Machine Learning-Driven Detection of Fraudulent Vehicle Insurance Claims

Ambrose Njeru
Strathmore University, Nairobi
School of Computing & Engineering Sciences

Evans A.K. Miriti, PhD
University of Nairobi, Nairobi
Department of Computer Science & Informatics

ABSTRACT
Fraudulent insurance claims pose a significant financial burden on the vehicle insurance industry, leading to increased premiums for honest customers and substantial losses for insurers. Traditional manual methods for detecting fraudulent claims are inefficient, time-consuming, and prone to errors. This study addresses these challenges by applying and evaluating multiple machine learning techniques to accurately distinguish between genuine and fraudulent vehicle insurance claims. The research specifically aims to characterize the nature of fraudulent vehicle insurance claims, identify key features relevant for model training, assess the performance of various classifiers on both balanced and unbalanced datasets, and develop a web-based system that automates the classification process using the optimal model. Experimental results demonstrate that ensemble methods, particularly AdaBoost and Extreme Gradient Boosting (XGBoost), outperform other classifiers, achieving a classification accuracy of 84.5%. Logistic Regression shows the poorest performance, while Artificial Neural Networks (ANN) perform better with unbalanced data but degrade with balanced data. Additionally, model scalability remains limited to smaller datasets for all evaluated classifiers. The study's outcomes provide a practical machine learning-driven framework to enhance fraud detection accuracy and processing efficiency, supporting insurers in mitigating losses and improving risk management.

Keywords
Classification Algorithms, Fraudulent Claims, Machine Learning (ML), Vehicle Insurance.

1. INTRODUCTION
Insurance fraud remains one of the most pervasive challenges confronting the insurance industry globally, characterized by intentional acts such as inflating claims or misrepresenting facts to receive undue financial benefits. [1] defined insurance fraud as ""knowingly making a fraudulent claim, inflating a claim, adding extra items to a claim, or being in any other way dishonest with the objective of collecting more than genuine entitlement." This justification pertains to deceitful, intentional, or fraudulent concealment, ultimately leading to an illegitimate financial benefit for the fraudulent claimant or policyholder. Insurance companies incur substantial financial losses because of costly fraudulent claims. Consequently, it is essential to discern between authentic statements and false ones.

This issue is particularly acute in motor vehicle insurance, a sector plagued by disproportionate losses. In Kenya, motor vehicle insurance incurs severe technical losses amounting to 68.92% for private vehicles and 60.72% for commercial vehicles [2], thereby significantly eroding insurer revenues. These figures reflect not only the high payout rates but also the considerable investigative costs, which represent 44.16% of total claim expenses, further diminishing net premium income and highlighting the financial strain imposed by potential fraudulent claims.

The magnitude of the problem is underscored by data from [3], indicating that approximately 35% of insurance claims are fraudulent, with motor vehicle insurance claims contributing to the lion's share of sector losses. Fraudulent activities within this domain span a spectrum from fabricated accidents to falsified claim details and deliberate misrepresentation of incidents [4]. Conventional fraud detection methods, often manual and rule-based, fall short in effectively identifying these sophisticated fraudulent behaviors, allowing many cases to evade detection. Consequently, insurers face escalating costs that threaten their financial health and drive-up premiums, affecting both consumers and the broader regulatory environment.

Given these challenges, this study proposes leveraging machine learning to detect fraudulent vehicle insurance claims through systematic analysis of claim data features. By training and comparing multiple classifiers, including Naïve Bayes, XGBoost, Random Forest, Decision Tree, AdaBoost, Logistic Regression, Artificial Neural Networks, and Support Vector Machines, this research aims to identify the most effective model for fraud detection. Furthermore, a novel predictive system incorporating the best-performing classifier has been developed to automate the classification of claims as genuine or fraudulent. This approach seeks to enhance detection accuracy, improve operational efficiency, and ultimately contribute to reducing the substantial losses driven by fraudulent claims in motor vehicle insurance.

1.1 General Aim and Specific Objectives
The general aim of this study was to explore the use of features extracted from vehicle insurance claim datasets to enhance the detection of fraudulent claims through machine learning algorithms. Building on this investigation, a novel system was developed to automatically predict and classify vehicle insurance claims as either genuine or fraudulent.

