A Novel Hybrid Method for Imbalanced Automobile Insurance Fraud Detection

Phannana Aiemsuwan
Faculty of Informatics
Burapha University
Chonburi, Thailand
66810024@go.buu.ac.th

Supawadee Srikamdee
Faculty of Informatics
Burapha University
Chonburi, Thailand
srikamdee@go.buu.ac.th

2024 16th International Conference on Knowledge and Smart Technology (KST) | 979-8-3503-7073-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/KST61284.2024.10499643

***Abstract*—This paper addresses fraud detection in the automobile insurance sector, which experiences the highest rate of fraudulent claims in the industry. The study proposes a methodology that combines resampling techniques and backward elimination for feature selection to tackle data imbalance. This approach significantly improves the accuracy of machine learning models in identifying fraudulent activities. The research conducts a comprehensive comparison of seven major machine learning algorithms: Logistic Regression, Decision Tree, Random Forest, k-nearest Neighbors, Naive Bayes, XGBoost, and Support Vector Machine. These models are evaluated based on precision, recall, and F1 score across three different datasets to determine their effectiveness in fraud detection. The findings reveal that the Random Forest algorithm is the most efficient, consistently achieving F1 scores above 0.92. This highlights the vital role of machine learning in detecting and preventing fraud in the automotive insurance industry.**

Keywords—fraud detection, resampling method, backward elimination

I. INTRODUCTION

Insurance is closely related to people's lives and covers various aspects. The insurance industry is, therefore, a significant financial institution and is susceptible to fraud. Especially, automobile insurance fraud has become a significant problem because it can lead to financial losses for insurers and higher premiums for customers, including insurance companies that incur substantial financial losses from overpaying excessive or fraudulent compensation [1]. Typically, insurance companies review a range of documents associated with claims to identify potential fraud. These documents often detail the incident's location, time, and the extent of the damage, aiding in the detection of fraudulent activities such as repeated or exaggerated claims. Additionally, repair invoices are scrutinized to reveal details and expenses related to repairs, which can highlight abnormalities, for instance, unnecessary repairs or discrepancies between claimed damages and actual repair work. Despite these measures, detecting complex fraud through document examination alone remains a challenge. This process not only demands a considerable workforce but also involves significant time and financial resources.

In addressing the challenges of detecting automobile insurance fraud through traditional document examination, the application of machine learning techniques to identify patterns or behaviors indicative of fraudulent activity has proven valuable ([2]-[4]). Ambrose Muchangi Njeru [2] found that both AdaBoost and XGBoost algorithms demonstrated the highest classification accuracy, reaching 84.5%, in identifying fraudulent insurance claims. In a related study, Ahmet Yucel [3] utilized five different models, with the Artificial Neural Network emerging as the most accurate, achieving a 76.56% accuracy rate in detecting automobile insurance fraud. Additionally, Yucel's study incorporated Singular Value Decomposition (SVD) to enhance feature representation, improving the model's accuracy. Expanding upon this research, Anisha Singhal and colleagues [4] conducted a comparative analysis of seven different classification algorithms specifically for car insurance fraud detection. Their results showed that the XGBoost algorithm surpassed other models, achieving an F1 score and an AUC (Area Under the Curve) score of 0.906. These findings underscore the significant potential of machine learning in effectively combating insurance fraud.

The comprehensive review of existing literature reveals that machine learning techniques offer a promising avenue for the rapid and efficient detection of fraud. However, a crucial challenge is the issue of imbalanced data, which is prevalent in datasets used for detecting automobile insurance fraud [5]. This imbalance is characterized by a disproportionate representation of various classes, with normal data often overshadowing fraudulent data. Such skewness results in machine learning models being predominantly trained on normal data. Consequently, their proficiency in accurately identifying fraudulent data is hindered. Therefore, addressing this imbalance in data is crucial for enhancing the efficacy and reliability of machine learning in fraud detection.

