IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 3, Issue 3, July 2023
Impact Factor: 7.301

An Explainable AI System for Fraud Identification in Insurance Claims via Machine-Learning Methods

Yeshwanth Macha and Sunij Kumar Pulichikkunnu
Independent Researcher
yeshwanthmacha97@gmail.com and spulichikkunnu@gmail.com

Abstract: *The estimation of insurance claims/fraud detection is significant to the stability and efficiency of the insurance industry. Effective estimation of claims assists the insurers in estimating risks more effectively, and cover compensation as fast as possible, and preventing fraud would prevent huge losses of finances that could undermine the stability of the world economies and the effectiveness of the capital markets. Trying to deal with these problems, the research paper discusses the application of the science of artificial intelligence (AI) in order to predict insurance claims with regard to accuracy, interpretability and decision-making support. On the basis of organized medical data, the given XGBoost algorithm was used to construct a strong predictive model. The experimental results show that the XGBoost model has a high performance with an accuracy of 98.78% which is much better than the traditional models, which incorporate the Logistic Regression (LR), AdaBoost and Naive Bayes (NB) models. In addition to that, with the introduction of explainable AI(XAI) approaches, including SHAP and LIME, the level of transparency is enhanced because it shows the role that potentially important features play in model forecasts. These findings confirm that integrating advanced machine learning (ML) with interpretability not only ensures predictive reliability but also fosters stakeholder trust, offering a scalable and practical framework for mitigating fraud and enhancing operational efficiency in insurance analytics.*

Keywords*:* Insurance claim prediction, fraud detection, machine learning, explainable AI (XAI), healthcare analytics, XGBoost

I. INTRODUCTION

In today's competitive environment, individuals are frequently exposed to stress, which can lead to both physical and mental health problems. In order to overcome these obstacles, it is crucial to obtain sufficient health insurance policies that cover both mental and physical sickness treatment [1]. If a patient is covered by health insurance at the time of treatment, financial difficulties can be avoided. Thus, more and more people are choosing to purchase health insurance these days after realising its significance [2][3]. However, alongside this positive trend in insurance adoption, the industry grapples with a pressing and costly issue: insurance fraud [4]. Fraudulent claims, whether exaggerated, falsified, or intentionally misrepresented, inflict significant financial losses on insurance companies and disrupt the overall efficiency of the system. Additionally, these dishonest tactics raise premiums for sincere clients, eroding confidence in the insurance industry [5][6].

Detecting such fraud presents a complex challenge [7]. The enormous volume of claims and the increasing sophistication of the methods employed by the fraudsters render the traditional methods of detection ineffective [8]. Certain and subtle fraud patterns concealed within large and varied data volumes are generally not easily uncovered by manual audits and rule-based approaches [9]. The recent advancements in the field of artificial intelligence (AI), namely, machine learning (ML) and deep learning (DL), have improved fraud detection in the insurance sector to a great extent, as the notion of processing an immense volume of claims data to identify hidden abnormalities and novel fraud patterns [10]. Many of them, however, are black boxes that cannot be seen clearly, which is a very critical matter

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 3, Issue 3, July 2023 Impact Factor: 7.301

in the insurance sector. To deal with that, explainable AI (XAI) provides such a solution, making ML models more interpretable without affecting their accuracy [11]. The rationale behind the proposed paper is to provide an explainable AI algorithm for spotting false insurance claims, achieving a high level of performance and transparency, and thereby enabling insurers to make sound and reliable decisions.

A. Motivation and Contribution

In the modern, stressful world, there has been an increased uptake of health insurance covers, which provides financial coverage in the event of medical emergencies due to growing health concerns. This good omen, however, is being spoiled by the increasing cases of insurance fraud, which involve fabricated claims or exaggerated claims that make insurers bear huge financial costs and inflate the price of premiums for honest policyholders, which is driving people out of the system. In large-scale, diverse claim data, such as human audits and rule-based systems, conventional fraud detection algorithms are unsuccessful at identifying more intricate, hidden patterns. Although DL and ML models have demonstrated potential in automating fraud detection, their interpretability issues provide a significant obstacle in operational and legal environments. Therefore, this study is motivated by the need for a transparent and accurate solution, proposing an XAI framework that not only improves the efficacy of fraud detection but also guarantees interpretability, allowing insurers to make prudent and responsible judgments.

