e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:12/December-2023 Impact Factor- 7.868 www.irjmets.com

COMPARATIVE ANALYSIS OF MACHINE LEARNING MODELS FOR AUTO INSURANCE FRAUD DETECTION

Akshay Pawar*1, Pallavi Thakur*2

*1,2Dept. Of Master Of Computer Applications Sardar Patel Institute Of Technology Mumbai, India.

DOI : https://www.doi.org/10.56726/IRJMETS47811

ABSTRACT

Auto insurance fraud is a significant problem in the insurance industry, with losses estimated to be in the billions of dollars annually. Insurance fraud can take many forms, including staged accidents, false injury claims, and exaggerated damage claims. The impact of auto insurance fraud is significant for both insurance companies and policyholders. Insurance fraud results in higher premiums for policyholders, as insurance companies pass on the costs of fraudulent claims to their customers. Fraudulent claims can also result in delays in claims processing and payouts, as insurance companies must investigate claims to determine their validity. Goal is to create a solution that analyzes auto insurance claim data to identify cases of fraud claims. Using machine learning algorithms, models are built to classify these claims. Also, This study compares the average accuracy, precision, recall, and other characteristics of all classification based machine learning methods. A machine learning model is created for fraudulent transaction validation using the scikit-learn Python Library.

Keywords: Machine Learning Algorithm, Scikit-Learn, Fraud Case Detection, Classifications.

I. INTRODUCTION

False claims for insurance benefits relating to motor accidents or theft are one sort of fraudulent activity known as "auto insurance fraud.". Auto insurance fraud is the practice of intentionally deceiving an insurance company to receive a financial benefit. It can take various forms, such as staging car accidents, submitting false claims, exaggerating the extent of damages or injuries, and misrepresenting information on insurance applications.

Insurance fraud is a serious crime that can have significant consequences for both the individual committing the fraud and the insurance industry as a whole. It can result in increased insurance premiums for everyone, as insurers may pass the cost of fraudulent claims onto their customers. Additionally, it can cause financial losses for insurance companies, which may result in decreased coverage options and higher deductibles. When a client attempts to obtain financial benefits by submitting fictitious documentation in response to the injury or damage of assets in fake incidents, or by claiming compensation for prior losses or excessive billing, auto insurance fraud occurs. Fraudsters frequently misrepresent the circumstances surrounding an occurrence in order to obtain higher rates. They do this by exaggerating the events and effects of the incident. [2]

Up to 1,388 fraud cases in 96 nations resulted in losses of 1.4 billion US dollars. The average annual loss for an organization is 7% of its gross income. [3]

There are many reasons why someone might commit auto insurance fraud. Some people do it for financial gain, while others may do it out of desperation or in an attempt to cover up other illegal activities. Regardless of the motive, insurance fraud is illegal and can result in criminal charges, fines, and even jail time.

II. LITERATURE REVIEW

There are several studies conducted on Auto insurance fraud detection with various machine learning algorithms, In (2018) [3] identify the insurance fraud by using Nearest Neighbour and Statistics Method, In (2022) [1] several machine learning classification algorithms are used on Auto insurance dataset for predictions and results are compared. In (2015) [7] Chun Yan and Yaqi Li, Based on Data mining, constructed and identified a machine learning algorithm and model to detect Automobile insurance fraud. S. Padhi and S. Panigrahi, in 2019 [8] used ensemble Classifiers based on Decision trees to detect Auto insurance frauds. N. Dhieb, H. Ghazzai, H. Besbes and Y. Massoud, in (2019) [9] uses Extreme gradient boosting Algorithms for safe auto insurance operations. Robust Fuzzy Rule based Technique to Detect Frauds in Vehicle Insurance (2017) [10] shows study conducted on Vehicle insurance frauds. In (2020) [11] Study shows Automobile insurance detection using Supervised classifier models.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:12/December-2023 Impact Factor- 7.868 www.irjmets.com

There are Many practical applications of machine learning that can be found in fields such as computer vision, natural language processing, and data analysis. It has the potential to transform many industries and improve human lives in numerous ways. Auto insurance fraud is classified into several types. The goal of this study is to predict whether a case is fraudulent or not using various Machine learning techniques. Various Machine learning algorithms are used to generate accurate results and to train the model.

Using ensemble learning techniques, this project seeks to identify both fraudulent and legitimate insurance claims. It also compares the average accuracy of several machine learning classification algorithms.

