Techno-Science

Scientific Journal of Mehmet Akif Ersoy University

www.dergipark.gov.tr/sjmakeu

Original Research Article

hosted by Turkish JournalPark ACADEMIC

IMPLEMENTATION OF XGBOOST METHOD FOR HEALTHCARE FRAUD DETECTION

Elvan DUMAN[1*]

[1] *Department of Software Engineering, Burdur Mehmet Akif Ersoy University, Turkey*

ARTICLE INFO

Article History

Received	:	22/12/2022
Revised	:	31/12/2022
Accepted	:	31/12/2022
Available online	:	31/12/2022

Keywords

Fraud detection, healthcare fraud, XGBoost, anomaly detection

ABSTRACT

The health care systems are quickly adapting digital health records, which will exponentially increase the quantity of medical data. The systems are generally faced with unsustainable costs and large volumes of electronic medical data. Therefore, more efficient research, practices, and real-world applications are needed to take advantage of all benefits of medical data. One strategy to cut back on the rising costs is the detection of fraud. In this paper, XGBoost, which is an implementation of gradient-boosted decision trees, was employed, along with supervised algorithms to include Random Forest, Logistic regression, and decision trees. The List of Excluded Individuals/Entities (LEIE) database, which contains excluded providers' information, was used to label as a fraud in the Medicare Part B dataset. Thus, the data has become available for use with supervised methods. According to the experimental results, the XGBoost algorithm outperformed traditional machine learning algorithms in terms of performance.

1. INTRODUCTION

As healthcare is a significant and rapidly expanding industry, it is essential to carefully plan for its growth. This industry provides a vital service to a growing number of facilities, and professionals, and continues to be a significant contributor to overall economic growth. Meanwhile, healthcare costs hold an important place in the total expenditures of both people and countries. More recently, National Health Expenditure (NHE) in the United States grew from 9.7% to 4.1 trillion dollars in 2020 [1]. NHE is projected to increase by an average annual rate of 5.4% and reach $6.2 trillion in the years 2019 to 2028 [1]. Due to the health-related high costs, many patients receive insufficient medical care. In order to improve this negative situation, governments cover the high costs of needed health services through the funds they have established as a necessity of the social state. However, these programs face many problems such as waste, abuse, and fraud that exploit the program's utility. The median loss for these offenses was 1,002,407 dollars for 2021 in the USA according to United States sentencing commission reports [1]. Being one of the most serious problems in the field, Fraud generally has the following characteristics:

- Submitting false claims intentionally in order to benefit from the Federal health care payment
- Deliberately offering remuneration to encourage referrals for items refunded by healthcare service programs
- Making unauthorized referrals for designated health services.

The volume of electronic healthcare data keeps growing dramatically as a result of the rapid increase in the use of internet technologies that make it possible to transmit and store large amounts of data. However, the techniques that enable the processing and analyzing of data cannot keep up with this rapid increase in data. The existence of fraud in healthcare is predominantly revealed by manual investigations by investigators and auditors. Because of the enormous amounts of data and complexity of processes, the detection of suspicious transactions with manual efforts can be time-consuming and extremely ineffective. Considering the size of the data that needs to be examined, it can be said that it is not possible to analyze whole data even in an inefficient way. For this reason, the method of detecting suspicious behavior with computer-aided systems using artificial intelligence techniques should be considered [2].

[*] Corresponding Author: eduman@mehmetakif.edu.tr

To cite this article: Duman E., (2022). Implementation of Xgboost Method for Healthcare Fraud Detection, Techno-Science, vol. 5, no. 2, p. 69-75

Machine learning and Deep learning methods can perform “Big data” which refers to extremely large, unstructured, and complex datasets that are beyond the ability of conventional data processing methods to store, transfer, analyze, and visualize in a timely and economical manner. In the healthcare system of the USA, Big datasets of Medicare are released by The Centers for Medicare and Medicaid Services (CMS). The collections include all transactions submitted by doctors and physicians in the USA. Each year, all records of the previous year are published by CMS. The datasets are publicly available and give information in similar formats. CMS datasets do not contain any judgments regarding records, such as abuse and fraud. Therefore, the List of Excluded Individuals and Entities (LEIE) dataset, which contains the fraud records published by the Office of the Inspector General, is used to label the data.

