

# Seminar

# Report on

**“Comparative Analysis Online Fraud Detection using Machine Learning Techniques”**

# By

*Shubham Shinde*

*1032222127*

*Under the guidance of*

***Dr. Madhuri Rao***

**Dr. Vishwanath Karad MIT-World Peace University**

**Department of Computer Engineering &**

**Technology School of Computer Science & Engineering**

**\* 2023-2024 \***



**Dr. Vishwanath Karad MIT-World Peace University (MIT-WPU) School of Computer Science & Engineering**

**CERTIFICATE**

This is to certify that Shubham Shinde of B.Tech. AIDS

Semester-VI, PRN. No. 1032222127, has successfully completed seminar on

***Comparative Analysis Online Fraud Detection Machine Learning Techniques  
Comparative Analysis Online Fraud Detection using Machine Learning Techniques  
   
Comparative Analysis Online Fraud Detection using Machine Learning Techniques***

This seminar is satisfactorily submitted & delivered during the academic year 2023-2024 towards the partial fulfillment of degree of Bachelor of Technology under Dr. Vishwanath Karad MIT- World Peace University, Pune.

Dr. Balaji Patil

Seminar Guide Name & Sign Program Director

Department of CET, SoCSE

Date & Place:

**ABBREVATIONS**

ML: Machine LearningAI: Artificial IntelligenceSVM: Support Vector MachineLR: Logistic RegressionDT: Decision TreeRF: Random ForestGRU: Gated Recurrent UnitLSTM: Long Short-Term MemoryRNN: Recurrent Neural NetworkCNN: Convolutional Neural NetworkANN: Artificial Neural NetworkXG Boost: Extreme Gradient BoostingOOB: Out-of-Bag (used in Random Forest to evaluate the model)ROC AUC: Receiver Operating Characteristic - Area Under the CurveFP: False PositiveFN: False NegativeKNN: K-Nearest NeighborsCoNLL: Conference on Computational Natural Language LearningSMOTE: Synthetic Minority Over-sampling Technique

**Acknowledgement**

I would like to take this opportunity to express my sincere appreciation to all those who have supported me throughout the completion of my individual project on "Comparative Analysis Online Fraud Detection using Machine Learning Techniques."

First and foremost, I extend my deepest gratitude to my supervisor, Dr. Madhuri Rao, for their invaluable guidance, encouragement, and expertise throughout this project. Their insights and feedback have been instrumental in shaping the direction of my research and achieving its objectives.

I am grateful for the assistance and support provided by MIT World Peace University, which has allowed me to access necessary resources and facilities for conducting my research effectively.

Special thanks to the creators of the CoNLL-2003 dataset for providing such a valuable resource, which served as a foundation for my project.

Finally, I would like to acknowledge the support and understanding of my family, friends, and colleagues, whose encouragement has been a constant source of motivation and inspiration throughout this journey.

Shubham Shinde

1032222127

**Index**

|  |  |  |
| --- | --- | --- |
| **Chapter No.** | **Name of Topic** | **Page No.** |
|  | Introduction  1.1 Fundamental  1.2 Objectives  1.3 Scope  1.4 Outline |  |
|  | Literature Survey  2.1 LITERATURE REVIEW |  |
|  | Methodology   * 1. System Architecture   3.2 Pros and Cons |  |
|  | Future Enhancement |  |
|  | Application |  |
|  | Conclusion |  |
|  | Reference |  |

**Abstract**

Online banking fraud occurs whenever a criminal can seize accounts and transfer funds from an individual’s online bank account. Successfully preventing this requires the detection of as many fraudsters as possible, without producing too many false alarms. This is a challenge for machine learning owing to the extremely imbalanced data and complexity of fraud. In addition, classical machine learning methods must be extended, minimizing expected financial losses. Finally, fraud can only be combated systematically and economically if the risks and costs in payment channels are known. We define three models that overcome these challenges: machine learning-based fraud detection, economic optimization of machine learning results, and a risk model to predict the risk of fraud while considering countermeasures. The models were tested utilizing real data. Our machine learning model alone reduces the expected and unexpected losses in the three aggregated payment channels by 15% compared to a benchmark consisting of static if-then rules. Optimizing the machine-learning model further reduces the expected losses by 52%. These results hold with a low false positive rate of 0.4%. Thus, the risk framework of the three models is viable from a business and risk perspective.

The rapid growth in E-Commerce industry has lead to an exponential increase in the use of credit cards for online purchases and consequently they has been surge in the fraud related to it. In recent years, For banks has become very difficult for detecting the fraud in credit card system. Machine learning plays a vital role for detecting the credit card fraud in the transactions. For predicting these transactions banks make use of various machine learning methodologies, past data has been collected and new features are been used for enhancing the predictive power. The performance of fraud detecting in credit card transactions is greatly affected by the sampling approach on data-set, selection of variables and detection techniques used. This paper investigates the performance of logistic regression, decision tree and random forest for credit card fraud detection

Keywords: Fraud detection, Credit card, Logistic regression, Decision tree, Random forest., Integration of machine learning and statistical risk modelling, Economic optimization machine learning outputs

**CHAPTER 1**

**1.Introduction**

* 1. **Fundamental**

With the expansion of e-commerce and the widespread adoption of online payment methods, fraudulent activities have seen a noticeable surge. Credible reports indicate a stark and rapid increase in financial losses attributed to credit and debit card fraud between 2020 and 2022 . What’s particularly striking is that while unauthorized purchases and the use of counterfeit credit cards make up a relatively small portion, approximately 12-17%, of the total reported The associate editor coordinating the review of this manuscript and approving it for publication was Tyson Brooks . 137188 fraud cases, they account for a disproportionately large share, ranging from 75% to 80%, of the overall financial losses. In light of these critical issues, private businesses and government organizations have substantially increased their funding for research and development projects. Their primary objective is to create more resilient and effective systems for detecting fraudulent activities. Implementing automated fraud detection systems has become essential for financial institutions that oversee credit card issuance and online transaction management. These systems not only help reduce financial losses but also play a crucial role in enhancing customer faith and assurance. Innovative big

