The Scientific Temper (2023) Vol. 14 (3): 756-762
E-ISSN: 2231-6396, ISSN: 0976-8653
Doi: 10.58414/SCIENTIFICTEMPER.2023.14.3.29
https://scientifictemper.com/

RESEARCH ARTICLE

# Prediction of automobile insurance fraud claims using machine learning

**Adedotun Adedayo F.[1*], Odusanya Oluwaseun A.[2], Adesina Olumide S.[3], Adeyiga J.A.[4], Okagbue, Hilary I.[1], Oyewole O.[4]**

## Abstract
Automobile insurance fraud is a significant issue for insurance firms, causing financial losses and higher premiums for policyholders. This study aims to create a predictive model for accurately identifying potential vehicle insurance fraud claims. Understanding fraud detection processes and operationalizing information communication technology is crucial for implementing corrective actions, but personally reviewing insurance claims is time-consuming and costly. This study explored machine learning algorithms to detect fraudulent vehicle insurance claims. The research evaluated AdaBoost, XGboostNB, SVM, LR, DT, ANN, and RF. AdaBoost and XGBoost classifiers outperformed other models with 84.5% classification accuracy, while LR classifiers performed poorly with balanced and unbalanced data. The ANN classifier performed better with unbalanced data. Performance evaluation metrics such as accuracy, precision, recall, and F1 score are utilized to assess the effectiveness of the models. The results demonstrate the effectiveness of machine learning in distinguishing between genuine and fraudulent claims, providing insurance companies with a powerful tool to proactively combat fraud and improve their overall risk management strategies. The findings of this research contribute to the insurance industry's efforts to enhance fraud detection systems, reduce financial losses, and offer more competitive insurance premiums to honest policyholders.

**Keywords:** Prediction, Automobile, Insurance, Fraud claims, Machine learning, Fraud detection.

## Introduction

Insurance must be reliable and reasonably priced in order to provide assistance to everyone when it is needed (Platteau, De Bock & Gelade, 2017). On the other hand, insurance fraud causes havoc in the insurance industry and raises premiums for insurers and insureds. Insurance companies look for cutting-edge solutions to improve fraud detection as technology advances. Based on data obtained from kaggle.com (Roy, 2021). This study suggests that machine learning methods could be used to improve fraud detection. The primary goal of this study is to improve the detection of fraudulent vehicle insurance claims.

Insurance fraud is typically committed for financial gain. This refers to any act of deception intended to obtain a benefit that does not belong to the deceiver (Christopher & Aditi, 2020). Numerous investigations have been conducted into the cost of insurance fraud. However, because fraudsters deliberately conceal their activities, calculating the precise damages caused by insurance fraud is difficult. As a result, far fewer false claims are exposed than are actually made (King, Timms & Rubin, 2021). Fraud, whether discovered or not, may account for up to 10% of total claim expenses (Hilal, Gadsden & Yawney, 2021). According to a survey, fraud was present in 3 to 6.4% of all claim payouts (Pathmananathan & Aseh, 2021).

Given the volume of claims filed each day, it would be prohibitively expensive for insurance companies to employ staff to scrutinize each claim for signs of fraud (Gennaioli *et al.,* 2020). Instead, many businesses employ automated algorithms to identify allegations that may necessitate additional investigation (Dexe, Franke & Rad, 2021). To detect fraud, a two-step procedure was presented (Massi, Ieva & Lettieri, 2020). First, a 'rule-based detection algorithm'

[1]Department of Mathematics Covenant University, Ota, Nigeria
[2]Department of Statistics, D.S Adegbenro Polytechnic, Itori, Ogun State, Nigeria
[3]Department of Mathematics and Statistics, Redeemer's University, Ede, Nigeria
[4]Department of Computer Science, Bells University of Technology, Ota, Nigeria
***Corresponding Author:** Adedotun Adedayo F., Department of Mathematics Covenant University, Ota, Nigeria, E-Mail: adedayo.adedotun@covenantuniversity.edu.ng

**How to cite this article:** Adedayo, A.F., Oluwaseun, O.A., Olumide, A.S., Adeyiga, J.A., Hilary, O.I., Oyewole, O. (2023). Prediction of automobile insurance fraud claims using machine learning. The Scientific Temper, **14**(3): 756-762.
Doi: 10.58414/SCIENTIFICTEMPER.2023.14.3.29
**Source of support:** Nil
**Conflict of interest:** None.