There have been several promising studies related to healthcare fraud detection. The studies can be grouped in terms of methods and datasets used. Considering the datasets used, the vast majority of the studies into Medicare insurance fraud have been conducted by using Part B dataset of CMS[3-9]. Mattgew Herland et al [10] created the new dataset which is combination of the datasets, Medicare Provider Utilization and Payment Data : Part B, Medicare Provider Utilization and Payment Data : Part D, Medicare Provider Utilization and Payment Data: Physician and Other Supplier (PUF) Refering Durable Medical Equipment, Prosthetics, Orthotics and Supplies (DMEPOS). In order to assess the performance of the methods, the real world fraud labels were used, which are from LEIE published by the Office of the Inspector General. In the study, Logistic regression, Gradient Boosted Trees, and Random Forest, which are supervised methods, are applied to the combined dataset. Logistic Regression showed the best performance according to AUC (Area under the Receiver Operating Characteristic) performance.

In another study [11], a Graph analytics framework was proposed for estimating healthcare fraud risk. Predictive variables were derived from three datasets. The first is PUF data for years of 2012, 2013, and 2014. The second is Part-D data for 2013 published by CMS and last one is LEIE dataset, which was used for obtaining the list of providers excluded from participation in Medicare. And presencing on the list was referred as exclusion and used for fraud label. National Provider Identifiers (NPIs) are used for identifying excluded providers in the datasets. However, the study emphasizes that only 5% of the excluded providers have National Provider Identifiers (NPIs) in LEIE dataset. Therefore, an identity-matching procedure was implemented to match excluded providers in the LEIE. Because of imbalanced data, the authors preferred to use randomly selected 12.000 non-excluded samples for equal class distribution in the study. However, the study was conducted with a limited dataset in this way. Sadiq et al. [12] have performed experiments on the CMS Part B, Part D, and DMEPOS datasets. However, the study focused on the state of Florida data only. One of the reasons why the authors chose the state of Florida is that medical expenses may differ from state to state. The second reason is that Florida has suffered severely from high-profile fraudulent malpractices. A novel fraudulent medical insurance claim detection framework based on Patient Rule Induction Method (PRIM) was proposed. PRIM had never been used in healthcare fraud detection problems before the study.

Although a significant amount of research has been done on fraud detection in healthcare programs, there are still some issues that need to be researched further to improve the performance of fraud detection. One of the problems is that it is not known exactly which of the recordings are fraudulent because Medicare datasets do not have any labels that indicate fraud. In Healthcare fraud studies, the list of doctors, physicians, and other healthcare workers who have been excluded from healthcare programs permanently or for a certain period due to fraud can be obtained from the published LEIE database. However, there may be many records that have not been identified by the regulators and have not been labeled as fraudulent. Moreover, National Provider Identifier (NPI) numbers used to match fraudulent records in CMS databases are often missing.

The second problem is that CMS databases are regarded as being extremely imbalanced, with less than 0.01% of instances being labeled as fraudulent. Unbalanced data naturally cause poor performance or over-learning of methods. In this study, first, we propose a XGBoost-based method for the imbalance data problem by preserving the distribution of labeled data. Second, we combined CMS datasets for the years 2013 through 2020 and handled the labeling data process uniquely.

2. THEORETICAL BACKGROUND

2.1. Datasets

The database, Provider Utilization and Payment Data Physician and Other Supplier Public Use File ,is maintained annually by CMS. The datasets consist of information about utilization, submitted charges, payment, and reimbursement of services provided by doctors, physicians and other medical workers. The data in the PUF dataset are final decisions after all objections are concluded and cover 2013 through 2020 [13]. The NPI for the performing provider is 10 digit identification number, the Healthcare Common Procedure Coding System (HCPCS) which is a group of codes that indicates operations, supplies, and services furnished to healthcare professionals, and the place of service give the spending and utilization information. Numerous health records with various HCPCS codes may have been billed and service provided for a given NPI because different price schedules apply depending on whether the location of service listed on the claim is a facility or not, data have been aggregated based on the place of service. The medical provider’s specialty is given as provider type in the

Dataset. Other information covers charges and average payments. Table 1 illustrates an example of a physician with a unique identifier (NPI) of 1003000126, which was sampled from the Part B dataset for the year 2020

Table 1. A selection of data from the Part B dataset

NPI	Last Name	First Name	Provider Type	HCPCS	Tot_Srvcs	Avg_Sbmtd_Chrg	...
1003000126	Cibull	Thomas	Pathology	88304	125.0	115.0	...
1003000126	Cibull	Thomas	Pathology	88305	4307.0	170.0	...
1003000126	Cibull	Thomas	Pathology	88312	346.0	88.0	...
1003000126	Cibull	Thomas	Pathology	88313	109.0	68.0	...
1003000126	Cibull	Thomas	Pathology	88341	122.0	92.0	...

