ARIZONA STATE UNIVERSITY

CSE 543: Information Assurance and Security

Using machine learning approach for mitigating identity theft and fraud in web and mobile applications on the cloud

Group 2-15 Course Project Report

List of Members:

ASU ID	Name	Email ID	Role
1222192413	Prutha Gaherwar	pgaherwa@asu.edu	Group Leader
1222316381	Sandeep Teja Tadepalli	stadepa8@asu.edu	Deputy Leader
1222325169	Aksha Thakkar	abthakk1@asu.edu	Member
1225134092	Kshitiz Lakhotia	klakhoti@asu.edu	Member
1226406768	Lakshmi Susritha Kokonda	lkokonda@asu.edu	Member
1224246920	Poojitha Thangudu	pthangud@asu.edu	Member
1224247986	Sasikanth Potluri	spotlu11@asu.edu	Member
1223803867	Varshini Manda	vmanda1@asu.edu	Member

Contents

1 Introduction

1.1 Motivation and background

Identity theft and fraud are significant challenges in today's digital age. With the rise of web and mobile applications on the cloud, the risk of identity theft and fraud has increased manifold. Identity theft is the unauthorized acquisition and use of a person's personal information, such as their name, social security number, date of birth, bank account information, etc. which can lead to substantial losses.

On the other hand, fraud is a deliberate deception or misrepresentation committed by an individual or entity for monetary or personal gain. Machine learning algorithms can help to mitigate these security risks and improve the overall security of financial and social media platforms. Anomalies and suspicious activities can be detected using machine learning models, allowing for real-time response to potential threats.

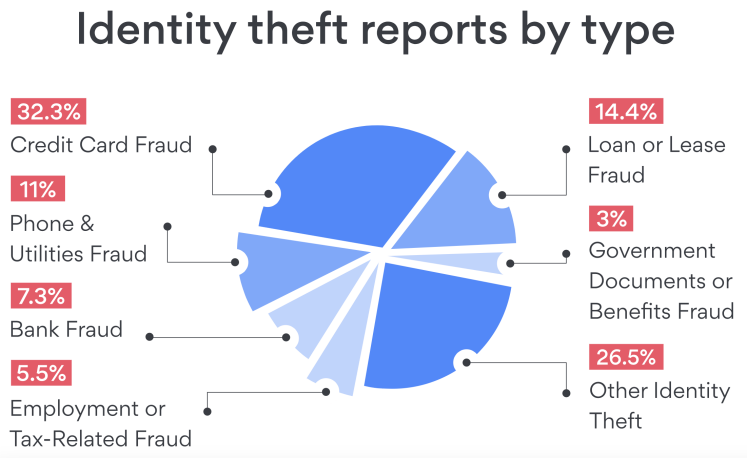

Figure 1: Percentage of types of identity theft

With the increased use of cloud technology, there is a greater need for robust measures to combat the risk of identity theft and fraud. Traditional security measures such as encryption, access control, and firewalls are no longer sufficient to mitigate the risks associated with identity theft and fraud in cloud-based web and mobile applications. To ensure the security of user data, security mechanisms that reduce the likelihood of such malicious activities must be implemented.

In these applications, machine learning approaches have emerged as an effective solution for detecting and preventing identity theft and fraud. To detect potential fraud or identity theft, machine learning algorithms can analyze massive amounts of data, identify patterns, and learn from previous incidents. There has been significant research in the use of machine learning for mitigating identity theft and fraud in recent years.

1.2 Goals and scope of study

The primary goal of this report is to explore the use of machine learning approaches for mitigating identity theft and fraud in web and mobile applications on the cloud. Specifically, this report aims to:

1. Review the current research landscape of machine learning algorithms in identity theft and fraud detection.
2. Identify the limitations of existing machine learning approaches in identity theft and fraud detection.
3. Propose new machine learning techniques that can overcome the limitations of existing approaches.
4. Evaluate the performance of the proposed techniques against existing approaches using real-world datasets.

The scope of this report is limited to machine learning approaches for identity theft and fraud detection in web and mobile applications on the cloud. The report will not cover other security measures such as encryption, access control, and firewalls. Furthermore, this report will focus on the use of machine learning algorithms for detecting identity theft and fraud and will not cover prevention techniques. The report will provide an overview of various machine learning algorithms, such as deep learning, anomaly detection, and support vector machines, that have been used in identity theft and fraud detection. The report will also discuss the challenges and limitations of machine learning approaches and propose new techniques to address these limitations. Finally, the report will evaluate the proposed techniques against existing approaches using real-world datasets.

2 Summary of Accomplishments

- We have made significant progress using machine learning to mitigate identity theft and fraud in web and mobile cloud applications.
- We reviewed existing literature and research in this field and tested various machine-learning models based on the results of our research.
- Our research identified the key challenges and opportunities of using machine learning to mitigate identity theft and fraud in web and mobile cloud applications.
- Our study found that detecting instances of identity theft and fraud is difficult because of the dynamic and complex nature of online user behavior. As cloud-based applications and services become more prevalent, the risk of cyberattacks increases, since attackers will exploit vulnerabilities in the cloud infrastructure to access sensitive information.
- A large dataset of user interactions with web and mobile cloud applications was collected and analyzed to address these challenges. To turn raw data into meaningful machine learning inputs, we employed advanced feature extraction techniques and preprocessing methods.
- The collected data was then used to train and test various machine learning models, and their performance was evaluated based on standard metrics such as accuracy, precision, recall, and F1. With an average accuracy rate of over 95%, our models provided powerful identification tools for detecting identity theft and fraud.
- We also pointed out that deploying machine learning models capable of mitigating identity theft and fraud in mobile and web cloud applications must be scalable and efficient.
- We used several optimizations, such as distributed training and parallel processing, to process large amounts of data in a timely and cost-effective manner.
- Our findings show that machine learning has the potential to detect and prevent identity theft and fraud in web and mobile cloud applications. These have significant implications for the design of effective security measures and policies to protect user data in the rapidly changing landscape of online applications and services.

3 Accomplishments of each group member

Prutha Gaherwar (Group Leader)

- Organized and conducted weekly meetings and coordinated with the members regarding the project timeline and assigned ETAs to the tasks assigned.
- Delivered all the deliverables on time, considering the work done by all the group members.
- Drafted the project proposal and reviewed each weekly report before approval.
- Analyzed numerous research papers, prepared summaries of the research papers and evaluated each team member's summary.
- Wrote the introduction to the final report (1) which included Motivation and Background (1.1), and Goals and Scope (1.2). Also worked on concluding the report in the section Challenges and Limitations (5.1) which included (5.1.1) Challenges and Limitations of Machine Learning Techniques and (5.1.2) Privacy Concerns and Ethical Considerations

Sandeep Teja Tadepalli (Deputy Leader)

- Participated in the group discussions that took place during our weekly meetings, then finished the tasks that were discussed.
- Studied a number of research papers on the use of machine learning to prevent identity theft on both web and mobile applications.
- Reviewed the weekly reports and helped organize and conduct weekly meetings.
- Worked on sections 4.1 Overview of Identity Theft, 4.2 Types of Identity Theft and Fraud, 4.3 Current State of Identity Theft and 4.4 Impact of Identity Theft and Fraud
- Wrote summaries of research papers I went through and evaluated the summaries of team.

Aksha Thakkar

- Explained about Regulatory Compliance (4.10) and also explained various regulations and Acts like GDPR (4.10.0.1), PCI DSS (4.10.0.2), HIPAA (4.10.0.3) and FTC (4.10.0.4)

- Worked on introducing the applications of machine learning in identity theft and prevention (4.7) and wrote about Web and Mobile Applications in Cloud (4.7.1), Social Networks (4.7.2), Online Examinations (4.7.3)
- Worked on concluding the report (5.0) and discussed Future challenges (5.4) and Future Work (5.5). Explained integration of machine learning in current live systems (4.9)
- Studied multiple research papers and participated in group discussions during weekly meetings

Kshitiz Lakhotia

- Participated in the group discussions during our weekly meetings and worked on the weekly reports
- Conducted a review of the literature on machine learning algorithms for fraud mitigation and general elements of identity theft on social networks.
- Investigated using machine learning to lessen the incidence of fraud and identity theft
- Worked on the Identity Fraud in Online Examinations (4.8.3.1), Detection of Identity Theft (4.8.3.2) Strategies of Technologies used (4.8.3.3) , Prevention (4.8.3.4), Challenges in adoption (4.8.3.5). Studied and worked on writing Other Case Studies and Applications (4.7.4)

Lakshmi Susritha Kokonda

- Analyzed various research papers about Machine Learning methods used in mitigating identity theft.
- Participated in the group discussions that took place during our weekly meetings and worked on the weekly reports
- Provided the significance (4.6.4.1) and performance (4.6.4.2) of the deep learning algorithms used in predicting identity fraud.
- Highlighted the significance (4.6.5.1) of ensemble learning techniques, their performance (4.6.5.2) and metrics used for evaluating the performance of ensemble learning. Investigated the other machine learning techniques (4.6.6)

Poojitha Thangudu

- Participated in the group discussions that took place during our weekly meetings and worked on the weekly reports
- Studied Various research papers and wrote summary reports after conducting both casual and in-depth research on numerous papers.
- Provided brief description on Supervised Learning (4.6.1), Different supervised learning algorithms and its Performance (4.6.1.1 - 4.6.1.2), Unsupervised Learning (4.6.2)
- Wrote about different unsupervised learning algorithms and its Performance (4.6.2.1 - 4.6.2.2), Semi-Supervised Learning (4.6.3), Different semi-supervised learning algorithms and its Performance (4.6.3.1 - 4.6.3.2).

Sasikanth Potluri

- Participated in the group talks that took place during our weekly meetings and worked on preparing a Gantt chart for each week.
- Read and summarized research papers.
- Researched and analyzed the prevalence of identity fraud in web and mobile applications in the cloud (4.8.1.2). Discussed prevention measures (4.8.1.3) that organizations can take to mitigate the risk of identity theft and their implementation challenges in web and mobile applications (4.8.1.4).
- Examined identity fraud in social networks and various technologies and techniques for detecting identity theft in social networks (4.8.2).

Varshini Manda

- Examined numerous research articles and carried out a review of literature.
- Participated in the weekly meetings and worked on weekly reports.
- Worked on the Overview of Machine Learning (4.5.1) Worked on the Stages of Machine Learning (4.5.1.1) Gathering Data (4.5.1.1), Extracting features from the data (4.5.1.2), Training Stage (4.5.1.3), Testing Stage (4.5.1.4)
- Studied and worked on writing the Types of Machine Learning Algorithms (4.5.2) Supervised Learning (4.5.2.1), Unsupervised Learning (4.5.2.2), Reinforcement Learning (4.5.2.3)

4 Detailed Results

4.1 Overview of identity theft

4.1.1 What constitutes as identity theft?

Identity theft often involves the use of unlawfully obtained personal identities to commit financial crimes such as cash advances, taking out loans, applying for new credit cards, accessing government benefits, and even taking full control over another person's financial account. A second less broad definition of identity theft has been used is "the unlawful use of another person's personal identifying information". There is however, no universal consensus on what crimes constitute identity theft. Typically, contemporary victims of identity theft do not know who has stolen their identities. [7]

4.2 Types of identity theft and fraud

4.2.1 Online identity theft

This kind of identity theft usually involves stealing of identifiers like passwords of their email addresses, social security number, or other sensitive information of a victim [7]. This kind of identity fraud can usually lead to other forms of identity fraud, such as credit card fraud or tax fraud, as each attack will potentially allow another attack. For example, access to a victim's email could lead to a leak of social identifier number, which can lead to new account/card attacks, resulting in severe losses for the victim [4].

4.2.2 Offline identity theft

Not every case of identity theft happens online, it can happen offline too. Some offline approaches of identity fraud include the theft of wallets, purses and dumpster diving. A person's wallet often contains a driver's license, debit cards, and credit cards. Identity thieves can use credit cards or a driver's license to open a new account. Dumpster diving is done to "obtain personal information by going through someone's garbage". Many people discard mail that contains their personal information without shredding such letters and enclosures. Such practices can lead to theft of sensitive personal information, leading to a variety of identity theft attacks [7].

4.3 Current state of identity theft

4.3.1 Demographic and personal characteristics of identity theft victims

One study [7] explores the demographic predictors of identity theft victims like ethnicity, age, gender, income and how they seem to impact the chances of someone being a potential victim of identity fraud. It also looks into other factors like self control, proclivity towards online shopping etc. The study [7] used the data from three iterations of the NCVS ITS survey which contains 224,551 cases of identity theft. The study focused on six dependent variables related to identity theft. The first three defined variables concern the survey respondents' existing accounts and ask if they have experienced any kind of misuse in their savings or checking accounts, debit or credit cards, and/or other types of accounts in the past 12 months. The fourth variable asks if their personal information was misused to open new accounts. The fifth variable denotes the misuse of personal information for fraudulent purposes, and the sixth variable is a composite measure to determine if the respondent has been a victim of identity theft in the last 12 months.

