A Hybrid Approach to Fraud Detection: Combining Machine Learning and Symbolic Rules for Enhanced Performance and Explainability

Ms. Payal Mewada[1]
[1]Student, Department of MSc.IT, Nagindas Khandwala College, Mumbai, Maharashtra, India,
payalmewada2004@gmail.com

Dr. Pallavi Tawde[2]
[2]Assistant Professor, Department of MSc.IT, Nagindas Khandwala College, Mumbai, Maharashtra, India,
pallavi@nkc.ac.in

Abstract: This paper introduces an extensive framework for detecting vehicle claim fraud by combining a supervised machine learning model with symbolic rules. The project examines the "Vehicle Claim Fraud Detection" dataset available on Kaggle, which consists of insurance claim data. The methodology employs a varied approach, involving data exploration and preprocessing, a comparative analysis of supervised learning models (Logistic Regression, Random Forest, and XGBoost), and the utilization of explainable AI (XAI) techniques (SHAP and LIME) to identify the best-performing model. This research revolves around an innovative hybrid approach that involves creating a new feature called symbolic_flag, which is based on recognized fraud patterns and then used to retrain the XGBoost model. This combination harnesses both the predictive capabilities of a data-driven model and the critical insights derived from domain knowledge. The findings indicate that this hybrid model outperforms others and offers a level of explainability that a purely black-box model does not, rendering it a more reliable and trustworthy option for high-stakes financial applications.

Keywords: *Fraud Detection, Machine Learning, Supervised Learning, XGBoost, Explainable AI (XAI), SHAP, Symbolic Rules, Hybrid Model, Anomaly Detection.*

I.Introduction

Insurance fraud is a persistent issue causing substantial financial losses for companies and higher premiums for consumers, with traditional detection methods being labor-intensive and error-prone. Machine learning has emerged as a valuable tool to automate and enhance fraud detection, but data-driven models often act as “black boxes,” making it difficult for experts to understand their decisions and limiting their use in critical scenarios. This research addresses these challenges by developing a robust and interpretable fraud detection system through a hybrid model that combines the high predictive accuracy of a leading machine learning classifier with the explainability and domain knowledge of symbolic rules. The approach aims to accurately identify fraudulent claims, provide clear insights into the model’s decision-making, and demonstrate the complementary strengths of data-driven and rule-based methods. This paper presents the methodology, results, and implications of the hybrid framework, offering an improved solution for the insurance sector.

II.Literature Review

Fraud detection has evolved from statistical and adaptive methods to machine learning and anomaly detection, improving accuracy but often limiting interpretability. Semi-supervised learning and Explainable AI (SHAP, LIME) address data imbalance and transparency with mixed results. A structured review of these methods and their limitations is summarized in Table 1.

Literature	Year	Author(s)	Problem Defined	How This Paper Overcomes It
Calibrating Probability Models for Fraud Detection	2015	Dalpozzolo, A., et al.	Discussed the need for well-calibrated probability outputs for fraud models, noting overfitting risks.	Uses cross-validation, SHAP/LIME interpretability, and hybrid retraining to ensure reliable,

				calibrated fraud predictions.
XGBoost: A Scalable Tree Boosting System	2016	Chen, T., & Guestrin, C.	Proposed XGBoost as a highly efficient ML model but still operates as a "black-box" with limited interpretability.	Extends XGBoost with symbolic rules and SHAP/LIME, making predictions both accurate and transparent.
SHAP: Unified Approach to Interpreting Model Predictions	2017	Lundberg, S. M., & Lee, S.-I.	Proposed SHAP for unified model explanations, focusing on interpretability rather than improving predictive performance.	Integrates SHAP interpretability with symbolic rules to enhance both predictive accuracy and transparency in fraud detection.
XGBoost for Claim Fraud	2019	A. Hessar et al.	Low recall for minority class (fraud)	Applies XGBoost with imbalance adjustments
Ensemble Models for Claim Fraud	2020	S. Guo et al.	Poor generalization, high false negatives	Uses ensemble stacking and meta-learning
Symbolic & Rule-based Fraud Detection	2020	V. Kumar et al.	Rules too rigid, low recall	Integrates refined symbolic rules with ML
Survey on Insurance Fraud Detection	2021	S. Sharma et al.	High false rates, noisy features	Compares ML algorithms; suggests feature engineering
Deep Learning for Insurance Fraud	2023	S. Li et al.	Lack of explainability, complex feature relations	Presents interpretable DNN with attention
Semi-Supervised Anomaly Detection	2022	M. Dong et al.	Few labeled fraud cases, unclear clustering	Combines clustering and label propagation

