Koc. J. Sci. Eng., 5(ICOLES2021): (2022) 25-31 https://doi.org/10.34088/kojose.1019125

Kocaeli University

Kocaeli Journal of Science and Engineering

http://dergipark.org.tr/kojose

SOBE: A Fraud Detection Platform in Insurance Industry

H. Onur ÖZCAN [1], İsmail ÇOLAK [2], Selin ERİMHAN [3], Vedat GÜNEŞ [4], Fatih ABUT [5,*], M. Fatih AKAY [6]

[1] *Anadolu Sigorta, Department of Business Intelligence and Analytical Solutions, Istanbul, Turkey,* ***ORCID:*** *0000-0002-2576-0212*
[2] *Anadolu Sigorta, Department of Business Intelligence and Analytical Solutions, Istanbul, Turkey,* ***ORCID:*** *0000-0002-2287-7183*
[3] *Anadolu Sigorta, Department of Business Intelligence and Analytical Solutions, Istanbul, Turkey,* ***ORCID:*** *0000-0001-5101-5235*
[4] *Anadolu Sigorta, Department of Business Intelligence and Analytical Solutions, Istanbul, Turkey,* ***ORCID:*** *0000-0002-5665-5909*
[5] *Department of Computer Engineering, Çukurova University, Adana, Turkey,* ***ORCID:*** *0000-0001-5876-4116*
[6] *Department of Computer Engineering, Çukurova University, Adana, Turkey,* ***ORCID:*** *0000-0003-0780-0679*

Article Info

Research paper

Received : November 04, 2021
Accepted : March 03, 2022

Keywords

Fraud Detection
Machine Learning
Social Network Analysis
KNIME

Abstract

Fraud detection identifies suspicious activities, false pretenses, wrongful or criminal deception intended to result in financial gain. Fraud is rare, well thought, effortful, and deceiving throughout claims. Detecting fraudulent claims is essential for the insurance industry. Therefore, most insurance companies must devote time and budget to fraud detection. Fraud detection can be divided into two categories; the main and most common type of fraud is individual fraud. Individual frauds can appear in many kinds of forms. For example, damage to an asset might be occurred before issuing a policy and be reported after. The second category is organized fraud which is much rarer and harder to detect than individual fraud. Especially motor insurance fraud is commonly attempted by organized crime rings. Counterparties involved in fraudulent claims change frequently, and changes make fraud detection difficult. According to Insurance Information and Monitoring Center findings, the fraudulent claim payment ratio is 10 to 30 %, and the detection success rate for an individual is at 1.4 to 5%. At the same time, the annual fraud cost is at 200 to 300 $ million. This study proposes a fraud detection platform called SOBE, which assists fraud departments' claim inquiry more easily and shorter than manual investigation made by employees. At its core, SOBE uses a rule engine approach. In order to support the rule engine, there is also a machine learning algorithm for fraud detection. In addition, the SNA module detects interconnected fraud counterparts among claim files. Consequently, the SOBE fraud detection platform allows Anadolu Sigorta to prevent improper payments from claiming participants. SOBE platform, the central fraud detection platform at Anadolu Sigorta, was developed in-house using different technologies and methods, including KNIME Analytics Platform, Python, graph methods, and web service methodologies.

1. Introduction

Insurance is a binding contract between the insurance company and the insurer to protect an asset against uncertain risks. In the insurance industry, fraud is one of the major problems for insurance companies. Insurance fraud may be committed by the policyholder or a third-party insurance policy claim. Fraud claim submission includes damages based on misleading or untruthful circumstances, including exaggeration of how accidents occur. On the other hand, in organized crime rings, car owners or drivers would be recruited to make false reports indicating false occurrences of vehicle accidents. These claims involve property damage or personal injuries as a result of the stated accidents [1-3].

Anadolu Sigorta offers motor and non-motor insurance policies such as health, fire, liability, marine, and car policies. Products coverages include motor policies, consisting of vehicle storage and safekeeping, towing, healthcare assistance, driver's coverage, passengers and those surrounding the vehicle, as well as legal defense expenses and minor repair services. Residential and workplace fire insurance offers services such as legal

* Corresponding Author: fabut@cu.edu.tr

consultancy and medical assistance in case of damages caused by theft, explosion, fire, internal water, vehicle impact, aircraft impact, and natural disasters such as lightning, flood, storm, landslide, and earthquake.