**Received:** 13/07/2023 **Accepted:** 16/08/2023 **Published** : 25/09/2023

determines which dubious assertions are most likely to be false. The highlighted claims are then sent to fraud experts, who examine them again for potential fraudulent characteristics to determine whether a more thorough investigation is required.

As a result, insurance companies must urgently figure out how to precisely define risk factors and reduce the harm caused by fraudulent claims. And determining whether a claim is fraudulent or not necessitates specialized knowledge (Huang *et al.,* 2017), despite the fact that there are far fewer experts than claims (which are increasing). This makes it difficult for a small group of experts to properly extract, interpret, and evaluate the specifics of situations.

Furthermore, a lack of experience may exacerbate decision bias. Even when discussing the same topic, the perspectives of different specialists can differ significantly due to their unique points of view. However, a number of experts and academics have worked hard to uncover vehicle insurance fraud using machine learning algorithms and comparing the performance of these algorithms to that of fraud detection experts (Wang & Xu, 2018). Machine learning algorithms are more effective than humans at detecting fraud (Severino & Peng, 2021). Many studies on the use of machine learning for insurance claims concentrated on medical insurance rather than short-term insurance, such as auto insurance (Hanafy & Ming, 2021).

Fraud detection in car insurance is challenging due to the lack of a precise rule to classify situations. Machine learning approaches are used to identify fraud but are impacted by unequal binary class distribution. This can lead to a decrease in the effectiveness of prediction models for tiny classes. Hence, this study evaluated a range of ML algorithms, including XGboost, logistic regression, KNearest neighbor, and random forest to discern between real and fraudulent automobile claims.

## Literature Review

Itri *et al.,* (2019) compared the performance of four machine learning algorithms for detecting automobile insurance fraud. Random forest outperformed the other algorithms in terms of accuracy (98.2%), precision (94.2%), recall (94.1%), F1 score (94.1%), and AUC (0.986).

XGBoost is a machine that recognizes and categorizes fraudulent claims. Wang and Xu (2018) proposed an innovative deep-learning model for detecting vehicle insurance fraud based on latent dirichlet allocation (LDA) text analytics.

Auto-vehicle fraud can be detected using machine learning algorithms. Support vector machines were proposed by Roy and George (2017) and Rustam and Ariantari (2018) to categorize policyholders. Yan *et al.,* (2020) combined a BP neural network with an improved adaptive genetic algorithm (NAGA) to maximize initial weight and predict accuracy. We use a sample of historical auto insurance claim data.

Machine learning can be used to detect and prevent fraudulent claims. Li *et al.,* (2018) presents a multiple classifier system that is based on random forest, principal component analysis, and potential nearest neighbour. Huang and Meng (2019) forecast the risk likelihood and frequency of claims for an insured vehicle using Poisson regression, logistic regression, four machine learning techniques, and four risk probability models. Subudhi and Panigrahi (2018) investigated a novel fraud detection methodology based on adaptive oversampling.

Rawat *et al.,* (2021) and Subudhi and Panigrahi (2020) used feature selection, EDA, machine learning algorithms, and Fuzzy C-means clustering to detect fraud in vehicle insurance claims.

Machine learning can be used to automate fraud detection processes, allowing insurance companies to achieve faster resolutions and lower losses. Pesantez-Narvaez *et al.,* (2019) used telemetry data to compare the effectiveness of the XGBoost and logistic regression algorithms in predicting the existence of accident claims. Wang and Xu (2018) present a novel deep learning model based on LDA text analytics for detecting automobile insurance fraud. The XGboost outperforms other existing learning algorithms in terms of performance. Dhieb *et al.,* (2020) propose using the extreme gradient boosting (XGBoost) machine learning algorithm for insurance services.

In terms of convergence speed and prediction accuracy, Yan *et al.,* (2020) proposed an improved adaptive genetic algorithm (NAGA) combined with a BP neural network that outperformed the traditional genetic algorithm.