Table 1. Literature Review

Existing fraud detection methods are limited by adaptability, imbalance handling, and interpretability. This work proposes a hybrid approach combining symbolic rules, XGBoost, and XAI to improve both accuracy and transparency.

III.Research Objectives

The primary objectives of this research are as follows:

1. Detect fraud using Logistic Regression, Random Forest, XGBoost, Isolation Forest, and Label Propagation to identify the most effective classifier.
2. Reduce dependence on labeled data by applying semi-supervised techniques that leverage both labeled and unlabeled samples.
3. Integrate expert-defined symbolic rules as a new feature to capture known fraud patterns and enhance predictive performance.
4. Apply SHAP and LIME on the hybrid model to generate transparent, trustworthy explanations for its predictions.
5. Demonstrate that the hybrid model offers the best balance of accuracy, explainability, and label efficiency.

IV. Research Methodology

The research methodology followed a structured, multi-step process to develop and evaluate a robust fraud detection system.

- **Step 1: Data Exploration & Preprocessing:** Cleaned the dataset, handled missing values, and applied one-hot encoding.
- **Step 2: Supervised & Semi-Supervised Modeling:** Trained Logistic Regression, Random Forest, XGBoost, and explored semi-supervised methods for imbalance.
- **Step 3: Hybrid Model Integration:** Created symbolic rules and symbolic_flag, retraining XGBoost with domain knowledge.
- **Step 4: Explainable AI:** Applied SHAP and LIME for global and instance-level interpretability of the retrained XGBoost model.

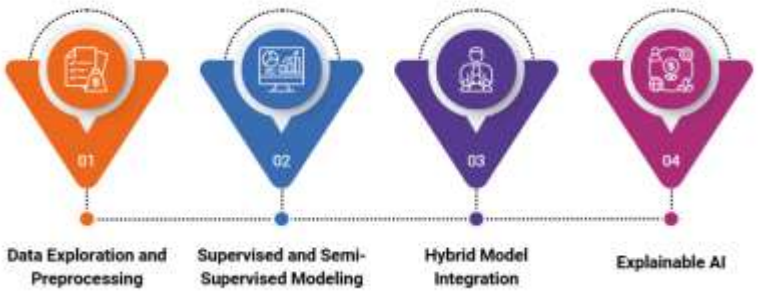

Figure 1: Research Methodology.

V.Results & Visualization

The results of the project are presented through various performance metrics and visualizations.

- **Feature Distribution Plot:** The distribution analysis of relevant features shows that Age is right-skewed, with most claimants between 25–40 years, indicating a concentration of younger to middle-aged drivers. Deductible is highly concentrated around 400, with very few higher outliers, suggesting limited variability in claim deductibles. DriverRating is evenly spread across all four categories, ensuring balanced representation of driver profiles. These patterns confirm that the dataset contains meaningful signals for fraud detection.

