B. Tech Major Project Work Report

On

Synthetic Identity and Synthetic Identity Fraud Detection Using Federated Learning

Submitted by

Ritik Kumar, Suresh Kamediya, Pooja G

(191CS150, 191CS158, 191CS233)

Under the Guidance of

Dr. Mahendra Pratap Singh

Asst. Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA,

SURATHKAL, MANGALORE - 575025

April, 2023

DECLARATION

We hereby declare that the U.G **Major Project Work** Report entitled **Synthetic Identity and Synthetic Identity Fraud Detection Using Federated Learning**, which is being submitted to the National Institute of Technology Karnataka, Surathkal, in partial fulfillment of the requirements for the award of the Degree of **Bachelor of Technology in Computer Science and Engineering** in the department of **Computer Science and Engineering** is a bonafide report of the work carried out by us. The material contained in this Report has not been submitted to any University or Institution for the award of any degree.

Ritik Kumar (191CS150)
Suresh Kamediya (191CS158)
Pooja G (191CS233)
Dept of Computer Science and Engineering
NITK, Surathkal

Place: NITK, Surathkal
Date: Monday 24th April, 2023

CERTIFICATE

This is to certify that the U.G. **Major Project Work** Report entitled **Synthetic Identity and Synthetic Identity Fraud Detection Using Federated Learning** submitted by

1. **Ritik Kumar** (191CS150)
2. **Suresh Kamediya** (191CS158)
3. **Pooja G** (191CS233)

as the record of the work carried out by them, is accepted as the U.G. Major Project work report submission in partial fulfillment of the requirements for the award of degree **Bachelor of Technology in Computer Science and Engineering** in the Department of **Computer Science and Engineering**, NITK Surathkal.

Guide
Dr. Mahendra Pratap Singh
Department of Computer Science and Engineering
NITK, Surathkal

Chairman - DUGC
Dr. Manu Basavaraju
Department of Computer Science and Engineering
NITK, Surathkal

ACKNOWLEDGEMENTS

First and foremost, We would like to express our deep sense of gratitude to our research guide **Dr. Mahendra Pratap Singh**, *Assistant Professor, Department of Computer Science and Engineering*, *NITK, Surathkal* for his constant guidance, support, and motivation during our research work. His meticulous and planned attitude always encouraged us to accomplish all of our tasks on time, including the research. Throughout the project, he provided excellent suggestions and feedback at the appropriate times, which considerably benefited the research process's seamless functioning. Motivation, gratitude, moral support, and the flexibility provided by the course all contributed to the successful completion of our research.

We must also thank **Dr. Manu Basavaraju** (HOD), faculty members, and the staff of the *Department of Computer Science and Engineering*, *NITK, Surathkal* for their kind help and support.

Place: Surathkal, India
Date: Monday 24th April, 2023.

Suresh Kamediya
Ritik Kumar
Pooja G

Abstract

Although identity theft has been an issue for many years, the strategies of its perpetrators have developed in recent years. In the past, actual identity fraud was a prevalent practice among criminals. Financial institutions, corporations, and individuals are all impacted by synthetic identity fraud. As a result, to protect themselves and their clients against synthetic identity theft, many financial institutions and businesses are investing in advanced fraud detection technology and implementing fraud prevention methods. In contrast, criminals have also moved onto a harder-to-detect scheme known as synthetic identity theft. One of the reasons synthetic identity fraud is so complex to detect is that the identity does not exist in any public records, making it challenging to verify its legitimacy. To crack down on this kind of activity, every customer requesting credit must undergo even more stringent ID checks than they do.

This report presents a unique method for analyzing application information and determining if they are synthetic and have been used to commit fraud. We generate Synthetic Identity Dataset and use SMOTE to deal with the imbalanced dataset. Since two parties can't share their information, we propose a Federated Learning-based Machine Learning model to deal with the data privacy issue. Our model helps identify whether the applicant's identity is synthetic. The model's performance demonstrates that the proposed technique can be a feasible solution to the synthetic identity challenge while preserving user data privacy.

Keywords: Synthetic Identity, Synthetic Identity Fraud, Credit Card Application Fraud, Machine Learning, Federated Leraning

Contents

LIST OF ABBREVIATIONS

SI	Synthetic Identity
SID	Synthetic Identity Detection
SIF	Synthetic Identity Fraud
PII	Personally Identifiable Information
Logistic	Logistic Regression
DT	Decision Tree
RF	Random Forest
ReLU	Rectified Linear Unit ANN
Artificial Neural Network	
OTP	One Time Password
SSN	Social Security Number
DOB	Date Of Birth
FE	Feature Extraction
TF-IDF	Term Frequency-Inverse Document Frequency
Model	Federated Learning Model
GB	Giga Bytes
CPU	Central Processing Unit
RAM	Random Access Memory
GPU	Graphics Processing Unit
TP	True Positives
TN	True Negatives
FP	False Positives
FN	False Negatives
ML	Machine Learning
DL	Deep Learning
OSINT	Open Source Intelligence Tools
FL	Federated Learning

List of Figures

List of Tables

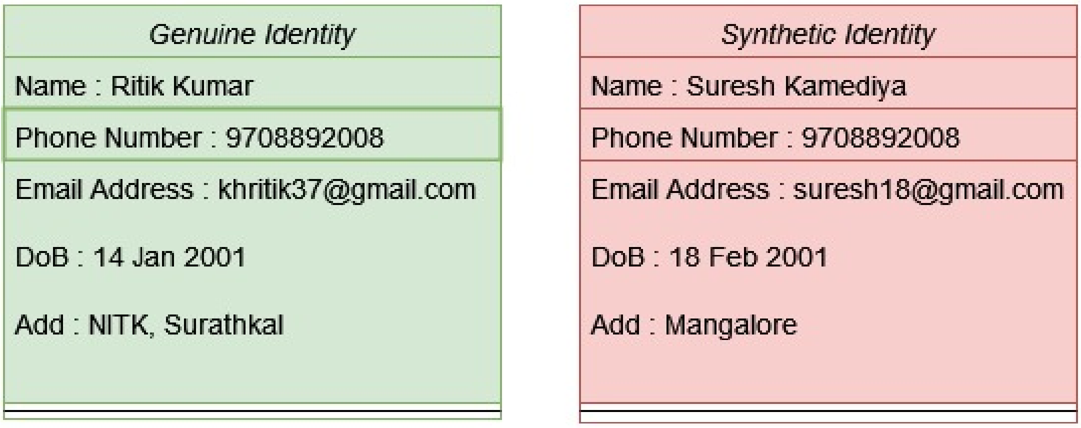
Chapter 1

Introduction

Financial institutions, corporations, and individuals are increasingly vulnerable to synthetic identity fraud. It entails constructing fictitious identities by mixing actual and fictitious personal information to open fraudulent accounts and gain credit or loans. Because the identities seem real and have no history of fraud, traditional fraud detection procedures are frequently unsuccessful against synthetic identity fraud. To fight this issue, we have used machine learning algorithms and federated learning approaches to identify and prevent synthetic identity fraud. We utilized data from numerous sources for training models that detect trends and anomalies suggestive of synthetic identities in this technique. Federated learning enables the safe and decentralized training of these models using data from numerous sources without jeopardizing individual data privacy. Combining machine learning and federated learning to identify synthetic identity fraud is a promising strategy that can drastically minimize fraud losses while protecting customers' financial data. There are two types of identity fraud; one is Traditional Identity Fraud, and the other is Synthetic Identity Fraud. Traditional Identity Fraud is explained as follows:

1.1 Traditional Identity Fraud

Traditional identity fraud is taking another person's personal information, such as their name, social security number, date of birth, and financial information, and then using it to open fraudulent accounts or perform unauthorized activities. It can occur through various means, including phishing, skimming, and hacking. Once the fraudster gets the victim's information, they can use it to make purchases, create



Genuine Identity
Name : Ritik Kumar
Phone Number : 9708892008
Email Address : khritik37@gmail.com
DoB : 14 Jan 2001
Add : NITK, Surathkal

Synthetic Identity
Name : Suresh Kamediya
Phone Number : 9708892008
Email Address : suresh18@gmail.com
DoB : 18 Feb 2001
Add : Mangalore

Figure 1.1: Traditional Identity Fraud

credit accounts, or secure loans in the victim's name, leaving them with financial obligations and credit score harm. Financial institutions employ two-factor or OTP-based authentication to determine if the person using the service is a genuine person or a fraudster.

1.2 Identity Theft Types

Fraudsters begin by gathering all of the information required to commit fraud. Once the data is collected, they attempt to build Synthetic Identities by combining numerous people's details into a single application's details. The various types of Identity Theft are as follows:

- **Financial Identity Theft:** Financial identity theft is when someone's financial information, such as bank account numbers, credit card information, or other sensitive financial data, is stolen. We can access this information through different means, including phishing schemes, data breaches, and the physical theft of financial papers. Once a fraudster obtains this information, they can use it to conduct unauthorized activities, create false accounts, or acquire loans and credit in the victim's name. Victims of financial identity theft may suffer substantial financial losses, lost credit ratings, and time-consuming efforts to recover their financial image.

- **Criminal Identity Theft:** When incarcerated by police for a crime, the offender, in this case, assumes the persona of a victim. We may confirm the perpetrator's claimed identity with government-issued identification documents obtained with credentials stolen from the victim. The conclusion is that the true persona of the perpetrator is unknown, and charges may eventually be filed against the victim rather than the criminal. Victims in these theft cases are typically ignorant that crimes have been committed in their names and are thus likely to be held guilty for any criminal activity.

- **Synthetic Identity Theft:** Synthetic identity theft is a sort of identity theft in which fictitious identities are created by mixing actual and false personal information to open fraudulent accounts or gain credit. The fraudster may construct a Synthetic Identity that seems authentic and has no history of deception by combining actual and fraudulent information. Synthetic identity theft may be difficult to detect, and typical fraud detection procedures are frequently useless. The implications for victims can be severe, including cash losses and credit score harm. Preventive steps such as checking credit reports and being wary of unusual behavior can help lessen the danger of Synthetic Identity theft.

- **Identity Cloning:**Identity cloning is a sort of identity theft in which we create a complete replica of someone's identity, including personal information, identification cards, and bank accounts. The fraudster may employ a variety of ways, such as stealing papers or hacking into accounts to access the victim's information. The fraudster can use a cloned identity to engage in illicit activities such as creating false accounts, seeking loans, and committing crimes. Identity cloning may be challenging to detect and can severely affect victims, including money losses, credit score damage, and legal issues. We can reduce identity cloning by taking preventive actions such as preserving personal information and being alert to questionable activities.

- **Child Identity Theft:** When attempting to gain credit, government assistance, or work, we take a child's personal information, such as their Social Security number. Identity theft can be challenging to identify and have long-term ramifications for the child's credit history and financial well-being. Preventive actions such as reviewing credit reports regularly and being wary of abnormal activity can help lessen the danger of child identity theft.

1.3 Synthetic Identities Creation

Synthetic identity is created by mixing actual and fictitious personal information, such as a valid Social Security number with a fictitious name or location. The fraudster may utilize multiple variants of personal information, such as a bogus date of birth or a different name spelling, to make the Synthetic Identity look real.

Financial institutions face several threats as a result of Synthetic Identity fraud, including:

- **Financial losses:** Financial institutions can suffer considerable financial losses due to Synthetic Identity theft, including losses from fraudulent loans, credit cards, and other financial goods.
- **Reputational harm:** Synthetic identity fraud may harm financial firms' reputations, which results in a loss of client trust and confidence.
- **Compliance risks:** Synthetic identity fraud may lead to compliance issues, such as breaches of rules and laws, which can lead to penalties and fines.
- **Operational expenses:** Synthetic identity fraud can raise financial institutions' operational costs, including fraud prevention, investigation, and remediation.
- **Legal hazards:** Synthetic identity fraud can result in legal issues such as customer lawsuits and regulatory proceedings.