This study aims to develop a fraud detection platform called SOBE to detect individual and organized frauds in the insurance industry. Anadolu Sigorta was using an external program for fraud detection. We decided to develop a new fraud detection platform internally and add more essential components and capabilities for detecting frauds more efficiently. The name "SOBE" comes from the hide and seek game "You're it" in Turkish. SOBE provides significant benefits such as improvement in manual investigation of individual organized fraud cases and automation of claim files in organized fraud suspicion.

The rest of the paper is organized as follows. First, the details of the proposed SOBE platform are given. Then, the methodology and the results are presented. Finally, the paper is concluded along with future directions.

2. Related Works

Fraud detection is one of the main and hot topics in the insurance industry. It is an open issue for new R&D fields and innovation ideas. In order to understand this issue, we need to discover major problems. The main purpose is to deceive insurance companies into paying false claims [4, 5].

In related studies in the literature, some techniques have already been proposed for detecting false claims in the insurance domain. For example, Sumalatha and Prabha [6] presented a system for collecting and analyzing insurance data, including current and past insurance claims, hospital records, patient data, and offered a single platform for checking and providing suspicious claims using Logistic Regression. Sowah et al. [7] proposed Genetic Support Vector Machine (SVM)-based models using the National Health Insurance Scheme claims dataset obtained from hospitals in Ghana to detect health insurance fraud and other anomalies. Kalwihura and Logeswaren [8] introduced a data pre-processing technique, particularly a fraud behavior feature engineering approach, to prevent fraud in the auto insurance industry. Gomes et al. [9] proposed a novel deep learning methodology to gain pragmatic insights into the behavior of an insured person with the help of a new unsupervised variable importance methodology. Severino and Peng [10] evaluated fraud prediction in property insurance claims using various machine learning models based on real-world data from a major Brazilian insurance company. Rukhsar et al. [11] conducted a comparative analysis on various classification algorithms, namely SVM, Random Forest (RF), Multilayer Perceptron (MLP), Decision Tree (DT), Adaboost, K-Nearest Neighbor (KNN), Linear Regression (LR), and Naïve Bayes (NB) to detect the insurance fraud. The performance of the classifiers has been evaluated based on precision, recall, and F1-Score metrics. Despite these studies, however, the domain of organized fake claims in insurance fraud detection has not been sufficiently investigated in the literature.

Rulesets in fraud detection systems are widely used. As mentioned in [12], rule generation and selection methodologies give a solid idea of how to effectively deal with fraud detection across industry boundaries, including applications in insurance fraud, credit card fraud, healthcare fraud, telecommunications fraud, and more. A genetic algorithm is another method of artificial intelligence that was designed so that each individual represents a possible behavioral model. This approach increases the detection rate and decreases the low false alarm rate [13].

If an applicant files a claim, the insurance company will perform various checks to flag the claim as suspicious or nonsuspicious. When the claim is considered suspicious, the insurance firm will first decide whether it's worthwhile to pursue the investigation. Obviously, this will also depend on the amount of the claim, such that small amounts of claims are most likely not further considered, even if they are fraudulent. When the claim is considered worthwhile to investigate, the firm might start a legal procedure resulting in a court judgment and/or legal settlement flagging the claim as fraudulent or not. Also, this procedure is not 100 percent error-proof, and thus nonfraudulent claims might end up being flagged as fraudulent or vice versa [14].

Our study aims to internalize traditional fraud detection processes in auto and nonauto branches. In addition to traditional methods like fraud detection rulesets and ML prediction, new features like text processing from expert reports, network analysis from organized fraud detection, and run-time anomaly detection alerts are additional key features.

3. Proposed "SOBE" Platform

The lifecycle of the SOBE project consists of six stages. The first stage is to get the data from the Anadolu Sigorta source system. We transformed, organized, and stabilized the data for the next stages via web service configuration. We used REST API methodology and KNIME Analytics Platform transformation capabilities for the stage. The input data includes 567 and 543 parameters for motor and non-motor claim files, respectively, which contain both categorical and quantitative parameters. We followed the general steps of Exploratory Data Analysis techniques. We employed data transfigurations such as data formatting, constant column filtering, correlation filtering, missing value, and outlier handling methods for different

variables. Also, we used one hot encoding method to transform some categorical variables into numerical ones. For outlier handling, we used Inter Quantile Range methodology. For normalizing the data, we used both min-max scaling and Z-score normalization according to the data. Also, we generated new variables using the existing parameters to group the data into meaningful categorical variables. We used the cardinality of the historical data, K-means claustration, and Silhouette Coefficient to determine the groups. We applied the weight of evidence methodology to generalize the quantitative parameters. We utilized forward and backward elimination for feature selection, and for dimension reduction, we preferred PCA techniques.

