International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Identity Theft Detection Using Machine Learning

Sri Nithya.Medapati [1]

Student[1], GMR Institute of Technology, Rajam, India

ABSTRACT

Identity theft is increasing as social and technological activities are increasing and the thief who are trying to get the data of a person are mainly focused on the personal information of the victim, because they targets the financial status of the victim. The victim doesn't even know that his identity is using for the unauthorized activity. As per current situations there are three main types of identity thefts are: identity cloning for concealment - in this type of theft the thief uses the personal information of the victim to hide from law enforcement or creditors synthetic identity theft - this type of theft is difficult to crack down because the thief uses the information that doesn't even exists Account take over identity theft - the thief uses the existing account of a victim to get the personal benefits . The major techniques for the thief to get the information of the victim are :mail theft using the mail to get the information shoulder surfing the thief uses the forms that are filled by the victim to get passwords and other information phishing the thief use the mails or messages by offering some offers or discounts to the victims there are many ways of stealing the information but the individuals must protect their data. Non-technical Protection methods Don't access personal accounts over unsecured wireless networks Protect your online data with multifactor authentication check for spyware or malware on your devices Don't give out personal information to unverified sources Regularly review bills and account statements for unusual activity Freeze your credit report if you find any suspicious activity. Algorithms that are used: K-means clustering algorithm, support vector machine.

Keywords: Identifying theft, Victim, Types and Techniques, Protect Measures.

INTRODUCTION

Identity theft, a pervasive crime, involves illicitly acquiring personal or financial data to exploit another's identity for fraudulent activities like unauthorized transactions. This crime leaves victims grappling with severe repercussions—damaged credit, financial instability, and tarnished reputation. Detecting identity theft employs numerous techniques, but machine learning algorithms stand out as the most reliable and precise models. Diverse algorithms such as Random Forest (RF), Logistic Regression (LR), K-Nearest Neighbour (KNN), and Decision Trees have emerged as formidable tools in detecting identity theft, consistently yielding accurate results. These algorithms effectively uncover fraudulent patterns, extending support to individuals and organizations to mitigate financial losses and safeguard sensitive information. Leveraging machine learning models significantly heightens accuracy, offering enhanced adaptability and reduced false positives, thereby augmenting their efficacy in predicting and identifying fraudulent patterns. Through these sophisticated models, a robust defence mechanism is established against the intricate and evolving landscape of identity theft, fortifying protection for individuals and institutions alike.

Random Forest (RF): RF is an ensemble learning method that operates by constructing multiple decision trees during training. Each tree in the forest gets a random subset of the data and makes its own individual prediction. When making a prediction, RF collects the predictions from each tree and averages them (for regression tasks) or takes a vote (for classification tasks). In identity theft detection, RF looks at various features (such as transaction history, user behaviour, etc.) and learns patterns that distinguish between legitimate and fraudulent activities.

Logistic Regression (LR): Despite its name, logistic regression is a classification algorithm. It works by analysing the relationship between the dependent variable (in this case, fraud or not-fraud) and the independent variables (features like user data, transaction details, etc.). LR estimates the probability of a certain event occurring based on given input data by using a logistic function. It's used in identity theft detection to predict the likelihood of fraudulent activity based on historical data and patterns.

K-Nearest Neighbour (KNN): KNN is a simple, instance-based learning algorithm. It classifies a new data point based on how its neighbour are classified. For identity theft, KNN looks at the similarity between new data (a transaction, user behaviour, etc.) and existing data points to determine if it aligns with known fraudulent patterns. It measures similarity using distance metrics.

Decision Trees: Decision trees partition data into smaller subsets based on various attributes and create a tree-like structure of decision nodes. Each node represents a feature and each branch a decision based on that feature. It continues until a prediction or outcome is reached. Decision trees in identity theft detection examine different aspects of transactions or behaviour to classify them as fraudulent or legitimate.

There are some outliers and each algorithm works differently on outliers for example Random Forest aggregates multiple decision trees to mitigate outlier impact. Logistic Regression models linear relationships but may be sensitive to outliers without regularization. K-Nearest Neighbor relies on neighbor similarity and can be influenced significantly by outliers, requiring normalization techniques. Decision Trees partition data, affected by extreme outliers but mitigated via pruning. Ensemble methods like Random Forest enhance robustness. Outliers impact each algorithm differently: RF is more resilient, LR needs regularization, KNN requires normalization, and Decision Trees might need pruning to counter outlier effects.

