Novelty Journals

ISSN 2394-7314

International Journal of Novel Research in Computer Science and Software Engineering
Vol. 9, Issue 3, pp: (16-23), Month: September - December 2022, Available at: www.noveltyjournals.com

Transaction fraud detector using KNN in deep learning

Ahmad Alammar[1], Yazeed Al Moayed[2], Nasir Ahmed Algeelani[3]

Faculty of computer science and Information technology

AL-Madinah International University

Kuala Lumpur, Malaysia

DOI: https://doi.org/10.5281/zenodo.7365019

Published Date: 26-November-2022

***Abstract*: Online credit and debit card purchases have increased bank fraud. Using deep learning, researchers analysed cutting-edge fraud-detection algorithms. Most online KNN implementations work with sequential data like time series, audio/speech, and language. KNNs process and store locally, unlike forwarding networks. Time series analysis often uses KNNs. Check the numbers for suspicious purchases. Malaysian customers made 278,309 credit card transactions in October 2022, according to creditcard.csv. According to creditcard.csv, Malaysians used 278,309 credit cards in October 2022. K-nearest neighbour simplifies data categorization. Choose k for accurate scenario classification. k could be any positive integer between 1 and the set's size. When k = 1, a new log with the same category is generated. When k is small relative to the training dataset's features, noise is introduced. A large value of k complicates the calculations so much that a specimen is incorrectly classified into the most frequent category during data training. k-adaptability, KNN's usefulness, and versatility give it many advantages.**

***Keywords*: KNN algorithm, Machine Learning, Bank fraud, Credit/Debit cards.**

I. INTRODUCTION

The human capacity to collect, store, and comprehend the various scientific advances and large volumes of data that have only recently been generated outnumbers the capacity of the appropriate tools, limiting businesses' ability to detect fraud. Credit or debit cards, which have gained widespread acceptance as a payment method for online purchases of goods and services, are one type of bank fraud. As a result, researchers used deep learning to examine the most popular anomaly detection algorithms for bank fraud. Most published RNN implementations are geared toward sequential data types such as time series, audio/speech, and language. In contrast to forward networks, KNNs store and process data locally. KNNs are frequently employed to analyze time series data [1].

II. SIGNFICANCE OF STUDY

This study uses K-N-N machine learning to identify fraudulent credit and credit card transactions.

III. LITERATURE REVIEW

The general use of cutting-edge technological systems has resulted in a worldwide perspective shift. The new digital environment has profoundly affected people's day-to-day lives, and as a result, organizations of all kinds face significant transformations. Big data is playing an essential role in the current social paradigm shift. Businesses that want to maximize their use now consider them confidential. Many organizations keep track of all business transactions in their databases. Considering the numerous scientific breakthroughs of recent years, as well as the massive amounts of data generated during that time reducing the ability of institutions to detect fraud [4, 6]

Data Mining is one approach that can be used to gain insight from data (MD). Anomaly detection is a Data Mining technique with numerous banking transaction and data analysis applications. This data can be obtained by searching for statistically reliable concepts, ideas, or patterns that are not obvious at first glance, are previously unknown and can be derived from the original data.

For a long time, the term I4 has been used to refer to the most recent trend in technological development, which focuses on the integration of intelligent automation and the sharing of production data. The concept envisions the development of "self-monitoring devices, sensors, and machines" capable of "performing condition-based, decentralized small tasks for continuous monitoring and self-diagnosis" [2, 9]). A unique or unusual occurrence is if the data is flawed or the measurements are inaccurate. "anomaly detection" refers to searching for and then identifying instances that significantly lead to different results. Unusual information usually indicates a problem or rare occurrence, such as bank fraud, medical problems, structural flaws, faulty machinery, Etc. As a result, recognising these events in real time is critical from a business standpoint. In addition to health [5]. Based on the available data, the following machine-learning methods for this detection [6]:

1. This type of problem requires supervised classification techniques, which involve two sets of data for training and testing. Since everything is known, data is annotated as anomalous or not, and a model that can tell the difference between anomalous and normal data is developed.

2. Semi-supervised techniques are used when outliers are expected but not observed in the data.

3. When anomalies are present in a dataset without being labelled, legitimate and anomalous behaviours coexist because it is impossible to tell, in advance, whether a given piece of data is an anomaly. Similarly, this is a domain with a diverse range of possible career paths.