Resampling approaches are essential techniques aimed at modifying the quantity of data in datasets to establish a balance across different classes. These techniques enable machine learning models to more effectively learn from and interpret patterns in both the majority and minority classes. The use of resampling techniques has shown significant promise in enhancing fraud detection capabilities. For example, Fuad A. Ghaleb and colleagues [6] utilized the Enhanced Synthetic Minority Over-Sampling Technique (ESMOTE) in detecting credit card payment fraud. Their method, which integrated Generative Adversarial Networks and the Random Forest Algorithm, achieved a notable average accuracy of 92.31%. In a similar vein, Matin N. Ashtiani and Bijan Raahemi [7] conducted a comparative analysis of different resampling methods for financial transaction fraud detection using the Multi-layer Feedforward Neural Network (MLFF). Their results indicated that SMOTE was particularly effective, achieving an average accuracy of 86.3%. In the context of insurance vehicle claims, Dhruvang Gondalia et al. [8] introduced the Adaptive Synthetic Sampling Approach (ADASYN), using the random forest algorithm to achieve a high accuracy of up to 97.1%. Moreover, G. Krishna Moorthy and K. Krishnaraja [9] conducted a comprehensive

979-8-3503-7073-7/24/$31.00 ©2024 IEEE

Authorized licensed use limited to: American University of Sharjah. Downloaded on November 20,2024 at 12:39:20 UTC from IEEE Xplore. Restrictions apply.

comparison of various resampling methods, finding that the SMOTE and SMOTE-Tomek methods were particularly effective, yielding an impressive accuracy rate of 93%.

Although resampling approaches provide a viable solution to the challenge of imbalanced data in machine learning, they also introduce a potential risk of overfitting. This risk emerges when machine learning models excessively learn the patterns of newly generated data, especially in cases where fraudulent data is created through oversampling techniques. To mitigate this, one effective strategy is the use of backward elimination for model variable selection. This method not only aids in reducing the likelihood of overfitting by favoring smaller models, which are generally less prone to overfitting than larger ones, but also simplifies the model by decreasing the number of parameters to be learned. Demonstrating the effectiveness of these strategies, Junzhang Wang and colleagues [10] applied a combination of feature selection techniques, including both forward selection and backward elimination, alongside a random forest algorithm for detecting anomalies in credit card usage. Their study achieved a noteworthy F1 score of 87.78%. Similarly, H. Onur Özcan and colleagues [11] developed the 'SOBE' platform for insurance fraud detection, incorporating backward elimination as a key component for feature selection.

In the field of fraud detection research, numerous studies have successfully integrated resampling methods with feature selection techniques to improve the effectiveness of their models. Moin Uddin and colleagues [12] implemented a comprehensive approach, utilizing Principal Component Analysis (PCA) for feature selection, the Synthetic Minority Over-sampling Technique (SMOTE) for oversampling, and the Random Forest algorithm for prediction. Their method achieved an impressive Area Under the Curve (AUC) score of 85.71, marking a significant 30% improvement compared to the standard performance of the random forest algorithm. In another notable study, Farhad Alam and Shariq Ahmad [13] developed a specialized framework for detecting financial fraud. They employed the PIKMCRos technique for resampling and the Halton Garra Rufa optimization algorithm for feature selection, centering their framework around the OC-LSTM model, which attained an F1 score of 91.45%. Additionally, Dan Zhao and colleagues [14] investigated a two-stage resampling method, initially using undersampling and ranking features by importance. Following this, they applied SMOTE and removed less significant features, achieving a peak fraud detection rate of 80%. Zainab Saad Rubaidi and colleagues [15] experimented with data oversampling using five distinct techniques. The resulting datasets were merged into a single balanced dataset, and feature selection was conducted using Pearson correlation alongside the Random Forest model. This approach reached a maximum F1 score of 97.5%, illustrating the effectiveness of these combined methods in enhancing fraud detection capabilities.

This comprehensive literature review underscores the efficacy of machine learning in fraud detection, specifically emphasizing the importance of resampling methods and feature selection in tackling issues related to data imbalance. It is imperative to evaluate machine learning models across multiple datasets to ascertain their generalizability. Furthermore, there is room for improvement in the accuracy of fraud detection across various studies, presenting opportunities for researchers to develop more effective methods. Consequently, this paper introduces a novel hybrid framework designed for the detection of automotive insurance fraud.