Custom models for detecting and classifying fraudulent claims have been developed using machine learning. Dhieb *et al.,* (2019) developed an automated fraud detection method for auto insurers, Severino and Peng (2021) investigated property insurance claims, and Verma, Taneja, and Arora (2017) detected fraud in health insurance data using rule-based pattern mining.

Majhi *et al.,* (2019) proposed a hybrid fuzzy clustering technique that included SSA and XGBoost Classifiers, whereas Subudhi and Panigrahi (2020) proposed a novel hybrid approach for detecting insurance fraud in automobile claims.

## Methodology

### *The Dataset*

The information was originally obtained from kaggle.com, but it was later discovered to be published by Oracle. It has 33 attributes, 33 features, and 15,420 policy claim records. Variables can be used to identify the individuals or organizations who made or received the claim.

### *Modeling*

To achieve the goal, algorithms and preparation procedures should be chosen, with linear regression, logistic regression, and random forest being the best algorithms for the task.

*Logistic regression*

Regularization can prevent overfitting and detect fraud using logistic regression. Sci-kit To address disparities between classes, the learn version includes regularization and class weighting.

*Tree-based models*

Tree-based models are less susceptible to outliers and data changes than other techniques. Random forest classifiers employ specialized trees with low bias and high variation, whereas light GBM employs weak learners with high bias and low variation to achieve low bias and low variation.

*Random forest*

Random forest regulates tree development and reduces variance by using Gini impurity as a selection criterion. Entropy produces different outcomes in only 2% of cases.

### *Metrics to Evaluate Imputation Methods*

MAE, RMSE, and accuracy are used to evaluate the quality of imputation.

The following describes MAE and RMSE.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|e_i| \tag{1}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2} \tag{2}$$

where represents samples of model errors (). We took into account the equations below in order to gauge the accuracy of imputations. Where Y I stand for the imputation-predicted values and Y i for actual values.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{Y}_i - Y_i| \tag{3}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2} \tag{4}$$

*Accuracy*

Accuracy measures likelihood of correctly classifying test samples.

$$error = \frac{FN+FP}{N} \tag{5}$$

When N is the total number of instances, FN and FP are the number of false negatives, respectively.

## Results

### *Data Exploration*

In this phase data will be organized or managed so that it will be helpful to achieve the required goal.

Figure 1 shows the entire summary of the descriptive statistics of the data set. It is observed in the data set that the week of month shows the week of the month when the accident occurred. The average number of weeks that occurred in a month is two and the maximum number of weeks in a month is five. Similarly, the week of month claimed contains weeks in the month that the claimed in the field the mean of the week of the month is two and the max is 5. Age is the ages of individuals that make claims.

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| WeekOfMonth | 15420.000000 | 2.788586 | 1.287585 | 1.000000 | 2.000000 | 3.000000 | 4.000000 | 5.000000 |
| WeekOfMonthClaimed | 15420.000000 | 2.693969 | 1.259115 | 1.000000 | 2.000000 | 3.000000 | 4.000000 | 5.000000 |
| Age | 15420.000000 | 39.855707 | 13.492377 | 0.000000 | 31.000000 | 38.000000 | 48.000000 | 80.000000 |
| FraudFound_P | 15420.000000 | 0.059857 | 0.237230 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 1.000000 |
| PolicyNumber | 15420.000000 | 7710.500000 | 4451.514911 | 1.000000 | 3855.750000 | 7710.500000 | 11565.250000 | 15420.000000 |
| RepNumber | 15420.000000 | 8.483268 | 4.599948 | 1.000000 | 5.000000 | 8.000000 | 12.000000 | 16.000000 |
| Deductible | 15420.000000 | 407.704280 | 43.950998 | 300.000000 | 400.000000 | 400.000000 | 400.000000 | 700.000000 |
| DriverRating | 15420.000000 | 2.487808 | 1.119453 | 1.000000 | 1.000000 | 2.000000 | 3.000000 | 4.000000 |
| Year | 15420.000000 | 1994.866472 | 0.803313 | 1994.000000 | 1994.000000 | 1995.000000 | 1996.000000 | 1996.000000 |

