JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 12(4), DECEMBER, 2024
ISSN: 2277-0011; Journal homepage: www.atbuftejoste.com.ng

Data Mining and Fraud Detection on E-Commerce Transactions Using Comparative Analysis of K-Nearest Neighbors and Deep Neural Network

Tobi Solomon, Ibrahim Mohammed, Sa'adatu Abdulkadir
Computer Science Department,
Faculty of Computing, Kaduna State University.

ABSTARCT

This work focuses on creating and assessing machine learning models algorithms and methods for identifying fraudulent e-commerce transactions. Features including transaction type, amount, balance, and date are extracted from transaction data analysis and classified as authentic or fraudulent based on transaction limits and balance consistency. K nearest neighbours and deep neural networks train and test the dataset using standardized features. Once it has been divided into training and testing sets; the models' recall, accuracy, and precision are assessed. According to the results, the deep neural network model outperformed then K-nearest neighbours with an accuracy of 98.12% and recall of 83.52% comparing our result to other research works the deep neural network also outperformed results from other research which used fraud detection tools, such as Logistic Regression, random forest and SVM so on in terms. This is because deep neural networks can handle complex changes in data, and it also provides useful probabilistic insights, which makes it a powerful tool. The visualization of prediction probabilities is done using ***MATLAB 9.0 R2016a*** *to improve interpretation and understanding. To further increase performance, future work will involve improving feature extraction, more real-world data sets and investigating more sophisticated models. This study shows how KNN and DNN may be used as machine learning models with an intuitive user interface to develop a reliable method for identifying fraudulent transactions and improving data mining and e-commerce transactions.*

ARTICLE INFO
Article History
Received: September, 2024
Received in revised form: November, 2024
Accepted: November, 2024
Published online: December, 2024

KEYWORDS
Economic, Fraud, K-nearest Neighbor, Deep Neural Network, Principal Component Analysis

INTRODUCTION

In the current digital era, fraudsters now have more ways to take advantage of weaknesses in financial systems due to the growth of online transactions, e-commerce, and digital transaction. With increasingly complex schemes targeting people, companies, and governments, cybercrime is on the rise (Świątkowska, 2020, Wainwright & Cilluffo, 2022). Among the many illicit behaviors that fall under the umbrella of e-commerce fraud include identity theft, credit card fraud, check fraud, account takeover, phishing schemes, and more(Jena et al. 2023). Unauthorized account access, fraudulent transactions, identity theft, and other fraudulent practices are examples of these acts. To stop these scams, banks use a multifaceted strategy. In order to lessen the possible harm, ecommerce fraud detection systems seek to identify and address these fraudulent acts in real-time or almost real-time. In order to quickly identify suspicious activity, financial firms use fraud detection systems that can analyze transaction data in real-time (Essien, 2019). Robust customer authentication techniques, such biometric verification and two-factor authentication (2FA), are used to confirm the identity of users gaining access to accounts.

Corresponding author: Tobi Solomon.
tobisolomon.i@gmail.com
Computer Science Department, Faculty of Computing, Kaduna State University, Kaduna.

Aiswarya, 2020 there are several difficult obstacles to overcome while trying to identify ecommerce fraud in ecommerce and financial organization. The fact that fraudulent techniques are always changing is one of the main obstacles. Fraudsters create novel ways to take advantage of system flaws, making it hard for conventional rule-based systems to stay up to date. The endeavor is made more difficult by the enormous volume of data and transactions that banks process. In addition to being time-consuming, manual analysis frequently fails to detect minute trends and irregularities that could be signs of fraud. Furthermore, it can be challenging to strike a balance between fraud detection and client ease; too stringent security measures may result in false positives and annoy legitimate customers, thereby turning them off. The consequences of financial fraud can potentially disrupt economies, raise living expenses, and undermine consumer confidence (Alfaiz and. Fati, 2022). Financial fraud can be termed as the deliberate employment of unlawful procedures or tactics to obtain financial gain (Ashfaq et al 2022). Zhu et al 2021 described the complex patterns and high dimensional nature of frauds and makes the statistical methods less effective, as such machine learning models were deployed,

Our capacity to generate and gather data has been growing significantly in recent years. We now have a vast amount of data thanks to the widespread adoption of bar codes and online data bases such as kaggle.com, Data.gov etc. for the majority of commercial goods, the computerization of many companies and government operations, and advancements in data collection techniques. Business management, government administration, scientific and engineering data management, and numerous more applications have made use of millions of databases. Due to the availability of strong and reasonably priced database systems, it is seen that the number of these databases continues to increase quickly. The rapid expansion of databases and data has created a pressing demand for innovative methods and resources that can automatically and intelligently turn processed data into knowledge and information. As e-commerce, e-retail, Ecommerce and online shopping, have grown in popularity, financial organization have seen an unprecedented surge in online customers, and new channels for online buyers have opened the door for fraud and abuse. By balancing security and privacy, these firms must overcome these obstacles and keep updating their fraud detection efforts. Dynamic fraudulent behavior patterns that mimic real-world operations, limited and unbalanced E-commerce transaction datasets, and the impact of testing methodologies, variable selection, and identification strategies on fraud detection effectiveness are some of the difficulties that fraud detection must overcome. Furthermore, it is difficult to discern between past legitimate operations and possible present or future fraudulent activity due to the continuously shifting nature of data throughout time. According to recent research, conventional manual detection carried out by humans can be with a lot of errors as a result of the databases' size. The inability of individual human specialists to identify recently discovered fraud patterns dispersed throughout the database and the inability to identify fraudulent activity as soon as it is performed .these are significant disadvantages of manual detection. This paper looks at the detection of Ecommerce and financial institution frauds so as to improve the security of the financial system. It also aims to investigate and assess how well data science methods—specifically machine learning and real-time monitoring—work at identifying and stopping fraudulent financial transactions. More so to analyze and compute an Ecommerce data which has fraudulent activity and identify fraudulent activity using k-nearest neighbor techniques and deep neural network with an objective of evaluating the performance of the k-nearest neighbor models and Deep neural Network in detecting fraudulent Ecommerce transactions using a set of predefined metrics, such as accuracy, precision, recall, proposing an innovative method that leverages the strengths of both models to improve the accuracy of e-commerce fraud detection, potentially leading to