The study's specific objectives were to:

i. Characterize fraudulent insurance claims within the vehicle insurance domain.
ii. Identify relevant features for training machine learning models to detect fraudulent vehicle insurance claims.
iii. Evaluate the performance of various machine learning models using both balanced and unbalanced datasets.
iv. Develop a classification system that categorizes vehicle insurance claims as genuine or fraudulent based on the best-performing machine learning model.

2. LITERATURE REVIEW
Vehicle insurance is defined as a contract where the insurer assumes the risk of loss resulting from property damage or injury caused by vehicle accidents [5]. Fraud within this sector manifests in diverse forms, including impersonation of

legitimate claimants, forgery of claim documents, misappropriation of funds, and internal manipulation by insurance employees to benefit undeserving parties, according to [6]. Accurate and timely verification of claims relies heavily on data collected at accident sites such as driver details, vehicle registration, and insurance coverage which are manually reviewed by insurance supervisors who use scoring systems and investigative follow-ups to determine claim validity [3]. While methodical, this manual process is labor-intensive and prone to errors, highlighting the need for automated, data-driven solutions.

Rule-based fraud detection systems, which use analyst-defined algorithms, offer simplicity but lack adaptability and can miss complex fraud patterns, necessitating frequent manual updates as investigated by [7]. Technological advances have broadened fraud detection capabilities, exemplified by biometric verification systems used in Ghana's health insurance sector that integrate social and technical measures to improve claim authenticity [8]. Moreover, data mining combined with expert knowledge has increasingly enabled automated fraud detection, extracting actionable insights from large datasets [9].

Machine learning (ML) has emerged as a promising approach for insurance fraud detection. Comparative studies demonstrate that algorithms such as Logistic Model Trees and Random Forests achieve high precision, recall, and F1 scores in classifying fraudulent claims, outperforming simpler models like Naïve Bayes [10]. Similarly, Extreme Gradient Boosting (XGBoost) has shown exceptional accuracy (up to 99.25%) in motor vehicle insurance fraud detection, surpassing traditional methods like Decision Trees and k-Nearest Neighbors, though at the cost of greater computational time [11]. However, these studies often lack transparency regarding feature selection, a critical aspect for model interpretability and generalization.

Challenges inherent in fraud detection are further compounded by dynamic fraud behaviors and highly imbalanced datasets, as demonstrated in credit card fraud detection research [12]. Hybrid sampling strategies and thorough feature selection markedly influence model performance, yet not all studies rigorously address these issues [13]. For instance, the AdaBoost algorithm displayed superior detection accuracy and F1 scores over Naïve Bayes, logistic regression, and artificial neural networks when applied to imbalanced transaction data [14]. Despite these advances, some logistic regression-based models in automotive insurance reported modest predictive accuracy (~59%), underscoring the limitations of traditional classifiers without extensive feature engineering [15].

Taken together, these findings highlight the critical need for comprehensive approaches that combine robust feature engineering, advanced machine learning algorithms, and scalable systems to effectively detect and reduce fraudulent vehicle insurance claims. This study builds on existing work by evaluating a broad range of classifiers with balanced and unbalanced data, aiming to develop an automated, reliable system for fraudulent claim identification that addresses both accuracy and operational efficiency gaps in current methodologies.

3. PROPOSED SYSTEM

Eight machine learning classifiers—Naïve Bayes (NB), Extreme Gradient Boosting (XGB), Random Forest (RF), Decision Tree (DT), AdaBoost (Ada), Logistic Regression (LR), Artificial Neural Networks (ANN), and Support Vector Machine (SVM)—were trained on features extracted from the selected dataset. The classification results from these models were analyzed and evaluated to identify the most accurate and best-performing algorithm for developing a system capable of detecting and classifying vehicle insurance claims as either legitimate or fraudulent.

The input variables were categorized into two groups to correspond with the target fraudulent status of the claims: insurance policy details and insurance claim details. Key data points included customer name, sex, age of the policyholder, vehicle category, vehicle make, vehicle age, sum insured, insurance cover specifics, policy start and end dates, claim logging date, incident date, incident location, and police report information. Figure 1 illustrates the proposed model architecture for handling both unbalanced and balanced datasets, while Figure 2 presents the design of the machine learning-powered web-based system aimed at detecting fraudulent vehicle insurance claims.