This framework effectively combines oversampling, feature selection, and machine learning algorithms to enhance predictive accuracy in binary classification tasks. It integrates random oversampling to address class imbalance and employs the backward selection algorithm for efficient feature selection. Additionally, a machine learning classifier is utilized for prediction. To assess the framework's effectiveness, it was applied to three distinct automobile insurance fraud datasets and then benchmarked against a conventional approach that does not incorporate feature selection or oversampling techniques. Our comparative analysis involved seven established machine learning models to determine the most generalizable and effective one. The hybrid framework offers considerable advantages for insurance managers and practitioners, notably in improving prediction accuracy and reducing financial risks in the insurance sector. Moreover, this method has potential applicability in fraud detection across various domains.

II. Related Works

In this section, we present techniques used to address the issue of imbalanced data, including resampling methods and algorithms for selecting features suitable for machine learning models. The details are as follows:

A. Resampling

Resampling is a method that entails generating a new dataset from an existing one by replicating or eliminating samples. It can be employed to tackle various issues, including imbalanced data. Imbalanced data refers to a dataset in which one or more classes are noticeably underrepresented in comparison to other classes. This situation can result in machine learning models showing bias toward the majority class, resulting in inferior performance on the minority class.

Resampling can be employed to mitigate imbalanced data by either oversampling the minority class or undersampling the majority class [16]. Various popular resampling techniques include:

- Random Oversampling involves randomly adding data to the minority class to match the number of samples in the majority class. New samples are either selected from the existing data in the minority class (existing data) or generated as new samples with different characteristics (synthetic samples).
- Synthetic Minority Over-sampling Technique (SMOTE) works by selecting data samples from the minority class and generating new samples by randomly choosing neighbors of the selected samples. Neighbors can be selected either randomly or with a specified distance.
- Random Undersampling involves removing data samples from the majority class to match the number of samples in each class. The reduction in sample numbers is achieved by randomly selecting and eliminating data samples from the majority class until reaching the desired number of samples in each class. Although this technique is fast and easy to implement,

Authorized licensed use limited to: American University of Sharjah. Downloaded on November 20,2024 at 12:39:20 UTC from IEEE Xplore. Restrictions apply.

it may result in data loss and inefficient utilization of the remaining data.

- Adaptive Synthetic Sampling (ADASYN) is a technique for generating synthetic data points without considering every instance within the minority group. It employs a weighted distribution based on the importance of sample data within the minority group. Synthetic data is generated depending on the difficulty of classifying the data; if a data point is challenging to classify, it receives a higher weight, and synthetic data is generated around that specific area.

It is often a good idea to experiment with different resampling methods to determine the most effective one for our dataset and machine learning model. Therefore, in this research, we conducted experiments comparing various resampling methods and found that random oversampling yielded the best results. Consequently, we adopted this technique in our fraud detection framework.

B. Backward elimination

Backward elimination, a technique utilized in regression analysis, serves to select a subset of explanatory variables for a model. This approach methodically removes less significant features, retaining only those that are crucial.

The implementation of backward elimination can vary, but a prevalent approach involves employing a p-value threshold. P-values are used to assess the significance of each feature. The process begins by calculating the p-values for each feature, which are then compared against the established threshold. Features whose p-values exceed this threshold are subsequently removed. This procedure is iteratively repeated until the remaining variables in the model all possess p-values below the set threshold, or until the model is refined to the desired number of features.

This technique is valuable as it helps reduce the risk of overfitting the data and enhances the performance of the machine learning model, particularly for the minority class [17]. Machinya Tongesai et al. [18] utilized Pearson correlation and chi-squared elimination to filter out significant features for Insurance Fraud Detection, revealing that XGBoost achieved the highest accuracy at 79%. Qazaleh Sadat Mirhashemi and colleagues [19] employed the K best method to select optimal features for credit card fraud detection models. The results revealed that Random Forest achieved the highest accuracy at 76%. Chandana Gouri Tekkali and Karthika Natarajan [20] utilized rough set theory for feature selection in detecting fraud in digital financial transactions, achieving an impressive accuracy of 94.96%, surpassing the accuracy obtained using Pearson correlation and chi-square methods.