data and artificial intelligence possibilities have opened up, giving intriguing potentials, particularly in utilizing powerful machine learning algorithms to combat financial crime. Modern fraud detection systems, aided by cutting-edge data analysis and advanced machine/deep learning algorithms, have demonstrated extraordinary efficacy . Typically, these algorithms are trained on large datasets of labeled transactions, allowing them to differentiate between reg ular and fraudulent activity. The ultimate result is the development of binary classification models capable of distinguishing between valid and fraudulent transactions. Detecting fraudulent transactions using classification algo rithms is a difficult task that requires constant innovation and flexibility. In the same way, innovation assures security, data availability, dependability, and resilience against cyber warfare assaults in the fight against wireless communication interference, the financial industry must continually innovate to stay ahead in the struggle against financial crime. Firstly, there’s the issue of an imbalanced dataset, where the number of fraudulent transactions is significantly lower compared to legitimate ones. Secondly, there is the issue of cost sensitivity, where misclassifying fraudulent and normal transactions carries dissimilar costs, potentially having severe consequences. Additionally, transactions exhibit temporal dependence, necessitating consideration of their temporal relationships. Moreover, concept drift exists over time, which means that class conditional distributions can evolve, mandating periodic updates to the classifier. Lastly, managing the dimensionality of the feature space is a significant challenge, demandingsophisticatedpreprocessing techniques. A thorough examination of existing research found that when it comes to artificial neural networks, supervised learning techniques, logistic regression, and decision trees are the most commonly employed methods. Researchers have been drawn to the remarkable achievements of deep learning in various domains like computer vision, translation, speech recognition, and complex time series forecasting. As a result, several studies have started to utilize recurrent neural network variants, such as GRU, to develop credit card transaction fraud detection systems, taking inspiration from the success witnessed in those areas. LSTM and GRU represent recurrent neural networks (RNNs) aimed at mitigating the issues of gradient vanishing and exploding gradients in RNNs . These architectures are designed explicitly for capturing temporal patterns within sequential data. Deep learning models are attractive because of the valuable information from unprocessed data. In many research areas, neural networks and deep learning methods have consistently shown better results than conventional algorithms. However, it’s important to note that these models haven’t been widely used in fraud detection of credit card fraud. The key to building effective and accurate fraud detection systems is to transform the input data from a fraud

dataset into a simplified, lower-dimensional form. This lower-dimensional representation is achieved using representation learning techniques, yielding a more detailed and informative depiction of the data. Prominently featured in this study, Autoencoders have gained prominence as practical tools within the repertoire of representation learning techniques. Their allure lies in their capacity to unveil latent patterns within input data before subjecting them to classification. An Autoencoder consists of two key components: an encoder and a decoder. The decoder attempts to reconstruct the original inputs using the condensed representation that the encoder has created from the input data.

* 1. **Objectives**

The objectives of the information provided in the given article on "Online Payment Fraud Detection using Machine Learning in Python" can be summarized as follows:

1.Understand the Problem of Online Payment Fraud:

* Highlight the increasing risk of fraud in online payment systems due to the rise of digital transactions.
* Emphasize the need for effective fraud detection mechanisms to ensure secure online transactions.

2.Data Exploration and Visualization:

* Explore the dataset structure to understand the nature of the features and the target variable (isFraud).
* Utilize visualization techniques to identify patterns, trends, and relationships among the different features.

3.Preprocessing and Data Preparation:

* Demonstrate the process of data cleaning, feature encoding, and dropping irrelevant columns to prepare the dataset for modeling.
* Split the dataset into training and testing subsets for model evaluation.

4.Model Selection and Training:

* Present a variety of machine learning models suitable for fraud detection, such as Logistic Regression, XGBoost, SVC, and Random Forest.
* Train these models to predict the likelihood of fraud in the test data.

5.Model Evaluation and Comparison:

* Evaluate the performance of different models based on metrics like ROC AUC to identify the most effective model.
* Compare these models to determine the best approach for detecting online payment fraud.

Overall, the main objective is to provide a comprehensive guide for building a machine learning-based fraud detection system in Python, showcasing the necessary steps from data exploration and preprocessing to model training, evaluation, and visualization.

* 1. **Scope**

The scope of this report centers on the application of machine learning in detecting online payment fraud. While the potential use cases for machine learning in fraud detection are extensive, this study specifically explores its effectiveness in combating fraud in online payment systems. We delve into the use of various machine learning algorithms to identify fraudulent transactions and examine the process from data collection to model evaluation. This report provides a high-level overview of the topic, focusing on key concepts, methodologies, and challenges, without diving into exhaustive technical details or exploring every possible fraud detection scenario.

1. Investigation of Machine Learning in Fraud Detection

Objective: The aim is to explore how machine learning can be utilized to detect online payment fraud. This section addresses the fundamental principles of machine learning and its application in analyzing large datasets to identify patterns indicative of fraudulent activities.

2. Addressing Challenges in Online Payment Fraud

Issues Addressed: The study examines common challenges in online payment systems, such as unauthorized transactions, identity theft, and financial scams. We investigate how machine learning can mitigate these risks by providing automated and scalable detection mechanisms.

3. Analysis of Data Preparation and Preprocessing

Data Handling: The scope includes data collection, preprocessing, and feature engineering necessary for building effective machine learning models. This section explores data imbalances, encoding techniques, and data splitting methods.

Data Scope: The data used for this study is limited to a specific dataset containing online payment transactions. The scope does not cover other datasets or cross-industry comparisons.

4. Model Selection and Implementation

Models Analyzed: The study covers various machine learning models, such as Logistic Regression, XGBoost, SVC, and Random Forest. This section discusses the rationale behind model selection and the training process.

Scope of Analysis: The focus is on training and evaluating these models for fraud detection. The study does not delve into intricate technical details of machine learning algorithms or explore advanced techniques like deep learning.

5. Evaluation of Model Performance

Performance Metrics: The scope covers the evaluation of models based on metrics such as ROC AUC score and confusion matrices. This section analyzes model accuracy and the implications for fraud detection.