- Developed a generalizable methodology applicable to other healthcare and insurance datasets.
- Performed exploratory data analysis to uncover feature relationships and guide preprocessing.
- Applied robust preprocessing (outlier detection, standardization, and train-test split) to improve data quality.
- Proposed XGBoost for prediction due to its efficiency, scalability, and regularization against overfitting.
- Evaluated model performance using a variety of criteria (accuracy, precision, recall, F1-score, and ROC) to provide a trustworthy evaluation.

B. Structure of Paper

The paper is structured as follows: Section II reviews related work on insurance claim prediction. Section III outlines the proposed methodology. Section IV provides the findings, a commentary, and a comparison with baseline models. Finally, Section V summarises the study's main conclusions.

II. LITERATURE REVIEW

The significance of using ML models in insurance is highlighted in recent research, especially in the domains of optimising premiums, analysing claims, detecting fraud, and assessing risk.

Ataabadi et al. (2022) suggest an approach that employs ML to predict the expenses associated with claims by analysing the medical records of other patients, as well as to identify claims that are significantly different from others and potentially fraudulent. In rare instances, the suggested data sampling method decreased the deduction rate's absolute error from 35 to 23 errors. The evaluation results showed that the dataset had around 0.5% of anomalous events with an absolute inaccuracy greater than 20%. It is possible to adjust the anomalous rates to a lower or higher range [12]. Yoo et al. (2022) Medicare beneficiaries and providers were placed as nodes in a heterogeneous graph. Consequently, the Graph SAGE model outperformed the accuracy, recall, and area under the receiver operating characteristic curve of the baseline model by 0.01, 0.35, 0.30, and 0.18, respectively [13].

Kaushik et al. (2022) developed and assessed employing AI networks in a model that forecasts health insurance prices using regression. The authors examined the model's performance utilising important performance measures after the trial results showed an accuracy of 92.72% [14]. Hanafy and Ming (2021) In order to predict how often claims filed, many ML methods can be employed, such as XGBoost, K-NN, naïve Bayes, decision trees, random forests, logistic regression, and XGBoost. Additionally, contrast and examine the mechanisms of various models. Measures like as accuracy (0.8677), kappa (0.7117), and area under the curve (0.840) demonstrated that RF outperforms other methods [15].

Dhieb et al. (2020) provide a foundation for an automated and safe insurance system that minimises human

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 3, Issue 3, July 2023 Impact Factor: 7.301

involvement. Applying XGBoost to a dataset including information on auto insurance results in impressive performance increases in comparison to other current learning algorithms, according to the acquired results. XGBoost outperformed decision tree models in identifying fraudulent claims on a car insurance dataset, obtaining a 7% better accuracy rate [16]. Rayan (2019) introduces an integrated system that integrates domain knowledge with supervised and unsupervised learning methods to find false claims among a certain group of unresolved claims. The initial case study with one insurer demonstrates an increase in hit-rate by 209.4% [17].

Despite growing interest in ML for insurance applications, key research gaps persist. Most studies focus on model accuracy but overlook interpretability, outlier handling, and real-world deployment challenges. Limited attention is given to hybrid architectures and secure integration frameworks like blockchain. As summarized in Table I, while various techniques show promising outcomes, future research should prioritize scalable, explainable, and domain-adapted ML solutions for broader industry adoption

Table 1: Summary of recent study on ML Applications in the fraudulent Insurance claim detection