1.3.1 Reasons for Synthetic Identity Frauds Popularity

Companies have difficulty adapting to Synthetic Identity Fraud, one of the fastest-growing scams. The following factors contribute to the popularity of Synthetic Identity fraud:

- The rising complexity of the dark web is the primary force behind easy access to stolen personally identifiable information from data breaches. The dark web has facilitated the selling and distribution of forged identities.

- Identity Synthesis Fraud systems can imitate "excellent" consumer behavior. According to a Federal Reserve analysis, 70% of suspected incidents exhibit traditional consumer habits for a brief period.

- Because it is challenging to detect Synthetic Identity Fraud, many firms write off cases as bad debt. This allows the fraudster to avoid prosecution. Furthermore, it makes it difficult to understand the scale of the problem, as well as the fraud's behaviors and telltale signs.

- Consumer inclination for faster and easier account sign-up procedures is another element aiding the growth of Synthetic Identity Fraud. Due to a decrease in in-person account installations and more convenient online sign-up, SIF usage has grown easier for fraudsters.

- The fact that Synthetic Identity Fraud usually goes undetected for months adds to the terror because the personal identification information exploited is typically simply one piece of information, not the complete person; an individual's discovery of the theft is frequently delayed or does not occur.

- Increased technology has aided crooks' ability to evolve and become more successful. The approach allows fraudsters to swiftly scale up for many synthetic Identity Fraud attempts, once we identify the victim.

- Cybercriminals utilize cutting-edge technologies to generate fresh scam versions to avoid detection.

- The increasing quantity and scale of data breaches are significant contributors to the emergence of Synthetic Identity Fraud.

1.3.2 Synthetic Identity Fraud Methodology

Synthetic identity fraud is frequently committed through a complex web of techniques, such as identity theft, social engineering, and manipulation of credit bureaus and financial organizations. The steps involved in it are as follows:

- Creating a Synthetic Identity by mixing actual and fictitious personal information, such as a valid Social Security number with a phony name and date of birth.

- Building the credit history of the Synthetic Identity by applying for credit products and making regular payments using the new identity.
- Making transactions or obtaining cash using the established credit.
- Abandoning the Synthetic Identity once it has been completely exploited or when there are indications of inquiry.

1.4 Problem Description

Once Synthetic Identities are complete, fraudsters apply to financial institutions. Using the fabricated persona, the fraudster applies for credit products like credit cards and loans. After obtaining credit products, the fraudster attempts to establish a favorable credit history for the Synthetic Identity, which may subsequently be used to qualify for larger loans or credit products. Once the phony persona has a solid credit history, the fraudster makes substantial purchases and flees. Due to a lack of third-party data or an effective way to validate whether the submitted application information is legitimate or fraudulent, financial institutions must undertake more strict screening to identify synthetic Identities and onboard new clients.

An approach for recognizing synthetic identities that employ third-party data can be useful. It can be based on the idea that individuals have histories left behind in various physical and digital data systems. Real people provide a consistent trail: the same address, email address, and phone number exist in many databases. These paths are difficult to replicate. They have access to massive volumes of data dating back years. Genuine ID may be too constant, whereas Synthetic Identity is fabricated. As a result, we choose to tackle this issue by utilizing our dataset, Machine Learning algorithms, and a Federates Learning model.

1.5 Motivation

Several companies have developed systems for Synthetic Identity Fraud Detection. Experian provides a portfolio of technologies, including advanced analytics, Machine Learning, and behavioral biometrics, to assist in detecting Synthetic Identity Fraud.

TransUnion offers multiple Synthetic Identity Fraud detection solutions, including identity verification, device identification, and Machine Learning-based fraud detection. FICO provides a Synthetic Identity Fraud detection service that employs Machine Learning to detect patterns of fraud and risk in credit applications. ID Analytics provides a Synthetic Identity Fraud detection system that uses a combination of behavioral analytics, Machine Learning, and network analytics to detect fraudulent behavior. But these techniques are insufficient to identify Synthetic Identity Fraud since the fraudster's method of committing Synthetic Identity Fraud changes daily. To protect it, we need a multi-pronged strategy. We attempt to offer a model that can identify Synthetic Identities and Synthetic Identity fraud in this study, and this technique may be useful for future Synthetic Identities detection.

1.6 Objectives

The objectives of our work are as follows:

- Generation of a Synthetic Identity Dataset using Open Source Intelligence Tools (OSINT) and Faker library.
- Design and implement a Federated Learning-based model using a Machine Learning algorithm to detect Synthetic Identities and Synthetic Identity Fraud.
- Experimental analysis of the proposed model using Synthetic Identity Dataset.

1.7 Organization of the Report

The remaining part of the report is divided into four chapters. The second chapter, Literature Review, examines the difficulties of synthetic identities and Synthetic Identity fraud. The third chapter, Proposed Synthetic Identity Fraud Detection Model, goes into further depth on the created Synthetic Identity and Synthetic Identity Fraud Detection. The fourth chapter, Experimental Results and Analysis, describes the environment setup and the comparison of various ML and FL models' accuracies. Conclusion and Future Work, the last chapter, summarises the proposed model and provides future research prospects.

Chapter 2

Literature Review

This chapter covers the research that went into building the Synthetic Identity Detection and Synthetic Identity Fraud Detection frameworks, as well as the contributions of numerous researchers.

2.1 Web Scraping

Web scraping, the act of mechanically gathering data from websites, has been an increasingly popular topic in recent years as the internet has grown and the amount of online data has increased. This literature review explores some of the research publications on web scraping.

Web scraping can simulate how people explore the internet using a full-featured web browser, such as Google Chrome or the low-level Hypertext Transfer Protocol (HTTP). In other words, online scraping allows computer programs to retrieve data more correctly and quickly than people. In contrast, it is to copy and paste information from the internet into a spreadsheet.

Understanding the legal and ethical ramifications of web scraping is one of the most significant parts of this practice. Krotov et al. [18] examine the legal and ethical implications of online scraping. While online scraping may be helpful for research and analysis, it can also violate intellectual property rights and privacy rules. The authors advise online scrapers to acquire website owners' permission and be open about their goals.

The technological problems in gathering data from websites are another essential part of web scraping. Singrodia et al. [26] evaluate the many approaches used for online scraping and highlight the issues involved with each method. They also advocate recommended practices for web scraping, such as utilizing a reliable web scraping application and adhering to website terms of service. In addition to technological difficulties, online scraping can be hampered by anti-scraping efforts implemented by website owners.

Pedro Marques et al. [23] review the strategies used by website owners to detect and prevent web scraping. They also discuss the efficacy of these strategies and offer tips for web scrapers to avoid discovery. Natural language processing (NLP) is one area where web scraping has shown significant potential. Stephanie Lunn et al. [22] analyze the various NLP approaches used for online scraping and present examples of how these techniques have been applied in real-world applications such as sentiment analysis and topic modeling.

In [16], different phases of web scraping are described as follows:

- **Fetching Phase:** The chosen website containing the relevant information must be accessed during the initial fetching step. The HTTP protocol is used for this.
- **Extraction Phase**: The critical information should be extracted after retrieving the HTML page. Regular expressions, HTML parsing libraries, and XPath queries are employed during the extraction step.
- **Transformation Phase**: Now that only relevant information remains, it may be organized into a structured manner for presentation or storage. We can extract information from stored data to help business intelligence make better judgments.

2.2 Synthetic Identity Fraud

Synthetic identity fraud has been a growing financial crime in recent years. Constructing a new identity by merging several bits of personal information from different persons is known as Synthetic Identity Fraud. This fictitious persona is then used to create fraudulent accounts or gain credit. This literature review will look at current survey publications on Synthetic Identity Fraud.

Bojilov et al. [5] look in-depth into Synthetic Identity Fraud, including definitions, kinds, and detection strategies. The lack of a unique identifier for individuals and the usage of complicated fraud rings are among the issues identified in the article in identifying and combating Synthetic Identity theft. The authors also highlight the significance of Machine Learning in determining Synthetic Identity Fraud and propose future research prospects.

Abdulalem Ali et al. [2] focus on utilizing Machine Learning to detect Synthetic Identity theft. The authors evaluate existing Machine Learning algorithms for Synthetic Identity Fraud detection and present a benchmarking framework for assessing their effectiveness. The research also cites many difficulties in detecting Synthetic Identity theft, such as a lack of labeled data and the dynamic nature of fraud trends.

Enriquez et al. [10] include an overview of Synthetic Identity Fraud and an examination of the difficulties in identifying and combating this fraud. The authors present a system for identifying Synthetic Identity theft that integrates Machine Learning approaches with data from various sources such as credit bureaus, public records, and social media. They also talk about how important it is for financial institutions and government organizations to work together to avoid and combat Synthetic Identity Fraud.

Walker-Moree et al. [29] examine Synthetic Identity Fraud in-depth, including the many ways to generate synthetic identities, the strategies used to identify and prevent this fraud, and the obstacles associated with countering Synthetic Identity theft. The authors also emphasize the need for coordination among financial institutions, law enforcement, and regulators to prevent and mitigate the consequences of Synthetic Identity theft.

2.3 Feature Extraction (TF-IDF)

Document clustering is an unsupervised Machine Learning technique that groups similar documents into clusters based on their content. One of the critical steps in document clustering is feature extraction, which involves selecting and transforming the raw text data into a set of meaningful features that can be used to represent the documents. TF-IDF stands for Term Frequency-Inverse Document Frequency, a common feature extraction technique used in document clustering.

Prafulla Bafna et al. [4] present a document clustering algorithm using the feature extraction term frequency-inverse document frequency (TF-IDF) approach. The approach involves transforming raw text data into a set of meaningful features that represent the documents. The paper demonstrates the effectiveness of the TF-IDF approach for document clustering and compares it to other feature selection methods. The proposed approach outperforms different clustering algorithms on benchmark datasets.

Shereen Albitar et al. [1] have proposed an efficient text clustering algorithm using cosine similarity based on the TF-IDF approach. The authors first preprocess the text data by removing stop words and stemming. They then use the TF-IDF approach to transform the text data into meaningful features. Finally, they use cosine similarity to cluster the documents based on their feature vectors. The authors demonstrate that their approach outperforms other clustering algorithms on several benchmark datasets.

In research [19], the TF-IDF characteristics and the fuzzy K-means method are used to propose a text document clustering technique. Kumbhar, Rutuja; Mhamane, Sneha et al. preprocessed the text data by eliminating stop words and stemming. They then employ the TF-IDF method to convert the text data into a collection of relevant characteristics. Finally, they cluster the papers based on their feature vectors using the fuzzy K-means technique. On numerous benchmark datasets, the authors show that their technique outperforms alternative clustering algorithms.

2.4 Feature Selection

Identifying and picking the most relevant characteristics from the given dataset is an important stage in Machine Learning. Several feature selection strategies have been developed in recent research publications to increase the performance of Machine Learning systems.

Asir A. Danasingh et al. [7] present a complete overview of feature selection approaches for high-dimensional data. The study examines several feature selection approaches, such as filtering, wrapping, and embedding. In addition, the authors examine the difficulties of feature selection in high-dimensional data and suggest future study paths.

Samina Khalid et al. [15] review feature selection strategies in Machine Learning. The study examines several feature selection strategies, such as filtering, wrapping, and embedding. The authors also present a comparative review of the various approaches, highlighting their benefits and drawbacks.