Corresponding author: Tobi Solomon.
tobisolomon.i@gmail.com
Computer Science Department, Faculty of Computing, Kaduna State University, Kaduna.

a more robust and reliable fraud detection system there by extending the works of Sharma, et al. (2021) going further to try out performance of other algorithms like K-Nearest neighbors and deep neural network

The rest of the paper is organized as follows, section II contains the related work done in the field of fraud detection section III contains the proposed model flow diagram, different algorithms used in the research section IV contains the experimental result and V conclusion and recommendation

RELATED WORK

The detrimental effects of ecommerce fraudulent activities, monitoring on consumers and financial institutions have received a lot of attention lately. Since dishonest behavior is unexpected, it is challenging for standard rule-based systems to deal with it. As a result, scientists are using machine learning to boost the accuracy and speed of detection. An overview of recent developments in machine learning-based credit card fraud detection research is provided.

Comparative Analysis and Combination of Algorithms in Fraud Detection

Sharma, et al. (2021) conducted research on "Machine Learning Model for Credit Card Fraud Detection - The study addressed the escalating issue of fraudulent activities in credit card transactions in today's cashless society. Recognizing the critical need for a systematic fraud detection system, the researchers utilized various machine learning algorithms, including Random Forest, Logistic Regression, Support Vector Machine (SVM), and Neural Networks. These algorithms were trained on a given dataset to develop a predictive model for fraud detection. The study conducted an in-depth comparative analysis of the accuracy and performance metrics of each algorithm, including F1 score, to evaluate their effectiveness in detecting fraudulent transactions. Through this analysis, the researchers aimed to identify the most suitable algorithm for fraud detection in credit card transactions. Ultimately, the study found that the Artificial Neural Network (ANN) exhibited the highest performance with an F1 score of 0.91, suggesting its efficacy in detecting fraudulent activities. Vaman et.al, 2024 used K-nearest neighbors, random forest, and logistic regression in a comparative study on credit card fraud detection. Their analysis revealed each algorithm's advantages and disadvantages, providing helpful guidance on algorithm selection for fraud detection applications. Yunlong et al. 2023 used random forest and logistic regression approaches to investigate credit card fraud detection. Their study highlighted how important machine learning technologies are used for identifying fraudulent transactions. Through an analysis of logistic regression and random forest outputs, the authors demonstrated the benefits of ensemble methods for fraud detection. Shukla and Tiwari 2020 presented a random forest model and deep learning artificial neural network (ANN). Their findings demonstrated the effectiveness of deep learning algorithms as a fraud detection tool. The authors improved detection accuracy and durability against fraud attempts by combining ANN with random forest.

Evaluation of Random Forest as a Fraud Detection Tool and Its Performance

Baig and Jai Sharma 2022 using an enhanced random forest approach combined with logistic regression. The research focuses on the improvements in accuracy that the best random forest approach provides. The scientists improved the efficacy of detection and reduced the false positive rates by modifying the settings of their system. Aiswarya 2020 presented an improved random forest method that incorporates user interface enhancements to identify credit card fraud. Improving the fraud detection systems' usability and efficacy was their main goal. The study came to the conclusion that efficient fraud detection systems need user-centered design. Lavanya, 2023 combined a random forest classifier with a logistic regression classifier. Using the performance indicators of the two algorithms, they were able to determine which algorithm should be used for fraud detection activities in order to achieve the goal of identifying credit card fraud.

Corresponding author: Tobi Solomon.
tobisolomon.i@gmail.com
Computer Science Department, Faculty of Computing, Kaduna State University, Kaduna.

JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 12(4), DECEMBER, 2024
ISSN: 2277-0011; Journal homepage: www.atbuftejoste.com.ng

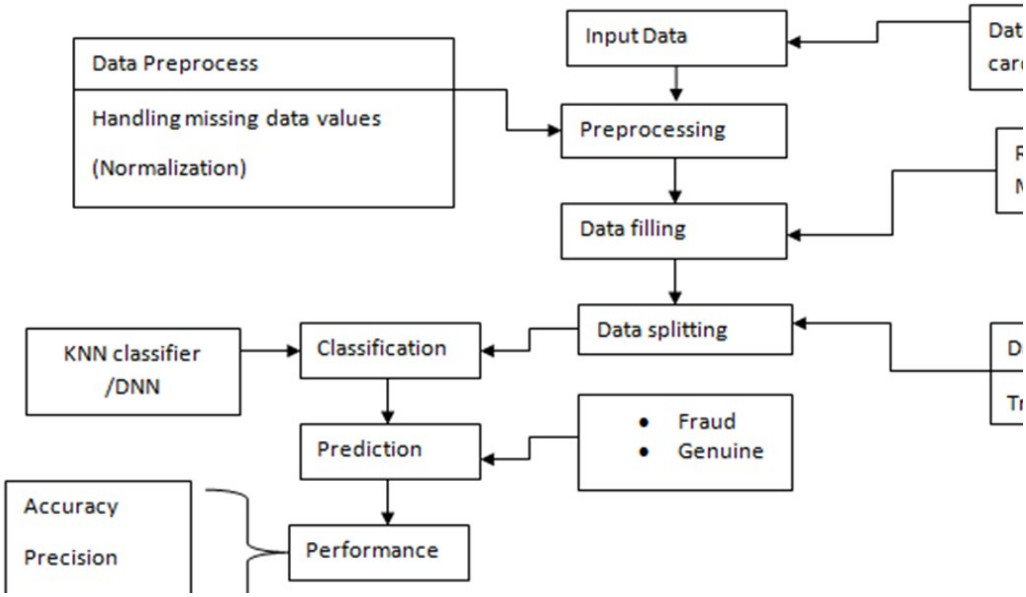
Data Balancing and Other Machine Learning Technique in Fraud Detection

