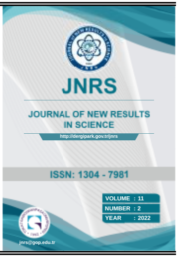
Journal of New Results in Science
https://dergipark.org.tr/en/pub/jnrs
Research Article
Open Access



E-ISSN:1304-7981 https://doi.org/10.54187/jnrs.1129440

A novel data processing approach to detect fraudulent insurance claims for physical damage to cars

Ahmet Yücel[1]

Keywords:
Artificial neural network, Tree-based decision systems, Support vector machines, Singular value decomposition, Data processing, Natural language processing

Abstract — Some automobile insurance companies use computerized auto-detection systems to expedite claims payment decisions for insured vehicles. Claims suspected of fraud are evaluated using empirical data from previously investigated claims. The main objective of this manuscript is to demonstrate a novel data processing system and its potential for use in data classification. The data processing approach was used to develop a machine learning-based sentiment classification model to describe property damage fraud in vehicle accidents and the indicators of fraudulent claims. To this end, Singular Value Decomposition-based components and correlation-based composite variables were created. Machine learning models were then developed, with predictors and composite variables selected based on standard feature selection procedures. Five machine learning models were used: Boosted Trees, Classification and Regression Trees, Random Forests, Artificial Neural Networks, and Support Vector Machines. For all models, the models with composite variables achieved higher accuracy rates, and among these models, the artificial neural network was the model with the highest accuracy performance at 76.56%.

Subject Classification (2020): 62C25, 62P25.

1. Introduction

The insurance industry is constantly evolving. New technologies, such as artificial intelligence (AI) and machine learning (ML), provide new opportunities for insurers to detect fraudulent claims. These technologies can analyse data from various sources, such as social media and credit card transactions, etc., to determine the risk ratio of clients. Fraudulent insurance claims are a significant problem for the insurance industry. In 2004, the Insurance Information Institute calculated that fraudulent property and casualty claims were more than $30 billion annually [1]. Insurance claims for physical damage to cars are usually filed by taking pictures of the car and uploading them as evidence. This has become a significant problem because it can lead to increased premiums for all drivers [2], not just those who have made fraudulent claims. There are also costs associated with people who are falsely accused of fraud, which can lead to serious consequences, including job loss and even imprisonment [3]. Baader

[1]ayucel@ybu.edu.tr (Corresponding Author)
[1]Department of Finance and Banking, Şereflikoçhisar Faculty of Applied Sciences, Ankara Yıldırım Beyazıt University, Ankara, Turkiye

Article History: Received: 11 Jun 2022 — Accepted: 22 Aug 2022 — Published: 31 Aug 2022

and Krcmar aim to decrease the number of false positives in internal fraud detection using a novel approach based on process mining [4].

Machine learning is a type of AI that uses algorithms to learn from data and make predictions or decisions [5]. Fraudulent insurance claims are one example of how machine learning can be used to make decisions by identifying patterns and trends in fraudulent claims, which will help insurers save money and reduce the number of fraudulent claims they pay out. Ma et al. use ML algorithms to detect radio frequency identification (RFID) reading problems. One of the technical problems hindering the effective and reliable use of RFID is false positive detection. In other words, it is a misclassified labelling problem. Based on the features obtained from the data, logistic regression (LR), SVM, and tree-based models were built to distinguish the false-positive reads. The outcomes indicate that SVM provides the best accuracy [6]. Chand and Zhang provide a fraud prediction model for property insurance claims using different ML models based on real data from a large Brazilian insurance company. The accuracy of the models was evaluated by counting the number of false positives and negatives. The main techniques used were Random Forest (RF), Gradient Boosted Trees (BT), and Artificial Neural Networks (ANN) [7].

This kind of AI system aims to detect fraudulent transactions with high accuracy and low false positives. This AI system can detect fraudulent transactions in different sectors, such as banking, insurance, healthcare, and retail. Mishra and Dash propose an AI-based approach to determine fraudulent transactions. This type of online transaction can be easily performed with a person's stolen bank card details. Decision tree and ANN models were used for detection [8]. Van Capelleveen et al. explain how unsupervised outlier techniques can be applied post-payment to determine fraudulent patterns in health insurance claims [9]. Sabetti and Heijmans apply a specific ANN model to detect abnormal payment activity in the Canadian retail payment system. The study examines the model's ability to detect irregular changes in transaction flows [10].

