Vehicle Insurance Claim Fraud Detection Using Machine Learning

David Naumovski, Mihajlo Naumoski, Hristijan Gjoreski
Faculty of Electrical Engineering and Information Technologies,
University of Ss. Cyril and Methodius in Skopje, Macedonia
davidnaumovski38@gmail.com, mihajlonaumoski07@gmail.com, hristijang@feit.ukim.edu.mk

Abstract — Detecting fraudulent insurance claims is a critical challenge in the vehicle insurance industry, as undetected fraud leads to significant financial losses and undermines trust in insurance systems. Traditional fraud detection methods are often labor-intensive and inefficient, highlighting the need for automated, data-driven solutions. With the growing availability of comprehensive claim datasets, machine learning offers promising new approaches for identifying suspicious claims. This paper presents a complete machine learning workflow for vehicle insurance claim fraud detection, including data preprocessing, class balancing, and the application of multiple classification algorithms such as Random Forest, Support Vector Machine, and XGBoost, as well as a stacking ensemble. The models were evaluated on a real-world dataset, with particular attention to handling class imbalance and optimizing recall for the minority (fraud) class. Experimental results demonstrate that the stacking ensemble performs better than individual models in achieving higher overall accuracy, but XGBoost slightly outperforms all other models when it comes to recall score. These findings suggest that advanced ensemble methods can significantly enhance the reliability and effectiveness of automated fraud detection systems in practical insurance applications.

Keywords — insurance fraud detection; machine learning; ensemble learning; vehicle claims; classification

I. INTRODUCTION

Today's technological advancements have fundamentally transformed the insurance industry, enabling companies to process vast amounts of data and automate complex tasks such as fraud detection. While these innovations have improved efficiency and customer experience, they have also introduced new challenges, as fraud perpetrators continually adapt their tactics to exploit digital systems [1]. Insurance fraud remains a significant global issue, resulting in substantial financial losses for insurers and increased premiums for honest policyholders [2]. The complexity and scale of modern insurance operations make manual fraud detection impractical, necessitating the adoption of automated, data-driven approaches.

Traditional rule-based systems and manual audits are often insufficient for identifying sophisticated or evolving fraud patterns [3]. The increasing availability of large, high-quality datasets, combined with advances in machine learning, has opened new avenues for automated fraud detection. Machine learning algorithms, such as Support Vector Machines [6], and ensemble methods like Random Forests [5], XGBoost [4], have demonstrated strong performance in detecting anomalies and uncovering hidden patterns within complex insurance data. These techniques enable insurers to identify suspicious claims more accurately and efficiently, reducing losses and maintaining the integrity of the insurance system [1,3].

A major challenge in insurance fraud detection is the highly imbalanced nature of the data, where fraudulent claims represent only a small fraction of all claims [3]. Addressing this imbalance is crucial for developing models that are both sensitive to fraud and robust against false positives. Recent research has shown that ensemble learning and stacking methods [7] can further enhance detection performance by combining the strengths of multiple classifiers.

This study aims to leverage state-of-the-art machine learning techniques to develop a robust workflow for vehicle insurance claim fraud detection. By systematically applying data preprocessing, class balancing, and advanced model training—including ensemble and stacking approaches—this work seeks to provide practical tools for improving fraud detection in real-world insurance applications [1,2,3,4,5,6,7].

II. RELATED WORK

Early approaches to insurance fraud detection relied on manual audits and rule-based systems, which were limited in their ability to adapt to evolving fraud patterns and often resulted in high false positive rates [1][2]. With the advent of machine learning, researchers began applying supervised learning algorithms to automate fraud detection and improve accuracy. Initial studies focused on traditional classifiers such as decision trees and support vector machines (SVM), which demonstrated the potential to uncover complex patterns in insurance data [3][6]. However, these single-model approaches often struggled with the highly imbalanced nature of fraud datasets, where fraudulent claims represent only a small fraction of all claims [3][8].

To address these challenges, ensemble methods such as Random Forests [5] and boosting algorithms like XGBoost [4] were introduced, offering improved performance by combining the strengths of multiple base learners. These methods have been shown to be particularly effective in handling class imbalance and capturing subtle anomalies indicative of fraud [3][5]. Recent research has also explored stacking ensemble techniques, where the predictions of several models are combined using a meta-learner to further enhance detection capabilities [7]. Studies comparing single classifiers to

ensemble and stacking approaches consistently report that ensembles achieve higher recall and overall accuracy, which is critical for minimizing undetected fraud [3][7][8].