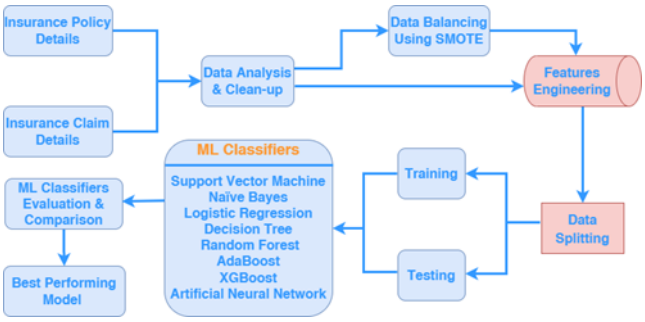

Figure 1: ML Model with Unbalanced and Balanced Datasets

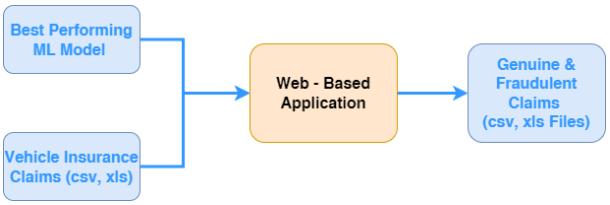

Figure 2: Machine Learning-Powered Web-Based System

4. METHODOLOGY

The claims content was initially analyzed to identify key features prior to developing the machine learning model. Vehicle insurance data was collected to gain a deeper understanding of its structure and to extract the essential characteristics needed for training the machine learning classifiers.

4.1 Dataset Understanding

The relevant dataset for this study was sourced, its quality evaluated, and key variables extracted to facilitate model development. Efforts were made to obtain data from several insurance companies in Kenya involved in motor vehicle policies, including Britam Insurance Kenya, Co-operative Insurance Company (CIC), and Jubilee Insurance. However, access to these datasets was not granted due to the sensitive nature and privacy concerns surrounding the information. Consequently, an alternative dataset was acquired from an online CSV file of vehicle insurance claims available on Kaggle (2018). This dataset comprised approximately 1,000 claims, with 247 (24.7%) labeled as fraudulent and 753 (75.3%) as genuine, providing a foundation for the study's analysis and modeling.

The dataset consisted of 1000 rows and 39 columns, where input variables were utilized to generate features for training the eight models applied in the study. Variables that did not meet the predetermined threshold were excluded. The characteristics of the data set are presented in Table 1.

Table 1. Dataset Features

Number of Claims	1000
Number of Attributes	39
Categorical Attributes	24
Genuine Claims	753
Fraudulent Claims	247
Fraudulent Claims Incidence Rate	24.7%
Number of Claims	1000
Number of Attributes	39

The extracted dataset was not entirely clean, as several input variables contained null values.

4.2 Data Preprocessing

The study employed a classical exploratory data analysis approach to identify and select quality features from real-world vehicle insurance claim datasets, which are often imperfect, inconsistent, and noisy. To improve data quality, data pre-processing techniques are applied to address missing values, reduce noise, and resolve inconsistencies. According to [16], data preparation involves cleaning, integrating, transforming, and reducing data to eliminate duplicates and irrelevant information, thereby retaining only valuable features for efficient and effective classification. This process follows the key stages below:

i. Data cleaning, which focuses on removing outliers and imputing missing values.
ii. Data transformation, such as normalization, can enhance the precision of distance-based mining algorithms.
iii. Data integration, combining information from multiple sources into a unified data warehouse.
iv. Data reduction through feature extraction and selection to eliminate redundant attributes and reduce data dimensionality.

Initially, the dataset was examined for duplicate records and missing values. Missing data were imputed using the fillna() method in Python. Exploratory data analysis began with the dependent variable, fraud reported, including the creation of heatmaps to visualize variables with at least a 0.3 Pearson's correlation coefficient alongside the dependent variable. This visualization helped highlight relevant relationships and guide further analysis.

A strong correlation of 0.92 was observed between the variable's month_as_customer and age, likely because vehicle insurance is typically obtained when an individual owns a car, and both values increase over time. Consequently, the age variable was removed to reduce redundancy. The total_claim variable was also removed due to its strong correlation with total_claim_amount, injury_claim, property_claim, and vehicle_claim variables. To avoid multicollinearity, several highly correlated variables were eliminated. However, the detailed claim variables were retained because they provided granular information not captured by total claims alone. Subsequent analysis confirmed no significant multicollinearity issues aside from the inherent correlation among claim variables.

