INTERNATIONAL JOURNAL OF SCIENTIFIC & ENGINEERING RESEARCH, VOLUME 14, ISSUE 1, JANUARY-2023
ISSN 2229-5518

A Deep Learning Model for Detecting Anomalies in The Banking Sector Using A Feed-Forward Neural Network

Ogochukwu Patience Okechukwu, Godson Nnaeto Okechukwu, Chinedu Emmanuel Mbonu, Roseline Uzoamaka Paul

Abstract- The evolution of anomalies in the banking sector has resulted in the execution of several criminal activities in cyberspace. Since several users go online to access the services offered by government and financial establishments, there has been a considerable growth in malicious website attacks over the past few years. In other to solve this problem, this paper presents a Deep-Learning model for detecting anomalies (e-banking phishing, and fraudulent transactions) on two datasets. The datasets were segmented into X_train and y_train, X_test and y_test which holds 60% data for training and 40% testing data. The system model was trained by using the Feed Forward Neural Network, which had a precision of approximately 97% on the phishing dataset and 99% on the fraudulent dataset. The trained model was exported to the web using flask, which is a suitable python framework for web applications so that users can check for malicious websites and legitimate website URLs. The work can be extended further by training other models and employing different machine learning algorithms to determine the algorithm with the optimum accuracy result.

Index Terms— AdaBoost, banking, credit card, cyberspace, dataset, deep-learning, e-banking, feed-forward neural network, machine learning, phishing.

1 Introduction

THE banking industry is the most valuable and economically diverse in the world. Every minute, there is the potential for an average of one billion transactions to take place.

When operating a system with this much variety, the level of security that is required for such an operation is extremely high. Each transaction may involve sums ranging from one rupee up to several crores, sometimes even more. These kinds of systems need an enormous amount of security built into them [1].

The process of locating data points within a dataset that do not conform to the typical patterns is known as anomaly detection. It has the potential to be useful in the solution of many problems, including the detection of fraud, medical diagnosis, and other issues. Anomaly detection can be made more efficient by using machine learning methods, which allow the process to be automated and work better even with large datasets. For this study, the type of anomaly detection to be looked at is that of e-banking phishing and credit card fraud [2].

E-banking Phishing is the most dangerous form of criminal activity that can be carried out online. Over the past few years, there has been a notable rise in the number of phishing attacks. This can be attributed to the fact that the majority of users go online to access the services that are provided by the government and financial institutions [3]. An email scam, a voice-over-internet protocol call, a text message, or any other communication forms could be used in a phishing attack. Users typically have multiple user accounts on a variety of websites, including social networking sites, email services, and even accounts for online banking. The fact that most people are unaware of the

- *Okechukwu Ogochukwu Patience is a lecturer in Department of Computer Science, Nnamdi Azikiwe University, Awka, Nigeria, E-mail: op.okechukwu@unizik.edu.ng*
- *Okechukwu Godson Nnaeto is of Department of Electronic/Computer Engineering, Nnamdi Azikiwe University, Awka, E-mail: gn.okechukwu@unizik.edu.ng*
- *Mbonu Chinedu Emmanuel is a lecturer in Department of Computer Science, Nnamdi Azikiwe University, Awka, Nigeria, E-mail: ce.mbonu@unizik.edu.ng*
- *Paul, Roseline Uzoamaka is a lecturer in Department of Computer Science, Nnamdi Azikiwe University, Awka, Nigeria, E-mail: ru.paul@unizik.edu.ng*

valuable information they possess is one of the factors that contribute to the success of this attack.

As a result, the innocent users of the internet are the targets, they are the most vulnerable to this attack. In most cases, phishing attacks make use of social engineering to trick victims into providing sensitive information by tricking them into clicking on spoofed links that take them to fake websites [4]. The victim will either receive an email with the spoofed link or the link will be posted on popular websites. The fake website is designed to look almost exactly like the real website. As a result, rather than directing the request made by the victim to the actual web server, it will be directed to the server belonging to the attacker.

