








Detection of Health Insurance Fraud using Bayesian Optimized XGBoost

Saravanan Parthasarathy^{*}, Arun Raj Lakshminarayanan, A. Abdul Azeez Khan, K. Javubar Sathick,
Vaishnavi Jayaraman

B. S. Abdur Rahman Crescent Institute of Science and Technology, Vandalur, Chennai 600048, Tamil Nadu, India

Corresponding Author Email: saravanan_cse_2019@crescent.education

<https://doi.org/10.18280/ijss.130509>

ABSTRACT

Received: 2 March 2023

Revised: 21 September 2023

Accepted: 13 October 2023

Available online: 10 November 2023

Keywords:

crime prediction, fraud detection, health insurance, insurance fraud, machine learning, XGBoost, Bayesian optimization

The mounting prevalence of health insurance fraud, propelled by a myriad of socioeconomic factors, presents significant hurdles to insurers, healthcare institutions, and individuals. In an attempt to counter this, insurance companies have begun harnessing the power of advanced technology, utilizing Machine Learning models to distinguish legitimate from fraudulent claims within expansive datasets. The present study conducts an in-depth examination of a health insurance dataset comprising 517,737 records, employing the Extreme Gradient Boosting (XGBoost) model as a potent tool for the detection of deceptive claims. In a noteworthy development, the performance of the model is markedly amplified through the integration of Bayesian optimization techniques, culminating in the Bayesian Optimized XGBoost (BOXGBoost) Model. The BOXGBoost Model is meticulously evaluated against an array of algorithms, which include Naive Bayes, Logistic Regression, Random Forest, K-Nearest Neighbor, and AdaBoost. A comparative analysis, focusing on key performance metrics such as accuracy, precision, recall, F1-Score, and the Area Under the Curve (AUC), is undertaken to discern the most effective algorithm. Remarkably, the proposed BOXGBoost model emerges as the superior performer, achieving an impressive accuracy rate of 98% and an AUC of 0.994. Additionally, the model exhibits high precision (98%), recall (97%), and F1-Score (97.5%), highlighting its exceptional capability in the prediction of health insurance fraud.

1. INTRODUCTION

The healthcare sector is intricately intertwined with government institutions, insurance companies, research institutions, and the public at large. The escalating financial investment into technological and scientific advancements by healthcare providers is directly linked to soaring treatment costs. Consequently, health insurance policies are increasingly being acquired by individuals as a countermeasure to these escalating expenses. These policies serve to protect against the financial strain associated with disease onset. However, the health insurance sector remains highly susceptible to fraudulent claims. As reported by the National Health Care Anti-Fraud Association (NHCAA), the United States experiences an annual financial impact of approximately \$68 billion due to health insurance fraud [1]. Some experts postulate that the true figure could significantly surpass the current estimate. These fraudulent activities primarily inflict financial losses upon insurance companies. In response, these companies often resort to raising the premiums of policies or incorporating new restrictions on treatments to recoup losses. Consequently, such deceptive activities indirectly exert a widespread impact on the general population, manifesting in increased healthcare costs and limitations on available treatments.

Fraudulent claims, potentially instigated by policyholders, healthcare providers, or third parties, manifest in diverse forms within the healthcare sector. These include medical bill fabrication, unbundling, upcoding, identity theft, collusion, drug diversion, kickbacks, multiple card usage, and eligibility

violations [2]. Medical bill fabrication, a widespread practice, involves the inflation or construction of fictitious bills to claim reimbursement for non-existent services. Upcoding is another prevalent approach, where billing codes are manipulated to overstate the severity of a disease or procedure, thereby maximizing insurance payouts. Identity theft refers to the unauthorized use of another's identity to secure medical services or insurance benefits fraudulently. Collusion between healthcare providers and patients typically results in overcharges for services or the prescription of unnecessary treatments. The illegal redirection of prescription medications is termed drug diversion. Kickbacks, or the illicit exchange of payments for referrals or services, can precipitate cost increases. The exploitation of insurance benefits through the use of multiple cards for the same patient, alongside eligibility violations involving fraudulent claims for non-insured services, further exacerbate the strain on the healthcare ecosystem. Individuals also contribute to this issue through false billing, claims manipulation, and submission of multiple claims for identical treatments. Healthcare providers are implicated in medical insurance fraud as well, charging insurance companies for non-performed treatments, superfluous procedures, and conditions not covered by policies [3]. Adding to these concerns, patient data theft by hackers who subsequently sell the stolen information to organized criminal groups is becoming increasingly prevalent. Such groups perpetrate insurance fraud through the use of false identities coupled with manipulated claims.

The methodologies and mechanisms employed for healthcare data collection and storage are manifold [4]. With

the ongoing technological advancements, such processes are becoming increasingly sophisticated, addressing their inherent complexities. Efforts are being made by researchers to thwart fraudsters, with the development of secure healthcare data environments being a key strategy [5, 6]. Additionally, Blockchain Technology is being deployed to safeguard patient records [7]. The healthcare and insurance sectors are generating vast amounts of data, the codification of which is beyond human capability. Traditional, manual methods of fraud detection are proving inadequate, plagued by issues of scalability, inefficiency, and delayed detection. In contrast, Machine Learning (ML) offers a scalable solution, with its inherent ability for pattern recognition, consistency, and efficiency, as well as continuous improvement in fraudulent claim identification. This contributes substantially towards a more secure and transparent healthcare ecosystem. Machine Learning approaches swiftly discern common properties from multiple attributes across various datasets. These approaches facilitate comparative analyses involving government regulations, existing transactions, healthcare provider histories, and policyholder credibility. Comprehensive reports about outliers are generated and provided to the authorities. Presently, researchers are exploring the development of a real-time scam prediction model, with the potential to identify fraudulent claims immediately upon their filing. This ongoing work in Machine Learning methodologies offers a glimmer of hope to healthcare and insurance providers, suggesting that health insurance fraud could be significantly mitigated in the foreseeable future.

This research endeavor introduces a methodology for predicting fraudulent health insurance claims using an optimized Extreme Gradient Boosting (XGBoost) method. The incorporation of the Bayesian hyperparameter optimization algorithm enables the identification of the XGBoost algorithm's hyperparameters. The primary focus of the proposed methodology is the classification of authentic and fraudulent transactions. The results derived from the proposed model are evaluated and contrasted with those obtained from other Machine Learning techniques, utilizing various performance metrics.