The Office of the Inspector General's List of Excluded Individuals released monthly and referred to as OIG Exclusion List, informs the healthcare sector about people and organizations that are prohibited from taking part in Medicare, Medicaid, and all other National healthcare programs. The effect of exclusion is that no payment will be made by any Federal health care program for any items or services furnished, ordered or prescribed by an excluded individual or entity. Individuals and entities who have been reinstated are removed from the dataset. Physicians are excluded for a number of reasons. Patient abuse or neglect, theft, and felony convictions relating to unlawful manufacture, distribution, prescription, or dispensing of controlled substances are in the category of mandatory exclusions. Some of the misdemeanor behaviors such as defaulting on health education loan; suspension, revocation, or surrender of a license to provide health care for reasons bearing on professional competence, professional performance, or financial integrity are classified as permissive exclusions. Mandatory exclusions, their codes, descriptions, and exclusion periods are given in Table 2.

Table 2. LEIE Mandatory Exclusions List

Exclusion code	Exclusion Detail	Period
1128(a)(1)	Conviction of program-related crimes	5 years
1128(a)(2)	Conviction relating to patient abuse or neglect	5 years
1128(a)(3)	Felony conviction relating to healthcare fraud	5 years
1128(a)(4)	Felony conviction relating to controlled substance	5 years
1128(c)(3)(G)(i)	Conviction of second mandatory exclusion offense	10 years
1128(c)(3)(G)(ii)	Conviction of third or more mandatory exclusion offenses	Permanent

A very beneficial study has been conducted with striking results by examining the LEIE dataset [14]. According to the study, 2222 physicians were excluded from Medicare insurance programs for the given periods listed top or permanently. Fraud, abuse, and crime exclusions of physicians increased by 20% per year, on average, between 2007 and 2017. While the exclusion rates were highest in West Virginia with 5.77 exclusions per 1000 physicians, no exclusion was observed in Montana. Considering the characteristics of excluded physicians, they are more likely to be male, physicians with osteopathic training, and older. They also were gathered in specific specialties such as internal medicine, surgery, psychiatry, and family medicine.

The LEIE dataset contains the following details: the cause of exclusion, the date of exclusion, and the date of reinstatement or waiver. Physicians who are on the list are excluded from healthcare services in the United States during the period of punishment. We detect all physicians who were excluded from Medicare Insurance Program in the United States frauds in the Provider Utilization and Payment Data Physician and Other Supplier Public Use File dataset using the LEIE dataset. To obtain excluded individuals and entities in the PUF dataset, national provider identifiers were used to match between LEIE and PUF datasets.

3. METHOD

Machine learning algorithms can learn from the data they are presented with and use that learning to make predictions about similar data. Machine learning algorithms can be applied to various problems that involve detecting anomalies, including cyber attack detection [15], fraud detection [16], email filtering [17], and others. This machine learning technique involves using input and output variables, and an algorithm (called a mapping function) to map the inputs to the output. This technique is useful because once it has been trained on a dataset, the mapping function can be used to easily classify new data or filter it. In other words, the machine learning model has learned how to map the input data to the output data and can use this knowledge to make predictions or decisions about new data.

Boosting is one of the machine learning algorithms that is used to improve the accuracy of a model by reducing the bias and variance of the data it is trained on. This is often done by combining the predictions of multiple weak models to create a more accurate overall prediction. Classification and regression are two supervised learning problems that can benefit from the usage of boosting methods. Boosting methods involve combining the predictions of multiple weak models to create a stronger overall prediction.

In ensemble learning methods like boosting, decision trees are frequently employed as the basic model. Boosting methods create a tree-like model of decisions based on the features of the data. The internal nodes of the tree represent decisions based on specific features, and the leaf nodes represent the predicted class or value. These models are easy to understand and interpret, and they can work with both categorical and numerical data. C4.5 and random forest (RF) are two different types of decision tree algorithms. C4.5 is known for its ability to handle enormous datasets and missing values in the data [18]. RF [19] involves creating multiple decision trees and using their predictions to make a final prediction. It does this by training each decision tree on a different subset of the data and then averaging the predictions of all the trees to make a final prediction. RF is effective at handling large datasets and can still make accurate predictions even if the individual decision trees are not highly accurate.