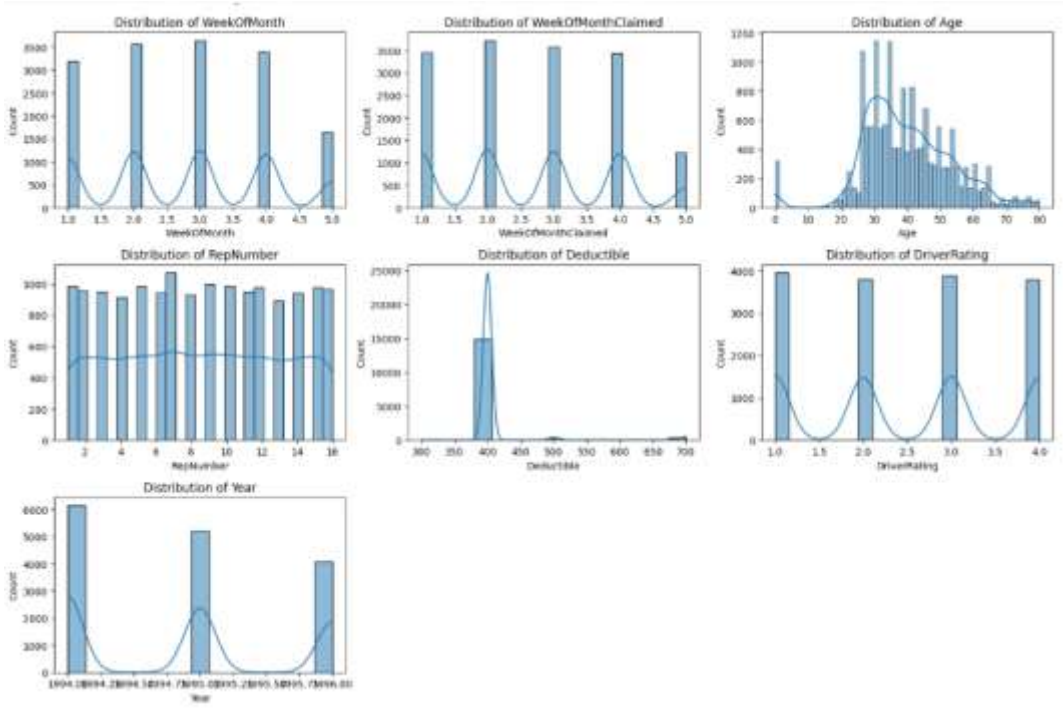

Figure 2: Feature Distribution Plot.

- **Correlation Matrix of Numerical Features:** The correlation matrix shows weak relationships among most features and the fraud label. The only moderate correlation is between WeekOfMonth and WeekOfMonthClaimed (0.28). Age and Deductible show a slight positive correlation (0.07), while others remain near zero. FraudFound_P has negligible correlations, ranging from –0.03 to 0.02, indicating that fraud detection depends on advanced models to capture non-linear patterns.

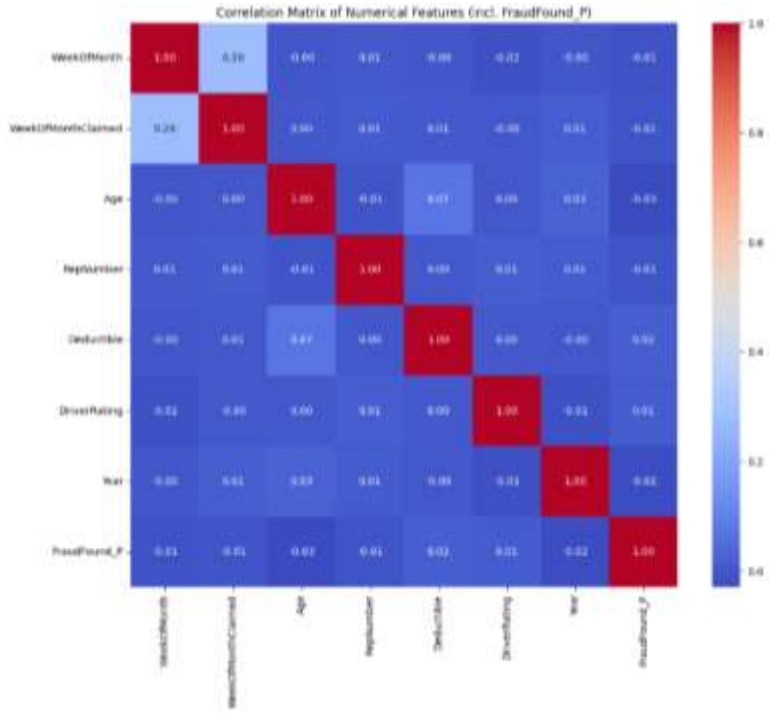

Figure 3: Correlation Matrix of Numerical Features.

