Synthetic Identity and Synthetic Identity Fraud Detection using Machine Learning

Thesis

submitted in partial fulfilment of the requirements for the award of the degree of

Master of Technology

(Computer Science and Engineering - Information Security)

by

Battu Vijaya Krishna

(Roll No: 202IS008)

Under the supervision of

Dr. Mahendra Pratap Singh

Assistant Professor

Department of Computer Science and Engineering

National Institute of Technology Karnataka, Surathkal

Mangaluru-575025, India

July-2022

DECLARATION

I hereby *declare* that the Report of the P.G. Project Work entitled **Synthetic Identity and Synthetic Identity Fraud Detection using Machine Learning** which is being submitted to the National Institute of Technology Karnataka Surathkal, in partial fulfilment of the requirements for the award of the Degree of **Master of Technology** in **Computer Science and Engineering - Information Security** in the department of **Computer Science and Engineering**, is a *bonafide report of the work carried out by me*. The material contained in this Report has not been submitted to any University or Institution for the award of any degree.

Battu Vijaya Krishna
Reg No. 202620IS008

.....
(Signature of the student)
Computer Science and Engineering

Place: NITK, Surathkal
Date: Monday 25th July, 2022

CERTIFICATE

This is to certify that the P.G. Project Work Report entitled **Synthetic Identity and Synthetic Identity Fraud Detection using Machine Learning** submitted by **Battu Vijaya Krishna** (Registration number: 202620IS008) as the record of the work carried out by him, is accepted as the P.G. Project Work Report submission in partial fulfillment of the requirements for the award of the degree of **Master of Technology** in **Computer Science and Engineering - Information Security** in the department of **Computer Science and Engineering**.

Dr. Mahendra Pratap Singh

Assistant Professor,
Computer Science & Engineering,
National Institute of Technology
Karnataka, Surathkal
Mangaluru - 575025, India.

Chairman - DPGC
(Signature with Date and Seal)

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deep sense of profound gratitude to my research supervisor **Dr. Mahendra Pratap Singh**, *Assistant Professor, Department of Computer Science and Engineering*, *NITK, Surathkal* for his constant guidance, support, and motivation during my research work. His organised and methodical approach continuously encouraged me to complete all of my tasks on time, including the research work. Throughout this project, he gave informative suggestions and ideas at the appropriate times, which greatly aided the smooth operation of the research process. Motivation, appreciation, moral support, and the flexibility allowed during the course all contributed to my effective completion of my research.

I must also thank the **Dr. Shashidhar G Koolagudi** (HOD), faculty members, and staff of the *Department of Computer Science and Engineering*, *NITK, Surathkal* for their kind help and support.

I owe my utmost gratitude to my parents **Mr. Venkata Veeraraghavaiah Battu** and **Mrs. Kanakaratnam Battu** for their love and support throughout my life. I appreciate them both for giving me the courage, wisdom, and freedom to achieve my goals.

Once more, I want to express my gratitude to my supervisor and my family for their support in helping me complete this project.

Place: Karnataka, India
Date: Monday 25th July, 2022. **Battu Vijaya Krishna**

Abstract

Even though identity theft has been an issue for many years, its perpetrators' tactics have evolved in recent years. Actual identity fraud was a common practice among criminals in the past. Organizations, like Banks, have improved their ability to control such frauds using technology, although crime has changed over time. In response, criminals have moved onto a harder-to-detect scheme known as synthetic identity theft. This new strategy involves creating a fake identity using accurate information from several victims, such as addresses, Social Security numbers, and other details. Due to the way fraudsters typically operate, there is a risk. Therefore, they use synthetic identity to obtain service (e.g., bank credit) without liability. It is simpler for criminals to avoid detection for more prolonged periods because there isn't a single clear-cut victim. More worrying still, significant losses are building up behind these IDs. To crack down on it, every customer requesting credit must undergo even more stringent ID checks than they do. This thesis presents a novel approach that analyzes application details and can identify whether they are synthetic and have been used to commit fraud. To determine whether the applicant's basic information matched a real person, we use LinkedIn profile data's personal information and create a dataset to train the machine learning model. The model's performance shows that the proposed approach can be a viable solution to address the synthetic identity issue.

