Vol. 6, No. 2, August 2024, pp. 119~133, p-ISSN 2685-2381, e-ISSN 2715-2626
http://dx.doi.org/10.28989/avitec.v6i2.2340

Comparative Study of Machine Learning Techniques for Insurance Fraud Detection

Navin Duwadi[1]*, Anita Sharma[2]
[1]Information and Communication Technology, Nepal Open University, Nepal
[2]Department of Management, Tribhuvan University, Nepal

Article Info

Article history:
Submitted July 11, 2024
Accepted July 29, 2024
Published August 6, 2024

Keywords:
Machine learning,
support vector machine,
random forest,
logistic regression

ABSTRACT

Insurance fraud has been a constant presence in the realm of insurance. However, as strategies and methods for committing insurance fraud have evolved, the frequency and volume of such fraudulent activities have also increased. An example of this is vehicle insurance fraud, which involves collaborating to fabricate false or exaggerated claims related to property damage or personal injuries resulting from an accident. Machine learning techniques seems to be more beneficial and great way to address the fraud in the insurance industry. This paper comprehensively examines existing research through a systematic literature review. This review aims to identify previously attempted approaches and evaluate which machine learning algorithm is best suited for this specific problem. This paper proposes a methodology for identifying fraudulent insurance claims. This approach can significantly improve efficiency and cost savings for insurance companies in handling such cases. The most popular traditional machine learning algorithms used to identify insurance fraud in the auto industry were found to be support vector machine, logistic regression, and random forest.

Corresponding Author:
Navin Duwadi,
Faculty of Science, Health and Technology, Nepal Open University
Manbhawan, Lalitpur, Nepal
Email: 76251002@nou.edu.np

1. INTRODUCTION

The financial fraud has created many problems and has much more effects on financial sector and daily life. Fraud seems to have the ability to undermine consumer confidence, damage economic stability, and increase cost of living. Traditional techniques for detecting fraud have relied on manual procedures like audits, which are unreliable and inefficient because to the complexity of the problem [1]. In traditional environment, an insurance agent would be able to examine each case and determine whether it is true. This technique, on the other hand, is not only time-consuming but also expensive. It is just difficult to find and afford the experienced personnel needed to review each of the hundreds of claims filed each day. These days, there are a lot of machine learning techniques has been introduced and different approach has been used to detect insurance fraud in insurance industry. Different algorithms have been used to detect insurance fraud and all of them has different performance based on accuracy, specificity and sensitivity.

Insurance company come into fraud on a regular basis in a part of businesses. It can be found in a variety of forms and models derived from previous fraud or scams, such as tax fraud, insurance fraud, to name a few, in which a large group of people gather together to commit such crimes [2]. These organized groups are found common in the automotive insurance industry. Scammers cause road accidents and file fake insurance claims in order to earn illegally from their general or auto insurance.

Fraud detection is a procedure used to track down and prohibit criminals from using deceptive means to obtain money or property. It consists of a series of steps taken to recognize and prevent offenders' attempts to steal money or property. In law enforcement agencies as well as the banking, insurance, healthcare, government, and public sectors, fraud detection is a frequent practice [3]. Having an effective system for detecting and preventing fraud within a company will result in higher levels of customer satisfaction. This will also eventually lead to the reduced loss and adjustment expenses. There are now numerous methods for identifying fraud claims. The most popular technique is data analysis using specific instructions. As a result, they require in-depth investigations that take a lot of time and deal with various fields of expertise to be able to resolve this issue [4].

There are many types of frauds such as cheque forgeries, credit card fraud, insurance fraud, and telecommunications fraud. In the auto and travel industries, insurance fraud is frequent. Three categories of offenders are involved in fraud detection: i) Criminals, ii) organized criminals who commit significant fraud, and iii) criminals who commit fraud while facing financial hardship [5] [6]. The cost of each suspected occurrence is typically larger than the cost of the fraud, making soft fraud the tough one to reduce.

Multiple researchers used various machine learning approaches to study insurance claims. They discovered that machine learning has the ability to raise security levels in the insurance sector [4]. The foundation of machine learning technology is the idea that by providing an algorithm with sufficiently vast datasets, it would be able to identify patterns within those datasets. Training the algorithm is the process of teaching the algorithm new patterns from the dataset. The outcome of unknowable cases is then predicted using the learned patterns. Supervised, unsupervised, and reinforcement learning are the three main subfields of machine learning [7].

Machine Learning is a field related to the Artificial Intelligence (AI). Artificial Intelligence's goal is to construct a computerized system that can perform complicated analysis and not only replace but also improve on human input [8]. Machine Learning involves artificial intelligence to provide systems the ability to learn and grow from their experiences without the need for additional programming. Systems analyze massive, labeled data sets to accomplish this. Menial duties may be taken over by AI, allowing human agents to focus on analysis that is more complicated. Classification algorithms have been shown to be particularly successful in detecting fraud and may thus be used to categorize crime data. To evaluate classification models, the distributed data-mining model employs a realistic cost model [9].

Algorithms can learn from the dataset thanks to machine learning, which is engaged in the design and development of algorithms. It focuses on analytical and statistical techniques for obtaining knowledge from data. Consequently, it requires statistics and data mining [10]. The accuracy of the method increased with the size of the dataset employed. According to Caruana & Grech [11], Real-time anomaly detection is possible with machine learning. Finding objects that are unique from others is the main goal of anomaly detection. When the labels of fraud and legitimate are present in the data set, one of the techniques that is frequently employed in anomaly detection is a supervised classifier. By creating a model from a labeled dataset and then labeling new data, it can indirectly solve complex issues [12]. The main objective of this paper is to use the multiple Machine learning algorithms to predict the occurrence of the fraud. Specifically, to review the recent literatures and find out the popular machine learning techniques used for fraud detection. Moreover, implement and compare the performance of most popular machine learning algorithms for insurance fraud detection.

In machine learning, a common problem is an unbalanced dataset. When dealing with unequal data sets, the general standard machine learning method is biased in favor of majority groups. Because of this, high data accuracy for majority of class and low for minority groups. The data level strategy, which involves undersampling and oversampling, can be employed by altering the data itself to get a balanced dataset. By removing a few instances of the majority classes at random, the undersampling technique can be created. While oversampling, raising the minority classes can result in a data set that is balanced [13].

A systematic literature review was conducted to comprehensively compare the performance of multiple machine learning algorithms in detecting insurance fraud across healthcare, automobile, and finance sectors. The analysis encompassed a detailed examination of datasets, including their attributes and sample sizes, as well as the validation techniques employed in previous studies. By evaluating the accuracy, precision, recall, and false positive rates of popular algorithms, this research aimed to identify the most popular and effective models for each insurance sector, considering the unique characteristics and challenges inherent to each domain. By synthesizing findings from these studies, the research identified key trends, gaps, and opportunities for future research in insurance fraud detection.

2. LITERATURE REVIEW

The literature review is divided into two different topics. The first topic is the review different algorithm for the insurance fraud detection. The second topic deals with the insurance fraud detection in multiple domains with special interest in automobile insurance using different algorithms and compare those algorithms to find out that perform the better result.