In addition to algorithmic advances, the availability of large, high-quality datasets and open-source machine learning libraries such as scikit-learn [9] and platforms like Kaggle [10] have accelerated progress in this field. These resources enable researchers and practitioners to experiment with a wide range of models and preprocessing techniques, fostering the development of robust fraud detection pipelines. Overall, the literature demonstrates that leveraging multiple models and diverse features—along with careful handling of class imbalance—yields the most reliable results for insurance fraud detection. This work builds on these findings by implementing and comparing advanced ensemble methods, including stacking, to maximize the detection of fraudulent vehicle insurance claims.

III. Data

For this research, we utilized a publicly available vehicle insurance claim dataset [10], which is widely used in fraud detection studies due to its comprehensive coverage of real-world claim scenarios. The dataset contains thousands of insurance claim records, each described by a diverse set of features related to the policyholder, vehicle, and claim circumstances. These features include both categorical and numerical variables, providing a rich context for machine learning analysis [1][2].

Key attributes in the dataset include:

- **Policyholder information:** Age, gender, marital status, and address change history, offering demographic and behavioral insights.
- **Vehicle details:** Make, category, price range, and age of the vehicle, which are relevant for assessing claim legitimacy.
- **Claim characteristics:** Accident area, fault assignment, police report status, witness presence, number of supplements, and past number of claims, all of which contribute to the risk profile of each claim [3][8].
- **Financial data:** Deductible amount and policy type, which can influence the likelihood of fraudulent activity [2][5].

The target variable, FraudFound_P, is a binary indicator specifying whether a claim was identified as fraudulent (1) or not (0). The dataset is characterized by a significant class imbalance, with fraudulent claims representing a small minority of the total records. No missing values were present, simplifying preprocessing and ensuring the integrity of subsequent analyses.

Each feature was carefully examined and preprocessed to maximize its utility for machine learning. Categorical variables were encoded using label, ordinal, or one-hot encoding as appropriate, while numerical features were standardized or normalized to ensure comparability [9]. Redundant or non-informative columns, such as unique identifiers and features with uniform distributions, were removed based on exploratory data analysis and domain knowledge [1][4]. This comprehensive dataset provides a robust foundation for developing and evaluating advanced fraud detection models in the insurance domain.

IV. Methods

The processed feature vectors, representing each insurance claim, are used as input to classification algorithms in order to learn models that can accurately distinguish between fraudulent and non-fraudulent claims. The target variable for prediction is FraudFound_P, a binary indicator where 1 denotes a fraudulent claim and 0 denotes a legitimate one. This binary classification task is central to automating fraud detection in the insurance domain.

For the development of our fraud detection models, we explored both single-model and ensemble-based approaches. In the single-model approach, five widely used machine learning algorithms were implemented using the scikit-learn and XGBoost libraries: Decision Tree, k-Nearest Neighbors (kNN), Support Vector Machine (SVM), Random Forest, and XGBoost [5][6][4][9]. Each algorithm was trained on the balanced dataset, with hyperparameters tuned using cross-validation (excluding the test data) to optimize performance, particularly recall for the minority (fraud) class.

The ensemble approach combines the predictions of multiple base learners to produce a more robust and accurate model. Three types of ensemble methods were employed: Bagging, Boosting, and Stacking. Bagging, as implemented by the Random Forest algorithm [5], builds multiple decision trees on random subsets of the training data and aggregates their predictions to improve stability and reduce variance. Boosting, exemplified by XGBoost [4], sequentially trains models so that each new model focuses on correcting the errors of its predecessors, resulting in a strong overall classifier. XGBoost also incorporates regularization and efficient handling of missing values, making it particularly well-suited for complex, real-world datasets.

Stacking, or meta-ensemble learning [7], combines the outputs of several diverse base models—such as SVM, Random Forest, and XGBoost—using a meta-learner to generate final predictions. This approach leverages the strengths of each base model, often outperforming individual classifiers by capturing complementary patterns in the data. In our workflow, the stacking ensemble was constructed with the best-performing tuned models as base learners and an XGBoost classifier as the meta-learner.