Gradient boosting is a method used for supervised learning tasks such as classification, regression, and ranking. XGBoost (Extreme Gradient Boosting) is a particular implementation of gradient boosting that is a widely used implementation of the gradient boosting method, particularly in machine learning competitions. It is renowned for its high performance and efficient implementation on multicore and distributed machines. XGBoost can be used for various supervised learning tasks, such as classification, regression, and ranking, and is particularly effective at handling large datasets and working with both dense and sparse data.

The XGBoost algorithm consists of multiple base classifiers that can be chosen from a variety of options including decision trees, KNN, SVM, logistic regression, and others. These base classifiers are then combined linearly to improve the overall performance of the algorithm. In supervised learning, we typically create both an objective function and a prediction function. The objective function is used to learn the necessary parameters by minimizing it during training. The prediction function, along with the learned parameters, is then used to classify or make predictions on new, unseen data.

In XGBoost, the model can be represented as follows for a dataset D containing *n* samples with *m* dimensions each:

$$D = \{(x_i, y_i)\}\ (x_i\ \in\ R^m, y_i\ \in R, i = 1,2,3, \dots, n) \tag{1}$$

when building an XGBoost model, we aim to find the optimal parameters that minimize the objective function. The objective function in XGBoost consists of two components: an error function term L and a model complexity term Ω. The objective function can be expressed as:

$$Obj(\Theta) = L(\Theta) + \Omega(\Theta) \tag{2}$$

$$L = \sum_{i=1}^{n} (y_i - \widehat{y_i})^{\ 2} \tag{3}$$

$$\Omega = {}_{\gamma}T + \frac{1}{2}\,\lambda\, \sum_{i=1}^{T} w_j^2 \tag{4}$$

The objective function in XGBoost consists of two regularization terms L1 and L2. These regularization terms help to prevent overfitting by introducing a penalty on the complexity of the model.

To optimize the model during training, we need to modify the objective function without changing the original model. One way to do this is by adding a new function *f* to the model that reduces the objective function as much as possible. The modified objective function can be expressed as:

$$Obj^{(t)} = \sum_{i=1}^{n} (y_i - (\widehat{y_i}^{(t-1)} + f_t(x_i)))^2 + \Omega \tag{5}$$

Finding the ideal values for the model parameters and f that minimize the objective function is the aim of training. The use of optimization algorithms like gradient descent or stochastic gradient descent can achieve this. We can increase the model's generalization performance and prediction accuracy on omitted data by minimizing the objective function. Additional information about the notations can be located at [19].

4. EXPERIMENTS

4.1. Data Processing

We combine the information for Part B datasets from 2013 to 2020. Then, we removed all instances which have missing important values such as National Provider Identifier (NPI). To encode a categorical string column of a DataFrame to a column of numerical indices for provider gender and provider type, the StringIndexer function provided by the pandas library was used.

In this study, we utilized the LEIE (List of Excluded Individuals/Entities) dataset, maintained by the United States Department of Health and Human Services Office of Inspector General, to generate fraud labels for physicians. Specifically, physicians listed in the LEIE dataset were classified as fraudulent, while those not listed were classified as nonfraudulent.

To ensure reliable and accurate matching between the Medicare datasets and the LEIE, we found that using the NPI value was the most reliable method. By using the NPI value, we were able to obtain exact matches between the two datasets. However, records with the same name, surname, and field of work as records without an NPI value in the LEIE dataset were identified as suspicious and removed from the dataset. In addition, we labeled the penalty period given in violation transactions as fraudulent.

4.2. Experiments

To process and evaluate our models, we utilized the Spark platform on Google Colab Pro, which is well-suited for handling large datasets. We applied three classification algorithms: logistic regression, gradient-boosted trees, random forests, and XGBoost. In our model evaluation, we employed stratified k-fold cross-validation, where k was set to 5, to assess the performance of our models. The data was divided into 5 folds and the model was trained and evaluated 5 times. This process allowed for a more robust estimate of the performance of the methods.