Due to the impressive results achieved across diverse domains, including image processing, numbers, text and fonts, and anomaly detection, deep learning has quickly become a popular machine learning technique amongst academics. The 1990s saw the application of machine learning technologies to a wide range of anomaly detection issues, including financial crime prevention. Deep learning is a branch of machine learning that has figured out how to achieve high performance and adaptability by representing data as a layered pyramid of constructs within layers of neurons. Artificial neural networks are frequently used for security tasks due to their well-known advantages, such as adaptability, parallelism, and learning; with this, the weight of the connections can be modified using training algorithms or learning rules, and with the updating of the weights, the network can optimize its connections adapting to various changes [7]. Artificial neural network architectures have come a long way from their biological forerunners, yet they continue to perform admirably.

A. Identification of Fraud using Machine Learning Machine Learning

Fraud detection involves looking for signs of wrongdoing in institutions like banks and stock exchanges and is one of the most studied anomaly detection applications. Fraud is the intentional misrepresentation of material facts for financial gain [8]. Half of the 7,200 companies surveyed in the 2018 Global Economic Crime Survey, Kaur (2020) reported experiencing fraud. Fraud in telecommunications, insurance, and banking is a significant problem for both public and private institutions. Since fraud is a dynamic offence, many tried-and-true machine learning methods have found usefulness in identifying instances of it.

Using a stolen credit or debit card to make an online purchase has become one of the most widespread forms of financial fraud. Payment card fraud is included in this category of fraud. The unpredictable nature of credit card fraud makes it challenging to detect. The standard method entails keeping a user's usage profile and monitoring for unusual behaviour. There are billions of cardholders, so user profiling is impractical; however, the problem's nature lends itself to a solution based on anomaly detection, which can spot attack patterns in real time by looking for out-of-the-ordinary behaviours indicative of bank fraud, such as stolen identities or unauthorized charges. Since preventing and detecting fraud in real time is usually necessary, it can be challenging.

Due to the abovementioned considerations, this paper focuses primarily on research into anomaly detection works employing the deep learning technique for detecting bank fraud. This paper gives a comprehensive overview of the current state of outlier identification because there are more options for misrepresentation.

IV. RESEARCH METHODOLOGY

This section discusses the current state of intrusion detection, with a particular emphasis on how financial institutions might employ machine learning algorithms to spot bank card fraud. At first, the fundamental ideas surrounding the topic and the operation of the techniques for spotting anomalies are laid out. Finding data points that are out of the ordinary is called "outlier detection." There are many domains in which outlier detection can be helpful. The term "outlier" refers to data patterns that deviate significantly from a standard measure of "normal" behaviour. A formidable obstacle in finding Outliers is probing the unknown. One method is to identify a "normal" range of data and label any observations that fall outside that range as "Outliers." This seemingly straightforward strategy [5] is, however, complicated for a number of reasons:

- Specifying a range of acceptable behaviour that encompasses every conceivable scenario is complicated.
- Frequently, it is challenging to distinguish between normal and abnormal behaviour. This implies that behaviour near the limit can be disguised as typical and vice versa.
- Depending on the context, "Outlier" can mean different things. Since each application domain has its own requirements and constraints, the outlier detection problem is formulated differently for each one.
- When designing an isolated occurrence recognition system, access to classified data for affirmation is often helpful.
- Malicious actors frequently cause outliers by altering the rules to make abnormal observations appear normal, which makes defining normal behaviour difficult.
- Due to noise mimicking authentic outliers' characteristics, it is often difficult to recognize and eradicate them from data.
- It is not unusual for "normal" behaviours to evolve, and tomorrow's "normal" may look very different.

The obstacles above make it difficult to find a general solution to the Exceptions segmentation task. Current detection strategies frequently oversimplify the issue by focusing on a single formulation. Several factors influence implementation, such as data type, availability of labelled data, the type of Outliers sought, etc. The technique's field of application typically determines these particulars.

1. *Data Set:*

We examine the data and label purchases as fraudulent or not fraudulent. The CSV (creditcard.csv) file contains 278,309 credit card transactions that Malaysian customers generated during the month of October 2022. Credit card purchases are either fraudulent or not based on the buyer's actions. Principal component analysis (PCA) transformation yields only numeric output variables for numeric input variables. Using the "Data-driven" dataset, we deploy colour labels models for detecting credit card fraud in this paper. Machine-zone signals were digitized and recorded based on location, transaction amount, device ID, audible noise, and ultrasonic acoustic emissions. These signals were captured using two data acquisition cards in a National Instruments LabView (NI LabView) environment (USA). Below is a diagram depicting the test stand's general layout and the measuring system's specifics.