Credit card operations have become popular among web payment gateways used globally. With the increase in digital payment systems, there is an increase in the number of fraudulent transactions, and this act is being carried out by cyber thefts using various approaches. Credit card fraud can be seen as the act of obtaining card information without proper authorization from the account holder for financial gain. There are many ways in which fraudulent transactions can take place, and these ways can be organized into many different categories. According to the Nilson Report published in October 2016, global online payment systems brought in more than $31 trillion in 2015, representing a 7.3% increase over the previous year's total. The global losses incurred as a result of credit card fraud reached $21 billion in 2015. There is a possibility of credit card fraud rising to $31 billion by 2020 [5][6].

2. RELATED WORKS

To analyse customer historical transaction information and extract behavioural patterns, [7] created a unique fraud detection approach for streaming transaction data. wherein cardholders are grouped according to the value of their transactions. The Europe credit card dataset is the one that was used in this study. They employed the machine learning techniques of the decision tree, random forest, and logistic regression. Their experimental findings indicated that Random Forest has the highest accuracy rating.

[8] gave a thorough guide for choosing the best algorithm based on the fraud type and using an appropriate performance measure to explain the evaluation. To determine if a specific transaction is legitimate or fraudulent, they used predictive analytics performed by the implemented machine learning models and an API module. They evaluated an innovative approach that successfully tackles the skewed distribution of data.

Machine learning techniques were utilised in the construction of a model by [9] to identify fraudulent transactions involving credit cards. The AdaBoost method and the Random Forest Classifier are the two algorithms that were utilised. The accuracy, precision, recall, and F1-score metrics are the ones that were used to evaluate the performance of the two methods. Their experimental result shows that the Random Forest and the AdaBoost algorithms were compared with other machine learning techniques, and it was demonstrated that the AdaBoost and Random Forest Classifier had the greatest accuracy, precision, recall, and F1-score and were considered to be the best algorithm that used to detect fraud.

[10] developed several methods that can be utilised to categorise transactions as either fraudulent or legitimate ones. This research made use of the Credit Card Fraud Detection dataset as its primary source of data. Oversampling was done using the SMOTE technique because the dataset contained a significant amount of imbalance. In addition to this, feature selection was carried out, and the dataset was then divided into two distinct parts: training data and test data. During the experiment, the following algorithms were utilised: Logistic Regression, Random Forest, Naive Bayes, and Multilayer Perception. According to the findings, each algorithm has the potential to be utilised for the accurate identification of credit card fraud. The proposed approach is capable of identifying a variety of additional types of anomalies.

On highly skewed credit card fraud data, [11] explored and compared the effectiveness of using decision trees, random forests, support vector machines, and logistic regression. This study made use of a dataset that consisted of credit card transactions obtained from European cardholders. This particular dataset has a total of 284,786 transactions. On both the raw and the pre-processed data, the machine learning approaches that have been discussed were utilised. Accuracy, sensitivity, specificity, and precision were the criteria that were used in the analysis of the performance of the machine learning techniques. According to the findings, the optimum levels of accuracy for logistic regression, decision trees, Random Forest, and support vector machine (SVM) classifiers are respectively 97.7%, 95.5%, and 97.5%.

[12] presented a deep learning algorithm as well as a machine learning method to safeguard credit card transactions; this would allow consumers to use e-banking in a secure and hassle-free manner. Deep learning, logistic regression, naive Bayesian, SVM, neural network, artificial immune system, k nearest neighbour, data mining, decision tree, fuzzy logic-based system, and genetic algorithm are the foundations of the deep learning method. The results of their experiments indicated that the machine learning algorithms demonstrate a higher degree of accuracy.

[3] provided an innovative method to identify phishing websites by utilising machine learning algorithms in their research. They evaluated and contrasted the performance of five different machine learning algorithms: Decision Tree (DT), Random Forest (RF), Gradient Boosting (GBM), Generalized Linear Model (GLM), and Generalised Additive Model (GAM). A table was created to compare the best three algorithms. According to the data presented in the tables of accuracy, recall, and performance, the Random Forest algorithm achieved the highest levels of accuracy (98.4%), recall (98.59%), and performance (97.70%).