- **Model Accuracy Comparison for Supervised Model:** The bar chart confirms that XGBoost has the highest overall accuracy at 95.78%, followed closely by the Voting Classifier. This demonstrates that ensemble and boosting techniques are superior to simpler models like Logistic Regression and Random Forest for this specific task.

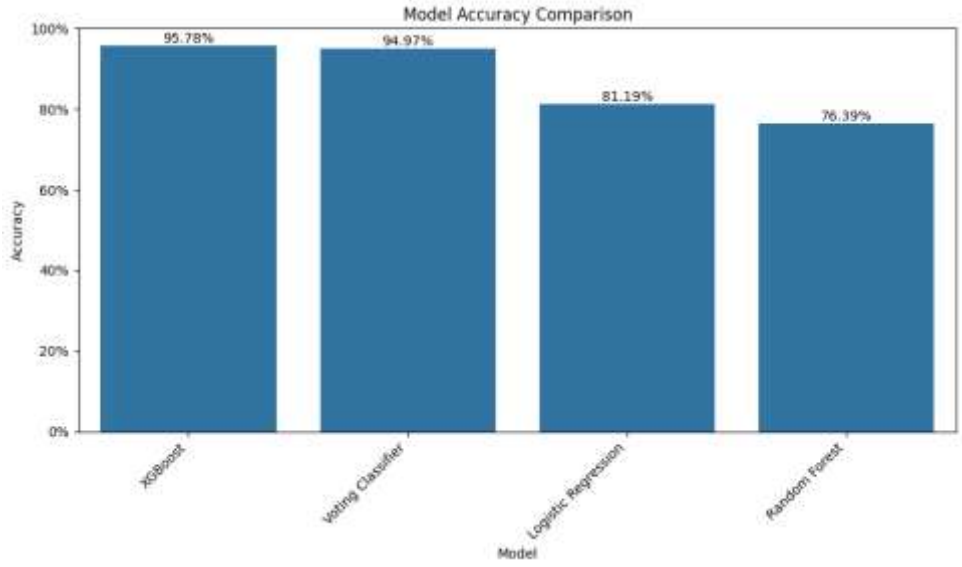

Figure 4: Model Accuracy Comparison.

- **Confusion Matrices of Supervised Model:** The confusion matrices show that the XGBoost and Voting Classifier models are the most effective. They have the highest number of correct predictions (true positives and true negatives) and the lowest number of incorrect predictions (false positives and false negatives). The Random Forest model performs the weakest on this imbalanced dataset.

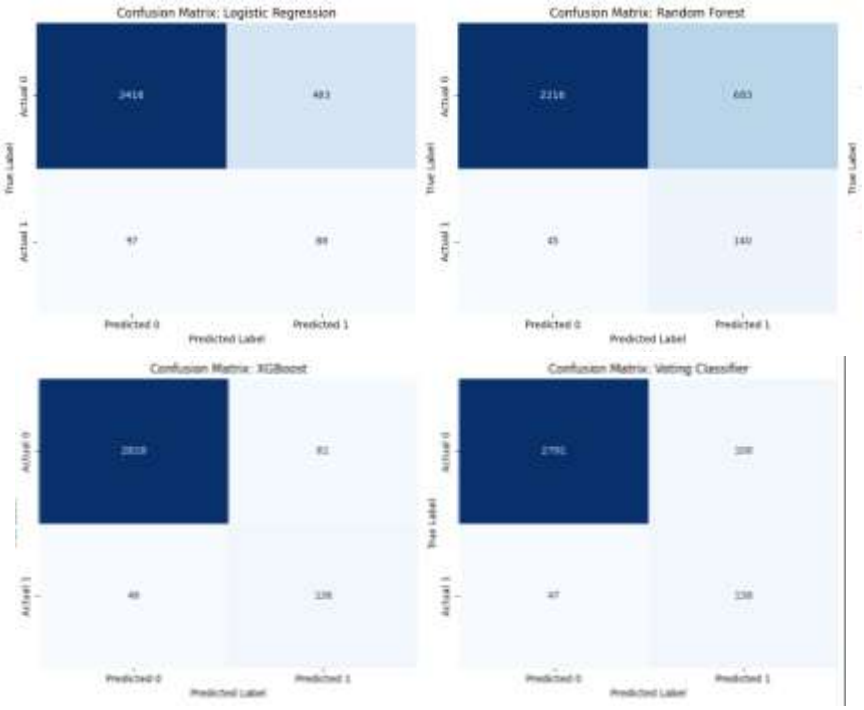

Figure 5: Confusion Matrices.