III. METHODOLOGY

In this section, we introduce a framework for predicting fraud in the automotive insurance domain. The process begins with data preprocessing, followed by balancing the data using resampling techniques, conducting feature selection through backward elimination, and ultimately culminating in fraud classification using machine learning models. An overview of the methods is visualized in Figure 1. The details of each sub-process are as follows:

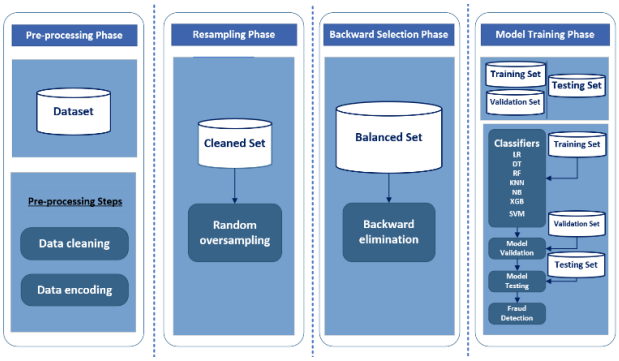

Fig. 1. The research methodology

A. Pre-processing Phase

This step involves data cleaning, which includes removing duplicate rows and handling missing values. In this research, missing values are addressed as follows: numerical data is filled with the mean, while nominal data is filled with the mode. Additionally, since some features contain categorical data, data encoding is required to convert text into numerical values before proceeding to the feature selection step. The data was imported and the data-cleaning process commenced. This involved the removal of features that were irrelevant to the label, a thorough examination of duplicate entries, and the handling of missing data. Subsequently, transformations were applied to deal with categorical data.

B. Resampling Phase

Random oversampling is a technique used to address the issue of imbalanced data. In the proposed framework, oversampling is performed within the minority class using a simple random sampling technique. Each member of the minority class has an equal chance of being randomly selected for duplication. This increases the number of instances in the minority class, resulting in a more balanced distribution compared to the larger majority class.

This research employs the 'RandomOverSampler' function from the 'imbalanced-learn' (imblearn) Python library to randomly increase the amount of data in smaller classes, aiming to achieve a more balanced distribution between both classes.

After completing the resampling phase, data from both classes will have an equal number, addressing the issue of data imbalance. Next, we proceed to the step of selecting important features to reduce the complexity of model learning and alleviate potential overfitting issues.

C. Backward Elimination Phase

If your model has several features, not all features may be equally important. To be sure that the framework has the optimal number of features, we explore various dimensionality reduction techniques including the chi-squared test, principal component analysis (PCA), and wrapper-based size reduction methods. Their process aimed at reducing the number of features or variables in a dataset while preserving its important information. For the chi-squared test, features that are deemed statistically

Authorized licensed use limited to: American University of Sharjah. Downloaded on November 20,2024 at 12:39:20 UTC from IEEE Xplore. Restrictions apply.

independent of the target variable (class labels) may be considered less informative and can be potentially removed. PCA is a linear dimensionality reduction technique that transforms the original features into a new set of uncorrelated variables called principal components. By selecting a subset of the top principal components, one can represent the data in a lower-dimensional space while retaining a significant portion of its original variability. Wrapper methods evaluate subsets of features by training and testing a model using a specific machine learning algorithm. Examples of wrapper methods include forward selection and backward elimination.

The results of comparing dimensionality reduction techniques on the selected dataset for this research reveal that backward elimination achieves the highest accuracy in predicting fraud. Therefore, this research adopts backward elimination as one of the components of the proposed hybrid method.

Our backward elimination method gradually eliminates unimportant features until only the essential ones remain. It begins by fitting a logistic regression model with all the independent variables. Then, it uses the logistic regression model to calculate the P-value for each independent variable. The independent variable with the highest P-value is eliminated from the model. Subsequently, a new logistic regression model is created with the remaining independent variables. This process is repeated until no more independent variables are eliminated.

The sklearn library in Python offers a convenient function specifically designed for conducting backward elimination in the context of logistic regression models. The function used is SequentialFeatureSelector, with various parameters set as depicted in Figure 2.

