predictions made by each base learner. The actual class labels were also included in this dataset. The Logistic Regression model was then trained on this dataset to identify the optimal weights for each base learner's predictions for any given sample. By training on the outputs of various base models, the meta-learner can effectively combine their strengths and mitigate their individual weaknesses. This stacking ensemble method leads to improved predictive performance compared to relying on any single base learner for binary fraud classification.

3) HYPER PARAMETER TUNING

Hyperparameters tuning plays a crucial role in optimizing the performance of machine learning models, including ensemble methods. This paper outlines the proposed approach for hyperparameters tuning in the context of using ensemble learning models for automobile insurance fraud detection.

Each classifier was optimized through hyperparameter tuning using randomized search, which systematically tests various hyperparameter values to identify the best combination for enhancing a machine learning model's performance. For the Random Forest classifier, the hyperparameters tuned were the number of trees (n_estimators), random state and the maximum depth of each tree (max_depth). The optimal values were found to be n_estimators = 50, random_state = 42 and max_depth = 7. The Support Vector Machine (SVM) classifier was configured to output class probabilities by enabling the probability parameter and hyperparameter tuned were the gamma and C. The best values were found to be gamma = 0.1 and C = 10. For the Logistic Regression classifier, the solver was selected to Liblinear, C = 10 and random state = 42. The AdaBoost classifier was tuned to use n_estimators = 400 and learning_rate = 0.5 for predicting the class of a new data point. In the case of the XGBoost classifier, several hyperparameters were tuned, including the number of threads (n_jobs), learning rate (learning_rate), number of trees (n_estimators), maximum depth of each tree (max_depth), objective function (objective), and the boosting algorithm type (booster). The optimal combination was determined to be subsample = 0.8, min_child_weight = 5, max_depth = 4, gamma = 2 and colsample_bytree = 1.0. For the meta-learner, the learning rate, number of estimators, and max-depth were increased to improve results and ensure greater stability in the final classifications.

4) PERFROMANCE EVALUATION

In binary classification tasks, model predictions can be categorized into four groups based on their alignments with the actual classes: True positives refer to the positive cases that are correctly identified, whereas false positives are cases wrongly identified as positive. True negatives are instances accurately recognized as not belonging to the positive class, and false negatives are cases mistakenly classified as not belonging to the positive class. False positives occur when negative instances are wrongly predicted as positive. True negatives represent correct predictions for negative instances, while false negatives involve positive instances being misclassified as negative.

To accurately evaluate model performance, various metrics are commonly used. One widely used metric is classification accuracy, which reflects the overall rate of correct predictions. This research also takes into account other significant metrics derived from the confusion matrix. Precision measures the model's ability to return only true positives out of all positive predictions. Recall measures the proportion of actual positives that are correctly identified by the model. Furthermore, the F1-score and AUC are also considered. The F1-score offers a balanced evaluation by combining both precision and recall, while the AUC represents the quality of the classification across different threshold levels. The evaluation criteria are as follows:

Accuracy: Accuracy measures the percentage of all classifications (both fraudulent and legitimate) that are correctly identified. A higher accuracy value signifies a stronger match between the predicted and actual classes.

Precision: Precision evaluates the model's ability to correctly identify only fraudulent transactions among those it classified as fraudulent. A high precision indicates that there are fewer cases of legitimate transactions being mistakenly labelled as fraudulent.

Recall: Recall, also referred to as sensitivity, measures the percentage of actual fraudulent transactions that are accurately detected as fraudulent. A high recall indicates that fewer fraudulent transactions are overlooked.

F1 score: The F1-score offers a balanced assessment of precision and recall by calculating their harmonic mean. Higher F1-scores, approaching 1, indicate superior overall classification performance of the model.

AUC: The Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve measures the model's accuracy across all possible classification thresholds. A larger AUC value indicates a greater ability of the model to distinguish between the two classes.