1. LITERATURE SURVEY

The above study explores various machine learning techniques for detecting credit card fraud, a prevalent and evolving financial threat. The research uses the European card benchmark dataset and applies machine learning techniques for initial fraud detection. The study uses various algorithms like random trees ,logistic regression and support vector machine to train the dataset and observe the patterns of fraud. The paper emphasises the importance of careful variation in model training. Overall, the proposed models outperform current fraud detection methods, offering improved accuracy, precision, and reduced false negatives, making them highly applicable to real-world scenarios.[1]

The study employs Latent Dirichlet Allocation (LDA) to analyze identity theft. Data collection and preprocessing steps are outlined, ensuring a clean dataset. The paper then explains the LDA algorithm's application in uncovering hidden themes in the text. The methodology section details how LDA was employed, including parameter choices and model evaluation. Results reveal key topics, such as prevention, detection, legal aspects, and technological advancements related to identity theft. The conclusion summarizes the findings, emphasizing the importance of comprehending diverse discussions on this issue. Future research directions, including trend analysis and sentiment evaluation, are suggested. Overall, this paper contributes valuable insights for cybersecurity experts, policymakers, and researchers interested in identity theft.[2]

The paper uses many machine learning algorithms like support vector machine , k nearest neighbour for credit card fraud. It employs a multistage deep learning model for enhanced accuracy and reduced false positives. In the first stage, a Convolutional Neural Network (CNN) captures spatial features from transaction data. The second stage employs a Recurrent Neural Network (RNN) to model temporal dependencies. A self-attention mechanism enhances feature weighting. The third stage deploys a deep autoencoder for unsupervised feature learning, aiding in the identification of subtle fraud patterns. The model's adaptability and scalability make it suitable for real-time fraud detection. Evaluation demonstrates superior performance compared to existing methods, with high detection rates and low false positives. The model holds significant promise for bolstering financial security and trust in payment transactions[3].

Healthcare fraud is a pervasive global issue that exploits inefficiencies in healthcare systems, depriving legitimate beneficiaries, particularly those covered by health insurance. This paper proposes a solution using machine learning and blockchain to combat healthcare fraud, specifically in claims processing. It employs a decision tree classification algorithm to categorize the original claims dataset. The knowledge extracted is then embedded in an Ethereum blockchain smart contract, enabling fraud detection and prevention. Comparative experiments reveal exceptional performance, with the best tool achieving a 97.96% classification accuracy and 98.09% sensitivity. This system significantly bolsters the blockchain smart contract's fraud detection accuracy, promising substantial improvements in safeguarding healthcare systems and ensuring the rightful provision of care to beneficiaries.[4].

The paper addresses the rising issue of financial fraud, particularly credit card fraud, in the context of advancing e-commerce and e-payment systems. The study uses Genetic Algorithm (GA) i.e used to optimize the features, enhancing the model's effectiveness. The paper then employs multiple machine learning classifiers, including Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), Artificial Neural Network (ANN), and Naive Bayes (NB), to detect credit card fraud.Using a dataset derived from European cardholders, the paper evaluates the performance of the proposed credit card fraud detection engine. The results demonstrate the system's superiority over existing approaches, indicating its potential to significantly improve fraud detection in the evolving landscape of financial transactions[5].

The paper uses the XGBoost algorithm for traning the data set on detecting the fraud.Experimental results demonstrate that XGBoost, achieves 0.91 precision score and 0.99 accuracy score when applied to imbalanced data. Various sampling techniques, including oversampling, undersampling, and SMOTE, are explored to enhance performance metrics like precision, recall, f1-score, and accuracy. Among these techniques, Random Oversampling emerges as the most suitable for addressing data imbalance, yielding a remarkable 0.99 precision and 0.99 accuracy when combined with XGBoost.Comparing the results across different classifiers underscores the superiority of XGBoost in handling imbalanced data scenarios for fraud classification. It defines the accurate results of XGBoost.[6].

The paper addresses the issue of online fraud, particularly in the area of credit card transactions. Common techniques explored in the study include Support Vector Machine, Gradient Boosting, Random Forest, and their combinations.In this comparative study, the authors examine the challenges associated with class imbalance in fraud detection and explore potential solutions. They note that algorithm effectiveness varies depending on the dataset and context. Despite exhaustive calculations, all algorithms exhibit certain imbalances at different stages of the study. The study recognizes the limitations of each approach, providing valuable insights for future research.Remarkably, while logistic regression displayed high accuracy, learning curves revealed potential underfitting issues. On the other hand, K-Nearest Neighbors (KNN) demonstrated robust learning capabilities, making it a superior classifier for credit card fraud detection. This research contributes to the ongoing efforts to combat financial fraud in the digital age.[7].