Beatriz Remeseiro et al. evaluate several feature selection approaches in medical diagnosis in [24]. The study examines the difficulties of medical diagnosis and the necessity of feature selection in increasing diagnostic accuracy and reliability. The authors also compare the various feature selection approaches and emphasize their efficiency in medical diagnosis.

Filter methods employ statistical approaches to assess and prioritize the relevance of characteristics to the target variable. These approaches are computationally efficient but need to consider feature interaction.

Wrapper approaches evaluate the model's correctness and pick the features that deliver the greatest performance. These approaches are computationally demanding, yet they can capture feature interaction.

Embedded approaches integrate the feature selection and model training steps. These approaches concurrently optimize feature selection and model performance, resulting in more efficient and accurate models.

Effective feature selection can increase model accuracy, and generalization performance, decrease training time, and improve model interpretability.

2.5 Over Sampling the Data (SMOTE)

Oversampling is a common strategy in unbalanced classification issues where the data is heavily biased towards one class. SMOTE (Synthetic Minority Over-sampling approach) is a well-known oversampling approach for dealing with unbalanced classification difficulties. In unbalanced datasets, one class may contain fewer samples than the other, causing typical Machine Learning techniques to perform poorly. SMOTE produces synthetic samples of the minority class to balance the dataset and improve the model's performance.

In [9], a complete description and evaluation of several SMOTE variations and their efficiency in boosting classification performance is given. In [25], the paper presents an ensemble learning strategy combining SMOTE with boosting approaches to improve performance on unbalanced datasets.

Dablain et al. [6] provide a deep learning-based oversampling strategy that outperforms existing ones in unbalanced classification situations. Dong et al. [8] provide an adaptive SMOTE algorithm that dynamically modifies the oversampling rate dependent on the degree of imbalance in the dataset.

Overall, current research publications thoroughly review SMOTE and its modifications and assess its efficiency in boosting classification performance on unbalanced datasets. They suggest numerous SMOTE upgrades and hybrid techniques for improved performance in diverse applications.

2.6 Machine Learning

Machine Learning is a technique for training computers to make predictions or decisions based on patterns in large sets of data without explicit programming for every possible scenario.

TP Latchoumi et al. [20] provide a unique framework for identifying SIF using Machine Learning approaches. The authors employ supervised and unsupervised learning approaches, such as logistic regression, random forests, and k-means clustering. According to the results, the proposed framework outperforms existing SIF detection approaches.

Abdulalem Ali et al. [2] describe a Machine Learning framework for SIF identification that employs supervised and unsupervised learning approaches, such as support vector machines, k-nearest neighbors, and hierarchical clustering. The authors demonstrate that their system can detect SIF with high accuracy.

Rahul Goyal et al. [11] provide a Machine Learning strategy for SIF identification that incorporates supervised and unsupervised learning techniques such as logistic regression, decision trees, and k-means clustering. The authors demonstrate that their technique can identify SIF with excellent accuracy.

Dejan Varmedja et al. [28] provide a deep learning-based SIF identification system based on a convolutional neural network (CNN). The authors demonstrate that their approach outperforms standard Machine Learning-based SIF identification methods. S. Sivanantham et al. [27] have described a hybrid SIF identification approach that employs Machine Learning and rules-based techniques. The authors demonstrate that their technique can identify SIF with excellent accuracy.

2.7 Federated Learning

Federated learning (FL) is a distributed approach for training Machine Learning models from disparate and isolated data.

Chaoyang He et al. [12] say three main differences exist between FL and data center-based distributed training: statistical heterogeneity, system constraints, and trustworthiness. Federated learning has become a multidisciplinary research topic due to the contributions required from various domains, including Machine Learning, wireless communication, mobile computing, distributed systems, and information security.

Shaoxiong Ji et al. [14] say that Federated learning safeguards data privacy by learning a shared model through distributed training on local client devices rather than gathering data on a central server. Distributed intelligent agents use a common global model from the central server's parameters as initialization to train their private models using personal data and make predictions on their own physical devices.

In the real world, Federated Learning has various applications, such as predicting which photographs a mobile user is likely to post on social websites [17], predicting the next word for mobile keyboards [3], retrieving important alerts, and identifying spam messages [21].

Kai Hu et al. [13] show that FL differs from distributed ML in that the information sent to the server by each participant is no longer the raw data but a trained submodel. At the same time, the FL supports asynchronous transmission, allowing communication that needs to be decreased suitably. Vertical Federated Learning focuses on data samples from each client in FL with overlapping training data. The data samples are the same among individuals, but the data attributes are not. Federated transfer learning is employed when each participant's training data attributes and data sample overlap are minor. Federated transfer learning is classified into case-based, feature-based, and model-based.

2.8 Summary

According to the research papers in the literature review, various Machine Learning techniques are helpful for Synthetic Identity Fraud Detection. But using only Machine-Learning techniques may intrude on the privacy of individuals and not safeguard them. To maintain and safeguard the privacy of individuals, our report mainly focuses on Synthetic Identity Fraud Detection using Federated Learning. We discuss our proposed architecture in the next chapter, which is based on Federated Learning to tackle the issue of Synthetic Identity Fraud.

Chapter 3

Proposed Methodology

This chapter presents a federated learning-based framework that uses an artificial neural network model and comprises three sections. The first section describes how the Synthetic Dataset is created and how Synthetic Identities are generated from the dataset. The second section describes the data preprocessing method, whereas the third section describes various ML models and a federated learning-based model.

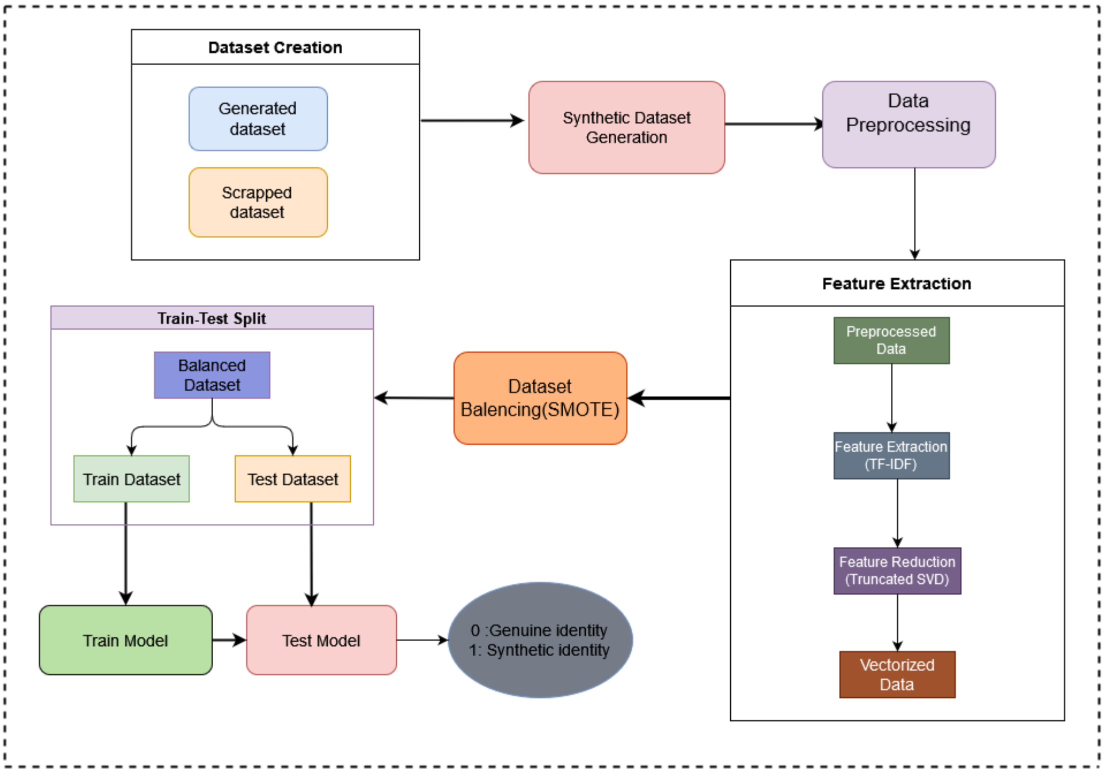


Figure 3.1: Synthetic Identity and Synthetic Identity Fraud Detection Framework

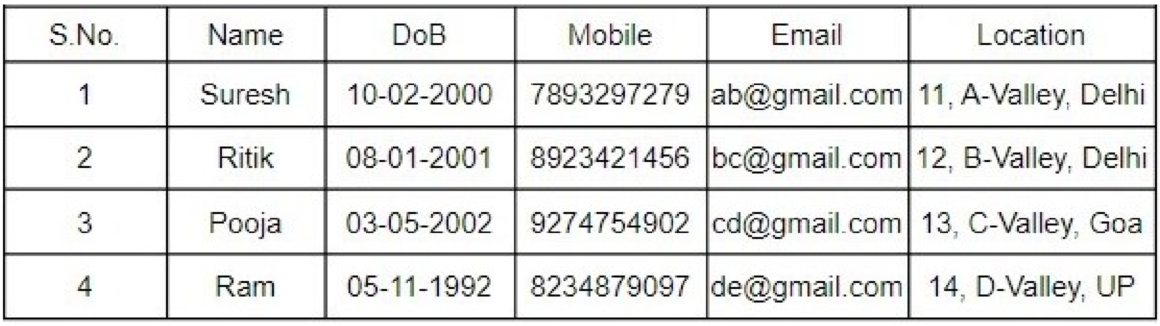
Fig 3.1 shows the framework of the proposed approach and contains the Dataset creation module, Synthetic identity generation module, Data Preprocessing module, Feature Extraction module, Data Balancing module, and Train-Test split module and the description of these modules are as follows:

3.1 Dataset Creation

The process of obtaining unprocessed data from multiple sources for analysis, research, or other reasons is called dataset creation. Our research collected data from the LinkedIn platform using various OSINT techniques and python modules. We have two kinds of datasets i.e original dataset and the synthetic dataset, by which we have generated synthetic identities.

3.1.1 Original Dataset

In this dataset, we have considered a person's full identification consisting of their name, email, phone number, date of birth, and address. For the initial dataset to contain only valid data points, we scraped information from LinkedIn. We used selenium, Beautiful Soup, and other Python modules to scrap each profile's contents. About 900 full LinkedIn profiles have been fetched. A sample of our initial database is shown in Fig 3.2.



S.No.	Name	DoB	Mobile	Email	Location
1	Suresh	10-02-2000	7893297279	ab@gmail.com	11, A-Valley, Delhi
2	Ritik	08-01-2001	8923421456	bc@gmail.com	12, B-Valley, Delhi
3	Pooja	03-05-2002	9274754902	cd@gmail.com	13, C-Valley, Goa
4	Ram	05-11-1992	8234879097	de@gmail.com	14, D-Valley, UP

Figure 3.2: Original Dataset samples

3.1.2 Synthetic Dataset

Synthetic data production is the process of producing fake data that replicates real-world data, to maintain the statistical characteristics and patterns of the original data while preserving sensitive information. The collection of IDs was artificially created using the Faker Python tool. Faker creates false data for many data kinds, including names, addresses, phone numbers, email addresses, and more, using a set of established criteria.

The resulting data may be modified and localized using Faker. To make the data more resemblant to our initial dataset, which was scraped from an Indian's LinkedIn profile, we recreated it in the context of Indian profiles. In this manner, the dataset's characteristics are closer to the original dataset's. Using the Faker package, we created a dataset of 50,000 profiles that we utilized to train our Machine Learning models.