IV. RESULT AND DISCUSSION

In this research on insurance fraud detection using machine learning, we evaluated the effectiveness of several classification algorithms, including Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM),

XGBoost (XGB), and AdaBoost (AB). These models were tested individually and in combination using voting and stacking ensemble method. The primary goal was to accurately detect fraudulent claims in a highly imbalanced dataset, where fraudulent cases are significantly outnumbered by legitimate ones. To address the class imbalance, both oversampling and undersampling techniques were applied. The dataset was first balanced using these techniques, and each base model was trained on the balanced data. Additionally, feature selection was performed using the Random Forest algorithm, which helped in identifying the most significant features contributing to fraud detection. This step reduced the dimensionality of the data and enhanced the models' performance by focusing on the most relevant attributes. To further optimize the models, hyperparameter tuning was conducted using RandomizedSearchCV. This method efficiently explored a wide range of hyperparameter combinations to identify the best configurations for each classifier. The optimized models were then evaluated, both individually and within the stacking ensemble framework.

The results of the study demonstrated that the stacking ensemble, particularly after feature selection and hyperparameter tuning, significantly outperformed the individual classifiers in terms of accuracy, recall, and Area Under the Curve (AUC). The ensemble approach effectively leveraged the strengths of each base model, such as LR's proficiency with linear relationships, RF's ability to handle non-linear interactions, SVM's robustness in high-dimensional spaces, XGB's power in complex data structures, and AB's enhancement of weak learners.

Table 2 and table 3 presents the evaluation of classifier parameters using accuracy, sensitivity, specificity, f1 score and AUC as performance metrics. The model's performance was evaluated by balancing the dataset using undersampling with a random undersampler and oversampling using SMOTE. The optimal classifier should exhibit the highest values in these metrics, as they reflect the sensitivity in classifying fraud instances. The stacking ensemble model (RF+LR+SVM+AB+XGB) with oversampling using SMOTE, feature engineering and hyperparameter tuning achieved the highest accuracy at 95%, sensitivity 96% and AUC ROC 99% with the fewest incorrectly classified instances and the most correctly classified ones. The ensemble stacking model also demonstrated the highest specificity at 95%.

Table 2: Result comparison using oversampling

Model	Accuracy	Precision	Recall	F1 score	ROC AUC
LR	0.799	0.783	0.826	0.804	0.885
RF	0.94	0.922	0.96	0.941	0.986
SVM	0.814	0.783	0.869	0.824	0.893
AB	0.852	0.826	0.891	0.858	0.929
XGB	0.932	0.913	0.955	0.933	0.979
VC(SVM+LR+RF)	0.881	0.852	0.922	0.886	0.952
SC(SVM+LR+RF)	0.94	0.943	0.950	0.947	0.986
VC(SVM+LR+RF+AB+XGB)	0.91	0.883	0.950	0.915	0.968
SC(SVM+LR+RF+AB+XGB)	0.95	0.945	0.96	0.950	0.99

Table 3: Result comparison using undersampling

Model	Accuracy	Precision	Recall	F1 score	ROC AUC
LR	0.747	0.711	0.832	0.766	0.796
RF	0.749	0.711	0.837	0.768	0.798
SVM	0.624	0.701	0.611	0.650	0.679
AB	0.753	0.707	0.865	0.777	0.791
XGB	0.707	0.688	0.755	0.719	0.779
VC(SVM+LR+RF)	0.756	0.718	0.843	0.775	0.800
SC(SVM+LR+RF)	0.750	0.706	0.853	0.772	0.807
VC(SVM+LR+RF+AB+XGB)	0.746	0.712	0.826	0.764	0.800
SC(SVM+LR+RF+AB+XGB)	0.758	0.713	0.862	0.780	0.801

International Journal of Engineering Applied Sciences and Technology, 2024
Vol. 9, Issue 04, ISSN No. 2455-2143, Pages 187-199
Published Online August 2024 in IJEAST (http://www.ijeast.com)

