Interpretable Credit Card Fraud Detection Using Deep Learning Leveraging XAI

***Abstract*—Due to the internet's widespread accessibility, more and more businesses are bringing their offerings online. Besides, because of the growth of E-commerce websites, both individuals and businesses that deal in finances are more dependent on internet administrations to handle their business. Since more and more people are using online banking and making purchases online, credit card fraud has increased. Fraudsters can also use anything to disrupt the existing fraud detection system's systematic operation. As a result, we took on the issue of improving the existing fraud detection system to the highest possible level. This research seeks to develop an efficient fraud detection system by utilizing deep learning (DL) as well as the machine learning methods that are responsive to shifting patterns of customer behavior and have a tendency to reduce fraud manipulation through the identification and filtering of fraudulent activity in real time. The techniques in our research include Artificial Neural Network, Convolutional Neural Network, Recurrent Neural Network, Logistic Regression, K-Nearest Neighbor, Naive Bayes, Meta-Learning, and Explainable Artificial Intelligence (XAI). This research suggests that the K-Nearest Neighbor is the most effective algorithm with an accuracy of 99.75% among many others.**

***Index Terms*—Artificial Neural Network, Convolutional Neural Network, Recurrent Neural Network, Logistic Regression, K-Nearest Neighbor, Naive Bayes, Meta-Learning, Explainable AI.**

I. Introduction

Electronic commerce has been applied by the majority of businesses, organizations, and government agencies in order to increase their levels of working. Electronic commerce platforms are increasingly vulnerable to large-scale fraud because of the fact that they are used by both legitimate customers and dishonest individuals. A crime that is performed with the intention of acquiring money through deception is referred to as fraud.

The purpose of a fraud control system is to protect sophisticated technology from fraud by avoiding its occurrence. However, this strategy is inadequate to avoid fraud. Detection of fraud is frequently advised as a means of enhancing a system's security. Here, the detection of credit card fraud identifies the fraudulent transactions and notifies the system administrator. Besides, credit card fraud detection by a machine or system is a tough phenomena. A system must be intensively trained with relevant data in order to achieve the process of detecting fraud. Moreover, deep learning and machine learning are the best approaches to solve these kinds of issues nowadays. The term "Deep Learning" (DL) refers to a specific type of machine learning that makes use of Artificial Neural Networks with several processing layers to extract more complex characteristics from raw input. We have moved to both deep learning-based and machine learning-based algorithms for detecting credit card fraud because it will give higher accuracy of detecting frauds. Besides, it will continuously processes and analyze new data. Manual work is less needed here. It has the ability to identify new patterns. It has the adaptability to change easily with the new environment. Indeed, "Deep learning" is an artificial intelligence (AI) and machine learning (ML) technology that aims to imitate how people learn certain knowledge. Deep learning is crucial in data science, which also includes areas like statistics and predictive modeling. Since deep learning and machine learning improves and speeds the processes involved in obtaining, analyzing, and interpreting enormous volumes of information, it is a valuable tool for data scientists. In contrast, not all datasets are precise. Data processing is thus an important factor in both deep learning and machine learning methods. Obviously, processing a dataset will increase its use for deep learning and machine learning. In addition, data processing involves the creation of a suitable data collection mechanism. As stated prior, to find error-free or accurate data is quite impossible. Therefore, the dataset which we are going to use, has also faults that can be addressed. It is possible that we will require some form of data shaping and modifying for the dataset.

II. Literature Review

A. Related Works

Engineers have been continuously looking for innovative solutions to provide more convenient, safe, and precise transactions in the field of Credit Card Fraudulence Discovery. This subject has become even more important in light of the recent development in machine learning and data science. Numerous significant research findings have been made in this area, serving as a foundation for ongoing and future research. Researchers have tried out different concepts and have worked with machine learning using different algorithms. As part of our background research, we came across studies in which tests were successfully carried out using a variety of machine learning (ML) and deep learning (DL) techniques.

