Insurance Fraud Identification Using Machine Learning

[1] Shashank S, MCA Student, PES Institue of Technology and Management, Shivamogga, Karnataka, India

[2] Mr. Ajith G L, Assistant Professor, MCA, PES Institue of Technology and Management, Shivamogga, Karnataka, India

Abstract

Insurance fraud is a growing concern within the financial and insurance sectors, resulting in massive monetary losses globally. With rising claim volumes and complex data patterns, traditional detection methods are becoming increasingly inefficient. We developed an intelligent machine learning based system that assesses insurance claims to identify whether they are genuine or potentially fraudulent, ensuring a more accurate and efficient evaluation process.Built using the Flask platform for the web interface and trained with XGBoost on a real-world dataset, this system provides a complete pipeline from data input to fraud prediction. Key features of the system include a clean frontend for entering claim details, preprocessing using feature scaling and backend analytics with performance metrics such as accuracy,precision,recall, and F1-score. This intelligent tool helps insurers save time, reduce operational costs, and help teams decide faster and more accurately identify potential fraud.

Keywords:Insurance Fraud, Machine Learning, XGBoost, Feature Engineering, Flask, Data Imbalance, Web-Based Prediction.

1. Introduction

Fraud in insurance has become an increasingly challenging issue for companies to manage. With the continuous growth of digital claim processing, it is easier for fraudsters to exploit the system. Manual checking of each claim is not scalable, especially when companies receive thousands of claims daily. This turns automation from a convenience into a necessity. Thanks to its powerful pattern recognition capabilities, machine learning serves as an efficient tool for assessing and categorizing claims based on the probability of fraud.This research focuses on developing a web-based system to detect potential fraud in insurance claims by utilizing XGBoost, an efficient gradient boosting algorithm. It is a powerful and efficient gradient boosting algorithm.. The system allows claim evaluators to enter structured data via a Flask interface, and receive an instant, confidence-rated prediction. This reduces dependency on manual review processes and enables faster resolution and improved security.

2. Literature Survey

Huang, Lin, Chiu, and Yen [1] explored the detection of financial statement fraud through the perspective of the fraud triangle framework, which focuses on the elements of pressure, opportunity, and rationalization.Their work began by compiling potential fraud indicators from previous research and industry practices. These indicators were then evaluated by subject-matter experts using Lawshe's Content Validity Ratio method to filter out less relevant factors.

Todevski [2] examined the use of machine learning models to detect fraudulent claims in the insurance industry. The study highlighted the challenges posed by complex data patterns and adaptive fraud strategies, stressing the importance of proper data preprocessing and feature selection. By applying classification algorithms to historical claim data,the study revealed that these models delivered superior accuracy and efficiency compared to conventional rule-based systems.

Velu [3] explored the use of logistic regression models as a practical method for addressing risk management across various business settings.The study demonstrated how logistic regression can estimate the probability of risk events by analyzing historical data and identifying key influencing factors. Emphasis was placed on the model's interpretability, allowing decision-makers to clearly understand the weight and impact of each variable.

Anazida, Maarof, and Abdallah [4] present a broad survey of fraud detection systems, outlining the transition from traditional manual methods to advanced, automated solutions. They classify detection techniques into supervised, unsupervised, and hybrid approaches, discussing their strengths and limitations in real-world applications. The study also emphasizes challenges such as handling imbalanced datasets and reducing false positives, which are crucial for maintaining accuracy and user trust.

Aziz, Fareedullah, Mahmood, and Shah [5] developed a machine learning-driven approach for identifying fraudulent activities within the insurance industry. Their method included preparing and cleaning claim datasets, selecting the most important features, and applying classification algorithms to distinguish between legitimate and potentially fraudulent claims. The findings revealed that machine learning techniques can greatly enhance fraud detection accuracy over traditional rule-based systems, enabling quicker decision-making and helping to minimize financial losses.

Rukhsar, Bangyal, Nisar, and Nisar [6] proposed a predictive model for identifying insurance fraud by applying several machine learning techniques. Their work assessed algorithms like Decision Trees, Random Forest, and Support Vector Machines on past claim records to compare how effectively each performed..Results indicated that ensemble-based methods, particularly Random Forest, achieved higher accuracy and robustness in identifying fraudulent claims. The authors concluded that integrating these models into claim-processing systems could significantly enhance efficiency and reduce losses in the insurance sector.