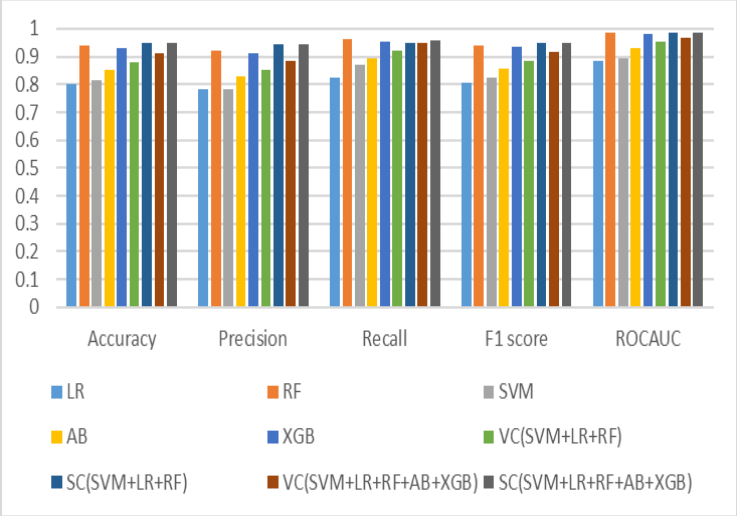

Fig.1. Model Comparison

The AUC values indicate that the model is considered excellent classifier. Figure 2 and 3 illustrates the AUC ROC curve of the voting and stacking classifier. The results clearly show that the stacking ensemble of all the five base classifier outperformed both the base classifiers, stacking model of random forest, support vector machine, and logistic regression and voting ensemble models. In conclusion, the stacking ensemble method, particularly when combined with data balancing techniques like oversampling, feature selection, and hyperparameter tuning via RandomizedSearchCV, proved to be the most effective approach in our insurance fraud detection study. It provided a robust and accurate solution, outperforming traditional individual classifiers and ensemble voting classifier offering a promising tool for detecting fraudulent activities in the insurance industry. Figure 1 illustrates a comparison of the models based on all metrics. The proposed ensemble stacking model demonstrates high accuracy with less variability compared to the other models.

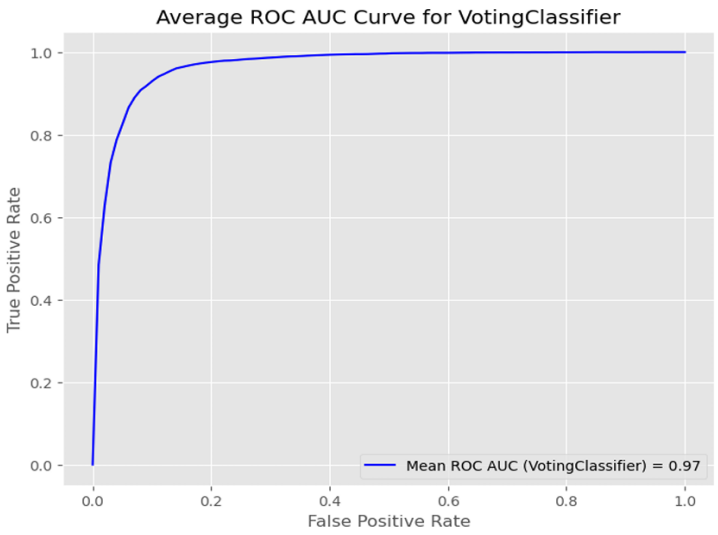

Fig. 2. Average ROCAUC Voting Classifier

International Journal of Engineering Applied Sciences and Technology, 2024
Vol. 9, Issue 04, ISSN No. 2455-2143, Pages 187-199
Published Online August 2024 in IJEAST (http://www.ijeast.com)

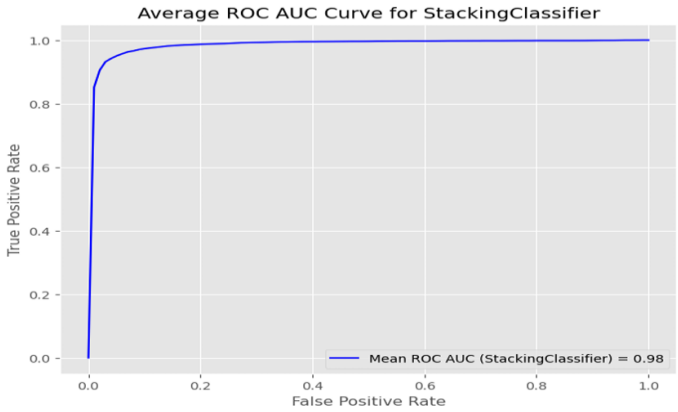

Fig.3. Average ROCAUC Stack Classifier

DISCUSSION

In this study, we explored the application of ensemble models for insurance fraud detection, comparing the performance of five base classifiers Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), AdaBoost, and XGBoost as well as two ensemble approaches Voting and Stacking. Through extensive experimentation, the Stacking ensemble of five base classifiers consistently outperformed both the individual base classifiers and the Voting ensemble across various evaluation metrics, including accuracy, precision, recall, and AUC-ROC. These findings suggest that leveraging the strengths of multiple base classifiers through the Stacking technique provides a more robust solution for detecting fraudulent insurance claims.