Author	Technique	Data	Outcomes	Implication	Recommendation
Ataabadi et al. (2022)	ML-based claim cost prediction + custom data sampling	700,000 claims from RASA web portal	Reduced error in exceptional cases; identified 0.5% abnormal claims	Enhances fraud detection and cost prediction accuracy	Use tailored sampling to improve ML performance on outlier cases
Yoo et al. (2022)	GraphSAGE on heterogeneous graph	Medicare provider-beneficiary relationships	Improved precision, recall, F1-score, and AUC over baseline	Graph-based modeling captures relational fraud patterns	Apply graph learning to exploit network structures in fraud detection
Kaushik et al. (2022)	Artificial Neural Network (ANN) regression	Health insurance data with demographic features	Achieved 92.72% accuracy in premium prediction	Enables personalized premium estimation	Use ANN for dynamic pricing based on individual risk profiles
Hanafy & Ming (2021)	Logistic Regression, XGBoost, RF, DT, NB, KNN	Automotive insurance big data	RF showed highest accuracy, kappa, and AUC	ML models can optimize claim prediction across insurance types	Prefer RF for robust performance in auto insurance analytics
Dhieb et al. (2020)	XGBoost + Blockchain framework	Auto insurance dataset	XGBoost outperformed other models; 7% higher accuracy than DT	Combines secure data sharing with effective fraud detection	Integrate ML with blockchain for secure, automated insurance systems
Rayan (2019)	Hybrid: Rule Engine + DT + Perceptron + Clustering	Outstanding claims from insurer	Increased hit-rate by 209.4% in fraud detection	Hybrid models enhance prioritization and investigation efficiency	Use ensemble and hybrid approaches for proactive fraud identification

III. METHODOLOGY

The proposed methodology for predicting insurance claims involves a systematic pipeline encompassing data analysis, preprocessing, model selection, and evaluation. This study begins with a structured healthcare dataset that undergoes to pre-processing involve cleaning, outlier detection with the IQR method, feature standardization to normalize numerical attributes, and partitioning the dataset into training and testing subsets (80:20 split). The XGBoost algorithm was then used for prediction due to its efficiency and regularization capabilities. To provide dependable and broadly applicable results, model performance was assessed using ROC curve analysis, accuracy, precision, recall, and F1-score, as shown

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 3, Issue 3, July 2023
Impact Factor: 7.301

in Figure 1.

The proposed methodology is explained step by step as follows:

A. Data Gathering and Analysis

The instance A structured healthcare dataset called the Insurance Claim Prediction Dataset was created to forecast a person's likelihood of filing an insurance claim by using demographic and health-related characteristics. A binary target variable insurance claim (1 = claim, 0 = no claim) and other data are included, parameters such as age, sex, body mass index (BMI), family size, smoking status, area of residence, average daily step count, and healthcare costs. A correlation heatmap, demonstrating a substantial positive connection between characteristics, was found using exploratory data analysis (EDA) to identify feature linkages and patterns.