2. RELATED WORK

Nalluri et al. [8] employed a set of ML algorithms, including Support Vector Machines (SVM), Decision Trees (DT), Random Forest (RF), and Multilayer Perceptron (MLP), to address the critical issue of medical insurance fraud. The primary goal of the study was to identify the most effective machine learning method for this task. MLP demonstrated superior performance in terms of accuracy but had the longest training time. In contrast, DT exhibited the shortest training time while achieving the second-best classification performance. MHAMFD, a novel health insurance fraud detection model that used an attributed heterogeneous information network (AHIN) to capture patient behavioral relationships across multiple visits [9]. MHAMFD incorporates a multilevel attention mechanism, outperforming existing methods in accuracy. This approach highlights the importance of considering patient behavioral relationships for effective healthcare fraud detection and efficient resource utilization. Lopo and Hartomo [10] analyzed the challenge of detecting healthcare insurance fraud in datasets with imbalanced cases. XGBoost model along with various

sampling methods, such as Random Oversampling and Undersampling were employed. Key features, like costs and diagnosis codes, are identified as crucial for accurate fraud detection. The health insurance claim documents are always filled with structured and unstructured data.

Farbmacher et al. [11] classified the German-based insurance data and drew out meaningful information. The proposed model identified the possible duplicitous activities better than existing models. It also delivered the perceptions about the outliers. Naidoo and Marivate [12] proposed an approach to ascertain anomalies using the concept of a Generative Adversarial Network (GAN). The model had been tested against two different datasets. The logistic regression and XGBoost algorithms yielded the best predictions. Singh and Urolagin [13] analyzed a health insurance dataset and evaluated the possibilities of rejections. The results indicated that KNN performed better than other methods by conceding 97% of accuracy, 92.5% of ROC, and 92.6% of F1-Score. Kapadiya et al. [14] published a thorough investigation on the identification of Health Insurance fraud and related security vulnerabilities in Health Insurance Claims. In addition, they developed a four-layer design for an intelligent HI fraud detection system. A comparative study was conducted by Rukhsar et al. [15] to detect insurance fraud. Among the eight machine learning algorithms used, the decision tree fared the best with an accuracy of 79%. In another study, CatBoost outperforms LightGBM by achieving an average Area Under the Curve (AUC) of 0.77452 [16]. The dataset obtained from Ardabil's Social Security Insurance Organization was used to determine fraud detection in health insurance by Parnian et al. [17]. The K-nearest neighbor classification along with squirrel optimization methodology outperformed the other employed models by acquiring an accuracy of 98.8%.

The Medicare program in the USA was abused by some policyholders and medical service providers. Since the program covered millions of people, it was very complex to locate the deceptions. Johnson and Khoshgoftaar [18] designed a neural network-based model for Medicare fraud detection. In that system, random oversampling, and random undersampling (ROS - RUS) methods were engaged to handle the imbalanced data. This study resolved the issue of data imbalance and helped health insurance providers to detect the swindles. Castaneda et al. [19] implemented a maxout network for discovering the deceits in medical insurance claims. The outcome of the same was compared with other activation functions like ReLU, LReLU, SeLU, and tanh. Since SeLU performed the operations faster than maxout with a ratio of 2.3 times, the latter one was meant to be the slowest activation function of all. Bauder et al. [20] applied unsupervised Machine Learning methods to ascertain fraudulent claims. Local Outlier Factor delivered a better outcome by producing 0.6298 of AUC, 0.5362 of Sensitivity, and 0.6768 of Specificity. Pandey et al. [21] availed statistical methods along with Machine Learning practices to envisage deceptions. Since the Neural Network based model returned a high ROC value, the authors recommended the same for fraud prediction.

Kareem et al. [22] propounded a Support Vector Machine algorithm-based framework for detecting duplicitous claims. Lasaga and Santhana [23] availed Restricted Boltzmann Machines to predict frauds related to overtreatment. The RBM approach resulted in AUCs of 0.95. Meanwhile, the model was validated with the manipulated dataset; the cogency of the outcome is questionable. Bauder and Khoshgoftaar [24] anticipated the incident of insurance fraud using the Naive

Bayes algorithm. The model was boosted with an 80-20 sampling technique and diagnosed the cross-functional transactions based on the specialty of the physician. Ahmadinejad et al. [25] offered an unsupervised model which proffered 96.97% accuracy in predicting the impostures. Bayerstadler et al. [26] have constructed a Bayesian Multinomial Latent Variable model to predict the intrigues. It worked based on the scorecard method and produced the AUCs as 0.85. A multi-stage methodology was postulated by Johnson and Nagarur [27] to capture the insurance-related scams. The anticipated model predicted the shams with an accuracy rate of 86%, which is better than the existing Neural Network equipped approaches. The studies frequently employ validation procedures utilizing synthetic datasets, which may not fully capture the intricacies of genuine fraud occurrences. Consequently, legitimate concerns may emerge regarding the accuracy and validity of the reported findings. It is pertinent to acknowledge that the real-world landscape of insurance claims entails the efficient and real-time processing of vast data volumes. In this regard, certain investigations within the literature may not comprehensively address the scalability aspects inherent to their proposed models, warranting further scholarly consideration. Consequently, we proposed a scalable Health Insurance Fraud Prediction (HIFP) method that makes use of a Bayesian Optimized XGBoost Model to address the challenges. The fraud prediction would be helpful to reduce the financial liabilities of the insurance companies and soothe the claiming process in real time.

3. PROPOSED WORK

This section is categorized into two subdivisions. Section 3.1 describes the nature of the dataset; Section 3.2 defines the proposed Bayesian Optimized XGBoost Model and Section 3.3 delineates the Performance Metrics used for the evaluation of the proposed model.

3.1 Dataset

A healthcare provider fraud detection dataset in Kaggle had been utilized in this study to predict the artifices in the health insurance claims [28]. The original patient dataset (D1) had more than half a million rows under 27 attributes. The other one with provider information (D2) contains two attributes with 5410 entries. An attribute named 'Fraud' was newly introduced in dataset D1 to indicate the genuineness of the healthcare providers. It was filled with binary values by considering the provider information available in dataset D2.