Insurance serves as a financial safeguard for both policyholders and their families in case of unfortunate events like fires, thefts, legal disputes, or car accidents. An insurance policy constitutes a legally binding agreement between an individual seeking coverage and an insurance company [14]. Additionally, insurance pays a chosen beneficiary, in accordance with the terms of the policy when insurer experience a loss that is covered by the policy and file a claim.

Insurance is an essential financial measure. It will help people to minimize anxiety and speed up the recovery if the financial aid is given after a disaster or accident [15]. Life insurance can keep a family from having to move or allow their children to go to the college. In case of auto insurance, it might indicate that there is extra cash on hand to help pay for repairs or a new vehicle after an accident. In addition, health insurance can provide assistance in case of serious health issues. After a negative event throws one’s life off course, insurance

can help, at least in part. Furthermore, economic theorists have been interested in the insurance business since it is one of the key drivers of financial development in both developed and developing nations. This is because the insurance influence coefficient is a crucial indicator of economic development [16].

Gupta et al. [17] presents fraud prediction methods to predict fraudulent behaviors from the data. Both supervised and unsupervised learning are connected to data mining techniques. Criminal offenders, organized criminal organizations, and soft fraud are all involved in fraud detection. Comparing the Bayesian Belief Networks algorithm to decision trees and back propagation, it is noticeably superior. It is more akin to an if-then structure in the decision tree algorithm. Information is separated out using algorithms, which are also used to create descriptive categorization rules that may be applied to new situations. The confusion matrix and the ROC graph are two different ways to categorize the model's performance. To identify how well a given data set's classifications performed, the confusion matrix and ROC graphs are utilized.

According to Pesantez-Narvaez et al. [18], Logistic Regression and XGBoost are examined based on their predictability that is, how effectively or accurately these algorithms forecast the claim amount. Telematics is the type of data used for prediction in this study that includes additional vehicle-related data. Without any model adjustment, XGBoost outperformed Logistic Regression in terms of predictive performance for training samples but performed much worse for testing samples. The prediction abilities of XGBoost and Logistic Regression were equalized by regularizing overfitting. This demonstrates that XGBoost requires more fine-tuning to produce a higher prediction.

Kowshalya & Nandhini [19] presents a fictitious dataset that was utilized in this case study to determine whether a claim was fraudulent. Due to the difficulty of obtaining the insurance dataset, a fictitious insurance dataset based on case studies of insurance fraud was developed. A data mining approach is used to estimate the premium percentage and predict fraudulent claims. Two categories—one for vehicle theft claim and the other for accident claim—are included in the dataset that was created. For predicting false claims, the Naive Byes, J48, and Random Forest classification algorithms are used. Data preprocessing was carried out in order to improve the model's accuracy.

Badriyah et al. [20] offers a model that uses statistics and the closest neighbor method to detect fraud. For identifying fraudulent claims, the nearest neighbor method uses two approaches, density-based and distance-based. The final model is contrasted with additional models created using the same dataset. They also used feature selection to boost accuracy and concluded that feature-selected datasets produced higher accuracy. If the dataset is large enough, this model cannot be applied because it is laborious to choose the attributes from it.

Subudhi & Panigrahi [21] purposed a model to identify automobile insurance fraud. Comparison study was conducted by comparing a number of classifiers employing fuzzy C-Means clustering based on genetic algorithms. The findings revealed that Support Vector Machine (SVM) outperformed the other techniques.

2.1 Logistic regression

Logistic regression is a valuable tool for insurance companies in combating fraud. It's a statistical technique that analyzes past claim data, including characteristics like policyholder age, type of claim, and previous accident history. By incorporating both legitimate and fraudulent claims, the model learns to identify patterns that differentiate them [22]. The logistic regression generates a probability score for each new claim, indicating the likelihood of fraud. This allows insurers to prioritize investigations, focusing resources on claims with the highest predicted fraud risk. While not a foolproof solution, logistic regression offers a data-driven approach to flag suspicious claims, improving efficiency and deterring fraudulent activity [23].

2.2 Support Vector Machines

A support vector machine is also known as supervised machine learning algorithm widely used in classification. An initial phase of the SVM algorithm is training phase. SVM can forecast which class the new incoming data will fall into once the training phase is complete [24]. SVM is an excellent prediction tool for a variety of learning issues, including handwritten digit identification, web page classification, and face detection. This approach can detect fraudulent conduct in the middle of a transaction. One of the most used supervised learning methods is the SVM algorithm. According to Kumar et al. [25], SVM is a binary classification method for locating a hyperplane between the two classes; this hyperplane is a decision boundary that is designed to stay as far away from the nearby training cases as feasible. Support vectors are the datasets that are closest to the decision border. In the SVM algorithm decision boundary, direction and location are set by support vectors. Due to the algorithm's dependence on the datasets that are closest to the boundary, it is less reliant on the datasets that are farther from the boundary, allowing it to perform well even when the dataset contains outliers. Additionally, because of the algorithm's excellent performance even in the absence of some of the datasets, this property lowers the model variance [26]. SVMs are capable of extracting even minute patterns from large datasets.

2.3 Random Forest

Random forest is also known as an ensemble classifier that includes multiple decision tree models. According to Li and Yun [27], by combining many single learners, ensemble classifiers can greatly increase performance and decrease mistakes. The majority of votes are used to determine the outcome. The advantages

of random forest include preventing overfitting, lowering variance, and enhancing model accuracy. The random forest technique has two components: full splitting and random sampling. The first component includes the random feature subspace and the bootstrap. Bootstrap randomly selects samples from a replacement dataset to ensure that certain samples appear more than once. The chosen features are no larger than the dataset's original features because they were chosen at random from the subspace of random features without any substitution [28]. Each tree is built without being trimmed throughout the splitting process, making each one unique.

2.4 Insurance fraud in multiple domain

2.4.1 Health Insurance Fraud

The majority of healthcare insurance fraud involves, allowing another person to acquire health care services using their name and insurance information, or utilizing benefits to pay for medications that were not written by their own physician. In this type of fraud, a health insurance company is given incorrect or misleading information in an effort to get them to pay illegitimate benefits to the policyholder, another client, or the business providing the services. This kind of activities can be done by the individual or the health service providers [29].

2.4.2 Financial insurance fraud

Generally, financial institution refers to the any bank, credit institution, undertaking for collective investment in securities, investment firm, investment advisory or asset Management Company. Fraud against banks or other financial institutions is frequently described as a white-collar crime. Fraud involving bank accounts and credit cards is one example of unlawful action. Use of another person's card, card cloning, gaining control of or sending instructions pertaining to another person's bank or card account, and use of another person's check are all examples of bankcard fraud, which involves the use of payment cards and bank accounts. Because of the advent of internet banking, frauds have gotten more sophisticated and difficult to detect in recent years. As a result, online banking frauds are rife. The bank should make every effort to return any stolen funds as quickly as feasible, but if it is suspicious that owner committing fraud or have been negligent, they have to investigate further and reclaim from the owner [30].

2.4.3 Automobile insurance fraud

A legally binding agreement between an insurance company and the owner of a vehicle to provide financial assistance in case of stolen or wrecked is known as an automotive insurance file. Automobile insurance fraud describes a scenario in which the insured obtains a financial gain by providing false documents to the firm and claiming that the vehicle was damaged in prearranged (fake) accidents or by making monetary claims for previous losses [31]. This form of fraud involves individuals trying to mislead an insurance company regarding a claim related to their own or a business-owned vehicle. This deception might encompass spreading false details or furnishing fake evidence to support the claim. Instances of auto insurance fraud frequently encompass staged collisions, contrived injuries, reporting of stolen vehicles, assertions that an accident happened after obtaining a policy or coverage, claims for pre-existing damage, and scenarios where individuals hide the fact that an uninsured individual was driving the vehicle during the accident [32].