In this study, we addressed the problem of identifying fraudulent activity among Medicare providers, which can be framed as a binary classification task. We considered fraud to be the positive class and non-fraud to be the negative class. Using the Spark platform, we generated confusion matrices for each of our models, which are frequently used to evaluate the accuracy of machine learning algorithms. These matrices compare the actual and predicted classifications. To quantify the performance of our fraud detection approach, we employed the Area Under the Curve (AUC) metric [20]. It is a measure used to assess the effectiveness of binary classification models. It evaluates the model's capacity to distinguish between the two classes. Recall is a metric used particularly in the context of imbalanced datasets where one class is more prevalent than the other. It is defined as Recall=TP/(TP+FN). True positive (TP) is the number of times the model correctly predicted that an instance was positive. True negative (TN) is the number of times the model correctly predicted that an instance was negative. False positive (FP) is the number of times the model incorrectly predicted that an instance was positive when it was actually negative. False negative (FN) is the number of times the model incorrectly predicted that an instance was negative when it was actually positive. These definitions can be calculated from a confusion matrix, which is a table that summarizes the performance of a classification model.

5. RESULTS AND DISCUSSION

This section presents the findings of our investigation into the effectiveness of machine learning algorithms for detecting Medicare fraud. We evaluated the performance of various datasets and learners to determine their suitability for this task. Table 3 illustrates the AUC (Area Under the Curve) scores for the combination of the Part B dataset with various machine learning algorithms. The highest AUC scores are indicated in italics, while the highest scores for each learner are underlined. These results demonstrate the performance of the Part B dataset and different learners in terms of their ability to accurately detect Medicare fraud.

Table 3. Average performance results for fraud detection

Methods	AUC	Precision	Recall	F1
Logistic Regression	77.362	0.78	0.77	0.77
Random Forest	78.690	0.84	0.71	0.80
Decision Trees	79.186	0.85	0.79	0.79
XGBoost	<u>84.563</u>	<u>0.86</u>	<u>0.83</u>	<u>0.84</u>

Based on the Table 3, it appears that the XGBoost method has the highest AUC, with a value of 84.563. It also has the highest precision and recall values, with 0.86 and 0.83 respectively. The F1 score for XGBoost is also the highest, at 0.84. This suggests that the XGBoost method is the most effective of the four methods in terms of these performance measures. The

Logistic Regression method has the lowest AUC, with a value of 77.362. It also has the lowest precision value, at 0.78. However, it has a relativey high recall value of 0.77 and an F1 score of 0.77. The Random Forest and Decision Trees methods have intermediate performance, with AUC values of 78.690 and 79.186 respectively. The Random Forest method has a higher precision value of 0.84, but a lower recall value of 0.71, resulting in a lower F1 score of 0.80. The Decision Trees method has a precision and recall value of 79 and an F1 score of 79.

Overall, the XGBoost method appears to be the most effective of the four methods in terms of these performance measures. However, it is important to consider the specific context and goals of the model when determining the most appropriate method to use. The confusion matrix of XGBoost method is given in Fig. 1.

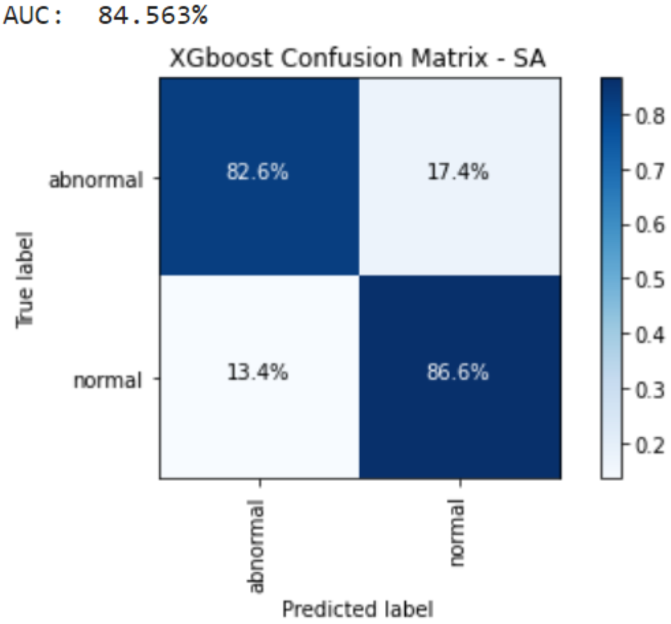

Fig.1. Confusion Matrix of XGBoost method

6. CONCLUSION

The Medicare program is a healthcare initiative that serves over 60 million Americans. It is believed that Medicare experiences significant financial losses each year due to fraud, waste, and abuse, which could range from $20 to $70 billion. This not only impacts taxpayers but also puts the welfare of beneficiaries at risk. To address this issue and improve transparency, the Centers for Medicare and Medicaid Services (CMS) has made several Medicare data sets accessible to the public. In this research, we explored the effectiveness of XGBoost and traditional machine learning algorithms in classifying an imbalanced dataset of Medicare claims data to detect Medicare fraud.