An Artificial Neural Network (ANN) is a computer system modelled on the human brain, composed of a large number of interconnected artificial neurons [11]. ANNs are well suited to solving problems that are either too complex for humans to solve or too vast in size. Therefore, ANNs can be used to detect fraudulent cases and make decisions or predictions. Ansari and Riasi propose an ANN-based approach to assess customer loyalty in newly established insurance companies [12].

A tree-based decision system is a type of decision system that uses a tree structure to represent the possible decisions and their outcomes. The tree is generated by considering all the possible combinations of input variables, and the decision node is used to evaluate each combination.

Frempong et al. present an estimation model that calculates the probability of insurance coverage based on some potential risk factors of vehicle insurance. The age of the vehicle, the age of the insured, etc., are the main risk factors that predict the occurrence of vehicle damage [13].

Support Vector Machines (SVM) are an algorithm that predicts a data set's binary classification. SVMs are one of the most popular algorithms for binary classification, which is the prediction of two possible outcomes for an input value. This can be seen in fraud detection, which predicts whether an individual transaction is fraudulent or not. Gyamfi and Abdulai examine various forms of fraud in the database of monetary transactions in banking systems to determine potentially fraudulent transactions. They also introduce an SVM model that represents regular and irregular transaction behaviour and then use it to assess transaction validity [14]. It can also be seen in cancer detection. Badr et al. propose an SVM model for breast cancer diagnosis with effective scaling methods [15]. Support vector machines are often confused with other machine learning algorithms because they use similar terminology and techniques; however, they have different purposes and training procedures [16].

Natural Language Processing (NLP) has been on the rise in recent years. NLP is a subfield of Text Mining (TM) that focuses on understanding text and speech and the language used to represent unstructured information [17]. TM is an umbrella term for technologies that can be used to extract meaning from unstructured data, such as sentiment analysis [18], semantic analysis [19], discourse analysis [20], and summarization [21]. NLP has been applied to many fields to extract information from complex unstructured data sets in different industries. Zhang et al. propose a model to detect potentially fraudulent financial activities through textual analysis of financial reports of some Chinese companies [22]. Fu et al. use an NLP algorithm to analyse a dataset based on periprosthetic joint disease [23].

Data processing is an essential part of data analytics. It converts raw data into a format that a computer can use. Data is often processed to make it more useful, for example, by converting it to the machine-readable binary format or compressing the data's size. Nourani et al. used three methods to pre-process the data. These are wavelet-based de-noising (WD), jitted data-training (JD), and the hybrid method resulting from integrating these two methods. The objective of this study was to observe the effect of different data pre-processing approaches on the estimated values of the ANN-based prediction interval analysis. It has been concluded that pre-processing methods significantly contribute to reducing the impact of uncertainty [24]. Zhang et al. present a novel approach to data pre-processing based on nonparametric kernel-based modelling. The performance of the proposed approach was observed on SVM. Experimental results show that the proposed data pre-processing approach provides more reliable and stable results than the regular data processing methods [25]. Chilipirea et al. propose a novel data processing model for the datasets operated by smart cities. The model consists of the entire data flow from the source to the user of the extracted information. In general, the model includes the phases of data collection, normalization, categorization, storage, analysis, visualization, and decision support [26]. Using real-world data, Hanafy and Ming evaluate 13 ML methods. Due to the imbalanced datasets in the field, insurance fraud prediction has become a major challenge. They propose an approach to improve the results of ML algorithms by using novel resampling methods such as random over sampler, random under sampler, and combination of these two methods to solve the problem of imbalanced data. And Hanafy and Ming compare them with each other. The results show that the resampling techniques significantly improve the efficiency of all ML classifiers [27]. Severino and Pend investigated the prediction of property insurance claim fraud using several ML models based on real data. The models were tested iteratively, and the average prediction results were compared. Results showed that RF, BT, and ANN produced the best results compared to logistic regression (LR) [28]. Roy and George propose a study on detecting automobile insurance claims fraud using ML methods [29].

Feature selection (FS) is one of the most important stages of a data analysis process. FS techniques reduce the number of input variables by eliminating redundant or irrelevant features. The goal of FS in ML is to identify the most useful features that can be used to build the best model for the case study. Especially in the field of TM, it simplifies the multidimensional structure and complexity of the data while eliminating the overfitting problem. Secondly, the advantage of such an approach is that it leads to a drastic reduction of the computational cost, which can become very important in the analysis of unstructured real-time data. Reducing the number of features to as few as possible increases the model's accuracy and speed of analysis and makes it seem less cumbersome. The new approach presented in this study has shown that speed and accuracy have increased for all models in the literature with different algorithmic structures.