In their paper, J. Galindo and P. Tamayo compared the effectiveness of KNN, Neural Networks, and CART model for analysis of credit risk using data on house mortgage loans supplied to them by Mexico's securities exchange commission. Each entry in the dataset which had around 4,000 in total—represented a customer account having a total of 24 attributes. It just needed small data pre-processing which

was necessary before they could begin using their preferred algorithms on it. Following the three chosen algorithms' predictions, they tabulated and graphically represented the findings and conducted a comparison. In comparison to the neural network and KNN model, it was observed that CART was the most accurate. However, it was also found that, to perform better the CART needs at least 22,000 entries. But it is fine when it comes to developing a risk prediction model for a company like CNBV [1].

It is important to separate the pertinent data and influencing aspects used in the risk calculation, as well as to choose the appropriate models for risk analysis, in order to effectively estimate risk while adhering to the essential criteria. Support Vector Machine or SVM, would be a useful technique in such a circumstance, as stated in a work by Gudas, S., Garsva, G. and Danenas, P. [2]. The main benefit of SVM over other AI-based solutions is that the solution produced by it will not be under local minima. Finding the unique data points that will be utilized as the solution's support vectors is important in order to put this strategy into practice.

Addo P.M. has developed a method to identify defaulters in this risk analysis by introducing the Elastic Net algorithm [3]. This technique has multinomial and logical functions as well as a strong error-checking system that improves accuracy while lowering error. Using an estimation process this approach provides information with a high degree of accuracy regarding loan repayment. Elastic net penalty is identified by constructing a graph in which x stands for predictors and y for response variables. Here, two elastic net algorithm equations are used. This has been taken assistance from gradient boosting machine and random forest modeling algorithms.

The Bayesian Classifier approach, where DAG (directed acyclic graph) strategy is used which helps to find loan repayment probability as per Pandey T.N. [4]. In this method, the nodes are random variables and the edges are dependencies. The accuracy is based on how well the datasets and dependencies are connected in the network. The Naive Bayesian classifier is another altered variation of the Bayesian classifier where the dataset attributes are represented as independent variables requiring less datasets. Additionally, with a non-parametric approach, KNN works with training sets having positive and negative cases. The testing and the training phase are the two divisions of the functionality. This approach computes the Euclidean Distance between the training points during the testing phase. The most equivalent instance is used as the output after producing instances using regression. The K-Means and SVM techniques are used in this research.

In terms of developing a credit rating system, LS-SVM and Neural Network algorithms perform better, according to Baesens B. [5]. They adopted three different SVM implementation methodologies for credit rating in their study. Additionally, they used two UCI datasets to assess the SVM's accuracy. This method's accuracy is almost comparable to the neural network and decision tree methods, having to require fewer input features. Additionally, the use of parameters can be reduced using genetic algorithms combined with SVM.

III. Proposed Model

A. Workflow:

The first step to do a complete thesis based on deep learning and machine learning is to choose a dataset that contains the right amount and type of data; the second step is to choose ML and DL algorithms that will be responsible for making predictions about the target variable; the third step is to pre-process the dataset which include splitting the dataset into training and testing sets by using splitting train test method. The fourth step is to train the algorithms on the dataset. In the fifth step, it's important to check if additional pre-processing improves the anticipated values in order to make sure the algorithms are operating as effectively as possible. Finally, it is necessary to determine the efficacy of the ML and DL algorithms. At the end, every model's results are explained through SHAP.

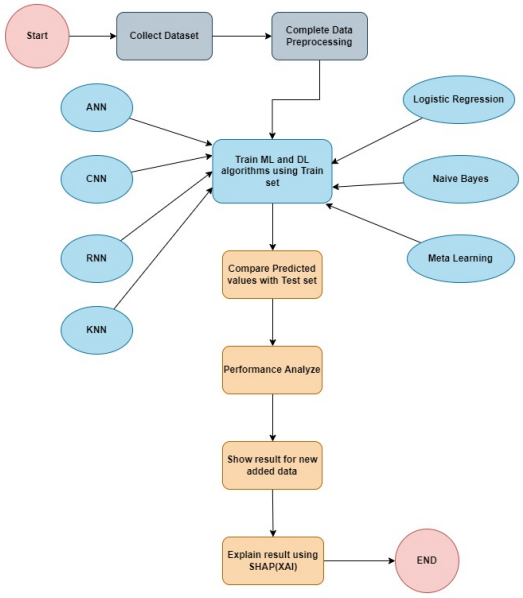

Fig. 1. Workflow Overview

In figure(l), we have showed the workflow and the steps of our implementation to reach the objectives.