Keywords: Synthetic Identity, Synthetic Identity Fraud, Credit Card Application Fraud, Machine Learning, Artificial Neural Network.

Contents

List of Figures **v**

List of Tables **vii**

List of Abbreviations **x**

1 Introduction **1**

1.1 Traditional Identity Fraud 1

1.2 Identity Theft Types 2

1.3 Synthetic Identities Creation 4

1.4 Synthetic Identity Fraud 5

1.4.1 Reasons for Synthetic Identity Frauds Popularity 5

1.4.2 Synthetic Identity Fraud Methodology 6

1.5 Problem Description 7

1.6 Motivation 8

1.7 Contributions 8

1.8 Organization of the Thesis 8

List of Figures

1.1 Traditional Identity Fraud 1
1.2 Synthetic Identity Fraud 2
1.3 Synthetic Identity Creation 4
1.4 Synthetic Identity Fraud Methodology 7

List of Tables

List of Algorithms

LIST OF ABBREVIATIONS

SI	Synthetic Identity
SID	Synthetic Identity Detection
SIF	Synthetic Identity Fraud
PII	Personally Identifiable Information
Logistic	Logistic Regression
GNB	Gaussian Naive Bayes
Multi-NB	Multi-Nomial Naive Bayes
DT	Decision Tree
RF	Random Forest
ANN	Artificial Neural Network
OTP	One Time Password
SSN	Social Security Number
DOB	Date Of Birth
NLTK	Natural Language Toolkit
FE	Feature Extraction
BoW	Bag Of Words
Model	Machine 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

Chapter 1

Introduction

Today, one of the crimes with the most significant growth is Synthetic Identity Fraud, which contributes significantly to credit losses for all Financial Institutions. In contrast to traditional identity theft, Synthetic identity fraud includes using accurate personal information, such as social security numbers, along with a false identity (produced by using several genuine documents) to produce an overall fraudulent identity. These Synthetic Identities are then used to apply for a credit card, purchase luxury, and apply for loans. The combination of fake and accurate info makes identifying and preventing fraud challenging for financial institutions difficult.

1.1 Traditional Identity Fraud

Figure 1.1: Traditional Identity Fraud

In Traditional identity fraud, the fraudster steals one person's personal identity information and tries to perform illegal activities. It does not include multiple people's

identity information. Fraudster pretends to be another person (victim) to get their financial details. For traditional fraud detection, financial institutions use two-factor or OTP-based authentication to know whether the person accessing the service is a real person or a fraudster. Figure 1.1 shows how true identity fraud happens.

1.2 Identity Theft Types

Fraudsters first gather all the necessary information to perform fraud. Once data is collected, they try to create Synthetic Identities using multiple person's details into single application details. Figure 1.2 shows different identity theft types, and their description is as follows.