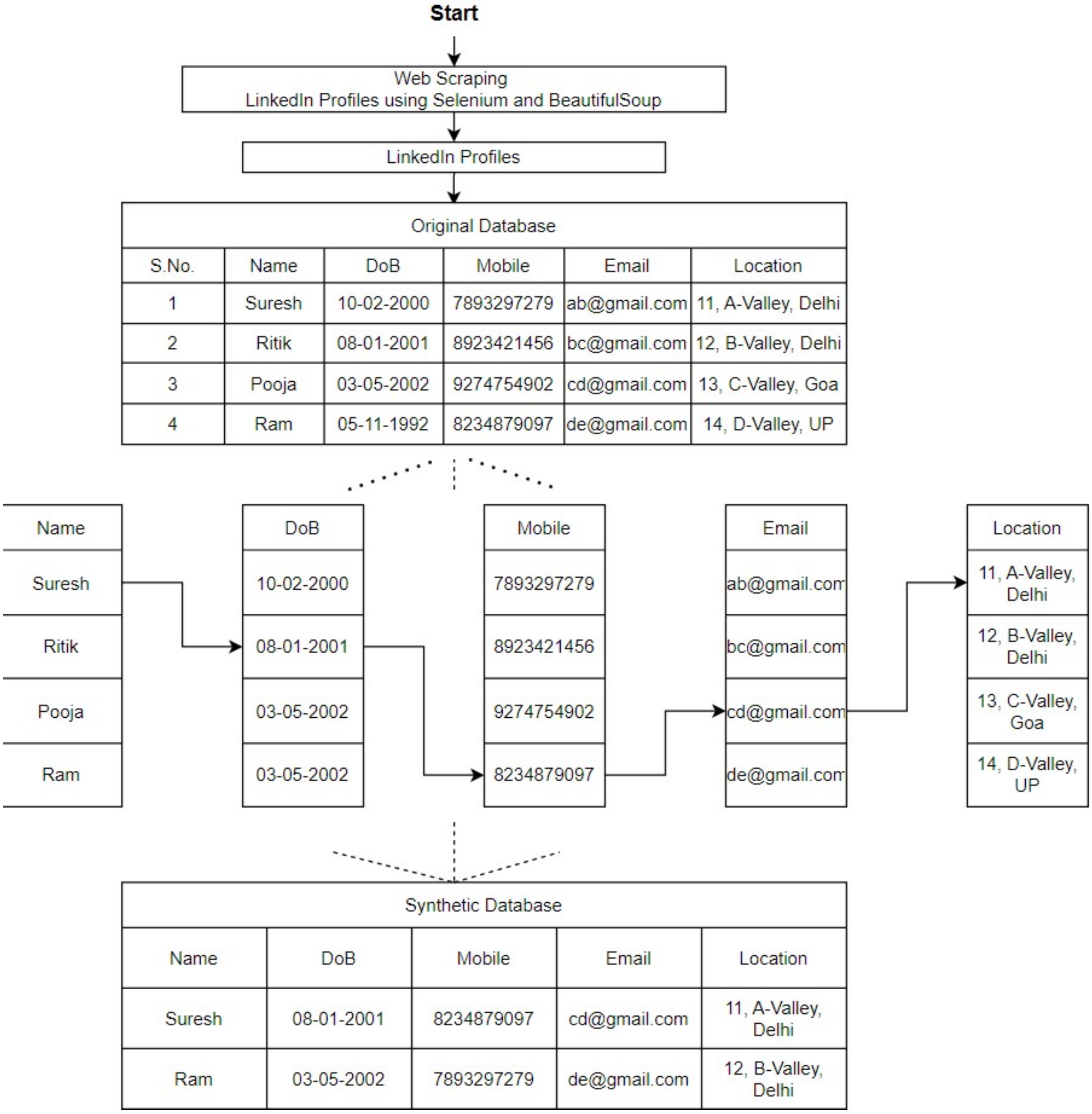
3.1.3 Synthetic Dataset generation

We can produce synthetic identities by combining several identities from the original or synthesized dataset, which does not correspond to any real identity. Figure 4.3 illustrates how several identities can be used to build synthetic identities. To create a new Synthetic Identity for each personally identifiable entity, an entry from the collected dataset is taken randomly and combined to create a new identity that does not belong to any real person. Synthetic identities are created to perpetrate frauds so in real-world situations, synthetic identities will be far less prevalent than genuine identities. Therefore, we have only produced 5% of synthetic identities using the whole dataset.

3.2 Data Preprocessing

Data preprocessing is preparing and cleaning raw data before we use it for analysis or modeling. This is an important step in Machine Learning, as preprocessing ensures the data is accurate, consistent, and ready for analysis.

The email and address of a profile may contain a few unusual characters; therefore, we utilized the NLTK library to remove them during data preparation. After removing special characters, we only kept capital and lowercase letters, numerals (a-zA-Z0-9), and spaces. We have utilized Python techniques to locate and eliminate all special characters to filter out such characters. Additionally, all strings were turned to lowercase. Fig. 4.4 shows a few examples of data before and after special characters removal.



Original Database

S.No.	Name	DoB	Mobile	Email	Location
1	Suresh	10-02-2000	7893297279	ab@gmail.com	11, A-Valley, Delhi
2	Ritik	08-01-2001	8923421456	bc@gmail.com	12, B-Valley, Delhi
3	Pooja	03-05-2002	9274754902	cd@gmail.com	13, C-Valley, Goa
4	Ram	05-11-1992	8234879097	de@gmail.com	14, D-Valley, UP

Name	DoB	Mobile	Email	Location
Suresh	10-02-2000	7893297279	ab@gmail.com	11, A-Valley, Delhi
Ritik	08-01-2001	8923421456	bc@gmail.com	12, B-Valley, Delhi
Pooja	03-05-2002	9274754902	cd@gmail.com	13, C-Valley, Goa
Ram	03-05-2002	8234879097	de@gmail.com	14, D-Valley, UP

Synthetic Database

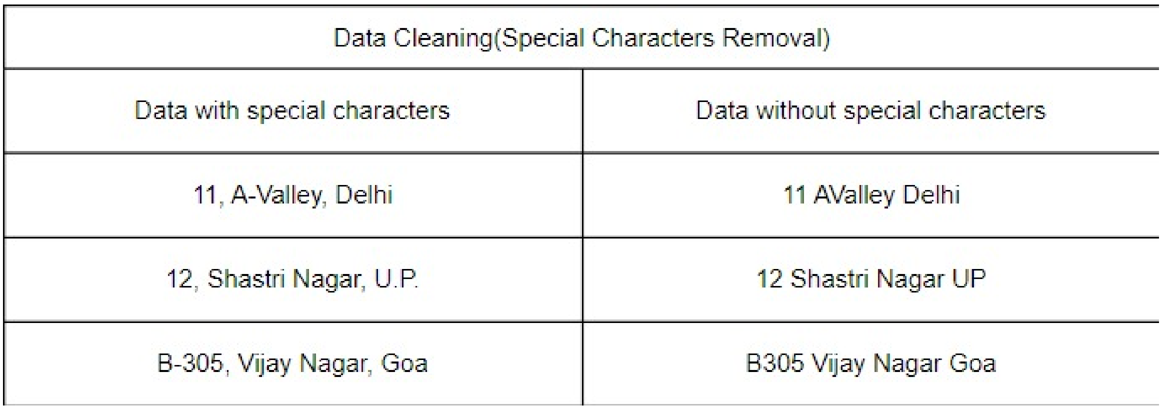
Name	DoB	Mobile	Email	Location
Suresh	08-01-2001	8234879097	cd@gmail.com	11, A-Valley, Delhi
Ram	03-05-2002	7893297279	de@gmail.com	12, B-Valley, Delhi

Figure 3.3: Synthetic Identity Generation Flow Chart

3.2.1 Feature Extraction

The process of selecting and transforming relevant information, or features, from raw data to make it more usable for a particular task or analysis is known as feature extraction. This procedure is crucial in the context of Machine Learning since it significantly impacts the algorithm's precision and efficiency.

We have used TF-IDF for feature extraction. An approach often used in natural language processing (NLP) for text analysis is TF-IDF (Term Frequency-Inverse Document Frequency). It assesses a term's (word or phrase's) significance inside a



Data Cleaning(Special Characters Removal)	
Data with special characters	Data without special characters
11, A-Valley, Delhi	11 AValley Delhi
12, Shastri Nagar, U.P.	12 Shastri Nagar UP
B-305, Vijay Nagar, Goa	B305 Vijay Nagar Goa

Figure 3.4: Data Cleaning

document or corpus of documents. The fundamental goal of TF-IDF is to give phrases greater weight when they often appear in one document but infrequently in another. This strategy is predicated on the idea that words that frequently appear in one document but infrequently in another are reliable indicators of the information contained in that document.

The TF-IDF score of a term is determined by multiplying two numbers, term frequency (TF) and inverse document frequency (IDF). The number of times a phrase appears in a document is known as frequency. The inverse document frequency, on the other hand, is calculated by dividing the total number of documents in a corpus by the number of documents that include the phrase. The formula for calculating the TF-IDF score is as follows:

$$TF - IDF = TF * IDF$$

where,

$$TF = \frac{(Number\ of\ times\ term\ t\ appears\ in\ a\ document)}{(Total\ number\ of\ terms\ in\ the\ document)}$$

and,

$$IDF = log_e \frac{(Total\ number\ of\ documents)}{(Number\ of\ documents\ with\ term\ t\ in\ it)}$$

If a term often appears in one document but not in other papers in the corpus, the phrase has a high TF-IDF score. The TF-IDF score indicates the term's importance to the document. TF-IDF is utilized in various NLP applications, including text mining, document categorization, and information retrieval. It may be used to find keywords in a document, compare papers, and even provide users recommendations for similar publications. Overall, TF-IDF is a powerful text analysis method that aids in extracting significant insights from huge amounts of text data.

3.2.2 Feature reduction

A collection of data that has been transformed into a numerical representation of vectors (arrays) that can be utilized for modeling and analysis is known as a vectorized dataset. The technique of reducing the number of features or variables in a dataset while keeping the most relevant information is known as feature reduction. This is crucial because datasets with many features can be computationally expensive and result in overfitting.

We obtain a vectorized dataset of 141142 dimensions after feature extraction. When the original dataset includes a lot of characteristics or dimensions, truncated SVD is necessary. The dataset may be noisy, redundant, or contain irrelevant characteristics under these circumstances, which can harm how well Machine Learning models perform. The dataset's dimensionality can be decreased while retaining as much of the original variance as feasible using truncated SVD. This might result in better model performance, quicker training times, and more effective resource usage. Additionally, latent semantic analysis (LSA), which is helpful in tasks like document categorization and information retrieval, may be carried out on text data using truncated SVD.

3.2.3 Truncated SVD

Singular Value Decomposition(SVD), is a mathematical technique for breaking down a matrix into its component elements to simplify calculations and minimize the complexity of data.

For dimensionality reduction in linear algebra and Machine Learning, truncated SVD is utilized. It is a modified form of SVD that truncates or decreases the dimensions of a certain dataset while retaining the majority of the original variance. In traditional SVD, the decomposition is applied to a matrix A, which is represented as a product of three matrices:

$$A = U\sigma Vt$$

U and V are orthogonal matrices, and σ is a diagonal matrix containing the singular values of A.

These singular values indicate the variance of the data along each of the principal axes. Truncated SVD works by only keeping the k largest singular values, where k is a chosen parameter. This results in a truncated version of the original decomposition:

$$A_k = U_k \sigma_k V_k^T$$

Where U_k and V_k is the truncated orthogonal matrices and σ_k is a truncated diagonal matrix containing the k largest singular values.