From Figure 2, the 'k_features' parameter represents the number of features to be selected. In this research, experiments were conducted by adjusting the number of features to four values: 5, 10, 20, and 30. It was observed that the number of features yielding the highest accuracy on the dataset under study was 20. Consequently, this parameter was chosen for use across all datasets in the model performance testing results presented in Section 4. The 'forward' parameter is set to false for backward selection, the 'scoring' parameter represents the accuracy evaluation criterion selected as the f1-score, and the 'cv' parameter denotes the number of folds in cross-validation, set to 5.

D. Model Training Phase

Once the data is prepared, we move on to the process of training the models to classify fraudulent and non-fraudulent data. In this research, seven well-known machine learning models have been collected, including Logistic Regression, Decision Tree, Random Forest, K Nearest Neighbors, Naive Bayes, XGBoost, and Support Vector Machine. We use scikit-learn in Python with default parameters to demonstrate the advantages of employing resampling and backward selection over the conventional approach. Each dataset will be evaluated using 5-fold cross-validation to study and compare the performance of each machine learning model.

```
sfs = SequentialFeatureSelector(LogisticRegression(),
                                k_features= 20,
                                forward=False,
                                scoring='f1',
                                cv=5)
```

Fig. 2. Using the SequentialFeatureSelector function to select important features

TABLE I. DATA CHARACTERISTICS

No.	Dataset Name	# Examples	#Feature	Fraud ratio
1	Auto Insurance Fraud	1000	40	25%
2	Fraudulent Claim on Cars Physical Damage	17998	25	16%
3	Car Insurance Fraud	15420	33	6%

IV. RESULTS AND DISCUSSTION

In this section, we present an experimental design to test the effectiveness of the proposed hybrid framework. These details encompass information about the datasets used to test the framework, the performance indicators utilized to evaluate fraud detection accuracy, benchmarking results of seven machine learning algorithms, and a performance comparison between the framework presented in this research and other studies on the same dataset. The specifics are as follows:

A. Datasets

To test the generalizability of the proposed framework, we collected three datasets related to automobile insurance from the Kaggle website: Auto Insurance Fraud [21], Fraudulent Claim on Cars Physical Damage [22], and Car Insurance Fraud [23]. Details of the data are provided in Table 1.

From Table 1, the second column displays the name of each dataset. The third column indicates the number of examples in each dataset. The fourth column represents the number of features in each dataset. Lastly, the fifth column shows the ratio of the amount of data in the fraud class to the total amount of data, indicating the degree of imbalance. Overall, the first dataset has a relatively large number of features. The second dataset had the largest amount of data, while the third dataset exhibited the highest level of imbalance. This demonstrates the diversity of the datasets chosen to test the framework.

B. Performance Metrics

In binary classification tasks, precision, recall, and the F1-score are crucial metrics that evaluate a model's prediction accuracy. These metrics are derived from the confusion matrix, a tool that encapsulates a classification algorithm's performance.

Precision quantifies the accuracy of a model's positive predictions. It is defined as the proportion of true positive observations to the total number of positive predictions made by the model. The calculation of precision is detailed in Equation 1.

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

Authorized licensed use limited to: American University of Sharjah. Downloaded on November 20,2024 at 12:39:20 UTC from IEEE Xplore. Restrictions apply.

Recall is a metric that evaluates the model's capability to accurately identify instances of the positive class among all true positive instances. The method for computing recall is presented in Equation 2.

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

The F1-score represents the harmonic mean of precision and recall, integrating both metrics into a single measure. It can be calculated as shown in Equation 3.

$$F1 - score = \frac{2 \times precision \times recall}{precision + recall} \quad (3)$$

C. Comparison of various classifiers

This section compares the performance of seven machine learning models: Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbors (KNN), Naive Bayes (NB), XGBoost (XGB), and Support Vector Machine (SVM). The results of evaluating fraud classification accuracy on the tested dataset are presented in Table 2.

