

**Healthcare Insurance Fraud Detection Through Data  
Mining Techniques**



STUDENT NAME: Zain Raza Hamid  
ENROLLMENT NO: 01-249211-017  
SUPERVISOR: Dr. Fatima Khaliq

A thesis submitted in fulfilment of the requirements for the award  
of degree of Masters of Science (Data Science)

Department of Computer Science  
BAHRIA UNIVERSITY ISLAMABAD

March 2023

## Approval of Examination

Scholar Name: Zain Raza Hamid

Registration Number: 74808

Enrollment: 01-249211-017

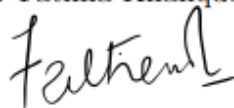
Program of Study: MS Data Science

Thesis Title: Healthcare Insurance Fraud Detection Through Data Mining Techniques

It is to certify that the above scholar's thesis has been completed to my satisfaction and, to my belief, its standard is appropriate for submission for examination. I have also conducted plagiarism test of this thesis using HEC prescribed software and found similarity index 18%. that is within the permissible limit set by the HEC for the MS/M.Phil degree thesis. I have also found the thesis in a format recognized by the BU for the MS/M.Phil thesis.

Principal Supervisor Name: Dr. Fatima Khalique

Principal Supervisor Signature:



Date: March 03, 2023

## **Author's Declaration**

I, Zain Raza Hamid hereby state that my MS/M.Phil thesis titled is my own work and has not been submitted previously by me for taking any degree from Bahria university or anywhere else in the country/world. At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw/cancel my MS/M.Phil degree.

Name of Scholar: Zain Raza Hamid

Date: March 03, 2023

## **Plagiarism Undertaking**

I, solemnly declare that research work presented in the thesis titled Healthcare Insurance Fraud Detection Through Data Mining Techniques is solely my research work with no significant contribution from any other person. Small contribution / help wherever taken has been duly acknowledged and that complete thesis has been written by me. I understand the zero tolerance policy of the HEC and Bahria University towards plagiarism. Therefore I as an Author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred / cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS/M.Phil degree, the university reserves the right to withdraw / revoke my MS/M.Phil degree and that HEC and the University has the right to publish my name on the HEC / University website on which names of scholars are placed who submitted plagiarized thesis.

Name of Scholar: Zain Raza Hamid

Date: March 03, 2023

## Dedication

This work is for the people who have made a big difference in my life and encouraged me to go after my dreams.

To my family, who have been my constant support system, providing me with love, encouragement, and unwavering support throughout my academic journey.

To my friends, who have been my pillars of strength, offering me their unwavering support, understanding, and encouragement. Your presence in my life has made all the difference.

To my mentors and professors, who have challenged me to think critically, provided me with invaluable guidance, and inspired me to aim higher.

To all the participants who contributed to this study, your willingness to share your experiences and perspectives has made this work possible. Your insights and feedback have been invaluable, and I am grateful for your participation.

Finally, I dedicate this work to myself, for all the hard work, late nights, and sacrifices that went into producing this thesis. This accomplishment would not have been possible without my dedication, passion, and persistence.

To all those who have played a role in my journey, I express my deepest gratitude.

This work is a tribute to you all.

## Acknowledgements

In preparing this thesis, I was in contact with many people, researchers, academicians, and practitioners. They have contributed towards my understanding and thoughts. In particular, I wish to express my sincere appreciation to my main thesis supervisor, Dr. Fatima Khalique, for encouragement, guidance, and critics. I am also very thankful to the former and current Director of CoE- AI Professor Dr. Imran Saddique and Dr. Summaira Kausar for their guidance, advices and motivation. Without their continued support and interest, this thesis would not have been the same as presented here.

Librarians at Bahria University also deserve special thanks for their assistance in supplying the relevant literatures. My fellow postgraduate students should also be recognised for their support. My sincere appreciation also extends to all my colleagues and others who have provided assistance at various occasions. Their views and tips are useful indeed. Unfortunately, it is not possible to list all of them in this limited space. I am grateful to all my family members.

## Abstract

Healthcare programs and insurance initiatives play a crucial role in ensuring that people have access to medical care. Countries as well as corporate companies around the world provide healthcare facilities to their citizen and employees for the balance and healthy life. However, despite the benefits of these programs, healthcare insurance fraud continues to be a significant challenge in the industry. Reports says, amount worth more than \$760 Billion wasted every years in terms of insurance fraud in United States.