As a pre-processing measure, the alphanumeric characters were transformed into numerical values. The Claim Start date and Claim End date columns were separated into the date, month, and year columns. All the parameters have been converted into numeric values and the 'Null' values were replaced with '0'. Since the ClmProcedureCode_5 and ClmProcedureCode_6 contained only '0', the two columns were removed from the dataset. The finalized HIFP Dataset contains 517737 rows under 30 attributes. The outcome of the models could be improved by choosing veracious features [29]. The SelectKBest method is usually employed to identify the top features by considering the best variance [30, 31]. The f_{classif} and K values had been interpolated to find the significance of each feature in the dataset. It calculated the ANOVA F-value and ranked the attributes based on the

specified K value. In this study, the K value was set as 10 and the attributes were selected accordingly. The finalized attributes are listed in Table 1.

Table 1. Attributes of the dataset

Attribute Name	Description	Data Type
Bene ID	Beneficiary identification number	Integer
Provider	Health care provider identification number	Integer
Attending physician	Attending physician	Integer
Other physician	Other physician	Integer
Clm admit diagnosis code	Claim admitted diagnostic code	Integer
ClmProcedureCode_3	Claim procedure code 3	Integer
ClmProcedureCode_4	Claim procedure code 4	Integer
ClmDiagnosisCode_2	Claim diagnostic code 2	Integer
ClmDiagnosisCode_4	Claim diagnostic code 4	Integer
CS year	Claim start year	Integer
Fraud	Genuineness status of provider	Integer

3.1.1 Correlation

The association amongst the attributes and representation of a linear relationship is demarcated using correlation. The correlation lies between -1 and +1, where -1, 0, and +1 indicate the perfect negative correlation, no correlation, and perfect positive correlation respectively. The coefficient of correlation can be calibrated mathematically using the following formula.

$$r = \frac{n \sum(xy) - (\sum x) (\sum y)}{\sqrt{n \sum x^2 - (\sum x)^2} \sqrt{n \sum y^2 - (\sum y)^2}} \quad (1)$$

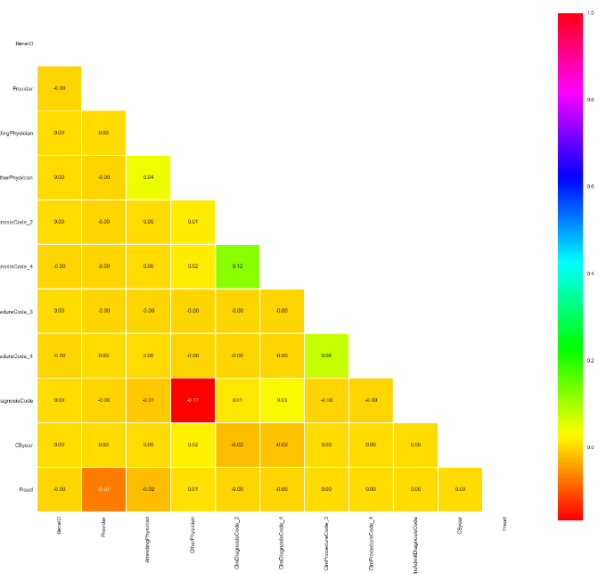


Figure 1. Post feature selection - the correlation of attributes in HIFP dataset

Figure 1 represents the correlation among post-feature selection attributes of the HIFP Dataset. The correlation between ClmDiagnosisCode_4 and ClmDiagnosisCode_2 is 0.12. The ClmProcedureCode_4 and ClmProcedureCode_3 have a correlation of 0.08. A negative correlation is found between Clm Admit Diagnosis Code and Other Physician (-0.17). The Fraud and Provider attributes have also had a negative correlation of -0.07.

3.2 Proposed BOXGBoost model

In this study, the XGBoost algorithm had been proposed to ascertain the deceitful health insurance claims. The XGBoost algorithm establishes the parallelization of tree construction and handles the missing values effectively. In the proposed approach, the hyperparameters of the XGBoost algorithm were optimized by the Bayesian-based method. The optimization techniques were usually employed to enhance the performance of the engineering applications [32]. Later, the same methodologies have been consumed in computer systems development. In the data mining operations, better models would be identified by appraising various algorithms. However, there is always a scope to improve the competency of the base models. The outcomes of ML models were mostly determined by the hyperparameters [33]. The researchers tussle to improvise the abilities of algorithms by availing the hyperparameterization methodologies. The hyperparameters could either be tuned by manual or automated approaches. Since manual tuning is a time taking process, automated methods like Random search, Grid search, Bayesian, and Tree-structured Parzen estimators were employed to ease the process.

Kotthoff et al. [34] upgraded the dexterity of the WEKA tool and proposed a new artifice called Auto-WEKA 2.0. The Auto-WEKA is equipped with the Bayesian optimization technique. Therefore, the package has regularly been updated as the improvisation happened in Bayesian optimization [35]. Bernard et al. [36] parameterized the Random Forest algorithm for feature selection. Subsequently, the algorithm measured the relevancy of each feature while execution, and the optimum level attribute selection helped to progress the prediction capacity of the model. Probst et al. [37] tried to improvise the performance of the Random Forest by parameter tuning. However, the Random Forest algorithm responded to the hyperparameterization methods with minimal improvement compared to the other algorithms.

Fauzan and Murfi [38] employed the XGBoost model to predict insurance claims. A 6-stage grid search scheme had been commissioned to optimize the hyperparameters. The optimized model performed better compared to the default one. Wang et al. [39] identified a better combination of features from the dataset using the feature selection methods. When the XGBoost model was equipped with the Bayesian Tree-structured Parzen Estimator (TPE) technique, it imparted a better outcome. A novel hyperparameter tuning algorithm named MeSH had been accoutered with XGBoost by Sommer et al. [40]. The collective approach was appraised with multiple datasets and produced better outcomes. Li et al. [41] tuned the hyperparameters of the XGBoost algorithm to predict the Gene Expression Value. In this case, the inversely proportionate relationship between 'n_estimators' and absolute error was reaffirmed. The hyper tuning of XGBoost resulted in better prediction of Gene Expression Values. Zhou et al. [42] exerted the XGBoost algorithm to predict the advance rate of a boring machine. The algorithm was trialed with default and Bayesian Optimized custom parameters. The proposed model demonstrated progression in predicting the advanced rate.