From Table 2, the first column displays the machine learning model names, the second column shows the performance measures, the third and fourth columns display the results for dataset1, while the fifth and sixth columns show the results for dataset2, and the seventh and eighth columns present the results for dataset3. The test results for each dataset are presented in two cases: 'None' displays the results without utilizing Resampling and Backward Elimination, while 'Hybrid' denotes the use of both these techniques.

The results show that combining random oversampling and backward elimination with machine learning significantly improves the performance of more accurate fraud identification across all datasets and models. For logistic regression and SVM models, using only machine learning, the models cannot detect fraud at all. The evidence is that the F1-score in the 'None' case is zero in every dataset. However, in the 'Hybrid' case, the F1-score increased by at least 31% in dataset1 and up to 75% in dataset3. Overall, out of the seven models, random forest is the best model with the highest F1-score across all datasets. Specifically, dataset3 was able to detect fraud with 100% F1-score, while dataset1 achieved an F1-score of 0.92, which is the same as XGBoost, and dataset2 achieved an F1-score of 0.98.

In the experimental results, it is noteworthy that the LR and SVM methods show an unexpected outcome with an F1-Score of 0 across all datasets. This occurrence may be attributed to the fact that LR and SVM aim to identify decision boundaries that effectively separate classes. However, when faced with imbalanced data, the decision boundary might be skewed towards the majority class, resulting in poor identification of the minority class (fraud). SVM, in particular, may place the hyperplane in a way that doesn't effectively capture the minority class. While the remaining methods yielded more satisfactory results, this may be due to their enhanced ability to capture non-linear relationships and complex patterns in the data. Fraud detection problems often involve intricate patterns, and models that can handle non-linearities may perform better. To improve the performance of LR and SVM on imbalanced datasets, techniques like resampling, or the use of ensemble methods can be considered.

TABLE II. COMPARISON OF FRAUD CLASSIFICATION PERFORMANCE WITH MACHINE LEARNING MODELS

		Dataset1		Dataset2		Dataset3	
ML	**Measure**	***None***	***Hybrid***	***None***	***Hybrid***	***None***	***Hybrid***
LR	**precision**	0.00	0.57	0.00	0.55	0.00	0.70
	recall	0.00	0.62	0.00	0.58	0.00	0.80
	F1-score	0.00	0.59	0.00	0.56	0.00	0.75
DT	**precision**	0.56	0.81	0.22	0.85	0.25	0.94
	recall	0.58	0.95	0.25	0.99	0.28	1.00
	F1-score	0.56	0.87	0.23	0.91	0.27	0.97
RF	**precision**	0.57	0.88	0.60	0.98	0.75	0.99
	recall	0.33	0.96	0.01	0.99	0.02	1.00
	F1-score	0.41	**0.92**	0.02	**0.98**	0.03	**1.00**
KNN	**precision**	0.30	0.63	0.16	0.69	0.11	0.84
	recall	0.12	0.77	0.03	0.90	0.01	1.00
	F1-score	0.17	0.70	0.05	0.78	0.01	0.91
NB	**precision**	0.61	0.82	0.31	0.61	0.25	0.71
	recall	0.55	0.68	0.00	0.60	0.03	0.65
	F1-score	0.58	0.75	0.00	0.60	0.05	0.68
XGB	**precision**	0.65	0.87	0.37	0.83	0.71	0.90
	recall	0.62	0.95	0.10	0.93	0.36	1.00
	F1-score	0.63	**0.92**	0.15	0.88	0.47	0.95
SVM	**precision**	0.00	0.54	0.00	0.53	0.00	0.64
	recall	0.00	0.23	0.00	0.50	0.00	0.28
	F1-score	0.00	0.31	0.00	0.52	0.00	0.39

TABLE III. COMPARISON OF FRAUD CLASSIFICATION PERFORMANCE WITH OTHER MODELS

Dataset	**Measure**	***Our Framework***	***Comparative Research***
Auto Insurance Fraud	**precision**	0.88	0.91
	recall	0.96	0.88
	F1-score	**0.92**	0.89
Fraudulent Claim on Cars Physical Damage	**precision**	0.98	0.74
	recall	0.99	0.79
	F1-score	**0.98**	0.76
Car Insurance Fraud	**precision**	0.99	0.99
	recall	1.00	0.94
	F1-score	**1.00**	0.97

D. Comparison with other methods on the same dataset

This section presents a comparison of the performance of the proposed framework with other studies tested on the same dataset. As the three datasets were obtained from distinct research studies, the dataset1 (auto insurance fraud) is compared to the model presented in article [4], the dataset2 (fraudulent claim on cars physical damage) is compared to the model introduced in article [3], and the dataset3 (car insurance fraud) is compared to the model presented in article [8]. The results of the random forest algorithm will serve as the representative basis for comparison, as shown in Table 3.