B. Overview of The Dataset

We are utilizing a hybrid approach to establish a system that can work properly because our dataset solely contains quantitative data. Some nameless institutions simulated transactions in our simulated dataset are both legitimate and fraudulent. The total dataset includes 10,00,000 transactions; 87,403 of which were genuine fraudulent, and 8 columns containing transaction information [6]. The eighth or last column indicated whether the transactions were fraudulent or non-fraudulent. We denoted this last column as the target variable or label. Moreover, we separated the dataset into X and Y, where X represents "Features" and Y represents "Label". We separated our entire dataset into training and testing sets. Training set containing 80% and testing set containing 20% of the entire data.

Features	Descriptions
distance_from_home	The distance from home where the transaction happened
distance_from_last_transaction	The distance from last transaction happened
ratio_to_median_purchase_price	Ratio of purchased price transaction to median purchase price
repeat_retailer	Is the transaction happened from the same retailer
used_chip	Is the transaction through chip (credit card)
used_pin_number	Is the transaction happened by using a PIN number
online_order	Is the transaction an online order
fraud	Is the transaction fraudulent

TABLE I
DETAILS ABOUT THE FEATURES

IV. EXPERIMENTATION

A. Confusion Matrix:

In order to make sense of the data generated by an algorithm, the findings are sometimes displayed in a graphical format known as a Confusion Matrix. Its primary applications are in the study of deep learning and machine learning, the solution of statistical classification issues. There are 4 factors in a confusion matrix produced by a deep learning algorithm. Those are True Positive(TP), True Negative (TN), False Positive (FP) and False Negative (FN).

B. Precision, Recall, Accuracy & F1-Score:

We calculated the training score and the testing score for a single algorithm as well as the precision, recall and accuracy metrics for different models. Any technique for pattern recognition that assists in locating a certain pattern within a given set of data should prioritize precision, recall, and accuracy as its primary metrics for measuring its performance.

Precision: Precision is a metric of the performance of a machine learning model and the quality of a model's accurate prediction. Precision is the ratio of the number of genuine positives to the overall number of positive forecasts. In mathematics, precision is indicated by:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

Recall: Similar to precision, recall is an essential component of pattern recognition, retrieval of information and classification. It is the percentage of appropriately returned instances. In math, it can be represented as:

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

Accuracy: The best measure for evaluating the results of a model simulation is accuracy. It is the proportion between all of the accurate predictions and all of the predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{3}$$

F1-Score: The F1 score is the last metric used. As a stronger metric, it is calculated utilizing the accuracy and recall values. The formula for determining the model's F1 score is provided below:

$$F1\_Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{4}$$

V. RESULT ANALYSIS AND DISCUSSION

A. Full Dataset:

We evaluated the models using our dataset of 10,00,000 transactions and found significant findings. However, there are 9,12,597 non-fraud transactions and 87,403 fraud transactions, making this dataset very imbalanced. Therefore, To resolve this issue we used a technique of using an equal proportion of both fraud and non-fraud transactions. Since there are only 87,403 fraud transactions, we utilized the same amount of fraud and non-fraud transactions to compare the metrics of all the models.

Category	Values
Fraud Transactions	87403
Non-Fraud Transactions	912597
Train set	800000
Test set	200000
Total Transaction	1000000

TABLE II
DESCRIPTION OF THE FULL DATASET

In order to run simulations of our models, we use the splitting train test method to split the full dataset into "train" and "test" sets. We know that supervised machine learning and deep learning models require both training and testing data; we chose 8,00,000 transactions and 2,00,000 transactions from the full dataset to train and test the model. After that, we used accuracy, precision, recall and F1-Score metrics in our entire dataset to evaluate how well the models performed. Table (III) shows that the results are excessively accurate and over-fitting, suggesting that an unbalanced dataset can not produce reliable results. The performance metrics for the full dataset model simulations are presented below in Table (III).