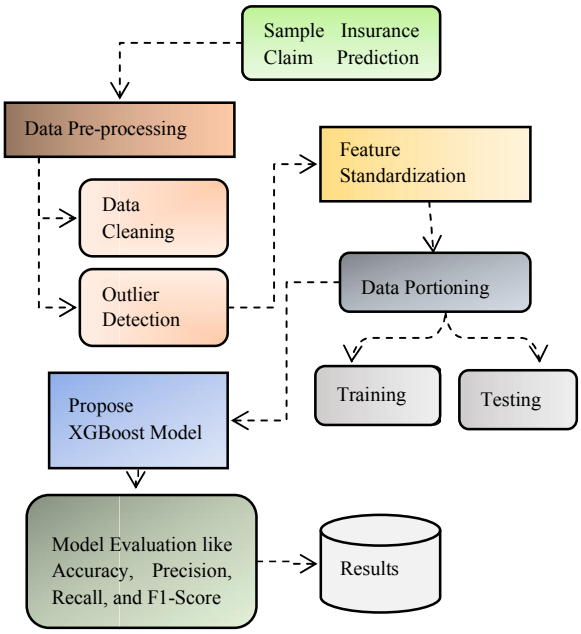

Fig. 1. Propose Flowchart For Insurance Claim Detection

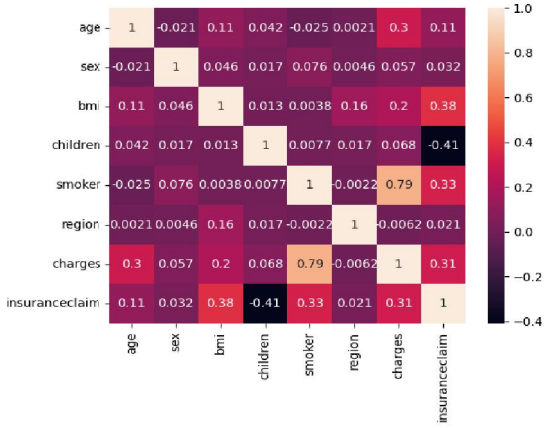

Fig. 2. Correlation Heamap of Features

Figure 2 shows the correlation heatmap of insurance features. "Smoker" and "charges" have the strongest positive correlation (0.79). "Insurance claim" is moderately correlated with "charges" (0.31) and "smoker" (0.33). A negative correlation with "children" (-0.41) suggests fewer claims among individuals with children.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 3, Issue 3, July 2023
Impact Factor: 7.301

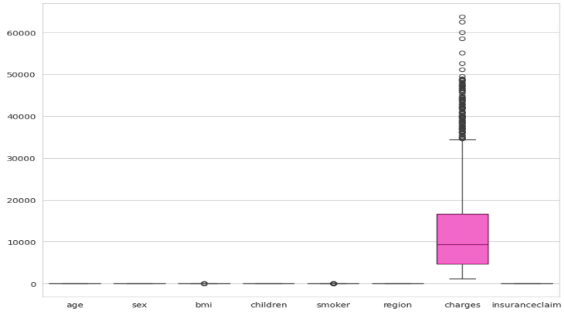

Fig. 3. Box Plot for outlier detection

A box plot visualisation of important insurance-related elements is shown in Figure 3, highlighting the distribution, central tendency, and outliers within each variable. Among them, "charges" exhibits the widest spread and numerous outliers above the upper quartile, indicating high variability in insurance costs. In contrast, variables like age, sex, BMI, children, smoker, region, and insurance claim show more compact distributions with fewer extreme values. This plot aids in identifying skewed data and potential anomalies, which are critical for preprocessing and model reliability.

B. Data Pre-Procesing

The procedures used to clean data and prepare it for usage in other contexts are referred to as data preparation. However, a series of steps must be taken to improve its quality before incorporating information into ML algorithms. This study uses various pre-processing pipelines that are listed in below:

- **Data Cleaning:** The dataset contained no duplicate rows and no missing values across all features. Hence, the data was already clean and required no further imputation or removal operations, making it suitable for further preprocessing and analysis.
- **Outlier detection:** Using the IQR method, 198 rows were detected as outliers across features like BMI, steps, and charges, representing extreme deviations that may be treated or retained based on analysis needs.

C. Feature Standardization

A feature standardisation procedure is carried out to ensure that all inputs are normalised when numerical inputs are utilised to feed data. This is crucial for models that rely on the scale of the features, as it implies that factors with huge relative sizes cannot significantly influence the learning process. In order to standardise a numerical property, determine the standardised value a_{ij}^{*} using Equation (1):

$$a_{ij}^{*} = \frac{a_{ij} - \mu_{aj}}{\sigma_{aj}} \qquad (1)$$

where σ_{aj} is the standard deviation of attribute a_{ij} over all projects and μ_{aj} is the mean.

D. Data Portioning

The data was first split into testing and training subsets before the model was built. The study's data was divided into an 80:20 ratio.

E. Propose XGBoost Model

The XGBoost algorithm was proposed by Chen and Guestrin and is based on the GBDT structure. It has received a lot of attention because of how well it performs in ML events' Kaggle contests [18]. To prevent overfitting, XGBoost, in contrast to the GBDT, adds the term regularisation to the objective function. The setting for the goal function is Equation (2):

$$Obj(\theta) = \sum_{i=1}^{n} \ell(\hat{y}_i, y_i) + \sum_{k=1}^{k} \Omega(f_k) \qquad (2)$$

In this equation:

- $Obj(\theta)$ is the function that XGBoost is trying to minimize.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 3, Issue 3, July 2023 Impact Factor: 7.301

- $\ell(\hat{y}_i, y_i)$ is a difference between the actual value y_i and the forecast $\hat{y}_i$ is measured using a differentiable convex loss function.
- $\Omega(f_k)$ is the word for regularisation to manage the model's complexity.