* 1. **Outline**

Chapter 1: Introduction

* What is Online Payment Fraud Detection?
* Why is it Important?
* Challenges of Developing Fraud Detection Systems
* Overview of the Seminar

Chapter 2: Literature Survey

* Review of Existing Fraud Detection Systems
* Identification of Gaps in Existing Systems
* Discussion on How Machine Learning Can Improve Fraud Detection

Chapter 3: Methodology

* System Overview
* Data Collection and Preprocessing
* Fraud Detection Models
* Model Training and Testing

Chapter 4: Performance and Results

* System Architecture
* Implementation Challenges
* Evaluation Results

Chapter 5: Application

* Experimental Setup
* Evaluation Metrics
* Results and Discussion

Chapter 6: Summary

* Summary of the Seminar
* Limitations of the Proposed System
* Directions for Future Work

**CHAPTER 2**

**LITERATURE SURVEY**

**Online Payment Fraud Detection Using Machine Learning by S. Lochan and H V Sumanth (26 January 2023):**

With the rise of web surfing and online shopping, so came the use of credit cards for online transactions, as did the prevalence of online financial fraud. Just in 2018, credit card theft cost the globe 24.26 billion USD. Many innocent individuals have lost a significant amount of money due to these scams, which have stopped them from ever engaging in online payment operations. Older folks, who are less skilled in technology, are the most frequent targets of these scams. Traditionally, To locate credit card fraud, automated learning techniques have been utilised through machine learning and deep learning approaches. However, automated learning approaches are labelled, which necessitates the acquisition of prior information. Our Hybrid technique outperformed the k-means clustering strategy in terms of accuracy, precision, and recall.

**Payment fraud detection using machine learning techniques by S.V.Juno , Bella Gracia, J.Godwin Ponsam, S Preetha( 18 February 2022):**

Credit card usage for online transactions has expanded tremendously as a result of the development and quick expansion of E-Commerce, resulting in an explosion in credit card fraud. There are many tools available in online to generate fake credit card numbers. Most fraudsters use an online credit card to generate fake cards to use for gaming platforms and E-Commerce platforms. We can use machine learning techniques like random forest, logistic regression for improve fraud detection in credit cards. The primary goal of this project is to improve current fraud detection processes by better predicting fake accounts.

**Fraud detection in Online Payment Transaction using Machine Learning Algorithms by Darshan Aladakatti, Gagana P, Ashwini Kodipalli( 22 May 2023):**

Online payment transaction is a transaction in which payment is made using digitalized currency. Customers all over the world prefer online payments to purchase almost everything from furniture to clothing, from food to medicines, from gadgets to appliances, and whatnot. the online transaction has now evolved into many platforms. It is one of the most efficient methods provided by many companies to its customer. It not only helps build the company's revenue but also impacts the growth of the company. Let us not forget that with pros come the cons with the greatest advantage of having tedious transactions done at the fingertips comes the fear of fraudulent activity in which our hard-earned money can get theft within seconds. These activities can be determined by machine learning algorithms by feeding adequate data about these transactions. We have used machine learning algorithms such as SVM (Support Vector Machine), LR (Logistic regression), Naive Bayes, Decision tree, and Random Forrest for the same. It is observed that Random Forest classifier has outperformed comparing to other classifiers with the accuracy of 99.94%

**Online Fraud Detection using Machine Learning by Diksha Dhiman, Amita Bisht, Anita Kumari, Dr Harishchander Anandaram, Shaurydeep Saxena, gana P,Ashwini Kodipalli,( 03 April 2023):**

Fraudsters find it easy to commit credit card fraud because it is an easy target. There has been an increase in online payment modes in due to e-commerce and other online platforms, there is now a higher danger of online fraud. Due to an increase in fraudulent online transactions, researchers have begun to evaluate and detect fraud using machine learning. In order to examine past transaction information and extract consumer behavioral patterns, our main goal in this study, a novel fraud detection algorithm for streaming transaction data is built and created. A system that clusters cardholders according to the amount of their transactions. In order to extract the behavioral pattern of the groups, we should aggregate the sliding window method transactions done by cards from various groupings. It is then decided which classifier with the best rating score can be chosen as one of the best methods to predict frauds after training different classifiers over the groups separately. The paper shows how the model is related to convolutional neural networks and afterward adding classifiers algorithms, for example, Isolation Forest, Local Outlier, and SVM can be utilized to recognize misrepresentation. As a result, concept drift can be solved via a feedback mechanism. We used the Kaggle credit card fraud dataset for this article.

**Online Transaction Fraud Detection Using Efficient Dimensionality Reduction and Machine Learning Techniques byRathan Kumar Chenoori, Radhika Kavuri ( 19 August 2022):**

In recent years, there has been a rapid increase in the number of online transactions. Substantial growth has been reported in e-commerce and e-governance in the past few years. Due to this the number of people using online payment methods has also increased. This has led to an exponential rise in the number of transactions that happen every day. This increase in online transactions has further led to an increase in the number of frauds in the transactions. There is an ever-growing need to detect these fraudulent transactions as early as possible so that appropriate actions could be taken and losses due to these frauds could be minimized. This work proposes machine learning models which could use the previously known data and try to predict frauds based on information learned through the old data. We propose a statistical based dimensionality reduction technique and various machine learning models were tried for classification purpose. We experimented our proposed method on IEEE-CIS Fraud Detection dataset and the best results were obtained on the XGBoost model which is demonstrated in this paper.

**CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING by Dr. Khaleel Ur Rahman Khan, A. Jahnavi, A. Nikitha, C. Ravishankar(4 December-2022 ):**

The detection of credit card theft is undoubtedly the most common problem in the modern day. This is a result of growing internet shopping and electronic commerce platforms. Since online payment methods save time and are more practical for transportation as digitalization grows more pervasive in today's world, more people are choosing them. The use of credit cards for e-commerce has skyrocketed, which has greatly boosted credit card theft. Fraudsters attempt to profit from the card and the transparency of online transactions. Credit card fraud typically happens when a card is misplaced or stolen, or when the card's details are known. Thus, it is essential to put a halt to fraudsters' actions. Today's population faces a wide range of credit card issues. use K-Nearest Neighbor (KNN) and A few approaches and machine learning algorithms can be utilized to address this issue including the can be utilized to address this issue include Knn classifier. Protecting electronic payments is the fundamental goal in order for people to use e-banking conveniently and safely

**Credit Card Fraud Detection Using Machine Learning by Deep Prajapati, Ankit Tripathi, Jeel Mehta, Kirtan Jhaveri (07 February 2022 ):**

From the day when payment systems emerged, there have been people willing to find novel ways to access someone’s finances illegally. This menacing hazards has grown in the current period, as the majority of transactions are now completed entirely online using credit card information. Frauds due to Credit Cards is a broad phrase that refers to any type of fraud involving a payment card, specifically a credit cards.The solitary purpose of such transgressions is usually to gain goods and services, or to make a huge payment to another account without the owner’s consent. According to the Nilson Report, By 2025, due to credit card fraud the United States has been projected to suffer losses up to 12.5 billion dollars. Using Machine learning algorithms to detect Credit card fraud is a process in which the data is investigated through various techniques to achieve the best possible outcomes in detecting and impeding fraudulent transactions. In order to evaluate different algorithms which accurately detect credit card fraud we have used techniques such as Random Forest, XGBoost, ANN (Artificial Neural Network). The results of these models can be used to effectively detect any credit card transaction happening whether a genuine one or fraudulent.

**2.1 LITERATURE REVIEW**

Machine learning has become an essential tool in the fight against online payment fraud. With the rapid growth of digital transactions, fraud detection has shifted from traditional rule-based systems to more sophisticated machine learning models capable of recognizing complex patterns. Online payment fraud encompasses various activities, such as credit card fraud, identity theft, and phishing, which necessitate a flexible and adaptable approach to detection.

The key advantage of using machine learning in fraud detection lies in its ability to learn from large datasets and continuously improve its predictions. Machine learning models, like Logistic Regression, Support Vector Machines (SVM), Random Forests, and Gradient Boosting (e.g., XGBoost), can analyze transaction data to identify anomalies that suggest fraudulent behavior. These models are particularly valuable because they can be retrained as fraud tactics evolve, ensuring that detection mechanisms remain effective.

However, several challenges persist in the application of machine learning for fraud detection. Data imbalance is a significant concern, as fraudulent transactions represent only a small fraction of all online payments. This imbalance can lead to models that are prone to false positives (flagging legitimate transactions as fraudulent) or false negatives (failing to detect actual fraud). Techniques like resampling and Synthetic Minority Over-sampling Technique (SMOTE) are often used to address this issue, but they introduce their own complexities.

Additionally, fraudsters are constantly adapting their methods, requiring fraud detection systems to evolve in response. Machine learning models need to be continuously updated with new data to stay ahead of these evolving tactics. Feature engineering, which involves creating new features from existing data, is a critical component in improving model accuracy and interpretability.

Another challenge is data privacy and security. The use of sensitive financial data in machine learning models raises concerns about data breaches and misuse. Developers must ensure robust data protection measures to maintain consumer trust and comply with regulations.

Despite these challenges, machine learning has demonstrated considerable success in fraud detection. Ensemble methods, which combine multiple models, have proven effective in reducing the risk of false positives and negatives. Moreover, continuous learning allows models to adapt to new trends in fraudulent activity, enhancing their effectiveness over time.

Future research in this field will likely focus on advanced machine learning techniques, such as deep learning, and real-time fraud detection, allowing for immediate response to suspicious activities. Cross-industry collaboration could also play a role in improving fraud detection capabilities, as financial institutions share insights and data to enhance their collective security.

Overall, while challenges remain, machine learning is poised to play a pivotal role in combating online payment fraud, offering a dynamic and effective approach to keeping digital transactions secure.

|  |  |
| --- | --- |
| **Online Transaction Fraud Detection Using Efficient Dimensionality Reduction and Machine Learning Techniques** | -Rapid increase in online transactions has led to a rise in frauds, necessitating early detection.-Machine learning models are proposed for fraud prediction using known data, with XGBoost showing best results.  -Efficient dimensionality reduction techniques and machine learning models like XGBoost are used for fraud detection . |
| **Online payment fraud: from anomaly detection to risk management** | -Machine learning-based fraud detection models significantly reduce expected and unexpected losses in payment channels by 15% compared to static if-then rules, with a low false positive rate of 0.4%  -Customer-related features like sociodemographic data, seniority of client relationship, and product usage were considered for risk assessment but not included due to data limitations and time constraints  -Validation results show that the machine learning model achieved a realistic detection rate of 18 true positives, which can be optimized to increase true positives by up to 45, with LOF performing best over the entire dataset |
| **An Analysis of the Most Used Machine Learning Algorithms for Online Fraud Detection** | -Various machine learning algorithms are used for online fraud detection, each with its own characteristics, advantages, and disadvantages  -Algorithms like SVM, DT, Naive Bayes, and KNN show high accuracy rates for fraud detection on a large volume of transactions  -The research reviews existing fraud detection techniques, applying a systematic quantitative literature review methodology to identify the most commonly used machine learning algorithms. |
| **Machine Learning For Credit Card Fraud Detection System** | - Machine learning techniques like logistic regression, decision tree, and random forest are used for credit card fraud detection.-Sensitivity, specificity, accuracy, and error rate are important performance metrics for fraud detection.  -Various methods of credit card fraud include application fraud, unauthorized use of lost/stolen cards, fake/doctored cards, skimming, site cloning, and false merchant sites . |
| **Online Payment Fraud Detection Model Using Machine Learning Techniques** | -Utilizes the ResNeXt-embedded Gated Recurrent Unit (GRU) model for real-time financial transaction data processing  -Addresses the rising threat of financial fraud by employing systematic artificial intelligence techniques, including SMOTE for data imbalance mitigation and feature extraction using an ensemble approach-Employs the RXT model fine-tuned with hyperparameters using the Jaya optimization algorithm, showcasing superior performance in financial transaction datasets  -Aims to enhance security, data availability, reliability, and stability against cyber warfare attacks in the financial industry. |

Top of Form

**Table 1: Key points from each research paper**

**CHAPTER 3**

**3. METHODOLOGY**

**3.1 System Architecture**

The proposed techniques are used in this is for detecting the frauds in credit card system. The comparison are made for different machine learning algorithms such as Logistic Regression, Decision Trees, Random Forest, to determine which algorithm gives suits best and can be adapted by credit card merchants for identifying fraud transactions. The Figure1 shows the architectural diagram for representing the overall system framework.

