MACHINE LEARNING ALGORITHMS FOR CREDIT CARD FRAUD DETECTION

Amarachi Blessing Mbakwe and Sikiru Ademola Adewale

Department of Computer Science, Virginia Tech, Blacksburg, VA, USA

ABSTRACT

Fraud is a critical issue in our society today. Losses due to payment fraud are on the increase as e-commerce keeps evolving. Organizations, governments, and individuals have experienced huge losses due to payment. Merchant Savvy projects that global losses due to payment fraud will increase to about $40.62 billion in 2027 . Among all payment fraud, credit card fraud results in a higher loss. Therefore, we intend to leverage the potential of machine learning to deal with the problem of fraud in credit cards which can be generalized to other fraud types. This paper compares the performance of logistic regression, decision trees, random forest classifier, isolation forest, local outlier factor, and one-class support vector machines (SVM) based on their AUC and F1-score. We applied a smote technique to handle the imbalanced nature of the data and compared the performance of the supervised models on the oversampled data to the raw data. From the results, the Random Forest classifier outperformed the other models with a higher AUC score and better f1-score on both the actual and oversampled data. Oversampling the data didn't change the result of the decision trees. One-class SVM performs better than isolation forest in terms of AUC score but has a very low f1-score compared to isolation forest. The local outlier factor had the poorest performance.

KEYWORDS

Credit card, fraud, detection, Isolation Forest, One-class SVM, Supervised algorithms.

1. INTRODUCTION

Advancement in technology has brought about rapid innovation in payment methods for goods and services for faster and more convenient ways of payment. We are no longer in the days when the world solely relies on physical cash alone. Credit card is one of the means of payment. Credit cards are small cards issued by credit companies to their approved customers as a means of payment. It gives the cardholders the privilege of paying for goods and services and repaying the company at a prescribed billing cycle.

Despite the ease and convenience associated with using credit cards, fraudsters increase the risk of credit card usage. Credit card fraud involves the unauthorized use of credit cards or accounts for transactions without the owner's permission. Different measures have taken place to increase the security of credit cards, but fraud in credit cards increases daily. The report from Merchant Savvy shows that global losses due to payment fraud have increased from $9.84 billion in 2011 to $32.39 in 2020. They projected that payment fraud will increase to about $40.62 billion in 2027 [1]. Among all the payment frauds, credit card fraud resulted in a higher loss.

Fraud can be avoided through prevention, preventing its occurrence, or through fraud detection when the action occurs. Banks and credit companies are applying different techniques to control payment fraud, such as monitoring the risk scores in real-time, physical biometrics (voice, facial), rules, and machine learning behavioral biometrics studying patterns in human behavior [1].

DOI:10.5121/mlaij.2022.9402

Though we have many fraud detection systems, this area still demands further contribution from researchers as fraudsters keep devising new ways of operations, resulting in an increase in loss due to payment fraud, especially credit card fraud. Therefore, this paper focuses on some machine learning algorithms comparison such as decision trees, random forest classifier, logistic regression, ensemble of Logistic regression, decision trees, random forest classifier, local outlier factor, isolation forest, and anomaly detection algorithms.

The paper is organized as follows: Section 2 presents an overview of related works. Subsequently, in section 3, we discuss the methodology. In section 4, we show the results. Finally, in section 5, we present the conclusion and future direction.

2. RELATED WORK

Fraud has become a major concern to different organizations and the government as it constitutes among the major causes of loss to organizations and government, including individuals. US payment fraud statistics show that 77% of companies in the US said they had suffered fraud during digital payment. 53% of customers said that they had been victims of fraud during digital payment, and 23% of online sales ended up being fraudulent compared to 18% in 2018 [1].

Credit card fraud can be defined as a situation where a person uses a credit card belonging to someone else for personal reasons without authorization from the owner.

Research has been ongoing in this area; different researchers have applied machine learning algorithms for fraud detection. This paper [2] compared the performance of Random Forest, Support vector machine, and Logistic regression in detecting credit card fraud. They used the SMOTE sampling method to handle imbalanced class sizes. Incremental learning was used by the authors to tackle the problem of ever-changing fraud patterns. Static and incremental learning was performed and evaluated using AUC and Average precision. The result showed that SVM had the poorest performance, with a slight difference in Random Forest and Logistic regression models.