The objective function consists of two main components:

- **Loss Function:** This section evaluates the degree to which the model's predictions agree with the actual values, the objective function.
- **Regularization Term:** This is one of the objective functions that is used to penalize more complicated models in order to eliminate their complexity. As a result, the model is less likely to overfit and can adapt to new data.

The ability to work with big data sets and employ various optimisation devices, such tree pruning or parallel processing, allows XGBoost to be effective and perform well. Its many useful features make it a top pick for many ML tasks.

F. Evaluation Measures

The performance of the models was measured with the use of a performance matrix, and the most important metrics that were used were accuracy, precision, recall, and F1-score. Calculation of various classes was done using the following measures: It is easier to understand that the positive cases have been correctly recognised when the number of True Positives (TPs) is larger than the number of True Negatives (TNs), which are defined as negative cases. False Negatives (FNs) are those that were wrongly identified, whereas False Positives (FPs) are those that are wrongly identified. These measures are also important in project risk management as it is used to predict the reliability. The values are established using the assistance of typical Equations (3) to (6):

$$Accuracy = \frac{\text{TP+TN}}{\text{TP+Fp+TN+FN}} \quad (3)$$

$$Precision = \frac{\text{TP}}{\text{TP+FP}} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

$$F1 - Score = \frac{2(Precision*Recall)}{Precision+Recall} \quad (6)$$

Accuracy refers to the general correctness, precision to the accuracy of positive predictions, and recall to the model's ability to detect actual dangers. The F1-score offers comparable performance while striking a balance between recall and accuracy. The ROC curve also indicates the model's ability to classify various criteria, ensuring that risk management decisions are informed.

IV. RESULTS ANALYSIS AND DISCUSSION

This section provides the results of the categorisation models, and shows the extent to which they predict the results in terms of assessment measures. All the testing was carried out with an Intel Core i7-vPro 7th gen processor with windows 10 installed and 16 GB of RAM. Table II shows the measures of the assessment of the XGBoost model, which shows superior predictive capabilities. The accuracy percentage of the model is 98.78% which implies that the model forecasts are accurate in practically all cases. The fact that it has a precision of 98.34% means that its rate of false positives is low, but the rate of its recall was 99% which confirms the fact that the test can be used to ascertain true claims. The model's strong F1-score of 98% suggests that it is very accurate and remembering, making it a viable option for use in making crucial decisions involving insurance analytics.

Table 2: Experiment Results of xgboost for insurance claim prediction

Matrix	XGBoost
Accuracy	98.78
Precision	98.34
Recall	99
F1-score	98

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 3, Issue 3, July 2023
Impact Factor: 7.301

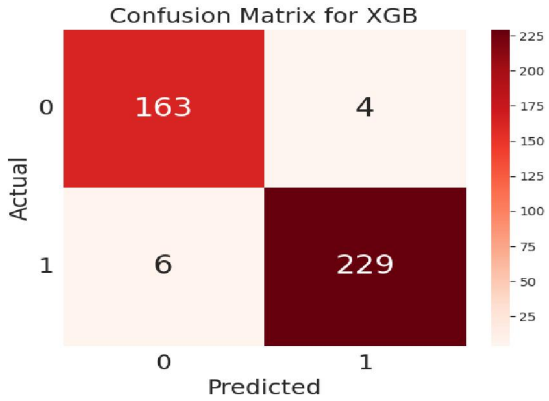

Fig. 4. Confusion Matrix of XGBoost Model

Figure 4 shows the confusion matrix of XGBoost classification model, providing a visual assessment of the prediction strength of the model. This matrix implies that the model had a high accuracy in classifying 163 TN and 229 TP and low accuracy in classifying 4 FP and 6 FN. This shows that there is good level of accuracy as well as categorisation ability of both classes. All in all, the matrix shows that the model was effective to separate the two target groups with minimal error.