3.2.1 BOXGBoost algorithm

Input: Health insurance claims dataset as D , and hyperparameters as Hp

Parameters: First order derivative $P(x, y)$,

Second order derivative as $P'(x, y)$

Base Learners as m

Left and right split as l and r respectively

Negative accuracy as z

Negative accuracy threshold as z^*

Gaussian mixtures as $q(x)$ and $v(x)$

Output: $F(x)$

1. Get the dataset

2. Divide dataset as Training and testing data

3. Init $F_0(x)$ with a constant.

4. For 1 to i :

4.1. Calculate $P = \frac{\partial L(y, F)}{\partial F}$ and $P' = \frac{\partial^2 L(y, f)}{\partial f^2}$

4.2. Evaluate the maximum gain with the correct

split:

$$G = \frac{1}{2} \left[\frac{P_l^2}{P'_l} + \frac{P_r^2}{P'_r} - \frac{P^2}{P'} \right]$$

4.3. Compute the base learners, m :

$$\hat{m}(x) = \underset{m}{\operatorname{argmin}} \sum_d m(x)P + \frac{1}{2} m^2(x)P'$$

4.4. Determine F_i , by bumping iteratively, $F_i =$

$F_0 + \hat{m}(x)$

5. End For

6. Result : $F(x)$

7. Bayesian_optimisation($Train_{data}$, $Test_{data}$, Hp_{list}):

8. $Hp_{best} = []$

8.1. For each Hp_{set} in Hp_{list} :

8.1.1. Set a surrogate probability model for the objective function, $p(z|Hp)$

8.1.2. Create a selection function:

$$p(Hp|z) = \begin{cases} q(x) & \text{if } z < z^* \\ v(x) & \text{if } z \geq z^* \end{cases}$$

8.1.3. Identify the Hp_{best} that Maximized Expected Improvement:

$$MEI_{z^*}(x) = \frac{\gamma z^* q(x) - q(x) \int_{-\infty}^{z^*} p(z) dz}{\gamma q(x) + (1 - \gamma)v(x)} \propto \left(\gamma + \frac{v(x)}{q(x)}(1 - \gamma) \right)$$

8.1.4. Substitute $MEI_{z^*}(x)$ in $p(z|Hp)$ and record the score

8.1.5. Update the surrogate model with Bayes' theorem

9. End For

10. $best_Hp_{set} = Hp_{list}[\max_index(Hp_{best})]$

11. $best_XGBoost_model =$

$train_model(Train_{data}, append(Test_{data}), best_Hp_{set})$

11.1 return ($best_Hp_{set}$, $best_XGBoost_model$)

12. End_BOXGBoost Algorithm

Figure 2 denotes the flow of the proposed approach. The XGBoost algorithm usually delivered better predictions by engaging the well-versed splitting techniques. The Bayesian optimization approach had been utilized in this study to tune the hyperparameters of the XGBoost algorithm. The `colsample_bytree`, `gamma`, `max_depth`, `min_child_weight`, `reg_alpha`, and `reg_lambda` are the hyperparameters tuned with the help of the Bayesian model. "colsample_bytree" represents the subsample ratio of the columns, "gamma" determines the minimum split loss of leaves per tree, "max_depth" signifies the maximum depth of the tree, "min_child_weight" connotes the minimum instance weight while partitioning, "reg_alpha" and "reg_lambda" values

denote L1 and L2 regularization terms on weight. As a boosting algorithm, the XGBoost is sequentially creating the trees. The consequent trees are being built up by learning the residual errors of the predecessors. The model would be fit by the gradient of loss created in the previous step.

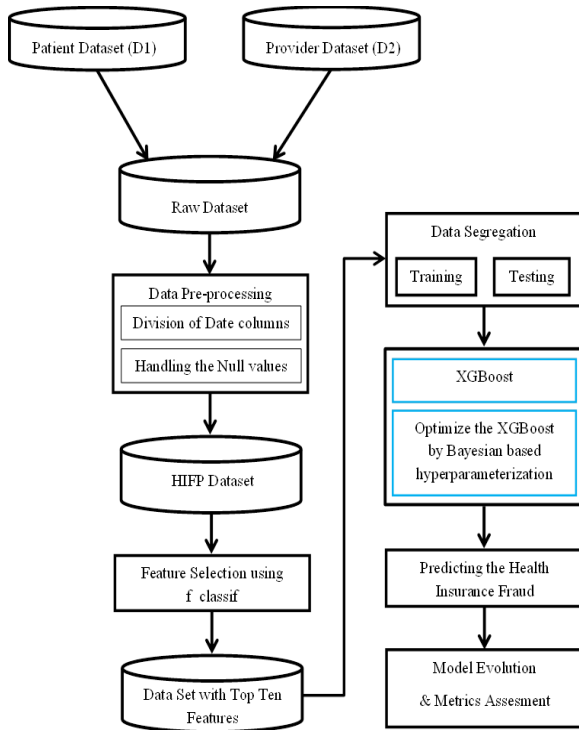


Figure 2. Flow of the proposed approach

The Grid and Random search algorithms explore the best parameters by generating the grids and random inputs. Instead, the Bayesian optimization would systematically explore the optimum hyperparameters by using the Gaussian processes. The search space, objective function, and surrogate and selection functions are the three core elements of the Bayesian optimization method. Obtain the samples by first initializing the search space. The objective function plays a crucial role in evaluating various settings for the hyperparameters. Since we perform classification, accuracy has been chosen as the evaluation metric as shown in algorithm 3.2.1. This BOXGBoost technique is useful for determining the best possible hyperparameters with which to train a computational model. The highest precision on the objective functions is extracted using a surrogate function and a selection function. Several techniques could be used to fine-tune the surrogate function's calibration. However, the Tree Parzen Estimator (TPE) or the Gaussian Processes are utilized in Bayes' optimization. Maximum predicted improvement is a widely applied criterion for selection processes. Over time, the derived score would be used to refine the surrogate function. Multiple combinations of hyperparameters would be applied to determine the loss. The performance metrics (e.g., accuracy) would be calculated at each of the iterations. The recommended optimum hyperparameters improve the potential of the XGBoost model.