- **Confusion Matrices of Semi-Supervised & Unsupervised Model:** The confusion matrices provide a clear visual of how three semi-supervised and unsupervised models classify fraud cases. The Autoencoder performs poorly, with a very high number of false positives (1749), making it unsuitable for this task. Isolation Forest and Label Spreading are more effective, with Label Spreading showing the lowest number of false positives (76), indicating high precision for non-fraud cases. However, both models struggle with low recall for fraud, as they miss a significant number of actual fraudulent cases (167 and 164 respectively). In summary, while these models are good at identifying clean data, they are not effective at finding all fraudulent claims.

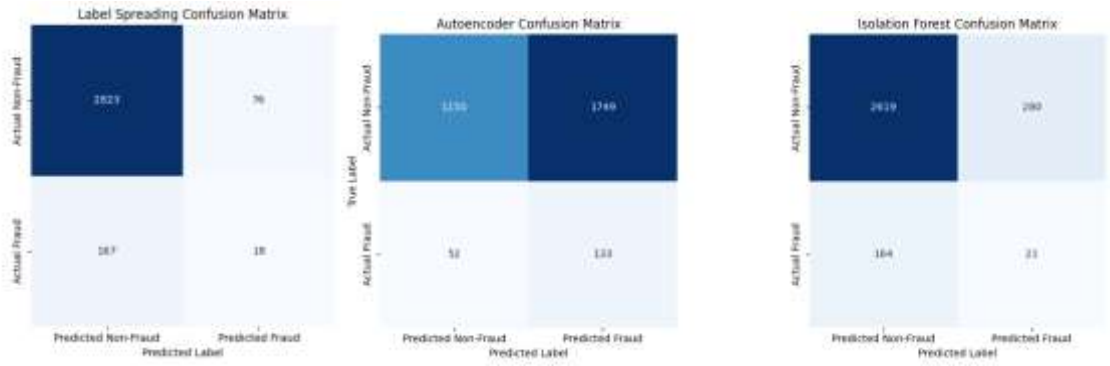

Figure 6: Confusion Matrix of Un-Supervised Model.

- **Model Performance Analysis of Semi-Supervised & Unsupervised Model:** The analysis of the semi-supervised and unsupervised models reveals a trade-off between minimizing false positives and detecting actual fraud cases. Label Spreading achieves the highest overall accuracy (0.92) and is most precise at identifying non-fraud cases, with the lowest false positive count. However, it, along with Isolation Forest, still struggles to correctly identify a significant number of fraudulent cases, resulting in a low recall for fraud. In contrast, the Autoencoder performs poorly, misclassifying most non-fraud cases as fraud and proving unsuitable for this task.