The novel approach in this study is about how we can use data processing and machine learning algorithms to generate correlation-based variables to improve the performance of AI-based decision-making models. The importance of the data processing approach in this study is that it is a kind of dimension reduction method that helps to make sense of the raw data, and it also helps to reduce the time needed for analysis. The ML models introduced in this study analyse the data of all the customers who have submitted a claim for physical damage to their cars to detect any suspicious activity. They are mainly used to find out if there are any fraudulent claims for physical damage to cars.

The rest of the article is organized as follows. Section 2 presents the material and methodology of the study. Section 3 introduces the results of the proposed approach and models. Section 4 provides the conclusions.

2. Methodology

The data used in this study were collected from kaggle.com. The dataset, consisting of a total of 25 variables and 5589 cases, was collected to identify physical damage fraud related to vehicle accidents and to explain the indicators of fraud claims. In addition to the accuracy of fraud detection, this study also aims to identify the main factors that cause fraud. The definitions of the variables in the data are presented in Table 1.

Table 1. Variable definitions

Variable	Type	Role	Description
accident_site	Categorical	Input	Accident site
address_change_ind	Categorical	Input	Whether the driver has changed home address in the last 1 year
age_of_driver	Continuous	Input	Driver age
age_of_vehicle	Continuous	Input	Age of first-party vehicle
annual_income	Continuous	Input	Driver's annual income
channel	Categorical	Input	Policy buying channel
claim_day_of_week	Categorical	Input	Day of the week of first claim notification
claim_est_payout	Continuous	Input	Estimated claim payment
fraud	Categorical	Target	Fraud indicator (0=no, 1=yes)
gender	Categorical	Input	Driver's gender
high_education_ind	Categorical	Input	Driver's higher education index
liab_prct	Continuous	Input	Responsibility percentage of the request
living_status	Categorical	Input	Driver's living situation, owner or rental
marital_status	Categorical	Input	Driver's marital status
past_num_of_claims	Continuous	Input	Number of claims filed by the driver in the last 5-years
policy_report_filed_ind	Categorical	Input	Policy statement filed indicator
safty_rating	Continuous	Input	Driver's safety rating index
vehicle_category	Categorical	Input	First-party vehicle category
vehicle_color	Categorical	Input	First-party vehicle colour
vehicle_price	Continuous	Input	First-party vehicle price
vehicle_weight	Continuous	Input	First-party vehicle weight
witness_present_ind	Categorical	Input	Evidence of the allegation

The data contains information about the vehicle, such as age_of_vehicle, vehicle_color, and personal information about the driver who reported an accident, such as age_of_driver and gender. The fraud variable is a binary variable consisting of 0 and 1 and is used as the dependent variable. 0 means that the accident report is genuine, and 1 means that it is fake. The frequencies of the categorical variables are presented in table 2, and the descriptive statistics of the continuous variables are presented in Table 3.

Table 2. Frequency table of categorical variables

		Count	Per cent			Count	Per cent
vehicle_color	white	391	7	vehicle_category	Compact	1910	34.17
	grey	392	7.01		Large	1813	32.44
	red	400	7.16		Medium	1866	33.39
	other	383	6.85	accident_site	Local	2869	51.33
	black	2140	38.29		Highway	1308	23.40
	blue	1465	26.21		Parking Lot	1412	25.26
	silver	418	7.48	policy_report_filed_ind	not reported	2171	38.84
living_status	Rent	2510	44.91		reported	3418	61.16
	Own	3079	55.09	marital_status	not married	1791	32.05
claim_day_of_week	Friday	753	13.47		married	3798	67.95
	Thursday	746	13.35	high_education_ind	highschool	1930	34.53
	Tuesday	830	14.85		undergrad	3659	65.47
	Saturday	818	14.64	address_change_ind	changed	3399	60.82
	Wednesday	817	14.62		not changed	2190	39.18
	Sunday	835	14.94	witness_present	not present	4446	79.55
	Monday	790	14.13		present	1143	20.45
channel	Broker	2929	52.41	gender	female	2755	49.29
	Online	821	14.69		male	2834	50.71
	Phone	1839	32.90				