**Algorithm steps:**

Step 1: Read the dataset.

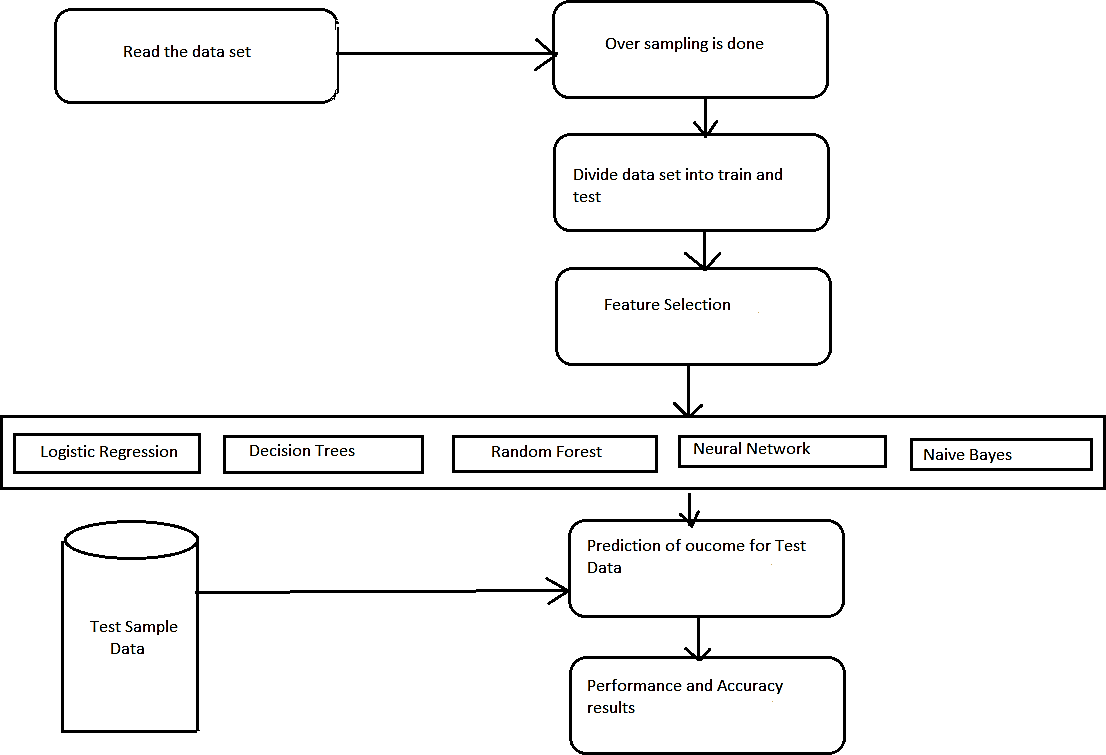
Step 2: Random Sampling is done on the data set to make it balanced.

Step 3: Divide the dataset into two parts i.e., Train dataset and Test dataset.

Step 4: Feature selection are applied for the proposed models.

Step 5: Accuracy and performance metrics has been calculated to know the efficiency for different algorithms. Step

6: Then retrieve the best algorithm based on efficiency for the given dataset.



**Figure1:** System Architecture

The image you sent depicts a simplified version of the architecture for online payment fraud detection using machine learning. Here's a breakdown of the steps involved:

1. **Read the Data Set:** Raw transaction data is uploaded into the system.
2. **Data Preprocessing (Over Sampling is Done):** This step might involve handling missing values, formatting data for the machine learning models, and balancing the dataset. Oversampling refers to a technique used to address imbalanced class distributions, where there are significantly more examples of one class (e.g., legitimate transactions) than the other (fraudulent transactions). In the context of fraud detection, this is important because fraudulent transactions are a minority but the ones we care most about detecting. By oversampling the minority class, we can ensure the model is trained on a more balanced dataset and improve its ability to detect fraud.
3. **Divide Data into Training and Test Sets:** The data is split into two sets: a training set used to train the machine learning models and a testing set used to evaluate the models' performance on unseen data.
4. **Feature Selection:** Relevant features are extracted from the transaction data. These features might include:
   * Transaction amount
   * Location (IP address)
   * Time of transaction
   * Device used (phone, computer, etc.)
   * Card details (billing address, card number)
   * Historical purchase behavior (past transactions)
5. **Model Training:** Various machine learning models are trained on the training data. The image shows these models:
   * Logistic Regression
   * Decision Trees
   * Random Forest
   * Neural Network
   * Naive Bayes

These are all common choices for fraud detection because they can learn complex patterns from data and identify anomalies that might indicate fraud.

1. **Prediction of Outcome for Test Data:** The trained models are used to predict the outcome (fraudulent or legitimate) for the transactions in the test set.
2. **Test Sample Data and Performance and Accuracy Results:** The performance of the models is evaluated on the test data set. This evaluation might involve metrics like accuracy, precision, recall, and F1 score.

* **Accuracy:** How often the model correctly predicts a transaction as fraudulent or legitimate.
* **Precision:** Out of the transactions the model flagged as fraudulent, how many were actually fraudulent.
* **Recall:** Out of all the actual fraudulent transactions, how many did the model identify correctly.
* **F1 Score:** A balance between precision and recall.

Based on the evaluation results, the best performing model is chosen for deployment.

1. **Deployment (not shown in the image):** The chosen model is integrated into the production environment to score new transactions in real-time and flag suspicious ones for further action (blocking, manual review).

This architecture provides a framework for online payment fraud detection using machine learning. By leveraging machine learning's ability to learn from data, businesses can significantly improve their ability to identify and prevent fraudulent transactions.