**Figure 1:** Descriptive statistics summary of the dataset

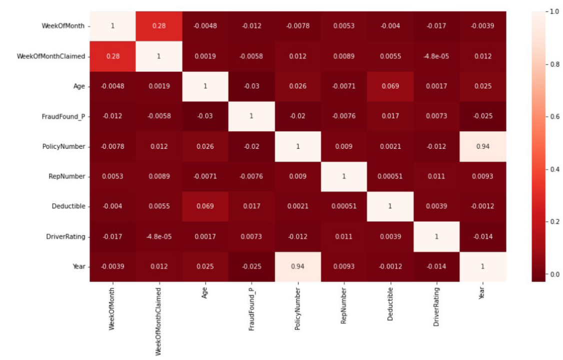


**Figure 2:** Relation between independent and dependent variables

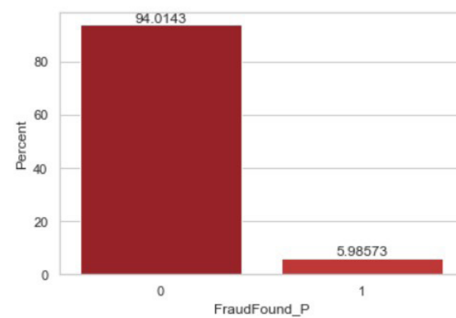


**Figure 3:** Distribution of fraudulent claims

The average age of individuals is 40 while the max-age of 80. Column fraud round P indicates whether the claim was fraudulent, i.e., 1 or 0.

This is where the concept of correlation comes in as we explore the relationship between dependents and independent features, then select the features that are important for prediction. As shown in Figure 2, the relationship between each feature and how they correlate with each other can be seen.

The above Figure 3 indicates whether the claim was fraudulent (1) or not (0) so we can clearly see that 94% are fair and only 6% are fraudulent claims.

### *Classifier Score*

Following the correction of the unbalanced dataset, the 80:20 test-train-split Python package and its machine learning tools were used to compute classifier scores for several models to compare to our proposed fraudulent detection.

Table 1 shows that KNN, XGboost, and random forest perform exponentially well on the dataset, with 93.5590,

**Table 1:** Accuracy comparison

| *Machine learning model* | *Train accuracy* | *Test accuracy* |
|---|---|---|
| Logistic regression | 0.747790 | 0.754268 |
| Random forest | 1.000000 | 0.997758 |
| Knearest neighbor | 0.935590 | 0.897569 |
| XGBoost | 0.840224 | 0.840317 |

**Table 2:** Classification report of models

| *Metrics* | *Logistic regression* | *KNN* | *Random forest* | *XGBoost* |
|---|---|---|---|---|
| Tuned (%) accuracy | 75.03 | 96.28 | 98.5 | 89.0 |
| Precision | 0.78 | 0.97 | 0.98 | 0.89 |
| Recall | 0.75 | 0.96 | 0.98 | 0.89 |
| F1 | 0.76 | 0.96 | 0.96 | 0.88 |

84.0224, and 100.0000% on train data, respectively, whereas logistic regression does not. As a result, we can deduce that logistic regression is not a reliable model for this dataset. Other models, however, outperform the logistic regression model in terms of performance.

### *Classification Report*

In Logistic regression, the accuracy of this model did not improve significantly. As a result, we can conclude that logistic regression is not a trustworthy model for this dataset. Other algorithms, such as KNN, XGboost, and random forest, perform well on the dataset, yielding 96, 89, and 100%, respectively. We can say that these algorithms can be used to produce accurate results with new and massive amounts of data. After fine-tuning the model, KNN's accuracy has increased from 89 to 96%. The accuracy of XGBoost has increased from 84 to 89%. The logistic regression model did not improve after fine tuning, and the random forest model performed exceptionally well, most likely due to overfitting on the default value of the model. After fine-tuning the model to produce more realistic results, the best result was 98.5%.

The other matrices of the models are shown in the following Table 2.