Fraudulent activities within the insurance sector, especially concerning insurance claims, have had a lasting impact. Establishing a robust strategy for fraud management is of most importance. This is because while many individuals pay their insurance premiums diligently, there are those who engage in deceitful practices to obtain reimbursements. In the context of auto insurance, fraudulent behavior can be classified into two main categories: hard insurance fraud and soft insurance fraud. Hard insurance fraud involves intentionally manipulating an accident scenario, often with the aim of causing damage or loss in order to file a fraudulent claim. On the other hand, soft insurance fraud transpires when a person submits a valid insurance claim but fabricates certain aspects of it to exaggerate the loss or damage incurred [33].

There are various examples of auto insurance fraud such as providing a false address is one of the major issues that eventually leads to the insurance fraud. In this scenario, people who live in the area where high auto theft is common so they falsified the address where their vehicle is parked and later file the compensation for that. In addition, leaving a car alone or wrecking it after reporting stolen and if it is not recovered, the insurance company declares the automobile a total loss and pays out the car's real monetary worth. Reporting an automobile as stolen after it has been sold, concealed, or otherwise disposed is committing a fraud against the auto insurance company. Furthermore, other types of fraud include filing the multiple claims for one accident [34].

3. DATA AND METHODS

In order to answer the questions based on insurance fraud, a literature review is conducted to examine the current state of research on financial fraud in various fields. The literature review approach is also appropriate for compiling and examining all studies that were specifically focused on a given research subject. In order to offer a review with high-quality evidence, it is used to find and combine information that focuses on specific concerns, as well as to examine the reviewers' judgments and conclusions.

This section is divided into two parts. First part deals with the systematic literature review method used to conduct the research and second part describes the comparative study on fraud detection in auto insurance industry. In addition, the method adopted for building a fraud detection model has been discussed. Multiple classification techniques such as Random Forest, logistic regression, and Support Vector Machine has been applied to chosen dataset. To discover which model is performing better, multiple validation techniques are used with training and testing the models [35].

3.1 Data

3.1.1 Peer reviewed literature

Multiple machine learning techniques is used to detect fraud in the industry. Some algorithms seem to be more effective in some domain and other are beneficial for other domains. In this paper, three domains have been considered for analyzing fraud detection such as healthcare, bank and finance, and automobile. Real-world problem solving has always benefited from artificial intelligence and machine learning. It is now widely utilized in all industries, including banking, insurance, and healthcare and so on. All of the reviewing tasks were previously completed manually. However, as computer technology has improved and statistical modeling has advanced, machine learning has become more widely accepted across all industries [36]. Furthermore, multiple machine learning and data mining techniques has been introduced and implemented. It is difficult to find which algorithm is better on performing fraud detection task. Therefore, in this paper review has been done to find out which algorithms performs better to detect fraud in multiple industries.

Different machine learning algorithms has been analyzed and evaluated based on the performance on fraud detection. These algorithms include random forest, support vector machine, decision tree, neural network, logistic regression, gradient boosting, k means, k- nearest neighbor and so on. Multiple research articles have found implementing multiple algorithms for different industry.

Validation in machine learning refers to the process of evaluating a trained model against a testing dataset. The training set is derived from the testing dataset, which is a separate portion of a similar dataset. The main goal of using the testing dataset is to evaluate a prepared model's capacity for prediction. Model training is followed by model validation. Model validation seeks to identify a perfect model with the best execution in addition to model training [37]. Different validation techniques such as cross validation, AUCROC, confusion matrix, anova test and so on have been found from the review articles to test and validate the model.

3.1.2 Implementation data

In this paper, a dataset that has the details of the insurance policy along with the customer details has taken from the internet. It also has the details of the accident based on which the claims have been made. Dataset has been taken to implement with the most popular machine learning algorithms. The dataset that contains 11565 rows and 34 columns. The column names are described by policy number, reference number, policy type, vehicle type, police report, etc. The huge sample size of this data collection is obviously an advantage. Any organization wishing to use data science must have the flexibility to work with what is already in place. The dataset is taken from the Kaggle Data Repository and this data set is used in this work to detect insurance claim fraud. All the features are used as input for mathematical computations of the models. The following attribute is an identification for reported fraud, with values of zero (0) representing no fraud and one (1) representing a fraud. The attributes of the choosen dataset are mentioned in Table 1 below.

Table 1. Attributes of dataset

Number	Attributes	Number	Attributes
1	Month	18	Rep Number
2	Week of Month	19	Deductible
3	Day of Week	20	Driver Rating
4	Make	21	Days Policy Accident
5	Accident Area	22	Days Policy Claim
6	Day of Week Claimed	23	Past Number of Claims
7	Month Claimed	24	Age Of Vehicle
8	Week Of Month Claimed	25	Age Of Policy Holder
9	Sex	26	Police Report Filed
10	Marital Status	27	Witness Present
11	Age	28	Agent Type
12	Fault	29	Number of Supplements
13	Policy Type	30	Address Change Claim
14	Vehicle Category	31	Number of Cars
15	Vehicle Price	32	Year
16	Fraud Found_ P	33	Base Policy
17	Policy Number	34	Claim Size

3.1.3 Data analysis

The dataset used for this article to implement fraud detection contains claims for a New Zealand based auto insurance company. The dataset contains 11565 individual claims. Fraud discovered in the dataset is the most significant variable of interest. If a certain claim is found to be fraudulent, this variable is labeled 1; otherwise, it is labeled 0. The dataset's 34 separate attributes, which are displayed as columns, are used to explain each claim. The insured party, the insured party's policy, the incident's description, and the features of the car involved in the occurrence can all be considered as separate categories of attributes. The data includes both categorical and numerical values. Some examples of these features are the insured's age, the insured's premiums and quantities, the insured's occupation, the quantity of vehicles involved in the accident, and the manufacturer of the car for which the claim was filed. There are both numerical and category variables in the data.

3.2 Methods

This section of the research follows two different methods to conduct the research. The first one represents the methods of literature review and second one is the method of implementation.

3.2.1 Methods of literature review

The main objective of this paper is to review and evaluate previous research on machine learning-based fraud detection techniques. Therefore, this article does not limit the search to any one field of knowledge. Multiple keywords such as fraud detection, fraud prediction, machine learning, insurance, healthcare, credit card, automobile, bank and finance have been used to search papers over the internet to find the scientific publication related to the fraud detection and process to identify the fraud in multiple industry. This paper has especially focused on the fraud detection using machine learning techniques and list out the paper with high sensitivity, f1-score, and accuracy. The literature review of the insurance fraud detection includes the multiple stages planning, conducting and reporting the review.

a. Planning the review

The initial phase of the systematic literature review involved outlining the research objectives and developing a structured review protocol. To identify relevant studies, a comprehensive search was conducted across major academic databases. These databases were selected due to their extensive coverage of scholarly articles and their high probability of containing research pertinent to the study's focus.

b. Conducting review

Following the development of the review protocol, the next phase involves conducting the review itself. This stage centers on defining the specific research questions that will guide the investigation. Once established, the process of searching for relevant studies, extracting data, and synthesizing the findings is undertaken.

i. Research question formulation.