Table 3. Descriptive statistics of continuous variables

	Valid N	Mean	Median	Minimum	Maximum	Std.Dev.
age_of_driver	5589	42.82	42	18	229.0	11.53
safty_rating	5589	0.73	0.76	0.01	1.0	0.16
annual_income	5589	37220.51	37382	28910	54333.0	2671.84
past_num_of_claims	5589	0.61	0	0	6.0	1.07
liab_prct	5589	49.27	50	0	100.0	33.23
claim_est_payout	5589	4951.37	4615.43	282.64	17218.4	2303.72
age_of_vehicle	5589	5.11	5	0	15.0	2.26
vehicle_price	5589	22960.15	20871.74	2722.86	100224.7	11988.58
vehicle_weight	5589	23180.90	20904	2713.47	95464.4	12096.46

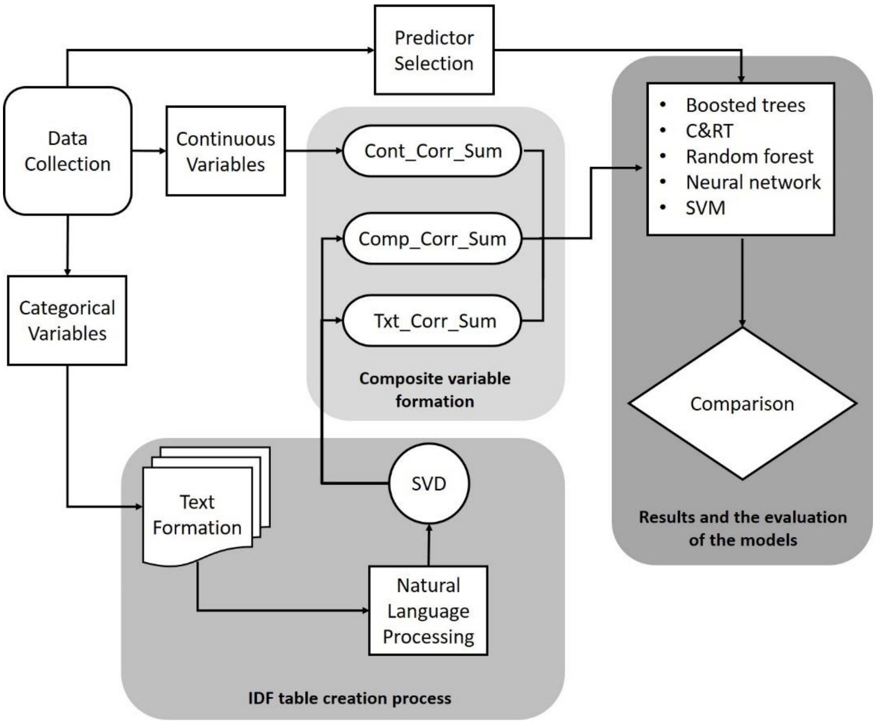


Figure 1. Overview of the proposed methodology

The proposed novel data processing approach (as depicted in Figure 1) consists of 3 basic steps. In the first step, the variables in the data are divided into a categorical and a continuous part, and the categories of the categorical variables are labelled with the words to which they correspond. Thus, for each case, a text is formed consisting of the categories of the categorical variables. In the second step, the texts were processed with Natural Language Processing (NLP), and then a table of frequency of terms and documents was created, where each category was a variable. The inverse document frequency (IDF) method was used to create the frequency table. Natural Language Processing (NLP) is the field of computer science and artificial intelligence that deals with the relations between computers and human (natural) language, specifically how to program computers to process and investigate large amounts of natural textual data. NLP has a wide range of applications, from search engines (e.g., Google) that can automatically find online information relevant to a query to speech recognition systems that can translate spoken words into text. Inverse document frequency (IDF) is a statistical measure of how important a word or phrase is in a document. It is calculated as the logarithm of the number of documents containing the term divided by the total number of documents in which it occurs [30].

Then, the new variables created based on the categories were reduced to composite variables (SVD components) using the singular value decomposition (SVD) method. SVD is a linear algebra technique used to generate the best approximation to a matrix [31].

The covariance value between x and y is calculated with the formula

$$cov(x,y) = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})/n \tag{2.1}$$

The values $\bar{x}$ and $\bar{y}$ are the arithmetic mean of the variables x and y. The correlation value between x and y is calculated with the formula

$$r_{xy} = cov(x,y)/S_x\, S_y \tag{2.2}$$

where the standard deviations S_x and S_y are any two non-zero variables, and r_{xy} takes a value in the range [0,1].

