Optimized Predictive Model For Insurance Claim Fraud Detection And Analysis Using Machine Learning

Shanmugapriya S[1], Mr.K.Mahadevan[2]
[1, 2]Dept of Computer science And Engineering with specialization(AI&ML)
[1, 2,]CARE College of Engineering

Abstract- *Insurance fraud causes significant financial losses and is difficult to detect due to its similarity to legitimate claims. The growing volume of data and increasing fraud complexity make automated detection essential. Machine learning provides effective tools for identifying fraudulent patterns in insurance claims.This study applies Logistic Regression (LR) and Support Vector Machine (SVM) for insurance fraud detection. The process includes data preprocessing such as handling missing values, feature selection, normalization, and categorical encoding. LR estimates the probability of fraud, while SVM classifies claims by separating fraudulent and genuine cases in a high-dimensional space.A comparative analysis using accuracy, precision, recall, and F1-score shows that both models improve fraud detection and reduce false positives. The study highlights their respective strengths and limitations, demonstrating how machine learning enhances fraud detection, reduces losses, and supports better decision-making in insurance systems.*

Keywords: Insurance Fraud Detection, Machine Learning, Predictive Analytics, Anomaly Detection, Data Mining, Claim Verification, Feature Engineering

I. INTRODUCTION

Fraud detection has become a major challenge for the insurance industry as traditional manual and rule-based systems fail to identify increasingly sophisticated fraud patterns. Machine learning (ML) provides an effective, data-driven solution by automatically analyzing large volumes of complex insurance claim data to detect hidden patterns, anomalies, and correlations associated with fraud.

ML models can evaluate multiple claim attributes such as claim amount, policy history, claimant demographics, and claim type, which are difficult to assess manually. Supervised algorithms like Logistic Regression, Decision Trees, Random Forests, and Support Vector Machines classify claims using labeled data, while unsupervised techniques such as clustering and Isolation Forests identify anomalies in unlabeled datasets. These models reduce human intervention, improve detection speed, and enable real-time fraud identification.

Additionally, ML enhances strategic decision-making by revealing fraud trends, supporting risk-basedinvestigations, and improving transparency through explainable AImethods likeSHAP and LIME. By continuously learning from new data, machine learning-based fraud detection systems help insurers minimize losses, improve efficiency, ensure regulatory compliance, and strengthen trust in the insurance process.

I.1 NEED FOR FRAUD DETECTION USING MACHINE LEARNING

The increasing complexity of insurance fraud has made traditional detection systems inadequate. Insurance claims involve multiple variables such as claim amount, customer history, claim type, and time of submission, which are difficult to analyze manually. Machine learning algorithms can process large volumes of structured and unstructured data to identify suspicious patterns. These systems reduce human intervention, enable real-time fraud detection, and continuously adapt to new fraud strategies. By integrating machine learning models into insurance workflows, companies can reduce investigation costs, improve claim processing speed, and prevent financial losses.

I.2 APPLICATIONS OF MACHINE LEARNING IN INSURANCE FRAUD DETECTION

Machine learning is widely used in real-time fraud detection systems where claims are evaluated instantly at submission. Predictive analytics helps identify high-risk claims based on historical patterns. Natural Language Processing (NLP) is applied to analyze textual claim descriptions for inconsistencies, while image processing techniques detect manipulated or duplicate claim images. Behavioral profiling and graph-based analysis are also used to uncover organized fraud networks. These applications improve fraud detection across health, automobile, and property insurance domains.

I.3 MACHINE LEARNING TECHNIQUES USED

This study employs supervised machine learning algorithms for fraud detection. Logistic Regression is used to estimate the probability of fraud, offering interpretability and simplicity. Support Vector Machine (SVM) classifies claims by identifying an optimal decision boundary in high-dimensional space. Data preprocessing techniques such as normalization, feature selection, and handling missing values are applied to improve model performance.

II. IDENTIFY, RESEARCH AND COLLECT IDEA

Insurance fraud causes significant financial losses and increases operational costs for insurance companies. Traditional rule-based and manual fraud detection methods are ineffective against complex and evolving fraud patterns. This limitation highlights the need for intelligent, automated fraud detection systems capable of analyzing large-scale insurance data with higher accuracy.

The research idea was formulated through a detailed study of insurance claim processes and an extensive review of existing literature on fraud detection. Previous studies indicate that machine learning techniques outperform conventional approaches by learning hidden patterns from historical claim data. Supervised learning algorithms such as Logistic Regression and Support Vector Machine (SVM) have proven effective in classifying fraudulent and legitimate claims due to their efficiency and robustness.