One of the key aspects of this research was the handling of imbalanced data, a common challenge in fraud detection. To address this, we employed both oversampling and undersampling techniques to balance the dataset, which had a significant impact on model performance. Among the two, oversampling proved more effective, allowing the classifiers to better differentiate between fraudulent and legitimate claims. This improvement in classification recall, particularly for the minority fraud class, was evident across all evaluation metrics.

In addition to resampling, hyperparameter tuning was conducted for each classifier using Randomized Search CV, which further optimized their performance. The combination of oversampling and hyperparameter tuning led to a substantial enhancement in model performance, particularly for the Stacking classifier. The best results were obtained when the Stacking classifier used Logistic Regression as the meta-learner, which effectively combined the predictions from the base classifiers RF, LR, Adaboost, XGBoost, and SVM. This meta-classifier leveraged the individual strengths of the base learners while mitigating their weaknesses, leading to a more informed and accurate final prediction.

The superior performance of the Stacking classifier, compared to both the individual base models and the Voting ensemble, can be attributed to its ability to intelligently combine diverse base classifiers. Unlike the Voting classifier, which simply averages predictions, the Stacking approach enables the meta-learner to learn from the patterns of errors made by the base classifiers, leading to more precise predictions. The highest accuracy, precision, recall, and AUC-ROC values were achieved with the Stacking ensemble, validating the effectiveness of this technique in detecting complex fraud patterns within the insurance dataset. The results of this study demonstrate the effectiveness of ensemble learning methods, particularly the stacking classifier, in detecting insurance fraud. When compared with previous studies, our approach shows notable improvements in key evaluation metrics such as accuracy, precision, recall, and AUC-ROC.

In the paper [45], the researchers employed traditional machine learning algorithms such as Logistic Regression, Support Vector Machine, Random Forest, Decision Tree, and KNN and ensemble stacking for healthcare fraud detection. While their stacking model achieved an accuracy of 88% and AUC-ROC of 88%, our stacking ensemble, which combines multiple base classifiers with a Logistic Regression meta-learner, outperformed these models with an accuracy of 95% and AUC-ROC of 99%. This improvement can be attributed to the stacking method's ability to leverage the strengths of diverse classifiers while mitigating their individual weaknesses, a feature that was not explored in the mentioned study.

Another study in [46], the author applied a soft voting ensemble approach for fraud detection, which achieved competitive results. However, our findings indicate that the stacking ensemble is more effective, as evidenced by our model's higher recall (96%) and AUC (99%) compared to their recall of 85% and AUC of 99%. Unlike the voting method, which averages predictions, the stacking approach allows the meta-learner to better capture relationships between

the base learners' predictions, resulting in superior recall, especially in identifying fraudulent cases.

In the work [47], the researcher explored the impact of data resampling techniques, including oversampling and undersampling, to address class imbalance in fraud datasets. Their research reported significant improvements in recall after oversampling. Our findings align with theirs, showing that oversampling, combined with feature selection and hyperparameter tuning, resulted in the best performance, particularly in terms of recall. However, our study extends this by incorporating a stacking ensemble, which further enhanced the model's robustness, especially when dealing with imbalanced datasets.

In summary, our research builds upon and advances existing work by incorporating a stacking ensemble approach, fine-tuning hyperparameters through Randomized Search CV, selecting important features through correlation and random forest and employing effective resampling techniques. These strategies contributed to superior model performance, highlighting the potential of ensemble methods in tackling the complexities of insurance fraud detection.