Each group action in the dataset and the first group action can be found in the 'Time' feature (The amount of time gap between each group action is also noted). The 'Amount' attribute represents the total number of participants in group activities. In instances of fraud, the 'Location' feature (the response variable) is assigned the value one, whereas, in all other instances, it is assigned the value zero.

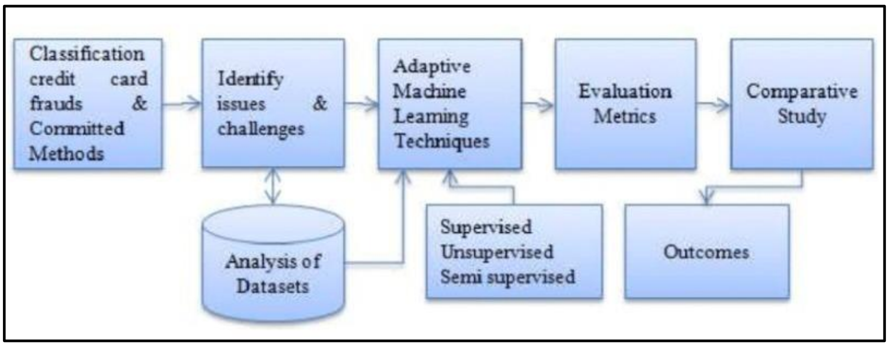

Classification credit card frauds & Committed Methods
Identify issues & challenges
Adaptive Machine Learning Techniques
Evaluation Metrics
Comparative Study
Analysis of Datasets
Supervised
Unsupervised
Semi supervised
Outcomes

Fig. Structure of measuring system used in Experiment

Experiments were carried out based on theoretical findings indicating that most pick strategies yield high accuracy rates when attempting to detect credit card fraud. To aid in further analysis of the hybrid models, the noise of approximately 100% and half an hour duration was added to the sample data. For half an hour of extra noise, many pick strategies received an honest score of 0.942. As a result, the picked approach demonstrated robust, noise-tolerant operation.

2. *System Architect*

More relevant to the calculation of credit card fraud rates is the fact that this system can accept credit card transaction data from customers in real time.

i. A statistical link Understanding the distinction between balanced and unbalanced knowledge and data, and separating the two into separate buckets, is essential when classifying a massive amount of information. Class-0 indicates that no fraud has occurred, while Class-1 indicates fraud.

ii. During k1 and k2 of the project, Class-1 indicates 492 fraud transaction samples across k28 iterations.

iii. The agglomeration method is utilized for the detection of outliers. This method involves calculating the distance between each similar bit of knowledge. Values that lie outside the bounds of the trained knowledge square are referred to as outliers. Before dividing the dataset into subsets, it standardizes the entire data set into a set appropriate for training and testing separated zero's quantitative connection.

iv. As a direct result of this, Train data will make up most of our overall comprehension and will serve as the primary basis for our understanding of the topic. The evaluation of the results will receive full credit.

3. *Design*

Classified models should make reliable and accurate predictions. The application must visualize the data and results. Some data appears to be excluded from the dataset due to restrictions on its use and privacy concerns.

4. *Training and Testing*

i. It is easier to analyze data when it is first divided into training and testing sets. Currently, the two data sets are entirely independent of one another. In the testing phase, data that has never been observed is utilized, whereas data that has been observed is used in the training phase to refine our model. The prediction model must be trained using a coaching set comprised of actual transactions for it to function correctly.

ii. Examine the dataset used to evaluate the prediction model's accuracy. The coaching staff set out to identify a trustworthy model for determining the legitimacy of a transaction.

iii. Used this method to determine the distance, in a multidimensional feature space, between the currently classified case from the test data set with an unknown colour label and each case in the training data set for which the colour was known. Then, the case without a colour label was automatically placed in the cluster containing the most similar instances from the training set, measured in terms of the number of k nearest neighbours.

iv. In order to construct the proposed machine, PYTHON is used. Dataset analysis showed that KNN had the highest accuracy rate.

V. DISCUSSION

1. *KNN Algorithm:*

K-nearest neighbour is a simple algorithm for classifying data. This technique is an example of a lazy algorithm, which has been around for a while (since Fix and Hodges 1951). Unlike some other classifiers, it does not construct an internal structure (prototype) of the training sample; instead, it simply looks for a solution when the classified object appears in the data. Using information about its k nearest neighbours, a brand-new x object can be assigned to the same category as the objects that make up the bulk of those neighbours. To assign a new item to a class, we compute the average distance between it and each item in the training data set. This is done in the n-dimensional feature space, for instance, with the Euclidean distance defined by Equation 1:

$$D(x,y)^2 = \sum_{i=1}^{n}(x_i - y_i)^2$$

x is the unlabeled object in the test dataset that is currently being classified;

y - attribute from the training sample;

n - a variety of characteristics of the object considered.

