Optimizing Healthcare Claim Fraud Detection Using Ensemble Learning Models and Modified SMOTE in Imbalanced Dataset with Application to the Egyptian Health Insurance

By

Dr. Asmaa Mohamed Saad Hussein

Lecturer of insurance

Department of Insurance and Actuarial Science

Faculty of Commerce, Cairo University

asmaa_msaad@yahoo.com

Scientific Journal for Financial and Commercial Studies and Research (SJFCSR)

Faculty of Commerce – Damietta University

Vol.6, No.2, Part 1., July 2025

APA Citation

Hussein, A. M. S. (2025). Optimizing Healthcare Claim Fraud Detection Using Ensemble Learning Models and Modified SMOTE in Imbalanced Dataset with Application to the Egyptian Health Insurance, ***Scientific Journal for Financial and Commercial Studies and Research***, Faculty of Commerce, Damietta University, 6(2)1, 1265-1294.

Website: https://cfdj.journals.ekb.eg/

Optimizing Healthcare Claim Fraud Detection Using Ensemble Learning Models and Modified SMOTE in Imbalanced Dataset with Application to the Egyptian Health Insurance

Dr. Asmaa Mohamed Saad Hussein

ABSTRACT

Healthcare fraud involves deliberately submitting inaccurate claims or distorting information to receive payment benefits. This results in the misallocation of healthcare funds and drives up overall healthcare costs. As a result, fraud represents a significant financial burden. Therefore, the machine learning Machine Learning (ML) Algorithms and Artificial Intelligence (AI) have essential role in detecting healthcare insurance fraud.

Detecting claims fraud in healthcare insurance datasets is a significant challenge due to severe class imbalance, where fraudulent cases are vastly outnumbered by legitimate ones**.** Traditional machine learning (ML) algorithms often underperform in such scenarios because they tend to favor the majority (non-fraud) class, leading to poor fraud detection rates. To address this issue, **resampling techniques**—such as **oversampling the minority class (fraud) or under sampling the majority class (non-fraud)**—are commonly employed to balance the dataset.

This study proposes a health model that helps in detecting health claim fraud in the Egyptian health market based on the **ensemble techniques**, including the XGBoost, Random Forest, bagging algorithms. This paper is the first paper uses the XGBOOST algorithm and tuning hyperparamters to optimize the performance accuracy of the classifier model especially, for the imbalanced data in the Egyptian Health Market. Additionally, this research used modified (SMOTE) algorithm to address the imbalance data issue that enhances the performance of the fraud claim detection model. This study is the first study that conducts comparison among different performance metrics, including AUC-ROC, $F1$-Score, Precision, and Recall for different ensemble classifiers models and the logistic regression in Egyptian Health Insurance Data. The findings show the effectiveness of SMOTE in building a more robust model for detecting the health claims frauds and prevention Techniques, reducing potential losses and enhancing the overall performance. Moreover, the ensemble learning technique greatly outperforms single learning algorithms (Logistic Regression) based on different performance metrics.

Key Words: SMOTE, Health claim fraud, imbalance Data, Oversampling, Under sampling Techniques, Area Under the Curve (AUC).

1. Introduction

The health care claims management and the financial sustainability of healthcare payers and providers are seriously affected by healthcare fraud. According to the NHACC, it is defined as the "deception or intentional misrepresentation that the person or entity makes and could result in an unauthorized benefit for the person, entity, or another part" ***(NHCAA 2018).***

Fraud in health insurance can take place at various points throughout the insurance process. It may involve individuals applying for coverage, third-party claimants, or even current policyholders. This type of fraud can drive up healthcare costs, reduce the quality of medical services and facilities, and result in significant financial losses.

Detecting insurance fraud is a challenging task, as fraud can be committed by a wide range of people, regardless of their education or profession. Many risk management solutions use rules to identify potential fraud, and some even evolve by learning from past cases. However, as fraudulent methods evolve, these systems often fail to detect new types of fraud. ***(Ekin et al.., 2021)***

Several models are utilizing Artificial Intelligence, Machine Learning, and deep learning algorithms to detect potential health insurance fraud. In the field of medical Binary classification models find extensive use. Binary classification is frequently employed in medical diagnostics, such as determining whether a patient requires surgery. Other applications of binary classification involve identifying cancers, diagnosing diabetes, forecasting heart disease risks, and testing for infections, among others. ***(Nabrawi& Alanazi , 2023)***

The Egyptian healthcare system currently exhibits a complex organizational structure. However, aligned with the national Vision 2030 strategy and the 2018 healthcare reform legislation, the system is undergoing significant transformation.

Egypt's healthcare claims management market is projected to increase from $75.76 Mn in 2022 to $443.76 Mn by 2030. Egypt's healthcare cost 4.81% of GDP, or $150 per person, in 2019. The improvement of healthcare and population health is a top priority for the Egyptian government in line with the Sustainable Development Goals (SDGs) of the UN.

The healthcare claims management market in Egypt is experiencing rapid growth due to increasing demand for healthcare services, rising costs, and the need for efficient claim management. Healthcare providers and insurance companies in Egypt are under pressure to reduce costs while maintaining high-quality care, which has driven the need for innovative solutions to streamline the claims process.

The health claims fraud is the most significant factor that limits the growth of the healthcare claims management market in Egyptian health system. The fraud detection involves two steps: **Firstly**, **identification the possible model** of fraud integrated with known fraudsters, and, **secondly prediction the probability** of a new transaction is being fraudulent. ***(Acharjya & Kun Ma, 2024)***

One of the challenging in the health care insurance is detecting claims fraud due to highly class imbalance between fraud cases and non-fraud cases. Health care datasets associated with healthcare perdition modeling or risk classification to detect the health fraud is imbalanced in nature. An imbalanced dataset refers to a classification dataset where one class is dominated significantly the other class. **In other words**, in imbalanced data, the minority classes have high weight, importance, and contain additional informative predictive relative to the majority ones.

Generally, classification problems associated with data imbalance occur in many fields and have long been studied using the machine learning Techniques, including fraud detection for credit card users , customer churn prediction and diagnosis prediction for rare diseases suffer from the imbalance data issue. ***(Ekin & Conversano,2021)***

ML algorithm has an essential tool for optimizing the performance accuracy of classifiers from imbalanced data, which has attracted a significant amount of interest in recent years, specifically, in medical diagnosis and health care.

The traditional ML algorithms: the Decision trees, Logistic regression, and ensemble models are performed poorly for detecting the health claims frauds due to imbalanced data of health care. As a result, many methodologies have been proposed to enhance prediction performance when developing prediction models using imbalanced data.

It is clear that the health care prediction models or risk classification using the imbalanced data may lead to biased towards the majority class and may be more likely to misclassify the smaller class the minority class. ***(Chen et el.., 2022)***

The common strategy to address imbalanced datasets is to either increase the number of samples in the minority class **(oversampling)** or reduce the number of samples in the majority class **(under sampling).** Approaches to handle class imbalance are generally grouped into four main categories.

The **first** is the **algorithm-level approach**, where adjustments are made to the learning algorithm itself to better recognize the minority class. This method tweaks the learning process to reduce bias toward the majority class without modifying the original training data. ***(Jain, 2020)***

The **second** is the **data-level approach**, which focuses on adjusting the dataset directly. Techniques such as oversampling and under-sampling are used here to balance the distribution between classes.

The **third** approach is **cost-sensitive learning**, where different costs are assigned to classification errors, particularly penalizing errors on the minority class more heavily. This encourages the model to pay more attention to the underrepresented class during training.