Models	Accuracy(%)	Precision(%)	Recall(%)	**F1Score**(%)
KNN	99.89	99.57	99.16	99.37
CNN	99.87	99.14	99.41	99.27
RNN	99.81	98.78	99.32	99.06
ANN	99.76	98.75	98.50	98.62
Meta-Learning	98.1	92.7	91.2	96.4
Logistic Regression	95.91	89.50	60.26	72.08
Naive Bayes	95.14	79.15	60.23	68.40

TABLE III
PERFORMANCE METRICS OF FULL DATASET

B. Balanced Dataset:

We have measured the models' efficacy using the precision, recall, accuracy and F1-score metrics. Because our dataset was so unbalanced, we changed our approach based on sampling approximately the same number of actual and fraudulent transactions so that we can quickly evaluate the highest performing models. To make our dataset a balanced one, we have taken the equal amount of fraud and non-fraud transactions. Here, we have taken 87,403 fraud transactions as well as 87,403 non-fraud transactions.

Category	Values
Fraud Transactions	87403
Non-Fraud Transactions	87403
Train set	139844
Test set	34962
Total Transactions	174806

TABLE IV
DESCRIPTION OF THE BALANCED DATASET

Now, for the balanced dataset, to run simulations of our models, we split the dataset into "train" and "test" sets. Here, we have modified our dataset into a balanced dataset. Given that supervised deep learning and machine learning models require training data, we decide to train 1,39,844 transactions on the model and to test 34,962 transactions on the model. For the whole dataset, we used accuracy, precision, recall and F1-Score metrics to evaluate how well those models performed. The performance metrics for the balanced dataset model simulations are shown in Table (V) .

Models	Accuracy(%)	Precision(%)	Recall(%)	F1-Score(%)
KNN	99.75	99.58	99.91	99.75
RNN	99.66	99.49	99.73	99.61
ANN	99.61	99.48	99.74	99.60
CNN	99.26	98.73	99.81	99.27
Meta-Learning	98.1	99.04	97.03	97.8
Naive Bayes	94.01	96.54	91.61	94.01
Logistic Regression	93.92	93.34	94.58	93.96

TABLE V
PERFORMANCE METRICS OF BALANCED DATASET

With all of the evaluation metrics mentioned above, KNN has achieved a maximum accuracy of 99.75%. Besides, the second highest is the RNN around 99.66%. ANN also gives very similar results like RNN, its accuracy is 99.61%, whereas CNN gives 99.26% and Meta-Learning gives 98.1% accuracy. However, we can see some decrease in accuracy in Naive Bayes(94.01%) and Logistic Regression(93.92%).

C. Confusion Matrix Analysis

Here, we have shown the confusion matrices for KNN, ANN, and RNN models below which we used in our paper.

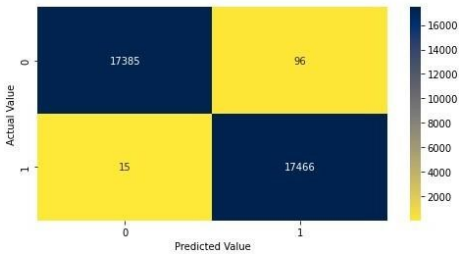

Fig. 2. KNN Confusion Matrix

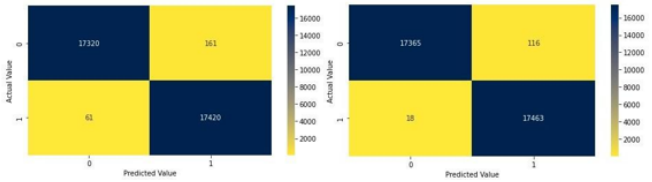

Fig. 3. ANN and RNN Confusion Matrix

D. ROC CURVE ANALYSIS

For the purpose of demonstrating the diagnostic efficacy of binary classifiers, a Receiver Operator Characteristic (ROC) curve is plotted graphically. In order to create a ROC curve, one must first plot the true positive rate (TPR) versus the false positive rate (FPR). A test's true positive rate is the fraction of tested positives that match the expected positives $\frac{TP}{TP+FN}$. The false positive rate $\frac{FP}{TN+FP}$ is the percentage of negative observations that are mistakenly projected to be positive. If a classifier only provides back the class it has predicted, it only has one point on the ROC plot. Instead, we have developed a curve by adjusting the score threshold for probabilistic classifiers, which provide a probability or score to each instance that represents the degree to which it belongs to one class rather than another [7].