A model with an answer was proposed by [13] for recognising phishing sites by applying a URL identification technique that utilised the Random Forest algorithm. Parsing, heuristic classification of the data, and performance analysis are the three stages that are included in Show. The feature set can be analysed by using parsing the collected information from the PhishTank database. Only 8 of the possible 31 features were taken into account throughout the parsing process. A degree of accuracy of 95% was achieved using the random forest method.

3. METHODOLOGY

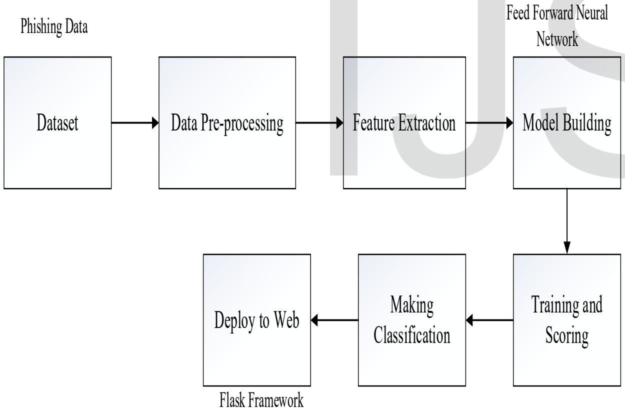


Figure 1: Architecture of the System

Figure. 1 shows the architecture of the system which is made up of different components. The components of the system are described as follows:

Dataset: The dataset used in this research is that of the e-banking phishing and fraudulent transaction dataset. The E-banking phishing contains 96012 of both websites that are phishing and websites that are not phishing. Whereas credit card fraud data comprises transactions made by credit cards in September 2013 by European cardholders. The dataset presented transactions which occurred in two days, and where there were 492 frauds out of 284,807 transactions [14]. It contained only numerical input variables which were the result of a principal component analysis (PCA) transformation [7][15]. Unfortunately, confidentiality issues did not allow, for provision of the original features and more background information about the data. Features V1, V2, … V28 were the principal components obtained with PCA, the only features which had not been transformed with PCA were 'Time' and 'Amount'. Feature 'Time' contained the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' was the transaction Amount, this feature could be used for example-dependent cost-sensitive learning. Feature 'Class' is the response variable and it takes a value of 1 for fraud and 0 otherwise [7][16].

Pre-processing: The dataset was pre-processed to remove redundant values, and infinite values and also convert the domain column to 0 and 1 for easy and efficient model training.

Feature Extraction: This has to do with the reduction of the features in the dataset that was employed in the system design process.

Model Building: The system model was built using a feed-forward neural algorithm in training the model for malicious website detection. A feed-forward neural network consists of inputs, hidden layers and an output layer which can be one output or two outputs etc., depending on classification. Here, the output layer is one (1).

Testing/Accuracy: The efficiency of the system model was determined by its accuracy, false and true positive.

Making Classification: This has to do with the deployment of our trained model to the web using the python flask framework so that users can input various websites into the system, in other to classify they are malicious ones or not.

Flask App: It is the python web framework. The System model was exported to the flask for easy execution.

4. RESULTS AND DISCUSSIONS

This system makes use of a deep learning algorithm in detecting anomalies in the bank sector. The anomalies that this research covers are e-banking phishing and fraudulent transactions that are done via credit cards. The first phase of the system starts with acquiring an e-banking dataset. The dataset comprises 48,006 legitimate website URLs and 48,006 phishing URLs making 96,012 website URLs. The dataset was pre-processed by removing all duplicate and Nan values, therefore making it fit for suitable training performance. After processing, feature extraction is performed with the sole reason of reducing the dataset dimension and some unwanted feature columns thereby reducing the dataset from 16 feature columns to 2 feature columns with the domain feature column (which contains the domain website URLs) and the label feature column (this contains binary values where 0 represent a legitimate website URL and 1 represent a Malicious website). CountVectorizer function was employed in converting text documents (Domain column) to a vector of term/token counts. CountVectorizer also enables the pre-processing of text data before generating the vector representation. The dataset was partitioned into X_train and y_train, X_test and y_test which holds 60% data for training and 40% testing data. The system model was trained using a Deep Feed Forward Neural Network, which had a precision of about 97% approximately as represented in figure 2. The trained model was exported to the web using flask, which is a suitable python framework for web applications so that users can check for phishing websites (harmful) and benign website URLs. The second phase of the experiment has to do with acquiring a fraudulent transaction dataset. The dataset was also pre-processed by removing null values. Feature extraction technique was also used in selecting important features on the dataset. The result of the extracted features was then used in training as a feed-forward neural network as input. The result of the feed-forward neural network on the fraudulent dataset can be seen in figure 3.