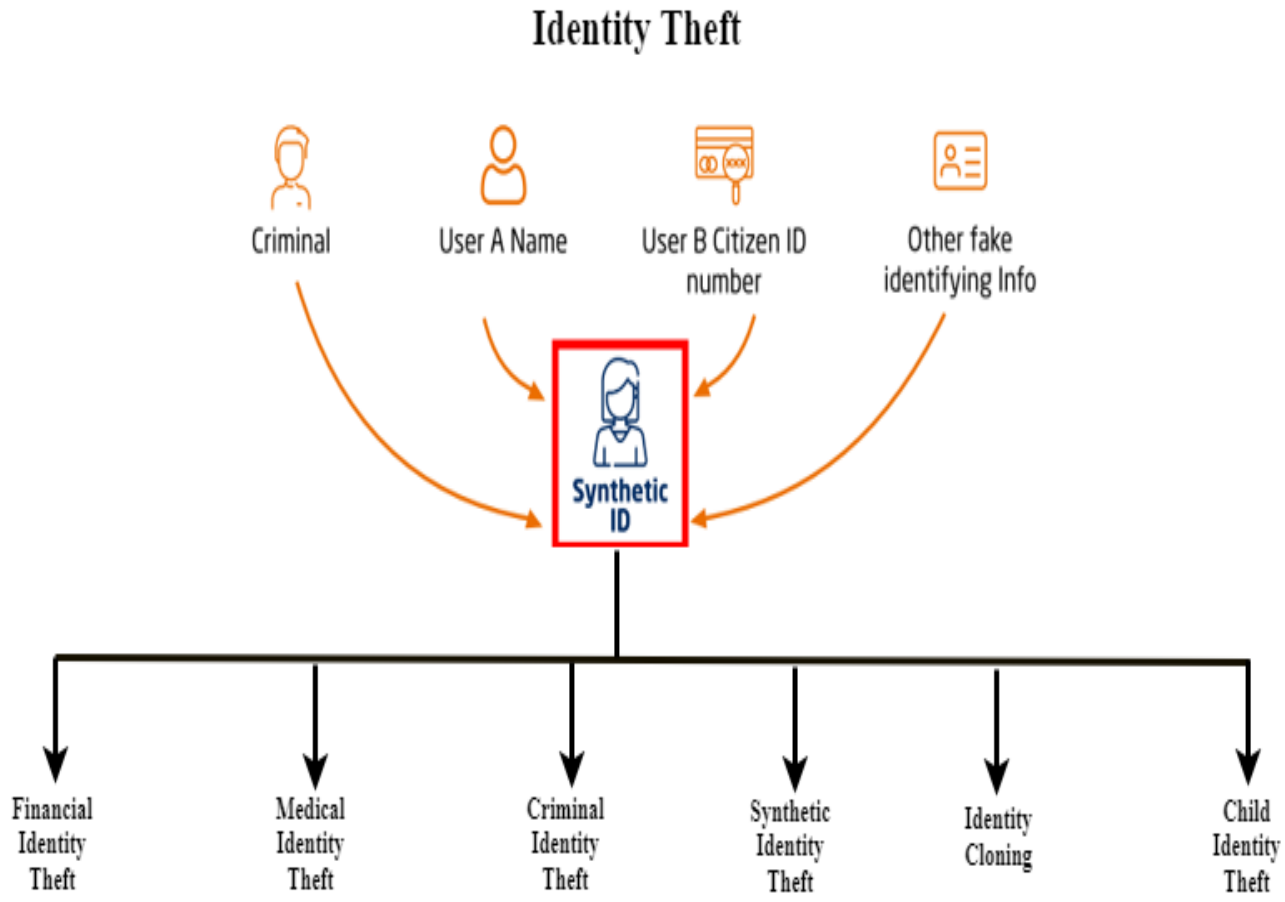

Figure 1.2: Synthetic Identity Fraud

- **Financial Identity Theft:** It involves accessing a bank account or obtaining benefits like products, services, and credit using another person's information. The criminal opens a new credit card account with access to credit or a bank

account with the ability to get blank cheques in the victim's name using the stolen personal information.

- **Medical Identity Theft:** When committing medical identity theft, the offender uses the victim's name without their knowledge or consent, in addition to other information, such as that of insurance, the securing of medical benefits involving goods or services, or the securing of fictitious reimbursements for claimed utilization of medical goods or services.

- **Criminal Identity Theft:** When being detained by the authorities for a crime, the criminal, in this instance, makes up a victim's identity. The identity claimed by the perpetrator may be supported by government-issued identification documents that were applied for and acquired using credentials taken from the victim. The conclusion is that the real identity of the culprit is hidden, and charges could ultimately be brought against the victim rather than the criminal. In these types of theft instances, victims are frequently unaware that crimes have been done in their names and are therefore likely to be held accountable for any illegal activity.

- **Synthetic Identity Theft:** It may involve the partial or complete fabrication of identity information. For example, a valid social security number can be used in conjunction with a name and birth date that are not those of the actual owner. This type of identity theft makes it harder to track down crimes committed. Due to that, they frequently do not appear clearly on the victim's record, such as his credit report. Instead, they could appear as whole new files or as an adjunct to the victim's current credit report. Typical victims of this type of identity theft are creditors who unintentionally give the perpetrator credit.

- **Identity Cloning:** In this scenario, the offender impersonates someone else, intending to hide their true identity. The offender's goal may occasionally be to remain anonymous for personal reasons, even though the typical motivation is to commit a crime. Identity cloning and concealment may never be discovered, mainly when the offender has acquired fake documents that allow him to pass regular authentication checks. In contrast, financial identity theft may only

become apparent after the victim has racked up large debts in their name.