Table 3 illustrates the data used in the study in its first column. The second column details the performance measures. The third column displays the results of fraud detection performance measurements obtained using our framework. Finally, the fourth column presents comparative results from other studies.

The results presented in Table 3 indicate that the framework introduced in this study outperforms other models in fraud detection accuracy across all datasets, with a minimum enhancement of 3% in the F1-Score. These results were obtained using the same machine learning model—random forests—but with varied data preparation techniques. This underscores the efficacy of employing random oversampling and backward elimination before training a

Authorized licensed use limited to: American University of Sharjah. Downloaded on November 20,2024 at 12:39:20 UTC from IEEE Xplore. Restrictions apply.

machine learning model. Such an approach not only improves fraud detection accuracy, especially in imbalanced datasets, but also ensures that valuable information is retained through the use of oversampling.

V. Conclusion

This study aims to enhance the effectiveness of fraud detection by focusing on the data preparation process, with automobile insurance as a case study. We achieve this by integrating random oversampling and backward elimination techniques to address the issue of imbalanced data, where fraud cases are typically fewer than normal cases. Experiments on three selected automobile insurance datasets were conducted to assess the generalizability of the proposed framework. A comparison of fraud classification outcomes across seven machine learning models revealed that the Random Forest model achieved the highest F1-score. Consequently, the outcomes from this Random Forest model were juxtaposed with findings from other studies employing different data preparation methods. The results indicate that our hybrid framework substantially improved fraud detection. The preprocessing techniques introduced in this research enable machine learning models to more accurately predict fraud with higher quality and better-balanced data. This approach is applicable in various real-world scenarios involving asymmetrical data, such as in banking, insurance, or other systems' fraud detection. Future research will explore alternative oversampling and feature selection methods and apply these techniques to real data from diverse organizations to evaluate their effectiveness in practical settings and refine them accordingly.

References

[1] Clarke, M. (1990). The control of insurance fraud: A comparative view. *The British Journal of Criminology, 30*(1), 1-23.

[2] Njeru, A. M. (2022). *Detection of Fraudulent Vehicle Insurance Claims Using Machine Learning* (Doctoral dissertation, University of Nairobi).

[3] Yucel, A., (2022). A Novel Data Processing Approach to Detect Fraudulent Insurance Claims for Physical Damage to Cars. *Journal of Results in Science, 11*(2), 120-131.

[4] Singhal, A., Singhal, N., & Sharma, K. (2023, June). Machine Learning Methods for Detecting Car Insurance Fraud: Comparative Analysis. In *2023 3rd International Conference on Intelligent Technologies (CONIT)* (pp. 1-5). IEEE.

[5] Baran, S., & Rola, P. (2022). Prediction of motor insurance claims occurrence as an imbalanced machine learning problem. *arXiv preprint arXiv:2204.06109*.

[6] Ghaleb, F. A., Saeed, F., Al-Sarem, M., Qasem, S. N., & Al-Hadhrami, T. (2023). Ensemble Synthesized Minority Oversampling based Generative Adversarial Networks and Random Forest Algorithm for Credit Card Fraud Detection. *IEEE Access,11*(2023), 89694-89710.

[7] Ashtiani, M. N., & Raahemi, B. (2023, July). An Efficient Resampling Technique for Financial Statements Fraud Detection: A Comparative Study. In 2023 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME) (pp. 1-7). IEEE.

[8] Gondalia, D., Gurav, O., Joshi, A., Joshi, A., & Selvan, S. (2022). Automobile Insurance Claim Fraud Detection using Random Forest and ADASYN. *International Research Journal of Engineering and Technology (IRJET), 9*(5), 104-107.