Paper [3] compared local outlier factors and isolation forests for credit card fraud detection. In paper [4], the authors proposed an ensemble learning for credit card fraud detection. They aggregated three feed-forward neural networks and two random forest algorithms. The authors combined the two algorithms based on the notion that the Neural network can be more accurate in detecting fraudulent instances while Random Forest is more accurate in detecting normal instances. They used Binomial Logistic Regression as prior art and compared their approach with it. They reported using oversampling (by replication) of minority class and SMOTE technique, but they didn’t yield encouraging results on the dataset. The 3-feedforward neural networks include L1: which consists of 3 hidden layers with 45, 68, and 102 neurons respectively, and L2 and L3 consist of 2 hidden layers of 15 and 8 neurons. The models used sigmoid activation functions. L1 and L2 were applied to 60% of normal transactions and 60% of fraudulent transactions. L3 was applied to 60% of fraudulent transactions and half of the normal transactions. The two Random Forests consist of L4 built using 300 decision trees and L5 built using 400 decision trees. The ensemble of the neural network and Random Forest yielded the best result.

[5] compares SVM, Decision Tree, Logistic Regression, and Random Forest using Kaggle dataset obtained from European cardholders that contain 284,786 transactions. The techniques were evaluated using accuracy, sensitivity, specificity, and precision. Their results show that Random

Forest performed best, with an accuracy score of 98.6%, compared to 97.7% for Logistic Regression, 97.5% for SVM, and 95.5% for Decision Tree.

In the paper [6], the authors designed an analytical framework interface with Hadoop that can read large volumes of data. It is made up of a Hadoop network for storing data from multiple sources in HDFS. SAS is used in reading the data from Hadoop and converting it into a raw data file before passing it to an analytical model for prediction. The authors compared three analytical models. Random Forest performed best in terms of accuracy, precision, and recall than Logistic Regression and Decision trees.

Paper [3] compares Isolation Forest and Local Outlier Factors using 284807 payments made by European consumers obtained from Kaggle. Isolation Forest achieved 99.72% accuracy with a precision of 0.28, recall of 0.29, and F1-Score of 0.28 while Local Outlier factor achieved an accuracy score of 99.62% with a precision of 0.02, recall of 0.02, and F1-Score of 0.02. Isolation Forest was proposed to perform better than the Local Outlier factor. The authors recorded that Isolation Forest had an accuracy of 97% in online transactions.

The approach that this paper [3] proposes, uses the Isolation Forest and Local Outlier Factor algorithms to detect anomalous activities, called outliers. The algorithm reached over 99.6% accuracy; Isolation Forest precision remains only at 28% when a tenth of the data set is taken into consideration. However, when the entire dataset is fed into the Isolation Forest algorithm, the precision rises to 33%. Local Outlier factor had a low precision and recall. The paper mentioned that the high percentage of accuracy was due to the huge imbalance between the number of valid and the number of genuine transactions. The precision of the algorithms increases when the size of the data set is increased.

In some of the papers, the authors achieved high accuracy in some of the models and based their judgment on high accuracy. This is interesting, but due to the imbalanced nature of the data, they might be considering high accuracy at the cost of misclassification of fraudulent transactions as normal transactions. Some of the works have either performed oversampling by replication or under-sampling or applied SMOTE technique to handle the imbalanced nature of the data.

In this paper, we will compare different machine learning models, Logistic Regression, Decision trees, Random Forests, Isolation Forests, Local Outlier Factors, and One-class support vector machines based on AUC and F1 scores.