3.1 Logistic Regression: Logistic Regression is one of the classification algorithm, used to predict a binary values in a given set of independent variables (1 / 0, Yes / No, True / False). To represent binary / categorical values, dummy variables are used. For the purpose of special case in the logistic regression is a linear regression, when the resulting variable is categorical then the log of odds are used for dependent variable and also it predicts the probability of occurrence of an event by fitting data to a logistic function. Such as O = e^(I0 + I1\*x) / (1 + e^(I0 + I1\*x)) Where, O is the predicted output I0 is the bias or intercept term I1 is the coefficient for the single input value (x). (3.1) Each column in the input data has an associated I coefficient (a constant real value) that must be learned from the training data. y = e^(b0 + b1\*x) / (1 + e^(b0 + b1\*x)) (3.2) Logistic regression is started with the simple linear regression equation in which dependent variable can be enclosed in a link function i.e.,to start with logistic regression, I’ll first write the simple linear regression equation with dependent variable enclosed in a link function: A(O) = β0 + β(x) Where A() : link function O : outcome variable x : dependent variable A function is established using two things: (3.3) 1) Probability of Success(pr) and 2) Probability of Failure(1-pr). pr should meet following criteria: a) probability must always be positive (since p >= 0) b) probability must always be less than equals to 1 (since pr <= 1). By applying exponential in the first criteria and the value is always greater than equals to 1. pr = exp(βo + β(x)) = e^(βo + β(x)) (3.4) For the second criteria, same exponential is divided by adding 1 to it so that the value will be less than equals to 1 pr = e^(βo + β(x)) / e^(βo + β(x)) + 1 (3.5) Logistic function is used in the logistic regression in which cost function quantifies the error, as it models response is compared with the true value. X(θ)=−1/m\*(∑ yilog(hθ(xi))+(1−yi)log(1−hθ(xi))) Where hθ(xi) : logistic function (3.6) yi : outcome variable Gradient descent is a learning algorithm

3.2 Decision Tree Algorithm: Decision tree is a type of supervised learning algorithm (having a pre-defined target variable) that is mostly used in classification problems. It works for both categorical and continuous input and output variables. In this technique, we split the population or sample into two or more homogeneous sets (or sub-populations) based on most significant splitter / differentiator in input variables. TYPES OF DECISION TREE 1. Categorical Variable Decision Tree: Decision Tree which has categorical target variable then it called as categorical variable decision tree. 2. Continuous Variable Decision Tree: Decision Tree has continuous target variable then it is called as Continuous Variable Decision Tree TERMINOLOGY OF DECISION TREE: 1. Root Node: It represents entire population or sample and this further gets divided into two or more homogeneous sets. 2. Splitting: It is a process of dividing a node into two or more sub-nodes. 3. Decision Node: When a sub-node splits into further sub nodes, then it is called decision node. 4. Leaf/ Terminal Node: Nodes do not split is called Leaf or Terminal node. 5. Pruning: When we remove sub-nodes of a decision node, this process is called pruning. You can say opposite process of splitting. 6. Branch / Sub-Tree: A sub section of entire tree is called branch or sub-tree. 7. Parent and Child Node: A node, which is divided into sub-nodes is called parent node of sub-nodes where as sub-nodes are the child of parent node. WORKING OF DECISION TREE Decision trees use multiple algorithms to decide to split a node in two or more sub- nodes. The creation of sub-nodes increases the homogeneity of resultant sub-nodes. In other words, we can say that purity of the node increases with respect to the target variable. Decision tree splits the nodes on all available variables and then selects the split which results in most homogeneous sub-nodes. 1. Gini Index 2. Information Gain 3. Chi Square 4. Reduction of Variance

3.3 Random Forest: Random forest is a tree based algorithm which involves building several trees and combining with the output to improve generalization ability of the model. This method of combining trees is known as an ensemble method. Ensembling is nothing but a combination of weak learners (individual trees) to produce a strong learner. Random Forest can be used to solve regression and classification problems. In regression problems, the dependent variable is continuous. In classification problems, the dependent variable is categorical. WORKING OF RANDOM FOREST: Bagging Algorithm is used to create random samples. Data set D1 is given for n rows and m columns and new data set D2 is created for sampling n cases at random with replacement from the original data. From dataset D1,1/3rd of rows are left out and is known as Out of Bag samples. Then, new dataset D2 is trained to this models and Out of Bag samples is used to determine unbiased estimate of the error. Out of m columns, M << m columns are selected at each node in the data set. The M columns are selected at random. Usually, the default choice of M, is m/3 for regression tree and M is sqrt(m) for classification tree. Unlike a tree, no pruning takes place in random forest i.e, each tree is grown fully. In decision trees, pruning is a method to avoid over fitting. Pruning means selecting a sub tree that leads to the lowest test error rate. Cross validation is used to determine the test error rate of a sub tree. Several trees are grown and the final prediction is obtained by averaging or voting.

Step 1: Import the dataset

Step 2: Convert the data into data frames format Step3: Do random oversampling using ROSE package

Step4: Decide the amount of data for training data and testing data

Step5: Give 70% data for training and remaining data for testing.

Step6: Assign train dataset to the models

Step7: Choose the algorithm among 3 different algorithms and create the model

Step8: Make predictions for test dataset for each algorithm Step9: Calculate accuracy for each algorithm

Step10: Apply confusion matrix for each variable

Step11: Compare the algorithms for all the variables and find out the best algorithm.

**3.2 Pros and Cons of Online Payment Fraud Detection Using Machine Learning**

**Pros:**

* **Increased Accuracy:** Machine learning models can analyze vast amounts of data and identify complex patterns that humans might miss, leading to more accurate fraud detection.
* **Adaptability:** These models can continuously learn and adapt to new fraud tactics as fraudsters develop new schemes.
* **Scalability:** The system can efficiently handle large volumes of transactions without compromising performance.
* **Efficiency:** Automates the fraud detection process, freeing up human resources for other tasks.
* **Cost Savings:** Prevents fraudulent transactions, potentially saving businesses significant financial losses.
* **Continuous Improvement**: Machine learning models can be continuously updated and retrained with new data, allowing for ongoing improvement in fraud detection accuracy and responsiveness.
* **Reduced False Positives and Negatives**: Machine learning models, especially ensemble methods like Random Forests and XGB Classifier, are designed to reduce the risk of false positives (legitimate transactions flagged as fraudulent) and false negatives (fraudulent transactions not detected). This balance is crucial for maintaining customer trust and minimizing disruption.