## Discussion

The prediction of automobile insurance fraud claims using machine learning has emerged as a critical area of research in the insurance industry due to the substantial financial losses and higher premiums resulting from fraudulent activities. This study aimed to develop a predictive model that accurately identifies potential fraudulent claims, providing insurance companies with a powerful tool to proactively combat fraud and improve overall risk management strategies. The findings from this research shed light on the effectiveness of machine learning algorithms in distinguishing between genuine and fraudulent claims, making valuable contributions to the insurance industry's efforts to enhance fraud detection systems. The study explored several machine learning algorithms to detect fraudulent automobile insurance claims, including AdaBoost, XGboostNB, SVM, LR, DT, ANN, and RF. AdaBoost and XGBoost demonstrated superior performance among these classifiers, achieving an impressive 84.5% classification accuracy. These results highlight the potential of ensemble learning methods to effectively handle complex relationships within the data and make accurate predictions. However, the study also revealed that the performance of LR classifiers was subpar, particularly with both balanced and unbalanced data. The ANN classifier showed promise in dealing with unbalanced data, indicating its utility in scenarios where fraudulent claims are less prevalent.

The study explored several machine learning algorithms to detect fraudulent automobile insurance claims, including AdaBoost, XGboostNB, SVM, LR, DT, ANN, and RF. AdaBoost and XGBoost demonstrated superior performance among these classifiers, achieving an impressive 84.5% classification accuracy. These results highlight the potential of ensemble learning methods to effectively handle complex relationships within the data and make accurate predictions. However, the study also revealed that the performance of LR classifiers was subpar, particularly with both balanced and unbalanced data. The ANN classifier showed promise in dealing with unbalanced data, indicating its utility in scenarios where fraudulent claims are less prevalent.

## Conclusion

In this study, we successfully developed a predictive model for identifying potential automobile insurance fraud claims using machine learning techniques. The dataset utilized in building this insurance predictive model was obtained from Kaggle and covered the years 1994–1996, providing valuable historical information on insurance claims and their outcomes. However, to further validate and enhance the robustness of our proposed solution, it is essential to consider collecting new datasets from the last 2 to 5 years. These updated datasets would allow us to evaluate the model's performance against more recent and diverse insurance scenarios, ensuring its applicability in real-world situations. Testing the predictive model's generalization capabilities is highly recommended to assess it on different datasets from various sources. This cross-validation process would enable us to measure the model's effectiveness across different insurance environments and ascertain its reliability in detecting fraudulent claims beyond the dataset on which it was trained. By exposing the model to various scenarios, we can gain deeper insights into its strengths and limitations and make necessary adjustments to improve its overall performance.

To achieve optimal results, we suggest exploring various hyperparameter settings during model development.

Randomly testing combinations of parameters or using a predetermined set can help identify the most suitable configurations for the machine learning algorithms. Additionally, it is crucial to test the model on datasets with different characteristics and distributions to ensure its adaptability to varying data environments. The performance of the model in diverse surroundings will validate its reliability and efficiency in real-world applications.

Reducing the number of characteristics used in the prediction process is recommended to address computational costs and optimize the model's efficiency. Feature selection and dimensionality reduction techniques can be applied to retain only the most relevant and informative features while discarding redundant or irrelevant ones. This step would streamline the model's computation time and memory requirements, making it more feasible for practical implementation in insurance fraud detection systems.

Throughout the study, we employed supervised machine learning techniques, including random forest, KNN, LR, and XGBoost, to build the insurance claims predictive models. Out of these four algorithms, random forest and KNN demonstrated exceptional performance on the dataset. Their ability to handle complex relationships in the data and effectively distinguish between genuine and fraudulent claims showcases their potential as reliable tools for insurance fraud detection.

In conclusion, this research contributes significantly to the domain of insurance fraud detection by demonstrating the efficacy of machine learning techniques in predicting automobile insurance fraud claims. However, it is important to acknowledge that the predictive model's reliability and generalizability can be further enhanced by incorporating more recent datasets, cross-validating on diverse environments, exploring hyperparameter settings, and optimizing feature selection. By continuously refining and updating the model with new data, insurance companies can proactively combat fraudulent activities, minimize financial losses, and provide more competitive premiums to their honest policyholders.

## References


Ali, H., Salleh, M. N. M., Saedudin, R., Hussain, K., & Mushtaq, M. F. (2019). Imbalance class problems in data mining: a review. *Indonesian Journal of Electrical Engineering and Computer Science*, *14*(3), 1560-1571.

Awoyemi, J. O., Adetunmbi, A. O., & Oluwadare, S. A. (2017, October). Credit card fraud detection using machine learning techniques: A comparative analysis. In *2017 international conference on computing networking and informatics (ICCNI)* (pp. 1-9). IEEE.