III. PROPOSED METHOD/ALGORITHM

Data preprocessing Applied to the complete dataset, before training the model through the dataset. The cleaned dataset is used to train a variety of classification algorithms. By using cross-validation models' average accuracy is determined and compared. After training, In order to determine if an insurance claim is fraudulent or not, a predictive model is created. Prediction-based binary classification yields a value between 0 (Non-Fraud) and 1 (Fraud), depending on the case.

A. Implementation Model

Following are proposed methods for model development.

1. Various classification models are used and compared to identify models with highest average accuracy.
2. Logistic regression, XGB, KNN, SVC, Decision Tree, Random forest, LDA and Guassian Naive Bayes, Algorithms are used for model comparison.
3. Average accuracy of models are calculated by taking the mean of 10-fold cross validation's result.
4. Algorithms are compared based on the Speed, Accuracy, predictors types, etc.
5. Binary Classification task takes place which gives answers between YES or NO.
6. Top 3 highest accurate models are selected in voting ensemble technique where a combination of these models is used to create a more accurate model.
7. Other ensemble techniques are also used to create the accurate and robust model with greater precision. Logistic regression, SVC, Random Forest and LDA used for binary classification only, while remaining algorithms are used for both binary as well as multiclass classification.

Logistic regression, XGB, KNN and LDA only accept Numeric Data whereas SVC, Decision Tree, Random forest and Gaussian Naive Bayes accept both numeric and categorical Data.

B. Ensemble Model

The idea is to train several models on the same dataset and combine their predictions to obtain a final prediction that is more accurate and robust than any individual model. This technique permits higher predictive performance. ensemble classifier is one of the proven models to boost prediction confidence. [14]

Voting, Bagging and Boosting are types of ensemble models whose Average Accuracy, precision and recall is calculated and compared.

1. Voting classifier

It operates by integrating predictions from two or more distinct models and choosing the prediction with the highest number of votes.

To get an accurate prediction, different, non-identical algorithms are combined in the voting method. In the proposed system Logistic Regression with maximum iteration of 5000 with 10 cross validation, XGB and Linear Discriminant Analysis (LDA) algorithms are combined to build the voting classifier.

2. Bagging

In Bagging, the combination of weak classifiers is used to improve the classification accuracy. [13] The idea of bagging is to fit several independent models and "average" their predictions in order to obtain a model with a lower variance. [5]

Bagging can be applied to any type of machine learning algorithm, but it is particularly effective with decision trees, which are known to have high variance. In a Proposed system multiple decision trees are combined and

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:12/December-2023 Impact Factor- 7.868 www.irjmets.com

trained on different subsets of the training data to reduce the variance. with 200 base estimators, max_sample is 0.8 and oob_score as True.

3. Boosting

It is a process of building a strong model by combining multiple weak models iteratively. Instead of picking classifiers at random, it transforms and combines a number of weak learners to create a stronger classifier. [15]

Ada boost and Gradient boost, boosting methods are used in given system with 100 maximum number of estimators

IV. DATA PRE-PROCESSING AND EXPERIMENTAL SETUP

Preprocessing processes may enhance the performance of machine learning algorithms by applying specific processing activities to ensure that the input data is prepared in the best possible way. The initial data set may have abnormalities or inaccurate values that impair the quality of the dataset. [12] Data pre-processing involves various steps such as Data cleaning, Encoding Categorical values, feature selection, Data normalization or standardization and Training and Testing.

- Since machine learning depends entirely on data, A dataset is the first thing required to train a Machine learning model.
- It is essential to import some predefined libraries in order to preprocess the data.
- The original dataset may have had imbalanced data, containing missing and null values. They need to be cleaned before use.

A. Handling missing data

In a given dataset, feature name _c39 contains mostly missing data; to handle this anomaly, the _c39 column is completely deleted from the dataset. Deletion of the entire column is only suitable if the population of missing data in that column is greater than 80%. If the missing data population is less, then Mean/Mode/Median technique is used to replace missing values with either mean, mode or median. However, this technique can lead to biased results, as it assumes that the missing values have the same distribution as the available values.

B. Encoding Categorical values

Categorical values are those that represent discrete categories or groups, such as insured sex, occupation and education level. Machine learning models require numerical inputs, so categorical values must be converted into numerical values.