3.2.4 Balancing Dataset

Balancing a dataset refers to the process of adjusting the class distribution of a dataset so that the number of instances in each class is roughly the same. This is crucial because many Machine Learning methods are built to work effectively on datasets that are balanced, meaning that each class contains a similar number of examples.

On the other side, unbalanced datasets may lead to Machine Learning algorithms that are biased towards the majority class and perform poorly on the minority class. A Machine Learning method could, for instance, perform well on the majority class but badly on the minority class in a dataset where 90% of the examples belong to one class and just 10% to another.

Our dataset is very imbalanced as it contains 50,000 original data points and only 2500 synthetic data points. So in order to get proper evaluation metrics, we have used techniques like SMOTE and ADASYN to balance our database.

3.2.4.1 SMOTE

SMOTE stands for Synthetic Minority Over-sampling Technique. It is a data augmentation approach that uses artificially created samples for the minority class to balance out unbalanced datasets. SMOTE works by interpolating between already existing minority class samples to create new synthetic samples.

A minority class sample and its k nearest neighbors are chosen at random, and the primary idea behind SMOTE is to construct synthetic samples along the line segments that link each of the k nearest neighbors to the minority class sample. As a consequence, a new set of synthetic samples is created that is comparable to the minority class samples already in existence but differ from them in some way.

By using artificial samples, the SMOTE approach oversamples the minority class in a dataset. This strategy concentrates on the feature space and creates new cases by interpolating between the positive examples that are near together. The binary class distribution is often selected as 1:1, and N is set up as the overall number of oversampling observations. Starting with a randomly chosen member of the positive class, the iteration gets its KNNs (default is five). Then, N is randomly picked from these K occurrences to create fresh synthetic samples. The distance between the feature vector and its neighbors is calculated through interpolation using any distance measure. Then, this difference is multiplied by the starting feature vector and a random number between [0, 1]. A visual representation of this process is shown in Fig. 4.5.

We have applied SMOTE to the Synthetic Identity dataset, which contains 50,000 original data points and 2500 synthetic dataset data points. We got 1,00,000 data points as a result.

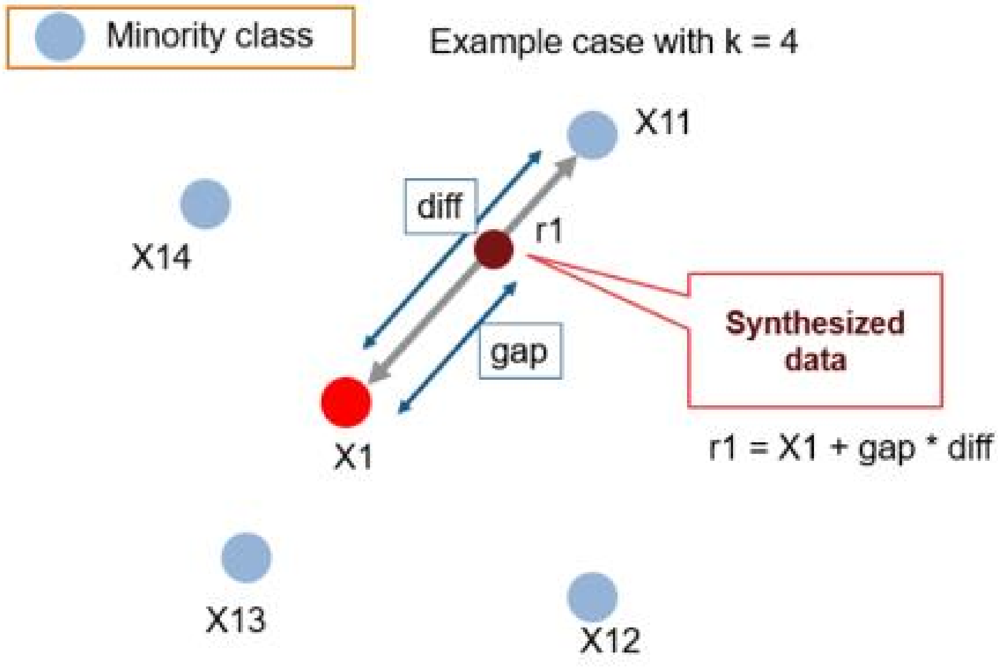


Figure 3.5: Data balancing using SMOTE

3.2.4.2 ADASYN

ADASYN stands for Adaptive Synthetic Sampling. Like SMOTE, it is a data augmentation technique that balances unbalanced datasets by producing artificial samples for the minority class. In contrast to SMOTE, ADASYN creates more synthetic samples for minority samples that are challenging and hard to learn.

To balance the class distribution, the ADASYN approach, which is used to manage unbalanced datasets in Machine Learning, generates synthetic instances of the minority class. In places where the class imbalance is more pronounced, ADASYN creates more synthetic instances adaptively than SMOTE.

The minority class examples' density distribution is calculated as part of the ADASYN process, after which synthetic examples are produced in low-density areas. This is accomplished by considering the degree of class inequality in the area while interpolating between minority class cases and their k-nearest neighbors.

When there are considerable fluctuations in the class imbalance across the feature space, ADASYN's adaptability is very helpful. As a result, it can provide more artificial situations in which the minority class is underrepresented and fewer in which the class is overrepresented.

When there is a high-class imbalance and high-dimensional features in the dataset, ADASYN has successfully enhanced classifier performance on unbalanced datasets.

3.2.5 Train-Test split

In Machine Learning, it is crucial to assess how well a model performs on data it has never seen before. To do this, the available data is often divided into training and testing sets. The testing set examines the model's performance and prevents overfitting the training data, while the training set is used to fit the model. The data has been divided into two halves using a 70:30 ratio, with 70% of the data used for training and 30% for testing.

3.3 Model

We have used two Machine Learning strategies for classification, as previously mentioned. The first strategy is regular learning, where we have evaluated the performance of several Machine Learning algorithms based on their accuracy. The second strategy is Federated Learning, in which an ANN model is used as the foundation for the Federated Learning architecture.

3.3.1 Regular ML algorithms

To see which model performs better in regular learning, we have used Logistic Regression, KNN, Random Forest, and ANN as Machine Learning models. A brief explanation of each model is given below.

3.3.1.1 Logistic Regression

Both classification and regression issues are addressed using logistic regression. The classification method of logistic regression is employed to forecast the likelihood of a categorical dependent variable. In a binary classification issue, a binary variable with data coded as 1 (Synthetic ID) or 0 (Genuine ID) is the dependent variable for the logistic regression procedure.

Discovering a correlation between traits and the likelihood of a particular result is the goal of logistic regression. Our project must determine if the provided dataset entries are Synthetic Identities or Genuine Identities based on our dataset. Two possible options for the answer variable are Synthetic Identity and Genuine Identity.

Similar to a linear regression model, logistic regression makes use of a cost function that is more complex, called the "sigmoid function" or "logistic function," rather than a linear function. We used logistic regression to train our Synthetic Identity balanced (SMOTE applied) and imbalanced data sets.

3.3.1.2 KNN

A straightforward but efficient Machine Learning approach called K-Nearest Neighbors (KNN) is utilized for classification and regression problems. The k closest training instances in the feature space are the input for a KNN, and the output is a prediction based on these k examples.

Euclidean distance, Manhattan distance, or other distance functions can be used as the distance metric to gauge how similar two data points are. There are several ways to find the ideal value of k, including cross-validation and employing a validation set.

Since KNN is a non-parametric algorithm, it makes no assumptions about the distribution of the data at its core. KNN may thus handle the intricate interactions between the characteristics and the output variable. It can, however, be computationally costly, particularly when working with huge datasets or high-dimensional feature spaces.

KNN has been demonstrated to perform effectively in several applications, including image recognition, text categorization, and recommendation systems, despite its simplicity.

3.3.1.3 Random Forest

Random Forest is employed for classification, regression, and other applications. It is an ensemble approach that creates several decision trees and then combines the output of each tree's analysis to produce predictions.

Supervised Machine Learning methods like random forest are widely used in classification and regression issues. It builds decision trees on several samples and classifies the data according to the choice made by the majority.

Its most important feature is the Random Forest Algorithm's ability to handle datasets with continuous variables for regression and categorical variables for classification. In terms of categorization issues, it delivers superior outcomes.

Based on the Bagging concept, Random Forest operates. Using replacement to build a distinct training subset from a sample of training data, bagging produces a result based on a majority vote. We used Random Forest to train our Synthetic Identity balance (SMOTE applied) and Imbalanced Datasets.

3.3.1.4 ANN

ANN stands for Artificial neural network. It is a particular kind of Machine Learning algorithm that takes inspiration from the design and operation of the human brain. An artificial neural network (ANN) comprises layers of linked artificial neurons that process input data and produce predictions as output.

The fundamental concept of an ANN is to learn a set of weights and biases that may be used with the input data to produce the desired output. An optimization technique repeatedly changes the weights and biases during training to reduce the difference between the anticipated and actual output.

The ANN architecture used in this project 3.6 comprises an input layer, three hidden layers, and an output layer. The first layer is a dense layer with 70 nodes which takes an input of 512 features and applies the ReLU activation function.

Rectified Linear Unit (ReLU) is a non-linear activation function that adds non-linearity to the model. ReLU is a simple but powerful function that computes the maximum of input and zero.

$$f(x) = max(x, 0)$$

After the first hidden layer, there are two more hidden layers, each dense layer with 50 nodes and ReLU activation function. The output layer uses Sigmoid activation function, producing a binary classification result.

Sigmoid activation function adds non-linearity in the output layer of ANN models. It maps input values into 0 and 1 based on probability.

$$f(x) = \frac{1}{1 + \exp -x}$$

The model uses the Adam optimizer (Adaptive Moment Estimation), an optimization algorithm used in deep learning to update the neural network's weights while training. To produce better results and speed up convergence, it combines the benefits of stochastic gradient descent (SGD) and RMSprop, two other well-known optimization methods. Adam is computationally more efficient, uses less memory, and is less sensitive to the initial learning rate. The model also incorporates an adaptive learning rate optimization approach with early stopping applied to prevent overfitting, and the *binary_crossentropy* loss function, generally used for binary classification problems.