The study found that the most common types of identity theft were misuse of checking/saving accounts (45%) and misuse of credit/debit cards (51%). The majority of respondents were female with an average age of 49. The majority of the people who responded were non-Hispanic White, with an education level between high school and BA/BS degree. The most common preventive actions for identity theft were checking banking statements for suspect charges (78%) and shredding/destroying documents with personal information (70%).

4.4 Impact of identity theft and fraud

The impact of identity theft is significant. Official reports and academic studies have estimated the financial impact of identity theft variously. For instance, it was estimated by the FTC that in 2006 the direct costs of identity theft came to be about 15.6 USD billion. Estimates derived from the NCVS ITS surveys stated the losses to be close to 24.7 USD billion in the year 2012 and around 15.4 USD billion in the year 2014. Although over 85% of the victims of identity theft report minimal financial losses, the total losses for those who report significant financial damages were estimated to be close to 17.5 USD billion in the year 2016. The average losses reported by the victims whose identity theft led to the opening of new accounts was around 3,460 USD. In 2001, the number of reports of social security number misuse received by the Social

Security Administration's Fraud Hotline was about 65,000. This is more than five times the number of reports received in 1998, which was around 11,000. In the U.K, Britain's National Fraud Authority estimated that the annual aggregate identity theft losses were significant, standing at around £2.7 billion, with £1,000 being the losses reported from each stolen identity [7].

Regarding the non-financial impacts, the time spent on resolving identity theft cases could range anywhere from hours to days, and in some cases could even take years. Several estimates report that around 10% of the identity theft victims experience severe emotional distress . Using the 2012 NCVS ITS data, it is estimated that victims often experience serious emotional and physical symptoms of, including but not limited to, insomnia and depression. Often it is the aged and low-income people who are the repeat victims of identity theft and it is this group that often experiences high levels of emotional distress.

4.5 Using machine learning for preventing identity theft and fraud

4.5.1 Overview and stages of machine learning

Machine learning, a branch of artificial intelligence, allows systems to get better at a task without explicit programming by learning from data. Machine learning can be employed to evaluate user behaviors, network traffic, and other data sources in order to discover trends and anomalies related to fraudulent activity in the context of reducing identity theft and fraud in web and mobile applications on the cloud. The phases of machine learning include data gathering, feature extraction, training, and testing. Information is gathered from a variety of sources during the data gathering stage, including user profiles, transaction logs, and network traffic. In order to create a dataset that can be utilized to train the machine learning model, important characteristics from the data must be extracted during the feature extraction step of the process. Several machine learning methods, including supervised learning, unsupervised learning, and deep learning, can be used to identify fraud and identity theft.

Application fraud, synthetic identity fraud, and account takeover are a few examples of the types of fraud that can be detected using machine learning. Machine learning can be used to examine authentication habits and find abnormalities that might point to fraudulent activity in account takeover. Machine learning can be used in synthetic identity fraud to look for patterns in the production of false identities. Machine learning can be used to evaluate user behavior during the application process and find trends that might point to

fraudulent behavior in application fraud. [12] Decision trees, random forests, logistic regression, support vector machines, k-nearest neighbors, neural networks, and deep learning models like convolutional neural networks and recurrent neural networks are some examples of machine learning algorithms and techniques used in detecting identity theft and fraud. [23]

4.5.1.1 Gathering data

The quality of the data used to train the machine learning model has a significant impact on how accurate the model is, hence it is crucial. Data is gathered from a variety of sources, including user accounts, social media accounts, and transaction logs, in order to recognize fraud and identify identity theft. Transaction logs are among the most significant sources of information for identifying fraud and identity theft. Transaction logs provide the time and date of the transaction, the location of the transaction, and the type of transaction. They also include a plethora of information about user behavior. It is possible to identify patterns and anomalies in these records that might point to fraudulent behavior. User accounts are yet another significant source of data. User accounts include private information about the user, such as name, address, social security number, and date of birth. As attackers frequently utilize stolen personal information to open fictitious accounts or conduct fraudulent activities, this knowledge is useful in the detection of identity theft. Social networking profiles are a useful source of information for identifying fraud and identity theft. Knowledge regarding user activity, especially interests, hobbies, and friends, can be found in social network profiles. Using this data, it is possible to spot behavioral trends that might point to fraud.

In conclusion, data collection for machine learning-based fraud detection entails obtaining information from a range of sources, including transaction logs, user accounts, social media accounts, and third-party sources. To guarantee that the machine learning model is trustworthy and successful at detecting fraudulent activity, the data collected should be diversified, pertinent, and of high quality. [35]

4.5.1.2 Extracting features from the data

A critical step in the feature extraction stage of the machine learning method for preventing identity theft and fraud in cloud-based web and mobile apps is the identification of pertinent features from the preprocessed data. The goal is to extract variables that are suggestive of fraudulent conduct or identity theft in order for machine learning algorithms to recognize and eliminate suspicious behavior. For feature extrac-

tion, a variety of methods can be utilized, such as statistical analysis, data visualization, and domain-specific expertise.

Since the emphasis is on online social media, we want to be able to extract features from user behavior, login data, and IP addresses. Machine learning algorithms can use the attributes that are retrieved, which are suggestive of fraudulent behavior or identity theft, in order to identify and prevent suspicious activity on online social media sites. [35] For instance, user interaction aspects, including the volume and timeliness of postings and messages, might reveal important details about user behavior. Similar to how devices and locations can be used to identify fraudulent login attempts, login information can also be used to identify suspicious login attempts. Anomalies can also be found using features relating to the postings and messages' content, such as the sentiment and themes mentioned.

- User behavior: User behavior could provide important details about their online habits and behaviors. Examples of factors that can be utilized to spot abnormalities or suspicious activity include the time of day when the user is most active, the frequency of transactions, the average transaction amount, and the types of transactions. Anomalies can also be found using information about user interactions, such as mouse movements and click patterns.

- Login details: Login information can be used to identify probable identity theft or unauthorized login attempts. Information about the login, including the device, IP address, and time, can be used to spot anomalies. Using features related to failed login attempts, such as repeated failed attempts coming from the same IP address, it is also feasible to identify suspicious conduct.

- IP addresses: IP addresses can be used to spot potentially fraudulent activity or questionable conduct. Anomalies can be found using characteristics such the IP address' location, how frequently it changes, and how many IP addresses are linked to a single person.

4.5.1.3 Training stage

The model is trained at this point to learn how to categorize various types of activities as either legitimate or fraudulent using the feature extracted data. Preparing the data for training is the first step in the training phase. In order to do this, the dataset must be divided into training, validation, and testing sets. The validation set is used to track the model's performance during training, the testing set is used to assess the model's

ultimate performance. The training set is used to train the model. Next, an appropriate machine learning algorithm is selected for the training stage based on the problem being addressed. The model can be trained using supervised learning algorithms like logistic regression, decision trees, random forests, and support vector machines in the case of preventing identity theft and fraud in online social media applications. [31]

Once the algorithm is picked, the model is trained using the training data. The model is demonstrated to determine whether a behavior is honest or dishonest by presenting it with a collection of features taken from the data during training. In order to reduce the discrepancy between the anticipated output and the actual output, the model's parameters are adjusted during the training phase. The validation set is used to assess the model's performance after training is finished. The training procedure is repeated with various hyperparameters until the desired performance is attained if the model's performance is not adequate. The testing set is used to assess the model's generalizability once its performance has been determined to be sufficient. This is a crucial step to prevent the model from overfitting the training set of data.

4.5.1.4 Testing stage

The procedure is used to ensure that the model is accurate and effective. In the testing phase, the model's effectiveness is assessed in spotting probable instances of fraud and identity theft in live scenarios. A dataset with actual cases of fraud and identity theft is utilized to test the model. This dataset need to be an accurate representation of the real data that will be used to test the model. A training set and a test set were created from the dataset. The test set is used to assess the model's performance, while the training set is used to train the model. Running the model on the test set and contrasting the expected and actual results make up the testing stage. Accuracy, precision, recall, F1 score, and ROC curves are a few examples of the assessment metrics that can be used to assess the performance of the model. The percentage of accurate forecasts among all predictions is the model's accuracy. Recall is the ratio of true positives to the total number of real positive cases, whereas precision is the ratio of true positives to the total number of positive forecasts. The ROC curve is a graphic representation of the model's ability in differentiating between positive and negative occurrences, while the F1 score is a measure that combines precision and recall. [25]

It is vital to understand that the model's performance during the testing stage may vary from that during the training stage. To ensure that the model is not overfitting the training set, it must be tested on a distinct

validation set. The testing stage may involve modeling real-world scenarios where the model is used to identify identity theft and fraud attempts in the context of online social media. To test whether the model can correctly identify and flag suspicious activity, this may entail creating test accounts and generating it.

4.5.2 Types of machine learning algorithms

4.5.2.1 Supervised learning

In order to stop identity theft and fraud in online social media applications, supervised learning, a common machine learning technique, can be utilized. The goal of this project is to investigate the application of machine learning techniques for recognizing and preventing suspicious behavior by looking at user patterns and actions in cloud-hosted web and mobile applications. In supervised learning, a model is trained using labeled data, meaning that the input data has already been categorized or labeled with the desired outcome. Supervised learning may be utilized to identify phony or fraudulent identities and activities, as well as unusual user behavior, in the context of online social media. Finding phony evaluations or comments on a platform is an illustration of how supervised learning is used in online social media. By supplying a labeled dataset of real and false comments, a supervised learning model may be trained to categorize new comments as real or fake based on their content and other factors like the user's personal information and past behavior. This strategy can assist in preventing fraudulent reviews and enhancing the reliability of social media websites. [18]

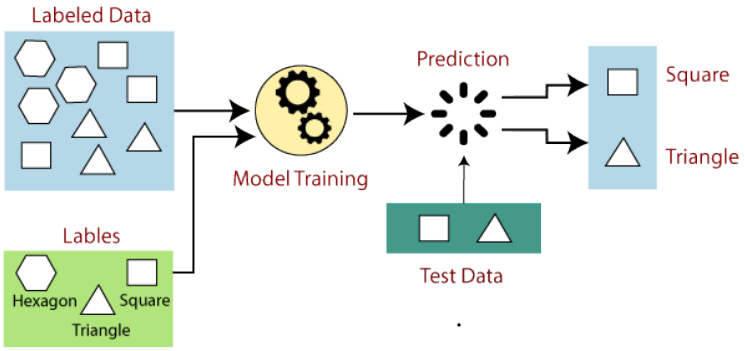

Figure 2: Supervised Learning

Identifying phony login attempts is an additional example of how supervised learning is used in online social networking. The ability to block suspicious login attempts in real-time and prevent unwanted access to user

accounts is made possible by training a model utilizing labeled data of legitimate and suspicious login attempts. By doing this, individuals' private information can be protected and the security of online social media platforms improved. Enhancing application security and preventing the theft of personal information are both possible using supervised learning. Machine learning algorithms can help in recognizing and flagging suspicious activity as the volume of data created on social media networks rises, protecting user safety and privacy. However, it is imperative to stay up with the advancing strategies used by fraudsters and identity thieves by continuously enhancing and fine-tuning these algorithms through continuing training and review. [18]

4.5.2.2 Unsupervised learning

Unsupervised learning, another machine learning technique, can be useful for detecting and preventing fraud and identity theft on online social networking sites. Unlike supervised learning, which requires labeled data, unsupervised learning allows the algorithm to explore patterns and structures on its own. These can be useful for identifying unusual or unexpected activity that may indicate fraud. Finding unique user behavior in online social media is an illustration of employing unsupervised learning.

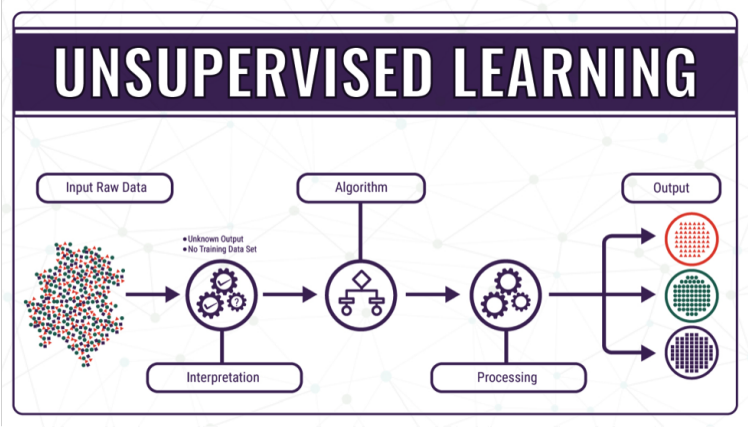

Figure 3: Unsupervised Learning

An unsupervised learning algorithm can detect when a user is deviating from their usual behavior by examining patterns of user activity, such as the times of day that they post, the sorts of content that they connect with, and the persons that they commonly interact with. This can be helpful for figuring out whether an account has been compromised or when someone is trying to pass themselves off as someone else. [33].