The second stage is the Social Network Analysis (SNA). This stage is running only for motor claims. We utilized graph theory to develop the capability to reveal claim participant/claimant relationships with insurance blacklists or assets on created claim files. We mainly employed this stage to discover organized crime relations on our full set of data by adding external data sources. We selected the participant's Turkish Identification Number as the main entity for constructing the network according to the regulations legally to protect the anonymity of the participants. We characterized networked structures in terms of nodes (i.e., individual actors, people within the network) and the ties, edges, or links (relationships or interactions and assets) that connect them. Finally, we used the shortest path algorithm to determine the direct path between the blacklisted people and the participants of the claim files.

The third stage is our machine learning model to predict the fraud potential for the claim file itself. Again, this stage is running only for motor claims. We employed two different machine learning algorithms, including RF and MLP. We examined the two model outputs and decided to continue with the RF algorithm. Currently, RF has enough accuracy and speed to predict the fraud potential of the claim file.

The fourth stage is the claim search history. We collect the history of a claim's participants and subjected asset history at this stage. In addition to these, we also search if any aspect of the claim, participant, or the asset, ever occurred in another fraud case. The fifth stage is the rule engine stage which includes business rules determined by our claims departments according to their experience in the field. We have 600 rules run for every claim. Every rule has a score given by our claims department. These are some of the examples of the mentioned rules: "Claim city is different from customer city", "Claim party has other policies in force for the same vehicle with the same or other insurance company at the time of the event", "There is more than five days between the date of the claim occurred and the date of the report". After all the stages, we calculate and send a file fraud score to our claim platform within 5 to 10 seconds.

The SOBE fraud detection lifecycle starts with a web service request from Claim Management System (CMS). As soon as the claim file is reported, CMS automatically sends fraud scoring requests at certain checkpoints. This request is sent to the KNIME server via Rest API and includes individual claim file details. KNIME server architecture accepts only JSON files on REST calls, so the requests and SOBE responses are sent in JSON format. To prevent load latency, a certain number of jobs are pooled on the KNIME server. So, when a request is sent from the CMS, the SOBE workflow is immediately executed in the server. Figure 1 illustrates external data integration via web services.

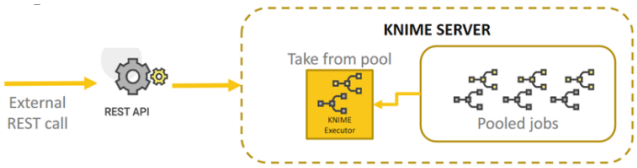


Figure 1. External data integration via web services

Figure 2 represents the overview of a SOBE workflow. The steps can be explained as follows:

1. Data transformation, standardization, and normalization step: Webservice request contains data from the CMS. Before using this input in the rule engine and ML prediction steps, several data transformations are applied.

2. SNA + Blacklist: The participants of the claim file (mainly policyholders) are searched for any existing connections in an organized fraud scheme. SNA results include Anadolu Sigorta and Insurance Information and Monitoring Center data. In this step, the company blacklist is also used to label participants of the claim file.

3. Machine learning model: The RF-based model for auto claims classifies the claim as fraudulent or not.

4. History search: To check recurring claims of a certain participant in the claim file, a history search is executed in this step (e.g., number of rejected claims for the claimant for motor). The results are used in the rule engine step.

5. Rule engine: The claim file is scored according to active rules. Also, ML model predictions, SNA, and blacklist results are taken into consideration in scoring.