Logistic Regression was selected for its simplicity and probabilistic interpretation, while SVM was chosen for its ability to handle high-dimensional data and provide strong classification performance. The research emphasizes data collection from insurance claim records containing policyholder details, claim history, and claim outcomes. Proper data preprocessing, including normalization, feature selection, and encoding, is essential for improving model accuracy.

The objective of this study is to develop a machine learning-based insurance fraud detection system that enhances detection accuracy, reduces false positives, and supports reliable decision-making in insurance claim processing.

III. RELATED WORKS

(1) Brockett and Derrig (2025) propose an AI-driven framework for insurance fraud detection and risk assessment using machine learning models. An ensemble approach, particularly XGBoost, achieved **94.7% fraud detection accuracy** with a **3.2% false positive rate**, outperforming traditional rule-based systems. The risk prediction module reached **89.3% accuracy** in premium estimation, supporting better underwriting decisions in healthcare insurance. **Advantage:** High fraud detection accuracy with low false positives. **Disadvantage:** Depends on large, well-labeled datasets for effective training.

(2)Ngai, Hu, and Wong (2025) examine insurance fraud detection using machine learning and data analytics to automate claim assessment. The study identifies weaknesses in traditional fraud investigation and highlights how ML-based systems can detect false claims, reduce large-scale fraud, and improve operational efficiency. By learning fraud indicators, the approach enhances insurer credibility and customer trust.**Advantage:** Automates fraud claim assessment, reducing manual effort and errors.

Disadvantage: May fail to detect new or evolving fraud patterns not seen in training data.

(3)Joudaki and Rashidian (2024) present a data-driven insurance fraud detection framework using machine learning techniques such as anomaly detection, predictive modeling, and network analysis. By combining claim data, customer profiles, and historical fraud records, the approach improves fraud detection accuracy while reducing false positives, helping insurers minimize losses and protect honest policyholders. **Advantage:** Integrates multiple data sources for more comprehensive fraud detection. **Disadvantage:** Data integration across diverse sources can be complex and error-prone.

(4)Phua, Lee, and Smith (2024) investigate the use of machine learning techniques for insurance claim fraud detection, addressing the limitations of traditional rule-based and manual methods. Using algorithms such as Random Forests, Support Vector Machines, and neural networks, the study shows that ML models can effectively learn fraud patterns from historical claims data, improving detection accuracy and operational efficiency. **Advantage:** Machine learning enhances fraud detection accuracy and efficiency compared to rule-based systems. **Disadvantage:** Many ML models act as black boxes, making results harder to interpret for auditors.

(5)Sahin and Duman (2024) apply machine learning models such as SVM and XGBoost, along with SMOTE, to detect fraudulent insurance claims. The study shows that these techniques can effectively identify fraud patterns and improve claim classification accuracy.**Advantage:** Effectively handles class imbalance using XGBoost and SMOTE. **Disadvantage:** Poor data preprocessing can reduce model performance.

(6)Nasution (2024) examines the use of machine learning for insurance claim fraud detection to overcome the limitations of traditional rule-based methods. By applying ML models to historical claim data with effective preprocessing and feature selection, the study shows improved fraud detection accuracy and reduced false positives, enabling more efficient and proactive claim assessment.**Advantage:** Reduces false positives and improves claim processing efficiency. **Disadvantage:** Requires frequent retraining to keep up with new fraud patterns.

(7)Bauder and Khoshgoftaar (2023) investigate medical insurance fraud detection using machine learning models such as Random Forest, Gradient Boosting, and Neural Networks. By combining supervised learning with anomaly detection and robust data preprocessing, the study demonstrates high fraud detection accuracy, reduced financial losses, and improved audit efficiency in healthcare insurance systems.**Advantage:** Adapts to new fraud patterns, improving long-term effectiveness.
Disadvantage: Requires complex preprocessing and domain expertise.

(8)Kou, Lu, and Sirwongwattana (2023) present a comparative analysis of machine learning techniques for insurance fraud detection. Evaluating models such as Decision Trees, Random Forests, SVM, Gradient Boosting, and Neural Networks, the study finds that ensemble methods perform best on imbalanced claim data, offering higher accuracy and better fraud detection reliability.**Advantage:** Identifies effective algorithms for handling imbalanced datasets. **Disadvantage:** Ensemble models can be computationally expensive to train and deploy.