Before employing the algorithm, the parameter k must be selected. There are not any generally applicable or practically relevant methods for determining k. [8]. It suffices to choose k to obtain desirable classification results in a given scenario. Theoretically, k can have any value between one and the set's maximum size. If k equals one, a new log with the same category as its nearest neighbour is created. When k is small for the total number of features present in the training dataset, noise is introduced. When k is too large, the calculations become too complex, and a specimen is incorrectly assigned to the category with the most instances within the training method of data. The flexibility, functionality, and responsiveness of the k-NN classifier to a wide range of situations are just a few of its many advantages.

2. *Data Reprocessing:*

Formatting, cleaning, and sampling are preprocessing steps for data. Due to asymmetry in the classification process within and between card transactions, the fraud rate in the data set is less than the total number of transactions. For this purpose, stratified sampling is employed. Typical MATLAB procedures were followed after the exploratory approach was finished. The k-NN algorithm is implemented after the generation of training and test data. Features from all signals recorded with a specific experimental setup comprised the test set of data, which was generated with each run of the detection algorithm. Microscopically detecting the signal's accurate phases allowed us to record all of its features, which we subsequently included in our training data set. It was determined when and where modifications took place by validating dataset signal characteristics using the k-nearest neighbour method. All the characteristics examined in this research were extracted using the MATLAB Signal Processing Toolbox and the MATLAB Wavelet Toolbox. Considering the proposed identifier scheme, the fundamental mathematical formula looked like this:

$$D(x,y)^2 = \sum_{i=1}^{455} (x_i - y_i)^2$$

3. *Ranking Alert*

The veracity of security-related queries that triggered an alert is assessed here. When unusual behaviour is detected, new security questions will be generated. The size of the squares depicts the likelihood of victimization based on social status. A queue is formed, and the suspected fraudster's location is half-tracked if an alert has a higher probability than the others.

4. *Approach*

The following approach is used for the proposed Algorithms.

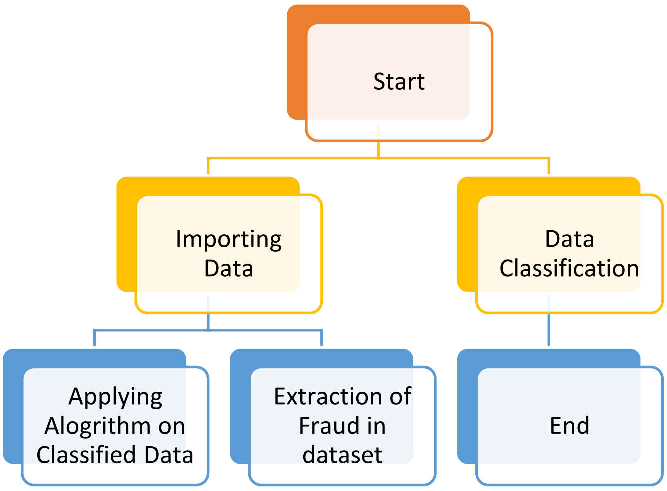

Start
Importing Data
Data Classification
Applying Alogrithm on Classified Data
Extraction of Fraud in dataset
End

Fig. Flowchart for Algorithm

5. *Evaluation Metrics*

Since precision is easily managed and generalizes to truly binary labels like accuracy, recall, and support, it is widely used as an objective evaluation metric to assess the overall performance of models across a wide range of tasks.

The preceding steps were executed multiple times, each time changing the value of the "k" variable. For each scenario, five unique test data sets were developed (for each k value). The accuracy of an evaluation can be gauged by dividing the number of correct labels by the total number of categories used in the assessment. Further, the following parameters are defined in a way that is independent of class:

- TN denotes the total number of correct negative identification
- TP denotes the total percentage of positive identifiers.
- FP denotes the number of false identifications.
- A false negative, or FN, is a misdiagnosis that occurs multiple times.