3.3 Performance metrics

The confusion matrix was used to analyze the performance measurement of the imbalanced classification. True Positive (TP), True Negative (TN), False Positive (FP), and False

Negative (FN) were the four distinct combinations of predicted value and real value of the confusion matrix (Table 2). This approach enumerated the Accuracy, Precision, Recall, and F1 scores, which were used to evaluate the outcome of this study.

Table 2. Confusion matrix

		Predicted Values	
		Negative (0)	Positive (1)
Actual Values	Negative (0)	True Negative (TN)	False Positive (FP)
	Positive (1)	False Negative (FN)	True Positive (TP)

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (3)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (4)$$

$$\text{F1 Score} = 2 \times \left[\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \right] \quad (5)$$

$$\text{AUC} = \int_0^1 \text{Pr}[TP](v) dv \quad (6)$$

4. RESULTS AND DISCUSSION

Figure 3 represents the steps involved in data segmentation and the sample size of data. The finalized dataset after the feature selection contains ten attributes with 517737 records. Refer to the list of attributes available in Table 1. The dataset was alienated into 70% of training data and 30% of testing data. The XGBoost algorithm with default parameters had been utilized to predict health insurance fraud. The model delivered an overall accuracy of 81%. The precision of the model in finding the nonfraudulent transactions was 79%. It is lower than the precision of recognizing fraudulent claims (87%). While the model was successfully identifying the fraudulent claims, the recall (55%) and F1-Score (67%) got sored. The AUC value accomplished by this model was 0.882. Hence there were scopes to improve the performance of the model by optimizing the parameters. In this study, the Bayesian method had been employed to discern the best hyperparameters. The parameters delivered by the approach are listed in Table 3.

The BOXGBoost yielded a better outcome by achieving 98% accuracy. The precision, recall, and F1-Score were significantly improved. The model's precision in identifying the fraud got enhanced from 87% to 99%. The recall of the same transaction elevated from 55% to 95%. The F1-Score was also raised from 67% to 97%. The model consummated 0.994 as an AUC value. The performance of the pre- and post-optimization approach is represented in Figure 4. The outcomes of the BOXGBoost model had been appraised with other state of art algorithms including Naïve Bayes (NB), Logistic Regression (LR), Random Forest (RF), K-Nearest Neighbor (KNN), and AdaBoost. The accuracy, recall, F1-Score, and AUC values were considered for this evaluation. The outcomes of these models are listed in Table 4.

Table 3. Optimal hyperparameters proposed by Bayesian model

Description	Data
colsample_bytree	0.8403
gamma	5.2145
max_depth	17.0
min_child_weight	10.0
reg_alpha	41.0
reg_lambda	0.7403

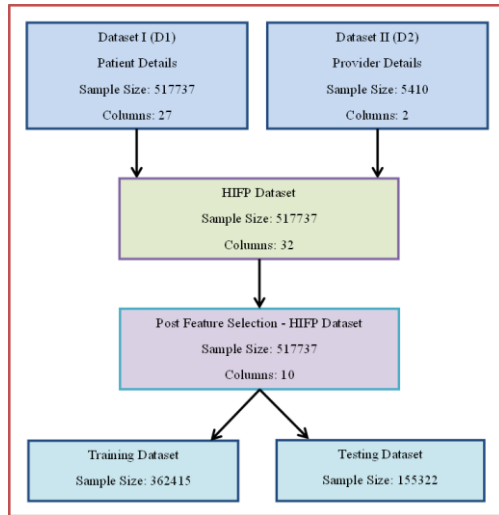


Figure 3. Formation of dataset

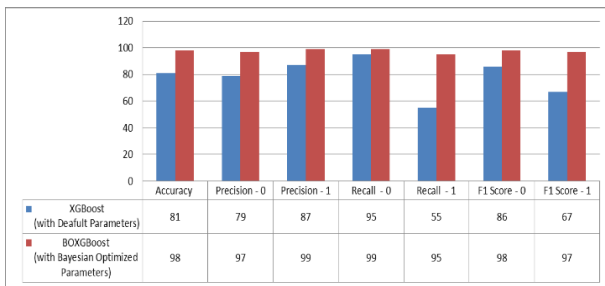


Figure 4. Performance of XGBoost and BOXGBoost

Table 4. Outcomes of the BOXGBoost and other state of art algorithms

Algorithm	Accuracy	Precision		Recall		F1-Score	
		0	1	0	1	0	1
Naive-Bayes	63	63	0	100	0	78	0
Logistic regression	63	63	47	100	0	78	0
Random forest	94	93	97	98	100	95	91
KNN	63	66	50	89	19	75	28
AdaBoost	71	71	72	87	57	82	64
BOXGBoost	98	97	99	99	95	98	97

By contemplating Table 4 and Figure 5, it is evident that the Naïve Bayes, Logistic Regression, and K-Nearest Neighbor (KNN) models have registered the lowest accuracy rates at 63%. Surpassing these models, the AdaBoost algorithm demonstrates a better performance with an accuracy of 71%. However, the Random Forest model shines with an impressive accuracy of 94%, securing the second-highest position. Notably, the proposed BOXGBoost model emerges as the clear leader, delivering an exceptional accuracy of 98%. Figures 6, 7, and 8 provide a comprehensive view of precision,

recall, and F1-Score metrics for all models. The analysis indicates that both Naïve Bayes and Logistic Regression models exhibit notably poor performance. Naïve Bayes fails to identify any True Negatives, indicating an inability to predict fraudulent claims. The logistic regression model, too, struggles with identifying fraudulent claims. As a result, these models rank at the lower end of the performance spectrum. On the other hand, KNN and AdaBoost models deliver a modest performance in these metrics. Despite Random Forest's commendable accuracy of 94%, it still falls short when compared to the superior performance achieved by the proposed BOXGBoost model.