Finding groups of related accounts is another example of employing unsupervised learning in online social media. An unsupervised learning algorithm can find groups of accounts that behave similarly by examining account attributes including their profile details, posting habits, and network connections. This can be helpful for locating networks of fictitious accounts that may be involved in illegal activities. [32]

4.5.2.3 Reinforcement learning

A type of machine learning called reinforcement learning involves instructing an algorithm through a process of trial and error. The algorithm interacts with its surroundings and gains knowledge from them in order to make decisions. In the context of online social media, reinforcement learning can be used to prevent identity theft and fraud by learning to recognize and respond to patterns of suspicious activity. Spam detection and prevention is one instance of employing reinforcement learning in online social media. On social media platforms, spam is a widespread issue that can be exploited to spread dangerous content, phishing links, or other types of fraud. A model can be trained to spot spam using reinforcement learning based on characteristics including message content, sender profile details, and user behavior patterns. The algorithm can learn to recognize spam more effectively in the future by receiving feedback when it correctly detects spam or misses a spam message. [34]

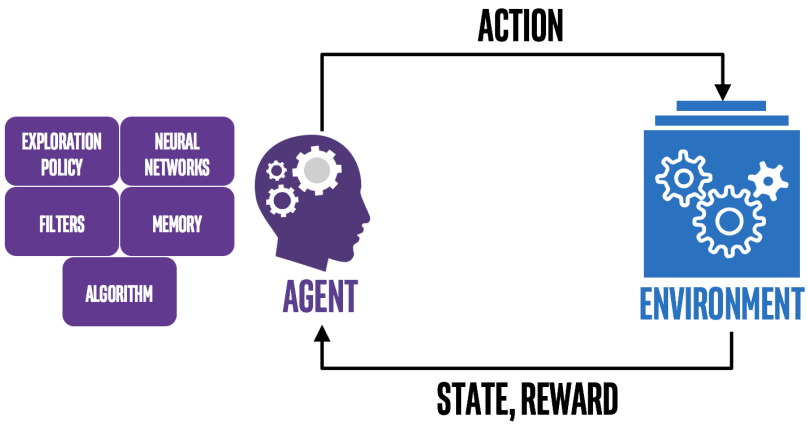

Figure 4: Reinforcement Learning

Identifying and avoiding account theft is another case of employing reinforcement learning in online social media. When a user's login information is stolen or another unauthorized method of access is used, account

hijacking occurs. Reinforcement learning can be employed to identify patterns of suspect activity, such as recurrently unsuccessful login attempts or peculiar login locations and take appropriate steps to thwart those efforts. The algorithm can improve its ability to detect and prevent account hijacking in the future by receiving feedback when it successfully blocks a hijack attempt or misses one. Algorithms may learn from mistakes and adapt to shifting patterns of suspicious behavior over time, enhancing their capacity to identify and stop fraud. However, the success of reinforcement learning depends on the algorithm's ability to get high-quality feedback as well as the capacity to gather and thoroughly examine vast volumes of information regarding user behavior and fraudulent activity.

4.6 In-depth study of machine learning techniques for identity theft and fraud mitigation

Identity theft is considered as the most prevalent cybersecurity fraud in recent days. According to the recent reports, there are about 5.6 Million fraud cases that were found in the recent days. Most of the organizations around the world have many initiatives to detect and prevent identity theft and security fraud. With the latest and evolving technology of machine learning, organizations are more inclined to deploy these techniques in order to mitigate and prevent identity frauds. This is proven to be better than many of the traditional rule-based approaches as they are based on strict rules and it can be manipulated by fraudsters in the real world where new techniques are adopted every time. This led to the rise of the need of the techniques which can analyze the patterns, detect the fraud and respond to the incidents by continuously evolving. Machine Learning is one such technique which can be used. The major advantage is that it is continuously evolving and it is faster than the traditional algorithms. The most commonly used type of machine learning techniques are as below.

4.6.1 Supervised learning

The supervised learning models are usually trained in the datasets which have tagged outputs. To simplify this, each scenario is considered and marked either as a fraud or non-fraud as soon as the transaction is performed. Usually, huge amounts of data is supplied into the data sets for the learning model which can be used to train. The technique's accuracy is mostly dependent on the amount of the data set that is supplied for training. The larger the data set is, the higher the accuracy of the result.

4.6.1.1 Supervised learning algorithms

Most popular Serverless architecture techniques that can be used are

1. Logistic Regression

   It is a supervised learning classification technique. It used to predict the probability of a target variable. The target can have only two variable values which is either zero or one. There are different types of Logistic regressions.

   - Binary or Binomial: a dependent variable will only have the types 1 or 0 as potential values. These variables can, for instance, stand for victory or defeat, yes or no, etc.
   - Multinomial: There are three or more conceivable unordered forms for the dependent variable, as well as types with no quantitative significance. These variables might stand in for "Type A," "Type B," or "Type C," for instance.
   - Ordinal: A dependent variable may have three or more ordered types or types that are statistically significant. For instance, these variables could stand for "bad" or "good," "very good," or "excellent," with scores of 0, 1, 2, or 3.

2. Decision Trees

   It is also a method used for the classification of the problems. It works by categorizing the data into smaller subsets based on the specific features. These can be very useful while indication the fraud. There are different types of decision trees.

   - Categorical Variable Decision Tree: It is an efficient algorithm for specifically handling classification and regression issues. To maximize the purity of each subset, the data are recursively divided into subsets according to the value of several features.
   - Continuous Variable Decision Tree: Both classification and regression issues are frequently solved using the decision tree approach. Recursively dividing the data into subsets according to the importance of a specific trait and aiming to maximize the purity of each subset is how it operates.

3. Random Forests

   Decision trees are expanded upon by random forests, which increase accuracy by building several decision trees and merging their output. For spotting more intricate fraud patterns, this can be helpful.

- Classification Random Forest: When the target variable is categorical, this kind of random forest is utilized to solve classification problems. Each decision tree in the ensemble of this particular Random Forest predicts the class label of a specific input data point, and the final prediction is determined by collecting the majority vote of all the decision trees.
- Regression Random Forest: For regression issues where the target variable is continuous, this kind of random forest is used. Each decision tree in the ensemble of this particular Random Forest predicts the continuous value of a specific input data point, and the final prediction is calculated by averaging all the decision trees.

4. Support Vector Machines

   SVMs are a kind of algorithm that can be applied to problems involving classification and regression. They operate by locating the best hyperplane that categorizes the data, which is helpful for spotting fraudulent transactions.

5. Neural Networks

   The structure and operation of the human brain served as the inspiration for this sort of algorithm. They are applicable to a variety of machine learning tasks, such as fraud detection. To increase accuracy, neural networks are frequently used in conjunction with other algorithms to recognize intricate patterns in data.

6. Gradient Boosting

   By merging the output of various decision trees, a technique called gradient boosting can be utilized to increase the accuracy of decision trees. This is helpful for spotting small patterns that point to fraud.

4.6.1.2 Performance of supervised algorithms

The effectiveness of supervised machine learning algorithms is influenced by a number of variables, including the quantity and quality of training data, the algorithm selected, the complexity of the model, and the hyperparameters applied. The effectiveness of supervised machine learning algorithms is typically measured using the metrics listed below:

- The most typical criterion for assessing how well classification algorithms perform is accuracy. The amount of instances in the test dataset that were correctly categorised is gauged.

- The effectiveness of binary classifiers is assessed using the metrics precision and recall. Recall assesses the proportion of real positives among all actual instances of positive data, while precision assesses the proportion of true positives among all positive predictions.
- An effective statistic to utilize when the dataset is unbalanced is the F1 Score, which is the harmonic mean of precision and recall.
- The performance of regression algorithms is frequently assessed using the Mean Squared Error (MSE) metric. Between the expected and actual values, it calculates the average squared difference.
- R-squared (R2) metric gauges how much of the target variable's variance is accounted for by the model. An improved model-data fit is indicated by a higher R2 value.
- Confusion Matrix can calculate the model's true positive, false positive, true negative, and false negative predictions are displayed in this table. It is used to figure out accuracy, recall, F1 score, and precision.

4.6.2 Unsupervised learning

Unsupervised learning models are developed to find unusual behavior in transactions that hasn't been seen before. Unsupervised learning models involve self-learning that helps in finding hidden patterns in transactions. In this type, the model tries to learn by itself, analyzes the available data, and tries to find the similarities and dissimilarities between the occurrences of transactions. This assists in detecting fraudulent actions.

4.6.2.1 Unsupervised learning algorithms

1. Clustering: clustering algorithms combine comparable data points. Clustering could be used in the context of identity fraud detection to group transactions or users with similar patterns of behavior, which could then be further studied to identify probable cases of fraud.
2. Anomaly detection: Data points that differ significantly from the other data points in the dataset are found using anomaly detection algorithms. Anomaly detection can be used to spot transactions or people who act strangely in the context of detecting identity fraud, such as when they make purchases in odd places or at odd hours.

3. Association rule mining: A dataset's patterns of co-occurring items are found using methods for association rule mining. The identification of user groups that are connected to one another based on traits or behaviors could be accomplished via association rule mining in the context of detecting identity fraud.

4. Dimensionality reduction: Using dimensionality reduction techniques, a dataset's most crucial details are preserved while fewer features are added. Dimensionality reduction may be used in the context of detecting identity fraud to lower the number of features connected to a user or transaction, making it simpler to spot patterns of behavior that might be fraudulent.

4.6.2.2 Performance of unsupervised algorithms

- True Positive Rate (TPR): the percentage of real abnormalities that the system accurately detects.
- False Positive Rate (FPR): The percentage of regular transactions that the algorithm wrongly incorrectly flags as anomalies.
- Precision: The percentage of real anomalies among highlighted anomalies.
- Recall: the percentage of real anomalies that the system successfully detects and flags.
- F1-Score: The harmonic mean of recall and precision, or the F1-score, strikes a compromise between the two measurements.

4.6.3 Semi-supervised learning

Unsupervised learning and supervised learning are both parts of semi-supervised learning. When labeling data necessitates the assistance of human experts and is either impossible or too expensive, it can be used. Even though it is unknown whose group the unlabeled data belongs to, a semi-supervised method for deep learning fraud detection maintains information about significant group parameters. It does so under the presumption that the patterns that have been found may still be useful.

4.6.3.1 Semi-supervised learning algorithms

1. Self-training

   A straightforward and well-liked method of semi-supervised learning is self-training. Self-training

involves training a classifier on labeled data first, after which it is used to label part of the unlabeled data. Once the freshly labeled data has been incorporated into the labeled dataset, the process is repeated until the classifier's performance is satisfactory.

2. Co-training

   Co-training is a semi-supervised learning strategy that employs two or more classifiers that have each been trained using a different perspective of the data. The labeled data is split into two sets, and each classifier is trained on a subset of the features. Each classifier is trained on a single set of labeled data before being used to label part of the unlabeled data. The freshly labeled data is then added to the matching labeled dataset, and the procedure is repeated until the classifiers' performance is sufficient.

3. Multi view learning

   Multiple perspectives of the data are combined in the semi-supervised learning method known as "multi-view learning" to enhance learning efficiency. Multi-view learning involves training a classifier on multiple views of the data, each of which is considered as an independent dataset. To arrive at a final forecast, the outputs of the classifiers are then combined.

4. Generative models

   A sort of semi-supervised learning method called generative models uses both labeled and unlabeled input to learn a probabilistic model of the data. After that, fresh data can be produced by the model to enhance its functionality.

4.6.3.2 Performance of semi-supervised algorithms

- Precision: Out of all the cases the model classified as fraud, this metric counts the percentage of cases that were accurately recognized as fraud. The model's high precision and low false positive rate indicate that it is successfully identifying fraud instances.

- Recall: Out of all the real fraud cases in the dataset, this metric measures how many were correctly recognized as fraud cases. A high recall indicates that a significant portion of the fraud instances in the dataset are being correctly identified by the model.

- F1 Score: It provides an overall evaluation of the model's success by combining precision and recall. It measures how well recall and accuracy are balanced and is the harmonic mean of both.

- The Receiver Operating Characteristic (ROC): Curve compares the true positive rate (TPR) and false positive rate (FPR) for various classification levels. Indicating a higher TPR and lower FPR, a good model will have a curve that is more pronounced in the upper left corner.
- Area Under the Curve (AUC): This metric measures the area under the ROC curve and provides an overall measure of the model's ability to distinguish between positive and negative cases. Model performance is greater when the AUC is higher.