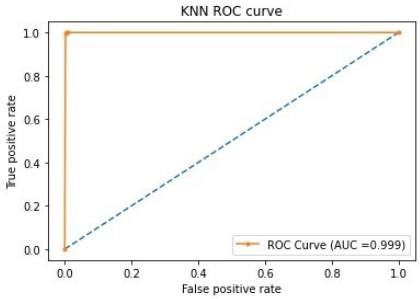

Fig. 4. ROC for KNN

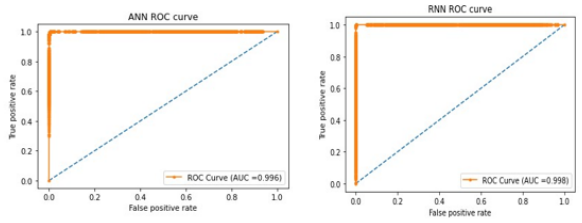

Fig. 5. ROC for ANN and RNN

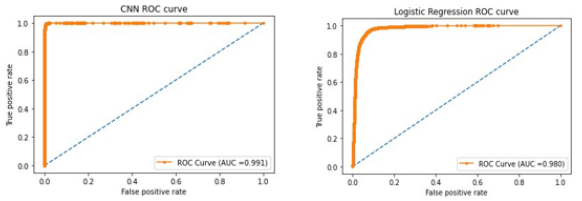

Fig. 6. ROC for CNN and Logistic Regression

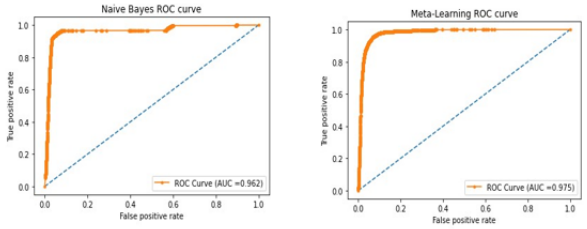

Fig. 7. ROC for Naive Bayes and Meta Learning

E. XAI Analysis (SHAP)

SHAP is an abbreviation for "SHapley Additive exPlanations". It is a technique for determining the effect of a given factor on the value of the target variable. Every model can be understood with the help of SHAP. One key idea is that a feature's significance is dependent not only on that feature but also on all of the features included in the dataset as a whole. We utilize SHAP in KNN, Logistic regression, Naive Bayes, ANN, CNN and RNN to determine the relative importance of each feature.

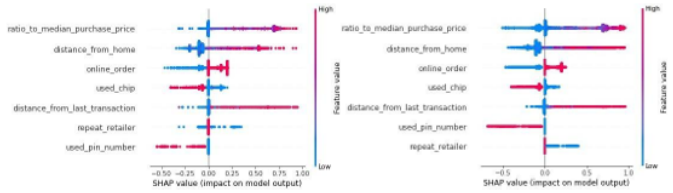

Fig. 8. XAI (SHAP) for KNN and Logistic Regression

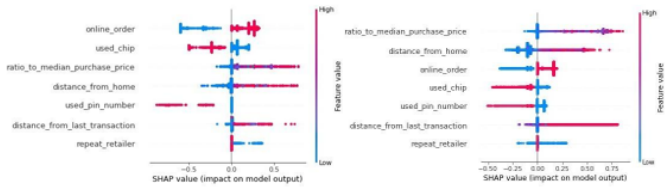

Fig. 9. XAI (SHAP) for Naive bayes and ANN

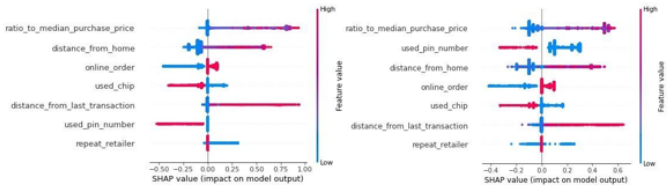

Fig. 10. XAI (SHAP) for CNN and RNN

Each column in every row represents a record in the dataset. There is a hierarchy established for the features, from most important to cheapest. For all models except Naive Bayes, the ratio to median purchase price variable stands out as the most crucial one. A higher value for this feature has a more advantageous effect on the target. This contribution will be increasingly negative as this value decreases. SHAP is a highly effective method when it comes to explaining models that can't grasp the value of features on their own [8] .