V. CONCLUSION

This research paper outlines a comprehensive strategy for enhancing automobile insurance fraud detection through the integration of ensemble learning models, imbalanced data handling techniques, feature engineering and hyperparameter tuning. The goal of this study was to develop a reliable fraud detection model that can accurately identify fraudulent insurance claim transactions. To achieve this, we implemented a stacking ensemble learning method that leverages the strengths of multiple base learners, including Random Forest, Support Vector Machine, Logistic regression, Adaboost and XGBoost, . We utilized a Logistic Regression model as the meta-learner to effectively integrate the predictions from these base learners. Our experiments on a real-world insurance fraud dataset demonstrated that the stacking ensemble model outperformed each individual base learner and combination of other base learners. The ensemble model achieved impressive performance metrics, including an accuracy of 95%, precision of 94%, recall of 96%, F1-score of 95%, and an AUC-ROC of 99%. These results highlight the effectiveness of our approach in tackling the challenges posed by imbalanced fraud data and the complexities involved in distinguishing fraudulent transactions from legitimate ones. The stacking ensemble technique allowed us to harness the strengths of individual models while compensating for their weaknesses. By combining the predictions from diverse base learners, the meta-learner could utilize their collective knowledge to make more informed decisions, significantly enhancing our fraud detection capabilities. This specially tailored approach, designed for the unique characteristics of insurance fraud data, provides a comprehensive methodology for fraud detection. In terms of future enhancements, adopting deep neural networks with stacking ensemble holds promise due to their superior accuracy, speed of classification and capability to handle interdependent attributes.

V. REFERENCE

[1]. Kamil, A., Hassan, I., & Abraham, A. (2016). Modeling Insurance Fraud Detection Using Ensemble Combining Classification. International Journal of Computer Information Systems and Industrial Management Applications, 8, 257–265.

[2]. Zheng, H., Peng, F., Tian, Y., Zhang, Z., & Zhang, W. (2023). Insurance fraud detection based on xgboost. Academic Journal of Computing & Information Science, 6(8). https://doi.org/10.25236/ajcis.2023.060808

[3]. Deogade, K. R., Thorat, D. B., Kale, S. V., Rajput, S., & Kaur, H. (2022). Credit card fraud detection using bagging and boosting algorithm. 2022 International Conference on Signal and Information Processing (IConSIP). https://doi.org/10.1109/iconsip49665.2022.10007446

[4]. Khandare, J., & Thakur, P. (2024). Review of Credit Card Fraud Detection Using Different Machine Learning Approach. https://doi.org/10.2139/ssrn.4874339

[5]. Gupta, P., Varshney, A., Khan, M. R., Ahmed, R., Shuaib, M., & Alam, S. (2023). Unbalanced credit card fraud detection data: A machine learning-oriented comparative study of balancing techniques. Procedia Computer Science, 218, 2575–2584. https://doi.org/10.1016/j.procs.2023.01.231

[6]. Sohony, I., Pratap, R., & Nambiar, U. (2018). Ensemble learning for credit card fraud detection. ACM International Conference Proceeding Series, July, 289–294. https://doi.org/10.1145/3152494.3156815

[7]. Kaler, K., & Kaur, K. (2021). Credit card fraud detection using imbalance resampling method with feature selection. International Journal of Advanced Trends in Computer Science and Engineering, 10(3), 2061–2071. https://doi.org/10.30534/ijatcse/2021/811032021

[8]. Vosseler, A. (2022). Unsupervised insurance fraud prediction based on anomaly detector ensembles. Risks, 10(7), 132. https://doi.org/10.3390/risks10070132

[9]. Mohanta, A., & Panigrahi, S. (2023). Health Insurance Fraud Detection using feature selection and Ensemble Machine Learning Techniques. Lecture Notes in Networks and Systems, 197–207. https://doi.org/10.1007/978-981-99-1203-2_17

[10]. Kunickaitė, R., Zdanavičiūtė, M., & Krilavičius, T. (2020). Fraud detection in health insurance using ensemble learning methods. CEUR Workshop Proceedings, 2698.