1. Label Encoding: This method involves assigning a numerical label to each category in a categorical variable."fraud_reported" column label encoded and 1 is assigned to YES and 0 assigned to No. However, this method may create a problem of the model thinking that there is an order to the categories when in fact there is not.
2. One-Hot Encoding: to avoid issues of label encoding One-Hot encoding is used, With this technique, a categorical variable's categories are turned into binary vectors. features such as policy state, insured sex, etc are encoded through this method.

C. Feature selection

Initially there are 39 features present in the dataset, the task is to remove features that are irrelevant, redundant, and contain major missing values.

3 features Selection techniques are used, namely Correlation matrix and variance threshold and chi-square test.

1. Correlation matrix

A correlation matrix in machine learning is a matrix that shows the correlation between two or more variables.

In this matrix, each cell represents the correlation between two variables, with values ranging from -1 to 1.

Correlation matrix only works on numeric data only. Original dataset contains 39 features, out of 35 features selected in the correlation matrix.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:12/December-2023 Impact Factor- 7.868 www.irjmets.com

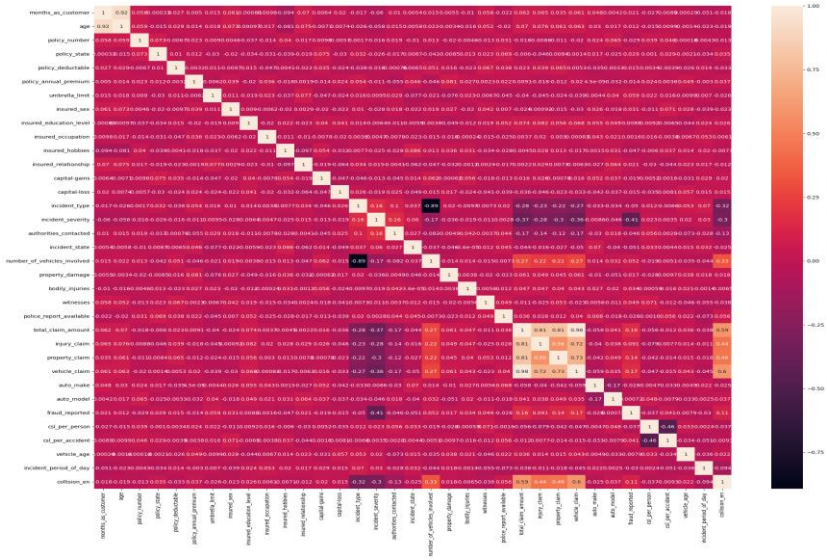

Fig 1: Correlation matrix

2. Variance Threshold

To remove the features with low variance, Variance Threshold technique used in machine learning to remove features with low variance. It is only applicable on numeric data, 35 features selected with threshold value 1, It removes all features whose variance doesn't meet the threshold. it removes all zero-variance features by default, i.e., features with the identical value throughout samples. Out of 35, nine features showed false, meaning they did not meet the threshold. all other features with true value are selected to train the model.

array([True, True, True, False, True, True, True, False, True, True, True, True, True, True, True, False, True, True, True, False, False, True, False, True, True, True, True, True, True, False, False, False, True, True, True])

3. Chi-square test

The chi-square test is performed to determine if there is a significant association between two categorical variables.

Features with a high chi-square statistic and a low p-values are considered more important and are therefore retained, while those with a low statistic and high p-value are removed. Claim amount, vehicle claim, capital gains, etc are some features with high chi value and auto model, insured occupation and education level are some features with p-value > 0.5 hence they are removed from training dataset.

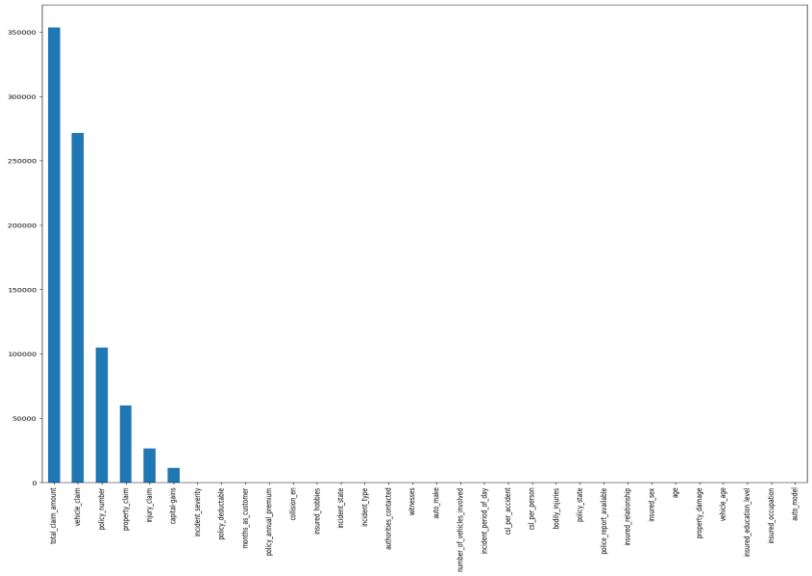

Fig 2: chi-value Graph

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:12/December-2023 Impact Factor- 7.868 www.irjmets.com

if p-value > 0.5, implies that column has less importance in the prediction process.