F. Comparison

Some related papers also showed their results based on the similar dataset. From our model, from Table (V), we showed that KNN (When k=3) has scored the highest accuracy and it's around 99.75%. Besides, the precision, recall, F1-score are also scored the highest value among the other models. However, we have made a comparison among some other relevant papers with our thesis research and displayed the result in Table (VI).

Reference	Best Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
[9]	Isolation Forest	97	99.8	99.45	99.75
[10]	ANN with GA	99.83	50.70	97.27	66.66
[11]	ANN	99.48	21.34	86.39	-
[12]	AE	99.90	89.55	57.14	69.77
[13]	Xgboost	96.44	96	97	96
[14]	Deep NN	98.12	-	83.52	87.65
Our Model	KNN	99.75	99.58	99.91	99.75

TABLE VI
COMPARISON BETWEEN RELEVANT PAPERS AND OUR MODEL

G. Discussion

Table (V) demonstrates the accuracy of our balanced datasets. K-Nearest Neighbor (KNN) has delivered the greatest accuracy for both processes among the two processes

we utilized for simulation and evaluation. The accuracy of KNN (when k=3), CNN, RNN, ANN, Meta-Learning, Logistic Regression and Naive Bayes Support in full-dataset are 99.89%, 99.87%, 99.81%, 99.76%, 98.1%, 95.91% and 95.14% respectively. Additionally, We also performed simulations using 87,403 fraud transactions to test the models' reliability. The accuracy of KNN (when k=3), RNN, ANN, CNN, Meta-Learning, Naive Bias and Logistic Regression in balanced-dataset are 99.75%, 99.66%, 99.61%, 99.26%, 98.1%, 94.01% and 93.92% respectively. Finally, we can easily get the accuracy of both datasets. Among them, the accuracy of full-dataset models is more than the accuracy of balanced-dataset models where the number of non-fraud is just 87,403. In order to evaluate the new balanced datasets, we obtained model performance metrics and compared them with some of the relevant papers. Isolation Forest has the best accuracy of 97% shown in [9]. Besides, in this paper [10], ANN with GA got 99.83% accuracy whereas its precision is around 50.70% which is very low here. With these outcomes in mind, the best algorithm we have is KNN, which achieves an 99.75% accuracy, 99.58% precision, 99.91% recall, and 99.75% F1-Score. When compared to other findings, our KNN performs best in terms of accuracy, recall, precision and F1-Score. Possible explanations include our effective approach of splitting the original unbalanced dataset into balanced datasets with varying quantities of non-fraud transactions.

VI. Conclusion

One cannot overstate the dishonesty of credit card fraud. Fraud using credit cards is a growing problem for banks. New fraud strategies are often developed by fraudsters. Due to the dynamic nature of fraud, a powerful classifier is necessary. This paper reviews recent developments in this sector and identifies the most prevalent types of fraud and how they might be detected. Along with the method, pseudocode, description of its implementation and experimental findings for detecting fraud, this paper also explains how machine learning and deep learning might be applied to achieve better outcomes. The primary goal of any fraud detection system should be to accurately forecast fraud situations while minimizing false positives. According to the specifics of each business scenario, ML and DL approaches may or may not be effective. When it comes to machine learning and deep learning, the nature of the input data is the most important determining element. The effectiveness of a model for identifying Credit Card fraud relies heavily on the amount of characteristics it utilizes, the number of transactions it processes, and the level of correlation between those features. With the right data trimming, noise reduction, feature extraction and model training, real-time credit card fraud detection is now a reality. K-Nearest Neighbor (KNN) had the highest accuracy (99.75%) out of the seven supervised machine learning and deep learning models. Here, we used a novel approach, Meta-Learning. While our strategy is effective and satisfactory in the end, it does take time to find enough legitimate data to train the model. There was little to no difference in performance among algorithms, but we can speculate that, with further training on additional real-world data, accuracy and precision might improve. We are not yet at 100% accuracy, despite having implemented a number of data mining techniques, but we are working to improve it by integrating several algorithms. We will try to reduce false negatives as our learning progresses, our accuracy may improve. Future experiments will include testing whether or not we can improve results by training models with additional data and using genetic algorithms [15].