**Cons:**

* **False Positives:** Models might flag legitimate transactions as fraudulent, causing inconvenience for customers.
* **False Negatives:** There's a possibility the model might miss some fraudulent transactions.
* **Data Dependence:** The model's performance relies on the quality and quantity of training data. Biases in the data can lead to biased models.
* **Complexity:** Implementing and maintaining machine learning models requires technical expertise and resources.
* **Explainability:** Complex models might be difficult to interpret, making it challenging to understand why a transaction is flagged.
* **Interpretability and Transparency**: Machine learning models, especially complex ones, can be challenging to interpret. This lack of transparency can pose problems when explaining decisions to stakeholders or addressing regulatory requirements.

**Overall, machine learning offers a powerful tool for online payment fraud detection. However, it's crucial to be aware of its limitations and implement it carefully to balance accuracy with customer experience.**

**CHAPTER 4**

**4. FUTURE ENHANCEMENT**

As the field of online payment fraud detection continues to evolve, there are several areas where future enhancements could lead to improved accuracy, efficiency, and reliability of fraud detection systems. These enhancements focus on addressing current limitations, adopting advanced technologies, and ensuring adaptability to changing fraud tactics. Here's a detailed look at the potential future enhancements for online payment fraud detection using machine learning.**1. Advanced Machine Learning and Deep Learning Techniques:**

Future fraud detection systems can leverage more advanced machine learning techniques, including deep learning, to improve pattern recognition and adaptability. Deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), are particularly effective at identifying complex patterns in large datasets. These models can offer greater accuracy and flexibility in detecting sophisticated fraud schemes.**2. Real-Time Fraud Detection:**Real-time fraud detection is crucial for minimizing financial losses and reducing the impact of fraud on consumers and businesses. Future enhancements could focus on improving the speed and efficiency of machine learning models, allowing them to process transactions in real-time. This requires optimizing the underlying algorithms and computational infrastructure to ensure rapid response times without compromising accuracy.3**. Enhanced Data Preprocessing and Feature Engineering**Effective data preprocessing and feature engineering are key to building robust machine learning models. Future enhancements could involve more sophisticated techniques for handling data imbalance, such as advanced resampling methods or synthetic data generation. Additionally, new features derived from transaction data, user behavior, or contextual information could provide additional insights for fraud detection models.4**. Integration of Additional Data Sources:**Integrating additional data sources into fraud detection systems can improve accuracy and reduce false positives. Future enhancements could involve incorporating data from external sources, such as credit bureaus, social media, or public records, to gain a more comprehensive view of potential fraud. This integration requires careful consideration of data privacy and compliance with regulations.5. **Cross-Industry Collaboration and Data Sharing:**Fraud detection systems could benefit from cross-industry collaboration and data sharing among financial institutions, payment processors, and regulatory agencies. Future enhancements could focus on creating secure frameworks for sharing information about fraud patterns and emerging threats. This collaboration can lead to better detection and prevention strategies while maintaining compliance with data protection laws.6. **Explainable AI and Model Transparency**:One of the challenges with advanced machine learning models is their lack of interpretability. Future enhancements could focus on developing explainable AI techniques that allow stakeholders to understand the decision-making process of fraud detection models. This transparency is essential for gaining trust from customers and regulators and for addressing compliance requirements.7. **Continuous Learning and Adaptability**:As fraud tactics evolve, fraud detection systems must remain adaptable. Future enhancements could involve implementing continuous learning mechanisms that allow models to update with new data and stay ahead of emerging fraud patterns. This adaptability requires robust data pipelines and ongoing monitoring to ensure consistent performance.8. **Advanced Security and Privacy Measures:**Given the sensitive nature of financial data used in fraud detection, future enhancements should prioritize security and privacy. This could involve advanced encryption techniques, secure data storage, and compliance with data protection regulations like GDPR and CCPA. Ensuring the security of machine learning models and data is crucial for maintaining consumer trust and preventing data breaches.These future enhancements outline the potential developments and improvements that could shape the next generation of online payment fraud detection systems using machine learning. By focusing on advanced techniques, real-time detection, cross-industry collaboration, and enhanced security, future fraud detection systems can provide even greater protection against evolving threats.

**CHAPTER 5**

**5.APPLICATIONS**

Applications of Machine Learning in Online Payment Fraud Detection

Machine learning has transformed the way online payment fraud is detected and prevented, offering a range of applications that enhance security, improve efficiency, and reduce financial losses. Below are some key applications of machine learning in the context of online payment fraud detection:

1. **Real-Time Transaction Monitoring:**

Machine learning models are applied to monitor online transactions in real-time, enabling the detection of fraudulent activities as they occur. By analyzing various features such as transaction amounts, frequency, location, and payment methods, these models can identify suspicious patterns and trigger alerts for further investigation.

2. **Fraud Detection Systems for Financial Institutions:**

Banks and financial institutions use machine learning-based fraud detection systems to secure their customers' accounts and transactions. These systems can detect unauthorized access, account takeover, or unusual spending patterns, helping to prevent financial losses and maintain customer trust. Machine learning models are especially valuable because they can adapt to new fraud techniques and provide scalable solutions for large volumes of transactions.

3. **Credit Card Fraud Prevention:**

Credit card companies employ machine learning algorithms to detect and prevent credit card fraud. These models analyze card usage patterns, detect anomalies, and identify potential fraudulent transactions. By using machine learning, credit card companies can reduce false positives and negatives, improving the accuracy of their fraud detection systems.

4**. Identity Theft Prevention:**

Machine learning is used to combat identity theft by analyzing user behavior and detecting inconsistencies that may indicate identity fraud. This application is particularly relevant for online payment systems, where users' identities must be verified before completing transactions. Machine learning models can identify unusual login patterns, multiple accounts with the same identity, or other suspicious behavior.