In this study, we present a model that utilizes five unsupervised learning techniques to detect healthcare insurance fraud. We used the Centers for Medicare and Medicaid Services (CMS) 2008-2010 DE-SynPUF dataset for our analysis. Our model began by implementing the Apriori Association Rule Mining Technique to extract frequent rules from the dataset. We then passed the extracted rules to fraudulent classifiers, such as IF, CBLOF, ECOD, and OCSVM, to identify fraudulent activity. However, while our model demonstrated potential, further research and testing are necessary to improve its effectiveness and accuracy. The healthcare industry generates vast amounts of data, and a more extensive analysis of multiple healthcare insurance datasets could improve our model's performance. Machine learning solutions offer the possibility of significantly reducing fraudulent activity in the healthcare industry, which could result in improved patient care and reduced healthcare costs.

In conclusion, our research demonstrates the potential for using data mining techniques to detect healthcare insurance fraud. By identifying fraudulent activity, we can take measures to prevent it, resulting in a more efficient and cost-effective healthcare system. Our study contributes to the growing body of literature on machine learning and fraud detection and underscores the importance of continued research in this area.

## TABLE OF CONTENTS

<b>AUTHOR'S DECLARATION</b>	<b>ii</b>
<b>PLAGIARISM UNDERTAKING</b>	<b>iii</b>
<b>DEDICATION</b>	<b>iv</b>
<b>ACKNOWLEDGEMENTS</b>	<b>v</b>
<b>ABSTRACT</b>	<b>vi</b>
<b>LIST OF TABLES</b>	<b>ix</b>
<b>LIST OF FIGURES</b>	<b>x</b>
<b>LIST OF SYMBOLS</b>	<b>xii</b>
<b>1 INTRODUCTION</b>	<b>1</b>
1.1 Financial Losses . . . . .	2
1.2 Fraud, Waste, and Abuse . . . . .	4
1.2.1 Fraud . . . . .	5
1.2.2 Waste . . . . .	6
1.2.3 Abuse . . . . .	7
<b>2 RELATED WORK</b>	<b>9</b>
<b>3 MATERIALS &amp; METHODS</b>	<b>14</b>
3.1 Dataset . . . . .	14



3.2	Proposed Methodology . . . . .	16
3.2.1	Apriori Algorithm . . . . .	17
3.2.2	Isolation Forest . . . . .	18
3.2.3	Cluster Based Local Outlier Factor . . . . .	20
3.2.4	One-Class Support Vector Machine . . . . .	22
3.2.5	Empirical Cumulative distribution based Outlier Detection	24
3.3	Evaluation . . . . .	25
<b>4</b>	<b>RESULTS &amp; ANALYSIS</b>	<b>28</b>
4.1	Descriptive Analysis . . . . .	28
4.2	Preprocessing . . . . .	33
4.3	Findings and Interpretations . . . . .	36
<b>5</b>	<b>CONCLUSION &amp; FUTURE WORK</b>	<b>55</b>
	<b>REFERENCES</b>	<b>56</b>

## LIST OF TABLE

4.1	Procedure Codes Similarity Check W.R.T. Diagnosis Codes . . .	33
4.2	Description of features chosen from the DE-SynPUF . . . . .	35
4.3	Unsupervised techniques applied independently on the dataset .	37
4.4	Unsupervised techniques applied on the rules extracted from Apriori algorithm . . . . .	37
4.6	Results of Apriori Association Rule Mining . . . . .	41
4.7	Rules Classification Through Applied Techniques . . . . .	42
4.8	Identification of fraudulent rules through applied techniques . .	49
4.9	Identification of fraudulent rules through applied techniques . .	51
4.10	Silhouette Score of applied techniques . . . . .	52