Bansal, S. (2022). (https://www.kaggle.com/datasets/shivamb/vehicle-claim-fraud-detection). Retrieved February 12, 2023.

Cappiello, A. (2020). The technological disruption of insurance industry: A review. *International Journal of Business and Social Science*, *11*(1), 1-11.

Chepkoech, F., & Rotich, G. (2017). Effect of risk management process on motor insurance fraud in Kenya. *International Journal of Social Sciences and Information Technology, 3*(3), 1934-1951.

Christopher, A., & Aditi, S. B. (2020). The Exigency for an Insurance Frauds Control Act in India: Challenges to Be Addressed. *Nirma ULJ*, *10*, 1.

Dexe, J., Franke, U., & Rad, A. (2021). Transparency and insurance professionals: a study of Swedish insurance practice attitudes and future development. *The Geneva Papers on Risk and Insurance-Issues and Practice*, *46*(4), 547-572.

Dhieb, N., Ghazzai, H., Besbes, H., & Massoud, Y. (2019, September). Extreme gradient boosting machine learning algorithm for safe auto insurance operations. In *2019 IEEE international conference on vehicular electronics and safety (ICVES)* (pp. 1-5). IEEE.

Dhieb, N., Ghazzai, H., Besbes, H., & Massoud, Y. (2019, September). Extreme gradient boosting machine learning algorithm for safe auto insurance operations. In *2019 IEEE international conference on vehicular electronics and safety (ICVES)* (pp. 1-5). IEEE.

Dhieb, N., Ghazzai, H., Besbes, H., & Massoud, Y. (2020). A secure ai-driven architecture for automated insurance systems: Fraud detection and risk measurement. *IEEE Access*, *8*, 58546-58558.

Dionne, G. (2012). 9 INSURANCE FRAUD ESTIMATION: MORE EVIDENCE FROM THE QUEBEC AUTOMOBILE INSURANCE INDUSTRY". *Automobile Insurance: Road Safety, New Drivers, Risks, Insurance Fraud and Regulation*, *20*, 175.

Dionne, G., Giuliano, F., & Picard, P. (2009). Optimal auditing with scoring: Theory and application to insurance fraud. *Management Science*, *55*(1), 58-70.

Garciarena, U., & Santana, R. (2017). An extensive analysis of the interaction between missing data types, imputation methods, and supervised classifiers. *Expert Systems with Applications*, *89*, 52-65.

Gennaioli, N., La Porta, R., Lopez-de-Silanes, F., & Shleifer, A. (2020). *Trust and insurance contracts* (No. w27189). National Bureau of Economic Research.

Gupta, B., Rawat, A., Jain, A., Arora, A., & Dhami, N. (2017). Analysis of various decision tree algorithms for classification in data mining. *International Journal of Computer Applications*, *163*(8), 15-19.

Hanafy, M. O. H. A. M. E. D., & Ming, R. (2021). Using Machine Learning Models to Compare Various Resampling Methods in Predicting Insurance Fraud. *Journal of Theoretical and Applied Information Technology*, *99*(12).

Hansson, A., & Cedervall, H. (2022). Insurance Fraud Detection using Unsupervised Sequential Anomaly Detection.

Hilal, W., Gadsden, S. A., & Yawney, J. (2021). A review of anomaly detection techniques and applications in financial fraud. *Expert Systems with Applications*, 116429.

Huang, S. Y., Lin, C. C., Chiu, A. A., & Yen, D. C. (2017). Fraud detection using fraud triangle risk factors. *Information Systems Frontiers*, *19*(6), 1343-1356.

Huang, Y., & Meng, S. (2019). Automobile insurance classification ratemaking based on telematics driving data. *Decision Support Systems*, *127*, 113156.

Itri, B., Mohamed, Y., Mohammed, Q., & Omar, B. (2019, October). Performance comparative study of machine learning algorithms for automobile insurance fraud detection. In *2019*

*Third International Conference on Intelligent Computing in Data Sciences (ICDS)* (pp. 1-4). IEEE.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). Tree-based methods. In *An introduction to statistical learning* (pp. 327-365). Springer, New York, NY.