5. **Risk Scoring and Customer Profiling:**

Machine learning algorithms are applied to generate risk scores for transactions and customers. These risk scores help payment processors and financial institutions assess the likelihood of fraud and take appropriate action. Customer profiling involves creating profiles based on historical data and behavior patterns, allowing systems to identify deviations that could indicate fraudulent activity.

6. **Adaptive Fraud Detection for Evolving Threats:**

Fraud tactics are constantly evolving, and machine learning provides an adaptive solution for fraud detection. By continuously updating models with new data, organizations can stay ahead of emerging threats. This adaptability is crucial in maintaining the effectiveness of fraud detection systems in dynamic online environments.

7. **Compliance and Regulatory Applications:**

Machine learning applications in fraud detection can also support compliance with regulatory requirements. By providing detailed audit trails and explainable decision-making processes, machine learning-based fraud detection systems help organizations comply with financial regulations and data protection laws.

8**. Enhanced Customer Experience:**

By reducing false positives and negatives, machine learning applications in fraud detection contribute to an enhanced customer experience. Customers are less likely to experience disruptions due to erroneous fraud alerts, and legitimate transactions can proceed smoothly. This improved accuracy fosters customer trust and loyalty.

These applications demonstrate the broad range of uses for machine learning in online payment fraud detection. From real-time monitoring and credit card fraud prevention to identity theft detection and risk scoring, machine learning offers a powerful toolkit for securing online payment systems. As fraud tactics continue to evolve, machine learning applications will play an increasingly important role in ensuring the safety and integrity of digital transactions.

**CHAPTER 6**

**CONCLUSION**

Machine learning has emerged as a powerful tool in the ongoing battle against online payment fraud. As digital transactions continue to grow in volume and complexity, the need for advanced fraud detection systems has never been greater. By leveraging machine learning algorithms, financial institutions and payment processors can significantly enhance their ability to detect and prevent fraudulent activities in real-time.

One of the key advantages of machine learning in this context is its ability to identify complex patterns within large datasets. This capability is particularly valuable when combating online payment fraud, where fraudulent activities often follow subtle trends that are difficult to detect through traditional methods. Machine learning models such as Logistic Regression, Random Forests, Support Vector Machines, and XGBoost offer robust solutions that can adapt to evolving fraud tactics, providing a level of flexibility and scalability that is essential for modern fraud detection.

Despite its advantages, there are challenges to consider. Data imbalance, where fraudulent transactions represent a small fraction of total transactions, can skew machine learning models and impact accuracy. Additionally, as fraud tactics evolve, continuous learning and model updates are required to maintain effectiveness. Issues related to data privacy and security also need careful attention, as the use of sensitive financial data raises concerns about compliance with regulations.

To address these challenges, organizations can employ advanced techniques such as ensemble methods, feature engineering, and real-time fraud detection. By focusing on reducing false positives and negatives, machine learning models can improve the accuracy and reliability of fraud detection systems. Integration of additional data sources and cross-industry collaboration can further enhance the effectiveness of these systems by providing a broader context for detecting fraudulent activities.

Machine learning applications in fraud detection have a significant impact on customer experience. By reducing the rate of false positives, customers experience fewer disruptions, leading to increased trust and satisfaction. This improved customer experience is critical for maintaining loyalty in a competitive financial landscape.

Looking ahead, the future of online payment fraud detection will likely focus on even more advanced machine learning techniques, such as deep learning, to further enhance pattern recognition and adaptability. Real-time fraud detection will become increasingly important, allowing organizations to respond to suspicious activities as they happen. Additionally, ensuring model transparency and explainability will be crucial for regulatory compliance and gaining stakeholder trust.

In conclusion, machine learning provides a dynamic and effective framework for online payment fraud detection. It offers the flexibility to adapt to changing fraud tactics, the scalability to handle large volumes of transactions, and the accuracy to minimize false positives and negatives. As technology continues to advance, machine learning will play a pivotal role in ensuring the security and integrity of online payment systems, protecting both consumers and businesses from the ever-present threat of fraud.

**CHAPTER 6**

**Reference**

[1] W. Wei, J. Li, L. Cao, Y. Ou and J. Chen, “Effective detection of sophisticated online banking fraud on extremely imbal anced data”, World Wide Web, 2013

[2] C. Phua, V. Lee, K. Smith, R. Gayler, “A comprehensive survey of data mining based fraud detection research”, Artificial Intelligence Review, pp 1 – 14, 2005

[3] K.N. Karlsen and T. Killingberg, “Profile based intrusion detection for Internet banking systems”, Norwegian University of Science and Technology, 2008

[4] K. Navanshu and S. Y. Sait, “Credit Card Fraud Detection Using Machine Learning Models and Collating Machine Learning Models”, International Journal of Pure and Applied Mathematics, Volume 118, No. 20, pp. 825-838, 2018

[5] S. J. Russell and P. Norvig, “Artificial In telligence: A Modern Approach”, 3rd edi tion. Prentice Hall, 2010

[6] K. R. Seeja and Z. Masoumeh, “Fraud miner. A novel credit card fraud detection model based on frequent item set min ing”, The Scientific World Journal, Arti cle ID 252797, 2014

[7] A. Krenker, M. Volk, U. Sedlar, J. Bester and A. kosh, “Bidirectional artificial neu ral networks for mobile phone fraud de tection”, Journal of Artificial Neural Net works, Volume 31, No. 1, pp. 92 - 98, 2009

[8] M. Mukesh Kumar and R. Dash, “A com parative study of Chebyshev functional link artificial neural network, multi-layer perceptron and decision tree for credit card fraud detection”, Information Tech nology (ICIT), International Conference on IEEE, 2014

[9] K. A. U. Shan, N. Akhtar and M. N. Qureshi, “Real-time credit-card fraud de tection using artificial neural network tuner by simulated annealing algorithm”, Proceedings of International Conference on Recent Trends in Information, Tele communication and Computing, ITC, 2014

[10] Y. Sahin, B. Serol and D. Ekrem, “A cost-effective decision tree approach for fraud detection”, Expert systems with ap plications, Elsevier, Volume 40, No. 15, 2013