3. METHODOLOGY

3.1. Dataset Description

The dataset is downloaded from Kaggle, a website for data science competition. The datasets parameters include the amount, time, class (label), and 28 principal component analysis (PCA) transformed features to protect user identities and sensitive features (v1-v28). In the dataset, there is no missing data, no NA, no empty row. The datasets show 284807 transactions of credit card holders. In this dataset, 492 transactions are found to be fraud and 284,315 transactions are found to be normal transactions. There are 31 columns and 284807 rows in this dataset. The summary statistics, exploratory analysis result, and graphical representation of the dataset are provided for clarity. Due to the structure of the data, we realized that the dataset is unbalanced. The fraud is 0.172% of all the transactions. Twenty-eight columns are transformed using principal component analysis (PCA), and three columns are not transformed by PCA. The three columns that are not transformed are time, amount, and class. After careful examination of the dataset, the class is the response while the other 30 columns are the independent variable/features/explanatory variables. The response variable takes the value of 1 if the transaction is fraud and 0 otherwise. Apart from time and amount, the other variable, v1 – v28 obtained using PCA dimensionality reduction to secure the cardholder's information.

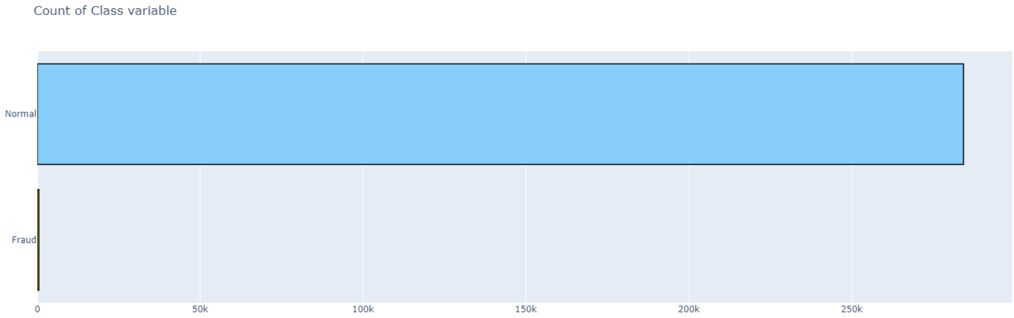

Figure 1. Distribution of the Class Labels

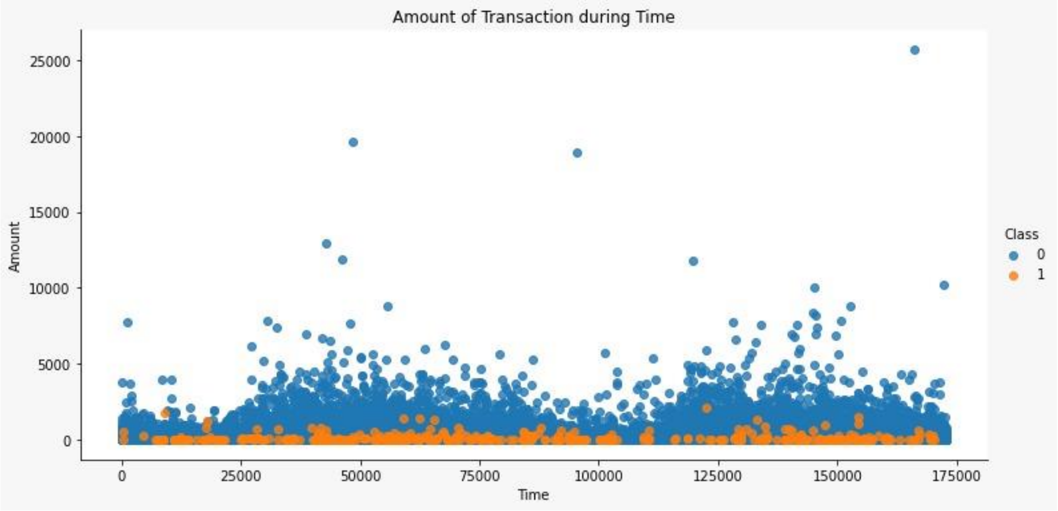

Figure 2. Amount of all transactions against Time

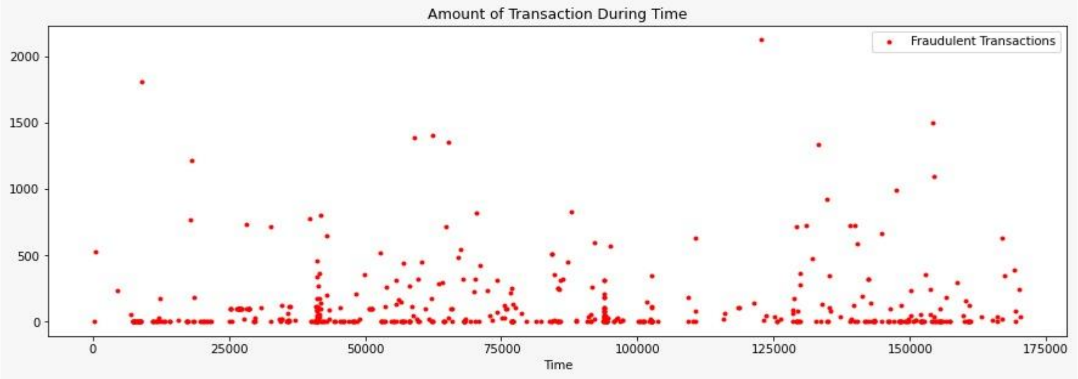

Figure 3. Fraudulent Amount transactions against Time

3.2. Models

We explored six machine learning models which include: three supervised models and three unsupervised models. The machine models were selected based on their explainability attribute and performance in several domains.

Logistic Regression is a supervised learning model that is used for classification problems. The goal is to estimate the posterior probability using continuous function and predict using discrete categorical value. Logistic regression (LR) is an arithmetic technique like linear regression since LR finds an equation that predicts the result of a twofold variable, Y, from one or more response variables, X. Nevertheless, unlike linear regression the response variables can be categorical or continuous, as the model does not firmly require continuous data. To predict group association, LR uses the log odds ratio rather than probabilities and an iterative maximum likelihood technique rather than least squares to fit the final model. This means the researcher has more freedom when using LR and the method may be more appropriate for non-normally distributed data or when the samples have unequal covariance matrices which is one of the assumptions for running regression analysis [7]. Logistic regression assumes independence among variables, which is not always met in some datasets. However, as is often the case, the applicability of the method (and how well it works, e.g., the classification error) often trumps statistical assumptions. One demerit of LR is that the technique cannot produce typicality probabilities (useful for forensic casework), but these values may be substituted with nonparametric methods such as ranked probabilities and ranked between similarity measures [8].