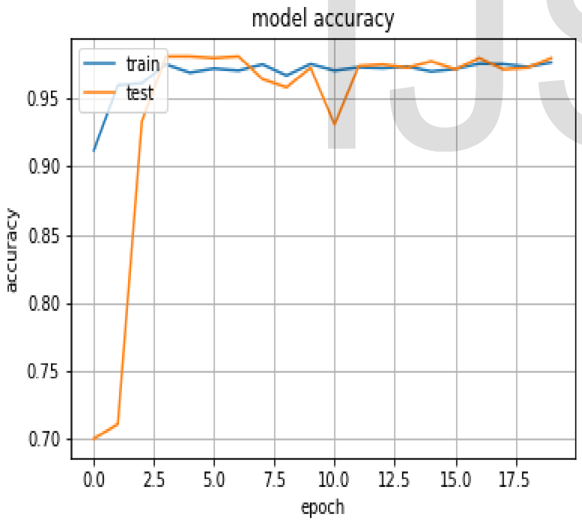
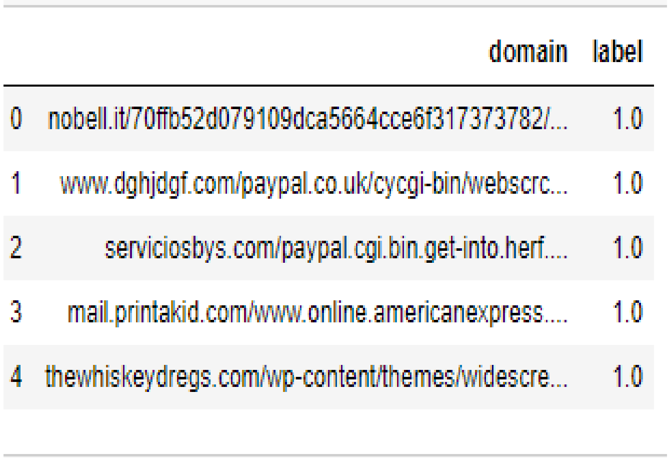


Figure 2: Accuracy Level of the Trained Model



	domain	label
0	nobell.it/70ffb52d079109dca5664cce6f317373782/...	1.0
1	www.dghjdgf.com/paypal.co.uk/cycgi-bin/webscrc...	1.0
2	serviciosbys.com/paypal.cgi.bin.get-into.herf....	1.0
3	mail.printakid.com/www.online.americanexpress....	1.0
4	thewhiskeydregs.com/wp-content/themes/widescre...	1.0

Figure 3: Malicious dataset of the first five rows of the trained dataset.

The label column represents the output of the system model where 1 represents a malicious website, and 0 represents a legitimate website.