The **fourth** is the use of **ensemble methods**, such as bagging, where multiple classifiers are combined to improve overall performance, especially in the context of imbalanced datasets. ***(Mohammed et al .., 2020)***

The most common different sampling methods: **are Random Under sampling, Random Oversampling examples (ROSE), SMOTE** (**Synthetic Minority Over-sampling Technique), and Instance Hardness Threshold.**

Random Oversampling is the common method involves instead keep all of the data, and oversampling (either generate more of or duplicate/sample with replacement) the positive class observations. Again, this provides balanced target classes, which will improve model performance.

Under Random sampling methods is Instead of using the full dataset, and under samples the majority class and keep all of the positive class (minority) observations. This means we end up with balanced classes when training our model, so it should better is able to pick up the signal leading to a claim occurrence***. (Vijay Singh et al.., 2024)***

To address this issue, SMOTE technique is developed, leading to an increased the accuracy of the model performance. SMOTE is a popular approach to the construction of classifiers from imbalanced datasets. SMOTE creates new samples from the minority class through interpolation in which new data points are added within the range of known observations. In this way new samples are different from the original ones and therefore it avoids duplication.

Currently, the linear sampling algorithm built on SMOTE is the most effective resampling technique. However, SMOTE technique has some drawbacks as follows***: (Yang et al.., 2024)***

(1) Inability to fit the distribution characteristics of the data set, (2) large variability in the sampling results since the selection of hyper parameters in the sampling performance. (3) Finally, Synthetic data points are generated without fully considering the underlying distribution of the minority class. This can lead to unrealistic examples that may not accurately represent the true characteristics of the data, potentially leading to over fitting or poor model generalization.

There are many types of Modified SMOTE. This study involves using the **SMOTE-Tomek Links** method and its applications using Python package, where this approach integrating over-sampling method from SMOTE and the under- sampling method from Tomek Links. The algorithms of SMOTE-Tomek Links will be discussed in the next section in details. ***(Khamesiana et al..,2022)***

Finally, many packages are developed to facilitate SMOTE calculation (i.e DMwR and caret Package , which provide a specific function to aid the estimation of a classifier in the presence of class imbalance, in addition to extensive tools for data mining problems (among others, functions to compute evaluation metrics as well as different accuracy estimators). In addition, selecting and validation regression and classification problems. ***(Nitesh et al.., 2002)***

2. Research Problem

Health care fraud Analytics market size is predicted to increase by USD 49.64 billion between 2025 and 2037. The industry size of the health care fraud analytics in 2025 is predicted to reach USD 3.7 billion.

The health claim fraud detection is a challenge for achieving universal Health Coverage in the Egyptian health system. According to Reinsurance Group of America, "at least 2% of the health claims in the Egyptian health system were declined due to fraud and abuse, as well additional 16 % reported that as many as 10 % of claims are declined". Fraud schemes in the Egyptian system include overbilling, billing for the services not rendered and the submission of fabricated documentations. From fabricated claims to unnecessary medical procedures, identifying fraudulent activity in the healthcare landscape can be a daunting task. Fraudulent claims, whether from patients, healthcare providers, or intermediaries, result in substantial financial losses, which significantly impact the cost management for the Egyptian health system.

Healthcare insurance datasets have a highly imbalanced class distribution, with fraudulent claims making up a very small percentage of the total claims. This results in poor performance of machine learning models, as they tend to be biased toward the majority class (non-fraudulent claims). In other words, Traditional fraud detection methods using single machine learning algorithms (e.g., Logistic Regression) often fail to accurately classify fraudulent claims, due to the class imbalance. ***(Shamitha & Ilango, 2020)***

Fraud detection is a common imbalanced classification Model, therefore, there is an urgent and critical need to develop a more effective, robust, and accurate fraud detection models specifically handling the class imbalance in healthcare claim data within the Egyptian context. Such models must not only achieve high predictive performance but also provide reliable identification of the minority fraudulent class, thereby mitigating financial losses and enhancing the integrity of the health insurance system.

This paper aims to develop a more effective and robust fraud detection model in the Egyptian health insurance market by proposing and evaluating advanced ensemble learning techniques combined with Modified resampling methodologies to optimize fraud detection accuracy in the Egyptian health insurance market.

3. Literature Review

Health claim fraud detection is evident by using the Machine Learning techniques in the area that have received a great concern from the academic disciplines such as: The health Economics, Technology and the information system, and the Data science.

Many previous studies concern on the claim Fraud detection using the supervised and unsupervised learning techniques and the deep learning, which are evidenced by Lu & Boritz (2005) on Canada; Kirlidog & Asuk (2012) and Kose, Gokturk & Kilic (2015) on Türkiye; Kumar, Ghani & Mei (2010) and Ekin et al. (2013) on the USA; Sun et al. (2020) on China; and Ortega, Figueroa & Ru (2006) on Chile.

All these studies shed on the advantages of using the Machine learning and artificial intelligence algorithms for the health claim detection, improving the efficiency, the performance accuracy and reducing the health care costs. However, all these previous studies ignored the health data imbalance, assuming the health data is balanced data.

On the other hand, the classification modeling is often encountered with an imbalanced dataset issue; this is evident that Fraud detection is a common imbalanced classification in which ensemble algorithms have been used to improve the model performance. ***The following studies investigated detecting the health care claims fraud in imbalanced dataset to build a more robust model:***

- **Shamitha & Ilango, 2020,** this research sheds light on a framework in detecting Health claims fraud with integration ML algorithms. Additionally, Data imbalanced classification of classes had also been taken into considerations in this study. To achieve this purpose, the principal components Analysis as a one of unsupervised Learning Technique to reduce the data dimensions and evaluate the effectiveness of the **SMOTE oversampling technique**, performance metrics such as **Artificial Neural Networks (ANNs), Precision, Recall, and the Area under the Curve (AUC)** were employed.

- **Severino and Peng, 2021** assessed different ML techniques to detect fraud in Brazilian insurance claims data. Comparing 9 models and the RF algorithm as ensemble technique achieved the most predictive performance. This model has a potential role in operational risk management techniques because of depending on actual data. Moreover, this model can be used as a probabilistic guide in predicting whether a policy will result in claims fraud.

- **Kaddi and Patil** , 2023, this research investigates a health model for healthcare claim fraud detection based on ensemble learning with SMOTE technique. Stack ensemble learning algorithm performance is used and compared to the following Supervised Leaning Techniques: Support vector classifier (SVC), Logistic Regression (LR), and decision tree model algorithm. The findings showed that for the detecting the healthcare claim fraud, the ensemble learning Technique is superior single learning algorithms and provides the most robust model.

From the previous papers, there is not any paper concern optimization detecting the Egyptian health claims fraud in the imbalanced data. This research is the first to propose a health model for optimizing the detection of healthcare fraud in the Egyptian health system to achieve a more robust model.

The main contribution of this paper can be summarized as follows :

(1) thorough research focusing on the most challenging issue in the classification model in the health insurance, which is imbalanced dataset;

(2) Comprehensive algorithms suitable for typical datasets collected in Egyptian health insurance companies to Detect claim health fraud.

(3) This paper provides a proposed model that incorporates the Ensemble techniques with tuning hyper parameters into in order to investigate the effect of the proposed model algorithms on achieving better classifier performance metrics based on (AUC-ROC), Precision, Recall, and F1-score, to compare the robustness of the different classifier models.

(4) Using the XGBoost algorithm and cross Validation for hyperparamters tuning in Egyptian health data in order to optimize model accuracy, particularly in the highly imbalanced dataset.

4. Research Objective

This paper proposes an a health model using an ensemble learning Techniques for detecting healthcare claim fraud in the Egyptian market in imbalanced data combining with the TONEK LINK SMOTE sampling to optimize performance accuracy of the model particularly, for the imbalanced data.