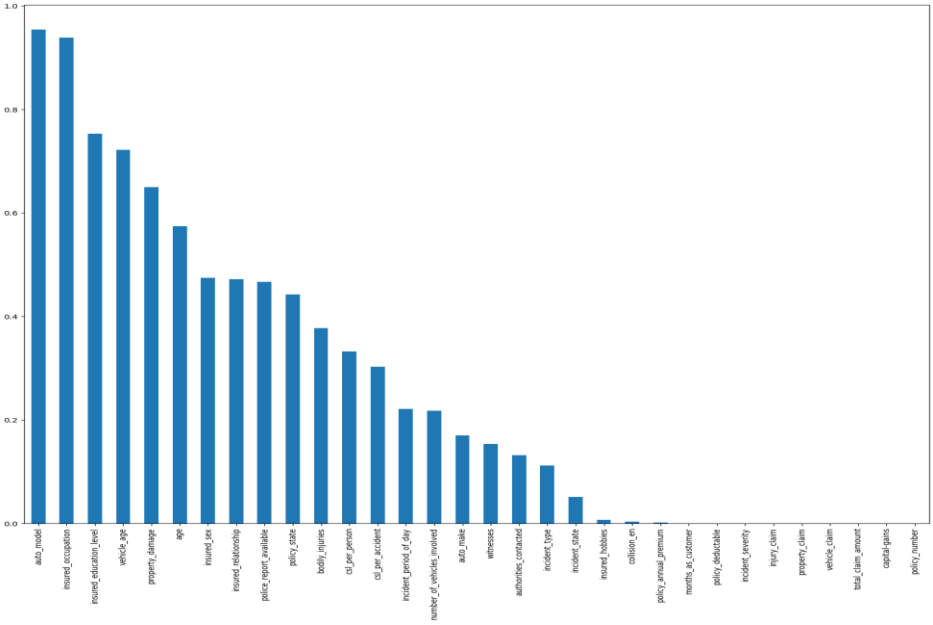

Fig 3: p-value Graph

D. Data standardization

Data standardization involves transforming the data in a way that puts all features or variables on a similar scale, which helps the algorithm perform better and converge faster during training. Standardization is especially useful when dealing with features that have vastly different scales, such as one feature measured in millions and another in single digits.

For each feature, the mean is subtracted from the data, and the result is divided by the standard deviation.

```
In [60]: from sklearn.preprocessing import StandardScaler
         scaler = StandardScaler(with_mean=False)
         X_train_scaled = scaler.fit_transform(X_train)
         X_test_scaled = scaler.transform(X_test)

In [61]: X_train_scaled

Out[61]: array([[0.        , 0.        , 2.08816082, ..., 0.90375361, 1.84093195,
                 2.63898796],
                [0.        , 0.        , 2.08816082, ..., 2.10117545, 1.60502305,
                 2.96886146],
                [0.        , 2.16950399, 0.        , ..., 2.69678424, 2.74665362,
                 1.64936748],
                ...,
                [2.11480423, 0.        , 0.        , ..., 1.26980485, 2.58657258,
                 4.61822894],
                [0.        , 2.16950399, 0.        , ..., 0.10960856, 0.22327092,
                 3.62860845],
                [0.        , 2.16950399, 0.        , ..., 3.29239303, 2.51495738,
                 1.97924097]])
```

Fig 4: Data Standardization

E. Training and testing Data

After feature selection techniques, Dataset used for Auto insurance fraud detection contains 1000 records and 17 columns. 80% of given data are used to train the model to learn the underlying patterns and relationships between the input features and output variables for training and 20% of data used for testing. The testing dataset should be representative of the population and have the same distribution of features as the training dataset. If the model performs poorly on the testing data, it may be overfitting to the training data.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:12/December-2023 Impact Factor- 7.868 www.irjmets.com