Let $r_i,\ i = 1,2,\dots,n$ be correlations between the continuous variables $C_1,\ \ C_2,\ \ C_3,\ \ \dots\ \ C_n$ and dependent variable 'fraud' and let R_{Cont} be a vector of the correlations such that

$$R_{Cont} = \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ \vdots \\ r_n \end{bmatrix} \tag{2.3}$$

Therefore,

$$\text{Cont\_Corr\_Sum} = [C_1 \quad C_2 \quad C_3 \quad \dots \quad C_n] \cdot R_{Cont} \tag{2.4}$$

Let $t_i,\ \ i = 1,2,\dots,m$ be correlations between the NLP-generated IDF-based variables $T_1,\ \ T_2,\ \ T_3,\ \ \dots\ \ T_m$ and dependent variable 'fraud' and let R_{Text} be a vector of the correlations such that

$$R_{Text} = \begin{bmatrix} t_1 \\ t_2 \\ t_3 \\ \vdots \\ t_m \end{bmatrix} \tag{2.5}$$

$$\text{Txt\_Corr\_Sum} = [T_1, \quad T_2, \quad T_3, \quad \dots \quad T_m] \cdot R_{Text} \tag{2.6}$$

Similarly, let $s_i,\ i = 1,2,\dots,k$ be correlations between the SVD-generated IDF-based components $K_1, K_2, K_3, \dots, K_k$ and dependent variable 'fraud' and let R_{Comp} be a vector of the correlations such that

$$R_{Comp} = \begin{bmatrix} s_1 \\ s_2 \\ s_3 \\ \vdots \\ s_k \end{bmatrix} \tag{2.7}$$

$$\text{Comp\_Corr\_Sum} = [K_1, \quad K_2, \quad K_3, \quad \dots \quad K_k] \cdot R_{Comp} \tag{2.8}$$

In the third step, weight was assigned according to the correlation values of the continuous variables with the dependent variable. The correlation matrices R_Cont, R_Text, and R_Comp, were used to assign the weights, resulting in the composite variables *Cont_Corr_Sum*, *Txt_Corr_Sum*, and *Comp_Corr_Sum*. The rank order of importance of the predictors of the composite variables is shown in Table 4. In Statistica, the importance of the estimator is calculated by subtracting or adding the change (delta) in impurity (in other words, the measure of homogeneity) of all nodes in the decision tree. The order of the values obtained corresponds to the order of importance of the estimators [32].

Table 4. Composite predictors' importance

Variable	Rank	Importance
Txt_Corr_Sum	100.00	1.00
Comp_Corr_Sum	94.00	0.94
Cont_Corr_Sum	12.00	0.12

All available variables were used to construct each composite variable, a size reduction method. This approach minimizes the loss of information in the raw data. To observe the difference in performance between the two approaches, standard and composite, all independent variables were used in the predictor selection step when constructing standard models. The variables labelled as important at a confidence level of alpha 0.05 are listed in order of importance in Table 5.

Table 5. Raw data predictor importance

Variable	Rank	Importance
vehicle_color	100	1.00
annual_income	43	0.43
age_of_driver	39	0.39
past_num_of_claims	33	0.33
claim_est_payout	29	0.29
gender	27	0.27
policy_report_filed_ind	21	0.21
high_education_ind	21	0.21
safty_rating	20	0.20
liab_prct	20	0.20
vehicle_weight	19	0.19
claim_day_of_week	19	0.19
accident_site	19	0.19
marital_status	19	0.19
age_of_vehicle	18	0.18
witness_present_ind	17	0.17
vehicle_price	15	0.15
address_change_ind	12	0.12
channel	10	0.10
living_status	5	0.05
vehicle_category	4	0.04

In the fourth step, machine learning models (ML) were designed with variables selected from the raw data using the standard predictor selection method, and composite variables were produced based on the original data processing approach. 5 different ML models were used, namely BT, C&RT, RF, ANN, and SVM. The models were trained and developed on a 70% training set and applied on a 30% test set.

3. Results and Discussion

C&RT is a type of decision tree used to predict values. The classification tree divides the data into two or more groups, while the regression tree predicts an output value for an input value. The number of non-terminal nodes is 17, and the number of terminal nodes is 18 in C&RT, which was designed using the standard method.