To achieve the objective, Different classifier models are taken into considerations: **logistics regression, random forest, bagging, and XGBOOST** to detect fraud with optimal accuracy and good evaluation metrics.

Due to the imbalance of the target variable, before applying the algorithms, TTONEK LINK SMOTE is used to create balanced datasets effectively that helps in detecting claims frauds the Egyptian health insurance companies to achieve better classifier performance based on ROC AUC than only under-sampling the majority class.

Additionally, other evaluation metrics, including, Accuracy, Specificity, Precision-score, and Recall, to compare the performance accuracy the classifier Logistic regression model with ensemble classifier models .

5. Research importance

Health claim fraud detection is important for the following parties:

1) **Insurance Companies**, through the following :

- Mitigating Financial Losses that direct payouts on fraudulent claims, protecting profitability.
- Preventing Premium Inflation: Curbs the need to raise premiums for honest policyholders.
- Enhancing Operational Efficiency through Frees up resources from investigating fraud to process legitimate claims faster.
- Reducing Investigation Costs: Efficient health claims detection minimizes the resources spent on manual investigations.

2) **For the Government, through the following** :

- Preserving Public Healthcare Funds that Protects resources allocated to the Universal Health Insurance System
- Supporting Universal Healthcare Goals that ensures the financial sustainability.
- Controlling Overall Healthcare Costs that Prevents fraud from contributing to inflation in the healthcare sector.
- Reducing Legal and Regulatory Burden that minimizes the need for extensive fraud investigations and enforcement.

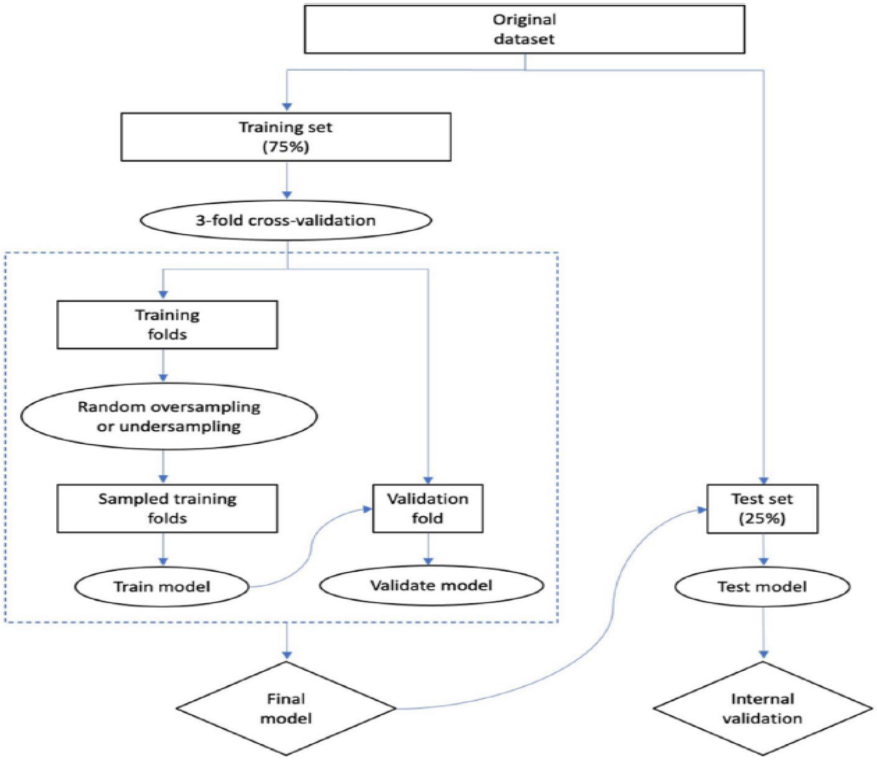

Figure (1): the stages of the research preprocess data and model research

Source: By the Researcher

▪ Data and methods

6.1 DATA

This section illustrates dataset, the methods, the algorithms and strategies in getting the results are provided. The algorithms are coded in Python and R programming languages. The data prepressing and stages of the model research is shown in the following figure. Validation data set can be achieved by removing Outliers and missing values are deleted.

The data is obtained from the Egyptian medical insurance company. Dataset contains **three parts**: The insurance policy, the claim application and the details of the risk assessment. The claim for reimbursement is defined as binary variable: indicating the occurrence/non-occurrence of the claim. Payment Claim Period **01/01/2023 to 23/12/2024.** The total number of observations in this data set is **26**, with **21** features.

The dataset consists of two distinct categories: the **majority class (Class 0)**, containing 24,394 observations, and the **minority class (Class 1)**, with only 1,836 instances. Class 0 corresponds to non-fraudulent cases (negative instances), whereas Class 1 indicates fraudulent claims (positive instances). A significant class imbalance exists, with the negative class making up **94%** of the data and the positive class representing just **6%**. For model evaluation, the dataset was split into **training (75%) and testing (25%)** respectively.

Using stratified sampling techniques can ensure accurate representative percentage for training and testing sets respectively. **The following figure shows the distribution of both classes:**

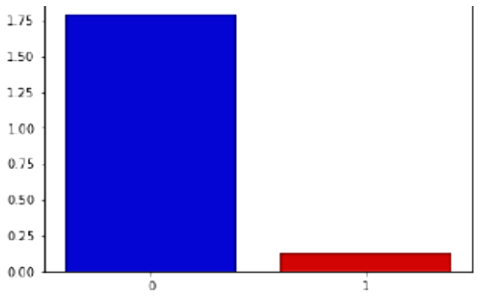

Figure (2): Distribution of both class imbalances in the dataset.

Source: R Algorithms output

6.2 STATISTICAL ANALYSIS METHODOLGY

This study used a Health dataset with the logistic model (Single supervised learning model) and Ensemble Learning : random forest, Bagging and XGBOOST algorithm with tuning hyperparamters to optimize the performance accuracy of the model algorithms and alongside data resampling techniques Modified SMOTE, Additionally, $F1$ -Score, Precision, and Recall as a model performance metrics approach is proposed to measure the model performance among different classifier models based on the receiver operating characteristic curve (AUROC).

6.2.1.1 The Synthetic Minority Over-sampling Technique SMOTE

In this section, we introduce the (SMOTE) as is a powerful method used to handle class imbalance in datasets. SMOTE creates new samples from the minority (positive) class through interpolation in which new data points are added within the range of known observations. In this technique new samples are different from the original ones and therefore it avoids duplication. **The steps of the SMOTE Algorithms can be summarized as follows:** ***(Gayeong &Byeon, 2023)***

1. Identify Minority Class Instances: SMOTE operates on datasets where one or more classes are significantly dominated compared to others. The first step is to identify the minority class or classes in the dataset.

2. Nearest Neighbor Selection: For each minority class observation (xi), SMOTE calculates its k-nearest neighbors within the feature space, where k represents a user-defined hyperparamters. This neighborhood determination employs distance metrics (typically Euclidean) to establish proximity dissimilarity relationships.

3. Synthetic Sample Generation: For each minority class instance, SMOTE randomly chooses one of its k nearest neighbors. It then generates synthetic samples along the line segment joining the minority class instance and the selected nearest neighbor in the feature space.

4. **Controlled Oversampling**: The synthesis process involves:
   - Random selection of one neighbor (xj) from the k-nearest set
   - Linear interpolation between xi and xj in feature space
   - Generation of new instances along the connecting vector:
     $$x\_new = xi + \lambda(xj - xi)$$
     where $\lambda \in (0,1)$ is a random **interpolation weight.**

5. **Iterate Through All Minority Class Instances:** Steps 2 to 4 are performed for every instance in the minority class within the dataset, creating synthetic samples to enhance the representation of the minority class.