The initial phase of a Systematic Literature Review (SLR) involves clearly defining the research questions. These questions serve as a roadmap for the entire review, guiding the search for relevant literature and the subsequent analysis of findings. There are multiple research questions has been identified to conduct the literature review. Such as,

1. What types of financial fraud are commonly addressed using machine learning techniques?
2. What machine learning methods are primarily employed for financial fraud detection?
3. Which performance metrics are commonly used to evaluate financial fraud detection models?
4. What are the current knowledge gaps, emerging trends, and potential future research directions in financial fraud detection?

The main objective of the review is to answer the research questions that inludes fraud type identification, machine learning algorithms categorization, performance metric analysis, and research gap and trend analysis.

ii. Search strategy and search method

Multiple search strategy has been used to search relevant papers from the internet. The popular strategies for the review techniques have been used. A search has been carried out in the following databases: IEEEXplore, ScienceDirect, and ACM Digital Library, the learning and technology library and Scopus (Elsevier) to select the most pertinent articles for the subject matter covered in this study. These databases have been selected because they provide the full-text publications and conference proceedings that are most important and have the greatest impact on the machine learning and fraud detection and prediction domains as a whole.

iii. Search criteria

The search string has been designed to search relevant papers from the internet. The multiple keywords such as deep learning or machine learning, and fraud detection or fraud prevention, and healthcare or automobile or bank and finance, and insurance have been used to create the relationship and combine the keywords and search over the databases using these search terms to find the relevant papers. Some of the paper found has only

abstract and limited information and some of the papers are irrelevant to the study. So the inclusion and exclusion criteria also have been used to filter the irrelevant papers. Some of the criteria to include or exclude the paper are article focus on fraudulent transactions, short papers, and abstract only paper, published language. On the basis of these criteria papers were filtered out for review as listed in Table 2.

Table 1. Exclusion and inclusion criteria

Exclusion criteria	Inclusion criteria
Articles that do not emphasize fraudulent activity.	The articles that are conducted from 2005 to 2022
Articles only containing abstracts, short papers, and book chapters.	Articles that focus on fraud detection and applied machine learning techniques.
Articles not related to the machine learning techniques.	A peer reviewed research article.
Studies that do not publish in English language.	Studies that published in English language.

c. Reporting review

The final phase of the literature review referes to the reporting the review that includes data extraction which serves as a structured tool to collect pertinent information from the included studies. The process of gathering data for analysis involved both automated and manual methods. Information extracted from the selected papers was categorized and assigned a specific purpose. By organizing this data based on fraud types, detection techniques, evaluation metrics, and research gaps, it became possible to identify patterns and trends among the studies.

3.2.2 Methods of Implementation

There are several machine learning approaches that can be used to create models for detecting insurance fraud, including Logistic Regression, k-Nearest Neighbors, Decision Trees, Naive Bayes, support vector machines, and many others. Therefore, choosing the classifier with the highest accuracy is a time-consuming process. In this paper, we have compared the various algorithms to find out the performance of the model and most popular machine learning algorithms for insurance fraud detection. According to the study, the Random Forest algorithm, logistic regression and support vector machine has found to be performing the best at spotting insurance fraud. The loading the dataset is the first step of the implementations. Then, data pre-processing will be completed, which comprises data cleaning and data normalization. The dataset is split into two segments, one for training the model and the other for testing its performance. Ultimately, the model's task is to classify transactions as either fraudulent or not. The applied machine learning techniques encompass random forest, logistic regression, and support vector machines. The study's methodology adheres to the supervised learning process, which unfolds as follows:

a. Data Collection

The dataset of this study is represented by data on insurance claims of vehicles is taken from kaggle data repository [38]. The dataset includes 11565 claims with 33 predictor variables and one target variable represents "Fraud" and "Non Fraud", the dataset has 10880 genuine samples (94.1 %) and 685 fraud instances (5.9 %). In addition, the dataset has 17 numeric entities and 17 categorical features.

b. Data Pre-processing

Preprocessing processes can enhance the effectiveness of machine learning techniques by applying specific processing tasks to the input data in order to prepare it in the best feasible way. The Dataset may contain anomalies or inaccurate values that damage the quality of the dataset. The data set for this study is in CSV format, and once it was imported, 17 categorical and 17 numerical features were discovered. Additionally, some kinds, such as resample, require a data set with nominal features and categorical types are not supported by classification models. Thus, it is necessary to convert all the features to numeric just after importing it. In this phase of data pre-processing, missing values in the dataset has been checked. Two of the attributes age and driver rating have found with missing values and treated them accordingly.

c. Feature Selection

By removing redundant, irrelevant, or missing data, feature selection effectively solves the issue while also speeding up calculation and increasing learning accuracy. In this research, we have purposed removing the features that were found unnecessary, including the policy number, reference number, accident month, week of accident month, day of accident week, day of claim week, and week of claim month. In addition, correlation heat map was plotted to find the multi-collinearity issues on the dataset and variables with high correlation has been removed from the further fraud detection.

d. Fraud Detection

The percentage of negative instances (fraud cases) represents just 5.9% of the total when compared to their positive counterparts; hence, the result of the learning models will not seem to be optimal. This is the common issue while working on the fraud detection in any industry because fraud is always less compared to

the non-fraud data in the dataset. To predict fraudulent activity in the automobile insurance industry following algorithms are implemented and other approaches is use to correctly classify the fraud.

3.2.3 Random forest implementation:

Random forest is a supervised machine learning technique based on ensemble learning. A method of learning called "ensemble learning" involves joining many algorithms or using the same method more than once to create a more potent prediction model [25]. The term "Random Forest" refers to the approach of combining numerous algorithms of the same kind, or various decision trees, into a forest of trees. The random forest algorithm can be applied to both classification and regression tasks.

A collection of basic trees and randomly chosen predictor variables are offered for the evaluation of classification issues. The margin function produced by random forest calculates how much the average number of votes for the correct class surpasses the average vote for any other class that is included in the dependent variable [26]. However, Random Forest is created with the growth of basic trees to evaluate regression difficulties. Every tree has the ability to offer a numerical response value. In the same distribution, the predictor variables are chosen at random. The average of the predictions made by each tree is used to calculate the forecast made by the Random Forest algorithm.

The random forest contains multiple hyperparameters. Some of these configured to enhance the predictive power of the random forest model in fraud detection,

a. n_estimators: the chosen n_estimators is 100, which represents the number of trees built by the algorithm before averaging the products.
b. max_features: the chosen max_features is 3, this refers to the maximum number of features random forest uses before considering splitting a node.
c. mini_sample_leaf: this refers to the minimum number of leaves required to split an internal node. This parameter is set to default.
d. Bootstrap: method for sampling data points. This parameter is set to true.