Decision tree is another supervised learning method that is used for classification. Decision trees (DTs) are a way to vividly establish a chronological decision process. A decision tree contains decision nodes, each with branches for each of the alternative decisions [9]. Unintended nodes (random variables) also appear in the tree, with the efficacy of each branch computed at the leaf of each branch. The expected efficacy of any decision can then be calculated based on the weighted addition of all branches from the decision to all leaves from that branch.

Types of decision trees are based on the type of response variable we have [10]. There are two types: Decision Tree which has a categorical response variable then it is referred to as a categorical variable decision tree and decision tree that has a continuous response variable is referred to as a continuous variable decision tree.

Random Forest is an ensemble method to discover the decision tree that best fits the training data by creating many decision trees and then determining the "average" one. The "random" part of the term refers to building each of the decision trees from a random selection of features; the "forest" refers to the set of decision trees. Random forest is another supervised learning algorithm. The "forest" built is an ensemble of decision trees, normally trained with the “bagging” method [9].

The general idea of the bagging method is that a combination of learning models increases the overall result. Most relevant merit of random forest is that it can be used for both classification and regression problems, which form most current machine learning systems. Random forest is also a very accessible algorithm because the default hyperparameters it uses often yield a good prediction result. Understanding the hyperparameters is forthright, and there are also not that many of them.

The main demerit of random forest is that a great number of trees can make the algorithm too slow and unproductive for real-time predictions. In general, these algorithms are fast to train, but slow to create predictions once they are trained. A more accurate prediction requires more trees, which results in a slower model. In most practical applications, the random forest algorithm is fast enough but there can certainly be situations where run-time performance is important and other approaches would be preferred.

Local Outlier Factor (LOF) is an unsupervised learning technique for spotting outliers which calculates the local density deviation of a given data value apropos its neighbors. It considers as outliers the subsets that have a significantly lesser density than their neighbors. The LOF of a point expresses the density of this point related to the density of its neighbors. If the density of a data value is considerably lesser than the densities of its neighbors (LOF $\gg$1), the data value is far from thick areas and, hence, an outlier. The result of our LOF is discussed in the next section.