6. **Achieve a Balanced Dataset:** Once the synthetic samples for the minority class have been generated, the final dataset achieves better balance, with a more even distribution of instances across all classes.

SMOTE balances the classes without considering the natural class reparability, potentially making the decision boundary unrealistic, and SMOTE generated samples may end up overlapping with the majority class , which may lead to misclassification.

6.2.1.2 Tomek Link SMOTE

In this approach one modifies the distribution of the data in a way to obtain either more observations in the minority class (over-sampling) or less data point in the majority class (under-sampling). In fact, Tomek-Link approach yields similar ratio for each class. This approach results in a better decision boundary for a classifier. In this method, we believe that the observations on the boundary are noises. The **Tomek-Link** technique

combines the capabilities of SMOTE to create synthetic data for the minority class with the functionality of **Tomek Link**, which removes data points identified as **Tomek-link** from the majority class. These **Tomek- links** refer to instances in the majority class that are nearest neighbors to instances in the minority class. This approach effectively balances the dataset by augmenting the minority class and cleaning up ambiguous or overlapping majority class samples.

The mathematical algorithms for the TOMEK TECHNIQUE can be as follows :

- Let $d(x_i, x_j)$ denotes the Euclidean distance between x_i, and x_j, where x_i, denotes sample that belongs to the minority class and x_j denotes sample that belongs to the majority class. If there is no sample x_k satisfies the following properties:

1. $d(x_i,, xk) < d(x_i, x_j)$.
2. $d(x_j, x_k) < d(x_i, x_j)$, then the pair of (x_i, x_j) is a **Tomek Link**.

Advantages of Tomek SMOTE OVER SMOTE can be concluded as follows: (***Zain Hamid et al… , 2024)***

- Minimizing **class overlap following the application of SMOTE**, in order to create clearer boundaries between the classes.
- Enhancing **data quality by eliminating Tomek links**, this helps improve the classifier's accuracy by removing borderline or noisy examples.

6.2.2 Ensemble Learning

The objective of the ensemble learning is building the block models that can often produce a much more powerful model. Thus, the concept of ensemble methods is reducing bias and/or variance by combining several of them together in order to achieve better performance and avoid over fitting. This technique has potential in the presence of noise, and better handling of the datasets in the case of class imbalance. ***(Dwi Hartomo & Agili Lopo, 2023)***

These simple building block models are sometimes known as weak learners, since they may lead to mediocre predictions on their own by aggregating many decision trees, using methods like bagging, random forests, and boosting, the predictive performance of trees can be substantially improved. (***James et al …, 2023)***

In this section, the researcher discusses the most common techniques for the Ensemble Learning Techniques: Bagging, random forests, boosting, and eXtreme Gradient Boosting (XGBOOST). These are ensemble methods can be used a regression or a classification tree.

6.2.2.1 Bagging

Bagging or aggregation bootstrapping, is a general-purpose procedure for reducing the variance (the model complexity) or improve the test accuracy of a model; by averaging the prediction of multiple base trees models. In other words, calculating the predictions $\hat{f}^{1}(x)$, $\hat{f}^{2}(x), \ldots\ldots\ldots\ldots\ldots$ $\hat{f}^{B}(x)$, using B separate training data and averaging them to obtain the less correlated base trees reducing the variance of the model , given by the following equation : ***(James et al.. , 2023)***

$$\hat{f}^{Ave}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{b}(x), \qquad \textbf{(equ.1)}$$

For the classification problem, and a given test observation, we can calculate the class predicted by each of the B trees, and take a majority vote for a given cut off ,Thus, the overall prediction is the most commonly occurring majority class among the B predictions.

6.2.2.2 The random forest

Random forests provide an improvement over bagged trees by way of a random decorrelates the trees. As in bagging, we build a number of decision trees on bootstrapped training samples. But when building these decision trees, each time a split in a tree is considered, a random sample of **m predictors**, which is treated as hyperparamters, which is chosen carefully as split candidates from the full set of **p predictors**.

The split is allowed to use only one of those m predictors that are taken at each split, and typically we choose m ≈ the squared root of the number of predictors considered at each split in the case of the classification tree. While, for the regression tree , choosing m = p/3.

6.2.2.3 Boosting

Recall that bagging involves using the bootstrapped sampled training data of based tree models, fitting a separate decision tree model, and then combining all of the trees in order to get a single predictive model. Notably, each tree is built on a bootstrap data set, independent of the other trees.

Boosting works in a similar way, except that the trees are grown based on iterative manner: each tree is grown using information from previously grown trees and reducing the bias.

Boosting is a supervised learning Techniques that combines a large number of simple models, called weak learners, into a final model with Superior predictive performance. Boosting does not take into consideration bootstrap sampling; instead each tree is fit on a modified version of the original data set.

In each iteration, each tree is fit to the residuals of the preceding tree and subtract scaled down version of the current's tree prediction from the residuals to form new residuals. Finally, output the boosted model, the overall prediction is***(Feng & Ambrose Lo, 2025) :***

$$\widehat{f}\ (x) = \ \sum_{b=1}^{B} \lambda \widehat{f^b}(x). \qquad \textbf{(equ.2)}$$

From the above, we can conclude **the hyper parameters of the boosting are the following :** ***(Zhang et al.. ,2020)***

1. **Number of Trees or Iterations**: Unlike bagging and random forests, boosting is prone to over fitting if the number of iterations (B) becomes excessively large.
2. **Shrinkage Parameter (Learning Rate λ)**: This is a small positive value that regulates the speed at which the boosting algorithm learns. Commonly used values for λ are 0.01 or 0.001, though the best choice often depends on the specific problem. A very small learning rate may necessitate a significantly larger value of B to achieve satisfactory performance.
3. **Number of Splits per Tree (d)**: This parameter determines the complexity of each individual tree in the boosted ensemble.

In the next section, the researcher shows **XGBoost** technique, which is the proper Ensemble technique. To improve the performance accuracy of the model, reducing the complexity, especially for the imbalanced data.

6.2.2.4 Extreme Gradient Boosting (XGBoost)

This method is a widely used approach in fraud detection due to its ability to manage class imbalance, which can lead to over fitting model if not addressed during training. At step *n*, the prediction for each learner is determined by the following equation:

$$\textbf{Prediction} = \widehat{Y} = \sum_{I=1}^{N} \mathbf{f}_{K\,(X_I)}, \qquad \textbf{(equ.3)}$$

Where,

f_k represents the base tree model, and $x_i$denotes the input features.

Secondly, measuring the performance of each learner *L*, XGBoost uses

$$L=\sum(\widehat{Y}i,\mathrm{Y}i)+\sum\gamma(fk) \qquad \textbf{(equ.4)}$$

Where,

- α**: is the Loss function.**
- γ**: the Regularization Penalty term.** ***(Rimanı̇ et al…, 2020)***

Finally, optimization the XGBoost algorithms. This paper applies hyper parameter tuning, which is set before running the predictive model, the hyperparamters tuning can be summarized in the following table:

Table (1) : The Hyperparamters Testing value of XGBoost algorithms

	Definition	**Impact**
Lambda and Alpha	L2 **(Ridge term):** it Penalizes the squared magnitude of the weights, promoting smaller weights.. L1 **(Lasso term):** it Penalizes the absolute value of weights, enforcing the coefficients toward zero.	Increasing lambda and alpha strengthens model shrinkage, helping to prevent over fitting by penalizing larger weight values.
Gamma	The minimum loss reduction to make a further split.	Larger values mean fewer splits, reducing model complexity.
Min-Child-Weight	The minimum sum of instance weight required in a child.	Higher values ensure splits only happen when there's sufficient data, reducing complexity.
Tuning alpha and lambda	L1 Penalty term on the bias, reducing the magnitude of features. L2 Penalty term on the bias, removing non-important variables.	Larger values will result in a more simple and interpretable model, which reduces complexity.
Max features	Maximum number of features	It has potential to tradeoff between variance and Bias, which results in improving the accuracy performance for the model.
Max_depth	Maximum depth of a tree.	It limits the depth of the decision trees. A greater depth allows the model to fit more complex relationships, but may also result in over fitting if the trees become more complex.