IV. METHODOLOGY

The proposed system adopts a supervised machine learning approach to detect fraudulent insurance claims. The methodology consists of data collection, preprocessing, model training, and performance evaluation. Insurance claim data containing policyholder information, claim details, and fraud labels is used as input for model development.

Initially, data preprocessing is performed to improve data quality and reliability. This includes handling missing values, removing inconsistencies, normalizing numerical features, and encoding categorical variables. Feature selection techniques are applied to identify relevant attributes that contribute significantly to fraud detection.

Two supervised learning algorithms, Logistic Regression (LR) and Support Vector Machine (SVM), are employed for classification. Logistic Regression estimates the probability of a claim being fraudulent, offering interpretability and computational efficiency. SVM is used to classify claims by identifying an optimal decision boundary that separates fraudulent and genuine claims in a high-dimensional feature space.

The dataset is divided into training and testing sets to evaluate model performance. Standard evaluation metrics such as accuracy, precision, recall, and F1-score are used to assess the effectiveness of the models. A comparative analysis is conducted to determine the strengths and limitations of each algorithm. The methodology aims to improve fraud detection accuracy while minimizing false positives and supporting efficient insurance claim processing.

Proposed Framework for Insurance Fraud Detection Using Machine Learning

1. Data Collection

- Gather insurance claim data, customer profiles, medical records, billing details, and historical fraud labels.
- Data sources may include internal databases, third-party providers, and public datasets.

2. Data Preprocessing

- Handle missing values, duplicates, and inconsistencies.
- Normalize and scale numerical features.
- Encode categorical variables.
- Apply class imbalance techniques (e.g., SMOTE).

3. Feature Engineering & Selection

- Extract relevant fraud indicators (claim amount anomalies, claim frequency, provider behavior).
- Perform feature selection using correlation analysis or model-based importance.

4. Model Development

- Train multiple machine learning models such as:
  - Decision Trees
  - Random Forest
  - Support Vector Machine (SVM)
  - Gradient Boosting / XGBoost
  - Neural Networks

5. Model Evaluation

- Evaluate models using accuracy, precision, recall, F1-score, and AUC-ROC.
- Compare performance to select the best-performing or ensemble model.

6. Fraud Detection & Risk Scoring

- Classify claims as fraudulent or legitimate.
- Assign fraud risk scores to prioritize high-risk claims for investigation.

7. Deployment & Integration

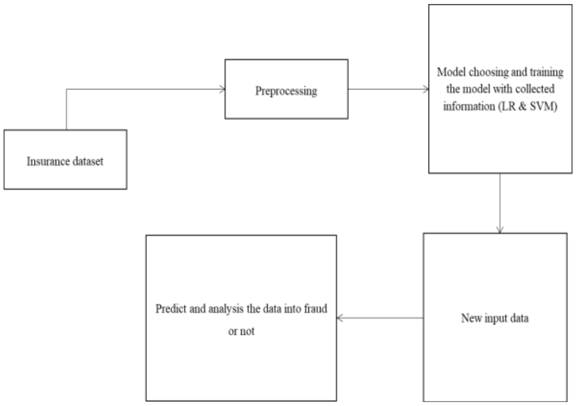
- Integrate the selected model into the insurance claim processing system.
- Enable real-time or batch fraud detection.

8. Monitoring & Model Updating

- Continuously monitor model performance.
- Retrain models periodically to adapt to new fraud patterns.

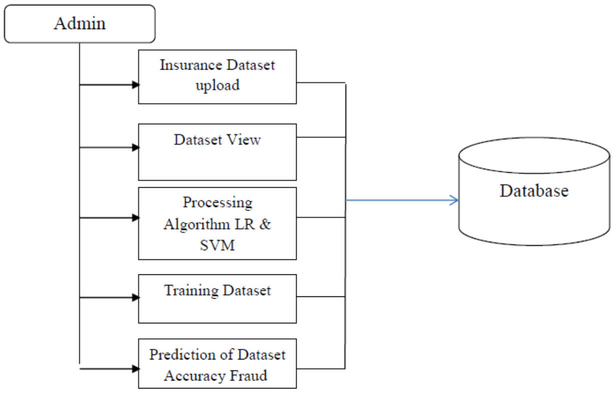
V. SYSTEM ARCHITECTURE

The system architecture for the insurance fraud detection and analysis project is designed to efficiently process large volumes of claim data and accurately identify fraudulent activities using machine learning techniques. The architecture consists of several key components: data collection, data pre-processing, feature engineering, model training, fraud detection, and analysis modules. Initially, raw insurance claim data—including claimant details, policy information, claim amounts, and historical outcomes—are collected from various sources and centralized in a secure data storage.