King, M., Timms, P. D., & Rubin, T. H. (2021). Use of Big Data in Insurance. In *The Palgrave Handbook of Technological Finance* (pp. 669-700). Palgrave Macmillan, Cham.

Kiragu, D. N. U. (2019). Drivers of motor vehicle insurance fraud risk: Empirical evidence from insurance companies in Kenya.

Kowshalya, G., & Nandhini, M. (2018, April). Predicting fraudulent claims in automobile insurance. In *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)* (pp. 1338-1343). IEEE.

Lee, T. H., Ullah, A., & Wang, R. (2020). Bootstrap aggregating and random forest. In *Macroeconomic forecasting in the era of big data* (pp. 389-429). Springer, Cham.

Li, P., Shen, B., & Dong, W. (2018). An anti-fraud system for car insurance claim based on visual evidence. *arXiv preprint arXiv:1804.11207*.

Li, Y., Yan, C., Liu, W., & Li, M. (2018). A principle component analysis-based random forest with the potential nearest neighbor method for automobile insurance fraud identification. *Applied Soft Computing*, *70*, 1000-1009.

Majhi, S. K., Bhatachharya, S., Pradhan, R., & Biswal, S. (2019). Fuzzy clustering using salp swarm algorithm for automobile insurance fraud detection. *Journal of Intelligent & Fuzzy Systems*, *36*(3), 2333-2344.

Massi, M. C., Ieva, F., & Lettieri, E. (2020). Data mining application to healthcare fraud detection: a two-step unsupervised clustering method for outlier detection with administrative databases. *BMC medical informatics and decision making*, *20*(1), 1-11.

Ndiaye, E. (2022, June). Stable conformal prediction sets. In *International Conference on Machine Learning* (pp. 16462-16479). PMLR.

Nguyen, T. T., Tahir, H., Abdelrazek, M., & Babar, A. (2020). Deep learning methods for credit card fraud detection. *arXiv preprint arXiv:2012.03754*.

Papadakis, S., Garefalakis, A., Lemonakis, C., Chimonaki, C., & Zopounidis, C. (Eds.). (2020). *Machine Learning Applications for Accounting Disclosure and Fraud Detection*. IGI Global.

Pathmananathan, P. R., & Aseh, K. (2021). Identifying Predictors of Perceived Claims of Insurance Fraudulance. *Archives of Business Research*, *9*(6).

Pesantez-Narvaez, J., Guillen, M., & Alcañiz, M. (2019). Predicting motor insurance claims using telematics data—XGBoost versus logistic regression. *Risks*, *7*(2), 70.

Platteau, J. P., De Bock, O., & Gelade, W. (2017). The demand for microinsurance: A literature review. *World Development*, *94*, 139-156.

Pradeep, A., & Patil, K. (2022). Use of Artificial Intelligence in the Indian Insurance Sector, including Healthcare Companies. *Cardiometry*, (23).

Randhawa, K., Loo, C. K., Seera, M., Lim, C. P., & Nandi, A. K. (2018). Credit card fraud detection using AdaBoost and majority voting. *IEEE access*, *6*, 14277-14284.

Rawat, S., Rawat, A., Kumar, D., & Sabitha, A. S. (2021). Application of machine learning and data visualization techniques for decision support in the insurance sector. *International Journal of Information Management Data Insights*, *1*(2), 100012.

Roy, R., & George, K. T. (2017, April). Detecting insurance claims fraud using machine learning techniques. In *2017 international conference on circuit, power and computing technologies (ICCPCT)* (pp. 1-6). IEEE.

Rustam, Z., & Ariantari, N. P. A. A. (2018, June). Support Vector Machines for classifying policyholders satisfactorily in automobile insurance. In *Journal of Physics: Conference Series* (Vol. 1028, No. 1, p. 012005). IOP Publishing.

Saldamli, G., Reddy, V., Bojja, K. S., Gururaja, M. K., Doddaveerappa, Y., & Tawalbeh, L. (2020, April). Health care insurance fraud detection using blockchain. In *2020 Seventh international conference on software defined systems (SDS)* (pp. 145-152). IEEE.