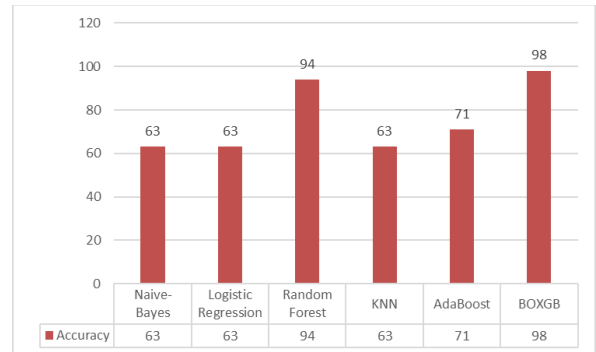


Figure 5. Accuracy of appraised ML models

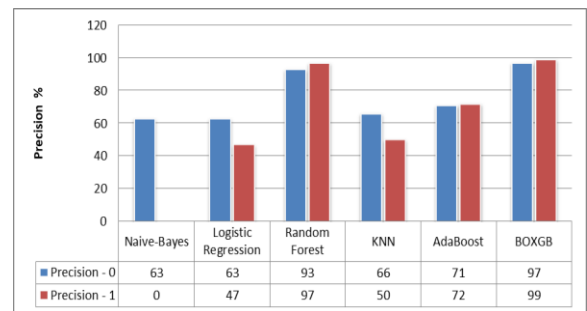


Figure 6. Precision of various ML models

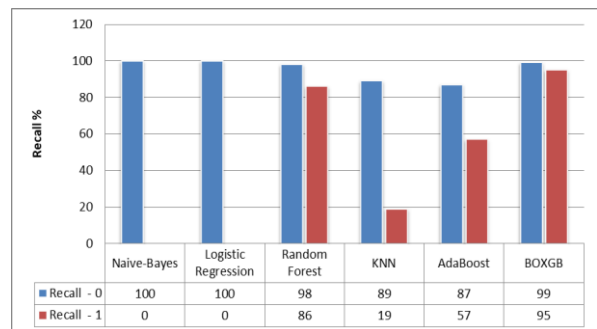


Figure 7. Recall of various ML models

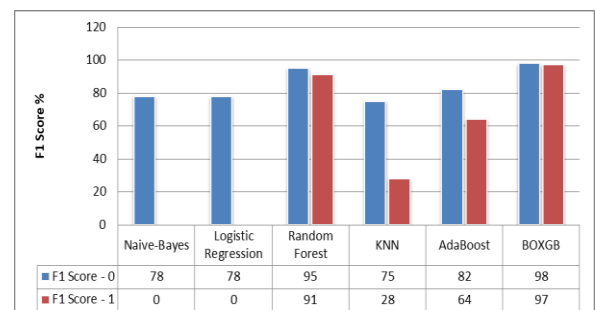


Figure 8. F1-Score of various ML models

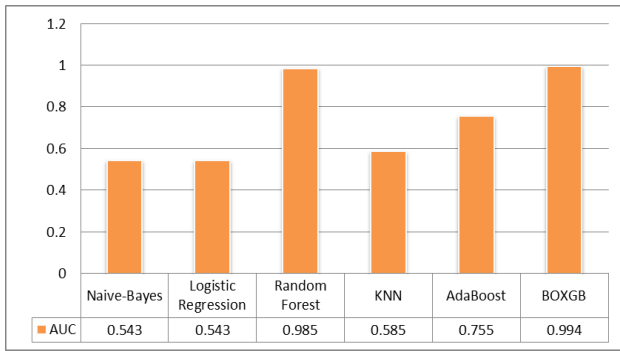


Figure 9. AUC values of ML models

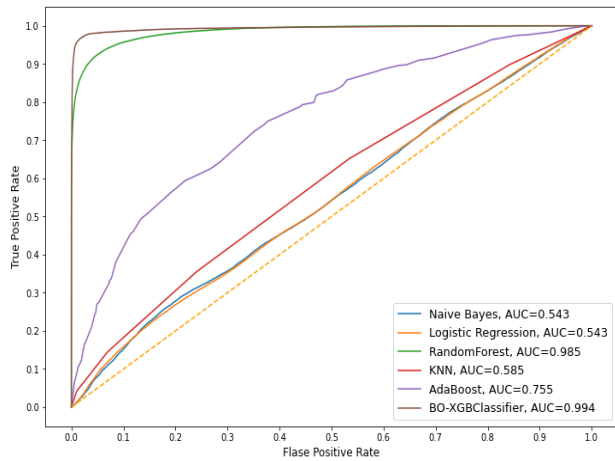


Figure 10. AUC-ROC plot of various ML models

Confusion Matrix

[[97748 798]
 [2986 53790]]

Classification Report

	precision	recall	f1-score	support
0	0.97	0.99	0.98	98546
1	0.99	0.95	0.97	56776
accuracy			0.98	155322
macro avg	0.98	0.97	0.97	155322
weighted avg	0.98	0.98	0.98	155322

Overall accuracy score: 98.0%

Figure 11. Confusion matrix of BOXGBoost model

In addition to the accuracy, precision, recall, and F1-Score metrics, AUC-ROC has also been a measure to ensure the capability of the proposed model. Figure 9 and Figure 10 represent the AUC-ROC of each appraised model. The trend of AUC-ROC indicated that the performances of Naïve Bayes (0.543), Logistic Regression (0.543), and KNN (0.585) models were low-lying. They conceded the AUC value which is proximate to the no-skill line. Compared with those models, AdaBoost performed better by generating an AUC of 0.755. Random Forest and BOXGBoost models effectuated the AUC values as 0.985 and 0.994. The Random Forest could only be assimilated to the second-highest position in all the metrics. The proposed BOXGBoost model outperformed every other

by clinched 98% of accuracy, 98% of precision, 97% of recall, 97.5% of F1 Score, and the AUC value as 0.994. However, the BOXGBoost model's performance, as illustrated in Figure 11, notably revealed a higher incidence of false negatives in contrast to false positives. This observation has practical implications, particularly in domains where minimizing false positives is of paramount importance to mitigate the potential for unwarranted interventions. The observed increase in false negatives could be attributed to factors such as class imbalance or a more conservative classification threshold applied to instances. To strike a more favorable balance between the occurrences of false positives and false negatives, it is prudent to consider threshold adjustment and dataset rebalancing as potential strategies. Modulating the classification threshold could offer a nuanced approach to achieve the desired trade-off, tailoring the model's behavior to the specific requirements of the application. Concurrently, addressing class imbalance through data augmentation or resampling techniques could further refine the model's predictive performance, reducing the prevalence of false positives without unduly elevating false negatives. These considerations underscore the importance of fine-tuning the model's settings to align with the specific objectives and constraints of the application context.