Throughout the modeling process, special attention was given to handling class imbalance, feature selection, and hyperparameter optimization. Model evaluation was conducted using metrics such as accuracy, precision, recall, F1-score, confusion matrices, and ROC curves, with a focus on maximizing the detection of fraudulent claims while minimizing false positives. This comprehensive methodology ensures that the resulting fraud detection system is both reliable and practical for deployment in real-world insurance applications.

V. Experimental Setup

To evaluate and compare the performance of different machine learning algorithms for insurance fraud detection, we

employed a stratified train-validation-test split. The dataset was first divided into training (56%), validation (24%), and test (20%) sets using stratified sampling to preserve the original class distribution, which is essential for imbalanced classification problems. This approach ensures that each subset is representative of the overall data and allows for fair model evaluation and comparison.

During model development, the training set was used to fit the models, while the validation set was reserved for hyperparameter tuning and model selection. The final evaluation was performed on the unseen test set to assess the generalization ability of each approach. For ensemble methods such as stacking, cross-validation was used within the training data to optimize the meta-learner and prevent overfitting.

For evaluation metrics, we focused on accuracy, precision, recall, F1-score, and the area under the ROC curve (AUC), as these are standard metrics for binary classification tasks [3][5]. The recall metric, in particular, was emphasized due to the critical importance of minimizing false negatives in fraud detection—missing a fraudulent claim can have significant financial consequences. Confusion matrices were also analyzed to provide insight into the types of errors made by each model.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\text{F1 Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Predictions}}$$

All experiments were conducted using the scikit-learn and XGBoost libraries [4][9], with hyperparameter optimization performed via randomized search and cross-validation. This rigorous experimental setup ensures that the results are robust, reproducible, and directly applicable to real-world insurance fraud detection scenarios.

VI. Experimental Results

The initial evaluation involved training five baseline classifiers—Decision Tree, K-Nearest Neighbors, SVM, Random Forest, and XGBoost—on the balanced training data. Their performance was assessed on the validation set using accuracy, precision, recall, F1-score, and confusion matrices. The results are summarized in **Table 1** and visualized in **Figure 1** (recall bar plot) and **Figure 2** (confusion matrices, two per row).

	Model	Precision	Recall	F1-score	Accuracy
0	DecisionTree	0.116541	0.701357	0.199871	0.664685
1	KNN	0.119112	0.800905	0.207381	0.634423
2	SVM	0.118664	0.932127	0.210526	0.582545
3	RandomForest	0.120984	0.800905	0.210214	0.640638
4	XGBoost	0.117188	0.746606	0.202578	0.649014

Table 1: Classification metrics (Accuracy, Precision, Recall, F1-score) for single-model approaches (Decision Tree, kNN, SVM, Random Forest, XGBoost)