The below graph depicts the performance of different machine learning techniques.

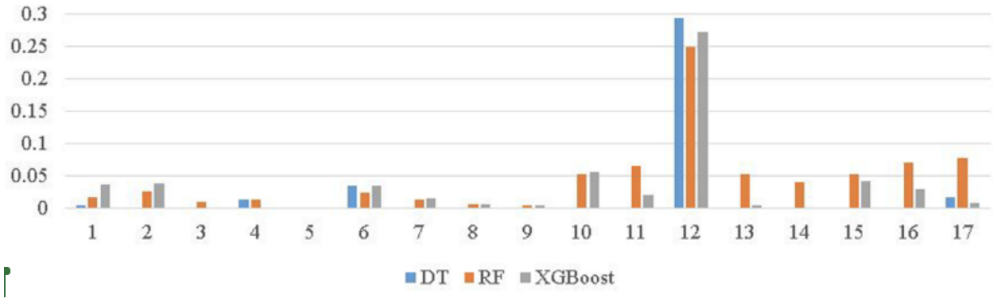

Figure 5: Comparison of Performance measurement between the machine learning techniques

4.6.4 Deep learning

Machine learning is a discipline that depends upon analyzing computer algorithms to learn and develop on its own, and can be viewed as a subset of deep learning. The study of computer algorithms to help computers learn and advance on their own is known as machine learning. It's considered to be a component of deep learning, which imitates human learning processes using artificial neural networks. Because of improvements in big data analytics, computers can now learn, observe, and react to complicated circumstances faster than people. Before, the intricacy of neural networks was constrained by the computational power available. Speech recognition, language translation, and image categorization are just a few of the fields where deep learning has been brought to use. Even without the intervention of humans, it can tackle pattern recognition issues. Common deep learning models typically involve CNNs for image identification, GANs for creating new data, and RNNs for processing natural language.

4.6.4.1 Overview of deep learning algorithms

Deep learning architecture includes a computational unit called a perceptron that enables the modeling of nonlinear functions, identical to how neurons serve as the basic building blocks of the brain. The fundamental perceptron is where deep learning's skill begins. The perceptron acquires a set of input signals and converts them into output signals in a manner identical to how a "neuron" in the human brain sends electrical pulses via our nervous system. The perceptron stacks numerous levels together to comprehend data representation, with each layer being in charge of perceiving a different aspect of the input. A neural network is said to be deep if it comprises several layers. A layer may be referred to as a group of computing components that learn to recognize values that recur over time. Deep learning systems require powerful hardware since they handle a substantial amount of information and carry out several intricate mathematical operations at once. Even with the most cutting-edge hardware, training a neural net could be challenging.

Deep learning algorithms are fed massive amounts of data as they require a great deal of information to get accurate results. The input is processed by artificial neural networks and classified using a series of binary true or false questions involving sophisticated mathematical calculations. An image recognition algorithm might, for example, acquire the ability to identify and detect the boundaries and curves of faces, then more significant facial features, and ultimately the overall depictions of faces. The likelihood of detecting the right answer goes up as the algorithm gains experience. In this case, time will allow the facial recognition algorithm to identify successfully.

- Choosing between machine learning and deep learning:
  - Depending on your application, the volume of data you're processing, and the kind of problem you're trying to address, machine learning offers a range of approaches and models from which to to choose. The model must be trained on numerous images, and a successful deep learning application needs GPUs, or graphics processing units, to analyze data efficiently. While deciding between machine learning and deep learning, consider factors like the high-performance GPU and the amount of labeled data you have available.
  - Machine learning may be more advantageous than deep learning if you lack either of those resources. In order to acquire accurate results with deep learning, we will need at least a few thousand images. Having a high performance GPU would be advantageous, as the time taken

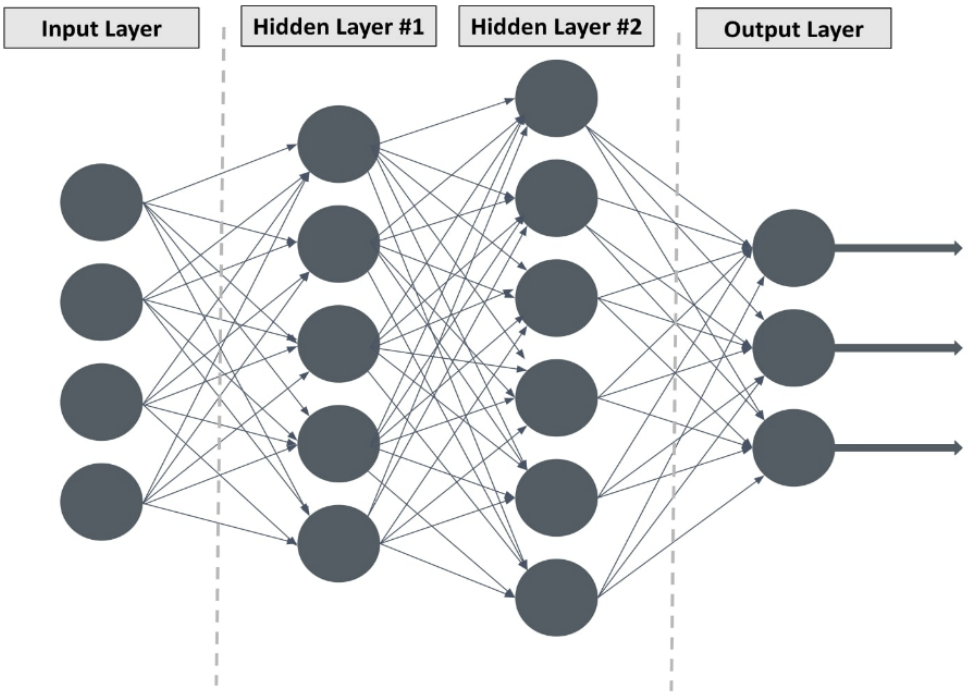

Figure 6: Neural Network

for analyzing the images reduces.

- Overview of deep learning in phishing:
  - Deep learning (DL) methods have gained popularity in the field of phishing URL detection in recent years. Researchers [42][43] have employed a variety of deep learning techniques, including Convolutional Neural Network (CNN), Auto-Encoder (AE), etc., to automatically extract abstract higher-level features from the raw URL. These algorithms are capable of extracting from raw data representations (features) that satisfy the requirements of the classifier. This helps to solve the issue of lexical-based feature extraction by removing the need for professionals to perform manual feature engineering. Moreover, DL models are adaptable to enormous amounts of data, and in fact, as the quantity of input data increases, so does their capacity for classification. DL models need the input data to be transformed into numerical vectors and the data needs to be preprocessed before feeding into the model. The time taken in training a neural network model is a significant burden as it involves training lots of parameters.
  - Deep learning algorithms have gained popularity due to their advantages over conventional ma-

chine learning classifiers. Deep learning algorithms have the capacity to learn features from an input automatically through repeated forward and backward data propagation. Few researchers have used neural network-based models for malicious URL detection in recent years.[41] To aid the unsupervised learning methodology, a variety of neural network-based techniques were applied to automatically extract inherent properties from raw URLs.

– A deep learning-based approach for predicting phishing attacks using time series data is presented by Mahmood et al. (2020)[24]. Phishing attacks pose a serious threat to internet security, and being able to predict them beforehand can help stop them from happening. The authors model the time series data of phishing assaults using a long short-term memory (LSTM) neural network. The likelihood of future assaults is predicted using the LSTM network, which was trained on a dataset of previous phishing attacks. In order to extract important characteristics from the time series data, such as the time of day, day of the week, and location of the attacks, the authors additionally employ feature engineering. The authors test the effectiveness of their strategy using a dataset of phishing assaults gathered from an enterprise network. The findings demonstrate that their strategy outperforms a number of benchmark methods in terms of forecasting precision and accuracy. The performance of the model is also the subject of a sensitivity study by the authors to look into the effects of various hyperparameters.

4.6.4.2 Performance of deep learning algorithms

The deep learning algorithms are very effective in handling enormous amounts of unstructured data which is present in the field of online identity fraud detection. The deep learning algorithms will learn the prominent features from the given raw data. This data will be used to detect the anomalies in the device usage and user activity. The algorithms like recurrent neural networks (RNNs) and convolutional neural networks (CNNs) would be useful in identifying fraudulent behavior.[4] The combination of the deep learning algorithms with other machine learning techniques may increase the accuracy of prediction of identity fraud cases and mitigating them beforehand. The machine learning techniques can be combined with Anomaly Detection or Clustering to achieve desirable results. Deep learning algorithms can be used to extract the prominent features from the raw data. This data can be further sent to other kinds of ML algorithms to enhance the accuracy of the predictions. The need to handle incomplete and noisy data should be addressed. Deep learning algorithms are effective in handling such drawbacks.

The necessity to take adversarial attacks into account is one of the difficulties with employing deep learning algorithms. In these attacks, fraudsters try to influence the algorithm by introducing noise or other types of interference into the data, which may result in predictions that are inaccurate or misleading. To increase the precision and resilience of the fraud detection system, deep learning algorithms that can successfully handle such adversarial attacks must be developed. The difficulty of balancing false positives and false negatives is a significant consideration when designing machine learning algorithms for online identity fraud detection. False positives happen when a valid transaction is mistakenly flagged as fraudulent, and false negatives happen when a valid transaction is not picked up as fraudulent. Since a high false positive rate can cause legitimate transactions to be blocked or delayed and a high false negative rate can cause fraudulent transactions to go undetected, striking the right balance between these two types of errors is essential. Hence, it's crucial to create deep learning algorithms that can tell the difference between genuine and fraudulent transactions precisely, while also reducing the likelihood of false positives and false negatives. These deep learning algorithms have the ability to handle huge amounts of unstructured data and to extract important features for finding fraudulent behavior patterns.

4.6.5 Ensemble learning

The implementation of machine learning for fraud and identity theft detection has grown in significance over the past few years as more private data is being stored online. It has been shown that ensemble learning is a useful method for enhancing the precision of fraud detection models. To create a reliable and accurate prediction, ensemble learning involves combining the predictions of various individual models. Multiple methods, including bagging, boosting, stacking, and hybrid ensembles, can be used to accomplish this.

4.6.5.1 Overview of ensemble learning

Many different independent models' predictions are combined by a family of algorithms known as ensemble machine learning approaches to get a more reliable and accurate forecast. Ensemble machine learning techniques have been employed by the authors of the publication "Current Research Landscape of Machine Learning Algorithms in Online Identity Fraud Prediction and Detection" [4] to increase the precision of online identity fraud prediction and detection. Machine learning ensemble techniques include bagging, boosting, stacking, and hybrid ensembles, among others. A technique called bagging entails building nu-

merous independent models from various subsets of the training data and averaging their predictions. While boosting, a technique entails developing a series of models, each of which focuses on the errors which were made by the previous models.

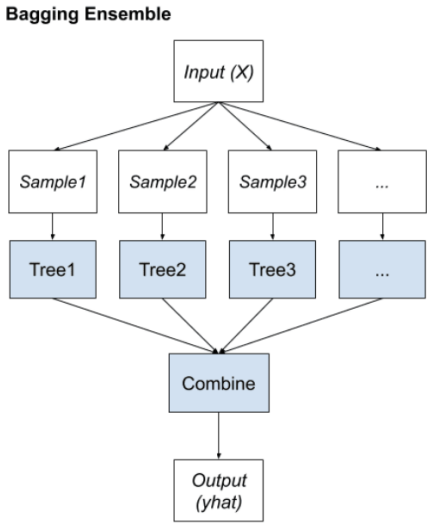

Figure 7: Bagging Ensemble

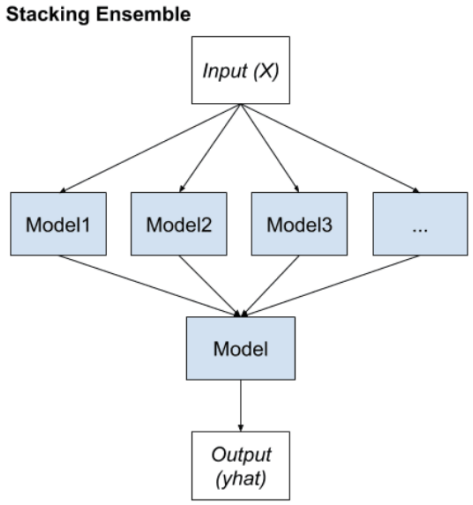

Figure 8: Stacking Ensemble

A technique called stacking incorporates training several models and then feeding the outputs of those

models into a meta-model. The final forecast is then created by the meta-model, which aggregates the predictions of the different models. As the name implies, hybrid ensembles combine two or more ensemble approaches to produce a stronger ensemble. Ensemble machine learning techniques have several benefits, including increased accuracy, robustness, and decreased overfitting. The performance of their models can be further enhanced by combining the ensemble approaches with additional machine learning techniques including feature selection, hyperparameter tuning, and class imbalance correction.