The paper addresses the persistent threat of cybercrimes, including cyberstalking, cyberbullying, hacking, data breaches, and identity theft, which plague the digital realm. Focusing on cyberstalking detection, the study explores various feature extraction techniques' impact on machine learning classifiers. Feature extraction methods include Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Random Forest (RF), Naive

Bayes (NB), and Decision Tree (DT).The paper evaluates the efficacy of each feature extraction technique in enhancing the detection model's performance. SVM records the accuracy of 95.2% with TF-IDF. Additionally, BERT and ELMo models demonstrate competitive accuracy rates of 90.9% and 90.5% for LR, and 90.7% and 90.2% for SVM, respectively.[8].

The study focuses on the increasing ranges of misrepresentation in online transactions due to the rising number of online customers. To control this issue, machine learning algorithms, including decision trees, naive Bayes, random forests, and neural networks, are being explored. The dataset used remains unchanged, and synthetic minority oversampling technique (SMOTE) is employed to address data imbalance.The results of the investigation indicate that the neural network model achieved the highest accuracy at 96%, followed by naive Bayes and random forest, both at 95%, and decision tree with 92%. These findings highlight the potential of machine learning algorithms, especially neural networks, in effectively addressing misrepresentation in online commerce.[9].

In summary, this article delves into the intricacies of Wangiri fraud patterns, detailing the implementation and assessment of ML algorithms in detecting this type of fraud. It emphasizes that the choice of the most suitable ML algorithm can vary depending on the specific Wangiri fraud patterns under consideration. The security analysis and experimental results underscore the potential of ML as a valuable tool in combating Wangiri fraud within the telecommunications industry.[10]

2. Methodology

The methodology used in the paper involves applying state-of-the-art machine learning and deep learning algorithms for credit card fraud detection. The performed a comparative analysis of both machine learning and deep learning algorithms to find efficient outcomes.

Initially machine learning algorithms such as Extreme Learning Method, Decision Tree, Random Forest, Support Vector Machine, Logistic Regression, and XG Boost are applied to the dataset to improve the accuracy of fraud detection. Later, three architectures based on a convolutional neural network (CNN) are applied to further enhance the fraud detection performance.

Machine Learning Models:

• **EXTREME LEARNING METHOD:** The extreme learning method (ELM) is a neural network for classification, clustering, regression and feature learning. It can be used with one or a multilayer of unseen notes. Given a single hidden layer of ELM, we assume that the output function of the junseen node is h(z) = G (p,q, z) wherever the parameters of the jth node are. The output function is as follows: fL (z) =summation of γihi (z) where summation till n, j=1. γi Is the weight of the output the ith hidden node?h (z.) = |Ghi (z), ... , hL (z)|

• **DECISION TREE**:As a result, the decision tree classifier is used to create the model, starting with the decision tree. We set the ‘max depth’ to ’4’ in the algorithm, which indicates that the tree can split four times, and the ‘criterion’ to ‘entropy,’ which is similarto ‘max depth’ but decides when to stop splitting the tree.We have thus finished installing and storing everything.

• **K-NEAREST NEIGHBOURS (KNN):**Supervised Learning is the learning that the amount or the result that we want or expect inside the training data (labelled data), and the amount in the data that we need to learn is known as the Target or the Dependent Variable. Next, forthe K-Nearest Neighbors (KNN), we build the model using the ‘K-Neighbours Classifier’ model and take the value of k,which represents the nearest neighbour, as ‘5’. The value of the ‘n-neighbours’ is arbitrarily selected, but it can be selectedpositively through iterating a range of values, surveyed byfitting and storing the predicted values into the ’knn-yhat’ variable

• **Random Forest:** Random Forest is an ensemble method that combines unpruned decision trees with feature randomness at each split. It aggregates predictions from individual trees to make the final prediction, leveraging the idea that no single algorithm is universally the most accurate. The algorithm seeks to enhance accuracy and robustness by incorporating diversity through random sampling of data and attributes in the decision tree construction.

