Fraud Detection And Analysis For Insurance Claims Using Machine Learning

Dr. M. Kishore Kumar[1],Golla Madhu[2], Guguloth Arun Kumar[3], Bhukya Dilip Kumar[4]
[1, 2, 3, 4] Dept of CSE(DS)
[1, 2, 3, 4] CMR Technical Campus(UGC Autonomous), Kandlakoya, Medchal,Telangana,India

Abstract- *Insurance fraud, particularly in claim processing, results in billions of dollars in losses annually for companies worldwide. Manual and rule-based detection mechanisms are often inefficient in detecting sophisticated fraud schemes. This study proposes an automated system that leverages machine learning (ML) algorithms to classify insurance claims as genuine or fraudulent. Supervised learning models, such as Logistic Regression, Support Vector Machine (SVM), Decision Tree, Naïve Bayes, and SGD, were trained and evaluated. An ensemble model using a Voting Classifier outperformed the individual classifiers. The system was deployed using Django on a WAMP server, which integrated real-time prediction and user access control.*

Keywords- Insurance Fraud Detection, Machine Learning, Classification Algorithms, Django Framework, Predictive Analytics, Supervised Learning, Voting Classifier, Data Preprocessing.

I. INTRODUCTION

The insurance sector plays a pivotal role in stabilizing the financial system. However, it has become increasingly vulnerable to fraud, with fraudulent claims accounting for a significant portion of global insurance payouts. Insurance fraud includes activities such as inflating claims, misrepresenting facts, and staging events to obtain insurance benefits. These deceptive practices not only lead to direct financial losses but also strain administrative resources and inflate premiums for honest policyholders.Traditional fraud detection systems rely heavily on manually defined rules, heuristics, and red-flag indicators. While such systems can detect obvious or well-known fraud patterns, they struggle to identify emerging or complex fraud schemes, especially in large datasets with high dimensionality. Machine learning (ML) introduces a data-driven approach that can learn patterns from historical claims data to identify subtle signals of fraud. By training models on labeled claim data, we can develop predictive systems that can classify incoming claims with greater speed and accuracy. Furthermore, the use of ensemble methods increases robustness by leveraging the strengths of multiple models. This study details the development of an intelligent fraud detection system using multiple ML models, compares their performance, and deploys the best-performing model through a web-based interface for real-world usability. This work includes preprocessing techniques, model evaluation, system architecture, and implementation using Django and MySQL.

II. RELATED WORK

Insurance fraud detection research spans statistical modeling, expert systems, and, more recently, machine learning and deep learning. Early methods focused on logistic regression, decision trees, and rule-based expert systems for pattern recognition in claims processing. While interpretable, such models were limited in handling nonlinear relationships and evolving fraud tactics. Belhadji et al. (2000) proposed a probabilistic model for auto insurance fraud detection, setting a foundation for statistical detection. Subsequently, research expanded into data mining and ML-based classification methods. For instance, Khadse et al. (2018) compared several supervised learning algorithms for IoT data and emphasized their application in fraud analytics. Similarly, Li et al. (2009) compared primitive classifiers with advanced learners such as Extreme Learning Machines (ELMs), emphasizing scalability and accuracy.Recent advances introduced ensemble methods like Random Forests and Gradient Boosting, which outperform single classifiers by reducing variance and bias. However, the challenge remains in real-time implementation, model interpretability, and adaptability to changing fraud patterns.Deep learning methods such as neural networks and autoencoders have also been explored for anomaly detection, but they often require extensive computational resources and large training datasets, which may not be feasible for all insurance providers.Our work builds upon these foundations by integrating classical machine learning models with web-based deployment, aiming to strike a balance between accuracy, speed, interpretability, and practicality. Additionally, we utilize ensemble methods and address issues such as data imbalance using SMOTE, providing a comprehensive solution suitable for real-world deployment.

III. METHODOLOGY

3.1 Data Collection and Preprocessing

- Source: Insurance claim datasets from online repositories.
- Cleaning: Null values imputed; categorical variables encoded; features normalized.
- Balancing: SMOTE used to address class imbalance in fraudulent vs. non-fraudulent claims.

3.2 Feature Engineering

Variables such as claim amount, driver age, vehicle type, policy duration, and others were selected based on domain knowledge. A correlation matrix was used to eliminate redundant features.