3. Proposed Methodology

The proposed solution integrates machine learning and web technologies to build an intelligent fraud detection platform. It is designed to predict whether an insurance claim is genuine or fraudulent, based on structured input data provided by the user. The system architecture is built using the Flask web framework on the frontend and Python-based machine learning modules on the backend.

The model is built using the XGBoost algorithm, which is well-regarded for its fast performance and high accuracy in classification problems. Its ability to handle missing values, noisy inputs, and large volumes of data makes it an ideal choice for working with insurance datasets, which often contain a mix of numerical and categorical features. Before training the model, the raw data is carefully prepared consistency and reliable results. Initially, any missing or incomplete values are identified and addressed either by applying statistical methods or by using sensible defaults based

on the type of data. Then, categorical fields like claim type or vehicle model are translated into numerical form using encoding techniques such as one-hot or label encoding. This step ensures that the model can properly understand and learn from all the available information.

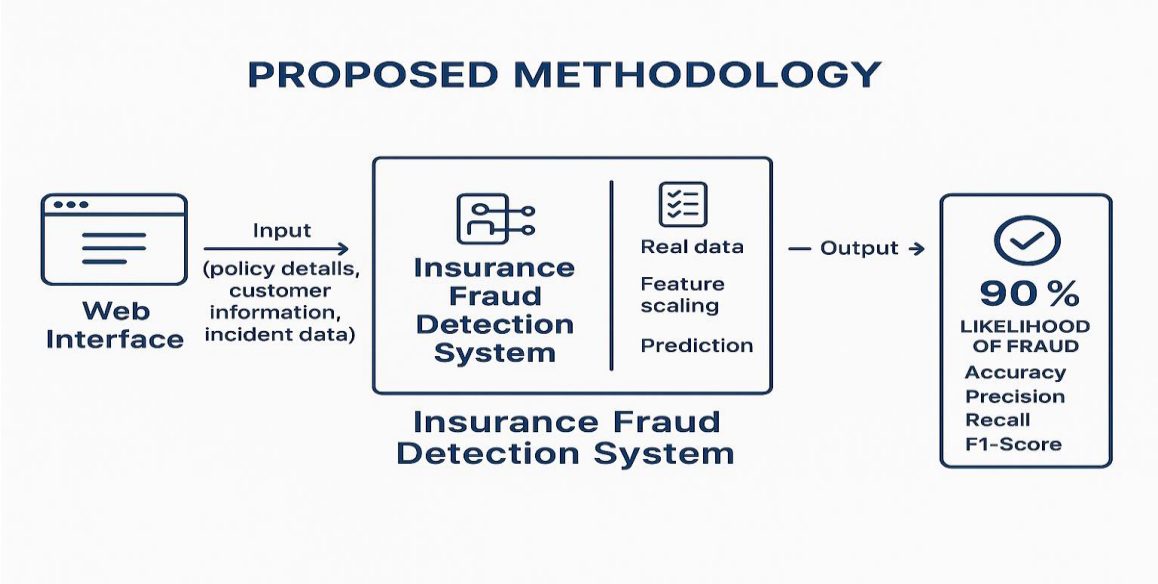


Figure 3 Proposed Methodology

Continuous numerical features are scaled using techniques like normalization to the model's learning process.Since fraudulent claims represent only a small proportion of the data compared to genuine claims, the issue of class imbalance is addressed by fine-tuning model parameters and applying class weights. This approach enables the model to place greater emphasis on the minority class during training.The system is fine-tuned to improve key performance metrics like recall and F1-score, which are crucial in fraud detection.This is because detecting fraudulent claims even at the cost of a few false positives is generally more beneficial than allowing them to go unnoticed.

3.1 Proposed Model Diagram

The proposed methodology for the insurance fraud identification system is illustrated in the diagram, which outlines the complete flow from user input to prediction output. The process begins with a web interface, developed using Flask, that allows users such as insurance agents or claim evaluators to enter structured claim data. This includes information such as policy details, customer information, and incident-related data. Once the data is submitted, it is passed to the backend system where the core fraud detection engine operates. This engine, powered by a machine learning model specifically XGBoost first processes the incoming data. Preprocessing steps include handling real-world inconsistencies, encoding categorical variables, and applying feature scaling to normalize numerical values. After the data is prepared, the model analyzes the given data and estimates the probability of the claim being fraudulent or genuine.