Source : By the researcher

In this paper , the researcher used a real-world of Egyptian Health Insurance Claim dataset with the Ensembles and XGBoost algorithms utilized with Tomek SMOTE, then using the K-fold cross validation for hyperparamters tuning of the model to increase the robustness of model performance, especially in the highly imbalanced dataset. Comparing the results with another ensemble models (Random Forest with SMOTE-Tomek Links method). The **procedures of these approaches can be classified in the following:**

Step 1: The following procedure describes how to assess the model performance of the XGBoost algorithm with tuning hyperparamters:

Input:

• Maximum number of n trees (T).

• Hyperparamters θi and values vi, a vector containing the hyperparamters to be tuned and the values to be tried, where i = 1,…,n.

• Training ($Xtrain$,) and validation ($Xval$,) data.

Steps for each trial:

1. Fit the model using training data: $Xtrain$,$ytrain$
2. Evaluate the model using **10 fold cross validation data**: $\hat{y}val$.
3. Predict and assess the performance of the model on multiple metrics.
4. Store the hyperparamters using the vector argument for this trial: ((t).

Step 2: Applying the algorithms for the random forest model With SMOTE-Tomek Links method using python coding.

6.3 The performance Measures for Model Evaluation

The performance of a binary classification model to distinguish between classes, and the best hyperparamters can be obtained from optimization the performance metric (High Scores). The performance of the classification model typically is assessed by a confusion matrix as illustrated in the following table:

Table (3) : The confusion metric for the classification model

	Predicted Negative	Predicted Positive
Actual Negative	TN	FP
Actual Positive	FN	TP

Source: Netsh et al…,2020

In the confusion matrix, TN is the number of negative observations Predicted Negative, FP is the number of negative eobservations incorrectly classified as positive (**False Positives**), FN is the number of positive observations incorrectly classified as negative **(False Negatives**) and TP is the number of positive observations correctly classified (**True Positives**). Predictive accuracy is the performance measure generally associated with machine Learning algorithms and is he proportion of the true observations and the total number of the cases :

$$\textbf{Accuracy} = \frac{TN+TP}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}. \qquad \textbf{(equ.5)}$$

The following terminology should be investigated in the confusion matrix:

(1) Error rate is the proportion of false observations (both false positives and false negatives) and the total number of cases.

$$\textbf{The error rate} = \frac{FN+FP}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}. \qquad \textbf{(equ.6)}$$

However, In the presence of imbalanced datasets with unequal error costs, it is more appropriate to use **the ROC curve** , known as the ”Receiver Operating Characteristics” is a plot that demonstrates the ability of a binary classification model to distinguish between classes. The ROC curve is a performance measure that can be used to examine the following such as cutoff, sensitivity and specificity. The ROC curve plots the True Positive Rate (TPR) versus the False Positive Rate (FNR). The vertical and horizontal axis of the curve indicates sensitivity and its (1-specificity) respectively.

(2) The true positive rate (**Sensitivity**) and the false negative rate (**Specificity**). The sensitivity of the classifier is the ration between the observations that correctly identified positive and the actual positive.

(3) **Both of them can be calculated in the following equations:**

$$\textbf{Sensitivity} = \frac{TP}{TP+FN} \qquad \textbf{(equ.7)}$$

$$\textbf{Specificity} = \frac{TN}{TN+FP} \qquad \textbf{(equ.8)}$$

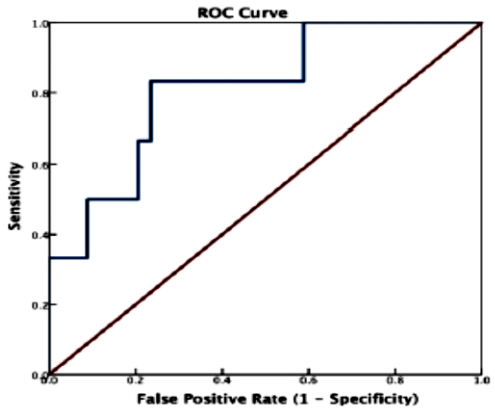

Figure (3) : Receiver Operating Characteristics (ROC) Curve

Source : James etal., 2023

As shown in the above figure, the ROC Curve is divided into three parts as follows:

(1) **Above the baselineline:** There are points in this area that have a higher sensitivity than soecificty The more ROC Curve approaches the top left of the plot, the better model perfor mance and more reliable results will be achieved (**AUC=1**).

(2) **On the baseline line**: In this area, the numerical values of the true positive rate (TPR) and the false positive rate (FPR) are equal. This indicates the model predicts the classes by chance (**AUC=0**).

(3) **Below the baseline line:** There are points in this area whose sensitivity is lower than specificity When the rock curve (ROC Curve) is below the baseline line (AUC=0.5), the model has a very poor performance. ***AUC can be caiculated based on thr following equaton :***

$$\mathbf{AUC} = \frac{sensitivity + specificity}{2} \qquad \textbf{(equ.9)}$$

Accuracy may be a falsely optimistic measure. Therefore, using the F1 score as a safeguard is helpful to balance precision and recall.F1-Score provides a single measure that is the average of the recall and precision scores.The percision is ameasure of the proportion of correctly made positive predictions to all positively observed. **The following equation shows the calculaton of the F1 measure:**

$$\mathbf{F= 2\ x\ \frac{Precision\ x\ Recall}{Precision\ +\ Recall}} \qquad \textbf{(equ.10)}$$

7. The Findings

In this research, the classification Supervised Leaning model (logistic regression) and ensemble model (random forest, Bagging, and X Boosting) are used to investigate the health claims fraud detections in the Egyptian health market. The dataset is highly imbalanced and requires intervention as shown in **Figure (2)**. Therefore, the synthetic minority oversampling technique (SMOTE) was applied. **Tomek Link SMOTE** technique has the ability to mitigate the class imbalance issue and provided a more representative training dataset **as shown in the following figure** , as a result, The dataset is resampled, effectively generating synthetic instances of the minority class (fraudulent claims) to balance with the majority class (non-fraudulent claims) as shown in the following figure:

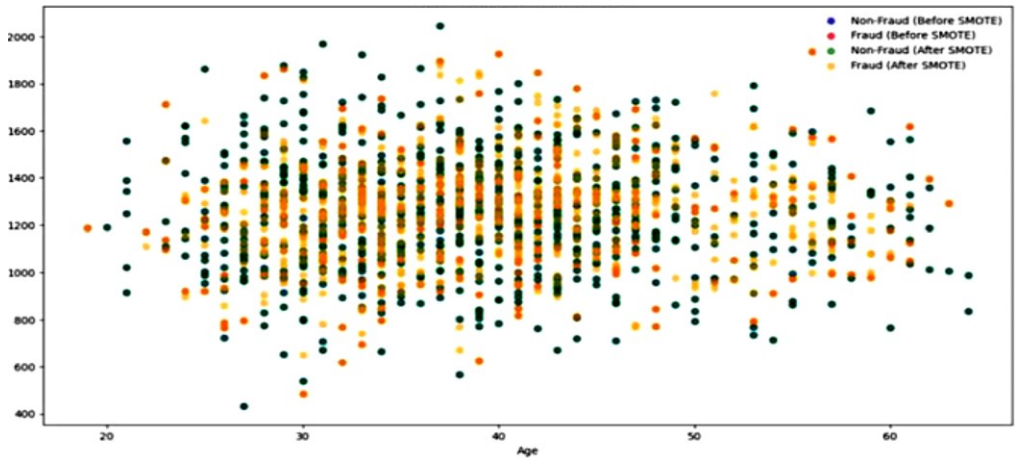

Figure (4): Data Set before and after Tomek Link SMOTE

Source: output of R Package Algorithms

The following Figure shows the trend for the positive accuracy and the negative accuracy based on SMOTE techniques with integration into varying degree of under-sampling of the majority class that is tuned by **Cross Validation**. The Y-axis represents the accuracy and the X-axis represents the percentage of majority class under-sampled.