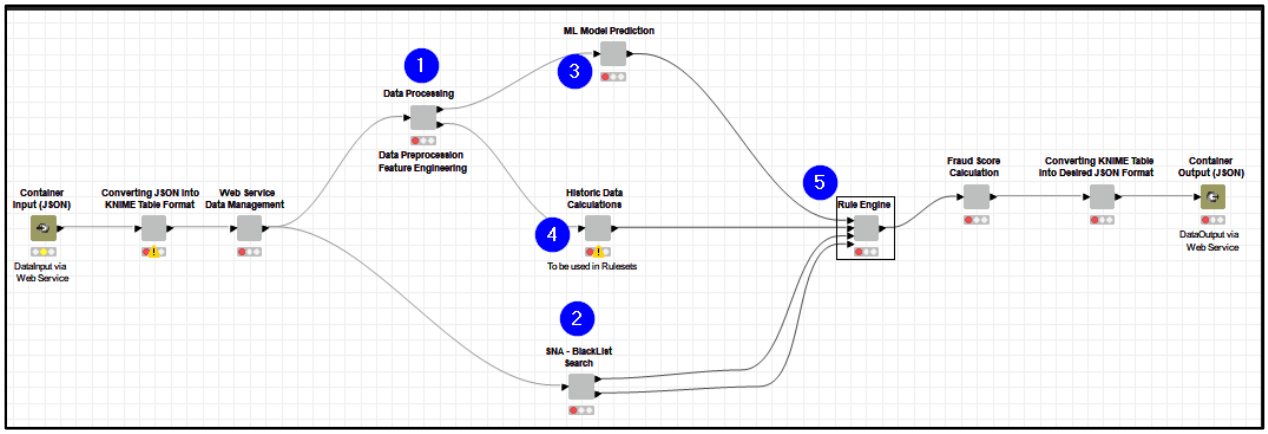


Figure 2. Overview of a "SOBE" workflow

4. Methodology

In insurance, a claim file has several phases called "touchpoint". Those phases contain several steps like file opening, expert opinion, file status changes. At the very beginning, the ML model tries to decide if the claim file has a fraud suspicious or not.

Since the fraud case is a good example of an imbalanced dataset, cleaning and selecting the right data to balance the target variable with the business side's opinions is crucial for modeling. After applying Exploratory Data Analysis (EDA) steps and balancing data as much as possible (i.e., 7% fraud cases in target variable), RF-based and MLP-based models have been created. With hyper-parameter optimizations, each model has been implemented on validation data that was not used during the training phase. For the RF-based model, we generated 100 trees using Gini Index and Information Gain Ratio and 100 different stratified sampled data partitioned via loops. In the MLP-based model build, epsilon is kept in 1e-8, and the number of maximum iterations is limited by 120 with an initial learning rate of 0.001. ReLU is used as the hidden layer activation function. In addition, several numbers of hidden layers ranging from one to six and different numbers of neurons varying between 8 and 35 have been tested during the MLP-based model training.

We evaluated the performance of the two models by calculating the precision, recall, F-Measure, and accuracy values, as defined in Eqs. (1) through (4), respectively, where tp is the ratio of true positives, tn is true negatives, fn shows false negatives, and fp represents false positives. In order to eliminate the overfitting danger while building the model, the 10-fold cross-validation method has been used to evaluate the generalization error of the models.

$$Precision = \frac{tp}{tp+fp} \tag{1}$$

$$Recall = \frac{tp}{tp+fn} \tag{2}$$

$$F - Measure = \frac{2*Precision*Recall}{Precision + Recall} \tag{3}$$

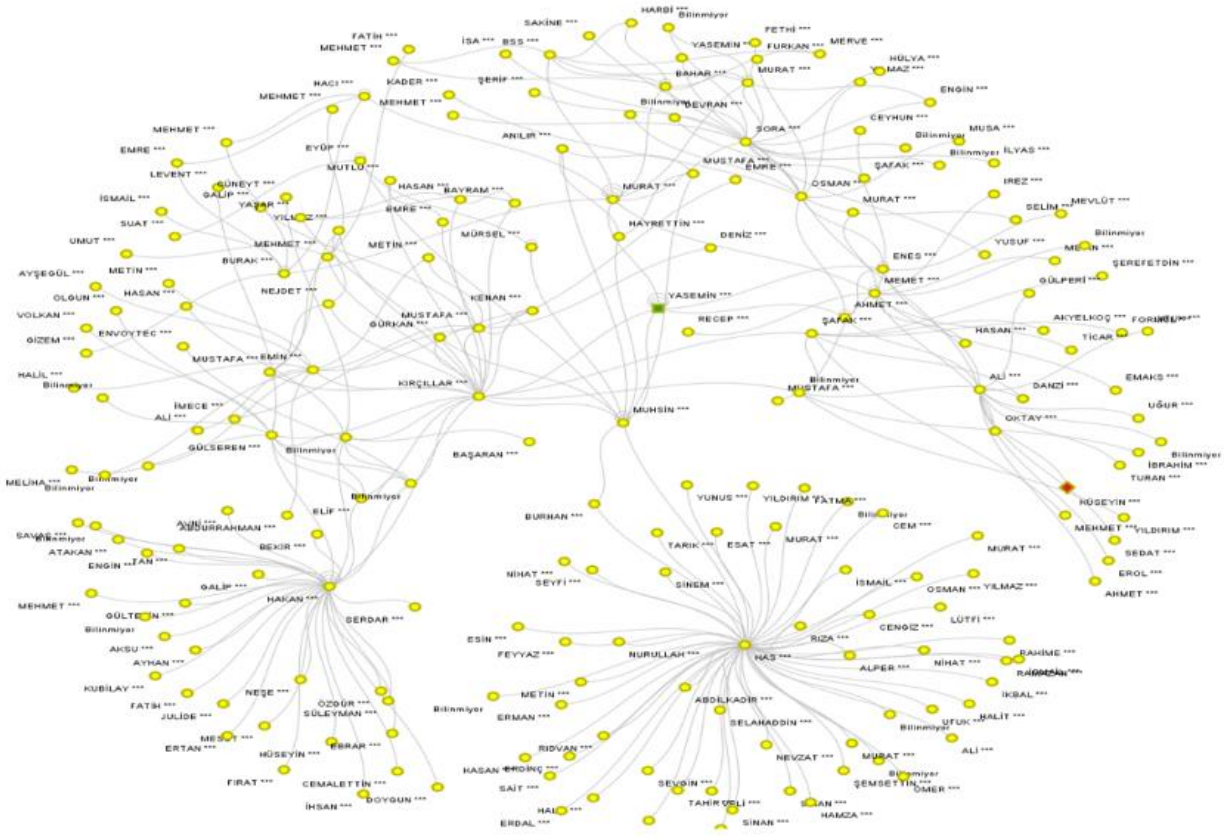
$$Accuracy = \frac{tp+tn}{tp++tn+fp+fn} \tag{4}$$