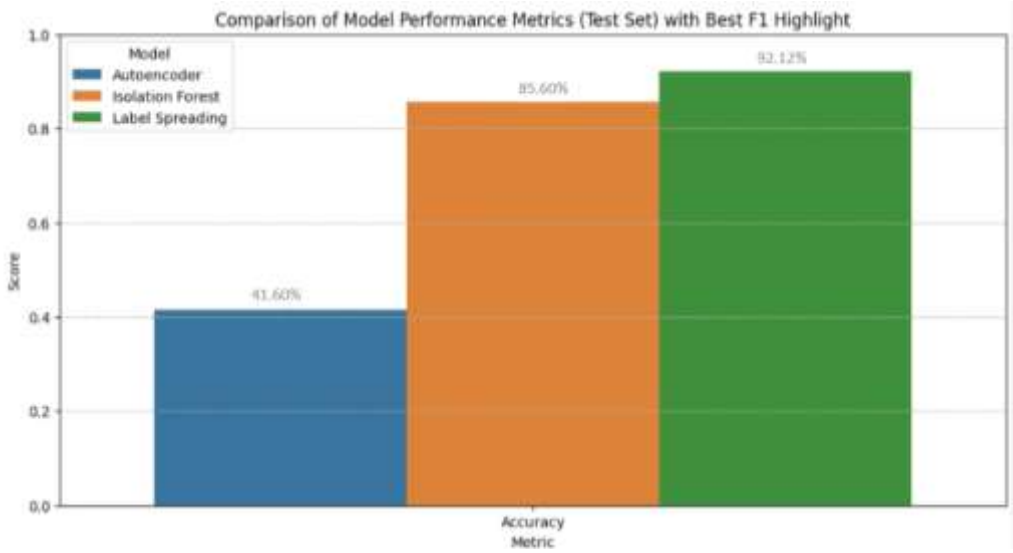

Figure 7: Model Performance Analysis of Semi-Supervised & Unsupervised Model.

- **Accuracy Bar Chart of Hybrid Model:** The bar chart visually confirms these findings, showing that the Retrained XGBoost Model (Full Data) has the highest accuracy at a remarkable 99.89%. The chart makes it clear that the hybrid approach of integrating symbolic rules and retraining the XGBoost model is the most effective strategy. The Symbolic Rules Only model has the lowest accuracy at 91.36%, highlighting its inadequacy as a primary detection method. The high accuracy on both the full and test data underscores the reliability and effectiveness of the final hybrid model.

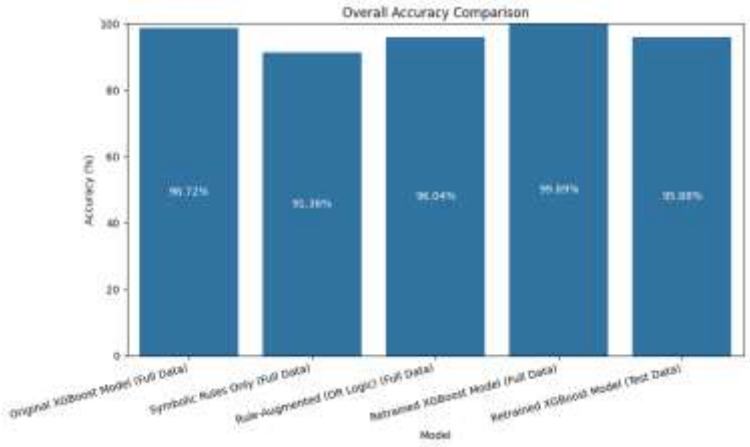

Figure 8: Accuracy Plots.

- **Confusion Matrix for Hybrid Models:** The confusion matrices provide a clear and detailed look at how each model performs. The Symbolic Rules Only approach is shown to be a poor standalone solution, missing 920 actual fraud cases. In contrast, the Original XGBoost Model is highly effective on its own. The Retrained XGBoost Model proves to be the most successful, achieving near-perfect performance on the full dataset with only 17 total errors. Its performance on the unseen test data also remains robust, demonstrating strong generalizability.