[9] Moorthy, G. K., & Krishnaraja, K. (2023). Addressing Class Imbalance Through Resampling and Achieving Explainable Machine Learning Using Permutation Important: A Case Study on Fraud Detection in Insurance Vehicle Claims. *The Indian Journal of Technical Education, 46*(Special), 260-270.

[10] Wang, J., de Moraes, R. M., & Bari, A. (2020, August). A predictive analytics framework to anomaly detection. In *2020 IEEE Sixth International Conference on Big Data Computing Service and Applications (BigDataService)* (pp. 104-108). IEEE.

[11] ÖZCAN, H. O., Çolak, İ., Erimhan, S., Güneş, V., Fatih, A. B. U. T., & Fatih, A. K. A. Y. (2022). SOBE: A Fraud Detection Platform in Insurance Industry. *Kocaeli Journal of Science and Engineering, 5*(ICOLES2021 Special Issue), 25-31.

[12] Uddin, M., Ansari, M. F., Adil, M., Chakrabortty, R. K., & Ryan, M. J. (2023). Modeling Vehicle Insurance Adoption by Automobile Owners: A Hybrid Random Forest Classifier Approach. *Processes*, *11*(2), 629.

[13] Alam, F., & Ahmad, S. (2023). Intelligent Fraud Detection Framework for PFMS Using HGRO Feature Selection and OC-LSTM Fraud Detection Technique. *SN Computer Science, 4*(4), 400.

[14] Zhao, D., Shen, Z., & Zhao, S. (2023). Feature Selection Based on Two-stage Resampling Technique for Imbalanced Dataset. *Procedia Computer Science, 221,* 316-321.

[15] Rubaidi, Z. S., Ammar, B. B., & Aouicha, M. B. (2023). Vehicle Insurance Fraud Detection Based on Hybrid Approach for Data Augmentation. *Journal of Information Assurance and Security, 18*, 135-146.

[16] Eltayeb, R., Karrar, A. E., Osman, W. I., & Mutasim, M. (2023). Handling Imbalanced Data through Re-sampling: Systematic Review. *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, *11*(2), 503-514.

[17] Khaire, U. M., & Dhanalakshmi, R. (2022). Stability of feature selection algorithm: A review. *Journal of King Saud University-Computer and Information Sciences*, *34*(4), 1060-1073.

[18] Tongesai, M., Mbizo, G., & Zvarevashe, K. (2022, November). Insurance Fraud Detection using Machine Learning. In *2022 1st Zimbabwe Conference of Information and Communication Technologies (ZCICT)* (pp. 1-6). IEEE.

[19] Mirhashemi, Q. S., Nasiri, N., & Keyvanpour, M. R. (2023, May). Evaluation of Supervised Machine Learning Algorithms for Credit Card Fraud Detection: A Comparison. In *2023 9th International Conference on Web Research (ICWR)* (pp. 247-252). IEEE.

[20] Tekkali, C. G., & Natarajan, K. (2023). RDQN: ensemble of deep neural network with reinforcement learning in classification based on rough set theory for digital transactional fraud detection. *Complex & Intelligent Systems*, 1-20.

[21] Shah, B., "Auto Insurance Claims Data", Kaggle, Available:https://www.kaggle.com/datasets/buntyshah/auto-insurance-claims-data , [Accessed 11-Aug-2023]

[22] SR, "Fraudulent Claim on Cars Physical Damage", Kaggle, Available:https://www.kaggle.com/datasets/surekharamireddy/fraudulent-claim-on-cars-physical-damage , [Accessed 25-Sept-2023]

[23] Gupta, S, "Car Insurance Fraud", Kaggle, Available:https://www.kaggle.com/code/jwilda3/classifying-fraud-by-decision-trees/data , [Accessed 25-Sept-2023

[24] Müller, A. C., & Guido, S. (2016). Introduction to machine learning with Python: a guide for data scientists. " O'Reilly Media, Inc.".

Authorized licensed use limited to: American University of Sharjah. Downloaded on November 20,2024 at 12:39:20 UTC from IEEE Xplore. Restrictions apply.