Accuracy:

Acc - accuracy,

$$Acc = \frac{TP + TN}{TP + FP + FN + TN}$$

F-score:

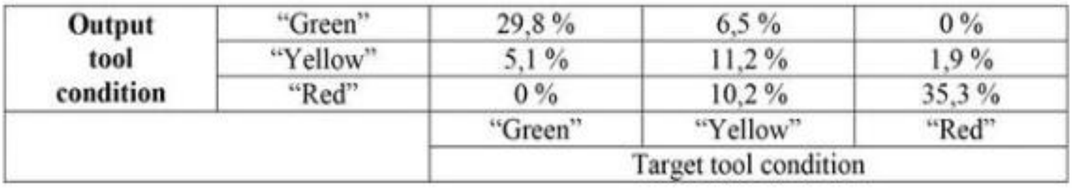
Acc - accuracy,

$$Acc = \frac{TP + TN}{TP + FP + FN + TN}$$

VI. RESULTS

The KNN algorithm was created specifically for this programme to detect Credit Card fraud. Any user can use the web app on any device to speculate on its future. The only requirements for a user are internet access and a data-display device. The only person authorized to enter data or make changes is the system administrator, hence the label "administrator-only." For the reasons stated above, we intend to establish a mechanism whereby a single trustworthy organization issues cards to customers. This organization would then access all necessary information about customers for local purposes. This confusion matrix's total number of classification occurrences is the independent variable.

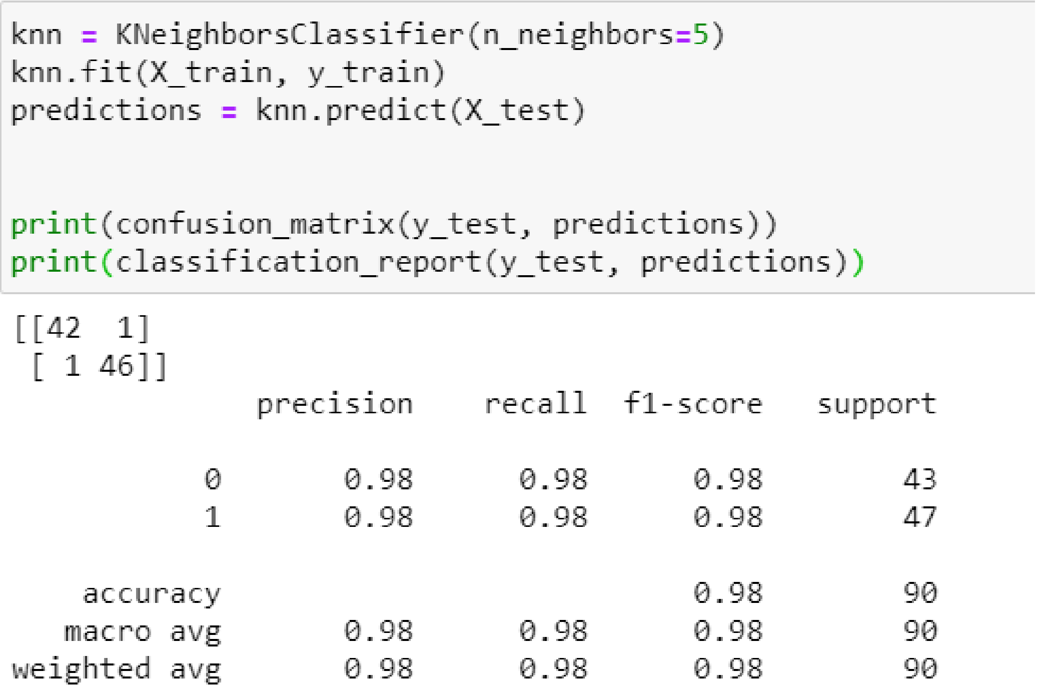
Table 1: Output Tools Condition



Output tool condition	"Green"	29,8 %	6,5 %	0 %
	"Yellow"	5,1 %	11,2 %	1,9 %
	"Red"	0 %	10,2 %	35,3 %
		"Green"	"Yellow"	"Red"
		Target tool condition		

This table is a percentage (rather than a numeric) representation of a typical confusion matrix, which is a chart used to examine how well a classification system works. The rows show the various classifications that the classifier's decisions led to (whether they were correct or not). However, the columns represent the actual divisions of societyThe first table presents the matrix confusion as a percentage, where the percentage represents the ratio of the number of cases assigned to a class to the total number of cases. That's right, the diagonal cells contain percentages that show how often the correct tool state was identified. The correct identification is indicated by a match between the expected and actual tool states. Table 1 reveals that 29.8 per cent of cases correctly belonged to the "Green" class, as evidenced by the value at the first column's crossing with the initial frame number. On the other hand, the frequency of different identification errors was displayed (also as a percentage) in all the other cells.