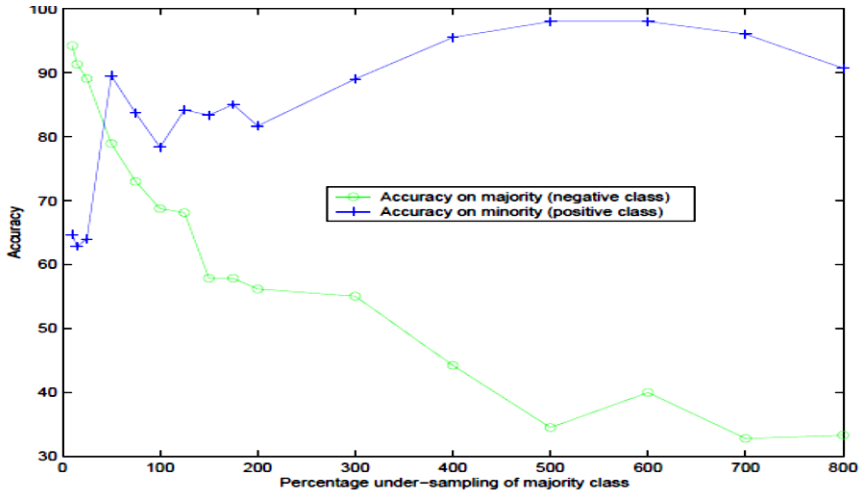

Figure (5): Accuracy Distribution comparison between minority and majority classes

Source: output of R Package Algorithms

The following tables illustrate the accuracy of performance Matrix of each classification model before applying SMOTE and after applying SMOTE based on AUC-ROC.

Table (4): Performance of different classification models before applying Tomek Link SMOTE

The classifier Model	Accuracy (%)	ROC-AUC (%)	Recall (%)	Precision (%)	Specificity (%)	F1-Score (%)
LR	**66.2**	**69.5**	**76**	**81.2**	**70**	**83**
RF	**88**	**75**	**82**	**88**	**75**	**82.65**
BAGGING	**80**	**73**	**79**	**84**	**73**	**84**
XGBoost	**86**	**63.2**	**69.4**	**74.8**	**68**	**79**

Table (5): Performance of different classification models after applying Tomek Link SMOTE

The classifier Model	ROC-AUC (%)	Accuracy (%)	Recall (%)	Precision (%)	Specificity (%)	F1-Score (%)
LR	**78**	**70**	**78.6**	**83**	**97.6**	**88**
RF	**82**	**96**	**97**	**100**	**98**	**99**
BAGGING	**79**	**94**	**95**	**99**	**97.8**	**98**
XGBoost	**76.5**	**95**	**71**	**81**	**93**	**96**

Source: output of Python Package Algorithms

From the above tables, the SMOTE technique has significantly improved the performance accuracy of the classification model, particularly, the random Forest and bagging in predicting accurately the minority class. In other words, it revealed a notable improvement in model performance, which illustrated enhanced accuracy in predicting the minority class.

The random forest has superior in the performance accuracy compared to the other classifier models based on the SMOTE technique. Moreover, hyperparamters are set as number of estimators (**=350**), maximum features (=**sqrt of predictors**) and maximum depth **(=6).**

The following figures can provide the best understanding of the performance accuracy between the random forest and the logistic regression models after applying the SMOTE techniques.

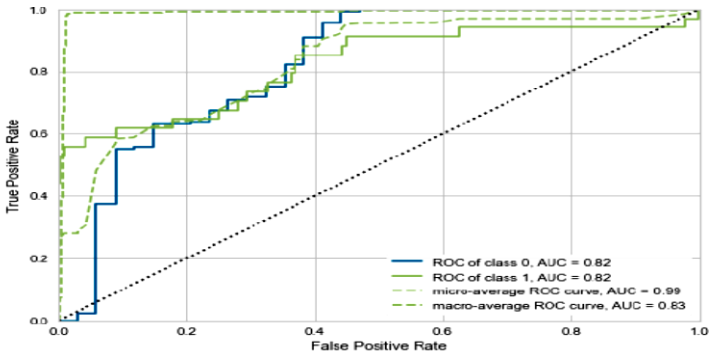

Figure (6): ROC CURVE For Random Forest Classifier

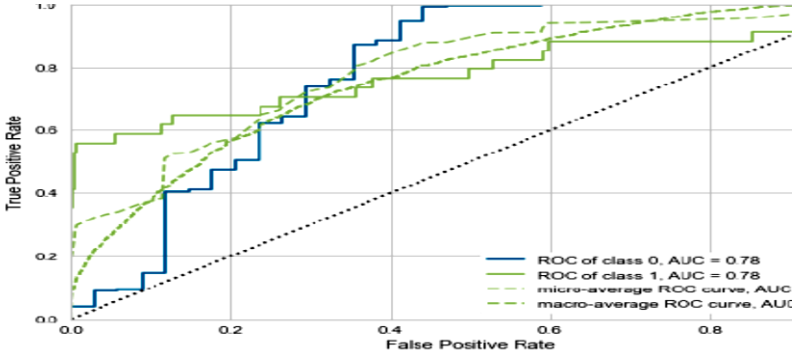

Figure (7): ROC CURVE For Logistic Regression (LR) Classifier

Source: output of Python Package Algorithms

Additionally, the following figure compares the performance of logistic regression and random forest with respect to the ROC AUC performance metric, illustrating that the random forest model as a one of ensemble model has achieved the best accuracy performance compared to logistic regression classifier.

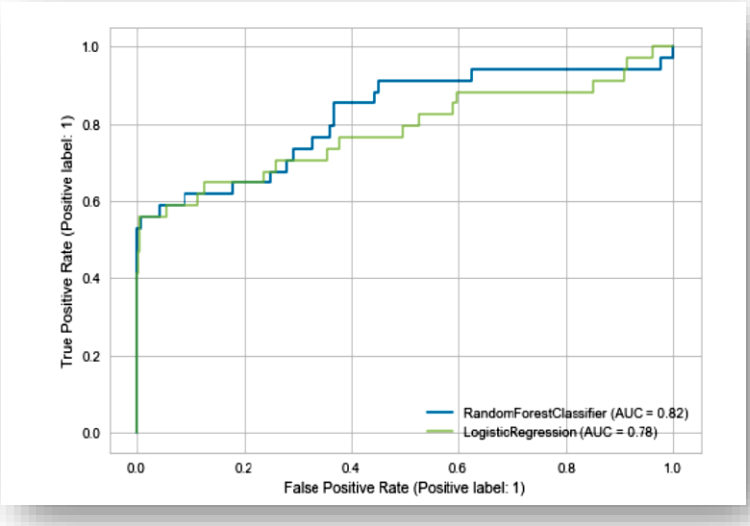

Figure (8): Comparison of logistic regression and random forest

Source: output of Python Package Algorithms

8. CONCLUSION

Healthcare insurance fraud has been depleting medical finances, but conventional, manual fraud detection methods require time and effort. ML and deep learning Algorithms offer a practical, cost-effective solution that detects healthcare insurance fraud effectively and automates the fraud detection process with minimal human intervention. This study should form the basis for future Health claims fraud detection investigations in Egyptian Market. It focuses on building a more robust model that aims to detect fraud in healthcare claims and determine significant factors that contribute to healthcare fraud. Additionally, this study provides insights on the health fraudulent acts in Egyptian health system.