Further feature assessment involved identifying unique values within remaining variables to determine their informational value. Features exhibiting excessive unique values, implying limited predictive value, were excluded. The removed features included policy_number, policy_bind_date, policy_state, insured_zip, incident_location, incident_date, incident_state, incident_city, insured_hobbies, auto_make, auto_model, and auto_year.

To enable machine learning models to process the data effectively, categorical variables were converted into integer formats, as classifiers cannot directly interpret text data. This categorical encoding enhanced model predictions, aligning with the definition from [17], which describes the process as transforming categorical attributes into numerical form suitable for algorithmic ingestion. The categorical columns transformed included policy_csl, insured_sex, insured_education_level, insured_occupation, insured_relationship, incident_type, collision_type, incident_severity, authorities_contacted, property_damage, and police_report_available.

4.3 Data Transformation

The final dataset used for training and testing the models was created by extracting numerical columns and combining them with the converted numerical values from categorical data. An analysis was then conducted to identify anomalies and outliers prior to splitting the dataset into training and testing subsets. Numerical columns containing outliers were subsequently scaled to ensure the data's suitability for model training.

Additionally, the dataset included several attributes derived from two key variables relevant to the study, specifically relating to insurance policies and claims aligned with the targeted fraud detection in vehicle insurance claims. The features retained for classification by the machine learning models, selected based on their relevance to these variables were: months_as_customer, policy_csl, policy_deductible, policy_annual_premium, umbrella_limit, insured_sex, insured_education_level, insured_occupation, insured_relationship, capital_gains, property_damage, bodily_injuries, witnesses, incident_hour_of_the_day, number_of_vehicles_involved, injury_claim, property_claim, vehicle_claim, and fraud_reported..

4.4 Modelling

For this study, eight machine learning models were trained and tested to identify the best-performing algorithm on both unbalanced and balanced datasets. These models were chosen based on prior research demonstrating promising accuracy performance. The claims dataset was initially split into 80% for training and 20% for testing the classifiers. The accuracy of each classifier was then evaluated to determine the most effective model.

Following training and testing on the unbalanced dataset, the data was analyzed to separate genuine and fraudulent claims, identifying majority and minority class instances. To address class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) as used by [4] was applied to oversample the minority class. The balanced dataset was again divided into training (80%) and testing (20%) subsets.

To optimize model performance and ensure robust evaluation, K-Fold Cross Validation with K=10 was employed. The dataset was partitioned into ten folds, with each fold used once as the testing set while the remaining nine folds were used for training. This iterative process, cycling through all folds, allowed for comprehensive utilization of the available data for both training and testing phases, enhancing the reliability of the model assessment.

4.5 Evaluation

After training all the classifiers on both balanced and

unbalanced datasets, their performance was evaluated using test data to determine their ability to categorize claims as genuine or fraudulent. The study conducted an analysis of key performance metrics to assess each model's efficiency and effectiveness, as well as to establish appropriate risk thresholds. Metrics employed included the confusion matrix, classification accuracy, and classification reports based on recall, precision, and F1 score. This evaluation identified the classifier with the highest prediction performance and classification accuracy.

4.6 Deployment

A novel and effective system was developed using the machine learning classifier that demonstrated the highest prediction performance and classification accuracy for identifying fraudulent vehicle insurance claims. This system is expected to significantly enhance the long-term profitability and customer satisfaction of insurance companies.

5. MEASURING PERFORMANCE

The model's performance was evaluated using both unbalanced and balanced datasets, with classification reports generated for each classifier. These reports included metrics such as accuracy, recall, precision, and F1 score to identify the best-performing classifier.

5.1 Confusion Matrix

According to [18], a confusion matrix is a classification performance metric used to assess the effectiveness of a machine learning algorithm based on the target classes. It consists of True Positives (TP), which represent correctly classified positive claims; True Negatives (TN), which are correctly classified negative claims; False Positives (FP), which occur when negative claims are incorrectly classified as positive; and False Negatives (FN), where positive claims are incorrectly classified as negative. To analyze the results more thoroughly, the confusion matrices for both unbalanced and balanced datasets were examined.