3.2.4 Support vector machine implementation:

The Support Vector Machine (SVM) is a supervised machine learning method applicable to both classification and regression tasks. Its foundation lies in statistical learning theory. The core principle of SVMs involves identifying an optimal hyperplane for classification [28]. This hyperplane is determined by utilizing support vectors, which are the data points from each class located at the margins of separation. SVM aims to find the hyperplane that maximizes the margin between these support vectors, thereby achieving effective classification. The support vector machine contains multiple hyperparameters. Some of these are configured to enhance the predictive power of the Support vector machine model in fraud detection,

a. C: It is the regularization parameter, C, of the error term. This parameter is set to one for the fraud detection classification.
b. Kernel: It specifies the kernel type to be used in the algorithm. It can be ‘linear’, ‘poly’, ‘rbf’, ‘sigmoid’, ‘precomputed’, or a callable. The chosen kernel in this paper is ‘linear’.

3.2.5 Logistics regression implementation:

It is a simplified form of "linear regression," an effective tool for data visualization. Logistic regression is used to determine the likelihood of a disease or other health concern because of a probable cause [30]. Exposures or the relationship between independent variables (X), Using both simple and multivariate logistic regressions, the relationship between predictors and a binary dependent (target) variable (Y), also known as the outcome or response variable, is investigated. It is frequently used to predict changes in dependent variables that will have multiple classes or be binary.

The following hyperparameters are used to enhance the predictive power of the Logistic regression model in fraud detection,

a. Class weight: can be used to offset the class imbalance. The custom class weight is chosen for the implementation i.e, (0:0.5, 1:4.5).
b. Random state: to avoid the unpredicted classification behavior of the algorithm the same random state 101 is used in the model.

3.2.6 Model Evaluation

Model evaluation is the practice of assessing models using different performance measurement metrics. Model evaluation gives a clear view of the model's effectiveness and supports in choosing the optimal model for making predictions. Three algorithms: support vector machine (SVM), logistic regression (LR), and random forest (RF) are used in this study, thus after the models are built, they go through a model evaluation phase. For concrete approach, this study makes substantial use of the Scikit-learn python libraries. The paper employ a variety of performance-measurement metrics, such as the Confusion matrix, which also facilitates the determination of Accuracy, Precision, Recall, and F1 score. Recall is an effective metric for evaluating the efficacy of the models. While accuracy may be good, high-class imbalance datasets often have low recall.

Precision will also be taken into account because lower accuracy suggests that the business trying to detect fraud will spend more money on transaction screening. As an alternative approach, we have also employed the Area Under Curve (AUC) metric within the Receiver Operating Characteristic (ROC) curve. However, relying solely on accurate detection of the majority of fraudulent transactions might not be sufficient. Therefore, to comprehensively assess the model's performance, we utilize the AUC to ensure a more comprehensive evaluation.

3.2.7 Validation Approach

The evaluation of classification performance in this study involved the utilization of both the confusion matrix and ROC curve. Table 3 illustrates the visualization of the confusion matrix. From the values within this matrix, metrics such as accuracy, precision, and sensitivity can be derived. Accuracy gauges the overall effectiveness of the machine learning model in identification. Sensitivity, also known as recall or the TP Rate, is the proportion of correctly identified positive cases to the total number of actual positive instances. Precision, on the other hand, quantifies the percentage of true positive cases in relation to the total predicted positive cases.

Table 3. Confusion matrix

True positive	False positive
False negative	True negative

In our methodology, we have employed the conventional Cross Validation technique. However, to address the issue of dataset imbalance, we have undertaken measures to readjust the proportions. By providing the learning algorithm with a dataset containing an equal representation of both positive and negative examples, we aim to enhance the model's performance. To achieve this, we adopted a stratified k-fold cross-validation approach. This approach ensures that both the training and test sets maintain the same proportion of the feature of interest as observed in the original dataset.

In order to prevent the models from overfitting the training set of data, stratified cross-validation is also necessary approach to implement. To make sure that the class imbalance is preserved in the validation sets, stratified K-fold cross validation approach with 10-fold has been used in the research. Moreover, the dataset is split into two parts: training data for the first part and test data for the second. About 70% of the total amount of data used is for training, while the remaining 30% is for testing. These models are trained using training data, and then tested using test data.

4. RESULTS AND DISCUSSIONS

4.1 Analysis of review/ research works

This section presents the search results obtained from the first stage of the review process, which involves selecting the relevant studies to be considered in this literature review study. We first present the description of the reviewed studies in this SLR and then later answer each of the research questions specified in the section.

The number of articles relating to insurance fraud detection using machine learning approaches from 2005 to 2022 is shown in the Figure 1, which provides a chronological summary of the published articles used in this review. The chart illustrates how research in this field has shown a growing trend in recent years, particularly since 2015 when the number of articles published began to rise significantly. Most of the articles used in this review were released after 2015. Within the study period, 2020 experienced the highest number of articles thirteen (13) published in this area, followed by six (6) papers in 2021. In addition, number of attributes used to detect insurance fraud in multiple domains has been described in Table 4. From the chart, it can be found that health care data has used maximum number of attribute while analyzing insurance fraud and minimum number of attributes used in the automobile industry.

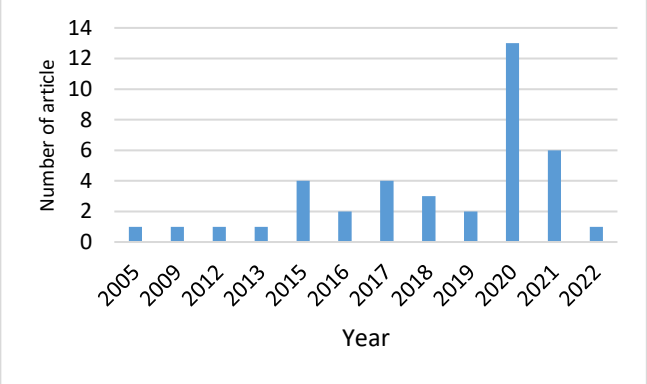

Figure 1. Number of articles by publication year

Table 4. Domain vs number of attributes

Domain	No of attribute	
	Max	Min
Automobile	32	11
Bank	31	11
Healthcare	98	20

This section presents the results of the data to address the research questions based on the selected papers. The first research question is about the popular machine learning technology used in the fraud detection in multiple domains. Three domains have been discussed in this section, auto industry, healthcare, bank, and financial institutions. Table 5 represents the domain and the popular machine learning techniques used to detect and predict the fraudulent activities in particular domain. From the literature survey, it was found that most common algorithms used in multiple domains in random forest followed by support vector machine, neural network, logistic regression, naïve bayes and XGBoost. From this analysis, it can be concluded that the most common popular machine learning algorithm used in fraud detection to answer the research question.