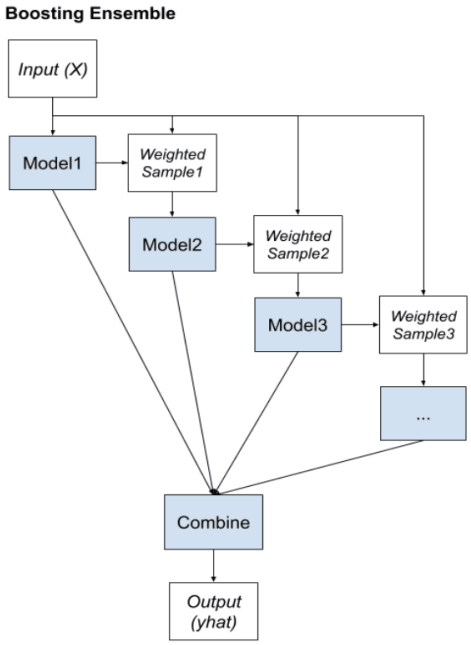

Figure 9: Boosting Ensemble

4.6.5.2 Performance of ensemble learning

According to studies, ensemble learning can greatly enhance the effectiveness of fraud detection algorithms. For instance, a study by Shi et al. [39] discovered that an ensemble of several deep learning models outperformed a single deep learning model in terms of detection rate and false positive rate. In a similar vein, Wang et al. [40] reported that a hybrid ensemble model mixing various machine learning techniques outperformed individual models in terms of accuracy and F1 score. Ensemble learning can assist increase the robustness of fraud detection models as well as their accuracy by lowering the possibility of overfitting. Due to the employment of many models with varying strengths and weaknesses in ensemble learning, which can aid in balancing the errors and individual biases.

The difficulties posed by imbalanced datasets and idea drift in fraud detection can be addressed with the use of ensemble learning. Ensemble learning can help to balance out individual biases and errors and evolve with changing fraud patterns over time by combining models with varied strengths and weaknesses. Ensemble learning has generally shown promise in reducing fraud and identity theft in web and mobile applications. Ensemble learning can help to safeguard private information and save financial losses for both individuals and companies by enhancing the precision, robustness, and dependability of fraud detection models.

4.6.6 Other machine learning techniques

In addition to ensemble and deep learning, the other machine learning techniques used for mitigating identity theft and fraud in web and mobile applications are: Decision Trees: Based on a set of if-then rules, decision trees are a common machine learning technique that may be used to categorize data into discrete categories. The most crucial characteristics for fraud detection can be found using decision trees, which are simple to understand.

- Random Forests: As an ensemble learning technique, random forests integrate various decision trees to increase the model's accuracy and robustness. Large datasets can be handled by random forests, which also lowers the chance of overfitting.

- SVMs (Support Vector Machines): SVMs are a potent machine learning method that can be applied to both classification and regression issues. With high-dimensional datasets, where there are many features that may be relevant to detecting fraud, SVMs are especially helpful for prediction.

- Naive Bayes: For classification issues, naive bayes is a straightforward probabilistic machine learning technique. For example, internet reviews or emails contain text data that can be used to detect fraud. Logistic Regression: To model the likelihood of a binary outcome, logistic regression is a statistical machine learning technique. To determine the probability that a transaction is fraudulent, logistic regression is frequently employed in fraud detection.

Ultimately, the dataset's unique properties and the objectives of the fraud detection system will determine which machine learning technique is used. Each method has its own advantages and disadvantages, there-

fore the best strategy may require combining several methods, such as using an ensemble of decision trees or hybrid models.

4.7 Application of machine learning in identity theft and prevention

4.7.1 Web and mobile applications in cloud

Using machine learning techniques to prevent fraud and identity theft in cloud-based online and mobile applications has great potential to improve the security and integrity of these programs. Machine learning algorithms can detect and stop suspicious activity, protect personal and financial information, and examine user patterns and behaviors to improve the overall security of cloud-based services.

As web and mobile applications in the cloud grow in popularity, so does the potential for identity theft. To combat this, machine learning approaches are used for fraud prevention and detection. To find patterns and predict future transactions, machine learning algorithms are trained on massive datasets of fraudulent and non-fraudulent transactions. For example, in the work of A. Roy et al. Created a deep learning model to identify fraudulent credit card transactions happening via cloud servers [2]. The accuracy of the model on the test device was 98.45%. It was trained on a dataset of 284,807 transactions.

Similarly, a machine learning-based adaptive fraud detection approach for fintech was presented in a study by Moon and Kim et al [6]. To identify fraudulent transactions in real time, the framework used supervised and unsupervised learning approaches. Results showed that the proposed approach is highly accurate in identifying fraudulent transactions.

4.7.2 Social networks

Identity thieves often target social networking websites. Identity theft on social networks is detected and prevented using machine learning technology. In a study explained in Villar-Rodriguez et al. [6] developed a new machine learning strategy to detect identity theft in social networks based on simulated attack cases and support vector machines. The proposed method was tested on a simulated attack data set and its accuracy was 98.9%.

Machine learning anomaly detection can also be used to preserve privacy in multimedia social networks [9]. The proposed method was based on a deep autoencoder model trained on a data set of typical and anomalous activities in social networks. The results showed that the proposed method is highly accurate in detecting abnormal behavior.

4.7.3 Online examinations

Online testing is growing in popularity due to its accessibility and simplicity. However, it is also vulnerable to identity theft. Identity theft in online exams is detected and prevented by machine learning algorithms. Location data can be used to identify and combat internal identity theft defaulters in online examinations [3]. The proposed method used a machine learning model to examine location data and identify anomalous activity patterns.

In the paper by Xiaochen Hu et al. (2021) [7], machine learning approaches were used to analyze characteristics and preventive actions for identity theft victims to correctly identify them. The proposed approach was based on a dataset of 9,214 identity theft victims and achieved an accuracy of 89.78% in predicting identity theft victims.

4.7.4 Other case studies and applications

There are several other case studies and applications of detection of identity thefts beyond the specific context of online examinations. One such example is in the field of financial fraud detection, where identity theft is a common tactic used by fraudsters. One study titled "Identity Fraud Detection using Machine Learning Techniques" explored the use of machine learning algorithms such as decision trees, random forests, and support vector machines for detecting identity fraud in financial transactions. The study achieved an accuracy of over 95% in detecting fraudulent transactions, suggesting that machine learning algorithms can be effective in detecting identity theft in financial transactions.

Another example is in the field of healthcare fraud detection, where identity theft is also a common tactic used by fraudsters. One study titled "Detecting Healthcare Fraud and Abuse using Machine Learning Techniques" explored the use of machine learning algorithms such as neural networks and decision trees for detecting healthcare fraud and abuse. The study achieved an accuracy of over 98% in detecting fraudulent

healthcare claims, indicating that machine learning algorithms can be effective in detecting identity theft in healthcare fraud. Overall, these case studies demonstrate the potential of machine learning algorithms for detecting identity theft in various contexts beyond online examinations. As the threat of identity theft continues to grow, the development and application of such algorithms can help prevent and detect fraud, protecting individuals and institutions from financial and reputational harm.

4.8 Detailed case studies of each application

4.8.1 Detect and prevent identity theft and fraud in web and mobile applications in the cloud

4.8.1.1 Identity fraud in web and mobile applications in the cloud

Identity fraud and scams are becoming more and more common in today's digital world. Web and mobile apps are particularly vulnerable to these kinds of attacks because of how frequently they manage private personal information like usernames, passwords, credit card information, and social security numbers. As fraudsters increasingly use technological methods, products, and channels to conduct crimes, businesses must use intelligent automated systems to reduce their exposure to fraud, and risk losses, and safeguard their organizations' reputations[35].

One method for detecting and preventing identity theft and fraud in cloud-based web and mobile apps is to use data analytics and machine learning techniques. By looking at user behavior, application usage trends, and other contextual data, anomalies and suspicious activity that might signify fraud can be identified. Machine learning algorithms can also be trained to recognize and prevent a specific type of assault, such as a phishing scam or a SQL injection attack[35].

4.8.1.2 Detection of identity theft

The development of computer technology and methods for data analysis has opened up new possibilities for the management and study of financial risks in the financial sector. Over the years, researchers have created a variety of anti-fraud measures and systems, including association rules, rule-based expert systems, and financial fraud modeling language. (FFML). These models, however, cannot be updated in real-time to reflect new frauds because they need adequate and precise expert knowledge.

Due to their effectiveness in reducing feature redundancy and producing deep feature representations for fraud detection, deep learning methods have gained popularity in the analysis of financial data. Deep networks can be useful in detecting fraud, according to research, when prior information is incorporated into them. Systems for detecting fraud can better spot possible fraudulent behavior and reduce the risk of identity theft by analyzing user characteristics [37].

Here the authors considered an Internet loan fraud detection model to demonstrate their approach The authors suggest using deep learning to identify fraud in a dataset of public lending. Customers' credit scores are examined because a low score suggests a larger chance of being a fraudulent user. The technique seeks to offer small financial credit companies a straightforward and practical model to enhance risk management and level of anti-fraud. They use the preprocessed data to train a deep neural network, then run tests to refine the network's design and hyperparameters. Finally, they run numerous tests to show that their suggested model performs better than widely employed models for detecting banking fraud [37].

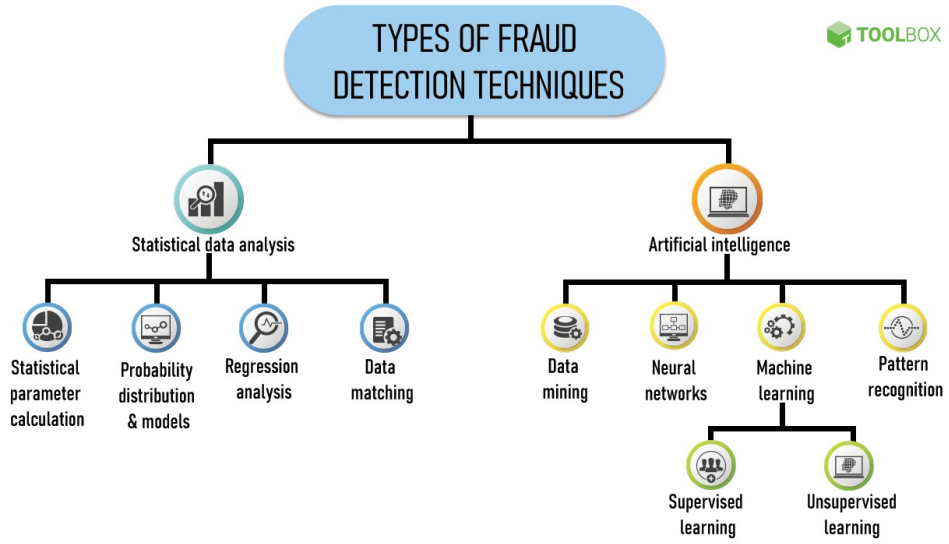

Figure 10: Types of Fraud detection techniques

4.8.1.3 Strategies of technologies used

The following frequent types of internet financial fraud include Identity theft, investment fraud, lending, and mortgage fraud, and widespread marketing fraud. Building an anti-fraud model requires careful consideration of data cleaning and feature selection. The writers deal with missing numbers and invalid or

unreasonable variables first. Then, they use an XGBoost model to select features and a random forest to generate missing variables, yielding 27 independent variable features and 1 target variable. The authors use a synthetic minority oversampling method to address the significant imbalance in the sample categories. After normalizing the data, the writers prepare it for modeling.

Building an anti-fraud model requires careful consideration of data cleaning and feature selection. The writers deal with missing numbers and invalid or unreasonable variables first. Then, they use an XGBoost model to select features and a random forest to generate missing variables, yielding 27 independent variable features and 1 target variable. The authors use a synthetic minority oversampling method to address the significant imbalance in the sample categories. After normalizing the data, the writers prepare it for modeling. An open lending dataset made available by Lending Club served as the experimental data for this research. The Internet fraud detection model was trained using 30 discriminant features that were chosen using XGBoost. Before being passed into a deep neural network with an input layer, six hidden layers, and an output layer, the feature data was normalized.