[11]. Yang, J., Chen, K., Ding, K., Na, C., & Wang, M. (2022). Auto Insurance Fraud Detection with

multimodal learning. Data Intelligence, 5(2), 388–412. https://doi.org/10.1162/dint_a_00191

[12]. Johnson, J. M., & Khoshgoftaar, T. M. (2019). Deep learning and data sampling with imbalanced Big Data. 2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI). https://doi.org/10.1109/iri.2019.00038

[13]. Krawczyk, B. (2016). Learning from imbalanced data: Open Challenges and Future Directions. Progress in Artificial Intelligence, 5(4), 221–232. https://doi.org/10.1007/s13748-016-0094-0

[14]. Viaene, S., Dedene, G., & Derrig, R. A. (2005). Auto claim fraud detection using Bayesian learning neural networks. Expert Systems with Applications, 29(3), 653–666. https://doi.org/10.1016/j.eswa.2005.04.030

[15]. Piyadasa, T. D., & Gunawardana, K. (2023). A review on oversampling techniques for solving the data imbalance problem in classification. International Journal on Advances in ICT for Emerging Regions (ICTer), 16(1), 22–31. https://doi.org/10.4038/icter.v16i1.7260

[16]. Salmi, M., & Atif, D. (2022). Using a data mining approach to detect automobile insurance fraud. Proceedings of the 13th International Conference on Soft Computing and Pattern Recognition (SoCPaR 2021), 55–66. https://doi.org/10.1007/978-3-030-96302-6_5

[17]. Hassan, AKI & Abraham, A. (2016). Modelling Insurance Fraud Detection Using Ensemble Combining Classification. International Journal of Computer Information Systems and Industrial Management Applications, 8, 257–265.

[18]. Lopo, J. A., & Hartomo, K. D. (2023). Evaluating Sampling Techniques for Healthcare Insurance Fraud Detection in Imbalanced Dataset. Jurnal Ilmiah Teknik Elektro Komputer Dan Informatika, 9(2). http://dx.doi.org/10.26555/jiteki.v9i2.25929

[19]. Polak, P. B., Prusa, J. D., & Khoshgoftaar, T. M. (2024). Low-shot learning and class imbalance: a survey. Journal of Big Data, 11(1). https://doi.org/https://doi.org/10.1186/s40537-023-00851-z

[20]. Nalluri, K., & Rekha, S. (2016). Detection of Health Insurance Fraud using Bayesian Optimized XGBoost. International Institute of Electronics and Telecommunications Engineers. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/cpe.7123

[21]. Gupta, A., Rajput, I. S., Gunjan, Jain, V., & Chaurasia, S. (2022). NSGA-II-XGB: Meta-heuristic feature selection with XGBoost framework for diabetes prediction. Concurrency and Computation: Practice and Experience, 34(21). https://doi.org/10.1002/cpe.7123

[22]. Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable AI: A Review of Machine Learning Interpretability Methods. Entropy, 23(1), 18. https://doi.org/10.3390/e23010018

[23]. Bishop, C.M. (2006) Pattern Recognition and Machine Learning. Springer, Berlin.

[24]. Bergstra, J., Yamins, D., & Dasgupta, D. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13, 281-305

[25]. Verikas, A., Correia, M. P., & Lopez, F. S. (2019). A survey on ensemble learning for fraud detection. ACM Computing Surveys (CSUR), 51(5), 1-30.

[26]. Zhou, L., Li, H., & Mao, C. (2020). A survey on multi-stage ensemble learning. IEEE Transactions on Knowledge and Data Engineering, 32(12), 2886-2904

[27]. Tax, D. M. J., Van Der Goot, M. B., & Verhoef, E. M. (2016). Ensembles of decision trees or fraud detection. Knowledge and Information Systems, 49(2), 497-512.

[28]. Hancock, J., & Khoshgoftaar, T. M. (2020). Performance of CatBoost and XGBoost in Medicare Fraud Detection. Proceedings - 19th IEEE International Conference on Machine Learning and Applications, ICMLA 2020, 572–579. https://doi.org/10.1109/ICMLA51294.2020.00095

[29]. Feng, H. (2021). Ensemble learning in credit card fraud detection using boosting methods. 2021 2nd International Conference on Computing and Data Science (CDS), 12, 7–11. https://doi.org/10.1109/cds52072.2021.00009

[30]. Nur Prasasti, I. M., Dhini, A., & Laoh, E. (2020). Automobile Insurance Fraud Detection using Supervised Classifiers. 2020 International Workshop on Big Data and Information Security, IWBIS 2020, 47–51. https://doi.org/10.1109/IWBIS50925.2020.9255426