Table 5. Domain vs algorithm used

Domain	Algorithms	Domain	Algorithms
Auto Industry	Artificial neural network	Healthcare	Random forest
Auto Industry	Naïve bayes	Healthcare	LEIE
Auto Industry	Neural network	Healthcare	K means clustering
Auto Industry	Random forest	Healthcare	Hierarchal clustering
Auto Industry	Stochastic Gradient Descent and Adaptive Boosting	Healthcare	Naïve bayes
Auto Industry	Decision Table	Healthcare	Decision trees
Auto Industry	Multilayer perception	Healthcare	Support Vector Machine
Auto Industry	Logistic Regression	Healthcare	ECM
Auto Industry	Partial Decision Trees (PART)	Healthcare	Unsupervised
Auto Industry	Support vector machine	Healthcare	Data mining
Auto Industry	Multiple linear regression	Healthcare	supervised learning
Auto Industry	Decision tree	Healthcare	machine learning
Auto Industry	XGBoost	Healthcare	80 - 20 approach
Auto Industry	K-Nearest Neighbours	Healthcare	oversampling approach
Auto Industry	GMDH	Bank and Finance	K-Nearest Neighbour
Auto Industry	MLP	Bank and Finance	C5.0
Auto Industry	Multilayer protection	Bank and Finance	Neural network
Auto Industry	C4.5	Bank and Finance	Logistics Regression
Healthcare	Random forest	Bank and Finance	Random Forest
Healthcare	XGBoost	Bank and Finance	Support vector machine
Healthcare	LightGBM	Bank and Finance	Random forest
Healthcare	Neural network	Bank and Finance	Logistic regression
Healthcare	GLM	Bank and Finance	SVDD
Healthcare	Gradient boosting	Bank and Finance	Artificial Neural Network
Healthcare	Logistic regression	Bank and Finance	Neo4j
Healthcare	Classification trees	Bank and Finance	Cypher query language

Table 6 describes the algorithms used in multiple domains with the dataset used and accuracy found from the analysis.

Table 6. Algorithm vs No of data vs Accuracy

Algorithm	No of data		Accuracy	
	Max	Min	Max	Min
Random forest	6000000	500	99.7	82
Logistic regression	6000000	418	99	78
Support vector machine	1000000	1000	98.7	65.1
Artificial neural network	837082	15419	82.9	82.9
Decision tree	837082	500	95.8	79
KNN	837082	246000	92	82
GBM	1000000	383587	99.6	99.6
Naïve bayes	837082	500	77.1	77.1
Neural network	1000000	418	95	75
XGBoost	837082	382587	95	76.81

Table 7 represents the validation techniques used to detect the insurance fraud in multiple domain such as healthcare, bank and financial institutions and automotive industry. From the review of multiple articles, it can be found that out of 94 algorithms, 49 algorithms have used confusion matrix as a validation technique.

Table 7. Validation technique used by algorithms

Validation	Count of Algorithm used
AUCROC	3
Confusion matrix	49
Cross validation technique	38
Two factor anova and turkey's HSD test	2
Wavelet and Permutation Pair Frequency Matrix (PPFM)	2
Total	94

Table 8 and Figure 2 represent the algorithms used in auto industry and accuracy percentage from these algorithms. From the literature review, it can be seen that most popular and effective algorithms are logistic regression, support vector machine and random forest. From the reviewed literature, logistic regression has higher percentage of accuracy with 99 percent and followed by support vector machine with 98.7 percent and random forest with 98.5.

Table 8. Algorithm vs accuracy in auto industry

Algorithm used	Domain	Accuracy
Logistic regression	auto industry	99
Support Vector Machine	auto industry	98.7
Random Forest	auto industry	98.5
Decision Tree	auto industry	95.8
variational auto encoder	auto industry	94.57
Spiking neural network (SNN)	auto industry	94
Kmeans	auto industry	92
Random forest	auto industry	91.01
Random forest	auto industry	89.75
Support vector machine	auto industry	83.4
SRA	auto industry	83
Artificial neural network	auto industry	82.9
Naïve bayes,	auto industry	77.1
XGBoost	auto industry	76.81
Neural network	auto industry	75

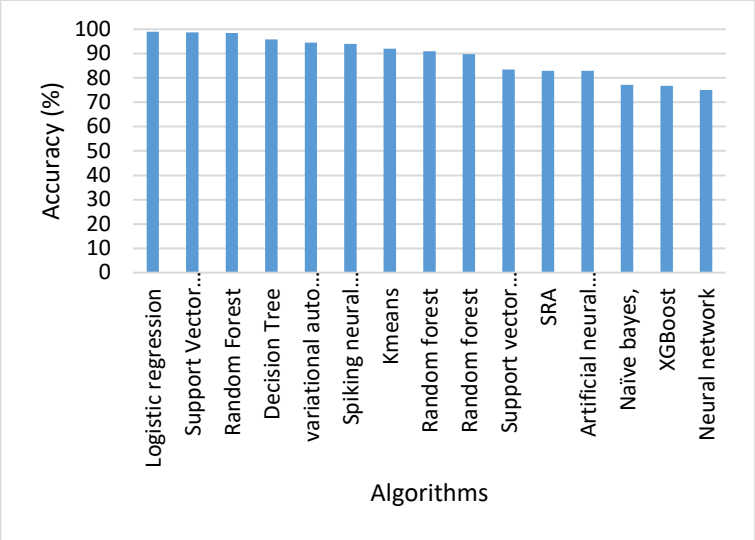

Figure 2. Algorithms vs accuracy in auto industry

4.2 Implementation Result

In this research, a comparison of the four-performance metrics precision, recall, accuracy, and AUC across the various classification models were perfomed. This section presents the findings of our research, which involved assessing the effectiveness of models generated from training data featuring diverse levels of fraud instances. The models under evaluation include Logistic Regression (LR), Random Forests (RF), and Support Vector Machines (SVM). From the analysis, random forest keeps the best score on accuracy with 95% followed by support vector machine 94% and logistic regression 93%. The two models produced 100% non-fraud recall score and only logistic regression has 95% on recall due to the similar characteristics of fraud and non-fraud instances. But the model has performed very poor recall while predicting fraud cases because of the imbalance properties of the dataset. The AUC values shows that all models categorized as excellent classification. The result clearly shows that random forest outperformed support vector machine and logistic regression in terms of accuracy. According to the result, it is proven that ensemble learner produces better performance compare to the single learner.

4.3 Discussions

In this paper, the publications related to the fraud detection with the focus on using machine-learning approach to detect the fraudulent activity on the insurance industry especially on automobile is discussed. In the first part, a systematic literature review is conducted to find out the trends of machine learning algorithm to detect fraudulent claim in different industries. Then, from the result of literature review the most popular three algorithm was compared with the chosen dataset to find the best algorithm to detect insurance fraud in automobile industry.

Using modern technology enables to analyze the behavioral aspects that might help the detection of fraudulent transactions by applying approaches related to the investigation of fraudulent behavior rather than conventional auditing process. It is clear from people's behavior that the human aspect is strongly connected to the fraud. We have evaluated fraud detection-related contributions in this study, with a focus on those that approach fraud detection in different domain such as healthcare institution, bank and finance and automobile industry.

On the other side, several studies have used machine learning approaches to detect fraud and forecast actions connected to this phenomenon. As a result, of our investigation, this strategy was covered in a considerable number of articles. In this situation, we discovered that the majority of algorithms utilized for fraud detection are supervised and unsupervised. With the help of fraudulent and non-fraudulent samples, fraud attempts can be blocked using the supervised technique. Using a labeled training set, rule-based detection automatically infers discriminatory rules. Additionally, our research revealed that supervised algorithms frequently deal with unbalanced classes, which could lead to unsatisfactory result on fraud detection. Additionally, these methods are unable to recognize new fraud trends. However, unsupervised learning focuses on the detection of suspicious behavior as a look for fraud detection and does not need prior knowledge about cases that have been proven fraudulent. So from the literature review, random forest, logistic regression and support vector machine was found most popular with higher percentage of accuracy to detect fraud in insurance claims. In addition, the confusion matrix, K-fold cross validation and area under curve was found to be more

popular for model validation. Therefore, in this research, these findings were utilized to compare the most popular three algorithms RF, LR and SVM and most popular validation approach to detect fraudulent claims.
Here, we exhibit an experimental comparison of the performance scores of three ML model-building techniques used to make predictions using our dataset. Machine learning algorithms has performed better result in dataset because this dataset has large amount of data. Using an evaluation technique, we discovered that random forest's accuracy (95%) is greater than SVM's (94%) and LR's (93%) However, since the confusion matrix revealed an unbalanced distribution of data, we cannot simply rate a model by its accuracy. In addition to that, we also used the Area Under Curve (AUC) to illustrate the degree of class separability.