The median loan sum is $20,000, with the majority of loans falling in the $10,000 range. The risk is greater the longer the loan period, as the likelihood of default rises. The most popular loan utilization type is debt restructuring, followed by credit card repayment, and customers who use loans in these ways run a higher risk of default. Customers' credit ratings are broken down into seven categories, ranging from A to G, with better credit ratings being associated with lower default rates. Customers in categories B, C, and A make up 81.12% of the total, while those in E, F, and G make up 5.42%. Applicants with lower credit scores are subject to stricter credit management by Lending Club. To implement the models described in this paper, we make use of the TensorFlow framework. It matters how many nodes there are in the hidden layers. To identify the nodes in the concealed layers, the empirical method shown below is employed [37].

$$N_h = \frac{N_s}{(\alpha \times (N_i + N_o))}$$

4.8.1.4 Prevention

In terms of AUC, KS, and ACC, the efficacy of a deep neural network model is assessed and contrasted to four widely used models (logistic regression, support vector machine, decision tree, and random forest). The

results reveal that the deep neural network works better than the other models in all three metrics, proving its suitability for spotting online fraud. The DNN model is a promising method for uses other than Internet fraud detection, such as traffic flow forecasting, recommendation systems, medical image processing, and intelligent computing. It has the highest KS and ACC as well as the best AUC value. The deep neural network performed better than the other models, according to the results, making it a potentially useful tool for identifying internet fraud in the financial sector. To improve the model's ability to detect fraud by incorporating blacklists, whitelists, and anti-fraud rules, the authors intend to work in conjunction with established internet financial technology companies and Chinese banks [37]. Thus by employing advanced methodologies like those described above using deep neural networks we can mitigate internet fraud.

4.8.1.5 Challenges in adoption

Using deep learning methods for fraud detection and prevention in cloud-based web and mobile apps presents a number of difficulties, including the following:

- Deep learning algorithms need a lot of high-quality data to train on, but this data may be hard to come by or not fully representative of fraudulent actions.
- When it comes to fraud detection, the proportion of fraudulent transactions to legitimate transactions is frequently very small, which can cause a class imbalance and make it challenging for the algorithm to accurately identify fraud.
- Attackers may use strategies like data poisoning, in which they purposefully ingest fraudulent transactions into the training data to bias the model in their advantage, to evade detection.
- Deep learning models can be difficult to interpret, which can make it challenging to find the root cause of a fraud detection decision and make it difficult to explain the decision to stakeholders.
- The use of personal information in fraud detection can give rise to privacy issues, and it is important to thoroughly evaluate and address the possibility of this information being misused.
- As the volume of data and complexity of fraudulent activities increase, deep learning models must be able to scale to handle the additional load while keeping high accuracy and low false positive rates[35].

4.8.2 Detection of identity theft and fraud in social networks

4.8.2.1 Identity fraud in social networks

Social media platforms are being used more frequently, making them a favorite target for scams and identity theft. In social networks, identity theft and fraud can take many different forms, including clickjacking, phishing scams, and fake accounts. Social networks are a goldmine for cybercriminals looking to take identities or con users out of money because they amass enormous amounts of personal information. Due to social networks' dynamic nature, where users' behavior and network structure are continuously changing, it is difficult to detect identity theft and fraud there. To detect novel and undiscovered scam types, conventional rule-based methods fall short. Deep learning in particular, along with machine learning, has surfaced as a promising method for spotting fraud and identity theft in social networks [36].

Deep learning methods, like deep neural networks, can be trained to extract useful features from social network data, making it possible to discover fraudulent activities that are challenging to spot using conventional techniques. Deep learning models, for instance, can examine social network data patterns to spot fake profiles, recognize bots, and spot irregularities in user behavior [36].

4.8.2.2 Detection of identity theft

To find duplicated or fake profiles, machine learning techniques can also be used. The algorithms are able to examine user behavior trends and identify any unusual behavior that might point to the existence of a forged or copied profile. To increase the algorithms' accuracy, they can be trained on big datasets of real user-profiles and fictitious profiles. Using network analysis is another method for spotting duplicated or fraudulent accounts. In order to look for any unusual trends that might point to the existence of a fake or cloned profile, network analysis can look at the connections between users. One sign that a user may be using a fake or duplicate profile is if they have numerous profiles with the same group of friends.

The increasing number of users accessing various social media platforms has resulted in large amounts of data being shared and stolen at a very large scale. To tackle these issues, researchers have proposed various mechanisms and approaches. Profile cloning, malware attacks, data leakage, Trojan applications, friend-in-the-middle threats, public information harvesting, social bots, and other threats have all been

addressed by experts in a variety of ways. Similarity index-based strategies, automated privacy management programs, security frameworks, and the use of preset privacy settings are a few of the methods that have been suggested. Studies on the importance of information security awareness and Islamic perspectives on social media security and privacy problems are also available [38].

4.8.2.3 Strategies of technologies used

The calculated Similarity Index parameter serves as the foundation for the suggested method. To identify a user attributes like name, location, profile picture, and date of birth are examined. In order to determine the Similarity Index, each trait is given a weight based on how significant it is. The weight placed on the profile picture is the highest. By using the Facebook graph API to find accounts with the same name, the same site profile cloning can be found. By looking up the user's profile on the OSN using their actual name, comparing it to the original profile, and calculating the similarity index, cross-site profile cloning is discovered. High-similarity index profiles are classified as copied profiles, and low-similarity index profiles are labeled as fake profiles.

Using a variety of tools like WEKA, Eclipse, and ARFF data files, the scientists have suggested a model. An open-source program called WEKA offers features for handling various machine learning algorithms. Applications based on constraint programming are created and deployed using Eclipse. The data files used to hold the information about the attributes and their types are called ARFF (Attribute Relationship File Format) files.

The data from the provided user's profile is extracted in order to determine whether their profile has been copied on the same social media platform. Then, from the same social networking site, the same characteristics for profiles with the same username are taken. To determine whether the biography is a clone or a fake, the Similarity Index is computed. Profiles with users who share the same name as the provided user are chosen, and attributes are extracted, in order to determine whether a profile has been copied on another social networking site such as Instagram, LinkedIn, or Google+. The similarity index is then computed, and the results are compared to determine whether or not the profile is fraudulent [38].

4.8.2.4 Prevention

Users of OSN can use the aforementioned method to prevent duplicated and fake profiles as well as future threats of identity theft in the current landscape of social networks. The newest social engineering methods will undoubtedly include fake profiles as they continue to develop. The quantity of information shared on online social networks has also been reduced in an effort to influence a person's popularity. Such assaults are readily curbed [38].

4.8.2.5 Challenges in adoption

- The accuracy of the ML model can be affected by the quality and quantity of data that are accessible for analysis. The model's performance may be compromised in some instances by incomplete, inconsistent, or inaccurate data.
- Social networks are dynamic in nature because users frequently add, remove, and change the information on their accounts. Because of this, it might be challenging to monitor alterations and spot fraud.
- Attackers may employ a variety of strategies to avoid discovery, including the creation of numerous false profiles, the use of identities that have been stolen, or the manipulation of data to make it seem real.
- Users of social networks may be hesitant to share personal information due to privacy concerns, which can reduce the amount of data that is accessible for analysis. Additionally, the use of specific kinds of data for ML analysis may be constrained by privacy laws.
- Depending on the data used to teach them, ML models may be biased. The model may not correctly represent real-world scenarios if the training data is skewed or lacking, which can result in false positives or false negatives.
- It can be challenging to comprehend how ML models came at a specific conclusion or prediction because they can be challenging to interpret. This may make it more difficult to spot and correct any possible biases or errors in the model [36].

4.8.3 Detection of cheating in the online examinations using deep learning approach

Due to the Covid-19 pandemic, there has been a surge in online learning and examinations, which has made it difficult for educational institutions to prevent academic dishonesty. In response to this issue, the authors of "Detection of Cheating at Online Examinations Using Deep Learning Approach" have proposed a solution that utilizes an e-cheating intelligence agent comprising of two major modules: the internet protocol detector and the behavior detector.[17]

The e-cheating intelligence agent provides a solution to the problem of academic dishonesty in online assessments and can aid in preserving the integrity of online exams. However, it is crucial to note that the proposed method should be regarded as a tool for preventing and detecting cheating practices, rather than a replacement for other measures that ensure the integrity of online exams, such as proctoring and academic integrity policies. One limitation of the method is that it requires a large amount of data to train the deep learning model. This can be challenging, especially for academic institutions with limited resources. Furthermore, the method is only effective at detecting cheating behavior that can be captured by digital means, such as keystroke patterns and mouse movements. It cannot detect more traditional forms of cheating, such as plagiarism or using external resources during an exam.

Future research in this area could focus on developing more sophisticated deep learning algorithms that can detect a wider range of cheating behaviors. Additionally, the article suggests that academic institutions should consider implementing multiple layers of detection and prevention measures to ensure academic integrity. These measures could include using proctoring software, designing exams that are difficult to cheat on, and providing students with education and training on academic honesty. The proposed method has the potential to be used in various online learning programs, and its implementation can help maintain the integrity of online assessments. Overall, the paper presents a promising solution to the issue of academic dishonesty in online assessments, which has become increasingly relevant during the Covid-19 pandemic.[17]

4.8.3.1 Identity fraud in online examinations

Identity fraud in online examinations refers to the act of a student assuming the identity of another individual to take an exam. This form of academic dishonesty has become increasingly prevalent in recent years,

particularly in the context of online exams where it is more difficult to verify the identity of students. To prevent identity fraud in online examinations, various measures can be implemented, such as using biometric authentication, requiring students to show identification documents during the exam, and conducting live proctoring. However, these measures can also have their limitations and may not be feasible for all institutions.

Using machine learning algorithms, such as deep learning, can also be a potential solution to the problem of identity fraud in online examinations. These algorithms can be trained to recognize specific patterns of behavior that are associated with fraudulent activity, such as variations in typing speed, mouse movements, and keystrokes. By analyzing these patterns, the algorithms can detect and flag suspicious behavior during the exam.[17] Overall, addressing identity fraud in online examinations requires a multi-faceted approach that includes the use of various measures, including machine learning algorithms. By taking proactive steps to prevent identity fraud, institutions can help ensure the integrity of online examinations and maintain the credibility of academic qualifications.

4.8.3.2 Detection of identity theft

Identity theft detection is the process of identifying and preventing fraudulent activity that aims to gain unauthorized access to an individual's personal and financial information. As mentioned in the previous paragraphs, identity theft is a growing concern in today's digital world, with the rise of credit card and mobile phone fraud, as well as attacks on social networks.[1] To detect identity theft, various technologies are being developed, including data mining, machine learning algorithms, and biometric identity checks. Fraud management systems have been implemented in the financial services and telecommunication industries to prevent credit card and mobile phone fraud. These systems use a combination of technologies, such as neural networks, case-based reasoning, and support vector machine ensembles, to identify patterns of fraudulent activity.[1] Moreover, in the context of social networks, identity theft detection is perceived as a binary hypothesis testing problem. A proposed identity theft detector works in three stages with alarm triggers. The first detector's objective is to sound the first alert of a potential identity theft attempt, which could then be passed on to the other two detection stages. These supplemental detectors would make use of the user's social network or the content itself. The detector works by training on two profiles generated by the real user and impostor's usage and pattern which undergoes feature extraction and is trained on a

machine learning algorithm.[12]

In conclusion, identity theft detection is crucial to protect individuals from fraudulent activity that aims to misuse their sensitive information. With the development of new technologies such as data mining, machine learning algorithms, and biometric identity checks, businesses can detect and prevent identity theft effectively. However, it is important to continue researching ways to increase trust and partnership with users and to predict rare events like fraud.

4.8.3.3 Strategies of technologies used

Machine learning is a powerful tool that can be used to prevent identity theft by analyzing patterns and data in real-time to detect potential fraud. One strategy is to implement fraud detection software which utilizes machine learning algorithms to analyze large amounts of data in real-time, looking for patterns that may indicate fraudulent behavior. By quickly identifying suspicious activity, this software can help prevent fraudsters from successfully stealing an individual's identity.[12] Another strategy is to use machine learning algorithms to analyze data from various sources to identify patterns and detect fraudulent activity. For example, machine learning algorithms can be used to analyze credit card transactions and flag suspicious activity, such as purchases made in different geographic locations or large purchases made outside of an individual's normal spending habits.[12]

In addition, machine learning can be used to improve user authentication processes. For example, machine learning algorithms can analyze an individual's browsing behavior, such as the websites they visit or the way they type, to verify their identity. By leveraging these unique behavioral patterns, organizations can help ensure that only authorized individuals have access to sensitive data. Contextually, the use of machine learning can significantly improve an organization's ability to prevent identity theft. By implementing fraud detection software and leveraging machine learning algorithms to analyze data and improve user authentication processes, organizations can quickly identify and respond to potential fraudulent activity. This can help protect individuals' personal data and prevent the devastating consequences of identity theft.

4.8.3.4 Prevention

The use of machine learning algorithms is a robust strategy for preventing identity theft, as they can quickly detect and prevent potential threats by analyzing patterns in data in real-time. To prevent identity theft using machine learning, organizations can implement fraud detection software that analyzes large amounts of data to identify patterns that may indicate fraudulent behavior. This can help organizations to quickly identify and block potential threats, safeguarding their customers' sensitive information.