Bouguettaya et al 2023 identified the need for machine learning and deep learning algorithms to drive a considerable number of datasets. Google Assistant, Alexa, Uber, and Siri are impressive existing services based on big data. The authors tried to Figure out all the possibilities of ML and DL in business intelligence. The big challenge is to design effective techniques to manage a huge amount of data. Jenipher, et al. 2021 developed data-balancing learning techniques for use in credit card fraud detection systems. Their goal is to address class mismatch issues so that fraud detection systems can identify fraud more accurately. Oakum and Özdemir 2023 looked studied the usefulness of isolation forests, auto-encoders, and one-class SVMs in credit card fraud detection investigations. They gained a significant deal of knowledge on the effectiveness of various anomaly detection technologies and how effectively they work to identify fraud.

Machine Learning Methodologies for Fraud Detection

Taylia et al 2024 did a work on machine learning algorithms and predictive traits to detect credit card fraud. Their work contributes to feature selection methodologies for fraud detection by examining the utility of predictive characteristics in detecting fraudulent transactions. Khatti, et. al, 2021 , using random under sampling improved the detection of credit card fraud. Their findings demonstrated the effectiveness of random under sampling techniques in handling unequal datasets in applications related to fraud detection. Thiyagarajan and Anbazhagan 2023 examined the confusion matrix of personal loan fraud detection using a novel random forest algorithm and a linear regression approach. Their investigation provides information on the effectiveness of various fraud detection techniques.

PROPOSED MODEL



Corresponding author: Tobi Solomon.
tobisolomon.i@gmail.com
Computer Science Department, Faculty of Computing, Kaduna State University, Kaduna.

JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 12(4), DECEMBER, 2024
ISSN: 2277-0011; Journal homepage: www.atbuftejoste.com.ng

Data set

For the purpose of fraud detection, data mining and fraud detection on e-commerce transactions dataset from kaggle.com was used. This dataset contains 284,807 credit card transactions, each transaction is represented by: 30 principal components extracted from the original data; the time from the first transaction; the amount of money.

The transactions have two labels: one for fraudulent and zero for genuine transactions. Only 492 transactions in the dataset are fraudulent.

Table 3.1 Sample Data Transaction data used for training set and testing from *www.kaggle.com*

N28 | 1

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Transaction ID	Customer ID	Transaction A	Transactio	Payment M	Product Catego	Quantity	Customer	Customer Loca	Device Used	IP Address	Shipping Address	Billing Address
2	15d2e414-8735	d1b87f62-51b2	58.09	########	bank tran	electronics	1	17	Amandaboro	tablet	212.195.49.1	Unit 8934 Box 0058	Unit 8934 Box
3	0bfee1a0-6d5e	37de64d5-e901	389.96	########	debit card	electronics	2	40	East Timothy	desktop	208.106.249.	634 May Keys	634 May Keys
4	e588eef4-b754	1bac88d6-4b22	134.19	########	PayPal	home & garde	2	22	Davismouth	tablet	76.63.88.212	16282 Dana Falls Suite	16282 Dana Falls
5	4de46e52-60c3	2357c76e-9253	226.17	########	bank tran	clothing	5	31	Lynnberg	desktop	207.208.171.	828 Strong Loaf Apt. 646	828 Strong Loaf
6	074a76de-fe2d	45071bc5-9588	121.53	########	bank tran	clothing	2	51	South Nicole	tablet	190.172.14.1	29799 Jason Hills Apt.	29799 Jason Hills
7	4e707452-7c8a	29616b04-2d5c	166.41	########	bank tran	toys & games	2	34	Herreramouth	tablet	202.237.29.5	5699 Brittany Villages	120 Kristi Dale
8	7ed952fe-8ae1	fe21ae29-ba4c	92.88	########	PayPal	toys & games	2	14	Ramosfort	tablet	13.45.27.192	727 Gibson Islands Apt.	727 Gibson
9	0b2fb5aa-7171	024257c3-5671	318.14	########	credit car	health & beau	4	42	Port Emily	desktop	131.141.230.	3914 Davis Union	3914 Davis Union
10	1f52366c-7f40	f17640ca-49da	47.92	########	bank tran	home & garde	4	38	Carneyfurt	desktop	210.148.17.2	47893 Maldonado	47893 Maldonado
11	3f10dfde-9c4c	aab93e75-582f	121.78	########	bank tran	health & beau	4	39	Brockburgh	mobile	174.32.252.2	2334 Briana Centers	2334 Briana
12	75f19b14-516c	6a2e1397-e24a	633.39	########	PayPal	toys & games	5	20	Craneport	tablet	201.188.209.	Unit 7360 Box 5180	55423 Henry
13	0dae14e6-aca	a30a5030-1c23	56.31	########	PayPal	home & garde	3	35	West Michael	tablet	81.95.103.20	3684 Morris Inlet Suite	3684 Morris Inlet
14	fb09ac9b-8c76	2d86cadf-184e	275.87	########	credit car	home & garde	5	45	Melindafurt	mobile	105.173.82.1	4197 Lewis Way	075 Monroe Court
15	3a25fa55-ec25	7be37ed2-b2d	178.94	########	debit card	clothing	4	27	Deniseburgh	mobile	211.46.251.2	298 Taylor Canyon	298 Taylor Canyon
16	c696ffee-f01d	bffedad1-43c1	374.04	########	bank tran	home & garde	5	51	Danielmouth	tablet	25.126.229.2	32232 Omar Glens	32232 Omar Glens
17	e3ac696e-978c	4c80f103-ce9c	169.04	########	debit card	home & garde	1	54	Lamburgh	tablet	1.199.155.11	Unit 4478 Box 3826	Unit 4478 Box
18	e3c1a8ee-6455	b4e261d2-2e81	254.48	########	bank tran	electronics	4	20	West Melissa	desktop	52.160.5.136	1117 Braun Courts Suite	1117 Braun Courts
19	08ded43b-24bf	ed5b5482-fa35	266.06	########	debit card	electronics	1	41	Port Rebecca	tablet	194.26.237.2	1228 Torres Squares	1228 Torres
20	0c0345fa-406b	e852cd05-3ee9	263.28	########	debit card	toys & games	3	30	East Jonathar	tablet	166.131.101.	12872 Kevin Creek	12872 Kevin Creek
21	8501d4c2-10e4	339683a0-3028	475.76	########	PayPal	clothing	2	29	Muellerstad	mobile	196.101.90.4	3496 Jason Ports	3496 Jason Ports
22	4b798a5c-8400	26695fcc-6a94	89.38	########	credit car	health & beau	4	35	Lake Elizabet	mobile	82.111.15.22	635 Danielle Parks Suite	635 Danielle