Across all performance criteria, random forests performed better overall. Random forests are appealing from a practical usage perspective since they are computationally efficient and have only two modifiable parameters that may be set at generally accepted default levels. Many practical machine learning applications have used the support vector machine as a standard technique. This relatively straightforward, well-known, and extensively used approach showed good performance in this investigation as well, frequently outperforming the LR models. As stated before, no intentional efforts were performed in this investigation to optimize the parameters of the procedures. When employing SVM, parameter adjustment can be crucial, and balanced sampling has been found to be helpful when utilizing Random Forests on unbalanced data. The study's dataset is also rather unbalanced, therefore resampling, undersampling, and oversampling may be able to enhance the models' performance. These have the potential to perform better than what is currently presented here and provide pertinent problems that merit additional research.

5. CONCLUSIONS

5.1 Conclusions

Fraud detection is challenging since it needs to know how people behave, but this is not the only problem. Fraud detection models require diverse and representative data to learn the patterns and features that distinguish between different classes of objects or events. When there is an unbalanced data, models might struggle to generalize well to unseen scenarios, leading to poor performance when confronted with new instances. Unbalanced datasets are typical in this field, even when data is available. As a result, there are many various ways to approach the issue of fraud detection. A systematic literature review was carried out in this paper with aim to address these issues from a wider angle. As a result, the goal of this research was to find papers that dealt with fraud detection using ML methods with special interest on automobile industry. From the literature review, it was found that the random forest, logistic regression and support vector machine were the most popular machine learning algorithm to detect insurance fraud in the automobile industry.

The field of fraud analytics is expanding, particularly in the realm of identifying fraudulent vehicle insurance claims. Researchers are dedicated to creating, deploying, and evaluating new models, software tools, and techniques to combat fraud effectively. To achieve this, various machine learning methods have been investigated. The research primarily focuses on three key classification algorithms: random forest, logistic regression, and support vector machine. These algorithms were employed to discern the outcomes of each approach. The study suggests adopting a supervised classifiers approach using LR, SVM, and RF to differentiate between genuine and false insurance claims. Through an evaluation methodology, it was determined that the accuracy of the random forest algorithm stands at 95%, surpassing the other algorithms.

5.2 Limitations and Further Scope

An enterprise-wide methodology for preventing, detecting, and managing insurance fraud helps decision-makers make more informed choices, improve capital efficiency, and boost business performance. The research's focus includes solution types such fraud analytics, authentication, and fraud detection.

Given the imbalanced nature of the dataset, various sampling techniques like under-sampling, oversampling, and SMOTE can be employed to mitigate the prevalent class imbalance, which is common in fraud detection scenarios. However, it is important to acknowledge that these techniques can come with computational demands, especially when dealing with large datasets. Considering this, a more viable approach might involve using a different and more recent dataset in the future, since the current dataset is both outdated and imbalanced.

When it comes to feature selection, it is crucial to carefully measure the impact of selecting specific features to maintain a balance between the overall set of features and the selected ones. Furthermore, enhancing the model's performance can be achieved by incorporating additional ensemble learning classifiers such as XGBoost, CatBoost, and AdaBoost. Optimization techniques can also contribute to model refinement.
In terms of future improvements, leveraging deep neural networks holds promise due to their superior accuracy, speed of classification, and capability to handle interdependent attributes. Additionally, their potential for incremental learning could be particularly advantageous in the realm of fraud detection.

REFERENCE

[1] J. West, M. Bhattacharya, R. Islam, "Intelligent Financial Fraud Detection Practices: An Investigation," *in International Conference on Security and Privacy in Communication Networks*, pp. 186–203, 2015. https://doi.org/10.1007/978-3-319-23802-9_16

[2] A. M. Caldeira, W. Gassenferth, M. A. S. Machado, D. J. Santos, "Auditing Vehicles Claims using Neural Networks," *in Procedia Computer Science*, vol. 55, pp. 62–71, 2015. https://doi.org/10.1016/j.procs.2015.07.008

[3] M. Kirlidog, C. Asuk, "A Fraud Detection Approach with Data Mining in Health Insurance," *Procedia - Social Behavioral Sciences*, vol. 62, pp. 989–994, 2012. https://doi.org/10.1016/j.sbspro.2012.09.168

[4] V. Rawte, G. Anuradha, "Fraud Detection in Health Insurance using Data Mining Techniques," *2015 International Conference on Communication, Information & Computing Technology* (*ICCICT*), Jan. 2015. https://doi.org/10.1109/ICCICT.2015.7045689

[5] M. Al Marri, A. AlAli, "Financial Fraud Detection using Machine Learning Techniques," *RIT Digital Institutional Repository*, Rochester Institute of Technology, Dubai, 2020.

[6] K. Nian, H. Zhang, A. Tayal, T. Coleman, Y. Li, "Auto Insurance Fraud Detection using Unsupervised Spectral Ranking for Anomaly," *The Journal of Finance and Data Science*, vol. 2, no. 1, pp. 58–75, Mar. 2016. https://doi.org/10.1016/j.jfds.2016.03.001

[7] Y. Wang, W. Xu, "Leveraging Deep Learning with LDA-based Text Analytics to Detect Automobile Insurance Fraud," *Decision Support Systems*, vol. 105, pp. 87–95, 2018. https://doi.org/10.1016/j.dss.2017.11.001

[8] J. O. Awoyemi, A. O. Adetunmbi, S. A. Oluwadare, "Credit Card Fraud Detection using Machine Learning Techniques: A Comparative Analysis," *2017 International Conference on Computing Networking and Informatics* (*ICCNI*), Oct. 2017. https://doi.org/10.1109/ICCNI.2017.8123782

[9] M. K. Severino, Y. Peng, "Machine Learning Algorithms for Fraud Prediction in Property Insurance: Empirical Evidence using Real-world Microdata," *Machine Learning with Applications*, vol. 5, p. 100074, 2021. https://doi.org/10.1016/j.mlwa.2021.100074

[10] A. Abdallah, M. A. Maarof, A. Zainal, "Fraud Detection System: A Survey," *Journal of Network and Computer Applications*, vol. 68, pp. 90–113, 2016. https://doi.org/10.1016/j.jnca.2016.04.007

[11] M. A. Caruana and L. Grech, "Automobile Insurance Fraud Detection," *Communications in Statistics: Case Studies, Data Analysis and Applications*, vol. 7, no. 4, pp. 520–535, 2021. https://doi.org/10.1080/23737484.2021.1986169

[12] V. Ambatipudi, "Machine Learning models for Automobile Fraud Detection - A literature Review Agenda," *20th Global Conference of Actuaries*, 2019.