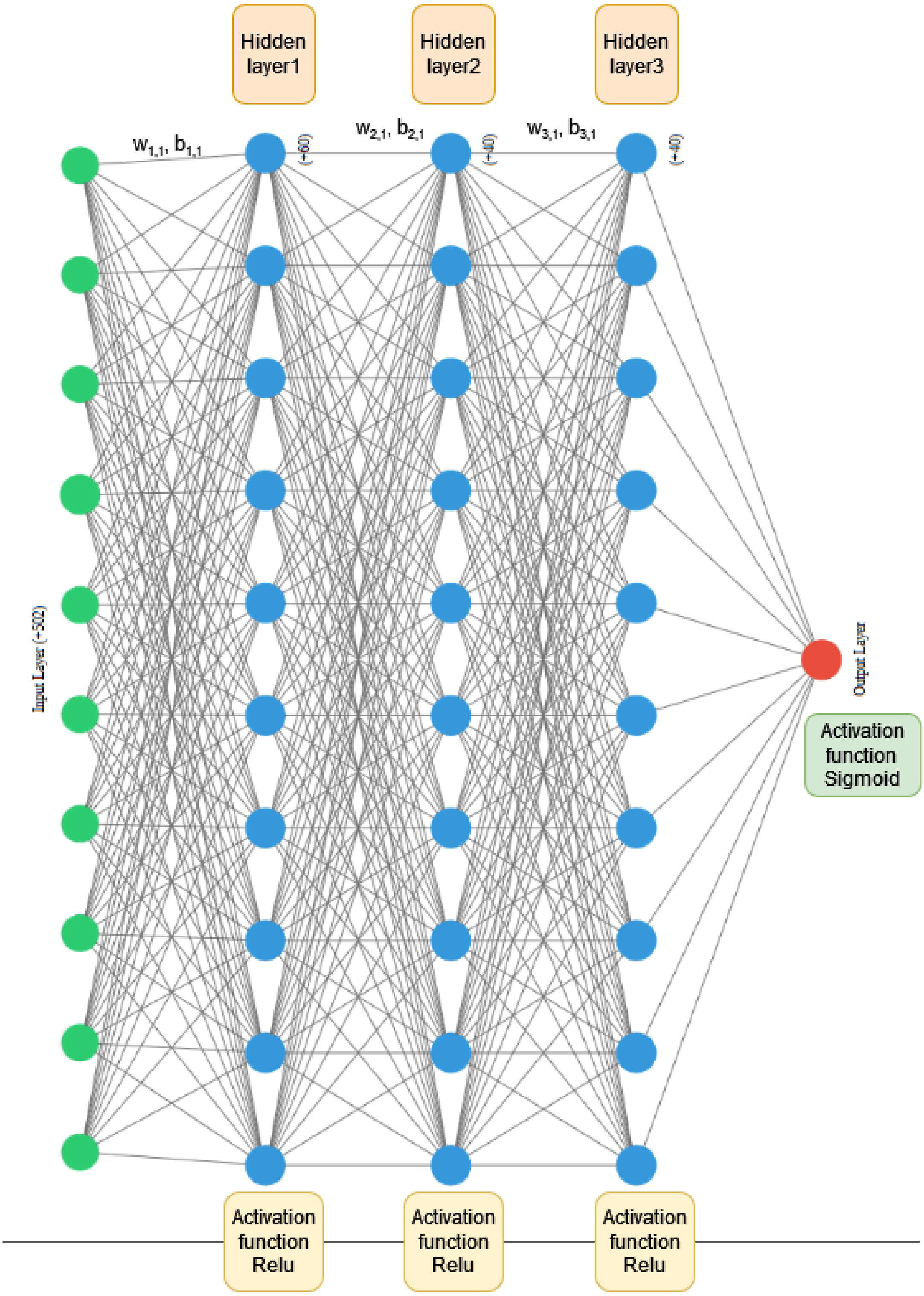


Figure 3.6: ANN Model

3.3.2 Federated Learning Architecture

A Federated Learning approach enables many parties to work together on developing a common Machine Learning model without disclosing their raw data. In Federated Learning, the model training occurs locally on each device or location while the training data is dispersed over several devices or locations. The global model is then shared among the parties by combining the local models.

The fundamental goal of Federated Learning is to get beyond the drawbacks of conventional centralized Machine Learning, in which all the data is gathered in one central location, and the model is trained using that data. Federated learning addresses data privacy, security, and ownership issues by enabling several parties to work together on developing a common model without disclosing their raw data.

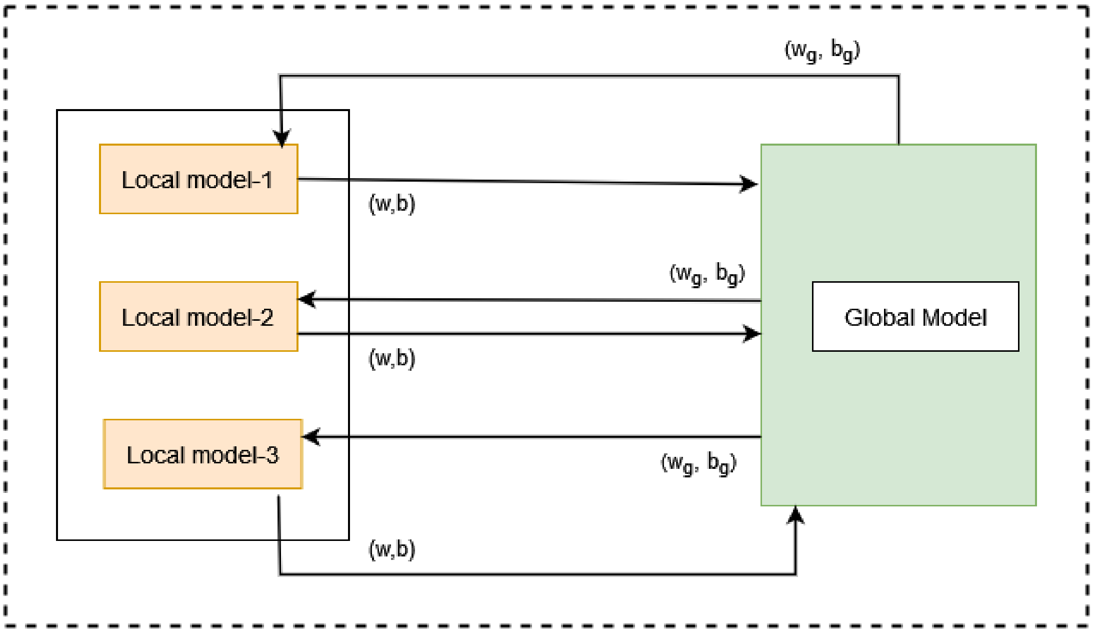


Figure 3.7: Depiction of Federated Learning

There are three major elements involved in the implementation of Federated Learning.

- **Client /Local model :** Client/local model refers to the devices or organizations participating in the learning process. Different individuals or organizations own these devices and are usually distributed geographically. Each client contributes its local data to the training process, and the global model is updated based on the aggregated results from all the clients.

- **Global model** : The global model is a central entity responsible for coordinating the learning process across multiple clients. The global model initiates the training process by sending a model to each client, then collects the updated models from each client after they perform local training on their data. The global model then aggregates the updated models to create a new global model and sends it back to each client to repeat the process. This process is repeated until the global model converges to a satisfactory level of accuracy. The global model is typically maintained by the organization or entity that owns the data and wants to train a Machine Learning model while maintaining data privacy.

- **Aggregator**: In Federated Learning, the weight and bias aggregator is a component of the global server that collects the model updates from participating clients, aggregates them, and computes the updated global model parameters. During each round of the Federated Learning process, the participating clients train their local models on their respective datasets and send the updates to the global model. The weight and bias aggregator collects the updated model parameters from all the clients and performs an aggregation process that combines them into a single set of global model parameters. The aggregation process can take various forms, such as averaging or weighted averaging, depending on the Federated Learning algorithm. Once the weight and bias aggregator has computed the updated global model parameters, it sends them back to the participating clients, and the process repeats for the next round of training. The goal of the weight and bias aggregator is to ensure that the updated global model accurately reflects the collective knowledge of all the participating clients while maintaining privacy and security.

Federated Learning frequently uses the FedAvg (Federated Averaging) method. On decentralized data, it is employed to train Machine Learning models. Each client in FedAvg has their dataset and builds a local model on top of it. A central server receives the local models and averages the model weights before sending the updated global model back to each client. The global model is iteratively repeated until the necessary degree of accuracy is achieved.

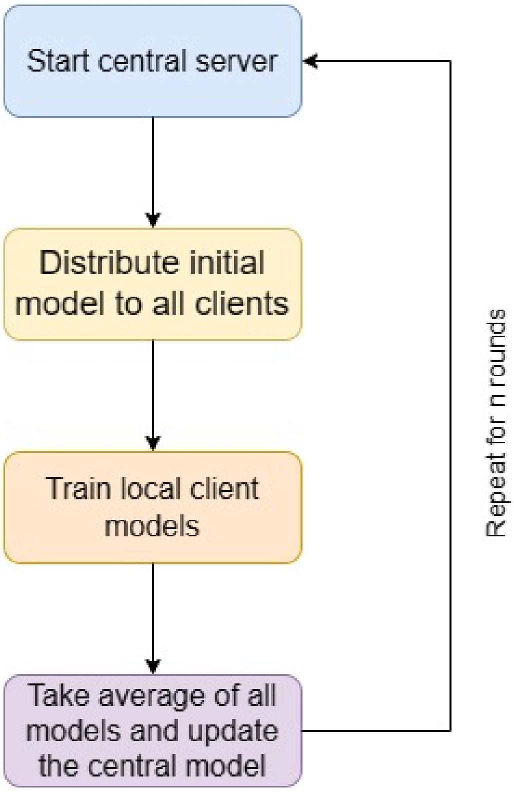

Start central server
Distribute initial model to all clients
Train local client models
Take average of all models and update the central model
Repeat for n rounds

Figure 3.8: Flowchart of Federated Learning

The training process in Federated Learning, as mentioned in Fig 3.8 can be broken down into the following steps:

- **Initialization**: The global server initializes a model with random weights and biases, typically a neural network.
- **Client Selection**: In each training round, a subset of clients is selected randomly or based on specific criteria, such as their computing power, data size, or data distribution. The selected clients download the current global model from the global server to their local devices.
- **Local Training**: Each selected client trains the global model on their local data using local optimization techniques, such as stochastic gradient descent (SGD) or Adam, to minimize the loss function and improve the model's accuracy.

During local training, the client updates the weights and biases of the model to fit their local data.

- **Model Aggregation**: After completing local training, each client sends the updated model parameters, i.e., weights and biases, back to the global server. The global server aggregates these model updates from multiple clients using a certain algorithm, such as Federated Averaging (FedAvg), to generate a new global model that averages the updated models received from the clients.

- **Global Model Update**: The global server sends the updated global model back to the clients, and the training process repeats from step 2. The clients continue to train the global model using their local data, and the global server aggregates the updated models from the clients to improve the global model.

3.4 Summary

Federated Learning is a decentralised approach to machine learning that makes it possible to train a global model without sending raw data to a centralised server. The server initialises the global model, which is then iteratively updated by client models, which use optimisation techniques to train the model locally on their own data. The modified models are then transmitted back to the server, where they are aggregated to produce a better overall model. Repeating this procedure until the appropriate level of precision is attained.

Chapter 4

Experimental Results and Analysis

This chapter is structured into two sections. The first section, Experimental Setup, provides an overview of the hardware and software requirements utilized during the Synthetic Identity Detection and the Synthetic Identity Fraud Detection experiments. The second section, Experimental Results, compares the performance of several traditional machine learning algorithms. Subsequently, the chapter focuses on the results of the proposed federated learning model based on ANN with varying numbers of clients, following the identification of ANN as the most effective method.

4.1 Experimental Setup

The hardware and software specifications for the experiment are as follows.

- Hardware: Intel(R) Core(TM) i7-8700 CPU @ 3.20GHz with 8 GB RAM and 14 GB of memory space and GPU access were used.
- Software: Windows 11 operating system with Python 3, Scikit learn, tensor flow, and Keras Machine Learning libraries are used for binary classification.

4.2 Experimental Results

We have created a dataset of 50,000 genuine profiles and 2500 synthetic profiles with Name, DoB, Mobile, Email, and Location as attributes. Later, after data processing, we extracted features using TfIdfVectorizer, which gave us a 52,500*1,41,142 matrix. This dataset contains noisy, redundant, or irrelevant characteristics, which can harm how well Machine Learning models perform. We have decreased the dataset's dimensionality while retaining as much of the original variance as feasible using truncated

SVD, which results in a 52,500*512 matrix. We have balanced this dataset using SMOTE, which gives us 50,000 genuine and 50,000 synthetic identities. We have analyzed the results of different Machine Learning approaches on unbalanced and balanced datasets in section 4.2.1 and the Federated Learning framework in section 4.2.2.

To evaluate the performance of the ML and FL models, we have used accuracy, precision, recall, and f1-score as parameters, where

- **True Positive**: An instance for which both predicted and actual values are positive.