- **Child Identity Theft:** In this type of identity theft, the perpetrator—often a predator of children or a relative—uses a Social Security number of a youngster to acquire access to something, like credit. Minors' social security numbers are helpful to identity thieves because they contain no personal information.

1.3 Synthetic Identities Creation

Synthetic Identity Fraud (SIF) can work with just one piece of personal identity data to generate a synthetic identity. Criminals might use various strategies to create fake identities to do the fraud. Below Figure:1.3 shows difference between traditional identity and synthetic identity.

Figure 1.3: Synthetic Identity Creation

- **Traditional Identity Theft:** In this fraudster uses actual personal identity information for all data elements, as shown in Figure 1.3.

- **Identity Manipulation:** For generating a new identity, actual personal identifying data is slightly modified, as shown in Figure 1.3.

- **Identity Compilation:** Many authentic and fake personal identification data elements are combined to create a new identity. The development of synthetic identities most frequently takes the form, as shown in Figure 1.3.

1.4 Synthetic Identity Fraud

Synthetic Identity fraud is perpetrated by combining fictitious and sometimes genuine information to create new identities to defraud financial institutions, businesses, and government agencies.

According to the US Federal Reserve, Synthetic Identity Fraud is a financial crime with the fastest global growth rate, and typical fraud models miss 85–95 % of possible synthetic identities. Currently, Synthetic Identity Fraud is responsible for 20% of credit losses, costing lenders $6 billion in 2016, with an average fraud loss of $15,000 per account.

Synthetic identity fraud presents numerous risks for financial organizations.

- **Regulatory:** greater vigilance as a result of compilation errors.
- **Operational:** operational disruptions and inefficient resource usage in the organisation.
- **Reputational:** consumer trust, money laundering, financing of crime or terrorism.
- **Financial:** fraud losses, increasing opeational expenses and fines.

1.4.1 Reasons for Synthetic Identity Frauds Popularity

Companies are having trouble adjusting to Synthetic Identity Fraud, one of the frauds with the most significant growth. The following are the drivers for the popularity of fraud:

- The growing number and size of data breaches are significant factors in the rise of Synthetic Identity Fraud.
- The increased sophistication of the dark web is the main driver behind the simple accessibility of stolen Personal identifying information from data breaches. The dark web has accelerated the sale and circulation of fake identities.
- Synthetic Identity Fraud systems can mimic "excellent" consumer conduct. According to a Federal Reserve investigation, 70% of suspected instances momentarily display conventional consumer tendencies.

- Due to the challenges in identifying Synthetic Identity Fraud, many businesses write off cases as bad debt. This enables the fraudster to escape punishment. Additionally, it makes it difficult to grasp the scope of the problem and the fraud's actions and telltale signals.

- Another factor facilitating the expansion of Synthetic Identity Fraud is consumer preference for quicker and easier account sign-up procedures. SIF usage has become easier for fraudsters due to a decline in in-person account setups and more straightforward online sign-up.

- The fact that Synthetic Identity Fraud frequently stays unnoticed for months makes it even more frightening. An individual's recognition of the theft is commonly delayed or does not occur since the personal identity information exploited is frequently only one piece of information, not the entire person.

- The improved technology used by crooks has facilitated the fraud's capacity to grow and be more effective. The technique enables the fraudsters to rapidly scale up for several Synthetic Identity Fraud attacks once they have chosen a target.

- Cybercriminals use cutting-edge technologies to create new fraud versions to make detection harder.

1.4.2 Synthetic Identity Fraud Methodology

Figure 1.4 shows Synthetic Identity Fraud methodology. In the first stage, the fraudster fabricates a synthetic identity using false personal identity information. In the second stage, the fraudster applies for credit, causing the credit bureau to create a credit account. In the third stage, the fraudster repeatedly applies for credit until it gets approved. In the fourth stage, the fraudster legitimizes the synthetic identity and increases its credit worthiness. In the fifth stage, the fraudster bursts out and vanishes without paying.