[13] R. Bhowmik, "Detecting Auto Insurance Fraud by Data Mining Techniques," *Journal of Emerging Trends in Computing and Information Sciences*, vol. 2, no. 4, pp. 156–162, 2011.

[14] J. Brownlee, "Tune Hyperparameters for Classification Machine Learning Algorithms," Machine Learning Mastery, 2020, [Online] Available: https://machinelearningmastery.com/hyperparameters-for-classification-machine-learning-algorithms, accessed Jul. 8, 2024.

[15] Sheethal H. D., P. Sai Pranavi, Sharanya S. Kumar, Sonika Kariappa, Swathi B. H. Gururaj H. L., "Comparative Analysis on Vehicle Insurances Fraud Detection using Machine Learning," *International Journal of Advance Research, Ideas and Innovations in Technology*, vol. 6, no. 3, 2020.

[16] P. Dua, S. Bais, "Supervised Learning Methods for Fraud Detection in Healthcare Insurance," *Machine Learning in Healthcare Informatics*, vol. 56, pp. 261–285, 2014. https://doi.org/10.1007/978-3-642-40017-9_12

[17] R. Y. Gupta, S. S. Mudigonda, P. K. Baruah, "A Comparative Study of using Various Machine Learning and Deep Learning-based Fraud Detection Models for Universal Health Coverage Schemes," *International Journal of Engineering Trends and Technology* (*IJETT*), vol. 69, no. 3, pp. 96–102, 2021. https://doi.org/10.14445/22315381/IJETT-V69I3P216

[18] J. Pesantez-Narvaez, M. Guillen, M. Alcañiz, "Predicting Motor Insurance Claims using Telematics Data—XGboost Versus Logistic Regression," *Risks*, vol. 7, no. 2, 2019. https://doi.org/10.3390/risks7020070

[19] G. Kowshalya, M. Nandhini, "Predicting Fraudulent Claims in Automobile Insurance," *2018 Second International Conference on Inventive Communication and Computational Technologies* (*ICICCT*), Coimbatore, India, pp. 1338-1343, 2018. https://doi.org/10.1109/ICICCT.2018.8473034

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

[21] S. Subudhi, S. Panigrahi, "Use of Possibilistic Fuzzy C-means Clustering for Telecom Fraud Detection," *Computational Intelligence in Data Mining*, pp. 633–641, 2017. https://doi.org/10.1007/978-981-10-3874-7_60

[22] R. A. Bauder, T. M. Khoshgoftaar, "Medicare Fraud Detection Using Machine Learning Methods," *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Cancun, Mexico, pp. 858-865, 2017. https://doi.org/10.1109/ICMLA.2017.00-48

[23] R. Roy, K. T. George, "Detecting Insurance Claims Fraud using Machine Learning Techniques," *2017 International Conference on Circuit, Power and Computing Technologies (ICCPCT)*, Kollam, India, pp. 1-6, 2017. https://doi.org/10.1109/ICCPCT.2017.8074258

[24] B. Itri, Y. Mohamed, Q. Mohammed, 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, pp. 1-4, 2019. https://doi.org/10.1109/ICDS47004.2019.8942277

[25] Y. Kumar, S. Saini, R. Payal, Y. Kumar, A. Professor, "Comparative Analysis for Fraud Detection using Logistic Regression, Random Forest and Support Vector Machine," *International Journal of Research and Analytical Reviews* (*IJRAR*), vol. 7, no. 4, 2020. http://dx.doi.org/10.2139/ssrn.3751339

[26] M. Mathew, N. M. Kunjumon, R. Maria Lalji, K. Susan Skariah, "Motor Insurance Claim Processing and Detection of Fraudulent Claims Using Machine Learning," *International Journal of Future Generation Communication and Networking*, vol. 13, no. 3, pp. 1855–1860, 2020.

[27] Y. Li, C. Yan, W. Liu, M. Li, "A Principle Component Analysis-based Random Forest with the Potential Nearest Neighbor Method for Automobile Insurance Fraud Identification," *Applied Soft Computing*, vol. 70, pp. 1000–1009, 2018. https://doi.org/10.1016/j.asoc.2017.07.027

[28] A. Sheshasaayee, S. S. Thomas, "A Purview of the Impact of Supervised Learning Methodologies on Health Insurance Fraud Detection," in *Advances in Intelligent Systems and Computing*, vol. 672, pp. 978–984, 2018. http://dx.doi.org/10.1007/978-981-10-7512-4_98

[29] D. Vineela, P. Swathi, T. Sritha, K. Ashesh, "Fraud Detection in Health Insurance Claims using Machine Learning Algorithms," *International Journal of Recent Technology and Engineering* (*IJRTE*), vol. 8, no. 5, pp. 2999–3003, 2020. http://dx.doi.org/10.35940/ijrte.e6485.018520

[30] G. G. Sundarkumar, V. Ravi, V. Siddeshwar, "One-class Support Vector Machine Based Undersampling: Application to Churn Prediction and Insurance Fraud Detection," *2015 IEEE International Conference on Computational Intelligence and Computing Research* (*ICCIC*), 2015. https://doi.org/10.1109/ICCIC.2015.7435726

[31] S. Subudhi, S. Panigrahi, "Detection of Automobile Insurance Fraud Using Feature Selection and Data Mining Techniques," *International Journal of Rough Sets and Data Analysis*, vol. 5, no. 3, pp. 1–20, Jul. 2018. http://dx.doi.org/10.4018/IJRSDA.2018070101

[32] C. Muranda, A. Ali, T. Shongwe, "Detecting Fraudulent Motor Insurance Claims Using Support Vector Machines with Adaptive Synthetic Sampling Method," *ITMS 2021 - 2021 62nd International Scientific Conference on Information Technology and Management Science of Riga Technical University*, 2021.

[33] T. Miyato, S.-I. Maeda, M. Koyama, S. Ishii, "Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1979–1993, Aug. 2019. https://doi.org/10.1109/TPAMI.2018.2858821

[34] M. Artís, M. Ayuso, M. Guillén, "Detection of Automobile Insurance Fraud with Discrete Choice Models and Misclassified Claims," *Journal of Risk and Insurance*, vol. 69, no. 3, pp. 325–340, 2002. https://doi.org/10.1111/1539-6975.00022

[35] R. R. Popat, J. Chaudhary, "A Survey on Credit Card Fraud Detection Using Machine Learning," *2018 2nd International Conference on Trends in Electronics and Informatics* (*ICOEI*), Tirunelveli, India, pp. 1120-1125, 2018. https://doi.org/10.1109/ICOEI.2018.8553963

[36] A. Kamil, I. Hassan, A. Abraham, "Modeling Insurance Fraud Detection Using Ensemble Combining Classification," *International Journal of Computer Information Systems and Industrial Management Applications*, vol. 8, pp. 257–265, 2016.

[37] J. M. Johnson, T. M. Khoshgoftaar, "Medicare Fraud Detection using Neural Networks," *Journal of Big Data*, vol. 6, no. 1, Dec. 2019. https://doi.org/10.1186/s40537-019-0225-0

[38] S. Bansal, "Vehicle Insurance Claim Fraud Detection," Kaggle, 2021. [Online] Available: https://www.kaggle.com/datasets/shivamb/vehicle-claim-fraud-detection/data, accessed Jul. 9, 2024.