The data contains principal components instead of the original transaction features for confidentiality reasons.

Normalization (Principal Component Analysis)

While processing the huge dataset, features selection is an important aspect, here selecting the variables, which are most relevant for future analysis. The quality of preprocessing reduces over fitting and training time and improves accuracy in case of implemented dataset. As most of the features are normalized we selected two features "Time" and Amount" these were observed It was concluded that "Time" feature have the same pattern in both fraudulent and genuine transactions. So it was concerned not import for the purpose of classification. Analysis of 'Amount' future showed that all the transactions that are more 10 000 are genuine. Therefore, this feature can be taken into consideration. Unwanted feature 'Time' is dropped. On this stage, the data normalization and scale of features were produced using principal component analysis

The principal component analysis

PCA is a linear combination of primary variables that geometrically this linear combination is a new coordinate system obtained from the rotation of the original system. The PCA method is advantageous to use if the data available has a large number of variables and has a correlation between the variables. The calculation of principal component analysis is based on the calculation of Eigen values and eigenvectors that represent the spread of data from a dataset. By using PCA, variables that were as many as n variables will be selected into k new variables called principal components, with the number of k less than n. Using only the k-principal component will produce the same value

Corresponding author: Tobi Solomon.
tobisolomon.i@gmail.com
Computer Science Department, Faculty of Computing, Kaduna State University, Kaduna.

JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 12(4), DECEMBER, 2024
ISSN: 2277-0011; Journal homepage: www.atbuftejoste.com.ng

using n variables. The outcome variable of the selection is called the principal component.

Data sampling algorithm implementation

As the dataset is highly imbalanced, the under-sampling techniques were implemented. Four following algorithms were tested in order to get the data that are more balanced:

• Repeated Edited Nearest Neighbors • Edited Nearest Neighbors • One-Sided Selection
• Neighborhood Cleaning Rule

K-Nearest Neighbor (KNN)

K-nearest neighbor algorithm is a classification algorithm that predicts the attributes of an informational point to other points based on its relative position. Its classification is based on similarity measures such as Euclidean distance, Manhattan distance measure. It is assumed that the data point in the training set that has the shortest Euclidean distance to the test point has the same unknown attribute as the test point.

The Euclidean distance measure for the KNN classifier is used in this study. The range between Euclidean (EC)'s two-point vectors (x1, x2) is determined by:

$EC = \sqrt{\Sigma (x1 - x2)2}$ k=1, 2,.....,n

Manhattan distance measure between two points (xi, yi) and (xn, yn) is a metric in which the distance between two points is the absolute difference of their Cartesian coordinate. $M = (xi - xn) + (yi - yn)$

Several methods can be utilized to select the nearest neighbors, including Euclidean distance, Manhattan distance, and cosine similarity. For instance, the Euclidean distance calculates the linear distance between two data points (Equation (2)):

$$d(x,x')=\sqrt{(x1-x'1)2+\ldots+(xn-x'n)2} \quad (2)$$

This model requires careful tuning of the parameter K. If K is set too low, the risk of over fitting increases; conversely, if K is set too high, the classification performance might become inaccurate.

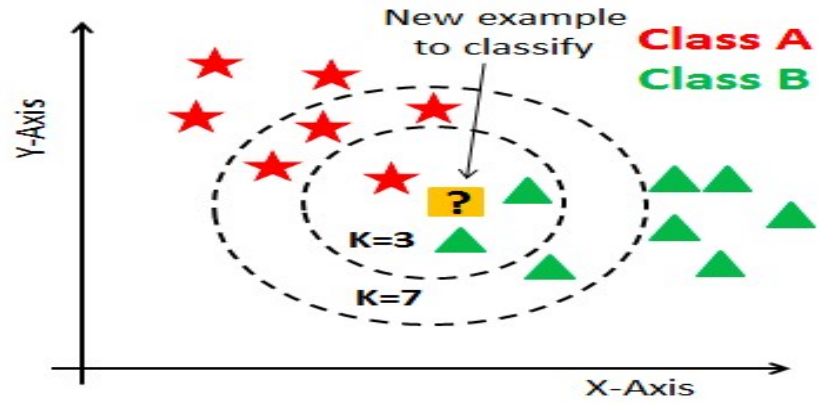


Figure 3.1 K-Nearest Neighbors

In the Figure 3.1, yellow is the query point, and we want to know which class it belongs to (red or green), With K=3, the 3 nearest neighbors of the yellow point are considered, and the class is assigned to the query point based on the majority (e.g., 2 green and 1 red — then it is of green class). Similarly, for K=5, 5 nearest neighbors are considered for the comparison, and the majority will decide which class the query point belongs to. One thing to notice here, if the value of K is even, it might create problems when taking a majority vote because the data has an even number of classes (i.e., 2). Therefore, choose K as an odd number when the data has an even number of classes

Corresponding author: Tobi Solomon.
✉ tobisolomon.i@gmail.com
Computer Science Department, Faculty of Computing, Kaduna State University, Kaduna.

JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 12(4), DECEMBER, 2024
ISSN: 2277-0011; Journal homepage: www.atbuftejoste.com.ng

and even number when the data has an odd number of classes.

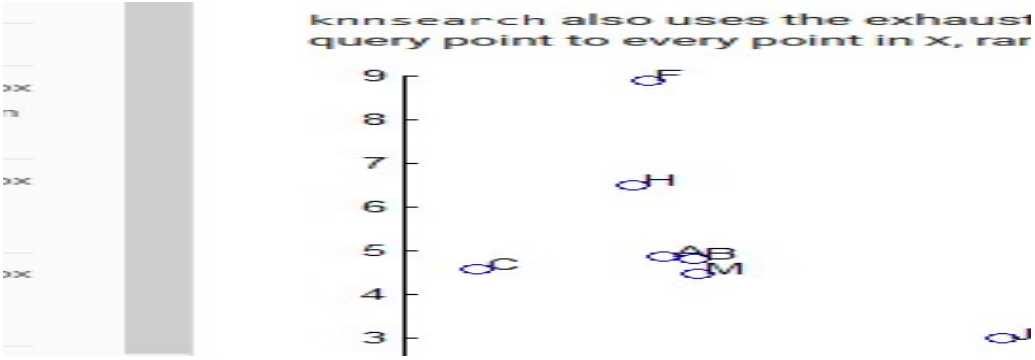
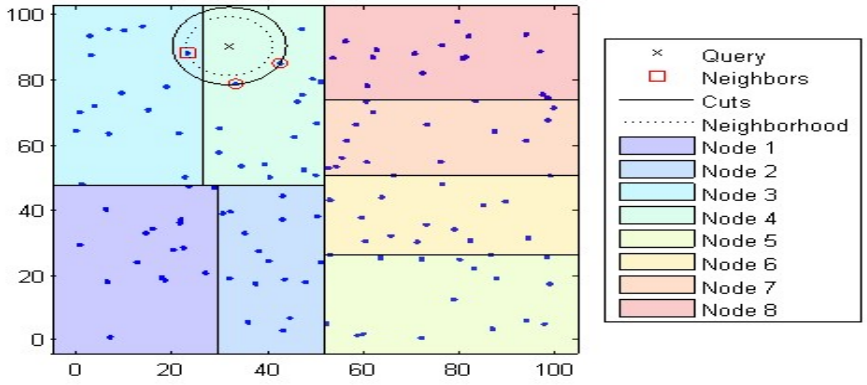


Figure 3.2 Euclidean distance b/w point (This diagram shows the k = 3 nearest neighbors)

The following is what KNN search does when you wish to determine the k-nearest neighbors to a particular query point. (1). Identifies the node that the query point is associated with. The query point (32, 90) in the example below is a part of Node 4. (2). Determines the distance between the query point and the nearest k points within that node. The points in red circles in the example below are the closest points to the query point in Node 4 and are equally spaced from it. (3). Selects every other node with an area that is the same distance from the query point to the kth closest point, in any direction. Only Node 3 crosses the solid black circle that is centered at the query point in this example.to the *k*th closest point. In this example, only Node 3 overlaps the solid black circle centered at the query point with radius equal to the separation between Node 4's nearest points. (4). Looks for any nodes that are closer to the query location within that range. The point in a red square in the example below is marginally nearer the query point than the points in Node 4.

Figure 3.3 Node to Which the Query Point Belongs



Corresponding author: Tobi Solomon.
tobisolomon.i@gmail.com
Computer Science Department, Faculty of Computing, Kaduna State University, Kaduna.

JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 12(4), DECEMBER, 2024
ISSN: 2277-0011; Journal homepage: www.atbuftejoste.com.ng

Figure 3.4 Nearest Neighbors

Deep Neural Network Model Architecture

Deep neural networks are a type of artificial neural network with multiple hidden layers, which makes them more complex and resource-intensive compared to conventional neural networks.

the model while enabling the model to process the inputs concisely for outpu solution.

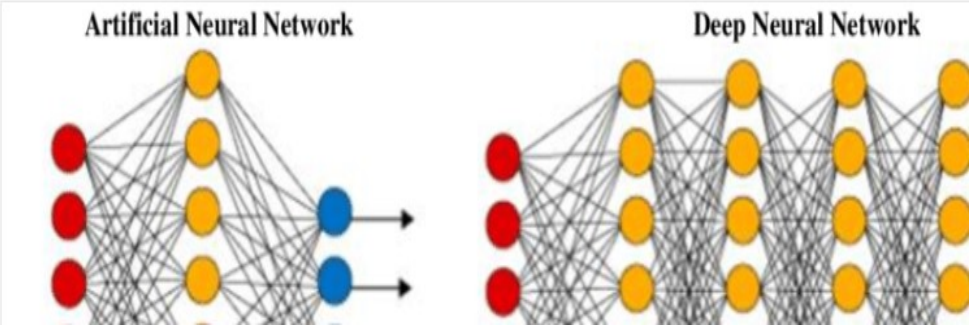


Figure 3.5 Difference between Artificial Network and Deep Neural Network

A feed-forward network with five hidden layers is the first deep learning network. A summary of the implemented network can be found in Figure 4. To avoid over fitting the model, two dropout layers and two dense layers are used. The dropout values for the first and second dropout layers are 0.5 and 0.3, respectively. The matching data set was subjected to 100 iterations in order to oversee the neural network. 32 was decided upon as the batch size. The matching data set was subjected to 100 iterations in order to oversee the neural network. 32 were decided upon as the batch size.