In this paper, a robust model is proposed that leverages ensemble learning techniques for the detection of fraudulent healthcare claims. The performance of various ensemble methods is investigated and compared with traditional classification algorithms, such as Logistic Regression (LR). Due to the data

is imbalanced, one of resampling techniques is used, this paper will focus on using the **SMOTE-Tomek Links method**, where this method combines oversampling method from SMOTE and the under sampling method from Tomek Links.

The results indicate that the ensemble learning approach achieves superior performance compared to Single machine learning models. Notably, Random Forest (RF) and Extreme Gradient Boosting (XGBoost) algorithms exhibited superior predictive accuracy for minority class identification in healthcare claim fraud detection in the Egyptian health care System.

Furthermore, the application of the Synthetic Minority Over-sampling Technique (SMOTE) significantly improved model performance by addressing class imbalance issues in the dataset models' performance as shown from AUC-ROC curve for different classifier models, ie comparative performance accuracy among single supervised learning technique (LR) and ensemble learning algorithms. This emphasizes the importance of tackling class imbalance and highlights the effectiveness of SMOTE in boosting model performance and providing a more robust model. Additionally, the proposed model showed that the **Age** has the highest Score (highest importance), followed by **number of claims** and claim **amounts.**

9. Recommendation

Based on the above results, some recommendation should be taken into consideration:

- **Incorporate Comprehensive Features**: It is recommended to expand the feature set used in health claim fraud detection models to encompass all relevant aspects that may not have been covered in the current research. By incorporating a wider range of features, the model will be more robust, capturing all possible dimensions of fraud detection in health claims by using Data mining.
- **Extend with Advanced SMOTE Variants**: To enhance the model's ability to handle imbalanced datasets, it is suggested to explore the use of advanced versions of SMOTE (Synthetic Minority Over-sampling Technique), such as Distance-based SMOTE (D-SMOTE) and Bi-phasic SMOTE (BP-SMOTE). These techniques, combined with selective classifiers, can improve prediction accuracy and help in the detection of health claim fraud by generating more representative synthetic data.

- **Adopt Deep Learning Techniques:** Integrating deep learning methods, particularly Artificial Neural Networks (ANN), should be prioritized for future studies. These techniques can potentially optimize the model's performance by learning complex patterns within large and noisy datasets, which are commonly encountered in health claim fraud detection.
- **Enhance Sampling Mechanisms:** It is recommended to improve the current sampling mechanisms to better balance noisy datasets. This can be achieved by generating high-quality synthetic data and developing adaptive, robust models that can cope with data quality issues, thus improving fraud detection capabilities.
- **The proposed model can be guide for developing more robust model for detecting claims fraud model for imbalanced health insurance data sets.**
- **Collaboration for Universal Health Insurance in Egyptian Market :** To support the goal of achieving universal health insurance in Egypt and simultaneously reduce healthcare costs, it is crucial to foster collaboration between regulators, healthcare providers, and insurance companies. By working together, these entities can help build more accurate and efficient machine learning models for health claim fraud detection, ensuring the sustainability of the healthcare system.

References

1. Acharjya D. P. & Kun Ma. (2024). ***"Computational Intelligence in Healthcare Informatics"***. Spinger -Studies in Computational Intelligence.
2. Alrais, A. I. (2022). **"*Fraudulent Insurance Claims Detection Using Machine Learning".*** Degree of Master of Science in Professional Studies: Data ANALYTICS.
3. Boritz, F. L. (2005).***" Dtecting faud in HealthInsurance Data : LEaning to Model Incomplete Benford's LAw Distributions".*** *Machine Learning : ECML* (pp. 633-640). Springer.
4. Chen, Z.; Duan, J.; Kang, L.; Qiu, G. (2022). "***Class-Imbalanced Deep Learning via a Class-Balanced Ensemble"***. *IEEE Trans. Neural Network Learn. System.*, Volume 33, 5626–5640.
5. Cynthia Yang , Egill A. Fridgeirsson, Jan A. Kors, Jenna M. Reps and Peter R. Rijnbeek . (2024). ***"Impact of Random Over Sampling and random undersampling on the performance of prediction models developed using observational health data"***. *Journal of Big Data*, Springer.
6. Divakar K. and K. Chitharanjan. (2019). ***"Performance evaluation of credit card fraud transactions using boosting algorithms".*** *Int. J. Electron. Commun. Comput. Eng. IJECCE*, 10 (6) , 262–270.
7. Ekin, T.,Frigau,L., & Conversano, C. (2021**) ."Health care Fraud classifier in practice"** . *Applied Stochastic Models in Business and industry* , 37(6) ,1182-1199.
8. Eman Nabrawi and Abdullah Alanazi.(2023). "***Fraud Detection in Healthcare Insurance Claims Using Machine Learning".*** *isk Journal* .
9. Gareth James ,Daniela Witten, and Robert Tibshiran . (2023). ***"An Introduction to Statistical Learning with Applications in R"***. Second Edition: Society of Actuaries .
10. Gayeong Eom and HaewonByeon. (2023). "***Searching for Optimal Oversampling to Process Imbalanced Data: Generative Adversarial Networks and Synthetic Minority Over-Sampling Technique"***. *Mathematics Journal* .
11. Jain, R. M. (2020). ***"Handling imbalanced data using ensemble learning in software defect pediction".*** *10th International Conference on Cloud Computing,* (pp. 300-304). Data Science & Engineering(Confluence), IEEE.
12. Justine Power, Marie-Pier Côté & Thierry Duchesne. (2024). "***A Flexible Hierarchical Insurance Claims Model with Gradient Boosting and Copulas".*** *North American Actuarial Journal*, 28(4), 772–800.
13. Khamesiana F., M. Esna-Asharia, E. D. Ofosu-Heneb and F. Khanizadeha,. (2022). ***"Risk Classification of Imbalanced Data for Car Insurance Companies: Machine Learning Approaches".*** *International Journal of*

Mathematical Modelling & Computations, Vol. 12, No. 03, Summer 2022, 153- 162.