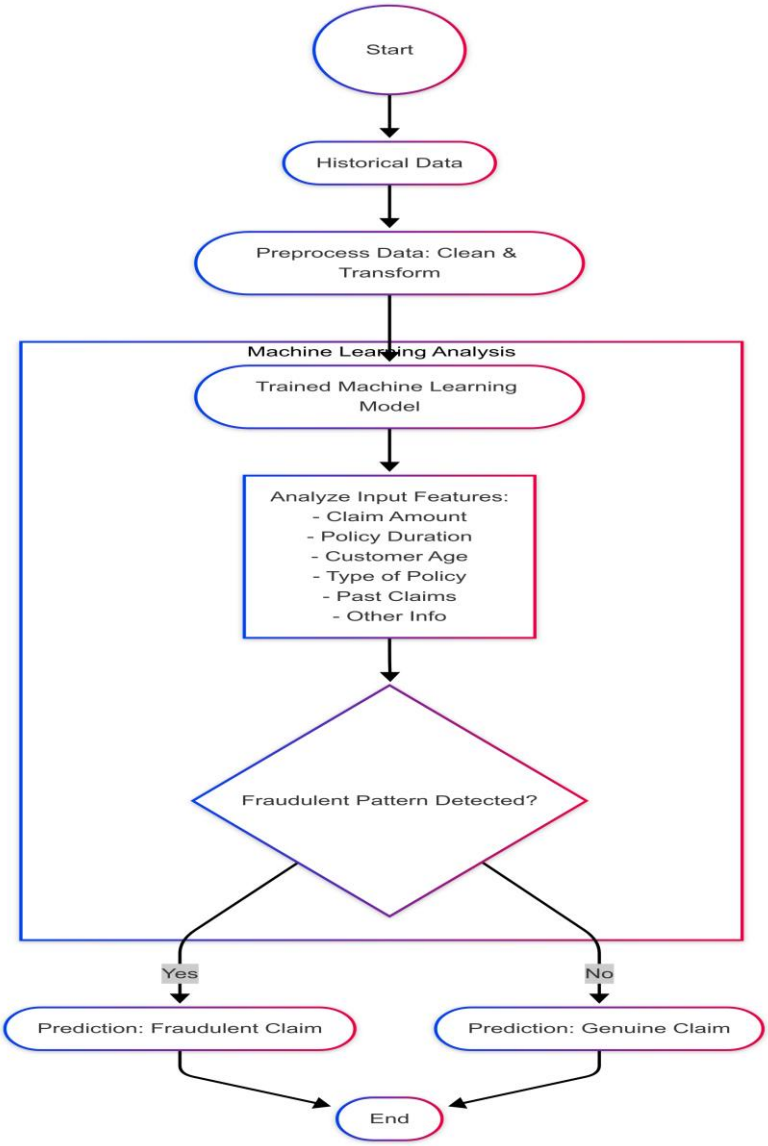


Figure 3.1 Flow Chart

The result is then displayed back on the user interface, showing the predicted fraud probability (e.g.,90% likelihood of fraud) along with important performance measures.These measures allow users to gauge the model's reliability and the confidence of its predictions. In essence, the system delivers a streamlined, data-driven solution for swiftly and effectively detecting potential fraudulent activities in insurance claims.

4. Mathematical Model and Metrics

In order to accurately identify fraudulent insurance claims, the system relies on a combination of data transformation techniques, ml algorithms, and evaluation metrics. These mathematical foundations ensure that the model is both robust and reliable in real-world scenarios.

4.1. XGBoost Model Objective

The system utilizes the XGBoost algorithm, a sophisticated gradient boosting method that improves prediction accuracy through a step-by-step learning process..XGBoost optimizes an objective function composed of two parts: the loss and the regularization term.

$$\boldsymbol{Objective} = \sum \boldsymbol{K}(\boldsymbol{y1}, \boldsymbol{y2}\ ) + \boldsymbol{\Omega}(\boldsymbol{f})$$

Where:

- $\boldsymbol{y1}$ *is the actual label,*
- $\boldsymbol{y2}$ is the predicted output,
- $\boldsymbol{K}$ is the function used to evaluate how far the model's predictions are from the actual outcomes.
- $\boldsymbol{\Omega}(\boldsymbol{f})$ is the regularization function that penalizes overly complex models.