Corresponding author: Tobi Solomon.
tobisolomon.i@gmail.com
Computer Science Department, Faculty of Computing, Kaduna State University, Kaduna.

work.
hange
non-
ctions
4,807
ers in
on is
iginal

Layer (type)	Output Shape
dense_124 (Dense)	(None, 16)
dense_125 (Dense)	(None, 24)
dropout_37 (Dropout)	(None, 24)
dense_126 (Dense)	(None, 20)
dense_127 (Dense)	(None, 15)
dropout_38 (Dropout)	(None, 15)
dense_128 (Dense)	(None, 24)

Figure 3.6 Deep Neural Network Structure

Investigation and Analysis

The information on the figure was taken from the www.kaggle.com online database and saved in a different format. To display the data, it was saved in a CSV file that could be easily downloaded as an extension. After converting the data to CSV format, we divided it into two folders: the training set and the test set. In order to reduce the dimensionality of the data for ease of computation, these data were normalized using Matlab 8.0. Use the principle component analysis, a statistical model, to facilitate computing. Examples from the majority class that were both redundant and unclear were eliminated after under-sampling procedures were applied. The new dataset's proportion was 1:10.

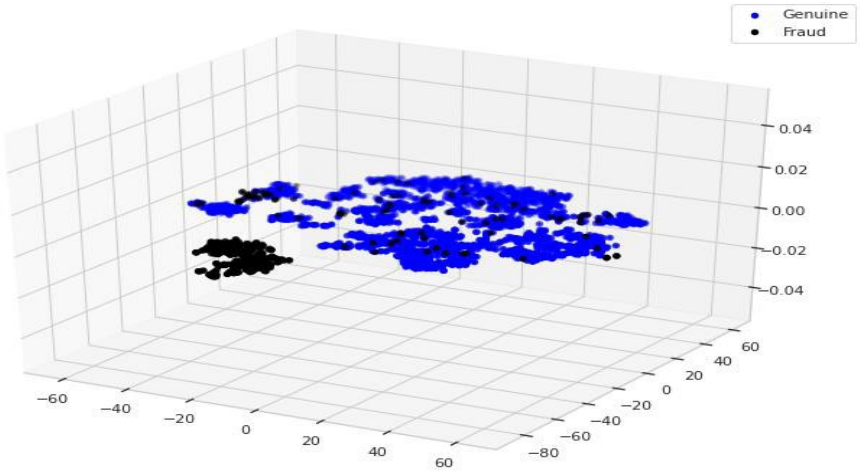


Figure 3.7 Data Points and Clusters Showing Either Genuine or Fraud (Visualization)

Figure 3.7 shows that the fraud data points have been shifted to a single cluster, whereas the normal transaction data points only contain a small number of fraud transaction data

Corresponding author: Tobi Solomon.
tobisolomon.i@gmail.com
Computer Science Department, Faculty of Computing, Kaduna State University, Kaduna.

points. After obtaining our data point, we can quickly plot a confusion matrix, a two-dimensional matrix used in classification experiments to assess a system's performance by displaying the number of data classes that were incorrectly and correctly classified. This helps us determine which data classes are most frequently misplaced. We looked at two distinct approaches to assess these machine learning models: (1) Classification accuracy, which is the ratio of accurate predictions to input samples. However, this works best when there are an equal number of samples in each class. Accuracy = total number of correct predictions /Total number of predicted made.

(2) Confusion Matrix: this gives a matrix as output and describes the complete performance of the model. Four essential measurements are utilized in evaluating the analyses, to be specific True Positive Ratio (TPR), True Negative Ratio (TNR), False Positive Ratio (FPR) and False Negative Ratio (FNR) rates metric individually. In which true positive, true negative, false positive and false negative are the quantity characterized by true positive, false positive, true negative, and false negative experiments, thus p and n are the absolute values of positive and negative class cases being tested. True positives are classes predicted to be positive and are, true negative classes are predicted as negative but are negative. False positive are classes predicted to be positive and are negative, false negatives are classes predicted to be negative but that are positive.

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

From the table 3.8 above, the rate of true positive and false negative was calculated using the parameters. Where: True positive – TP, False Negative – FN, False positive – FP, True Negative - TN

The true positive rate (TPR), (sensitivity) is giving as --- $TPR = \frac{TP}{TP+FN}$

The True negative rate (TNR) is giving as ---------- $TNR = \frac{TN}{TN+FP}$

The false positive rate (FPR) is giving as ---------- $FPR = \frac{FP}{TN+FP}$

The precision (positive predictive value) ---------- $precision = \frac{TP}{TP+FP}$

For our accuracy
Accuracy = TN +TP / TN + FP +FN +TP
Then the Precision
Precision = TP/ (TP +FP)

Corresponding author: Tobi Solomon.
✉ tobisolomon.i@gmail.com
Computer Science Department, Faculty of Computing, Kaduna State University, Kaduna.

JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 12(4), DECEMBER, 2024
ISSN: 2277-0011; Journal homepage: www.atbuftejoste.com.ng

Table 4.1 Confusion matrix for KNN

TP = 970	FP = 11
FN = 22	TN = 63

Table 4.2 Confusion matrix for DNN

TP = 975	FP = 6
FN = 34	TN = 71

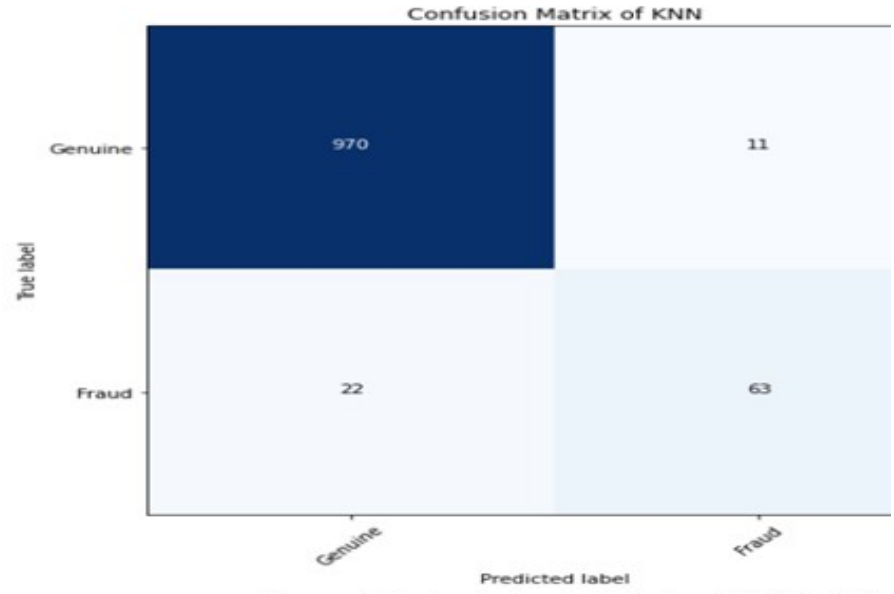


Figure 3.9 Confusion Matrix for KNN Model

In Figure 3.9 and 3.10 above , it can be observed that deep neural network demonstrated higher accuracy and recall rate comparing to KNN algorithm in which we had DNN 87.65% We will then calculate our f1 score- F1 Score then harmonizes these metrics and its given by – F1 = 2 * (precision * recall) / (precision + recall).

Corresponding author: Tobi Solomon.
tobisolomon.i@gmail.com
Computer Science Department, Faculty of Computing, Kaduna State University, Kaduna.

JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 12(4), DECEMBER, 2024
ISSN: 2277-0011; Journal homepage: www.atbuftejoste.com.ng

Table 4.3 performance measure

Performance	DNN	KNN
Accuracy	98.12%	96.90%
Recall	83.52%	74.12%
F-1 score	87.65%	79.24

Based on the research technique, the model's success rate was determined using the Area of under Curve. We discovered an unsupervised learning rate with a true positive rate on our model since the percentage of AUC of both applied algorithms is high (91.54% for the KNN model and 94.02% for the DNN model). Figures 3.9 and 3.10 are shown. The applied DNN model, however, is more perceptive in identifying fraudulent transactions. The performance of our machine learning classifier will now be visualized using the AUC ROC curve. An ROC plot is produced by ordering all predictions according to their level of confidence, even though it is only effective for binary classification tasks. Next, you begin in the lo were left corner and proceed to the right for each point

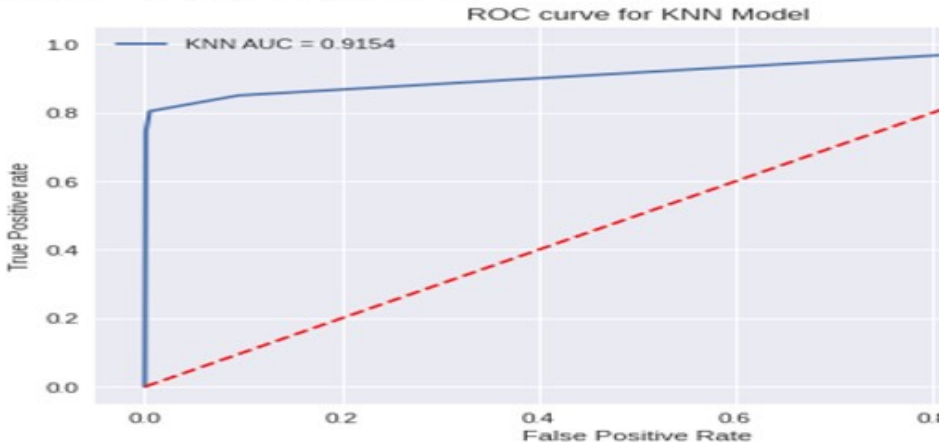
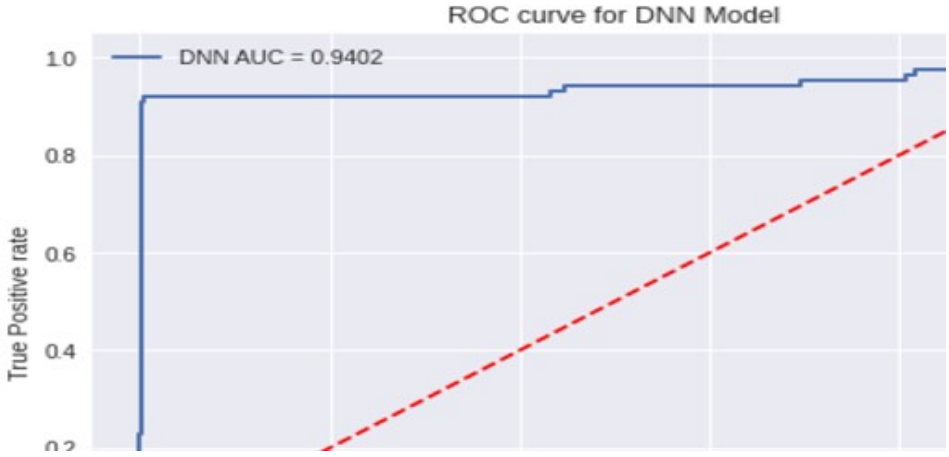


Figure 3.12 ROC curve for KNN model



Corresponding author: Tobi Solomon.
✉ tobisolomon.i@gmail.com
Computer Science Department, Faculty of Computing, Kaduna State University, Kaduna.