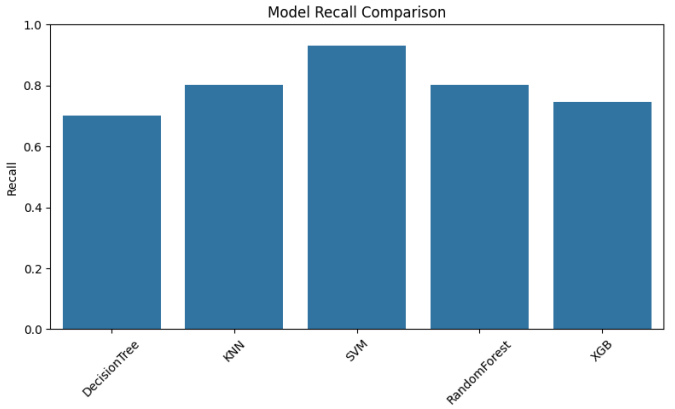

Figure 1: Bar plot comparing recall scores for each single-model algorithm

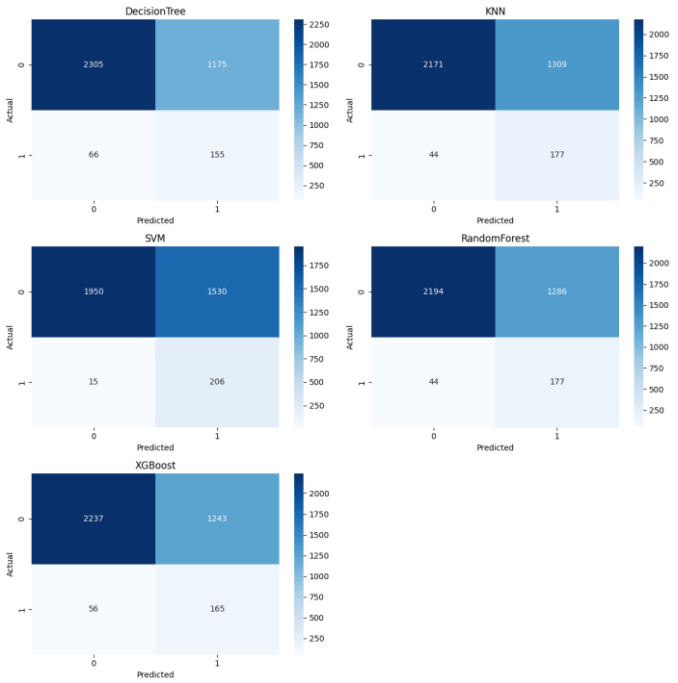

Figure 2: Confusion matrices for each single-model algorithm

The results indicated that SVM, Random Forest, and XGBoost achieved the highest recall for the fraud class, which is critical for minimizing undetected fraudulent claims. Based on these findings, these three models were selected for further optimization.

Next, hyperparameter tuning was performed for SVM, Random Forest, and XGBoost using randomized search with

cross-validation. The best configurations for each model were identified by maximizing recall, precision, F1-score, and accuracy. The optimized models were then re-evaluated on the validation set, and their updated performance is shown in **Table 2** and **Figure 3** (recall bar plot for tuned models).

	Model	Precision	Recall	F1-score	Accuracy
0	SVM (Tuned)	0.080627	0.954751	0.148696	0.347203
1	RandomForest (Tuned)	0.122425	0.941176	0.216667	0.593623
2	XGBoost (Tuned)	0.121474	0.954751	0.215526	0.584977

Table 2: Classification metrics for tuned SVM, Random Forest, and XGBoost

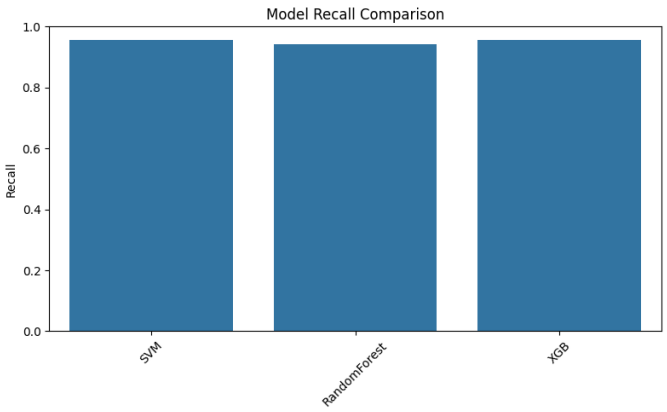

Figure 3: Bar plot of recall scores for tuned models

This process ensured that each model was not very accurate but were really robust in identifying the minority (fraud) class.

Building on the optimized models, a stacking ensemble was constructed using SVM, Random Forest, and XGBoost as base learners, with XGBoost as the meta-learner. The stacking model was trained on the balanced dataset and evaluated on the validation set. Its performance is presented in **Table 3** and visualized in **Figure 4** (confusion matrix for the stacking model).

	Model	Precision	Recall	F1-score	Accuracy
0	Stacked	0.125937	0.760181	0.216077	0.67063

Table 3: Classification metrics for the stacking ensemble

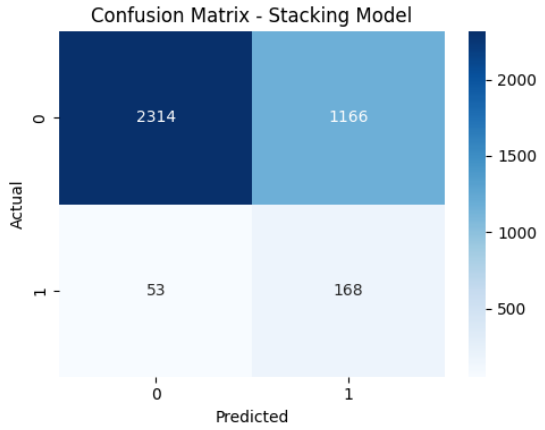

Figure 4: Confusion matrix for the stacking model

The stacking ensemble scored better accuracy than all individual models, achieving acceptable recall and F1-score for the fraud class. This improvement is attributed to the ensemble's ability to leverage the strengths of each base learner.