14. Kose,I., Gokturk,M.,Kilic,K. (2015***). "An Interactive Machine-Learning Based Electronic Fraud and abuse Detection System in Health Care Insurance"*** . *Applied Soft Insurance*, 36, 283-299.
15. Kristoko Dwi Hartomo & Joanito Agili Lopo. (2023). ***"Evaluating Sampling Techniques for Healthcare Insurance Fraud"*** . *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)* .
16. Mary Sowjanya and Owk Mrudula. (2023). *"**Effective treatment of imbalanced datasets in health care using modified SMOTE coupled with stacked deep learning algorithms"**. Applied Nanoscience*, 13:1829–1840 .
17. Melih Kirlidog & Cuneyt Asuk . (2012). ***"A Fraud Detection Approach with data Mining in health insurance"*** . *Procedia -Social and behavioral Science* , 989-994.
18. Menshchikov , A., V. Perfilev, D. Roenko, M. Zykin, and M. Fedosenko. (2022). "***Comparative analysis of achine Learning methods application for financial faud detection*****"** . *32nd Conference of open Innovations Assocoation* (pp. 178-186). FRUCT.
19. Mohammed R., J. Rawashdeh, and M. Abdullah .(2020). ***"Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results".*** 11th International Conference on Information and communication System (pp. 243–248, 2020, https://doi.org/10.1109/ICICS49469.2020.239556.). ICICS .
20. NHCAA. (2018). ***"The Problem of Health Care Fraud: A Serious and Costly Reality for All Americans".*** National Health Care Anti-Fraud. *National Health Care Anti-Fraud (NHCAA)*, http://www.nhcaa.org/resources/health-care-anti-fraud-resources/the-challenge-of.
21. Nitesh V. Chawla , Kevin W. Bowyer, Lawrence O. Hall & W. Philip Kegelmeyer. (2002). ***"SMOTE: Synthetic Minority Over-sampling Technique"***. *Journal of Artificial Intelligence Research* , 321–357.
22. Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall and W. Philip Kegelmeyer. (2022). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, Volume-16 , 321–357.
23. Pedro A. Ortega , Cristian J. Figgueroa & Gonzalo .A. RUZ. (2006). ***"A medical claim fraud/abuse dtection System based om data mining : Acase Study in Chile"***. International Confrence of Data Mining (pp. 224-231). Las Vegas , USA: DBLP.
24. Priscilla, C. V. and Prabha, D. P. . (2020). ***"Influence of optimizing XGBoost to handle class imbalance in credit card fraud detection".*** *Third*

International Conference on Smart Systems and Inventive Technology (ICSSIT), (pp. 1309-1315). IEEE.

25. Rimanɨ eKunickaiɨ ea , Monika Zdanavičiuɨ eaband & Tomas Krilavičiusa. (2020). ***"Fraud Detectionin Health Insurance Using Ensemble Learning Methods".*** *Information SocietyandUniversityStudies*.
26. Runhuan Feng & Ambrose Lo. (2025**). "*ACTEX Learning SRM Exam Manual"*** . 8th Edition: SOA.
27. Shamitha K. and V. Ilango. (2020). ***"A time-effeicent model for detecting graudulent health Insurance claims using Artificial nueal networks".*** *international confrence of system computer networkings* (pp. 1-6). ICSCAN.
28. Shweta S. Kaddi & Malini M. Patil . (2023). ***"Ensemble learning based health care claim fraud detection in an imbalance data environment"*** . *Indonesian Journal of Electrical Engineering and Computer Science*.
29. Sobanadevi, V. and Ravi, .G. (2021). "**Handling data imbalance using a heterogeneous bagging-based stacked ensemble (HBSE) for credit card fraud detection"**. *Intelligence in Big Data Technologies Springer*, 517–525.
30. Sun,Y., Zhang ,J., Li , H & Luo, X. (2020). "**Combating health insurance faud: Adeep Leraning Approach"** .*Journal of health care Engineering* .
31. Tian C., L. Zhou, S. Zhang, and Y. Zhao. (2020). "***A New Majority Weighted Minority Oversampling Technique for Classification of Imbalanced Datasets"***. *Int. Conf. Big Data, Artif. Intell. Internet Things Engineering* (pp. pp. 154–157). ICBAIE.
32. Vijay Singh Rathore, Joao Manuel R. S. Tavares, & B. Surendiran. (2024). "***Universal Threats in Expert Applications and Solutions".*** Springer: Proceedings of 3rd.
33. YI DENG & MINGYONG LI , (2023). **"An Adaptive and Robust Method for Oriented Oversampling With Spatial Information for Imbalanced Noisy Datasets**". *IEEE Journal of Biomedical and Health Information* , VOLUME 11.
34. Yuxuan Yang, Hadi Akbarzadeh Khorshidi, and Uwe Aickelin. (2024). **"A review on over-sampling techniques in classification of multi-class imbalanced data sets : Insights for Medical Problems"** . *Frontier in Digital Health* .
35. Zain Hamid,Fatima Khalique, Saba Mahmood, Ali Daud, Amal Bukhari and Bader Alshemaimri . (2024). **"Healthcare insurance fraud detection using data mining"**. *BMC Medical Informatics and Decision Making* .

تعظيم كشف الاحتيال في المطالبات الصحية باستخدام نماذج Ensemble Learning ونموذج SMOTE المعدلة في حالة البيانات غير المتوازنة مع التطبيق على بيانات التأمين الصحي في السوق المصري

الملخص

يعد الاحتيال في الرعاية الصحية من القضايا البارزة التي تؤثر سلبًا على كفاءة تخصيص الموارد، حيث يتمثل في تقديم مطالبات زائفة أو تحريف متعمد للمعلومات بهدف الحصول على مزايا مالية غير مستحقة. يُسفر ذلك عن هدر كبير في الإنفاق الصحي وزيادة التكاليف العامة للرعاية الصحية، مما يجعله عبئًا ماليًا حقيقيًا على المنظومة . تلعب تقنيات التعلم الآلي (ML) والذكاء الاصطناعي (AI)دورًا محوريًا في كشف الاحتيال الطبي، إلا أن عملية اكتشاف المطالبات الاحتيالية تظل تحديًا كبيرًا نظرًا لعدم توازن الفئات في البيانات، حيث تعاني النماذج التقليدية للتعلم الآلي من ضعف في الأداء في مثل هذه السيناريوهات، إذ تميل إلى الانحياز للفئة الأكبر (غير الاحتيالية)، ما يؤدي إلى معدلات كشف منخفضة.

وللتغلب هذه المشكلة، تُستخدم تقنيات إعادة أخذ العينات بشكل شائع، مثل زيادة عينات الفئة النادرة (الاحتيال) أو تقليل عينات الفئة السائدة (غير الاحتيالية) بهدف تحقيق التوازن في البيانات. تقترح هذه الدراسة نموذجًا قائمًا على تقنيات التعلم التجميعي(Ensemble Learning) ، يتضمن نماذج تنبؤية مثل XGBoost ،RANDOM FOREST، لاكتشاف الاحتيال في المطالبات الصحية ضمن السوق المصري، حيث تُعد هذه الدراسة الأولى من نوعها في استخدام نموذج XGBoost مع ضبط المعاملات الفائقة (Hyper parameter Tuning) لتحسين دقة اكتشاف الاحتيال في البيانات غير المتوازنة في قطاع التأمين الصحي المصري. كما تم تطبيق نموذج SMOTE المُعدل لمعالجة عدم توازن البيانات، مما ساهم في تعزيز أداء نموذج الكشف.

بالإضافة إلى ذلك، تُعد هذه الدراسة أول محاولة لإجراء مقارنة شاملة بين مؤشرات الأداء المختلفة، بما في ذلك منحنىAUC-ROC ، ومعامل1F ، والدقة(Precision) ، والحساسية (Sensitivity) ، عبر نماذج التجميع المختلفة والانحدار اللوجستي. وقد أظهرت النتائج فعالية أسلوب SMOTE في بناء نموذج أكثر كفاءة وقوة في اكتشاف المطالبات الاحتيالية، كما أظهرت النماذج التجميعية تفوقًا ملحوظًا على نموذج الانحدار اللوجستي، سواء من حيث مقاييس الأداء أو دقة التنبؤ.

الكلمات الدالة :الاحتيال في المطالبات الصحية، البيانات غير متجانسة ، نظام التأمين الصحي ، المساحة تحت المنحني.