[31]. Rubaidi, Z. S., Ammar, B. Ben, & Aouicha, (2022). Fraud Detection Using Large-scale Imbalance Dataset. International Journal on Artificial Intelligence Tools, 31(8). https://doi.org/10.1142/S0218213022500373

[32]. S. Patil, K., & Godbole, A. (2018). A survey on machine learning techniques for insurance fraud prediction. HELIX, 8(6), 4358–4363. https://doi.org/10.29042/2018-4358-4363

[33]. Hancock, J. T., Bauder, R. A., Wang, H., & Khoshgoftaar, T. M. (2023). Explainable machine learning models for Medicare fraud detection. Journal of Big Data, 10(1). https://doi.org/10.1186/s40537-023-00821-5

[34]. Fernando, E.N.R. (2021). Machine Learning Approaches on Motor Insurance Fraud Detection. University of Colombo School of Computing.

[35]. Pranavi, P. S., D, sheethal H., Kumar, S. S., Kariappa, S., & H, S. B. (2020). Analysis of vehicle insurance data to detect fraud using machine learning. International Journal for Research in Applied Science and Engineering Technology, 8(7), 2033–2038. https://doi.org/10.22214/ijraset.2020.30734

[36]. Edgar, T. W., & Manz, D. O. (2017). Exploratory study. Research Methods for Cyber Security, 95–130. https://doi.org/10.1016/b978-0-12-805349-2.00004-2

[37]. Francis, C., Pepper, N., & Strong, H. (2011). Using support vector machines to detect medical fraud and abuse. 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society. https://doi.org/10.1109/iembs.2011.6092044

[38]. Muranda, C., Ali, A., & Shongwe, T. (2020). Detecting fraudulent motor insurance claims using support vector machines with adaptive synthetic sampling method. 2020 61st International Scientific Conference on Information Technology and Management Science of Riga Technical University (ITMS). https://doi.org/10.1109/itms51158.2020.9259322

[39]. Randhawa, K., Loo, C. K., Seera, M., Lim, C. P., & Nandi, A. K. (2018). Credit card fraud detection using ADABOOST and majority voting. IEEE Access, 6, 14277–14284. https://doi.org/10.1109/access.2018.2806420

[40]. Chen, T., & Guestrin, C. (2016). XGBoost. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. https://doi.org/10.1145/2939672.2939785

[41]. Mohbey, K. K., Khan, M. Z., & Indian, A. (2022). Credit Card Fraud Prediction Using XGBoost: An Ensemble Learning Approach. International Journal of Information Retrieval Research (IJIRR), 12(2), 1-17. http://doi.org/10.4018/IJIRR.299940

[42]. Xie, Y., Li, A., Gao, L., & Liu, Z. (2021). A heterogeneous ensemble learning model based on data distribution for credit card fraud detection. Wireless Communications and Mobile Computing, 2021(1). https://doi.org/10.1155/2021/2531210

[43]. Soleymanzadeh, R., Aljasim, M., Qadeer, M. W., & Kashef, R. (2022). Cyberattack and fraud detection using ensemble stacking. AI, 3(1), 22–36. https://doi.org/10.3390/ai3010002

[44]. Iqbal, A., Amin, R., Alsubaei, F. S., & Alzahrani, A. (2024). Anomaly detection in multivariate time series data using Deep Ensemble Models. PLOS ONE, 19(6). https://doi.org/10.1371/journal.pone.0303890

[45]. Ghasemieh, A., Lloyed, A., Bahrami, P., Vajar, P., & Kashef, R. (2023). A novel machine learning model with stacking ensemble learner for predicting emergency readmission of heart-disease patients. Decision Analytics Journal, 7, 100242. https://doi.org/10.1016/j.dajour.2023.100242

[46]. Azim Mim, M., Majadi, N., & Mazumder, P. (2024). A soft voting ensemble learning approach for credit card fraud detection. Heliyon, 10(3). https://doi.org/10.1016/j.heliyon.2024.e25466

[47]. Rubaidi, Z. S., Ammar, B. B., & Aouicha, M. B. (2022). Fraud detection using large-scale imbalance dataset. International Journal on Artificial Intelligence Tools, 31(08). https://doi.org/10.1142/s0218213022500373