In addition to the confusion matrix, other evaluation metrics such as precision, recall, and F1 score were used to gain deeper insights into the models' effectiveness. Precision measures the proportion of correctly identified class members among all instances predicted to belong to that class. Recall indicates the percentage of actual class members that were correctly predicted by the model. The F1 score combines precision and recall into a single metric, providing a balanced measure that is high only when both precision and recall are strong [18]. The results of these evaluations are presented in Tables 2 and 3.

Table 2. Unbalanced Dataset Evaluation Report

	TPs	FPs	TNs	FNs	Pre	Rec	F1
SVM	45	115	36	4	0.28	0.92	0.43
NB	13	27	121	39	0.33	0.25	0.28
LG	49	137	14	0	0.26	1.00	0.41
DT	32	38	110	20	0.46	0.62	0.52
RF	24	14	134	28	0.63	0.46	0.53
Ada	36	20	128	16	0.64	0.69	0.67
XGB	36	20	128	16	0.64	0.69	0.67
ANN	0	0	158	42	0.0	0.0	0.0

The results presented in the confusion matrix indicate that the ANN outperformed other classifiers on unbalanced data, achieving the highest percentage of true negatives. Random Forest ranked closely behind ANN, followed by AdaBoost and XGBoost. In contrast, the Logistic Regression classifier showed the poorest performance, exhibiting the lowest percentage of true negatives and the highest number of false positives, which implies misclassification of fraudulent claims as genuine. Given the inherent imbalance of fraudulent claims in the dataset, these findings suggest that AdaBoost, XGBoost, and ANN are more effective in handling unbalanced data, whereas Logistic Regression struggles under these conditions.

Table 3. Balanced Dataset Evaluation Report

	TPs	FPs	TNs	FNs	Pre	Rec	F1
SVM	40	32	124	106	0.58	0.73	0.64
NB	36	36	119	111	0.50	0.24	0.33
LG	146	147	8	1	0.50	0.99	0.66
DT	82	45	110	65	0.65	0.56	0.60
RF	110	46	138	37	0.87	0.75	0.80
Ada	126	19	136	21	0.87	0.86	0.86
XGB	126	19	136	21	0.87	0.86	0.86
ANN	147	155	0	0	0.49	1.00	0.65

The results on balanced data from the confusion matrix indicated that the ANN performed poorly, as it failed to identify any fraudulent claims, evidenced by the absence of true negative values. In contrast, the AdaBoost, XGBoost, and Random Forest classifiers achieved the highest true negative rates, demonstrating their superior ability to differentiate between genuine and fraudulent claims. Although the Logistic Regression classifier exhibited the lowest true negative rate and the highest false positive count after ANN, it consistently performed poorly on both balanced and unbalanced datasets. Overall, AdaBoost, XGBoost, and Random Forest attained the highest F1 scores, confirming their superiority over the other models.

5.2 Accuracy Evaluation

Accuracy is defined as the proportion of correctly predicted observations (True Positives) to the total number of observations, including True Positives, False Positives, False Negatives, and True Negatives [18]. In the classification accuracy performance tests conducted on eight classifiers using unbalanced data, the AdaBoost and XGBoost classifiers outperformed the others, each achieving an accuracy rate of 84.5%. The ANN classifier ranked second with an accuracy of 76.5%. Random Forest, Support Vector Machine (SVM), Naïve Bayes, and Decision Tree classifiers attained accuracy rates of 73.5%, 68.5%, 68.0%, and 63.5%, respectively. Logistic Regression performed the worst, with an accuracy of 26.0%, indicating it was ineffective at detecting fraudulent insurance claims.

Following this, the dataset was balanced using the SMOTE, after which all classifiers were retrained and retested. It was observed that the AdaBoost, XGBoost, Random Forest, and Logistic Regression classifiers improved their predictive accuracy, whereas the SVM and Naïve Bayes classifiers experienced a decline in performance. The Decision Tree classifier maintained a consistent accuracy rate of 63.5%, demonstrating robustness against data balancing. Notably, the ANN classifier's accuracy dropped significantly, suggesting it is less suited for identifying fraudulent insurance claims in balanced datasets. After balancing, AdaBoost and XGBoost continued to lead with accuracy rates of 86.75%, followed

closely by Random Forest at 82.1%.