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, train_size=0.8, random_state=7)
print('length of X_train and X_test: ', len(X_train), len(X_test))
print('length of y_train and y_test: ', len(y_train), len(y_test))
```

```
length of X_train and X_test:  800 200
length of y_train and y_test:  800 200
```

Fig 5: Train Test Split

V. RESULT AND ANALYSIS

The average accuracy of different classification Algorithms are compared. K-fold Cross-validation is applied to all models, which divides the dataset into K folds. It is used to evaluate how well the model performs when presented with new data. K indicates how many groups are created from the data sample. 10-fold Cross validation performed on a given dataset.

For each classification model, 10-fold cross validation is applied and results are stored into the array, which is then used to find Average Accuracy of that given model.

Table 1: Model comparison

Model	Average Accuracy (%)
Logistic Regression	82
XGB	81.7
KNN	72.3
SVC	77.5
Decision Tree	80
Random Forest	78.5
LDA	83.2
Gaussian Naive Bayes	61.8

The distribution of accuracy values of each model is visualized using a box plot, where the average accuracy values of each model are compared.

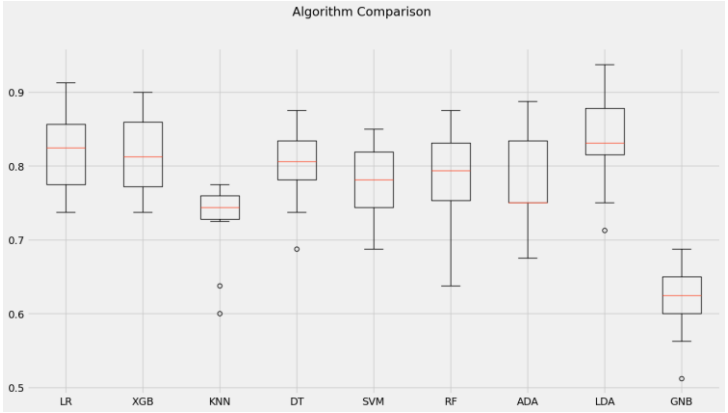

Fig 6: Algorithms Average Accuracy distribution

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:12/December-2023 Impact Factor- 7.868 www.irjmets.com

To increase the accuracy, prediction and overall performance of Model Voting, Bagging and Boosting ensemble methods are used.

- In Voting LDA, Logistic Regression and XGB, these machine learning algorithms are combined to increase average accuracy of the model.
- In Bagging multiple Decision trees are used to improve accuracy of the model.
- for Boosting, AdaBoost and Gradient Boost, These two Algorithms are used.

Table 2: Ensemble models Accuracy Calculation

Model	Average Accuracy
Voting	83.5
Bagging	84.1
Adaboost	79.5S
Gradient boost	82

Confusion matrix of all above Ensemble techniques are visualized and compared.

Confusion matrix is used to determine how well each method performs. Information on false positives (fp), false negatives (fn), true positives (tp), and true negatives (tn) is provided by the confusion matrix. [6]

Fig 7: Voting confusion matrix

Fig 8: Bagging confusion matrix

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:12/December-2023 Impact Factor- 7.868 www.irjmets.com

Fig 9: Adaboost confusion matrix

Fig 10: Gradient Boost confusion matrix

Sensitivity, specificity, precision, accuracy, and error rate are computed using the confusion matrix.

Table 3: Sensitivity and Specificity

Model	Sensitivity (%)	Specificity (%)
Voting	86.5	66.7
Bagging	83.8	80.3
Adaboost	90.6	58.8
Gradient boost	84.5	74.5

Table 4: Precision, Accuracy and Error rate

Model	Precision (%	Accuracy (%	Error rate (%
Voting	88.3	81.5	18.5
Bagging	92.5	83	17
Adaboost	86.5	82.5	17.5

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:12/December-2023 Impact Factor- 7.868 www.irjmets.com

Gradient boost	90.6	82	18

From Above table and by considering all the values, conclusion is made that Bagging shows Highest precision (92.5%) with Accuracy of 83% and shows minimum error rate (i.e 17%) amongst all other models.

VI. CONCLUSION

The "Fraud Detection in Auto Insurance Using Ensemble Learning Methods" study aimed to develop an effective approach to detect fraudulent claims in the auto insurance industry using ensemble learning techniques. Through the use of various ensemble learning algorithms and machine learning algorithms, This study evaluate the performance of four popular ensemble learning techniques, including the Bagging, Gradient Boost, AdaBoost, and Voting Classifier, and compared their results against a baseline models, the study was able to analyze a dataset of insurance claims and identify fraudulent claims with a high degree of accuracy.