5. CONCLUSION

The quantities and dimensions of insurance fraud are getting multifold every day. Since manual embezzlement detection is not viable, the presence of artificial intelligence-based methodologies is inevitable. To address the issues in this discipline, the researchers are building Machine Learning models. In this study, the Bayesian Optimization (BO) method was employed to improve the outcome of the XGBoost model. The proposed BOXGBoost model was designated as the best one to predict the deceptions by producing 98% accuracy. The BOXGBoost-based framework would be helpful to predict health insurance frauds and minimize claim adjudication time. In the future, we plan to implement balancing techniques aimed at enhancing the model's overall performance. Additionally, we aim to explore the application of transfer learning strategies, allowing us to assess and compare the model's performance across diverse health insurance claim datasets. We also planned to construct a common platform for the health insurance business which would contain a secured data repository equipped with Machine Learning based fraud detection models. The real-world legal and functional challenges would also be studied as a part of this project. We hope that the Machine Learning based fraud detection platform will ensure victory over the war on health insurance frauds.

REFERENCES

- [1] National Health Care Anti-Fraud Association (NHCAA). (2018). The challenge of health care fraud. <https://www.nhcaa.org/tools-insights/about-health-care-fraud/the-challenge-of-health-care-fraud/>, accessed on Jan. 27, 2023.
- [2] Chen, Z.X., Hohmann, L., Banjara, B., Zhao, Y., Diggs, K., Westrick, S.C. (2020). Recommendations to protect patients and health care practices from Medicare and Medicaid fraud. *Journal of the American Pharmacists*

- Association, 60(6): e60-e65. <https://doi.org/10.1016/j.japh.2020.05.011>
- [3] Simborg, D.W. (2008). Healthcare fraud: Whose problem is it anyway? *Journal of the American Medical Informatics Association*, 15(3): 278-280. <https://doi.org/10.1197/jamia.M2672>
- [4] Nerenz, D.R., McFadden, B., Ulmer, C. (2009). *Race, Ethnicity, and Language Data: Standardization for Health Care Quality Improvement*. The National Academies Press, Washington, D.C.
- [5] Shakil, K.A., Zareen, F.J., Alam, M., Jabin, S. (2020). BAMHealthCloud: A biometric authentication and data management system for healthcare data in cloud. *Journal of King Saud University-Computer and Information Sciences*, 32(1): 57-64. <https://doi.org/10.1016/j.jksuci.2017.07.001>
- [6] Hathaliya, J.J., Tanwar, S., Tyagi, S., Kumar, N. (2019). Securing electronics healthcare records in healthcare 4.0: A biometric-based approach. *Computers & Electrical Engineering*, 76: 398-410. <https://doi.org/10.1016/j.compeleceng.2019.04.017>
- [7] Saldamli, G., Reddy, V., Bojja, K.S., Gururaja, M.K., Doddaveerappa, Y., Tawalbeh, L. (2020). Health care insurance fraud detection using blockchain. In 2020 Seventh International Conference on Software Defined Systems (SDS), Paris, France, pp. 145-152. <https://doi.org/10.1109/SDS49854.2020.9143900>
- [8] Nalluri, V., Chang, J.R., Chen, L.S., Chen, J.C. (2023). Building prediction models and discovering important factors of health insurance fraud using machine learning methods. *Journal of Ambient Intelligence and Humanized Computing*, 14(7): 9607-9619. <https://doi.org/10.1007/s12652-023-04633-6>
- [9] Lu, J., Lin, K., Chen, R., Lin, M., Chen, X., Lu, P. (2023). Health insurance fraud detection by using an attributed heterogeneous information network with a hierarchical attention mechanism. *BMC Medical Informatics and Decision Making*, 23(1): 1-17. <https://doi.org/10.1186/s12911-023-02152-0>
- [10] 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 (JITEKI)*, 9(2): 223-238.
- [11] Farbmacher, H., Löw, L., Spindler, M. (2020). An explainable attention network for fraud detection in claims management. *Journal of Econometrics*, 228(2): 244-258. <https://doi.org/10.1016/j.jeconom.2020.05.021>
- [12] Naidoo, K., Marivate, V. (2020). Unsupervised anomaly detection of healthcare providers using generative adversarial networks. In *Conference on e-Business, e-Services and e-Society*, Springer, Cham, pp. 419-430. https://doi.org/10.1007/978-3-030-44999-5_35
- [13] Singh, J., Urolagin, S. (2021). Use of artificial intelligence for health insurance claims automation. In *Advances in Machine Learning and Computational Intelligence*, Springer, Singapore, pp. 381-392. https://doi.org/10.1007/978-981-15-5243-4_35
- [14] Kapadiya, K., Patel, U., Gupta, R., Alshehri, M.D., Tanwar, S., Sharma, G., Bokoro, P.N. (2022). Blockchain and AI-empowered healthcare insurance fraud detection: An analysis, architecture, and future prospects. *IEEE Access*, 10: 79606-79627. <https://doi.org/10.1109/ACCESS.2022.3194569>
- [15] Rukhsar, L., Bangyal, W.H., Nisar, K., Nisar, S. (2022). Prediction of insurance fraud detection using machine learning algorithms. *Mehran University Research Journal of Engineering & Technology*, 41(1): 33-40. <https://doi.org/10.22581/muet1982.2201.04>
- [16] Hancock, J.T., Khoshgoftaar, T.M. (2021). Gradient boosted decision tree algorithms for Medicare fraud detection. *SN Computer Science*, 2(4): 1-12. <https://doi.org/10.1007/s42979-021-00655-z>
- [17] Parnian, K., Sorouri, F., Souha, A.N., Molazadeh, A., Mahdavi, S. (2021). Fraud detection in health insurance using a combination of feature subset selection based on squirrel optimization algorithm and nearest neighbors algorithm methods. *Future Generation in Distributed Systems Journal*, 3(2): 1-11.
- [18] Johnson, J.M., Khoshgoftaar, T.M. (2019). Medicare fraud detection using neural networks. *Journal of Big Data*, 6(1): 1-35. <https://doi.org/10.1186/s40537-019-0225-0>
- [19] Castaneda, G., Morris, P., Khoshgoftaar, T.M. (2019). Maxout neural network for big data medical fraud detection. In 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService), Newark, CA, USA, pp. 357-362. <https://doi.org/10.1109/BigDataService.2019.00064>
- [20] Bauder, R., da Rosa, R., Khoshgoftaar, T. (2018). Identifying Medicare provider fraud with unsupervised machine learning. In 2018 IEEE International Conference on Information Reuse and Integration (IRI), Salt Lake City, UT, USA, pp. 285-292. <https://doi.org/10.1109/IRI.2018.00051>
- [21] Pandey, P., Saroliya, A., Kumar, R. (2018). Analyses and detection of health insurance fraud using data mining and predictive modeling techniques. In *Soft Computing: Theories and Applications*, Springer, Singapore, pp. 41-49. https://doi.org/10.1007/978-981-10-5699-4_5
- [22] Kareem, S., Ahmad, R.B., Sarlan, A.B. (2017). Framework for the identification of fraudulent health insurance claims using association rule mining. In 2017 IEEE Conference on Big Data and Analytics (ICBDA), Kuching, Malaysia, pp. 99-104. <https://doi.org/10.1109/ICBDAA.2017.8284114>
- [23] Lasaga, D., Santhana, P. (2018). Deep learning to detect medical treatment fraud. *Proceedings of Machine Learning Research*, 71: 114-120.
- [24] Bauder, R.A., Khoshgoftaar, T.M. (2017). Medicare fraud detection using machine learning methods. In 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), Cancun, Mexico, pp. 858-865. <https://doi.org/10.1109/ICMLA.2017.00-48>
- [25] Ahmadinejad, H., Norouzi, A., Ahmadi, A., Yousefi, A. (2016). Distance based model to detect healthcare insurance fraud within unsupervised database. *Indian Journal of Science and Technology*, 9(43): 1-16. <https://doi.org/10.17485/ijst/2016/v9i43/104971>
- [26] Bayerstadler, A., van Dijk, L., Winter, F. (2016). Bayesian multinomial latent variable modeling for fraud and abuse detection in health insurance. *Insurance: Mathematics and Economics*, 71: 244-252. <https://doi.org/10.1016/j.insmatheco.2016.09.013>
- [27] Johnson, M.E., Nagarur, N. (2016). Multi-stage methodology to detect health insurance claim fraud. *Health Care Management Science*, 19(3): 249-260. <https://doi.org/10.1007/s10729-015-9317-3>