The final evaluation compared the best-tuned individual models and the stacking ensemble on the unseen test set. **Table 4** summarizes their performance, while **Figure 5** shows ROC curves for each model.

	Model	Precision	Recall	F1-score	Accuracy
0	XGBoost	0.127782	0.962162	0.225602	0.603761
1	SVM	0.082560	0.962162	0.152072	0.356355
2	RandomForest	0.130627	0.956757	0.229870	0.615435
3	Stacked	0.131193	0.772973	0.224314	0.679313

Table 4: Test set metrics for SVM, Random Forest, XGBoost, and Stacking

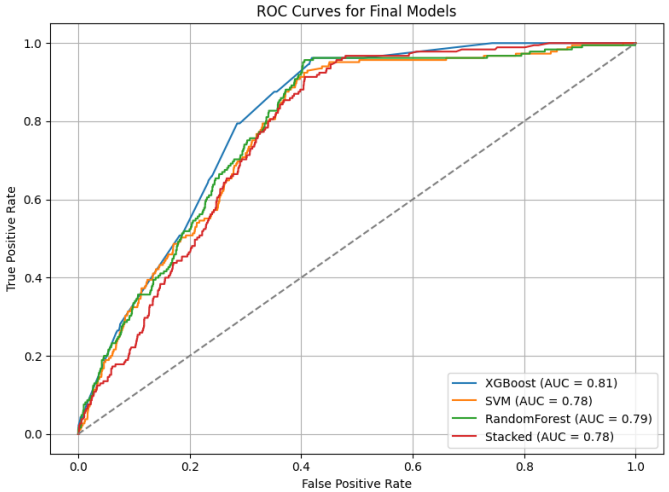

Figure 5: ROC curves for all final models on the test set

While the stacking model demonstrated strong generalization and maintained high AUC, **XGBoost achieved the highest recall among all models**, making it particularly effective at identifying fraudulent claims. Therefore, XGBoost is recommended as the primary model for deployment in real-world fraud detection scenarios where maximizing fraud detection is critical.

VII. Conclusion

In this study, we developed and compared several machine learning approaches for detecting fraudulent vehicle insurance claims using structured data from claim records, policyholders, and vehicles. Both single-model classifiers (Decision Tree, K-Nearest Neighbors, SVM, Random Forest, and XGBoost) and ensemble learning methods were evaluated.

Among all the models tested, XGBoost consistently achieved the highest recall, making it the most effective model for identifying fraudulent claims—a critical requirement in real-world fraud detection, where minimizing false negatives is

essential. While ensemble approaches such as stacking showed competitive overall performance, XGBoost stood out for its ability to correctly identify the majority of fraudulent cases, thus reducing the risk of undetected fraud.

Our analysis also emphasized the importance of careful feature selection, class balancing, and hyperparameter optimization in building robust fraud detection systems. Although the models performed well on the available dataset, future work should focus on validating these approaches on larger and more diverse datasets, and exploring additional feature engineering techniques to further enhance detection accuracy.

Overall, the findings of this study support the use of advanced machine learning models—particularly XGBoost—for effective and reliable insurance fraud detection..

REFERENCES

[1] Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. *Artificial Intelligence Review*, 34(1), 1-14.

[2] Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, 50(3), 559-569.

[3] Carcillo, F., Dal Pozzolo, A., Le Borgne, Y. A., Caelen, O., Mazzer, Y., & Bontempi, G. (2021). Machine learning for imbalanced datasets: Applications in fraud detection. *IEEE Transactions on Neural Networks and Learning Systems*, 32(8), 3461-3477.

[4] 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).

[5] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.

[6] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273-297.

[7] Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, 5(2), 241-259.

[8] West, J., Bhattacharya, M., & Islam, R. (2014). Intelligent financial fraud detection practices: An investigation. *Information Management & Computer Security*, 22(5), 450-464.

[9] Scikit-learn: Machine Learning in Python. https://scikit-learn.org/.

[10] Kaggle: Vehicle Insurance Claim Dataset. https://www.kaggle.com/datasets/