REFERENCES

[1]. Hartman, M., et al. (2022). "National Health Care Spending In 2020: Growth Driven By Federal Spending In Response To The COVID-19 Pandemic: National Health Expenditures study examines US health care spending in 2020." Health Affairs 41(1): 13-25.

[2] Loh, H. W., et al. (2022). "Application of Explainable Artificial Intelligence for Healthcare: A Systematic Review of the Last Decade (2011–2022)." Computer Methods and Programs in Biomedicine: 107161.

[3] Bauder, R. A. and T. M. Khoshgoftaar (2016). A novel method for fraudulent medicare claims detection from expected payment deviations (application paper). Information Reuse and Integration (IRI), 2016 IEEE 17th International Conference on, IEEE.

[4] Bauder, R. A., et al. (2016). Predicting medical provider specialties to detect anomalous insurance claims. Tools with Artificial Intelligence (ICTAI), 2016 IEEE 28th International Conference on, IEEE.

[5] Chandola, V., et al. (2013). Knowledge discovery from massive healthcare claims data. Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM.

[6] Bauder, R. A. and T. M. Khoshgoftaar (2018). The detection of medicare fraud using machine learning methods with excluded provider labels. The Thirty-First International Flairs Conference.

[7] Bauder, R. A. and T. M. Khoshgoftaar (2016). A probabilistic programming approach for outlier detection in healthcare claims. Machine Learning and Applications (ICMLA), 2016 15th IEEE International Conference on, IEEE.

[8] Khurjekar, N., et al. (2015). Detection of Fraudulent Claims Using Hierarchical Cluster Analysis. IIE Annual Conference. Proceedings, Institute of Industrial and Systems Engineers (IISE).

[9] Herland, M., et al. (2017). Medical provider specialty predictions for the detection of anomalous medicare insurance claims. 2017 IEEE International Conference on Information Reuse and Integration (IRI), IEEE.

[10] Herland, M., et al. (2018). "Big Data fraud detection using multiple medicare data sources." Journal of Big Data 5(1): 29.

[11] Branting, L. K., et al. (2016). Graph analytics for healthcare fraud risk estimation. Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, IEEE Press.

[12]. Sadiq, S., et al. (2017). Mining Anomalies in Medicare Big Data Using Patient Rule Induction Method. Multimedia Big Data (BigMM), 2017 IEEE Third International Conference on, IEEE.

[13] Data, P. (2014). "Medicare Fee-For Service Provider Utilization & Payment Data Physician and Other Supplier Public Use File: A Methodological Overview."

[14] Chen, A., et al. (2018). "Characteristics of Physicians Excluded From US Medicare and State Public Insurance Programs for Fraud, Health Crimes, or Unlawful Prescribing of Controlled Substances." JAMA network open 1(8): e185805-e185805.

[15] Söğüt, E., et al. (2021). "Detecting Different Types of Distributed Denial of Service Attacks." Gazi University Journal of Science Part C: Design and Technology 9(1): 12-25.

[16] Duman, E. A. and Ş. Sağıroğlu (2017). Heath care fraud detection methods and new approaches. 2017 International Conference on Computer Science and Engineering (UBMK), IEEE.

[17] Gangavarapu, T., et al. (2020). "Applicability of machine learning in spam and phishing email filtering: review and approaches." Artificial Intelligence Review 53(7): 5019-5081.

[18] Quinlan, J. R. (2014). C4. 5: programs for machine learning, Elsevier.

[19] Belgiu, M. and L. Drăguţ (2016). "Random forest in remote sensing: A review of applications and future directions." ISPRS journal of photogrammetry and remote sensing 114: 24-31.

[20] Jaskowiak, P. A., et al. (2022). "The area under the ROC curve as a measure of clustering quality." Data Mining and Knowledge Discovery 36(3): 1219-1245.

Techno-Science *Paper ID: 1223234*