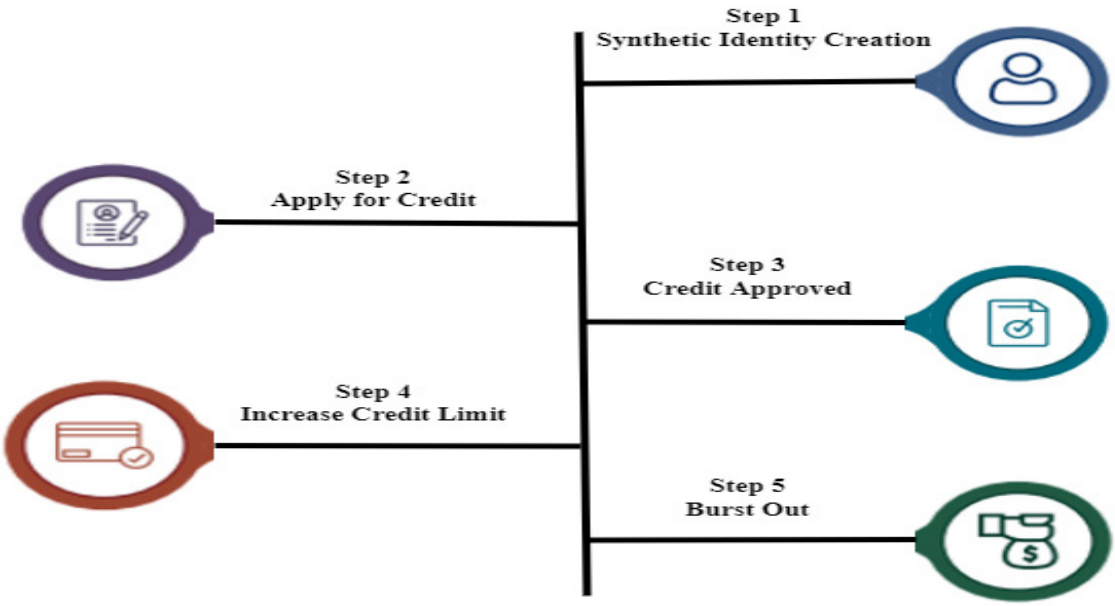

Figure 1.4: Synthetic Identity Fraud Methodology

1.5 Problem Description

Fraudsters apply to financial institutions once Synthetic Identities are ready. Using for credit produces a credit file in the name of the synthetic Identity at the financial institution, allowing the fraudster to open accounts in this name and start building credit. The credit file looks exactly like many real people just starting their credit histories. Financial institutions fail to conduct more stringent screening to identify synthetic Identities and onboard new customers due to a lack of third-party data or have no efficient mechanism to confirm whether the given applicant details are authentic or fraudulent.

A method for detecting synthetic identities that involve the use of third-party data can be a helpful tool. It can be built on the concept that real people have histories, which they left behind in numerous physical and digital data systems. Real-life persons provide a consistent trail: the same address, email account, and phone number appear in many databases. These trails are hard to replace. They have the breadth or enormous amounts of data going back years. The Genuine ID may be too consistent, whereas the synthetic Identity is fabricated. So we prefer to tackle this issue using people's LinkedIn information and machine learning models.

1.6 Motivation

For Synthetic Identity Detection, McKinsey Company developed a scoring-based system that assigns density values to each field of the application. It also sets a consistency score for each application attribute if that attribute value existed in different internal databases of McKinsey. Ultimately, it will give both density and consistency scores for each application. All the applications are segregated into four categories based on the threshold value. It is not enough to detect Synthetic Identity Fraud because the fraudster's way of performing Synthetic Identity fraud is changing daily. We need a multi-layered approach to defend it. With this work, we try to present a model that can detect Synthetic Identities and Synthetic Identity fraud, and this approach could be helpful for future Synthetic Identities Detection.

1.7 Contributions

Our contributions are the following.

- Created a Synthetic Identity Dataset using OSINT.
- Designed and developed a Machine Learning based model to detect Synthetic Identities and Synthetic Identity Fraud.

1.8 Organization of the Thesis

The remaining portions of the thesis are divided into four chapters. The first chapter, Literature Review, walks you through the existing Synthetic identities and Synthetic Identity Fraud challenges in the financial sector. The second chapter, Proposed Synthetic Identity Fraud Detection Model, elaborates on the details and work of the developed Synthetic Identity and Synthetic Identity Fraud Detection. The third chapter, Experimental Results and Analysis, describes the environment setup. The fourth and last chapter, Conclusion and Future Scope, summarizes the work and presents future research directions.