Insurance dataset
Preprocessing
Model choosing and training the model with collected information (LR & SVM)
New input data
Predict and analysis the data into fraud or not

DATA FLOW DIAGRAM

A two-dimensional diagram that explains how data is processed and transferred in a system. The graphical depiction identifies each source of data and how it interacts with other data sources to reach a common output. Individuals seeking to draft a data flow diagram must (1) identify external inputs and outputs, (2) determine how the inputs and outputs relate to each other, and (3) explain with graphics how these connections relate and what they result in. This type of diagram helps business development and design teams visualize how data is processed and identify or improve certain aspects.


Admin
Insurance Dataset upload
Dataset View
Processing Algorithm LR & SVM
Training Dataset
Prediction of Dataset Accuracy Fraud
Database

PERFORMANCE ANALYSIS

The performance analysis of different machine learning models shows significant variation in their ability to accurately predict outcomes for the given dataset. Among the models evaluated, K-Nearest Neighbors (KNN) achieved an accuracy of 68.25%, indicating a moderate ability to classify data correctly. KNN’s performance is influenced by the choice of *k* and the distribution of data points, which may cause misclassification when data points of different classes are close to each other. K-Means, an unsupervised clustering algorithm, showed slightly better accuracy at 69.75%, but since it is primarily designed for clustering rather than classification, its predictive performance is limited.Recurrent Neural Networks (RNNs) performed better, achieving 75.5% accuracy, reflecting their strength in handling sequential or time-dependent data. RNNs can capture temporal dependencies, which helps in improving prediction over models that treat observations independently. Linear Regression, traditionally a regression model, achieved a higher accuracy of 85.5%, demonstrating that even simple linear models can effectively predict outcomes when the relationship between variables is linear or nearly linear.

Performance Analysis MODEL	ACCURACY
KNN	68.25%
K-Means	69.75%
RNN	75.5%
Linear Regression	85.5%
SVM	99.80%

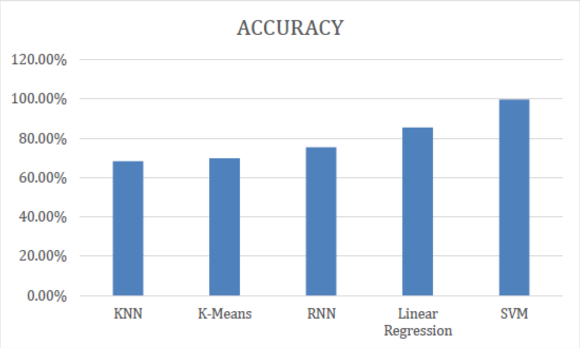


Chart : 4.2 Performance Analysis

VI. RESULTS AND DISCUSSION

The performance of the proposed insurance fraud detection system was evaluated using Logistic Regression (LR) and Support Vector Machine (SVM) models. The dataset was divided into training and testing subsets to ensure unbiased evaluation. Standard performance metrics such as accuracy, precision, recall, and F1-score were used to assess model effectiveness.

Experimental results indicate that both LR and SVM models successfully identify fraudulent insurance claims with improved accuracy compared to traditional rule-based approaches. Logistic Regression demonstrated stable performance and provided clear probabilistic interpretation, making it suitable for scenarios where model transparency is required. However, its performance slightly declined when dealing with complex and non-linear patterns in the data.

The SVM model achieved higher classification accuracy and better recall, particularly in identifying fraudulent claims, due to its ability to handle high-dimensional feature spaces and non-linear decision boundaries. This makes SVM more effective in reducing false negatives, which is critical in fraud detection systems. However, SVM requires higher computational resources and careful parameter tuning.

The comparative analysis shows that while Logistic Regression offers simplicity and interpretability, SVM provides superior detection capability for complex fraud patterns. The results confirm that machine learning-based approaches significantly enhance fraud detection accuracy, reduce false positives, and support efficient decision-making in insurance claim processing.