Isolation Forest is another outlier detection algorithm in machine learning. Isolation forest was first introduced by [11]. They took advantage of two quantitative properties of anomalous data points in a sample: Minor - they are the minority consisting of fewer instances; and Dissimilar - they have attribute values that are very dissimilar from those of normal instances. Since anomalies are "minor and dissimilar", they are easier to “isolate” compared to normal points. Isolation Forest builds an ensemble of “Isolation Trees” (iTrees) for the data set, and anomalies are the values that have lesser average path lengths on the iTrees [11]. This is an unsupervised algorithm that learns a decision function for novelty detection: classifying new data as similar or different to the training set. The algorithm is based on the assumption that fewer instances of anomalies result in shorter path length [12]. This suggests that data instances that have shorter path length are most likely to be anomalies. The necessary input parameters for building the Isolation Forest algorithm: are the subsampling size, the number of trees and the height of the tree [12].The subsampling size was suggested to be smaller for the machine learning algorithm to function faster and yield a better detection result [13].

One Class Support Vector Machine is an unsupervised algorithm that learns a decision function for abnormal observation detection: classifying new data as similar or different to the training set. It has different kernels which include radial basis, polynomial, linear and sigmoid.

4. RESULT AND EVALUATION

In this section, we will compare supervised models; Logistic regression, decision trees, and random forest classifier models, including unsupervised models; Isolation Forest, Local outlier factor, and One-class Support vector machine models. We conducted several experiments on the credit card data set.

4.1. Dataset Processing

The dataset was splitted into training and test sets in the ratio of 70:30 respectively. The choice for the ratio division is based on empirical studies the best results are obtained if we use 20-30% of the data for testing, and the remaining 70-80% of the data for training. The models were built, and the performance was evaluated using 10-Fold cross validation to ensure the models were not overfitting on the training data set. Then, the model performance on unseen data was tested using the test data.

The supervised models were first implemented on the imbalanced data set. Afterward, to handle the imbalanced nature of the data, the data was oversampled using Smote (Synthetic Minority Over-sampling) technique. This works by generating synthetic samples.

4.2. Evaluation Metrics

Most machine learning algorithms are evaluated using predictive accuracy, but this is not appropriate in fraud detection because they are mostly imbalanced. In terms of imbalanced data, we mean that the proportion of data points in each class are not approximately equal. In fraud detection, we are very much interested in correctly predicting the minority class (fraudulent class).

The evaluation metrics adopted in this paper for evaluation and comparison of our models are standard area under the ROC curve (AUC), ROC and F1-score. The Receiver Operating Characteristic (ROC) curve was proposed as a standard technique for evaluating the performance of classifiers over a range of trade-offs between true positive and false positive error rates [14,15]. The ROC curve accepted performance metric is Area Under the Curve [15,16].

4.3. Results

Testing the predictive performance on the test data which constitutes 30 percent of the entire data, the result of each model is given in table 1.

Table 1. Results of AUC and F1-Score Comparison Among Models.

Models	Dataset	AUC	FI-Score for Fraudulent Class	F1-Score for Normal Class
Logistic Regression	Original data	0.978	0.74	1.00
Decision trees	Original data	0.897	0.74	1.00
Random Forest Classifier	Original data	0.962	0.86	1.00
Logistic Regression	Oversampled data	0.981	0.10	0.99
Decision trees	Oversampled data	0.897	0.74	1.00
Random Forest Classifier	Oversampled data	0.987	0.87	1.00
Isolation Forest	Original data	0.626	0.25	1.00
Local Outlier Factor	Original data	0.499	0.00	1.00
One-Class SVM	Original data	0.721	0.07	0.99

Analysing the obtained results, we can see that the supervised algorithms achieved a high AUC score and good f1-score. Random forest classifier has the highest AUC score and f1-score. The result obtained by the random forest classifier on the oversampled data shows that oversampling can help correct imbalance data. Oversampling can help improve model performance, avoid overfitting, and reduce computational time. Local outlier factor has the lowest performance.