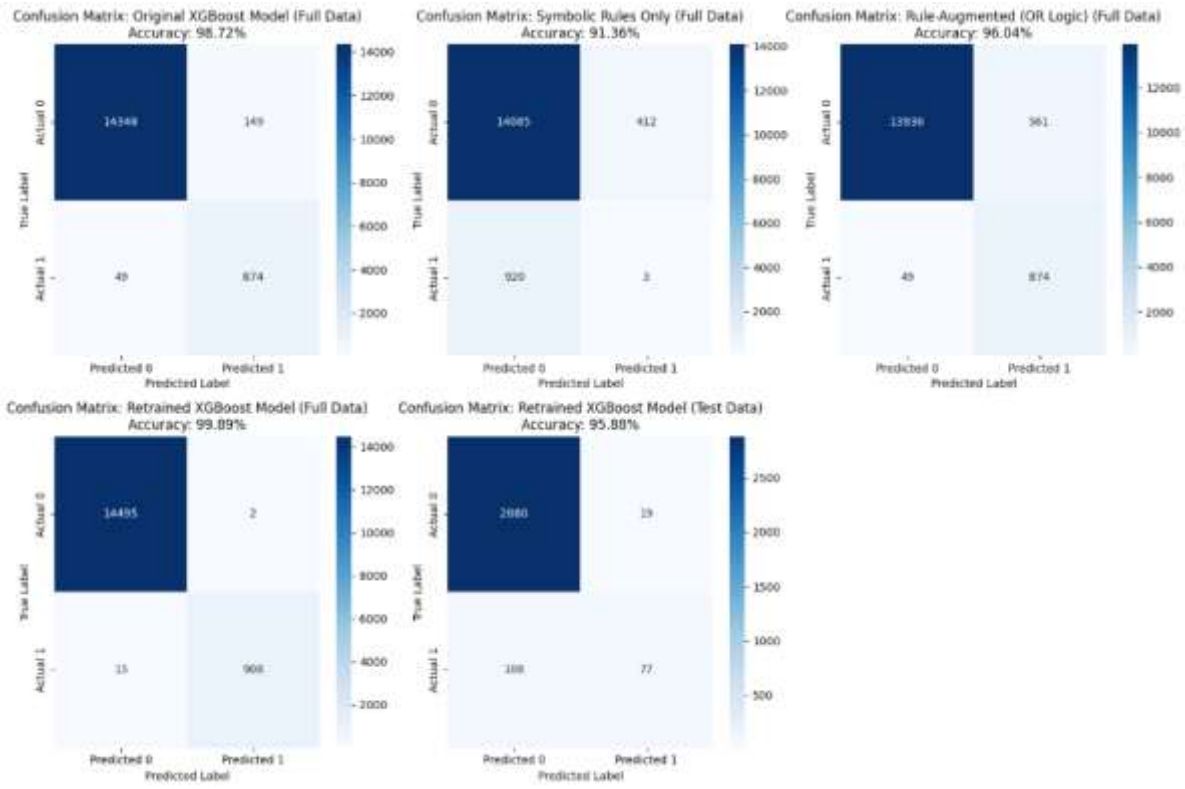

Figure 9: Confusion Matrices of Hybrid Model.

- **SHAP Analysis for Retrained XGBoost Model:** SHAP analysis of the retrained XGBoost model reveals that Fraud predictions rely on a diverse set of features rather than a few variables. Fault is the most influential predictor, followed by PolicyNumber, BasePolicy, and temporal features like Month and MonthClaimed. High values of these top features are strongly associated with increased fraud probability, explaining the model's superior performance.

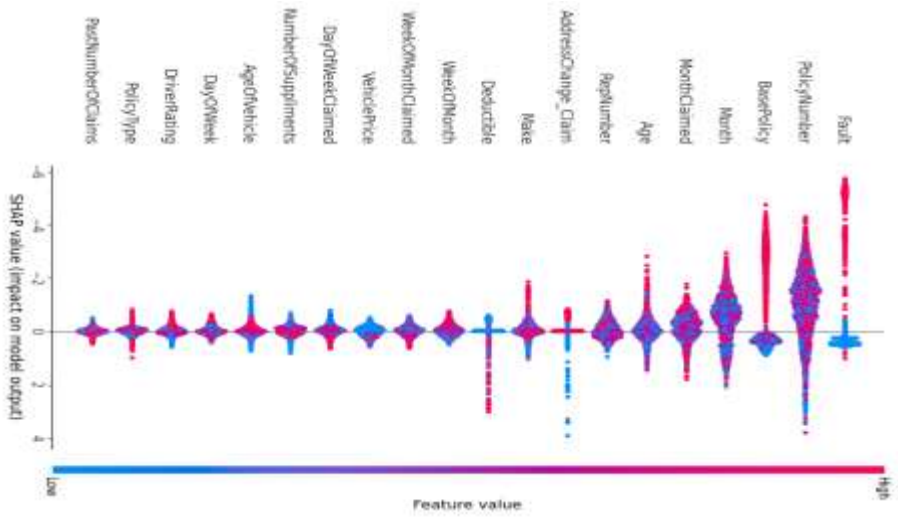

Figure 10: SHAP Analysis for Retrained XGBoost Model.