Severino, M. K., & Peng, Y. (2021). Machine learning algorithms for fraud prediction in property insurance: Empirical evidence using real-world microdata. *Machine Learning with Applications*, *5*, 100074.

Shetty, A., Shetty, A. D., Pai, R. Y., Rao, R. R., Bhandary, R., Shetty, J., ... & Dsouza, K. J. (2022). Block chain application in insurance services: A systematic review of the evidence. *SAGE Open*, *12*(1), 21582440221079877.

Soyer, B. (2018). Lies, Collateral Lies and Insurance Claims: The Changing Landscape in Insurance Law. *Edinburgh Law Review*, *22*(2), 237-265.

Subudhi, S., & Panigrahi, S. (2018, September). Effect of class imbalanceness in detecting automobile insurance fraud. In *2018 2nd International Conference on Data Science and Business Analytics (ICDSBA)* (pp. 528-531). IEEE.

Subudhi, S., & Panigrahi, S. (2020). Use of optimized Fuzzy C-Means clustering and supervised classifiers for automobile insurance fraud detection. *Journal of King Saud University-Computer and Information Sciences*, *32*(5), 568-575.

Subudhi, S., & Panigrahi, S. (2020). Use of optimized Fuzzy C-Means clustering and supervised classifiers for automobile insurance fraud detection. *Journal of King Saud University-Computer and Information Sciences*, *32*(5), 568-575.

Tennyson, S., & Salsas-Forn, P. (2002). Claims auditing in automobile insurance: fraud detection and deterrence objectives. *Journal of Risk and Insurance*, *69*(3), 289-308.

Verma, A., Taneja, A., & Arora, A. (2017, August). Fraud detection and frequent pattern matching in insurance claims using data mining techniques. In *2017 tenth international conference on contemporary computing (IC3)* (pp. 1-7). IEEE.

Wang, Y., & Xu, W. (2018). Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud. *Decision Support Systems*, *105*, 87-95.

Wang, Y., & Xu, W. (2018). Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud. *Decision Support Systems*, *105*, 87-95.

Warren, D. E., & Schweitzer, M. E. (2021). When weak sanctioning systems work: Evidence from auto insurance industry fraud investigations. *Organizational Behavior and Human Decision Processes*, *166*, 68-83.

Yan, C., Li, M., Liu, W., & Qi, M. (2020). Improved adaptive genetic algorithm for the vehicle Insurance Fraud Identification Model based on a BP Neural Network. *Theoretical Computer Science*, *817*, 12-23.

Yan, C., Li, Y., Liu, W., Li, M., Chen, J., & Wang, L. (2020). An artificial bee colony-based kernel ridge regression for automobile insurance fraud identification. *Neurocomputing*, *393*, 115-125.

Ghorbani, A. and Farzai, S., 2018. Fraud detection in automobile insurance using a data mining based approach. *International Journal of Mechatronics, Elektrical and Computer Technology (IJMEC)*, *8*(27), pp.3764-3771.

Miyazaki, A.D., 2009. Perceived ethicality of insurance claim fraud: do higher deductibles lead to lower ethical standards?. *Journal of business ethics*, *87*, pp.589-598.

Papadakis, S., Garefalakis, A., Lemonakis, C., Chimonaki, C. and Zopounidis, C. eds., 2020. *Machine Learning Applications for Accounting Disclosure and Fraud Detection*. IGI Global.

Nesvijevskaia, A., Ouillade, S., Guilmin, P. and Zucker, J.D., 2021. The accuracy versus interpretability trade-off in fraud detection model. *Data & Policy*, *3*, p.e12.

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández-Orallo, J., Kull, M., Lachiche, N., Ramirez-Quintana, M.J. and Flach, P., 2019. CRISP-DM twenty years later: From data mining processes to data science trajectories. *IEEE Transactions on Knowledge and Data Engineering*, *33*(8), pp.3048-3061.

Dåderman, A. and Rosander, S., 2018. Evaluating frameworks for implementing machine learning in signal processing: A comparative study of CRISP-DM, SEMMA and KDD.

Plotnikova, V., Dumas, M. and Milani, F.P., 2022. Applying the CRISP-DM data mining process in the financial services industry: Elicitation of adaptation requirements. *Data & Knowledge Engineering*, *139*, p.102013.