- **True Negative**: An instance for which both predicted and actual values are negative.

- **False Positive**: An instance for which the predicted value is positive but the actual value is negative.

- **False Negative**: An instance for which the predicted value is negative, but the actual value is positive.

Accuracy is defined as:

$$Accuracy = \frac{True_{positive} + True_{negative}}{True_{positive} + True_{negative} + False_{positive} + False_{negative}}$$

Precision is defined as:

$$Precision = \frac{True_{positive}}{True_{positive} + False_{positive}}$$

The recall is defined as:

$$Recall = \frac{True_{positive}}{True_{positive} + False_{negative}}$$

F1-Score is defined as:

$$F1 - Score = \frac{2 * (Precision * Recall)}{(Precision + Recall)}$$

Model	Accuracy Before Applying SMOTE	Accuracy After Applying SMOTE
Logistic Regression	95.24%	62.83%
KNN	95.22%	54.86%
Random Forest	95.21%	96.73%
ANN	-%	97.25%

Table 4.1: Performance Comparison of Different ML Models with and without SMOTE in Machine Learning

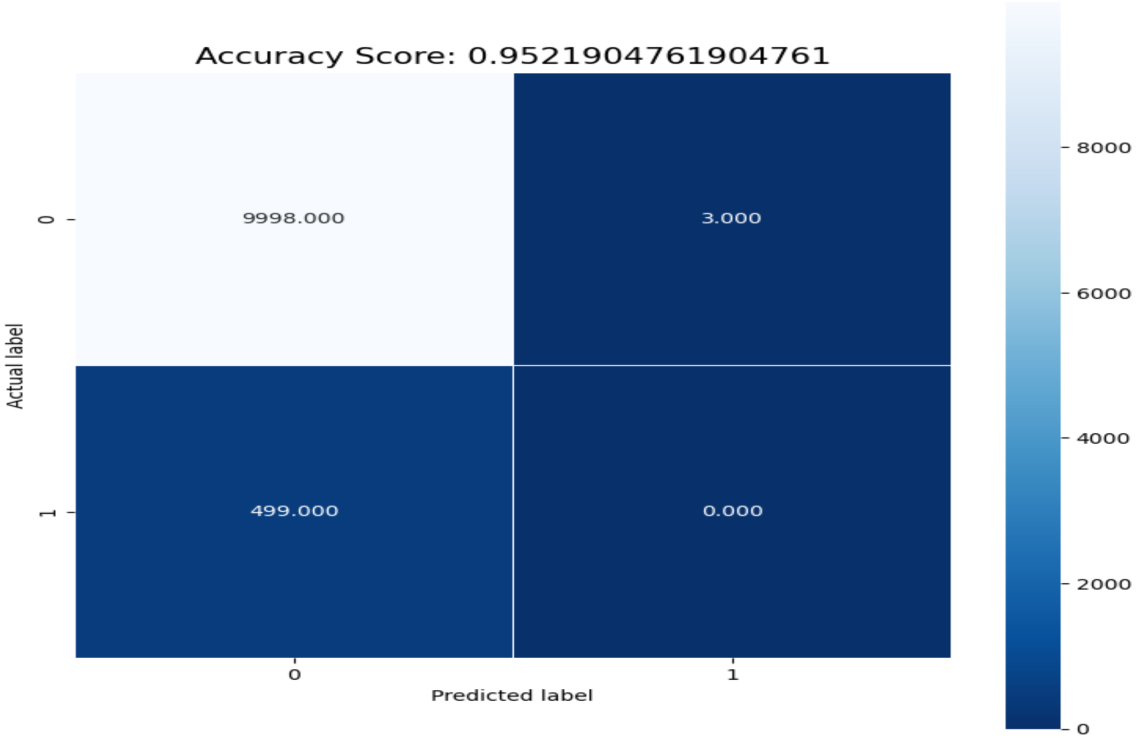


Figure 4.1: Confusion Matrix of Random Forest on Imbalanced Dataset

4.2.1 Analysis of Performance of ML Models

In this subsection, we analyze the results of different Machine Learning models to get the most effective method for Federated learning. We have implemented Logistic Regression, Random Forest, KNN, and ANN, as discussed in Section 3.3.

The observation of Table 4.1 are as follows -

- We are getting unexpected results with an imbalanced dataset on all models, as shown in Fig 4.1. In the case of the imbalanced dataset, we have 50,000 genuine identities and only 2500 synthetic identities. So, each model classifies almost every input as 0 (genuine), only that is why the model has given 95% accuracy.

- We are getting better results on a balanced dataset with 50,000 genuine and 50,000 synthetic identities, as shown in Fig 4.2.
- ANN gives the best result among all the regular learning approaches.

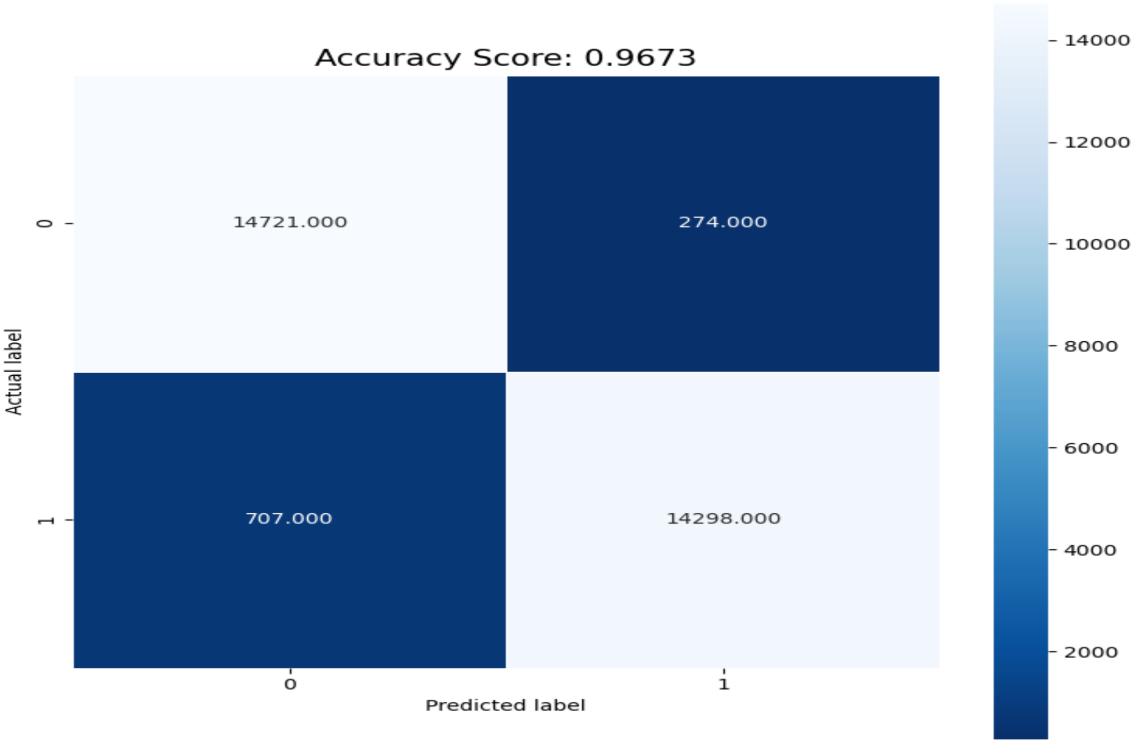


Figure 4.2: Confusion Matrix of Random Forest on Balanced Dataset

4.2.2 Federated Learning

Federated learning is a type of Machine Learning where multiple devices or clients collaboratively train a shared model without sharing their data with a central server. So, we have proposed the Federated Learning model based on ANN with varying numbers of clients(3, 5, and 10 clients).

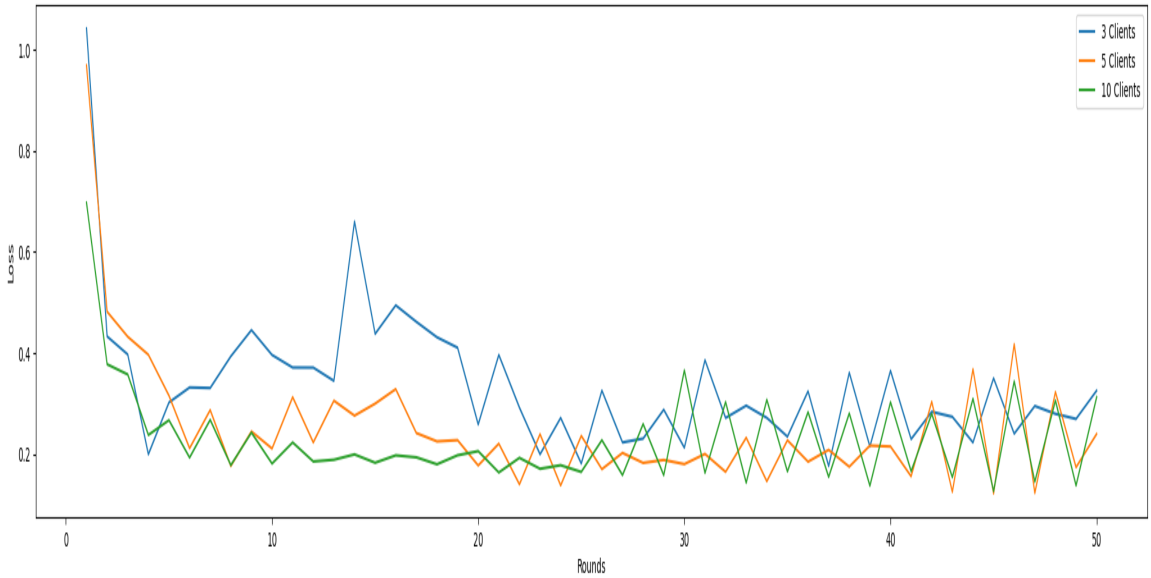


Figure 4.3: Loss Curve of Federated Learning on Different Clients

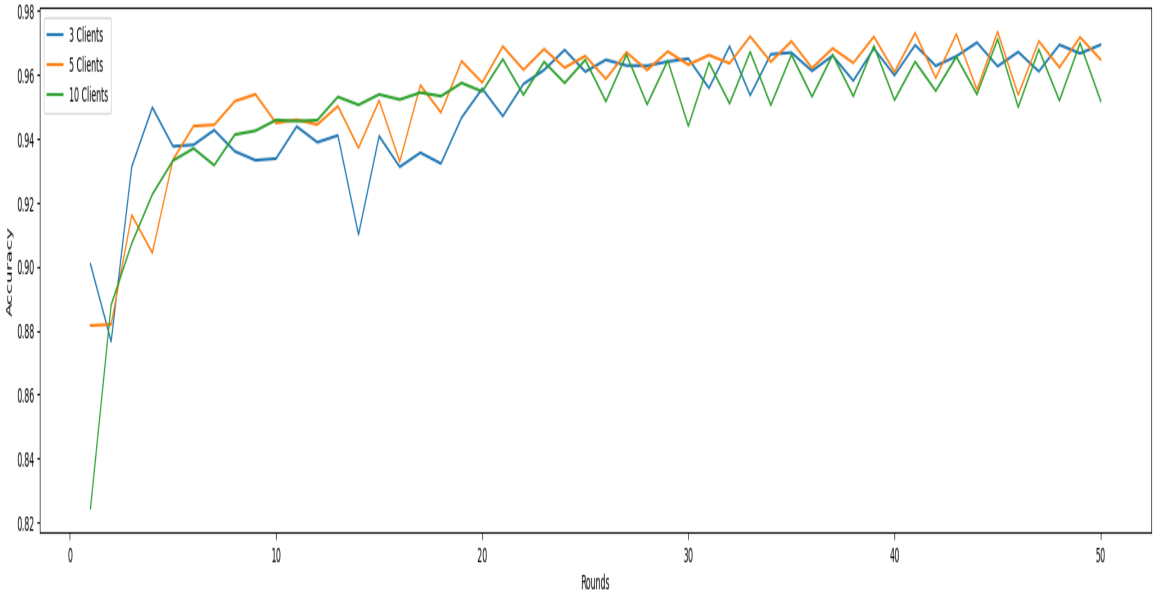


Figure 4.4: Accuracy Curve of Federated Learning on Different Clients

Accuracy	98.0%
Precision	96.67%
Recall	99.16%
F1-Score	97.06%

Table 4.2: Results of Proposed FL Framework

Each local client is trained on ANN with the client's own data. The local model has an input layer with 512 features, three hidden layers with 70, 50, and 50 nodes each, and a Rectified Linear Unit (ReLU) activation function. The ReLU activation function is applied to the output of each hidden layer to introduce non-linearity. The output layer of the local model has one node with a sigmoid activation function, which determines whether the identity is synthetic or genuine. Each client trains their local model on their dataset for 5 epochs using binary cross-entropy as the loss function and the Adam optimizer. The global model is initialized with the same architecture as the local models. It runs for 50 rounds, where on each round, it takes the weights and bias updates of each local client and updates the global weights and biases by aggregating those local weights and biases updates.