Organized fraud detection is a different discipline that uses other techniques than traditional ML methods. Using graph theory is one of the solutions to that problem. In that phase of fraud detection, the aim is to detect the relationships of blacklisted persons with claim file participants in a huge network (e.g., a network consisting of 8 million nodes and 12 million connections). The network has been constructed not only with Anadolu Sigorta claim data but with external data sources like accident report files (KKT-Kaza Tespit Tutanağı) and Insurance Information and Monitoring Center (SBM-Sigorta Bilgi Merkezi) insurers data. Network connections have been constructed basically with vehicle chassis numbers and participants' identification numbers.

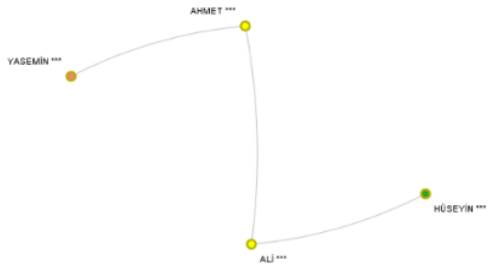
After the network has been established, blacklisted people have been labeled within the network to see the relationships with other participants. In Figure 3, the red-colored node is a blacklisted person, and the green-colored node is the participant being investigated for an organized fraud scenario.

Once the connections have been established, the next challenge is to detect the relationships in terms of closeness, strengths, and organized crime suspicious. The main solution to that problem is finding the shortest path between participants. When the shortest path algorithm has been applied to the network, the shortest path between two participants can be revealed easily. Within daily incoming

claim files, first and second-degree connections of participants with blacklisted people, using the same chassis number among participants and their degrees, and a participant's number of connections are investigated and scored with respect to importance to support the fraud detection system.



(a)



(b)

Figure 3. The output view of the SNA module

5. Results

The performance of the models has been confirmed by applying the cross-validation step for model generalization and calculating the precision and recall values. The priority is to keep recall as high as possible while also keeping precision at an acceptable rate. Considering this priority, the RF-based model produced precision, recall, F-measure, and accuracy values of 0.274, 0.374, 0.316, and 95.83%, respectively, whereas the MLP-based model yielded precision, recall, F-measure, and accuracy values of 0.537, 0.298, 0.383, and 97.51% respectively.

SOBE platform has replaced the purchased fraud application, and it is the central application used by the Anadolu Sigorta Claim department. The platform handles over 2500 claim cases daily and predicts the probability of fraud for each file in less than 10 seconds. Integrated module

architecture makes the platform simple and manageable for both business and technology-wise.

All the fraud detection platform stages, the rule set determination, score prediction with a machine learning algorithm, and SNA results give the fraud score of the claim record. Our operational process is triggered based on this score, and further steps run until the file is closed.