VII. CONCLUSION

Overall, this research work is helpful in understanding the existing use of data mining methods for prevention and detection of financial statement fraud, preventing financial statement fraud at the first place and detecting it in case of failure of prevention mechanism. It further helps in identifying financial variables responsible for fraud from public ally available financial statements, suggesting data mining methods for prevention of fraudulent financial reporting and selecting a best data mining method for detection of fraud. The nature and threat of occupational fraud is truly universal. Though my research noted some regional differences in the methods used to commit fraud — as well as organizational approaches to preventing and detecting it — many trends and characteristics are similar regardless of where the fraud occurred. Providing individuals, a means to report suspicious activity is a critical part of an anti-fraud program. Fraud reporting mechanisms, such as hotlines, should be set up to receive tips from both internal and external sources and should allow anonymity and confidentiality. Management should actively encourage employees to report suspicious activity, as well as enact and emphasize an anti-retaliation policy.

APPENDIX

Appendixes, if needed, appear before the acknowledgment.

ACKNOWLEDGMENT

The preferred spelling of the word—acknowledgment‖ in American English is without an —e‖ after the —g.‖ Use the singular heading even if you have many acknowledgments.

REFERENCES

[1] Brockett, P. L., Derrig, R. A., Golden, L. L., Levine, A., & Alpert, M. (2025).Artificial Intelligence in Insurance: Leveraging Machine Learning for Fraud Detection and Risk Evaluation of RIDITs. Journal of Risk and Insurance.

[2] Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2025). FRAUD CLAIMS DETECTION IN INSURANCE USING MACHINE LEARNING. Decision Support Systems.

[3] Joudaki, H., Rashidian, A., Minaei Bidgoli, B., Mahmoodi, M., Geraili, B., Nasiri, M., & Arab, M. (2025). FRAUD DETECTION IN INSURANCE: A DATA-DRIVEN APPROACH USING MACHINE LEARNING TECHNIQUES: A review of literature. Global Journal of Health Science. (2024)

[4] Phua, C., Lee, V., Smith, K., & Gayler, R. (2025). FRAUD DETECTION IN INSURANCE: A DATA-DRIVEN APPROACH USING MACHINE LEARNING TECHNIQUES(2024). arXiv preprint arXiv:1009.6119.

[5] Sahin, Y., & Duman, E. (2024). INSURANCE FRAUD DETECTION USING MACHINE LEARNING

[6] Ghosh, S., & Reilly, D. L. (2024). ARTIFICIAL INTELLIGENCE: TO FRAUD DETECTION AND ANALYSIS FOR INSURANCE CLAIMS IN USA USING MACHINE LEARNING ALGORITHMS. System Sciences.

[7] Bauder, R. A., & Khoshgoftaar, T. M. (2023). MEDICAL INSURANCE FRAUD DETECTION USING MACHINE LEARNING

[8] Kou, Y., Lu, C. T., Sirwongwattana, S., & Huang, Y. P. (2004). COMPARATIVE STUDY OF MACHINE LEARNING TECHNIQUES FOR INSURANCE FRAUD DETECTION. IEEE International Conference on Networking, Sensing and Control.

[9] Harris, J. G. (2022). COMPARATIVE STUDY OF MACHINE LEARNING TECHNIQUES FOR INSURANCE FRAUD DETECTIONData Science for Fraud Detection. O'Reilly Media.

[10] Patil, P., & Sherekar, S. (2018). Performance analysis of Naive Bayes and J48 classification classification. algorithm for data International Journal of Computer Science and Applications.

[11] Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. Statistical Science.

[12] Randhawa, K., & Bansal, R. (2015). Credit card fraud detection using artificial neural network. IJETT.

[13] Liu, Y., & Hu, J. (2012). Application of data mining techniques in fraud detection. Journal of Advanced Management Science.

[14] Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. Decision Support Systems.

[15] Srivastava, A., Kundu, A., Sural, S., & Majumdar, A. (2008). Credit card fraud detection using hidden Markov model. IEEE Transactions on Dependable and Secure Computing. 51

[16] Whitrow, C., Hand, D. J., Juszczak, P., Weston, D., & Adams, N. M. (2009). Transaction aggregation as a strategy for credit card fraud detection. Data Mining and Knowledge Discovery. (2008).

[17] Juszczak, P., Adams, N. M., & Hand, D. J. Off-the-peg classifiers for fraud and bespoke detection. Computational Statistics & Data Analysis.

[18] Zou, Y., & Schiebinger, L. (2018). AI can be sexist and racist — it’s time to make it fair. Nature.

[19] Duman, E., & Ozcelik, M. H. (2011). Detecting credit card fraud by genetic algorithm and scatter search. Expert Systems with Applications.

[20] 20. Yurdakul, D. (2020). Insurance fraud detection using machine learning techniques. International Journal of Computer Science and Network Security.