## LIST OF FIGURE

1.1	Average sources of waste in US health care every year . . . . .	2
1.2	Healthcare insurance ecosystem involves patient, hospital, and services providers . . . . .	4
1.3	Healthcare insurance ecosystem involves patient, hospital, and services providers . . . . .	5
3.1	All the table and their attributes of DE-SynPUF . . . . .	15
3.2	Proposed Ensemble Technique . . . . .	17
3.3	Association Rule Mining Process Flow Chart . . . . .	18
3.4	Isolation Forest . . . . .	20
3.5	Cluster Based Local Outlier Factor . . . . .	21
4.1	Provider institutions occurrence in entire dataset . . . . .	29
4.2	Attending physicians occurrences in entire dataset . . . . .	30
4.3	Operating physicians occurrences in entire dataset . . . . .	31
4.4	Unique & common physicians in attending physicians and operating physicians . . . . .	31
4.5	All diagnosis codes with their occurrences in the dataset, 4019 occurred the most . . . . .	32
4.6	All procedure codes with their occurrences in dataset . . . . .	33
4.7	Codes Comparison Summary . . . . .	34
4.8	Fraudulent Classification Through Classifiers . . . . .	42
4.9	Fraudulent Classification Through Isolation Forest . . . . .	53
4.10	Fraudulent Classification Through CBLOF . . . . .	53

4.11 Fraudulent Classification Through ECOD . . . . .	54
4.12 Fraudulent Classification Through OCSVM . . . . .	54

## LIST OF SYMBOLS

$\rightarrow$  – With respect to

$\cup$  – Union

$\sum$  – Summation

$w$  – vector

$\delta$  – Empirical Cumulative Distribution Functions

$\log$  – logarithm

# CHAPTER 1

## INTRODUCTION

The healthcare system plays a crucial role in maintaining the health and well-being of society, and many countries provide health insurance to their citizens to ensure they have access to medical care when needed. Health insurance can be provided by both public and private entities, and it helps to cover the cost of medical treatments, procedures, and medications. This system also helps to protect people from the financial burden of unexpected medical expenses that can arise due to illness or injury. The Sehat Sahulat Program was a health insurance initiative launched by the government of Pakistan in partnership with provincial governments, aimed at providing health coverage for needy people to minimize or eliminate out-of-pocket expenses and reduce poverty [1]. The program covers emergency and inpatient services requiring secondary and tertiary care but does not include outpatient services. The financial range for overall treatment coverage varies from 720,000 to 1,000,000 PKR and includes transportation for maternal care, referrals to tertiary care, and funeral allowances [2]. Similarly United States has its own Federal Government sponsored national healthcare program, Medicare, which provides affordable health insurance to individuals 65 years and older, and other select individuals with permanent disabilities [3]. Other than United States, countries like Canada, UK, France and many other also provide such facilities to their citizens.

## 1.1 Financial Losses

Advancements in medical sciences and technology have led to significant improvements in the health and well-being of the general public. However, the cost of quality healthcare can be high, and this is where Health Insurance Plans come in handy. Despite this, there are still individuals who engage in fraudulent activities to benefit themselves, which can have negative consequences for the healthcare system. Healthcare insurance frauds are causing billions of dollars loss in healthcare funds around the world. By 2010, it costs up to 10% of total health care expenditure worldwide [4]. According to some reports, the US healthcare system losses around \$505 billion to \$850 billion every year. This percentage is from 9% to 19% of the total healthcare expenditure [5]. It can be easily seen that this additional burden leads to increased taxes and higher health insurance plans for individuals.

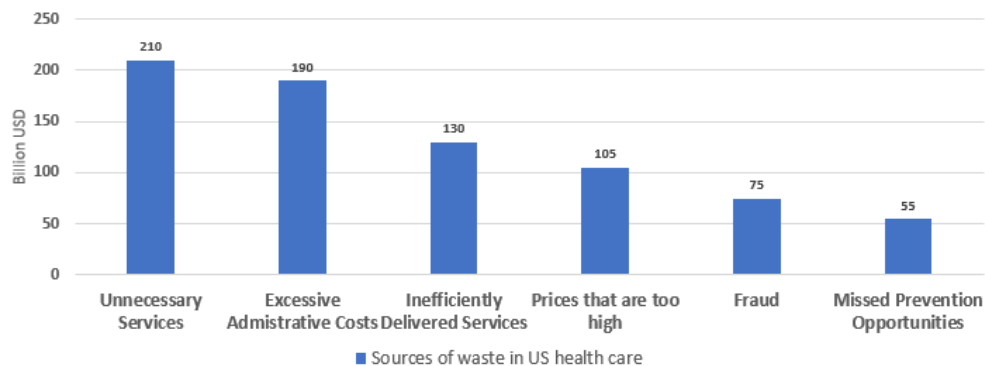


Figure 1.1: Average sources of waste in US health care every year

The Government Accountability Office of the United States (U.S. GAO) estimated over \$51 million Medicare beneficiaries in 2013 with services costing somewhere over \$600 billion. On the other side, it also costs \$50 billion in improper payments, including some basic technical and human errors, which there may be fraudulent cases [6]. According to RGA data, published in 2017, The countries of Australia, Singapore, Malaysia, Thailand, the Philippines, Viet-