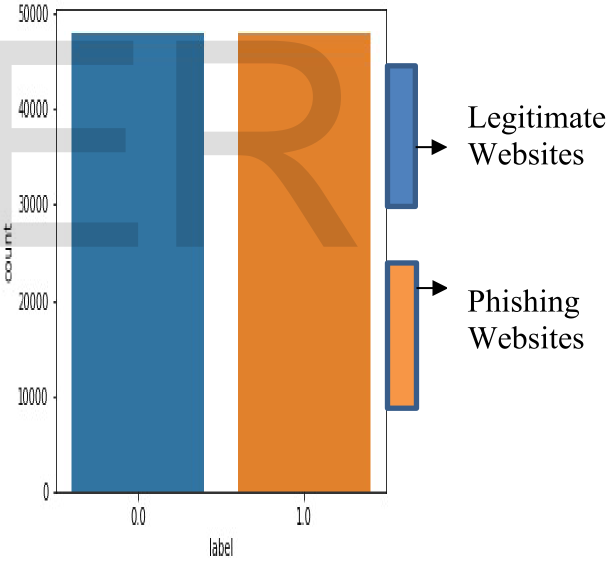
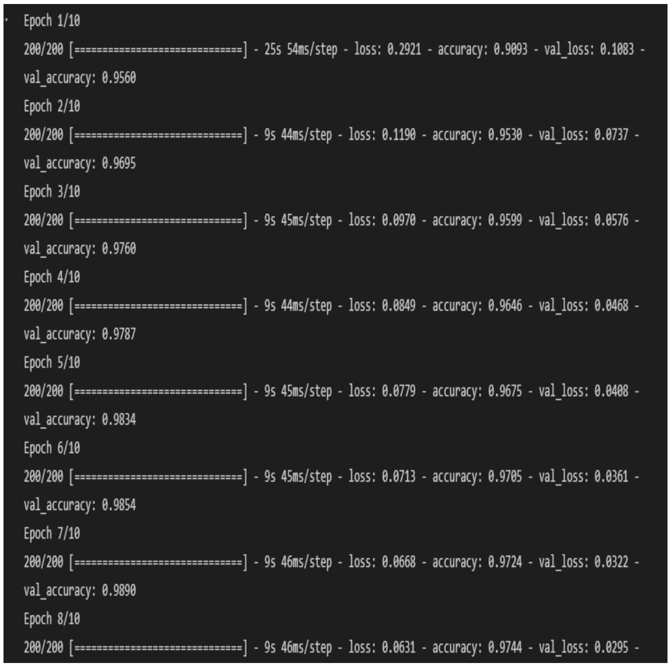


Figure 4: Dataset analysis of both Phishing and Legitimate Websites

The figure 4 gives a graphical view of the number of websites which are in the phishing category and legitimate category. The countplot showed that over 45000 were both benign and malicious websites. This signified that the dataset was balanced.



```
Epoch 1/10
200/200 [==============================] - 25s 54ms/step - loss: 0.2921 - accuracy: 0.9093 - val_loss: 0.1083 -
val_accuracy: 0.9560
Epoch 2/10
200/200 [==============================] - 9s 44ms/step - loss: 0.1190 - accuracy: 0.9530 - val_loss: 0.0737 -
val_accuracy: 0.9695
Epoch 3/10
200/200 [==============================] - 9s 45ms/step - loss: 0.0970 - accuracy: 0.9599 - val_loss: 0.0576 -
val_accuracy: 0.9760
Epoch 4/10
200/200 [==============================] - 9s 44ms/step - loss: 0.0849 - accuracy: 0.9646 - val_loss: 0.0468 -
val_accuracy: 0.9787
Epoch 5/10
200/200 [==============================] - 9s 45ms/step - loss: 0.0779 - accuracy: 0.9675 - val_loss: 0.0408 -
val_accuracy: 0.9834
Epoch 6/10
200/200 [==============================] - 9s 45ms/step - loss: 0.0713 - accuracy: 0.9705 - val_loss: 0.0361 -
val_accuracy: 0.9854
Epoch 7/10
200/200 [==============================] - 9s 46ms/step - loss: 0.0668 - accuracy: 0.9724 - val_loss: 0.0322 -
val_accuracy: 0.9890
Epoch 8/10
200/200 [==============================] - 9s 46ms/step - loss: 0.0631 - accuracy: 0.9744 - val_loss: 0.0295 -
```

Figure 5: Training process of the fraudulent transactions

This shows the training process of the feed-forward neural network on the fraudulent data. The training process comprised the accuracies and loss values obtained by the model during training and testing.