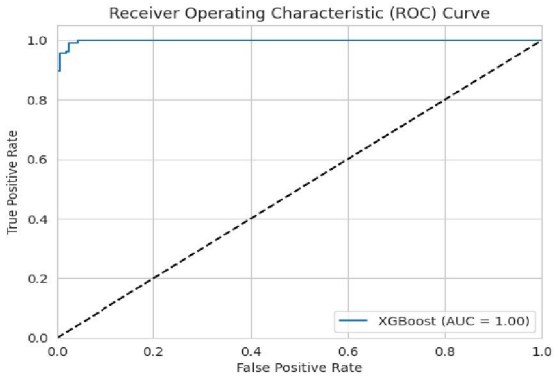

Fig. 5. ROC Curve of XGBoost Model

Figure 5 illustrates the ROC curve of XGBoost classification model and it shows that the model possesses an outstanding discriminative power. The curve has plotted the model's TPR and FPR, the blue line shows the model's performance, and the dashed diagonal shows the random classifier's baseline. The Area under the Curve (AUC) = 1.00, that is, no cases of FP or FN. This ideal AUC score confirms that the model achieves maximum for binary classification problems, and its sensitivity and specificity make it extremely dependable.

Explainable AI(XAI)

A brief explanation of the XAI methods employed in this study is provided below. The purpose of this study is not to provide an accurate forecast, but rather to demonstrate and explain the artificial intelligence techniques, including the SHAP and LIME tools, that are used to test the AI model.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 3, Issue 3, July 2023
Impact Factor: 7.301

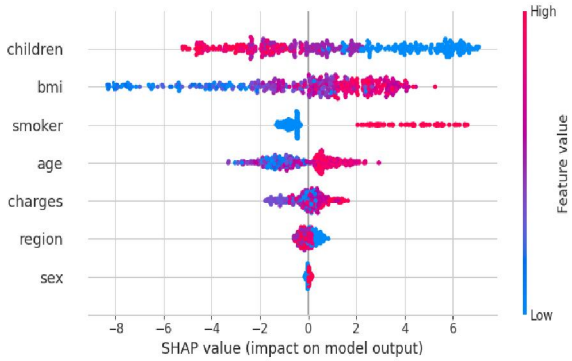

Fig. 6. SHAP summary plot for the XGBoost model

The impact of each attribute on the model's predictions is seen in Figure 6. High values (in red) for features such as "smoker," "BMI," and "age" typically drive projections upward, demonstrating their considerable impact. One forecast is symbolised by each dot, and the blue-to-red colour gradient signifies the feature value, helping visualize how feature magnitude affects model output. This plot provides clear insight into feature importance and directionality, enhancing model interpretability.

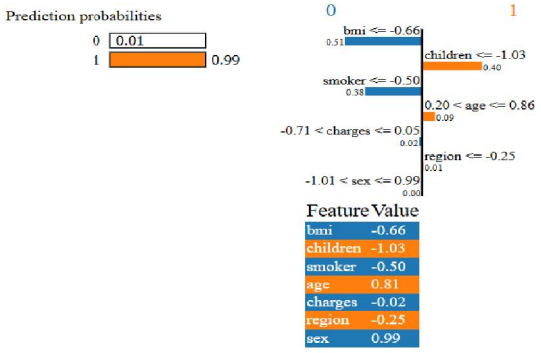

Fig. 7. Lime Graph for the XGBoost model

Figure 7 illustrates the prediction explanation for a classification model, showing that the instance was classified as Class 1 with a probability of 0.99. The horizontal bar graph highlights the dominance of Class 1, while the decision rules and feature values, such as BMI, children, smoker, age, charges, region, and sex, reveal how each input contributed to the final prediction. Color-coded feature values and thresholds indicate the directional impact of each variable, offering a transparent view of the procedure by which the model makes decisions.

Comparative Analysis

The performance of many models for predicting insurance claims is compared in this section, as indicated in Table III, clearly demonstrating the superiority of the proposed XGBoost model. XGBoost achieves the highest accuracy (98.78%), significantly outperforming Logistic Regression (75.03% accuracy), AdaBoost (82.73% accuracy), and Naïve Bayes (60.56% accuracy). These findings justify why XGBoost is robust and reliable in the accurate determination of insurance claims with limited false prediction.