References

[1] Jorge Galindo and Pablo Tamayo."Credit risk assessment using statistical and machine learning:basic methodology and risk modeling applications".In:Computational economics 15.1 (2000), pp. 107–143. doi:https://doi.org/10. 1089 / big. 2014 . 0018. url: https:// link. springer. com/ article/ 10 . 1023 / A: 1008699112516?error=cookies_not_supportedcode=b5845b26-7112-45cc- a601-7a5b71457440.

[2] Paulius Danenas, Gintautas Garsva, and Saulius Gudas. "Credit risk evaluation model development using support vector based classifiers". In: Procedia Computer Science 4 (2011), pp. 1699–1707. doi: https://doi.org/10.1016/j. procs.2011.04.184.

[3] Peter Martey Addo, Dominique Guegan, and Bertrand Hassani. "Credit risk analysis using machine and deep learning models". In: Risks 6.2 (2018), p. 38. doi: 10.3390/risks6020038.

[4] Trilok Nath Pandey et al. "Credit risk analysis using machine learning classifiers". In: 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS). IEEE. 2017, pp. 1850–1854. doi: 10.1109/ICECDS.2017.8389769.

[5] Bart Baesens et al. "Benchmarking state-of-the-art classification algorithms for credit scoring". In: Journal of the operational research society 54.6 (2003), pp. 627–635. doi: 10.1057/palgrave.jors.2601545.

[6] DHANUSH NARAYANAN R. Credit Card Fraud. 2022. url: https://www. kaggle.com/datasets/dhanushnarayananr/credit-card-fraud.

[7] Carmen Chan. What is a ROC curve and how to interpret it. Aug. 2022. url: https://www.displayr.com/what-is-a-roc-curve-how-to-interpret-it/.

[8] Gianluca Malato. How to explain neural networks using shap. Nov. 2021. url: https:// www. yourdatateacher. com/ 2021 / 05 / 17 / how- to-explain- neural- networks-using-shap/.

[9] Shubham Jaiswal, R Brindha, and Shubham Lakhotia. "Credit Card Fraud Detection Using Isolation Forest and Local Outlier Factor". In: Annals of the Romanian Society for Cell Biology (2021), pp. 4391–4396.

[10] Anuruddha Thennakoon et al. "Real-time credit card fraud detection using machine learning". In: 2019 9th International Conference on Cloud Computing, Data Science Engineering (Confluence). IEEE. 2019, pp. 488–493.

[11] John O Awoyemi, Adebayo O Adetunmbi, and Samuel A Oluwadare. "Credit card fraud detection using machine learning techniques: A comparative anal- ysis". In: 2017 international conference on computing networking and infor- matics (ICCNI). IEEE. 2017, pp. 1–9.

[12] Arjwan H. Almuteer et al. "Detecting Credit Card Fraud using Machine Learn- ing". In: International Journal of Interactive Mobile Technologies (iJIM) 15.24 (Dec. 2021), pp. 108–122. doi: 10 . 3991 / ijim. v15i24 . 27355. url: https:// online-journals.org/index.php/i-jim/article/view/27355.

[13] Krishna Kumar Mohbey, Mohammad Zubair Khan, and Ajay Indian. "Credit Card Fraud Prediction Using XGBoost: An ensemble Learning Approach". In: International Journal of Information Retrieval Research (IJIRR) 12.2 (2022), pp. 1–17.

[14] Dinara Rzayeva and Saber Malekzadeh. "A Combination of Deep Neural Net- works and K-Nearest Neighbors for Credit Card Fraud Detection". In: arXiv preprint arXiv:2205.15300 (2022).

[15] D Tanouz et al. "Credit card fraud detection using machine learning". In: 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS). IEEE. 2021, pp. 967–972.