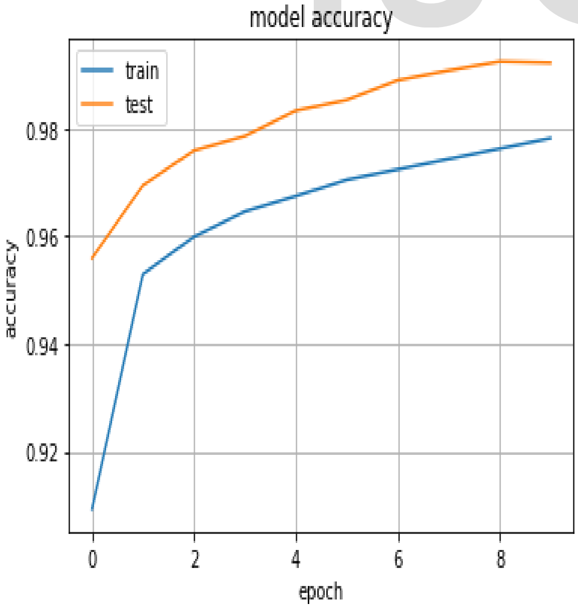


Figure 6: Training accuracy of the credit card transaction model.

This shows the accuracy achieved at each training step that the model completed.

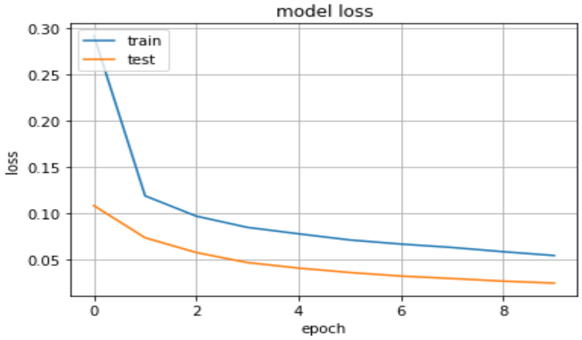
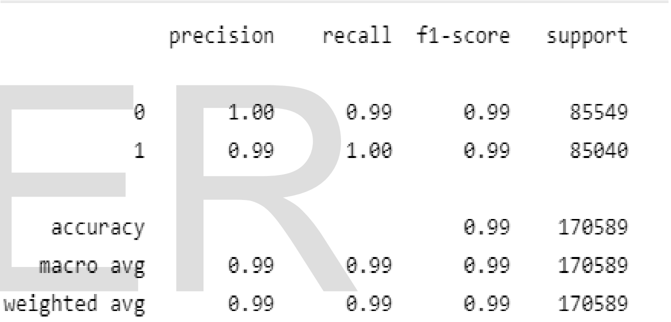


Figure 7: Training loss of the credit card transaction model.

This shows the loss gotten by the model at each training step that the model completed.



```
              precision    recall  f1-score   support

           0       1.00      0.99      0.99     85549
           1       0.99      1.00      0.99     85040

    accuracy                           0.99    170589
   macro avg       0.99      0.99      0.99    170589
weighted avg       0.99      0.99      0.99    170589
```

Figure 8: Classification Report

This shows the accuracy and loss achieved by the model.

5. CONCLUSION

The evolution of anomalies in the banking sector has resulted in the execution of several criminal activities in cyberspace. Since several users go online to get access to the services offered by government and financial establishments, there has been a considerable growth in malicious website attacks over the past few years. Cyber attackers start getting financial benefits and they see this as a lucrative business. Different methods are employed by cyber criminals to target vulnerable users and such methods include VOIP (voice over internet protocol), spoofed link and counterfeit websites, and messaging. It is quite simple to create spurious websites, which look exactly like the real website concerning age layout and contents. Even, the contents of these websites would be similar to benign websites. This presents a Deep-Learning model for detecting anomalies (e-banking phishing, and fraudulent transactions) on two datasets. The datasets were segmented into X_train and y_train, X_test and y_test which holds 60% data for training and 40% testing data. The system model was trained using Feed Forward Neural Network, which had a precision of about 97% approximately on the phishing dataset and 99% on the fraudulent dataset. The trained model was exported to the web using flask, which is a suitable python framework for web applications so that users can check for malicious websites and legitimate website URLs. The work can be extended further by training other models and employing different machine learning algorithms to determine the algorithm with the optimum accuracy result.

REFERENCES

[1] V. C Sharmila, K. Kumar, R. Sundaram, D. Samyuktha D, and R. Harish, "Credit Card Fraud Detection Using Anomaly Techniques," *2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT)*, pp. 1-6, April 2019, doi:10.1109/ICIICT1.2019.8741421. IEEE.