Support Vector Machines (SVM) is an ML algorithm for classification and regression. An SVM consists of two collections of data points: The first collection is called support vectors, and the other collection is called margins. The margin is created by selecting a point on either side of the dividing line and then finding all points on either side of that line. In the models created according to standard and composite variables, the Radial Basis Function (RBF) was used as the kernel function, the fixed value was set to 1, and the gamma value was set to 3. In addition, in the standard method, a total of 1000 separating lines (SVs) in 500 categories 0s and 500 categories 1s were used to estimate the categories (0/1) of the dependent variable. In the composite variable-based method, a total of 890 SVs (452 (0s) + 438 (1s)) were used.

ANN is a computer model inspired by the brain's biological neural networks. They are used for various applications, from medical diagnosis to image mining [33,34]. ANN is also increasingly used for prediction tasks. A typical ANN consists of many neurons arranged in layers, with connections between the layers. These connections form a pattern known as the synaptic weight for that neuron and its neighbours. The process of training an artificial neural network is called backpropagation. In this process, the synaptic weights are iteratively adjusted to improve the predictive ability of the system for future data points [35].

The multi-layer perceptron (MLP) ANN model was used for both approaches. There are 46 predictors and 2 outputs (prediction) in the model ANN, which was built using the standard approach, and there are 3 predictors and 2 outputs in the composite-based ANN model. In addition, both models have 9 hidden layers. When training the models based on the training data, the Broyden-Fletcher-Goldfarb-Shanno (BFGS) 6 training algorithms were used in the standard approach, and BFGS 18 training algorithms were used in the composite approach. In the standard and composite approaches, SOS and entropy were used as error functions, identity as the hidden layer activation function for both, and Tanh and softmax as the output activation functions, respectively.

RF is an ML method used for classification and regression. It is an ensemble model that combines multiple decision trees to make more accurate predictions. The RF can be used for many tasks, from predicting the stock market to classifying documents. In this study, the maximum number of trees is 100 in RF, which are created using standard and composite methods.

Boosted trees (BT) are a type of classification algorithm. The algorithm is based on the idea that classification accuracy can be improved by training a set of decision trees on different subsets of the data and combining their decisions into a final prediction. The optimal number of trees is 88, and the maximum tree size is 13 in the model BT, which was built using the standard method. In the BT model, which was created with composite variables, the optimal number of trees is 49, and the maximum tree size is 13.

Table 6. Cross-tabulation and accuracy rates

Model Method		Boosted Trees Prediction		CandRT Prediction		Random Forest Prediction		ANN Prediction		SVM Prediction	
		0	1	0	1	0	1	0	1	0	1
Standard	0	478	358	382	454	364	472	314	522	476	360
	1	162	674	160	676	169	667	151	685	378	458
Accuracy Rates		68.9		63.28		61.66		59.75		55.86	
Composite	0	599	237	467	369	608	228	634	202	543	293
	1	182	654	168	668	172	664	190	646	170	666
Accuracy Rates		74.94		67.88		76.08		76.56		72.31	
Improvement rate (%)		8.77		7.27		23.39		28.13		29.45	

The aggregate results of the cross-tabulation prediction in a confusion matrix are shown in Table 6. As can be seen from the results, among the models based on the standard procedure, Boosted Trees outperformed the other models with an accuracy of 68.9%, while CandRT, RF, ANN, and SVM achieved accuracy rates of 63.28%, 61.66%, 59.75%, and 55.86%, respectively. In addition, as is represented in the results, among the models built based on the SVD components and composite variable procedure, ANN outperformed the other models by achieving an accuracy rate of 76.56%, while BT, C&RT, RF, and SVM achieved accuracy rates of 74.94%, 67.88%, 76.08%, and 72.31%, respectively.

4. Conclusion

The main objective of this manuscript is to demonstrate a novel data processing framework and its potential for use in data classification. The framework was used to develop an ML-based sentiment classification model to describe the physical damage fraud in vehicle accidents and the indicators of fraud claims. This study is not only concerned with the accuracy of fraud detection but also with determining the main factors that cause fraud.

For this purpose, SVD-based components and correlation-based composite variables were created. Machine learning models (ML) were then developed, with predictors and composite variables selected based on standard feature selection methods. 5 different ML models were used, namely BT, C&RT, RF, ANN, and SVM. For all models, the models with composite variables achieved higher accuracy rates, and among these models, ANN was the model with the highest accuracy performance at 76.56%.