Table 3: Comparison between propose and existing model performance for insurance claim prediction

Models	Accuracy	Precision	Recall	F1 Score
LR[19]	75.03	78	75	76
AdaBoost[20]	82.73	92.55	82.61	87.29
NB[15]	60.56	65.58	72.73	68.97
XGBoost	98.78	98.34	99	98

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 3, Issue 3, July 2023 Impact Factor: 7.301

The XGBoost-based solution proposed is both effective and novel in insurance claims prediction. It has better predictive performance than conventional models do. This can be attributed to the fact that XGBoost can support complex combinations of features and when the model is overfitting, it can be regularized. The combination of SHAP and LIME is new as it incorporates interpretability and accuracy. The significance of key features to the stakeholders is easy to comprehend and this increases accountability and trust. The major novelty is the end-to-end incorporation of a robust ML algorithm and XAI into a framework, which can be more precise, understandable, and useful in practice than black-box models.

V. CONCLUSION AND FUTURE WORK

The insurance fraud is also another significant issue in the world insurance industry, and the premiums being paid to the policyholder are excessive and thus incurring a lot of losses to the policyholder. In an effort to address this issue, this study examined the use of AI in claims processing, with a focus on predictive models for detecting both fraudulent and authentic claims. The suggested XGBoost model had enormous predictive behaviour, having an accuracy of 98.78%, precision of 98.34%, recall of 99% and F1-score of 98% which is immeasurably improved compared to the traditional models of LR, AdaBoost and NB. Moreover, the XAI techniques, including SHAP and LIME, made the results of the models more understandable, as they provided the impact of such variables as smoking status, BMI, and age on the predictions of the model, which enhanced the degree of trust and transparency in decision-making. Though these findings are good, the study is constrained by the fact that it utilizes a single dataset, which may not be quite representative of the larger and more varied populations, and the consideration of the outliers, which may jeopardize the strength of the models. The research directions in the future might be considered the utilization of more extensive and heterogeneous data, the integration of time and behavioural data to assess dynamic risk, and the discussion of DL and advanced ensemble techniques to enhance the predictive quality and adaptability in the context of practical insurance.

REFERENCES

[1] N. Patel, “Quantum Cryptography In Healthcare Information Systems: Enhancing Security In Medical Data Storage And Communication,” *J. Emerg. Technol. Innov. Res.*, vol. 9, no. 8, pp. 193–202, 2022.

[2] M. Ghosh and R. Gor, “Health Insurance Premium Prediction Using Blockchain Technology and Random Forest Regression Algorithm,” *Int. J. Eng. Sci. Technol.*, vol. 6, no. 3, pp. 74–82, Jun. 2022, doi: 10.29121/ijoest.v6.i3.2022.346.

[3] T. Takeshima, S. Keino, R. Aoki, T. Matsui, and K. Iwasaki, “Development of Medical Cost Prediction Model Based on Statistical Machine Learning Using Health Insurance Claims Data,” *Value Heal.*, vol. 21, Sep. 2018, doi: 10.1016/j.jval.2018.07.738.

[4] N. Malali, “Using Machine Learning To Optimize Life Insurance Claim Triage Processes Via Anomaly Detection In Databricks: Prioritizing High-Risk Claims For Human Review,” *Int. J. Eng. Technol. Res. Manag.*, vol. 6, no. 6, 2022, doi: 10.5281/zenodo.15176507.

[5] S. Rawat, A. Rawat, D. Kumar, and A. S. Sabitha, “Application of machine learning and data visualization techniques for decision support in the insurance sector,” *Int. J. Inf. Manag. Data Insights*, vol. 1, no. 2, Nov. 2021, doi: 10.1016/j.jjimei.2021.100012.

[6] N. Malali, “The role of Devsecops in Financial AI Models: Integrating Security at Every Stage of AI/ML Model Development in Banking and Insurance,” *Int. J. Eng. Technol. Res. Manag.*, vol. 6, no. 11, p. 218, 2022.