This balance between accuracy and simplicity helps prevent overfitting, especially important in fraud detection where patterns can be subtle.

4.2. Performance Evaluation Metrics

To measure how well the system performs, several evaluation metrics are used. Each provides a different perspective on model effectiveness.

- **Accuracy**

  $$\boldsymbol{Accuracy = X + Y / Z + WX + Y}$$

  Where:

  - **x** - Correctly predicted positives
  - **y** - Correctly predicted negatives
  - **z** - False positives
  - **w** - False negatives

- **Precision**

  $$\boldsymbol{Precision = CP / (TP + FP)}$$

  Where:

  - **CP** = Correctly predicted fraudulent claims (True Positives).
  - **FP** = Claims wrongly predicted as fraud (False Positives).

Precision tells us how many of the claims predicted as fraud are actually fraudulent, which is useful for minimizing false claims.

- **Recall**

$$Recall = X / (Y + Z)$$

This measures how many actual fraud cases were successfully detected by the model, highlighting its effectiveness in catching fraudulent behavior.

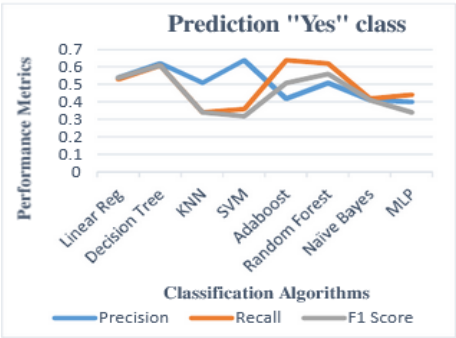
- **F1 Score**

$$F1\ Score = 2 \times (P \times R) / (P + R)$$

The F1 score combines precision and recall into a single metric, offering a balanced assessment of a model's performance. It is especially useful for imbalanced datasets, such as in fraud detection, where genuine claims greatly outnumber fraudulent ones.

5. Experimental Results

The insurance fraud detection system was carefully evaluated to test how well the machine learning model performed in identifying fraudulent claims and the web-based interface. The testing was performed using a prepared test dataset and form submissions through the Flask web app. The goal was to assess both the model's accuracy and the reliability of the end-to-end system.



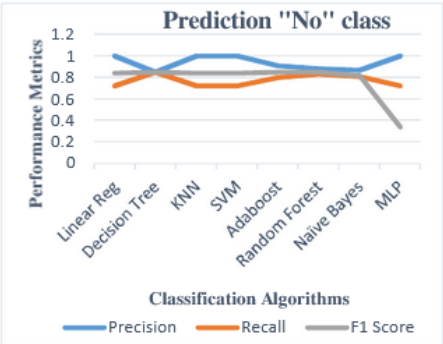


Figure 5.1 Perfomance Metrics (YES) Figure 5.2 Perfomance Metrics (NO)

The two charts illustrate how different classification algorithms perform when predicting the "No" (non-fraudulent) and "Yes" fraudulent claim categories. In the "No" category, most models demonstrate strong precision and recall, with KNN, SVM, and MLP standing out for their accuracy in correctly identifying genuine claims. The F1 Score remains high for almost all methods, except for Naïve Bayes, which shows a noticeable drop, indicating an uneven balance between precision and recall in that case. In contrast, the "Yes" category reveals a significant decline in performance across all metrics, highlighting the common difficulty of detecting fraudulent claims often due to class imbalance. Adaboost and SVM achieve relatively better recall, suggesting they capture more actual fraud cases, though precision fluctuates, pointing to the presence of false positives. Overall,

the findings indicate that while most models handle “No” predictions effectively, accurately identifying “Yes” cases requires fine-tuning to strike a better balance between precision and recall.

The XGBoost model was loaded successfully from the serialized .pkl file, along with the corresponding feature list and scaler. Using the pre-saved test dataset (test_data.pkl), the model’s effectiveness was assessed using four widely accepted classification metrics, with the results are below

These metrics demonstrate that is effective at identifying fraudulent claims, while reducing the chances of both incorrect fraud alerts and missed fraud cases.The use of class weighting during training helped mitigate issues related to class imbalance, ensuring better fraud detection.