We will recollect the confusion matrix and the classification report to see if we can more precisely classify the misaligned points.



```
knn = KNeighborsClassifier(n_neighbors=5)
knn.fit(X_train, y_train)
predictions = knn.predict(X_test)


print(confusion_matrix(y_test, predictions))
print(classification_report(y_test, predictions))
```

```
[[42  1]
 [ 1 46]]
              precision    recall  f1-score   support

           0       0.98      0.98      0.98        43
           1       0.98      0.98      0.98        47

    accuracy                           0.98        90
   macro avg       0.98      0.98      0.98        90
weighted avg       0.98      0.98      0.98        90
```

Fig. KNN Algorithm results

1. *Tools Condition*

Indicators of the k-NN algorithm's performance in distinguishing between various states of the tool are provided in Table 2. As a result, we can say that the "Green" and "Red" groups performed well on the Accuracy and F-scores measures.

Table 2:

Tool Condition	Accuracy	F-score
Green	0.88	.84
Yellow	.76	.48
Red	0.88	.85

KNNs, on the other hand, have disadvantages such as high prediction costs, which are exacerbated by large datasets. KNNs are sensitive to outliers because they affect the nearest points. Furthermore, they need help with categorical features and high-dimensional datasets. The KNN algorithm becomes slower as more data is added because the model needs to keep track of all the data points to calculate the distance between them.

VII. CONCLUSION

One form of fraud against financial institutions is the fraudulent use of credit and debit cards, which have replaced cash as the primary payment method for online purchases. Keeping a usage profile for each user and monitoring them for deviations is the traditional method for detecting card fraud; however, this is impractical and expensive given the billions of card users. Deep learning systems have become increasingly important in this field due to their efficiency and accuracy, making the world safer. This form of learning can now be incorporated into most projects involving pattern recognition and data mining. Several network topologies, including the KNN algorithm, have been developed to prevent the identification of credit card fraud. The most pressing challenge for developers of deep learning-based bank fraud prevention models is the creation of adaptable, predictive models. The expectation for the future is that they will be able to recognize uncommon or novel events.

REFERENCES

[1] Awoyemi, J. O., Adetunmbi, A. O., & Oluwadare, S. A. (2017, October). Credit card fraud detection using machine learning techniques: A comparative analysis. In *2017 international conference on computing networking and informatics (ICCNI)* (pp. 1-9). IEEE.

[2] Asha, R. B., & KR, S. K. (2021). Credit card fraud detection using artificial neural network. *Global Transitions Proceedings*, *2*(1), 35-41.

[3] Dighe, D., Patil, S., & Kokate, S. (2018, August). Detection of credit card fraud transactions using machine learning algorithms and neural networks: A comparative study. In *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)* (pp. 1-6). IEEE.

[4] Kiran, S., Guru, J., Kumar, R., Kumar, N., Katariya, D., & Sharma, M. (2018). Credit card fraud detection using Naïve Bayes model based and KNN classifier. *International Journal of Advance Research, Ideas and Innovations in Technology*, *4*(3), 44.

[5] Kaur, D. (2020). Machine Learning Approach for Credit Card Fraud Detection (KNN & Naïve Bayes). In *Machine Learning Approach for Credit Card Fraud Detection (KNN & Naïve Bayes)(March 30, 2020). Proceedings of the International Conference on Innovative Computing & Communications (ICICC)*.

[6] Kumari, P., & Mishra, S. P. (2019). Analysis of credit card fraud detection using fusion classifiers. In *Computational Intelligence in Data Mining* (pp. 111-122). Springer, Singapore.

[7] Mehbodniya, A., Alam, I., Pande, S., Netware, R., Rane, K. P., Shabaz, M., & Madhavan, M. V. (2021). Financial fraud detection in healthcare using machine learning and deep learning techniques. *Security and Communication Networks*, *2021*.

[8] Pun, J. and Lawryshyn, Y., 2012. Improving credit card fraud detection using a meta-classification strategy. *International Journal of Computer Applications*, *56*(10).

[9] Zareapoor, M., & Shamsolmoali, P. (2015). Application of credit card fraud detection: Based on bagging ensemble classifier. *Procedia computer science*, *48*(2015), 679-685.

[10] Ahmad Alammar, Yazeed Al Moayed, Nasir Ahmed Algeelani (2022). Debit And Credit card fraud detection using k-n-n in machine learning.