3.3 Algorithms Used

- Logistic Regression (LR)
- Naive Bayes (NB)
- Support Vector Machine (SVM)
- Stochastic Gradient Descent (SGD)
- Decision Tree (DT)
- Ensemble Method: Voting Classifier (combining multiple models)

IV. SYSTEM DESIGN

4.1 Architecture

A three-tier architecture was used:

- Frontend: HTML/CSS/JavaScript
- Backend: Django (Python)
- Database: MySQL (WAMP Server)

4.2 Modules

- Remote User Registration/Login
- Insurance Claim Entry
- Real-time Prediction
- Admin Dashboard with Graphs
- Dataset Upload and Model Retraining

V. RESULTS AND EVALUATION

Algorithm	Accuracy (%)	Precision	Recall	F1-score
Naive Bayes	~85.0	0.84	0.86	0.85
SVM	~89.0	0.88	0.89	0.89
Logistic Regression	~87.5	0.87	0.87	0.87
SGD Classifier	~86.5	0.86	0.87	0.86
Decision Tree	~88.0	0.88	0.88	0.88
Voting Classifier	~91.2	0.91	0.92	0.91

The ensemble model outperformed individual models, with the highest accuracy and balanced metric performance.

VI. IMPLEMENTATION

The project was deployed with a user-friendly web interface. Key features include:

- Real-time fraud prediction
- Role-based access control
- Graphical dashboards for performance analysis
- Email alerts for high-risk claims

Sample code integrates model prediction with Django views using Voting Classifier.

VII. *CONCLUSION AND FUTURE SCOPE*

7.1 Conclusion

This project successfully demonstrates the use of machine learning for detecting fraudulent insurance claims. Various supervised learning algorithms were implemented and compared, with the ensemble Voting Classifier achieving the best performance. The system was deployed using Django, enabling real-time predictions and user interaction through a web interface. Overall, the approach improves accuracy, reduces manual effort, and supports faster claim processing.

7.2 Future Scope

Future enhancements include adding explainable AI techniques for model transparency, enabling real-time fraud detection, and exploring deep learning for more complex fraud patterns. The system can also be extended to other types of insurance and integrated with mobile and cloud platforms for broader usability.

REFERENCES

[1] K. Ulaga Priya and S. Pushpa, "A Survey on Fraud Analytics Using Predictive Model in Insurance Claims," International Journal of Pure and Applied Mathematics, vol. 114, no. 7, pp. 755–767, 2017.

[2] E. B. Belhadji, G. Dionne, and F. Tarkhani, "A Model for the Detection of Insurance Fraud," Geneva Papers on Risk and Insurance - Issues and Practice, vol. 25, no. 4, pp. 517–538, 2000.

[3] F. C. Li, P. K. Wang, and G. E. Wang, "Comparison of the Primitive Classifiers with Extreme Learning Machine in Credit Scoring," IEEE International Conference on Industrial Engineering and Engineering Management, 2009.

[4] V. Khadse, P. N. Mahalle, and S. V. Biraris, "An Empirical Comparison of Supervised Machine Learning Algorithms for Internet of Things Data," 2018 4th International Conference on Computing, Communication, Control and Automation (ICCUBEA), 2018.

[5] S. Ray, "A Quick Review of Machine Learning Algorithms," Proceedings of International Conference on Machine Learning, Big Data, Cloud and Parallel Computing, 2019.

[6] G. Dal Pozzolo, O. Caelen, Y. A. Le Borgne, S. Waterschoot, and G. Bontempi, "Learned Lessons in Credit Card Fraud Detection from a Practitioner Perspective," Expert Systems with Applications, vol. 41, no. 10, pp. 4915–4928, 2014.

[7] M. Phua, V. Lee, K. Smith, and R. Gayler, "A Comprehensive Survey of Data Mining-Based Fraud Detection Research," arXiv preprint arXiv:1009.6119, 2010.

[8] L. Van Vlasselaer et al., "APATE: A Novel Approach for Automated Credit Card Transaction Fraud Detection Using Network-Based Extensions," Decision Support Systems, vol. 75, pp. 38–48, 2015.

[9] J. West and M. Bhattacharya, "Intelligent Financial Fraud Detection: A Comprehensive Review," Computers & Security, vol. 57, pp. 47–66, 2016.