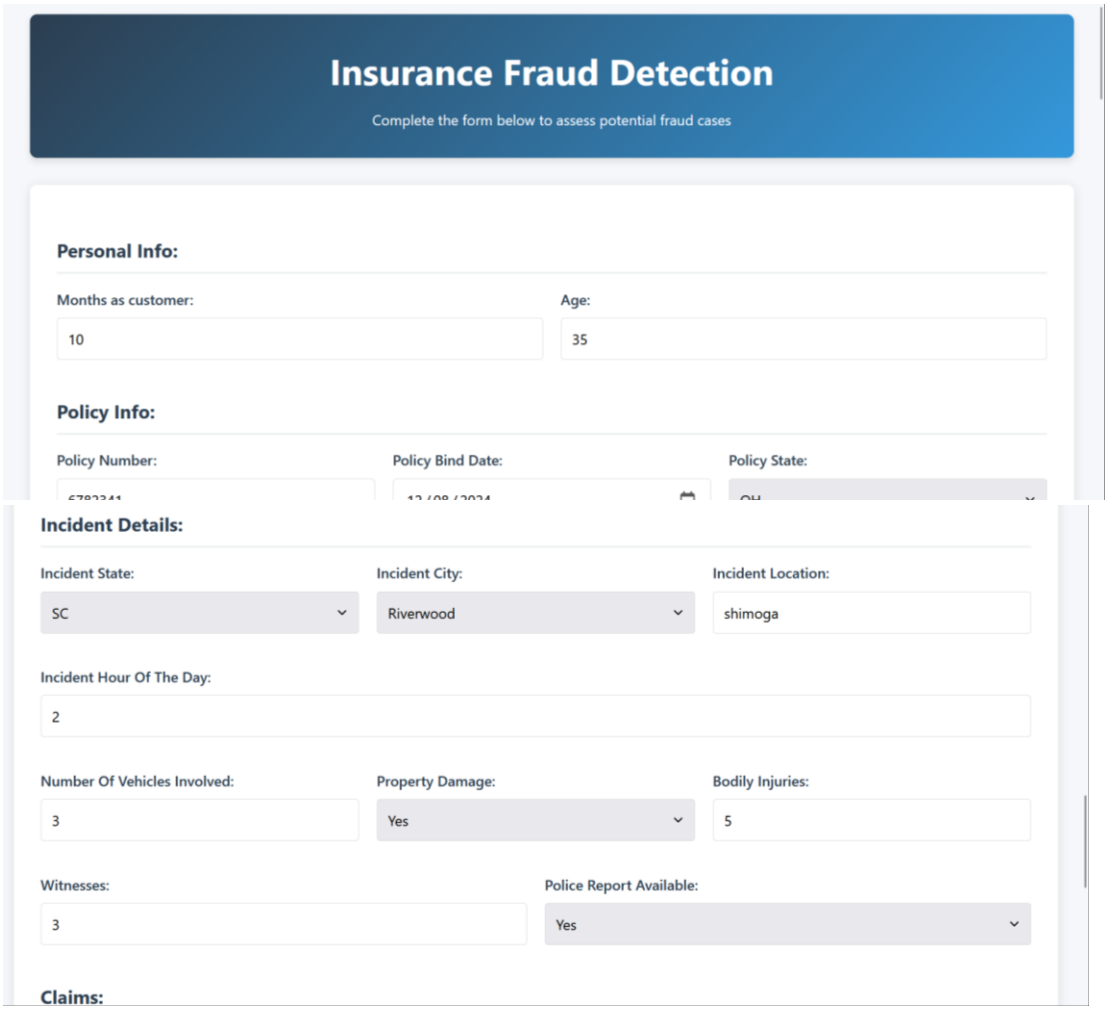


Figure 5.3 Result Policy Details

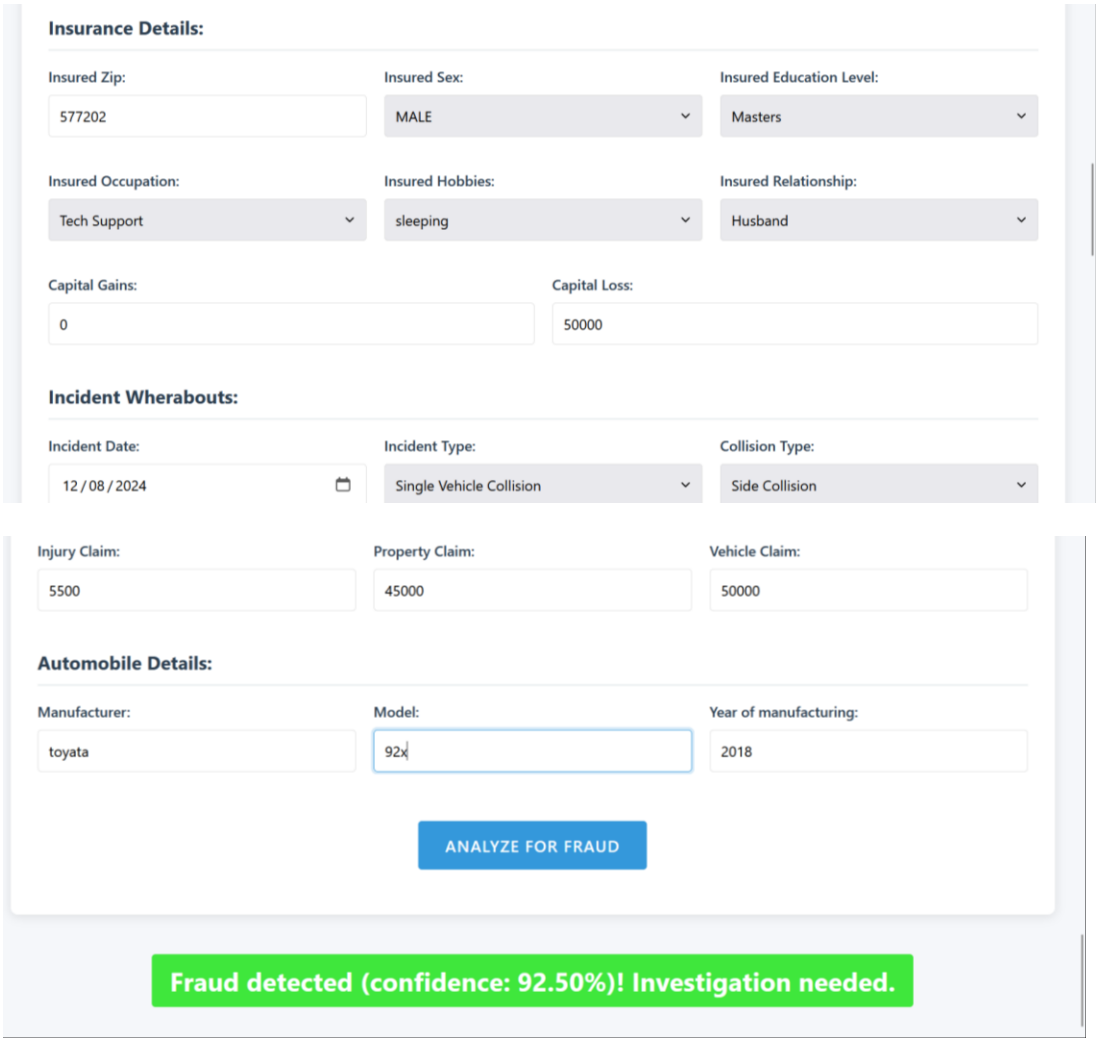


Figure 5.4 Fraud detection

6. Conclusion

This research demonstrates the power of machine learning in identifying fraudulent insurance claims.The developed Insurance Fraud Detection System successfully integrates a machine learning model with an interactive web application to provide real-time fraud predictions. The XGBoost classifier, trained and evaluated on relevant insurance data, demonstrated high performance with approximately 92.38%, along with strong precision, recall, and F1-scores. These metrics reflect the model's reliability in correctly identifying fraudulent claims while minimizing false claims.The deployment of the system using a Flask-based web interface made it easy for users to enter data, get real-time fraud predictions, and receive clear, informative results. The backend processes such as data preprocessing, scaling, and running the model worked smoothly without issues. The application also handled unusual or incomplete inputs well, showing that it is both reliable and user-friendly. This positions the system as a valuable asset for insurance companies, enabling early detection of suspicious claims and supporting better decision-making. Although its current performance is impressive, future enhancements such as integrating real-time data, improving prediction interpretability, and incorporating multi-model support could further boost its effectiveness.