CONCLUSION

To sum up, our work's comparison analysis provides insight into how well the DNN and KNN algorithms identify credit card fraud. Even in situations where every algorithm performed as expected, the stark differences in the various performance metrics exposed the little benefits and drawbacks of each method. By outperforming K-nearest neighbors in terms of accuracy, recall, and F1-Score, DNN demonstrated its resilience as an adversary. The distinct approach taken by DNN, which concentrates on recovering anomalies from high-dimensional data landscapes, may account for this improved performance. DNN's ability to quickly segregate massive data sets and identify outliers makes it excellent at spotting subtle trends that indicate fraudulent activity. It has higher detection accuracy for fraudulent transactions since it produces fewer false positives. Furthermore, DNN's outstanding recall demonstrates its ability to identify a sizable portion of actual fraudulent occurrences in the dataset. The balanced F1- Score demonstrates its ability to withstand skewed datasets, which are commonly found in fraud detection systems. Despite its poor precision and performance, KNN performed similarly to DNN in terms of accuracy and recall. Although KNN uses community knowledge to make classification decisions, accuracy may be significantly impacted by its ensemble method, thereby increasing the false positive rate. KNN is still doing rather well, as evidenced by its competitive accuracy and recall rates. This is particularly true when it comes to accurately identifying the vast majority of transactions and a sizable portion of fraudulent instances. As a whole. The differences in performance metrics show how well the KNN and DNN methods work together in fraud detection applications.

RECOMMENDATION AND FURTHER WORK

In order to increase the external validity of the findings, future research should try to address these limitations of adopting data by doing tests on real-world datasets rather than adapting data from online databases. Additionally, efforts ought to focus on incorporating additional domain knowledge and thoroughly assessing how well they function in actual fraud detection scenarios. Notwithstanding these drawbacks, the study contributes to the ongoing conversation over the combination of explainable use of K-Nearest neighbors and deep neural networks and represents a significant advancement in the usage of algorithms in fraud detection.

REFERENCE

Aiswarya, C. J. "Optimized Random Forest for Credit Card Fraud Detection with User Interface." (2020).

Baig, M. Shahid Saif Ali, and K. Jaisharma. "Comparison of Novel Optimized Random Forest Technique and Logistic Regression for Credit Card Fraud Detection with Improved Precision." Journal of Pharmaceutical Negative Results (2022): 723-727.

J Essien -A Synergistic Approach for Enhancing Credit Card Fraud Detection using Forest and Naïve Bayes Models, International Journal of Innovative Science and …, 2019

Jena, Amrut Ranjan, Santanu Kumar Sen, Madhusmita Mishra, Shrutarba Banerjee, Nupur Dey, and Ipsita Saha. "A comparative analysis of financial fraud detection in credit card by decision tree and random forest techniques." In AIP Conference Proceedings, vol. 2876, no. 1. AIP Publishing, 2023.

Jenipher, V. Nisha, J. Dafni Rose, M. Sabharam, and M. Nithin. "Learning Algorithms with Data Balancing in Credit Card Fraud Detection Application." In 2021 Fifth International Conference on ISMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), pp. 1-6. IEEE, 2021.

Kumar, K. Yashwanth, and B. Vani. "An optimal approach for fraud detection by comparing random forest algorithm and support vector machine algorithm for

Corresponding author: Tobi Solomon.
tobisolomon.i@gmail.com
Computer Science Department, Faculty of Computing, Kaduna State University, Kaduna.

credit card transaction with improved accuracy." In AIP Conference Proceedings, vol. 2821, no. 1. AIP Publishing, 2023.

Lavanya, K. "A Comparison of Logistic Regression Classifier and Random Forest Classifier for the Accurate Classification of Credit Card Fraudulent Transactions." Journal of Survey in Fisheries Sciences10, no. 1S (2023): 2008-2017.

N. S. Alfaiz and S. M. Fati, ‘‘Enhanced credit card fraud detection model using machine learning,’’ Electronics, vol. 11, no. 4, 662, 2022. Innovative Technologies in Economy. 10.31435/rsglobal_ijite/30062022/7828.

Rafi, D. Mahammad, Macha Mahipal Reddy, and Kunduru Ashwini. "Securing Digital Transactions: A Random Forest-Based Approach for Online Credit Card Fraud Detection and Accuracy Assessment." Journal of cardiovascular disease research ISSN: 0975-3583, 0976-2833 VOL 13, ISSUE 02, 2022

Shukla, Nitin, and Pragya Tiwari. "ANN Deep Learning and Random Forest Model for Fraud Detection ofCredit Card Users in Ecommerce System." (2020).

T. Ashfaq, R. Khalid, A. S. Yahaya, S. Aslam, A. T. Azar, S. Alsafari, and I. A. Hameed, ‘‘A machine learning and blockchain based efficient fraud detection mechanism,’’ *Sensors*, vol. 22, no. 19, p. 7162, Sep. 2022.

Thiyagarajan, A., and K. Anbazhagan. "Confusion matrix analysis of personal loan fraud detection usingnovel random forest algorithm and linear regression algorithm." In AIP conference Proceedings, vol.2822, no. 1. AIP Publishing, 2023, using machine learning,’’ *Electronics*, vol. 11, no. 4, 662, 2022.

X. Zhu, X. Ao, Z. Qin, Y. Chang, Y. Liu, Q. He, and J. Li, ‘‘Intelligent financial fraud detection practices in post-pandemic era,’’ *Innovation*, vol. 2, no. 4, Nov. 2021, Art. no. 100176, doi: 10.1016/j.xinn.2021.100176

Corresponding author: Tobi Solomon.
tobisolomon.i@gmail.com
Computer Science Department, Faculty of Computing, Kaduna State University, Kaduna.