5.3 Fraudulent Claims Detection System

Based on the results obtained, either the AdaBoost or XGBoost classifier was selected as the preferred model for integration into a web-based application designed to detect fraudulent vehicle insurance claims. The application was developed using the Streamlit framework, which, according to [19], is an open-source, Python-based platform enabling the creation and deployment of interactive data science dashboards and machine learning models as web applications. The system allows users to upload a CSV or Excel file containing vehicle insurance claims data, which the chosen model then processes to classify claims as either genuine or fraudulent. Upon uploading the data and applying the saved AdaBoost or XGBoost model, the application accurately categorizes the claims and generates a downloadable CSV file containing the classification results for user convenience.

6. DISCUSSION

The literature review by [20] indicates that XGBoost outperformed Logistic Regression, Support Vector Machine, and Random Forest classifiers. This study, which found XGBoost to be the superior classifier, aligns with and supports that conclusion. Additionally, the research demonstrated that ANNs are not the most effective classifiers when using balanced data for vehicle insurance claim classification. This finding is consistent with [21], whose neural network-based model also underperformed due to its inability to reliably detect fraudulent claims, mirroring the results of the current study. Furthermore, this study corroborates the work of [22], where features were derived from insurer information, vehicle type, maximum tonnage, insurance branch code, insured amount, paid loss, claim details, vehicle age, and similar factors. These elements were likewise incorporated in the present research to construct the feature set.

7. CONCLUSION

Machine learning techniques play a pivotal role in extracting valuable insights from large volumes of data, enabling the development of robust predictive models. In response to the growing challenge of fraudulent insurance claims, technological advancements have spurred extensive research and application of various detection methods. Despite these efforts, fraudsters continuously adapt and employ increasingly sophisticated tactics to bypass existing defenses, escalating the threat to organizational security and the broader economy.

Addressing the gap in literature, this study presents a practical and efficient web-based system that leverages the best-performing machine learning classifiers to categorize insurance claims as genuine or fraudulent. Among the evaluated models, AdaBoost and XGBoost emerged as the most effective, achieving the highest classification accuracy of 84.5% on both balanced and unbalanced datasets. Their strong performance suggests that the selected feature set effectively captures key characteristics of fraudulent claims, justifying their integration into the proposed web application.

Conversely, Logistic Regression consistently underperformed, yielding the lowest accuracy across datasets, highlighting its limitations in this complex classification task. The ANN demonstrated better results with unbalanced data but deteriorated when applied to balanced datasets, indicating sensitivity to data distribution. A significant operational limitation was observed: all eight classifiers evaluated were only feasible on smaller datasets due to computational constraints, as none could efficiently process large-scale data without overloading the available GPU resources.

Overall, this analysis underscores the strengths and weaknesses of diverse machine learning approaches in vehicle insurance fraud detection, emphasizing the necessity for scalable methods capable of handling real-world, large-volume datasets while maintaining high predictive accuracy.

8. REFERENCES

[1] Gill, K. M., Woolley, A., & Gill, M. (2005). Insurance Fraud: The Business as a Cictim? In *Palgrave Macmillan UK eBooks* (pp. 73–82). https://doi.org/10.1007/978-1-349-23551-3_6

[2] Association of Kenya Insurers, (2023). *Insurance Industry Market Report 2023.* Available at: https://www.akinsure.com/content/uploads/documents/Insurance_Industry_Market_Report_2023.pdf?t=0838

[3] Insurance Regulatory Authority (2023). *Insurance Industry Annual Report*. Available at: https://libraryir.parliament.go.ke/items/ba83e0b3-b7a2-4583-b671-c022e6da6ec5

[4] Subudhi, S., & Panigrahi, S. (2018b). Effect of Class Imbalanceness in Detecting Automobile Insurance Fraud. *2018 2nd International Conference on Data Science and Business Analytics (ICDSBA),* 528–531. https://doi.org/10.1109/icdsba.2018.00104

[5] Caruana, M. A., & Grech, L. (2021). Automobile Insurance Fraud Detection. *Communications in Statistics Case Studies Data Analysis and Applications*, 7(4), 520–535. https://doi.org/10.1080/23737484.2021.1986169