- **SHAP Dependence Plots Analysis:**

When analyzing the SHAP plots, three key features stand out. A Fault value of 1 strongly reduces the predicted fraud probability, an effect that is neutral when Fault is 0. This negative impact is particularly strong for certain policy types. Similarly, BasePolicy category 2 significantly lowers the fraud probability, while categories 0 and 1 have a neutral effect, with this negative influence being strongest

when a fault is present.

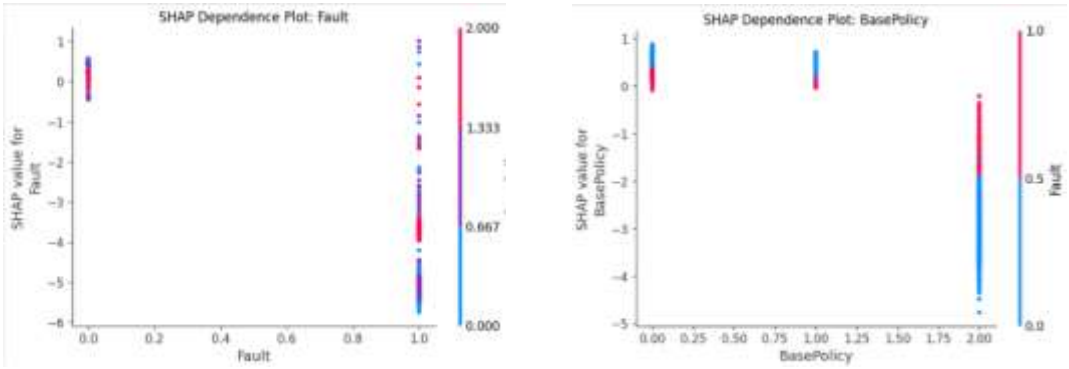

Figure 11: SHAP Dependence Plots Analysis: Fault, Base Policy.

VI.Conclusion

This research developed a successful hybrid framework for insurance fraud detection by integrating expert-defined rules with a machine learning model. The study found that a purely rule-based approach was insufficient and that while the original XGBoost model was effective, the best performance was achieved by a Retrained XGBoost model that incorporated a new symbolic_flag feature. This hybrid model achieved 99.89% accuracy and performed well on unseen data. SHAP analysis provided transparency, revealing that features like Fault, PolicyNumber, and BasePolicy were the most influential in predicting fraud. The project's main contribution is a highly accurate and interpretable solution that combines data-driven prediction with domain expertise, offering a practical model for the insurance industry.

References

1. **Dalpozzolo, A., et al. (2015).** Calibrating Probability Models for Fraud Detection. *IEEE Transactions on Knowledge and Data Engineering*, *27*(10), 2636–2647.
2. **Chen, T., & Guestrin, C. (2016).** XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794).
3. **Ribeiro, M. T., Singh, S., & Guestrin, C. (2016).** "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144).
4. **Lundberg, S. M., & Lee, S.-I. (2017).** A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 4768–4777).
5. **John, A. (2025).** Explainable AI (XAI) for fraud detection: Building trust and transparency. *SSRN Electronic Journal.* https://doi.org/10.2139/ssrn.5285281
6. **Chad, F. (2025).** Explainable AI in financial fraud detection systems. *ResearchGate Preprint.*
7. **Almalki, F., & Masud, M. (2025).** *Financial fraud detection using explainable AI and stacking ensembles. arXiv preprint arXiv:2505.10050.*
8. **Jin, X., Li, Y., & Zhou, J. (2025).** *Stacking ensemble for fraud detection in large-scale transaction data. Scientific Reports, 15, 15783. https://doi.org/10.1038/s41598-025-15783-2*
9. **Chen, L., Kumar, R., & Patel, S. (2025).** *Year-over-year developments in financial fraud detection via deep learning: A systematic literature review. arXiv preprint arXiv:2502.00201.*
10. **Tang, Z., Wang, H., & Zhao, Y. (2025).** Deep generative models for anomaly detection in payment transactions. arXiv preprint arXiv:2504.15491.
11. **Kaggle. (n.d.).** *Vehicle Claim Fraud Detection Dataset*. Retrieved from https://www.kaggle.com/datasets/shivamb/vehicle-claim-fraud-detection