4.4. Model Comparison

The models are compared using ROC curve, AUC, and F1-score. The ROC curve helps to visualize the performance of the model based on AUC. Random forest classifier achieves the highest AUC and f1 score, followed by Logistic regression while local outlier factor has the poorest performance.

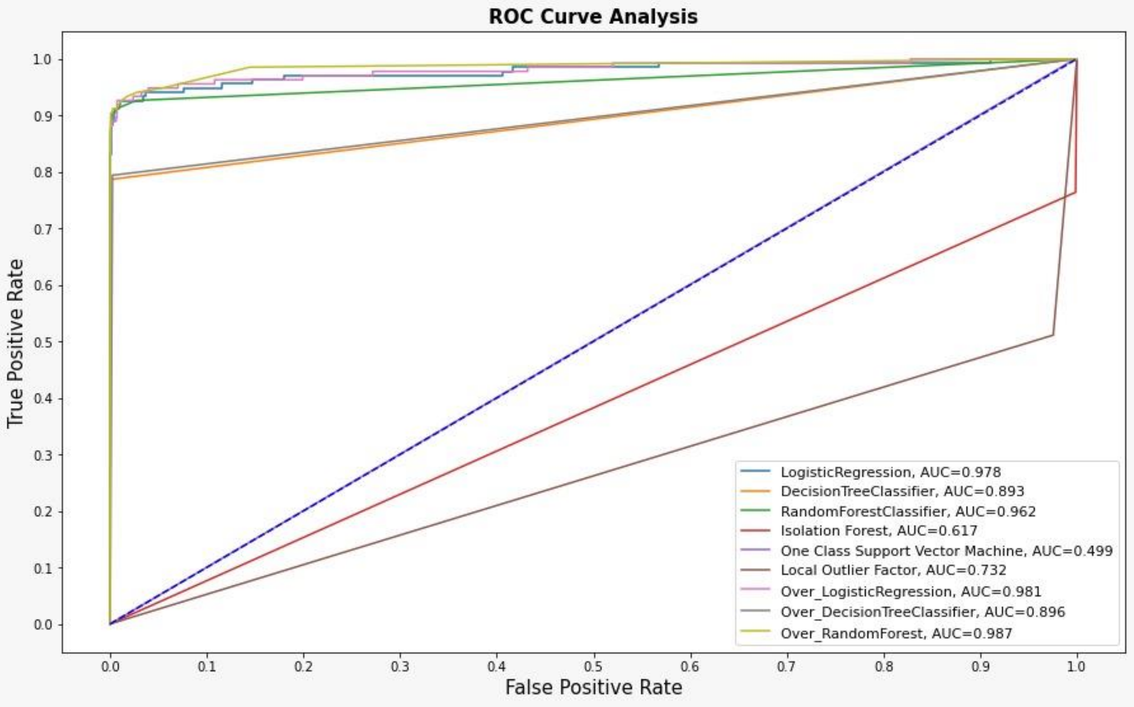

Figure 4. Model Comparison using ROC curve

5. Conclusion and Future Work

Credit card fraud is an important problem that calls for efficient solutions. Having an efficient solution will drastically reduce the loss incurred by the government, companies, and individuals. In this paper, we compared both supervised and unsupervised models. Random Forest classifier outperformed all the other models based on the AUC, followed by logistic regression, decision trees, one-class support vector machine, isolation forest and local Outlier factor.

Oversampling the data improved the performance of the random forest classifier on unseen data. Also, we can see here that the supervised algorithms performed better than the unsupervised algorithms, which shows that supervised algorithms can be used in fraud detection with the advantage of less computational time and ease of interpretation.

For a better performance of predicting the fraudulent transactions, a better well-informed data might improve the models generally. The data used in this paper is a bit challenging because some information has been transformed due to confidentiality of customer’s information. For future research, we will try to obtain well-informed data or use synthetic data that imitates real-life transactions. Also, for future work, we will try a context-aware learning approach following the idea in [17].

REFERENCES

[1] Merchant savvy, (2020) Global Payment Fraud Statistics, Trends and Forecasts.https://www.merchantsavvy.co.uk/payment-fraud-statistics/. Updated: October 2020.

[2] Puh, M., & Brkić, L. (2019). "Detecting Credit Card Fraud Using Selected Machine Learning Algorithms," 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 1250-1255.