As shown in Table 6, the novel data processing approach proposed in this study resulted in an increase in the accuracy of the models ranging from 7.27% to 29.45%. This represents a successful and significant improvement of 22.06% on average.

Author Contributions

The author read and approved the last version of the manuscript.

Conflicts of Interest

The author declares no conflict of interest.

Appendix 1. Data

The raw data used in this article can be found online at

https://www.kaggle.com/datasets/surekharamireddy/fraudulent-claim-on-cars-physical-damage

Appendix 2. Coding

The coding of the statistical procedures used in this article can be found online at

CODES - A novel data processing approach to detect fraudulent insurance

References

[1] S. Viaene, M. Ayuso, M. Guillen, D. V. Gheel, G. Dedene, *Strategies for detecting fraudulent claims in the automobile insurance industry*, European Journal of Operational Research, 176(1), (2007) 565–583.

[2] T. Baldock, Insurance fraud. Australian Institute of Criminology: Trends and issues in crime and criminal justice, 66, (1997).

[3] I. Akomea-Frimpong, C. Andoh, E. Ofosu-Hene, *Causes, effects and deterrence of insurance fraud: evidence from Ghana*, Journal of Financial Crime, 23(4), (2016) 678–699.

[4] G. Baader, H. Krcmar, *Reducing false positives in fraud detection: Combining the red flag approach with process mining*, International Journal of Accounting Information Systems, 31, (2018) 1–16.

[5] J. Nahr, H. Nozari, M. E. Sadeghi, *Artificial intelligence and machine learning for real-world problems (A survey)*, International journal of innovation in Engineering, 1(3), (2021) 38–47.

[6] H. Ma, Y. Wang, K. Wang, *Automatic detection of false positive RFID readings using machine learning algorithms*, Expert Systems with Applications, 91, (2018) 442–451.

[7] S. Chand, Y. Zhang, *Learning from machines to close the gap between funding and expenditure in the Australian National Disability Insurance Scheme*, International Journal of Information Management Data Insights, 2(1), (2022) 1–15.

[8] M. K. Mishra, R. Dash, *A comparative study of Chebyshev functional link artificial neural network, multi-layer perceptron and decision tree for credit card fraud detection*, in: S. P. Mohanty, R. K. Patnaik, M. Gomathisankaran, B. S. Panda (Eds.) International Conference on Information Technology 2014, Bhubaneswar, India, 2014, pp. 228–233.

[9] G. van Capelleveen, M. Poel, R. M. Mueller, D. Thornton, J. van Hillegersberg, *Outlier detection in healthcare fraud: A case study in the Medicaid dental domain*, International Journal of Accounting Information Systems, 21, (2016) 18–31.

[10] L. Sabetti, R. Heijmans, *Shallow or deep? Training an autoencoder to detect anomalous flows in a retail payment system*, Latin American Journal of Central Banking, 2(2), (2021) 1–14.

[11] J. Jiang, P. Trundle, J. Ren, *Medical image analysis with artificial neural networks*, Computerized Medical Imaging and Graphics, 34(8), (2010) 617–631.

[12] A. Ansari, A. Riasi, *Modelling and evaluating customer loyalty using neural networks: Evidence from startup insurance companies*, Future Business Journal, 2(1), (2016) 15–30.

[13] N. K. Frempong, N. Nicholas, M. A. Boateng, *Decision tree as a predictive modeling tool for auto insurance claims*, International Journal of Statistics and Applications, 7(2), (2017) 117–120.

[14] N. K. Gyamfi, J. D. Abdulai, *Bank fraud detection using support vector machine*, in: V. Leung, S. Vuong, S. Chakrabarti (Eds.), IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON) 2018, Vancouver, BC, Canada, 2018, pp. 37–41.

[15] E. Badr, S. Almotairi, M. A. Salam, H. Ahmed, *New sequential and parallel support vector machine with grey wolf optimizer for breast cancer diagnosis*. Alexandria Engineering Journal, 61(3), (2022) 2520–2534.

[16] G. Tolan, T. Abou-El-Enien, M. Khorshid, *A comparison among support vector machine and other machine learning classification algorithms*, IPASJ International Journal of Computer Science (IIJCS), 3(5), (2015) 25–35.

[17] A. Kao, S. R. Poteet, Natural language processing and text mining, Springer Publishing Company, 2006.