Various comparisons between ensemble classifiers have been done in this study. Out of all, Bagging ensemble model outperforms all others in terms of accuracy, precision, and error rate. The results of this study can be used to develop more accurate fraud detection systems, ultimately reducing the financial burden placed on insurance companies and policyholders.

VII. REFERENCES

[1] A. Urunkar, A. Khot, R. Bhat and N. Mudegol, "Fraud Detection and Analysis for Insurance Claim using Machine Learning," 2022 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES), THIRUVANANTHAPURAM, India, 2022, pp. 406-411, doi: 10.1109/SPICES52834.2022.9774071.

[2] D. K. Patel and S. Subudhi, "Application of Extreme Learning Machine in Detecting Auto Insurance Fraud," 2019 International Conference on Applied Machine Learning (ICAML), Bhubaneswar, India, 2019, pp. 78-81, doi: 10.1109/ICAML48257.2019.00023.

[3] T. Badriyah, L. Rahmaniah and I. Syarif, "Nearest Neighbour and Statistics Method based for Detecting Fraud in Auto Insurance," 2018 International Conference on Applied Engineering (ICAE), Batam, Indonesia, 2018, pp. 1-5, doi: 10.1109/INCAE.2018.8579155.

[4] R. Roy and K. T. George, "Detecting insurance claims fraud using machine learning techniques," 2017 International Conference on Circuit ,Power and Computing Technologies (ICCPCT), Kollam, India, 2017, pp. 1-6, doi: 10.1109/ICCPCT.2017.8074258.

[5] C. Yan and Y. Li, "The Identification Algorithm and Model Construction of Automobile Insurance Fraud Based on Data Mining," 2015 Fifth International Conference on Instrumentation and Measurement, Computer, Communication and Control (IMCCC), Qinhuangdao, China, 2015, pp. 1922-1928, doi: 10.1109/IMCCC.2015.408.

[6] S. Padhi and S. Panigrahi, "Decision Templates based Ensemble Classifiers for Automobile Insurance Fraud Detection," 2019 Global Conference for Advancement in Technology (GCAT), Bangalore, India, 2019, pp. 1-5, doi: 10.1109/GCAT47503.2019.8978332.

[7] N. Dhieb, H. Ghazzai, H. Besbes and Y. Massoud, "Extreme Gradient Boosting Machine Learning Algorithm For Safe Auto Insurance Operations," 2019 IEEE International Conference on Vehicular Electronics and Safety (ICVES), Cairo, Egypt, 2019, pp. 1-5, doi: 10.1109/ICVES.2019.8906396.

[8] Supraja, K., & Saritha, S. J. (2017). Robust fuzzy rule based technique to detect frauds in vehicle insurance. 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), doi:10.1109/icecds.2017.8390160.

[9] I. M. Nur Prasasti, A. Dhini and E. Laoh, "Automobile Insurance Fraud Detection using Supervised Classifiers," 2020 International Workshop on Big Data and Information Security (IWBIS), Depok, Indonesia, 2020, pp. 47-52, doi: 10.1109/IWBIS50925.2020.9255426.

[10] B. Itri, Y. Mohamed, Q. Mohammed and B. Omar, "Performance comparative study of machine learning algorithms for automobile insurance fraud detection," 2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS), Marrakech, Morocco, 2019, pp. 1-4, doi: 10.1109/ICDS47004.2019.8942277.

[11] D. P. Gaikwad and R. C. Thool, "Intrusion Detection System Using Bagging Ensemble Method of Machine Learning," 2015 International Conference on Computing Communication Control and Automation, Pune, India, 2015, pp. 291-295, doi: 10.1109/ICCUBEA.2015.61.

[12] J. Dutta, Y. W. Kim and D. Dominic, "Comparison of Gradient Boosting and Extreme Boosting Ensemble Methods for Webpage Classification," 2020 Fifth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), Bangalore, India, 2020, pp. 77-82, doi: 10.1109/ICRCICN50933.2020.9296176.

[13] V. F. Adegoke, D. Chen, S. Banissi and E. Banissi, "Predictive Ensemble Modelling: Experimental Comparison of Boosting Implementation Methods," 2017 European Modelling Symposium (EMS), Manchester, UK, 2017, pp. 11-16, doi: 10.1109/EMS.2017.13.