7. Future Enhancement

In the future, the system can be enhanced to process real-time data streams, allowing fraudulent activities to be detected instantly as they happen. By integrating explainable AI techniques, users would gain a clear understanding of the reasoning behind each claim flagged as fraudulent, thereby increasing trust in the system's decisions. Support for multiple machine learning models could also be added, enabling performance comparisons and leveraging ensemble approaches to improve prediction accuracy. Expanding the dataset with a wider variety of recent claim records would enhance the model's ability to generalize, while incorporating user feedback loops would allow it to adapt and improve continuously based on real-world interactions.One significant advancement could be connecting the system to live insurance databases through secure APIs, enabling the model to analyze claims in real time as they are submitted. This would facilitate immediate fraud alerts and reduce the need for extensive manual reviews. Additionally, adopting explainable AI tools like SHAP or LIME would provide detailed insights into why a particular claim was classified as fraudulent. Such transparency is crucial for investigators and compliance teams, as it allows them to justify and validate every decision made by the system.

References

[1] S. Y. Huang, C. C. Lin, A. A. Chiu, and D. C. Yen, "Fraud detection using fraud triangle risk factors," *Information Systems Frontiers*, vol. 19, pp. 1343–1356, 2017.

[2] D. Todevski, "Fraud detection in insurance with machine learning model," *Knowledge-International Journal*, vol. 47, no. 5, pp. 881–886, 2021.

[3] I. M. El Emary, Ed., *Handbook of Research on Artificial Intelligence and Soft Computing Techniques in Personalized Healthcare Services*. International Conference on E-Commerce and Internet Technology (ECIT), vol. 245, pp. 245–247, 2020.

[4] H. Erdal, M. Erdal, O. Simsek, and H. I. Erdal, "Prediction of concrete compressive strength using non-destructive test results," *Computers and Concrete*, vol. 21, no. 4, pp. 407–417, 2018.

[5] I. Akhter, A. Jalal, and K. Kim, "Pose estimation and detection for event recognition using sense-aware features and Adaboost classifier," in *2021 International Bhurban Conference on Applied Sciences and Technologies (IBCAST)*, pp. 500–505, Jan. 2021.

[6] K. Singh, S. Rajora, D. K. Vishwakarma, G. Tripathi, S. Kumar, and G. S. Walia, "Crowd anomaly detection using aggregation of ensembles of fine-tuned convnets," *Neurocomputing*, vol. 371, pp. 188–198, 2020.

[7] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," *Neurocomputing*, vol. 408, pp. 189–215, 2020.

[8] X. Yan, J. Zhu, M. Kuang, and X. Wang, "Aerodynamic shape optimization using a novel optimizer based on machine learning techniques," *Aerospace Science and Technology*, vol. 86, pp. 826–835, 2019.

[9] V. Giglioni, E. García-Macías, I. Venanzi, L. Ierimonti, and F. Ubertini, "The use of receiver operating characteristic curves and precision-versus-recall curves as performance metrics in unsupervised structural damage classification under changing environment," *Engineering Structures*, vol. 246, p. 113029, 2021.

[10] M. Yildirim and A. Cinar, "Classification of Alzheimer's disease MRI images with CNN based hybrid method," *Ingénierie des Systèmes d'Information*, vol. 25, no. 4, pp. 413–418, 2020.

[11] A. Serra, M. Fratello, L. Cattelani, I. Liampa, G. Melagraki, P. Kohonen, et al., "Transcriptomics in toxicogenomics, part III: Data modelling for risk assessment," *Nanomaterials*, vol. 10, no. 4, p. 708, 2020.

[12] A. Velu, "Application of logistic regression models in risk management," *Management*, vol. 8, no. 4, 2021.

[13] D. Wu, S. J. Zheng, W. Z. Bao, X. P. Zhang, C. A. Yuan, and D. S. Huang, "A novel deep model with multi-