[18] N. Chintalapudi, G. Battineni, M. D. Canio, G. G. Sagaro, F. Amenta, *Text mining with sentiment analysis on seafarers' medical documents*, International Journal of Information Management Data Insights, 1(1), (2021) 1–9.

[19] R. Alfrjani, T. Osman, G. Cosma, *A hybrid semantic knowledgebase-machine learning approach for opinion mining*, Data and Knowledge Engineering, 121, (2019) 88–108.

[20] E. Teso, M. Olmedilla, M. Martínez-Torres, S. Toral, *Application of text mining techniques to the analysis of discourse in eWOM communications from a gender perspective*, Technological Forecasting and Social Change, 129, (2018) 131–142.

[21] O. Rouane, H. Belhadef, M. Bouakkaz, *Combine clustering and frequent itemsets mining to enhance biomedical text summarisation*, Expert Systems with Applications, 135, (2019) 362–373.

[22] Y. Zhang, A. Hu, J. Wang, Y. Zhang, *Detection of fraud statement based on word vector: Evidence from financial companies in China*, Finance Research Letters, 46, (2022) 1–7.

[23] S. Fu, C. C. Wyles, D. R. Osmon, M. L. Carvour, E. Sagheb, T. Ramazanian, H. M. Kremers, *Automated detection of periprosthetic joint infections and data elements using natural language processing*, The Journal of Arthroplasty, 36(2), (2021) 688–692.

[24] V. Nourani, M. Sayyah-Fard, M. T. Alami, E. Sharghi, *Data pre-processing effect on ANN-based prediction intervals construction of the evaporation process at different climate regions in Iran*, Journal of Hydrology, 588, (2020) 1–15.

[25] W. Zhang, T. Liu, L. Ye, M. Ueland, S. L. Forbes, S. W. Su, *A novel data pre-processing method for odour detection and identification system*, Sensors and Actuators A: Physical, 287, (2019) 113–120.

[26] C. Chilipirea, A. C. Petre, L. M. Groza, C. Dobre, F. Pop, *An integrated architecture for future studies in data processing for smart cities*, Microprocessors and Microsystems, 52, (2017) 335–342.

[27] M. Hanafy, R. Ming*, Using machine learning models to compare various resampling methods in predicting insurance fraud*, Journal of Theoretical and Applied Information Technology, 99(12), (2021), 2819–2833.

[28] M. K. Severino, Y. Peng, *Machine learning algorithms for fraud prediction in property insurance: Empirical evidence using real-world microdata*, Machine Learning with Applications, 5, (2021) 1–14.

[29] R. Roy, K. T. George, *Detecting insurance claims fraud using machine learning techniques*, in: K. P. Isaac, A. Rahiman, G. P. Padmakumar (Eds.), International Conference on Circuit, Power and Computing Technologies (ICCPCT) 2017, Kollam, India, 2017, pp. 1–6.

[30] G. Miner, D. Delen, J. Elder, A. Fast, T. Hill, R. A. Nisbet, *Conceptual foundations of text mining and pre-processing steps, practical text mining and statistical analysis for non-structured text data applications*, Academic Press. (2012) 43–51.

[31] A. K. Menon, C. Elkan, *Fast algorithms for approximating the singular value decomposition*, ACM Transactions on Knowledge Discovery from Data, 5(2), (2011) 1–36.

[32] TIBCO product documentation, Data Science Textbook, https://docs.tibco.com/data-science/GUID-4C6F72C1-F4F8-48A9-83C7-D4C72A66A3AC.html (Accessed on 14.08.2022)

[33] C. Peña-Bautista, T. Durand, C. Oger, M. Baquero, M. Vento, C. Cháfer-Pericás, *Assessment of lipid peroxidation and artificial neural network models in early Alzheimer disease diagnosis*, Clinical Biochemistry, 72, (2019) 64–70.

[34] R. Azadnia, K. Kheiralipour, *Recognition of leaves of different medicinal plant species using a robust image processing algorithm and artificial neural networks classifier*, Journal of Applied Research on Medicinal and Aromatic Plants, 25, (2021) 1–10.

[35] C. Li, R. Chen, C. Moutafis, S. Furber, *Robustness to noisy synaptic weights in spiking neural networks*, in: A. Roy (Ed.), International Joint Conference on Neural Networks (IJCNN) 2020, Glasgow, UK, 2020, pp. 1–8.