Another strategy is to use predictive analytics to identify potential risks before they occur. By analyzing common patterns and trends in data, predictive analytics can help organizations identify potential vulnerabilities and take proactive steps to prevent fraud before it happens.[12] Additionally, machine learning algorithms can be used to improve authentication processes. By analyzing user behavior, such as the way they interact with websites or mobile applications, these algorithms can identify potential fraudsters and prevent them from accessing sensitive information.

By investing in machine learning technologies, organizations can improve their security posture and protect individuals' personal data. Implementing a multi-layered approach to security, including using fraud detection software, predictive analytics, and improving authentication processes, can help prevent identity theft before it occurs.[12] In summary, preventing identity theft using machine learning requires a proactive approach. Organizations must invest in effective technologies and take a multi-layered approach to security to detect and prevent potential threats. By doing so, they can protect their customers' personal and financial information and maintain their trust in the organization.

4.8.3.5 Challenges in adoption

While technology can be an effective tool for preventing identity theft, there are several challenges that may hinder its widespread adoption. One significant challenge is the cost of implementing these technologies. Smaller organizations may not have the financial resources to invest in expensive fraud detection software or other security measures, leaving them vulnerable to cyber attacks.[12] Another challenge is the potential for false positives. Fraud detection software relies on complex algorithms that analyze patterns and data in real-time to detect potential fraud. While these algorithms can be highly effective, they may also flag legitimate transactions as fraudulent, leading to customer frustration and mistrust.

Additionally, there is a lack of standardization across different industries and organizations when it comes to identity theft prevention. This can create confusion and inefficiencies for organizations that must navigate a patchwork of different regulations and best practices. It can also make it difficult for consumers to know which organizations are using the most effective technologies to protect their data. Data privacy is another major concern. With the increasing amount of personal data being collected, there is a growing risk of data breaches and identity theft. Many individuals are hesitant to share their personal information due to concerns about privacy and data security.

Finally, the rapid pace of technological change presents a challenge for organizations that must continually invest in updating and upgrading their security systems to stay ahead of fraudsters. To address these challenges, organizations must be willing to invest in the most effective technologies for preventing identity theft while also addressing concerns around cost, privacy, and data security. This may require working with industry partners and regulatory agencies to develop common standards and best practices. It may also require investing in ongoing training and education for employees to ensure they are equipped to detect and prevent fraud. Ultimately, the adoption of these technologies will require a concerted effort from all stakeholders, including individuals, organizations, and governments, to protect sensitive data and prevent the devastating consequences of identity theft.[12]

4.9 Integrating machine learning solutions into live systems

Fraud and impersonation detection and prevention benefit greatly from machine learning. There is great promise to integrate machine learning techniques into operating systems to reduce fraud and identity theft in cloud-based online and mobile applications. There are various steps that must be followed to integrate machine learning techniques into live systems [23].

The first step is to collect data from various sources such as financial institutions, social media sites, mobile and web applications. To get better results from machine learning algorithms, data must be well-researched, accurate, and complete. Data preprocessing or cleaning, transformation, and normalization are the second phase. This step is essential as it ensures that the data is in a form that can be easily processed by machine learning algorithms. Choosing the right machine learning algorithm is the third step. Machine learning

algorithms come in many forms, including supervised, unsupervised, semi-supervised, deep learning, and ensemble learning. Depending on your data type and your current question, you should choose the right algorithm. In the fourth step, the selected machine learning algorithm is trained on the preprocessed data. You need to select features and train the algorithm using the training set. Testing the trained algorithm on the test set is the fifth stage. This process helps evaluate algorithm performance and find model errors. Then, as a final step, you need to integrate your machine learning model into your operational system. This step requires careful evaluation of system requirements such as data throughput, accuracy, and speed [26] [27].

Integrating machine learning solutions into live systems can be challenging due to architectural complexity and the need for real-time processing. However, if properly planned and executed, machine learning technology can greatly improve system accuracy and reduce the risk of fraud and identity theft. In summary, integrating machine learning techniques into the operating system can significantly reduce fraud and identity theft in cloud-based web and mobile services. Data, machine learning algorithms, training, testing and integration should be carefully considered. Future research may explore more complex methods for preparing data, choosing optimal algorithms, and integrating machine learning models into operating systems.

4.10 Regulatory compliance

Regulatory compliance issues with cloud-based web and mobile applications that employ machine learning to combat fraud and identity theft must be handled. Compliance with privacy and security laws is necessary to prevent legal repercussions and reputational harm because they differ by region and nation. We'll examine some of the most important compliance concerns in this part when utilizing machine learning to thwart fraud and identity theft.

4.10.1 General Data Protection Regulation (GDPR)

The European Union (EU) Directive, the General Data Protection Regulation (GDPR), defines the principles of protection of personal data of people living in the EU. All companies that process the personal data of EU citizens, regardless of their location, must comply with the GDPR [4]. One of the key aspects of the GDPR is the requirement that any processing of personal data must be based on the express consent of the data subject. This includes detecting and preventing fraud and identity theft using machine learning

algorithms. The processing of personal data must be secure and organizations must implement the necessary organizational and technical precautions to prevent unauthorized access or disclosure [9].

4.10.2 Payment Card Industry Data Security Standard (PCI DSS)

Businesses that receive, send or store credit card data must meet the security requirements set forth in the Payment Card Industry Data Security Standard (PCI DSS) [2]. Companies that process credit card data must comply with PCI DSS, which requires machine learning based fraud detection and prevention. Organizations must maintain secure networks, regularly review and test security systems and procedures, and comply with PCI DSS to ensure fraud detection and prevention [12].

4.10.3 Health Insurance Portability and Accountability Act (HIPAA)

Patient health information security requirements are established by a US law known as the Health Insurance Portability and Accountability Act (HIPAA). HIPAA applies to many organizations, including healthcare providers, insurers and clearinghouses [9]. The processing of protected health information may be necessary to apply machine learning to detect and combat medical fraud. To protect the privacy, security, and availability of protected health information and to prevent unauthorized access and disclosure, HIPAA requires covered entities to implement necessary technical and administrative measures [13].

Figure 11: Comparison between GDPR, PCI DSS and HIPAA

4.10.4 Federal Trade Commission Act (FTC)

Federal Trade Commission (FTC) regulations prohibiting unfair and deceptive trade practices must be enforced. applies to all companies that collect or maintain records of US personally identifiable information. Some of the regulatory data security challenges are data encryption, access restrictions and data breach notifications [10]. Companies must comply with FTC regulations when using machine learning-based fraud detection and prevention solutions for online transactions. In conclusion, the use of machine learning in cloud-based websites and mobile applications requires strict compliance with regulations to reduce fraud and identity theft. Companies must ensure that their solutions comply with a wide range of regulations, including FTC regulations, GDPR, HIPAA and PCI DSS. By adhering to these standards, organizations can protect their users' financial and personal information while avoiding financial and legal consequences.

5 Conclusion and Recommendations

To conclude, you can increase the security of websites and apps that live in the cloud by using complex machine learning methods to avoid cyber crime. To boost up the overall security of everything in the cloud, those machine learning technologies can block suspicious activities, safeguard personal and financial information, and examine user habits and behaviors. Plus, they can be great at identifying and stopping fraudulent transactions, detect and prevent suspect login attempts, and examine user behavior for signs of possible identity theft.

5.1 Challenges and limitations

5.1.1 Challenges and limitations of machine learning techniques

Machine learning techniques have been shown to reduce identity theft and fraud in cloud-based web and mobile applications. These methods, however, have their own set of challenges and limitations. In this section, we will look at the challenges and limitations of machine learning in the context of preventing identity theft and fraud.

- Data quality: The quality of the data used by machine learning models is critical to their success and poor data quality can lead to inaccurate and unreliable predictions, which can be detrimental in the detection of identity theft and fraud. Many factors influence data quality, including data accuracy, completeness, and consistency. [8]

- Limited explainability: Machine learning models can be difficult to interpret, making it difficult to understand how they make predictions. This lack of interpretability can make identifying and addressing potential model issues difficult which can make explaining the model's predictions to stakeholders difficult.[14]

- Limited scalability: Machine learning models are computationally expensive and may not scale well to large datasets or high-throughput environments. This can make deploying these models in real-world scenarios where speed and scalability are critically difficult.[14]

- Limited applicability: Limited generalizability: Machine learning models can be sensitive to changes in data distribution, limiting their generalizability to new and unseen scenarios. This is particularly

challenging in the context of identity theft and fraud, where attackers constantly adapt their strategies and tactics.

- Fairness and bias: Machine learning models are prone to bias and unfairness, particularly if the data used to train them is biased. As this can have discriminatory consequences and exacerbate existing inequalities using machine learning to detect identity theft and fraud, fairness, and bias are critical considerations. These issues have serious legal and ethical ramifications.

- Resource Constraints: Machine learning models are often computationally expensive and need a lot of storage, which could be a problem in contexts with restricted resources like mobile devices. Techniques like model compression and quantization can be utilized to shrink the model while boosting its effectiveness.[8]

We can say that machine learning techniques have the potential to reduce identity theft and fraud in cloud-based web and mobile applications. As these techniques, however, have their own set of challenges and limitations it is critical to keep these challenges and limitations in mind when using machine learning for identity theft and fraud detection, to ensure that models are effective, dependable, and fair.

5.1.2 Privacy concerns and ethical considerations

The use of machine learning approaches for mitigating identity theft and fraud in web and mobile applications on the cloud has raised a number of privacy concerns and ethical considerations. This section provides an overview of some of the key issues that need to be addressed in order to ensure that the use of machine learning techniques in this context is both effective and ethical.

- Data privacy: To mitigate the risks associated with the use of machine learning techniques for the collection and processing of large amounts of personal data privacy laws and regulations must be followed. This data may include social security numbers, banking numbers, and other sensitive information. [8]

- Adversarial Attacks: Adversarial attacks involve purposefully manipulating data in order to fool machine learning models which is a significant issue in the detection of identity theft and fraud, as fraudsters may attempt to circumvent the machine learning model's defenses. Adversarial training and detection techniques can be used to strengthen the model's resistance to such attacks.

- Transparency: The use of machine learning algorithms for mitigating identity theft and fraud in web and mobile applications on the cloud needs to be transparent. This implies that users should be informed about the type of data that is being collected, how it is being used, and who has access to it. This will help build trust among users and ensure that their privacy is protected.

- Fairness and Bias: Machine learning models are prone to bias and unfairness, particularly if the data used to train them is biased. As this can have discriminatory consequences and exacerbate existing inequalities using machine learning to detect identity theft and fraud, fairness, and bias are critical considerations. These issues have serious legal and ethical ramifications.

- Accountability: The use of machine learning techniques for mitigating identity theft and fraud in web and mobile applications on the cloud requires accountability. This means that those responsible for the development and deployment of these techniques should be accountable for their actions and that there should be mechanisms in place for users to seek redress if their rights are violated.

In summary, the use of machine learning techniques for mitigating identity theft and fraud in web and mobile applications on the cloud requires careful consideration of a number of privacy and ethical issues. It is important to ensure that data privacy laws and regulations are followed, that algorithms are transparent and unbiased, and that fairness and accountability are maintained throughout the development and deployment process. By taking these factors into account, it is possible to use machine learning techniques in a way that is both effective and ethical.

5.2 Further challenges

Although improvements have been made in the development and application of machine learning methods to reduce fraud and identity theft, there are still many issues that need to be resolved. This part highlights these upcoming problems.

- Explainability and Interpretability of Machine Learning Algorithms: Without a full understanding of how they work, machine learning algorithms can be perceived as mysterious "black boxes." To understand how machine learning models work and identify potential biases, it is important to analyze them. There is a need for more interpretable and explainable models that clearly explain how to arrive at results [1].

- Data quantity and quality: The performance of machine learning algorithms is highly influenced by the quality and quantity of data used to train the algorithm. The quantity and quality of data poses significant challenges in preventing fraud and identity theft. Data is often difficult to obtain, and privacy concerns can make it difficult to obtain quality data. Addressing these issues requires creating new approaches to data acquisition and processing [8].

- Scalability: Another issue is the scalability of machine learning techniques. Machine learning models can be computationally intensive when used in real-time or large-scale computing environments. As a result, we need to develop a paradigm that scales well and effectively [14].

5.3 Future work

Despite the above challenges, there is still much room for progress in machine learning-based fraud and identity theft prevention. The following are potential areas for future research.

- Hybrid Machine Learning Techniques: Anti-fraud and identity theft systems may work more effectively using hybrid machine learning techniques. For example, combining supervised and unsupervised learning strategies can produce more reliable and accurate models [28].