The model would be evaluated on a test dataset to determine its accuracy in identifying synthetic and genuine identities. This process would continue for 50 rounds, with the global model becoming increasingly accurate as it incorporates the knowledge and insights from each local client. It is scalable with different no of clients as shown in Fig 4.4.

Federated Learning based on ANN, as shown in Table 4.2, gives the best results as accuracy is 98.04%, and recall is 99.16% which shows that the false negatives value is minimum (there will be only a few cases where model prediction for identity will result in genuine, while actual identity is synthetic). So Federated Learning-based model is more efficient for training models collaboratively while preserving the privacy of user data

Chapter 5

Conclusion and Future Work

Using a Federated Learning-based strategy for Synthetic Identity and Synthetic Identity Fraud Detection is a viable solution for combating Synthetic Identity Fraud. By harnessing Machine Learning, Federated learning enables data to be processed and analyzed without jeopardizing data privacy or security. This method enables numerous institutions to interact and share data while protecting the anonymity of their datasets. Therefore, our proposed Federated Learning based model holds up as an effective means to combat Synthetic Identity Fraud.

As the financial industry faces new difficulties, the application of Federated Learning can aid in detecting and preventing fraudulent conduct while protecting data privacy and security. Federated Learning can assist in uncovering more complex patterns and anomalies in data that may be suggestive of fraudulent behavior by using sophisticated Machine Learning techniques such as deep learning and reinforcement learning. Continued research and development in this field can aid in detecting and preventing fraudulent conduct in the financial industry while protecting the privacy and security of sensitive data. Furthermore, the ideas of Federated Learning may be extended to different sectors, boosting cooperation and data sharing while protecting privacy.

Bibliography

[1] Albitar, Shereen ; Fournier, Sébastien ; Espinasse, Bernard: An effective TF/IDF-based text-to-text semantic similarity measure for text classification. In: *Web Information Systems Engineering–WISE 2014: 15th International Conference, Thessaloniki, Greece, October 12-14, 2014, Proceedings, Part I 15* Springer (Veranst.), 2014, S. 105–114

[2] Ali, Abdulalem ; Abd Razak, Shukor ; Othman, Siti H. ; Eisa, Taiseer Abdalla E. ; Al-Dhaqm, Arafat ; Nasser, Maged ; Elhassan, Tusneem ; Elshafie, Hashim ; Saif, Abdu: Financial Fraud Detection Based on Machine Learning: A Systematic Literature Review. In: *Applied Sciences* 12 (2022), Nr. 19. – URL `https://www.mdpi.com/2076-3417/12/19/9637`. – ISSN 2076-3417

[3] Arnold, Kenneth C. ; Gajos, Krzysztof Z. ; Kalai, Adam T.: On Suggesting Phrases vs. Predicting Words for Mobile Text Composition. In: *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. New York, NY, USA : ACM, 2016 (UIST '16), S. 603–608. – ISBN 978-1-4503-4189-9

[4] Bafna, Prafulla ; Pramod, Dhanya ; Vaidya, Anagha: Document clustering: TF-IDF approach. In: *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, 2016, S. 61–66

[5] Bojilov, Mario: *Methods for Assisting in Detection of Synthetic Identity Fraud in Credit Applications in Financial Institutions*, Central Queensland University, Diplomarbeit, 2023

[6] Dablain, Damien ; Krawczyk, Bartosz ; Chawla, Nitesh: DeepSMOTE:

Fusing Deep Learning and SMOTE for Imbalanced Data. In: *IEEE Transactions on Neural Networks and Learning Systems* PP (2022), 01, S. 1–15

[7] Danasingh, Asir A. ; Balamurugan, Suganya ; EPIPHANY, JEBAMALAR L.: Literature Review on Feature Selection Methods for High-Dimensional Data. In: *International Journal of Computer Applications* 136 (2016), 02

[8] Dong, Yanjie ; Wang, Xuehua: A New Over-Sampling Approach: Random-SMOTE for Learning from Imbalanced Data Sets. In: *Knowledge Science, Engineering and Management*, 2011

[9] Elreedy, Dina ; Atiya, Amir F.: A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance. In: *Information Sciences* 505 (2019), S. 32–64. – URL https://www.sciencedirect.com/science/article/pii/S0020025519306838. – ISSN 0020-0255

[10] Enriquez, Clarice: *Synthetic Identity Theft: Prevention and Detection*, Utica University, Diplomarbeit, 2022

[11] Goyal, Rahul ; Manjhvar, Amit K.: Review on Credit Card Fraud Detection using Data Mining Classification Techniques & Machine Learning Algorithms. In: *IJRAR-International Journal of Research and Analytical Reviews (IJRAR), E-ISSN* (2020), S. 2348–1269

[12] He, Chaoyang ; Li, Songze ; So, Jinhyun ; Zeng, Xiao ; Zhang, Mi ; Wang, Hongyi ; Wang, Xiaoyang ; Vepakomma, Praneeth ; Singh, Abhishek ; Qiu, Hang u. a.: Fedml: A research library and benchmark for federated machine learning. In: *arXiv preprint arXiv:2007.13518* (2020)

[13] Hu, Kai ; Li, Yaogen ; Xia, Min ; Wu, Jiasheng ; Lu, Meixia ; Zhang, Shuai ; Weng, Liguo ; Cheong, Siew A.: Federated Learning: A Distributed Shared Machine Learning Method. In: *Complex.* 2021 (2021), jan. – URL https://doi.org/10.1155/2021/8261663. – ISSN 1076-2787

[14] Ji, Shaoxiong ; Pan, Shirui ; Long, Guodong ; Li, Xue ; Jiang, Jing ; Huang, Zi: Learning Private Neural Language Modeling with Attentive Aggregation. In:

2019 International Joint Conference on Neural Networks (IJCNN). IEEE, jul 2019. – URL `https://doi.org/10.1109%2Fijcnn.2019.8852464`

[15] KHALID, Samina ; KHALIL, Tehmina ; NASREEN, Shamila: A survey of feature selection and feature extraction techniques in machine learning. In: *2014 science and information conference*. 2014, S. 372–378

[16] KHDER, Moaiad A.: Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application. In: *International Journal of Advances in Soft Computing & Its Applications* 13 (2021), Nr. 3

[17] KIM, Eunice ; LEE, Jung-Ah ; SUNG, Yongjun ; CHOI, Sejung M.: Predicting selfie-posting behavior on social networking sites: An extension of theory of planned behavior. In: *Computers in Human Behavior* 62 (2016), S. 116–123. – URL `https://www.sciencedirect.com/science/article/pii/S0747563216302503`. – ISSN 0747-5632

[18] KROTOV, Vlad ; SILVA, Leiser: Legality and Ethics of Web Scraping. In: *Americas Conference on Information Systems*, 2018

[19] KUMBHAR, Rutuja ; MHAMANE, Snehal ; PATIL, Harshada ; PATIL, Sukruta ; KALE, Shubhangi: Text Document Clustering Using K-means Algorithm with Dimension Reduction Techniques. In: *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, 2020, S. 1222–1228

[20] LATCHOUMI, TP ; KANNAN, VV: Synthetic Identity of Crime Detection. In: *International Journal* 3 (2013), Nr. 7, S. 124–129

[21] LEI, John Z. ; GHORBANI, Ali A.: Improved competitive learning neural networks for network intrusion and fraud detection. In: *Neurocomputing* 75 (2012), Nr. 1, S. 135–145. – URL `https://www.sciencedirect.com/science/article/pii/S0925231211003900`. – Brazilian Symposium on Neural Networks (SBRN 2010) International Conference on Hybrid Artificial Intelligence Systems (HAIS 2010). – ISSN 0925-2312

[22] LUNN, Stephanie ; ZHU, Jia ; ROSS, Monique: Utilizing Web Scraping and Natural Language Processing to Better Inform Pedagogical Practice. In: *2020 IEEE Frontiers in Education Conference (FIE)*, 2020, S. 1–9

[23] MARQUES, Pedro ; DABBABI, Zayani ; MIRONESCU, Miruna-Mihaela ; THONNARD, Olivier ; BESSANI, Alysson ; BUONTEMPO, Frances ; GASHI, Ilir: Detecting Malicious Web Scraping Activity: A Study with Diverse Detectors. In: *2018 IEEE 23rd Pacific Rim International Symposium on Dependable Computing (PRDC)*, 2018, S. 269–278

[24] REMESEIRO, Beatriz ; BOLON-CANEDO, Veronica: A review of feature selection methods in medical applications. In: *Computers in Biology and Medicine* 112 (2019), S. 103375. – URL `https://www.sciencedirect.com/science/article/pii/S0010482519302525`. – ISSN 0010-4825

[25] ROUT, Neelam ; KUHOO ; MISHRA, Debahuti ; MALLICK, Manas: Ensemble learning for handling imbalanced datasets with the combination of bagging and sampling methods. In: *Indian Journal of Public Health Research Development* 9 (2018), 09, S. 1412

[26] SINGRODIA, Vidhi ; MITRA, Anirban ; PAUL, Subrata: A Review on Web Scrapping and its Applications. In: *2019 International Conference on Computer Communication and Informatics (ICCCI).* 2019, S. 1–6

[27] SIVANANTHAM, S. ; DHINAGAR, S. R. ; KAWIN, P. ; AMARNATH, J.: Hybrid Approach Using Machine Learning Techniques in Credit Card Fraud Detection. In: SURESH, P. (Hrsg.) ; SARAVANAKUMAR, U. (Hrsg.) ; HUSSEIN AL SALAMEH, Mohammed S. (Hrsg.): *Advances in Smart System Technologies.* Singapore : Springer Singapore, 2021, S. 243–251. – ISBN 978-981-15-5029-4

[28] VARMEDJA, Dejan ; KARANOVIC, Mirjana ; SLADOJEVIC, Srdjan ; ARSENOVIC, Marko ; ANDERLA, Andras: Credit Card Fraud Detection - Machine Learning methods. In: *2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH)*, 2019, S. 1–5

[29] Walker-Moree, Kimberly: *Analysis of the Evolution and Impact of Synthetic Identity Fraud: Is It Preventable?*, Utica College, Diplomarbeit, 2018