[2] O. Gorodetskaya, Y. Gobareva, and M. Koroteev, "Forecasting Time Series in the Banking Sector Using a Machine Learning Pipeline," *2021 14th International Conference Management of large-scale system development (MLSD)*, 2021, pp. 1-5, doi: 10.1109/MLSD52249.2021.9600170. IEEE.

[3] R. Kiruthiga, D. Akila, "Phishing Websites Detection Using Machine Learning," *International Journal of Recent Technology and Engineering (IJRTE)*, vol. 8, no. 8, pp. 111-114, Sep 2019.

[4] R. Palanisamy, M. T. Shaikh, S. Jayapal, and D. Thomas, "Analyze of Phishing Violence and Alleviation," *European Journal of Business and Management Research*, vol. 4, no. 6, pp. 1-3, Dec 2019.

[5] J. O. Awoyemi, A. O. Adetunmbi, S. A. Oluwadare, "Credit Card Fraud Detection Using Machine Learning Techniques: A Comparative Analysis," *International Conference on Computing Networking and Informatics (ICCNI)*, pp. 1-9, Oct. 2017, doi: 10.1109/ICCNI.2017.8123782

[6] B. Buonaguidi, "Credit Card Fraud: What You Need to Know," 2017, https://www.bbc.com/worklife/article/20170711-credit-card-fraud-what-you-need-to-know, accessed 2nd July, 2022.

[7] V. N. Dornadula, and S. Geetha, "Credit Card Fraud Detection Using Machine Learning Algorithms," *Procedia Computer Science*, vol.165, pp. 631-641, 2019, doi: 10.1016/j.procs.2020.01.057

[8] A. Thennakoon, C. Bhagyani, S. Premadasa, S. Mihiranga and N. Kuruwitaarachchi, "Real-time Credit Card Fraud Detection Using Machine Learning," *9th International Conference on Cloud Computing and Data Science & Engineering (Confluence)*, Jan 2019, pp. 488-493, doi: 10.1109/CONFLUENCE.2019.8776942.

[9] R. Sailusha, V. Gnaneswar, R. Ramesh, and G. R. Rao, "Credit Card Fraud Detection Using Machine Learning," *4th International Conference on Intelligent Computing and Control Systems (ICICCS)* Jun, 2020, pp. 1264-1270, doi: 10.1109/ICICCS48265.2020.9121114

[10] D. Varmedja, M. Karanovic, S. Sladojevic, M. Arsenovic and A. Anderla, "Credit Card Fraud Detection - Machine Learning Methods," *2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH)*, May 2019, pp. 1-5, doi: 10.1109/INFOTEH.2019.8717766.

[11] N. Khare, S. Y. Sait, "Credit Card Fraud Detection Using Machine Learning Models and Collating Machine Learning Models," *International Journal of Pure and Applied Mathematics*, vol. 118, no. 20, pp. 825-838, 2018, https://acadpubl.eu/hub/2018-118-21/articles/21b/90.pdf accessed 4th July, 2022.

[12] H. Paruchuri, "Credit Card Fraud Detection using Machine Learning: A Systematic Literature Review," *ABC Journal of Advanced Research*, vol. 6, no. 2, 113-120, Sep. 2017

[13] S. Parekh, D. Parikh, S. Kotak, and S. Sankhe, "A new method for Detection of Phishing Websites: URL Detection," *Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, Apr. 2018, pp. 949-952, doi: 10.1109/ICICCT.2018.8473085

[14] J. Brownlee, "Imbalanced Classification with the Fraudulent Credit Transactions Dataset," 2020, https://machinelearningmastery.com/imbalanced-classification-with-the-fraudulent-credit-card-transactions-dataset/ accessed 3rd July, 2022

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

[16] S. P. Maniraj, A. Saini, S. Ahmed, and S. D. Sarkar, "Credit Card Fraud Detection Using Machine Learning and Data Science," *International Journal of Engineering Research & Technology (IJERT)*. vol. 8 no. 9, pp. 110-115, Sep 2019.