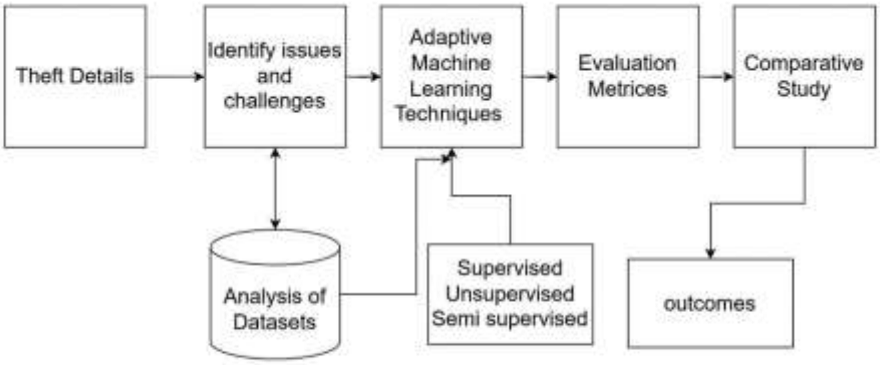
• **Support Vector Machine:** SVM (Support Vector Machine) is a widely used model for both binary and multi-class classification problems. It operates by finding a hyperplane that separates instances in a binary classification, with the equation w^T x + b = 0, where w is the coefficient weight vector and b is the bias term. The goal is to determine the values of w and b. In the linear case, these can be found using a Lagrangian function and support vectors. The decision function in SVM can be expressed as f(x) = sign(∑(i=1 to n) (λ _i * y_i * K(x_i, x)) + b), where K(x_i, x) represents the kernel trick. The polynomial kernel is represented by K(x, x_i) = ((x^T x_i) + 1)^d, and the Gaussian kernel is given by K(x, x_i) = exp(-||x - x_i||^2). The parameters C and γ are essential for SVM and need to be defined for the specific problem.

METHODS DEFINITION:

• It is one of the supervised algorithms that is utilized for the purpose of classifying the dataset into distinct categories. The value of a categorical or numerical variable, dependent on each other, may be predicted with the help of this classifier. Logistic Regression is the method that has the capacity to categorize fresh data by making use of both continuous and discrete datasets at the same time. This ability is what gives the algorithm its name. Due to the fact that it possesses this quality, it is considered to be one of the essential machine learning algorithms. This classifier's primary function is to provide predictions on the probability associated with a variety of scenarios.

• Gradient Boosting is used to create this classifier. In many contests, the model generated after using this classifier is a clear winner. Weights play an important part in the XGBoost algorithm. A decision tree is created. Weights are assigned to certain independent factors, which are then input into the decision tree, which predicts the outcomes.XG Boost provides a number of advantages, including the reduction of overfitting, which is why it is frequently referred to as a regularized boosting strategy. XGBoost also accommodates missed values in the data with ease, and it features a built-in cross-validation mechanism that executes at each step.

• This classifier can be utilized for classification-based difficulties as well as regression-based issues; nevertheless, for the most part, it is recommended for classification applications. The structure of this classifier resembles a tree, with the core nodes representing the attributes of the dataset, the branches representing the decision rules, and the leaf nodes of the tree representing the final outputs. Therefore, we can also say that this classifier gives a graphical method for finding all of the potential answers to a specific problem based on the conditions that have been provided.


Theft Details
Identify issues and challenges
Adaptive Machine Learning Techniques
Evaluation Metrices
Comparative Study
Analysis of Datasets
Supervised Unsupervised Semi supervised
outcomes

UML FLOWCHART

The research study focuses on detecting credit card frauds using machine learning and deep learning algorithms. It compares the performance of different algorithms and applies variations in the number of hidden layers and epochs to improve fraud detection accuracy. The evaluation of the research work shows improved results in terms of accuracy, f1-score, precision, and AUC curves. The proposed model outperforms state-of-the-art machine learning and deep learning algorithms for credit card detection problems. The study also highlights the use of feature selection algorithms to rank the top features from the credit card transaction dataset, which aids in class label predictions. It emphasizes the application of deep neural networks, specifically the CNN model, for identifying credit card fraud. The research paper focuses on supervised and unsupervised learning approaches and addresses the problem of class imbalance in machine learning. The research methodology involves performing steps such as feature selection, data balancing, and applying various supervised machine learning and deep learning models for fraud detection. The study utilizes a transaction information table of credit cards, which contains important features accessible for fraud detection modeling. The related work section discusses different research studies on credit card fraud detection, including approaches such as deep learning, machine learning, ensemble methods, and feature ranking. It also mentions the problem of class imbalance in classification tasks and the use of authentication methods for credit card authorization.

3. CONCLUSION

The study omits discussion on algorithm implementation challenges and specific country data for identity theft, while solely emphasizing autoencoders and CNNs for fraud detection without comparing alternate methods. It overlooks limitations in credit card fraud detection using genetic algorithms and prioritizes algorithm accuracy over dataset constraints like limited data availability, potential anonymization impact, and a short time frame of transaction records. These limitations underscore the models' dependence on quality data, affecting their broader applicability to diverse global fraud landscapes beyond the specific Brazilian insurance dataset. The absence of broader methodological discussions and dataset limitations hampers a comprehensive understanding of real-world implementation challenges and limits the study's generalizability to varied fraud contexts and regulations globally.