[7] S. N. Pushpak and S. Jain, “An Implementation of Quantum Machine Learning Technique to Determine Insurance Claim Fraud,” in *2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, IEEE, Oct. 2022, pp. 1–5. doi: 10.1109/ICRITO56286.2022.9964828.

[8] R. Roy and K. T. George, “Detecting insurance claims fraud using machine learning techniques,” in *2017 International Conference on Circuit, Power and Computing Technologies (ICCPCT)*, IEEE, Apr. 2017, pp. 1–6. doi: 10.1109/ICCPCT.2017.8074258.

[9] H. O. Özcan, İ. Çolak, S. Erimhan, V. Güneş, F. Abut, and F. Akay, “SOBE: A Fraud Detection Platform in

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 3, Issue 3, July 2023 Impact Factor: 7.301

Insurance Industry," *Kocaeli J. Sci. Eng.*, vol. 5, no. ICOLES2021 Special Issue, pp. 25–31, Nov. 2022, doi: 10.34088/kojose.1019125.

[10] I. Kose, M. Gokturk, and K. Kilic, "An Interactive Machine-Learning-Based Electronic Fraud and Abuse Detection System in Healthcare Insurance," *Appl. Soft Comput.*, vol. 36, pp. 283–299, Nov. 2015, doi: 10.1016/j.asoc.2015.07.018.

[11] H. Farbmacher, L. Löw, and M. Spindler, "An explainable attention network for fraud detection in claims management," *J. Econom.*, vol. 228, no. 2, pp. 244–258, Jun. 2022, doi: 10.1016/j.jeconom.2020.05.021.

[12] P. E. Ataabadi, B. S. Neysiani, M. Z. Nogorani, and N. Mehraby, "Semi-Supervised Medical Insurance Fraud Detection by Predicting Indirect Reductions Rate using Machine Learning Generalization Capability," in *2022 8th International Conference on Web Research (ICWR)*, IEEE, May 2022, pp. 176–182. doi: 10.1109/ICWR54782.2022.9786251.

[13] Y. Yoo, D. Shin, D. Han, S. Kyeong, and J. Shin, "Medicare fraud detection using graph neural networks," in *2022 International Conference on Electrical, Computer and Energy Technologies (ICECET)*, IEEE, Jul. 2022, pp. 1–5. doi: 10.1109/ICECET55527.2022.9872963.

[14] K. Kaushik, A. Bhardwaj, A. D. Dwivedi, and R. Singh, "Machine Learning-Based Regression Framework to Predict Health Insurance Premiums," *Int. J. Environ. Res. Public Health*, vol. 19, no. 13, Jun. 2022, doi: 10.3390/ijerph19137898.

[15] M. Hanafy and R. Ming, "Machine Learning Approaches for Auto Insurance Big Data," *Risks*, vol. 9, no. 2, Feb. 2021, doi: 10.3390/risks9020042.

[16] N. Dhieb, H. Ghazzai, H. Besbes, and Y. Massoud, "A Secure AI-Driven Architecture for Automated Insurance Systems: Fraud Detection and Risk Measurement," *IEEE Access*, vol. 8, pp. 58546–58558, 2020, doi: 10.1109/ACCESS.2020.2983300.

[17] N. Rayan, "Framework for Analysis and Detection of Fraud in Health Insurance," in *2019 IEEE 6th International Conference on Cloud Computing and Intelligence Systems (CCIS)*, IEEE, Dec. 2019, pp. 47–56. doi: 10.1109/CCIS48116.2019.9073700.

[18] W. Liang, S. Luo, G. Zhao, and H. Wu, "Predicting Hard Rock Pillar Stability Using GBDT, XGBoost, and LightGBM Algorithms," *Mathematics*, vol. 8, no. 5, May 2020, doi: 10.3390/math8050765.

[19] A. I. Alrais, "Fraudulent Insurance Claims Detection Using Machine Learning," 2022.

[20] M. Hanafy and R. Ming, "Using machine learning models to compare various resampling methods in predicting insurance fraud," *J. Theor. Appl. Inf. Technol.*, vol. 99, no. 12, pp. 2829–2833, 2021.