- Integration with Blockchain: Solutions for preventing fraud and identity theft can gain from the enhanced security and anonymity provided by blockchain technology. A more organized and secure method of preventing fraud and identity theft can be made possible by combining blockchain technology with machine learning [29].

- Collaboration: Collaboration between academics and experts in fields such as computer science, psychology, criminology and law enforcement will create more efficient systems to prevent identity theft and fraud. Collaboration could also help identify new research areas and challenges that need to be addressed.

- Explainable AI: Explainable AI helps solve the interpretability problem of machine learning systems. By creating machine learning models that can provide accurate insight into decision making, identity theft and fraud prevention systems can be made more reliable and effective [30].

6 References

[1] R. Bose, "Intelligent Technologies for Managing Fraud and Identity Theft," Third International Conference on Information Technology: New Generations (ITNG'06), Las Vegas, NV, USA, 2006, pp. 446-451, doi: 10.1109/ITNG.2006.78.

[2] A. Roy, J. Sun, R. Mahoney, L. Alonzi, S. Adams and P. Beling, "Deep learning detecting fraud in credit card transactions," 2018 Systems and Information Engineering Design Symposium (SIEDS), Charlottesville, VA, USA, 2018, pp. 129-134, doi: 10.1109/SIEDS.2018.8374722.

[3] Y. Cho and S. Lee, "Detection and Response of Identity Theft within a Company Utilizing Location Information," 2016 International Conference on Platform Technology and Service (PlatCon), Jeju, Korea (South), 2016, pp. 1-5, doi: 10.1109/PlatCon.2016.7456790.

[4] B. Conlin and U. Ruhi, "Current Research Landscape of Machine Learning Algorithms in Online Identity Fraud Prediction and Detection," 2021 IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD), Marrakech, Morocco, 2021, pp. 1-6, doi: 10.1109/ICTMOD52902.2021.9739308

[5] A. A. Yilmaz, "Intrusion Detection in Computer Networks using Optimized Machine Learning Algorithms," 2022 3rd International Informatics and Software Engineering Conference (IISEC), Ankara, Turkey, 2022, pp. 1-5, doi: 10.1109/IISEC56263.2022.9998258.

[6] Moon, Woo Young, and Soo Dong Kim. "Adaptive fraud detection framework for fintech based on machine learning." Advanced Science Letters 23.10 (2017): 10167-10171.

[7] Xiaochen Hu, Xudong Zhang Nicholas P. Lovrich (2021) Forecasting Identity Theft Victims: Analyzing Characteristics and Preventive Actions through Machine Learning Approaches, Victims Offenders, 16:4, 465-494

[8] R. Banerjee, G. Bourla, S. Chen, M. Kashyap and S. Purohit, "Comparative Analysis of Machine Learning Algorithms through Credit Card Fraud Detection," 2018 IEEE MIT Undergraduate Research Technology Conference (URTC), Cambridge, MA, USA, 2018, pp. 1-4, doi: 10.1109/URTC45901.2018.9244782.

[9] Randa Aljably, Yuan Tian, Mznah Al-Rodhaan, "Preserving Privacy in Multimedia Social Networks Using Machine Learning Anomaly Detection", Security and Communication Networks, vol. 2020, Article ID 5874935, 14 pages, 2020. https://doi.org/10.1155/2020/5874935

[10] A. L. Buczak and E. Guven, "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection," in IEEE Communications Surveys Tutorials, vol. 18, no. 2, pp. 1153-1176, Secondquarter 2016, doi: 10.1109/COMST.2015.2494502.

[11] Jang, M., Song, J. Kim, M. A Study on the Detection Method for Malicious URLs Based on a Number of Search Results Matching the Internet Search Engines Combining the Machine Learning. J. Electr. Eng. Technol. 17, 617–626 (2022).

[12] Villar-Rodríguez, Esther, et al. "A novel machine learning approach to the detection of identity theft in social networks based on emulated attack instances and support vector machines." Concurrency and Computation: Practice and Experience 28.4 (2016): 1385-1395.

[13] Y. Yang, M. Manoharan and K. S. Barber, "Modelling and Analysis of Identity Threat Behaviors through Text Mining of Identity Theft Stories," 2014 IEEE Joint Intelligence and Security Informatics Conference, The Hague, Netherlands, 2014, pp. 184-191, doi: 10.1109/JISIC.2014.35.

[14] Chandra, J.V., Challa, N., Pasupuleti, S.K. (2020). Detection of Deceptive Phishing Based on Machine Learning Techniques. In: Fiaidhi, J., Bhattacharyya, D., Rao, N. (eds) Smart Technologies in Data Science and Communication. Lecture Notes in Networks and Systems, vol 105. Springer, Singapore. https://doi.org/10.1007/978-981-15-2407-3_2

[15] W. Fang, X. Li, P. Zhou, J. Yan, D. Jiang and T. Zhou, "Deep Learning Anti-Fraud Model for Internet Loan: Where We Are Going," in IEEE Access, vol. 9, pp. 9777-9784, 2021, doi: 10.1109/ACCESS.2021.3051079.

[16] M. E. Rothrock,"Comments, with reply, on 'Is copyright law steering the right course' by P. Samuelson," in IEEE Software, vol. 5, no. 6, pp. 12-, Nov. 1988, doi: 10.1109/52.9998.

[17] Tiong, Leslie Ching Ow, and HeeJeong Jasmine Lee. "E-cheating Prevention Measures: Detection of Cheating at Online Examinations Using Deep Learning Approach–A Case Study." arXiv preprint arXiv:2101.09841 (2021).

[18] Al-Sarraj, W., Al-Ayyoub, M. (2020). Identity Theft Detection and Prevention Approaches: A Survey. Journal of Information Privacy and Security, 16(1), 1-24.

[19] Stojanović, B.; Božić, J.; Hofer-Schmitz, K.; Nahrgang, K.; Weber, A.; Badii, A.; Sundaram, M.; Jordan, E.; Runevic, J. Follow the Trail: Machine Learning for Fraud Detection in Fintech Applications. Sensors 2021, 21, 1594. https://doi.org/10.3390/s21051594

[20] Copes, Heith, et al. "Differentiating identity theft: An exploratory study of victims using a national victimization survey." Journal of Criminal Justice 38.5 (2010): 1045-1052.

[21] Villar-Rodriguez, Esther Del Ser, Javier Salcedo-Sanz, Sancho. (2015). On a Machine Learning Approach for the Detection of Impersonation Attacks in Social Networks. 10.1007/978-3-319-10422-5_28.

[22] Q. Ye et al., "Modeling Access Environment and Behavior Sequence for Financial Identity Theft Detection in E-Commerce Services," 2022 International Joint Conference on Neural Networks (IJCNN), Padua, Italy, 2022, pp. 1-8, doi: 10.1109/IJCNN55064.2022.9892383

[23] Vapnik, V. N. (2013). The nature of statistical learning theory (Vol. 13). Springer Science Business Media.

[24] S. H. Amin Mahmood, S. Mustafa Ali Abbasi, A. Abbasi and F. Zaffar, "Phishcasting: Deep Learning for Time Series Forecasting of Phishing Attacks," 2020 IEEE International Conference on Intelligence and Security Informatics (ISI), Arlington, VA, USA, 2020, pp. 1-6, doi: 10.1109/ISI49825.2020.9280509.

[25] N. Kumar, P. Dabas and Komal, "Detection and Prevention of Profile Cloning in Online Social Networks," 2019 5th International Conference on Signal Processing, Computing and Control (ISPCC), Solan, India, 2019, pp. 287-291, doi: 10.1109/ISPCC48220.2019.8988394

[26] Dahee Choi1and Kyungho Lee, Center for Information Security Technologies (CIST), Korea University, Seoul 02841, Republic of Korea ; An Artificial Intelligence Approach to Financial Fraud Detection under IoT Environment: A Survey and Implementation

[27] Y. Zhang, M. Li, Y. Li, and J. Li, "An Efficient Phishing Website Detection Method Based on Random Forest and Decision Tree," in Proceedings of the 2018 5th International Conference on Information Science and Control Engineering (ICISCE), 2018, pp. 1389-1393. DOI: 10.1109/ICISCE.2018.00254

[28] Haddad, M., Abed, A. (2019). Detection of social engineering attacks using machine learning techniques in social media. Journal of Cybersecurity and Mobility, 8(2), 141-164. doi: 10.13052/jcsm2245-1439.821

[29] O. Hussein, "A Proposed Anti-Fraud Authentication Approach for Mobile Banking Apps," 2022 4th Novel Intelligent and Leading Emerging Sciences Conference (NILES), Giza, Egypt, 2022, pp. 56-61, doi: 10.1109/NILES56402.2022.9942402.

[30] Singh, Ashutosh Kumar, and Deepika Saxena. "A cryptography and machine learning based authentication for secure data-sharing in federated cloud services environment." Journal of Applied Security Research 17.3 (2022): 385-412.

[31] Rathore, H.; Mohamed, A.; Guizani, M. A Survey of Blockchain Enabled Cyber-Physical Systems. Sensors 2020, 20, 282. https://doi.org/10.3390/s20010282

[32] B. B. Shrestha and K. Yoo. (2020). An Improved Unsupervised Machine Learning Algorithm for Botnet Detection in Social Media Networks. Journal of Computational Science, 44, 101165.

[33] K. Torkilsheyggi, M. Jensen, and C. B. Nielsen. (2018). Detecting abnormal user behavior in social media: An unsupervised learning approach. Journal of Computational Science, 28, 207-215

[34] S. S. Vakilinia and R. N. Khushaba, "A Reinforcement Learning-Based Approach for Identifying Fake News on Social Media," in 2020 IEEE Symposium Series on Computational Intelligence (SSCI), 2020, pp. 3258–3265.

[35] M. Turk, E. N. Ayday and A. Almog, "Detection of Profile Cloning in Social Networks," in IEEE Transactions on Dependable and Secure Computing, vol. 3, no. 4, pp. 301-311, Oct.-Dec. 2006, doi: 10.1109/TDSC.2006.52.

[36] R. Yan, Z. Shang, Q. Sun, W. Wang, and Y. Zhou, "An Ensemble Learning Framework for Identity Theft Detection in Social Networks," Security and Communication Networks, vol. 2020, Article ID 5874935, 14 pages, 2020, doi: 10.1155/2020/5874935.

[37] P. Hui, M. Li, J. Pan and J. Huang, "Detecting Identity Fraud in Social Networks with Improved Deep Forest," 2020 IEEE International Conference on Computational Science and Engineering (CSE), Beijing, China, 2020, pp. 189-194, doi: 10.1109/CSE49834.2020.00039.

[38] L. Nigusse and B. Ramadhan, "Detecting Social Engineering Attacks in Social Networks using Graph Neural Networks," 2021 IEEE 18th Annual Consumer Communications Networking Conference (CCNC), Las Vegas, NV, USA, 2021, pp. 1-7, doi: 10.1109/CCNC42901.2021.9369455.

[39] Shi, X., Yang, C., Zhang, J., Zhao, X. (2021). Deep learning-based online fraud detection for mobile payment. IEEE Access, 9, 25908-25920.

[40] Wang, J., Zhang, M., Han, S. (2021). A hybrid ensemble approach for fraud detection in mobile payments. Journal of Ambient Intelligence and Humanized Computing, 12(1), 1057-1066.

[41] Prabakaran, M.K., Meenakshi Sundaram, P., Chandrasekar, A.D.: An enhanced deep learning-based phishing detection mechanism to effectively identify malicious URLs using variational autoencoders. IET Inf. Secur. 1– 18 (2023). https://doi.org/10.1049/ise2.12106

[42] Yang, P., Zhao, G., Zeng, P.: Phishing website detection based on multidimensional features driven by deep learning. IEEE Access 7, 15196–15209 (2019). https://doi.org/10.1109/ACCESS.2019.2892066

[43] Bu, S.J., Cho, S.B.: Deep character-level anomaly detection based on a convolutional autoencoder for zero-day phishing url detection. Electronics 10(12), 1492 (2021). doi.org/10.3390/electronics10121492

[44] Asha S, Shanmugapriya D, Padmavathi G, Malicious insider threat detection using variation of sampling methods for anomaly detection in cloud environment, Computers and Electrical Engineering, Volume 105, 2023, 108519, ISSN 0045-7906

[45] İ. Erdoğan, O. Kurto, A. Kurt and Ş. Bahtıyar, "A New Approach for Fraud Detection with Artificial Intelligence," 2020 28th Signal Processing and Communications Applications Conference (SIU), Gaziantep, Turkey, 2020, pp. 1-4, doi: 10.1109/SIU49456.2020.9302374.

[46] Catal, C., Giray, G., Tekinerdogan, B. et al. Applications of deep learning for phishing detection: a systematic literature review. Knowl Inf Syst 64, 1457–1500 (2022).