- [28] Kaggle Data, v1. <https://www.kaggle.com/rohitrox/healthcare-provider-fraud-detection-analysis>, accessed on Sep. 30, 2023.
- [29] Rezaeijoo, S.M., Abedi-Firouzjah, R., Ghorvei, M., Sarnameh, S. (2021). Screening of COVID-19 based on the extracted radiomics features from chest CT images. *Journal of X-ray Science and Technology*, 29(2): 229-243. <https://doi.org/10.3233/XST-200831>
- [30] Alaskar, L., Crane, M., Alduailij, M. (2019). Employee turnover prediction using machine learning. In *International Conference on Computing*, Springer, Cham, pp. 301-316. https://doi.org/10.1007/978-3-030-36365-9_25
- [31] Sheluhin, O.I., Ivannikova, V.P. (2020). Comparative analysis of informative features quantity and composition selection methods for the computer attacks classification using the unsw-nb15 dataset. *T-Comm-Телекоммуникации и Транспорт*, 14(10): 53-60. <https://doi.org/10.18372/2310-1766-2023-14-10-53-60>
- [32] Tsai, J.F., Carlsson, J.G., Ge, D., Hu, Y.C., Shi, J. (2014). Optimization theory, methods, and applications in engineering 2013. *Mathematical Problems in Engineering*, 2014: 319418. <https://doi.org/10.1155/2014/319418>
- [33] Hutter, F., Hoos, H., Leyton-Brown, K. (2014). An efficient approach for assessing hyperparameter importance. In *International Conference on Machine Learning*, pp. 754-762.
- [34] Kotthoff, L., Thornton, C., Hoos, H. H., Hutter, F., Leyton-Brown, K. (2017). Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA. *Journal of Machine Learning Research*, 18: 1-5.
- [35] Kotthoff, L., Thornton, C., Hoos, H.H., Hutter, F., Leyton-Brown, K. (2019). Auto-WEKA: Automatic model selection and hyperparameter optimization in WEKA. *Automated Machine Learning*, Springer, Cham, 81-95. https://doi.org/10.1007/978-3-030-05318-5_4
- [36] Bernard, S., Heutte, L., Adam, S. (2009). Influence of hyperparameters on random forest accuracy. *International Workshop on Multiple Classifier Systems*, Springer, Berlin, Heidelberg, 171-180. https://doi.org/10.1007/978-3-642-02326-2_18
- [37] Probst, P., Wright, M.N., Boulesteix, A.L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(3): e1301. <https://doi.org/10.1002/widm.1301>
- [38] Fauzan, M.A., Murfi, H. (2018). The accuracy of XGBoost for insurance claim prediction. *International Journal of Advances in Soft Computing and Its Applications*, 10(2): 159-171.
- [39] Wang, Y., Ni, X.S. (2019). A XGBoost risk model via feature selection and Bayesian hyper-parameter optimization. *arXiv preprint arXiv:1901.08433*. <https://doi.org/10.48550/arXiv.1901.08433>
- [40] Sommer, J., Sarigiannis, D., Parnell, T. (2019). Learning to Tune XGBoost with XGBoost. *arXiv preprint arXiv:1909.07218*. <https://doi.org/10.48550/arXiv.1909.07218>
- [41] Li, W., Yin, Y., Quan, X., Zhang, H. (2019). Gene expression value prediction based on XGBoost algorithm. *Frontiers in Genetics*, 10: 1077. <https://doi.org/10.3389/fgene.2019.01077>
- [42] Zhou, J., Qiu, Y., Zhu, S., Armaghani, D.J., Khandelwal, M., Mohamad, E.T. (2021). Estimation of the TBM advance rate under hard rock conditions using XGBoost and Bayesian optimization. *Underground Space*, 6(5): 506-515. <https://doi.org/10.1016/j.undsp.2020.05.008>