[6] Viaene, S., & Dedene, G. (2004). Insurance Fraud: Issues and Challenges. *The Geneva Papers on Risk and Insurance Issues and Practice*, 29(2), 313–333. https://doi.org/10.1111/j.1468-0440.2004.00290.x

[7] Moon, H., Pu, Y., & Ceglia, C. (2019). A Predictive Modeling for Detecting Fraudulent Automobile Insurance Claims. *Theoretical Economics Letters, 09*(06), 1886–1900. https://doi.org/10.4236/tel.2019.96120

[8] Owusu-Oware, E., Effah, J., & Boateng, R., (2018). Biometric Technology for Fighting Fraud in National Health Insurance: Ghana's Experience. *Americas Conference on Information Systems*.

[9] Dull, R. (2014). What Gets Monitored Gets Detected. *Journal of Accountancy, Feature Fraud/Technology*. http://www.journalofaccountancy.com/issues/2014/feb/20137694.html.

[10] Burri, R.D., Burri, R., Bojja, R.R., & Buruga, S.R. (2019). Insurance Claim Analysis Using Machine Learning Algorithms. *International Journal of Innovative Technology and Exploring Engineering 5*(6), Special Issue 4, pp.577-582.

[11] Dhieb, N., Ghazzai, H., Besbes, H., & Massoud, Y. (2019). Extreme Gradient Boosting Machine Learning Algorithm For Safe Auto Insurance Operations. *2019 IEEE International Conference on Vehicular Electronics and Safety (ICVES),* 1–5. https://doi.org/10.1109/icves.2019.8906396.

[12] Awoyemi, J.O., Adetunmbi, A.O., & Oluwadare, S.A., (2017). Credit Card Fraud Detection Using Machine Learning Techniques: A Comparative Analysis. *2017 International Conference on Computing Networking and*

Informatics (ICCNI), 1–9. https://doi.org/10.1109/iccni.2017.8123782.

[13] Gedela, B., & Karthikeyan, P. R. (2022). Credit Card Fraud Detection using AdaBoost Algorithm in Comparison with Various Machine Learning Algorithms to Measure Accuracy, Sensitivity, Specificity, Precision and F-score. *2022 International Conference on Business Analytics for Technology and Security (ICBATS)*, 1–6. https://doi.org/10.1109/icbats54253.2022.9759022.

[14] Mishra, A. (2021). Fraud Detection: A study of AdaBoost Classifier and K-Means Clustering. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3789879

[15] Wilson, J.H. (2009). An Analytical Approach To Detecting Insurance Fraud Using Logistic Regression. In *Journal of Finance and Accountancy.* 1–3. https://www.aabri.com/manuscripts/08103.pdf

[16] Pandey, P., (2019). Machine Learning Data Preprocessing: Concepts. Available at: https://towardsdatascience.com/data-preprocessing-concepts-fa946d11c825.

[17] Bolikulov, F., Nasimov, R., Rashidov, A., Akhmedov, F., & Cho, Y. (2024). Effective Methods of Categorical Data Encoding for Artificial Intelligence Algorithms. *Mathematics, 12*(16), 2553. https://doi.org/10.3390/math12162553

[18] Parab, R., (2020). Performance Evaluation Metrics for Machine Learning Models with Python Code. Available at: https://medium.com/swlh/performance-evaluation-metrics-for-machine-learning-models-ad0dd480d5af.

[19] Patil, S., and Lokesha, V., (2022). Live Twitter Sentiment Analysis Using Streamlit Framework. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4119949

[20] Tongesai, M., Mbizo, G., & Zvarevashe, K. (2022). Insurance Fraud Detection using Machine Learning. *2022 1st Zimbabwe Conference of Information and Communication Technologies (ZCICT), 3*, 1–6. https://doi.org/10.1109/zcict55726.2022.10046034

[21] Jalali, B., (2020). Detecting Fraudulent Claims – A Machine Learning Approach. In *Risk Insights: Vol. No. 1–2020.* https://www.genre.com/content/dam/generalreinsuranceprogram/documents/ri20-1-en.pdf

[22] Sunita, M., Prasun, G., & Parita, S. (2018). Management of Fraud: Case of an Indian Insurance Company. *Accounting and Finance Research, Sciedu Press, vol. 7*(3), 1-18. https://ideas.repec.org/a/jfr/afr111/v7y2018i3p18.html