[3] Hyder John, Sameena Naaz, (2019) "Credit Card Fraud Detection using Local Outlier Factor and Isolation Forest," International Journal of Computer Sciences and Engineering, Vol.7, Issue.4, pp.1060-1064, 2019.

[4] Ishan Sohony, Rameshwar Pratap, and Ullas Nambiar, (2018) " Ensemble learning for credit card fraud detection," In Proceedings of the ACM India Joint International Conference on Data Science and Management of Data (CoDS-COMAD '18). Association for Computing Machinery, New York, NY, USA, 289–294. https://doi.org/10.1145/3152494.3156815

[5] Campus, kattankulathur, (2018). "Credit card fraud detection using machine learning models and collating machine learning models." international journal of pure and applied mathematics 118.20 (2018): 825-838.

[6] Patil, S., Nemade, V., & Soni, P.K., (2018) "Predictive Modelling For Credit Card Fraud Detection Using Data Analytics," Procedia Computer Science, 132, 385-395.

[7] Elizabeth A. DiGangi and Joseph T. Hefner, (2013) Chapter 5 - Ancestry Estimation. In Research Methods in Human Skeletal Biology, Elizabeth A. DiGangi and Megan K. Moore (Eds.). Academic Press, 117 – 149. https://doi.org/10.1016/B978- 0-12-385189-5.00005-4.

[8] Joe Hefner and Stephen Ousley, (2005) "Morphoscopic Traits and the Statistical Determination of Ancestry ".

[9] Banfield, Robert E et al, (2007) "A comparison of decision tree ensemble creation techniques," IEEE transactions on pattern analysis and machine intelligence vol. 29,1 (2007): 173-80. doi:10.1109/tpami.2007.250609.

[10] Nagesh Singh Chauhan, (2020) "Humility in AI: building trustworthy and ethical AI system," https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html#:~:text=Types\%20of\%20decision\%20trees\%20are,a\%20Categorical\%20variable\%20decision\%20tree

[11] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou, (2008) "Isolation Forest" In Proceedings of the 2008 Eighth IEEE International Conference on Data Mining (ICDM '08). IEEE Computer Society, USA, 413–422. https://doi.org/10.1109/ICDM.2008.17.

[12] Nari S Arunraj, Robert Hable, Michael Fernandes, Karl Leidl, and Michael Heigl, (2017) "Comparison of supervised, semi-supervised and unsupervised learning methods in network intrusion detection system (NIDS) application" Anwendungen und Konzepteder Wirtschaftsinformatik 6 (2017).

[13] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou, (2012) "Isolation-Based Anomaly Detection" ACM Trans. Knowl. Discov. Data 6, 1, Article 3 (March 2012), 39 pages.https://doi.org/10.1145/2133360.2133363

[14] John A Swets. (1988) "Measuring the accuracy of diagnostic systems" Science 240, 4857 (1988), 1285–12933.

[15] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer, (2002) "SMOTE: synthetic minority over-sampling technique," Journal of artificial intelligence research 16 (2002), 321–357.

[16] Richard O Duda, Peter E Hart, et al., (2006) "Pattern classification," John Wiley & Sons.

[17] Karwande, G., Mbakwe, A.B., Wu, J.T., Celi, L.A., Moradi, M., Lourentzou, I, (2022) "CheXRelNet: An Anatomy-Aware Model for Tracking Longitudinal Relationships Between Chest X-Rays," In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. MICCAI 2022. Lecture Notes in Computer Science, vol 13431. Springer, Cham.

AUTHORS

Amarachi Blessing Mbakwe:
I am a third-year Ph.D. student in the department of Computer Science at Virginia Tech and a member of the PLAN lab. My research interests lie at the intersection of artificial intelligence (AI), natural language processing (NLP), machine learning, and multimodal learning. My previous work introduced an anatomy-aware model for tracking longitudinal relationships between images.

Sikiru Ademola Adewale:
I am a third year Ph.D. student of computer science and applications at Virginia Tech. Most of my research work is in the area of machine learning & applications, artificial intelligence, and computer vision. As a member of the Perception + LANguage (PLAN) Lab, I am currently working on scene graphs.