The process of developing the SOBE platform has resulted in various gains. The benefits of the application can be outlined as follows:

- We tailor our own fraud platform from the beginning of the process: data preparation, data enrichment, quality study, rule definitions, and SNA platform integrated file investigation. This is the first platform ever developed in the insurance sector.
- The purchased platform has been eliminated, and there is no subscription/maintenance cost anymore.
- Our claims department can determine faster for the claim investigation on a suspicious file.
- The system is in-house, so implementation, maintenance, and development of the platform can be done easier.
- Platform health check process will be done internally, and there is no dependency on a vendor company.

The SOBE platform went live in Sept 2021, and its efficiency has been compared to our previous platform based on quarter 4 (Q4) of the years 2020 and 2021. Table 1 shows the comparison results.

Table 1. Comparing the SOBE and our previous platforms

	Previous Platform	**SOBE Platform**
Total savings	2020 Q4: 8.3 million TL	2021 Q4: 12.8 million TL
Fraud detection ratio	2020 Q4: 8.36%	2021 Q4: 12.69%
SNA output	n/a	5 organized fraud rings
SNA savings	n/a	2.2 million TL

6. Conclusion and Future Work

This study proposed SOBE, a fraud detection platform that allows Anadolu Sigorta to prevent improper payments from claiming participants. SOBE improves manual investigation of individual organized fraud cases and automates processing claim files in organized fraud suspicion. We managed to perform a detailed analysis and determine if a claim has fraud risk or not throughout a claim lifecycle.

In the following steps, the study can be extended in multiple ways. SOBE can also be extended to consider anomaly detection scenarios. Machine learning algorithms can be developed for non-motor claims. The collection of historical data could be expanded by adding external data sources. From all kinds of reports in the CMS, keyword extraction and expert fraud opinions can be extracted from the reports. Also, additional ML models could be implemented in critical touchpoints to detect fraud cases more precisely with the enriched data.

Declaration of Ethical Standards

The authors of this article declare that the materials and methods used in this study do not require ethical committee permission and/or legal-special permission.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Ribeiro R., Silva B., Pimenta C., & Poeschl G., 2020. Why do consumers perpetrate fraudulent behaviors in insurance?. Crime, Law and Social Change, **73**(3), pp. 249-273.

[2] Abdallah A., Maarof M. A., & Zainal A., 2016. Fraud detection system: A survey. Journal of Network and Computer Applications, **68**, pp. 90-113.

[3] Hargreaves C. A., & Singhania V., 2015. Analytics for Insurance Fraud Detection: An Empirical Study. American Journal of Mobile Systems, Applications and Services, **1**(3), pp. 223-232.

[4] Liu X., Yang J. B., Xu D. L., 2020. Fraud detection in automobile insurance claims: a statistical review. In: Developments of Artificial Intelligence Technologies in Computation and Robotics: Proceedings of the 14th International FLINS Conference, pp. 1003-1012.

[5] Patil K. S., Godbole A., 2018. A survey on machine learning techniques for insurance fraud prediction. Helix, **8**(6), pp. 4358-4363.

[6] Sumalatha M. R., Prabha M., 2019. Mediclaim fraud detection and management using predictive analytics. In: Proc. of Intl. Conference on Computational Intelligence and Knowledge Economy, pp. 517-522.

[7] Sowah R. A., Kuuboore M., Ofoli A., Kwofie S., Asiedu L., Koumadi K. M., Apeadu K. O., 2019. Decision support system for fraud detection in health insurance claims using genetic support vector machines. Journal of Engineering, Article ID 1432597.

[8] Kalwihura J. S., Logeswaran R., 2020. Auto-insurance fraud detection: a behavioral feature engineering approach. Journal of critical reviews, **7**(3), pp. 125-129.

[9] Gomes C., Jin Z., Yang H., 2021. Insurance fraud detection with unsupervised deep learning. Journal of Risk and Insurance, **88**, pp. 591–624.

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

[11] Rukhsar L., Bangyal W. H., Nisar K., & Nisar S., 2022. Prediction of insurance fraud detection using machine learning algorithms. Mehran University Research Journal of Engineering & Technology, **41**(1), pp. 33-40.

[12] Baesens B., Van Vlasselaer V., Verbeke W., 2015. Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques. John Wiley & Sons, Inc

[13] Katoch S., Chauhan S.S. & Kumar V., 2021. A review on genetic algorithm: past, present, and future. Multimedia Tools and Applications, **80**, pp. 8091-8126.

[14] Dokas P., Ertoz L., Kumar V., Lazarevic A., Srivastava J., & Tan P. N., 2002. Data mining for network intrusion detection. In: Proc. of NSF Workshop on Next Generation Data Mining (pp. 21-30).