Accuracy: Accuracy is defined as out of all the predictions we made, how many were true. Accuracy is calculated as no of true values divided by no of false values or summation of true positive and true negative divided by summation of true positives, true negatives, false positives and false negatives. Accuracy=(true positives + true negatives)/(true positives+ true negatives+ false positives+ false negative)

Precision: Precision is defined as out of all the positive predictions we made, how many were true. Precision is calculated as true positive divided by true positive+ false positive.

Precision = true positives / true positives + false positives

Recall: Recall focuses on how good the model is at finding all the positives. Recall is also called true positive rate and answers the question out of all the data points that should be predicted as true, how many did we correctly predict as true.

Recall= true positives / true positives + false positives

F1 Score: F1 Score is a measure that combines recall and precision. As we have seen there is a trade-off between precision and recall, F1 can therefore be used to measure how effectively our models make that trade-off.

F1=2.((precision. recall)/(precision+ recall))

The study discussed about fraud detection using machine learning where it uses different algorithms to know the hidden patterns of the frauds in the data set if any patterns that are detected as outliers then it is a frauded transaction. The correlation between the features selected will play a key role in prediction the accuracy of the algorithm. And there are many algorithms to apply but we have used the best algorithms that has best accuracy ,precision, recall andF1score and other metrices for their best performance. As per future works we can apply some imbalanced techniques to balance data . And we can apply the models on different data sets to for its perfoemance. Other popular machine learning algorithms such as deep belief networks and restricted Boltzmann machines can also be applied in similar experiments on fraud detection.

References

1. Matsui, Takuro, and Masaaki Ikehara. "Low-light image enhancement using a simple network structure." IEEE Access (2023).
2. Zhang, Feng, et al. "Unsupervised low-light image enhancement via histogram equalization prior." arXiv preprint arXiv:2112.01766 (2021).
3. Liu, Fangjin, et al. "Dual UNet low-light image enhancement network based on attention mechanism." Multimedia Tools and Applications 82.16 (2023): 24707-24742.
4. J. Hai, Z. Xuan, R. Yang, Y. Hao, F. Zou, F. Lin, and S. Han, ''R2RNet: Low-light image enhancement via real-low to real-normal network,'' J. Vis. Commun. Image Represent., vol. 90, Feb. 2023, Art. no. 103712
5. Gasparyan, H., Hovhannisyan, S., Babayan, S., & Agaian, S. (2023). Iterative Retinex-Based Decomposition Framework for Low Light Visibility Restoration. IEEE Access.
6. Zhang, Zhijia, et al. "Continuous learning deraining network based on residual FFT convolution and contextual transformer module." IET Image Processing 17.3 (2023): 747-760.
7. H. Tang, H. Zhu, L. Fei, T. Wang, Y. Cao, and C. Xie, ''Low-illumination image enhancement based on deep learning techniques: A brief review,'' Photonics, vol. 10, no. 2, p. 198, Feb. 2023.
8. Su, Y., Wu, M., & Yan, Y. (2023). Image Enhancement and Brightness Equalization Algorithms in Low Illumination Environment based on Multiple Frame Sequences. IEEE Access.
9. Wang X, Chen L (2018) Contrast enhancement using feature-preserving bi-histogram equalization. Signal Image Video Process 12(4):685–692
10. Yu, W., Zhao, L., & Zhong, T. (2023). Unsupervised Low-Light Image Enhancement Based on Generative Adversarial Network. Entropy, 25(6), 932.
11. C. Wei, W. Wang, W. Yang, and J. Liu, ''Deep Retinex decomposition for low-light enhancement,'' 2018, arXiv:1808.04560.
12. Zhang, Q., Zou, C., Shao, M., & Liang, H. (2023). A Single-Stage Unsupervised Denoising Low-Illumination Enhancement Network Based on Swin-Transformer. IEEE Access.
13. Zhang, D., Huang, Y., Xie, X., & Guo, X. (2023). A variational Retinex model with structure-awareness regularization for single-image low-light enhancement. IEEE Access.
14. Liu, X., Zhang, C., Wang, Y., Ding, K., Han, T., Liu, H., ... & Ju, M. (2022). Low Light Image Enhancement Based on Multi-Scale Network Fusion. IEEE Access, 10, 127853-127862.
15. Garg, A., Pan, X. W., & Dung, L. R. (2022). LiCENt: Low-light image enhancement using the light channel of HSL. IEEE Access, 10, 33547-33560.