nam, and Korea, as well as Japan and Indonesia also faces healthcare fraud issues [7]. As per limited figures, the European continent has at least €56 billion in losses annually over fraud practices [8]. The Swedish insurance industry pays out SEK 70 billion loss to its customers in more than \$3 million in claims; unfortunately, 5% of these payments turn out as fraudulent [9].

Insurance scams in the healthcare industry are resulting in losses of billions of dollars for public healthcare systems all over the world. As technology is getting advances on a day-to-day basis, it also generates more data. A huge amount of data can be found in the insurance providers' databases, and it continues to grow. Data mining techniques in combination with different analytical approaches i.e., machine learning techniques are today recognized as a key practice to identify fraud [10].

The figure 1.2 explains the most popular classification of the frauds in healthcare insurance system. Fraud can be identified through the services availing as well as providing patterns. Availing patterns such as repetition of services, age inconsistency, gender inconsistency, and visit frequency can leads towards fraud and waste of healthcare insurance. These patterns are performed by the patients. On the sides hospitals, providing patterns such as Billing, Unnecessary treatments, unnecessary procedures, charging multiple times, and misuse of credentials can leads towards fraud and abuse of system [11].

The healthcare insurance system involves three parties - the insured, medical institutions, and insurance providers, as in figure 1.3. Each party may have different interests that can lead to fraud, such as over-diagnosis and treatment from hospitals, fake medical treatment from insured individuals, and insufficient review of medical insurance settlement data from insurance providers. Medical fraud can cause significant losses to the insurance fund and threaten its normal operation. It is important to detect and prevent medical insurance fraud to ensure the normal operation of the insurance fund in the



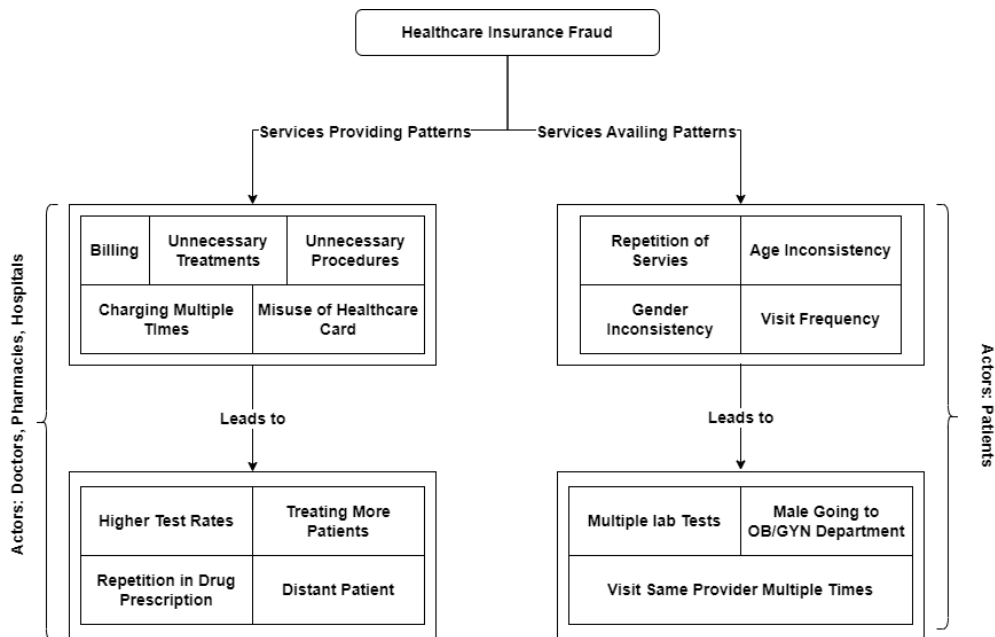


Figure 1.2: Healthcare insurance ecosystem involves patient, hospital, and services providers

long term. Measures should be taken to detect and report fraud, waste, and abuse in the system, including errors and abuse by providers, unnecessary costs to the payer, and exploitation of weaknesses in internal control mechanisms.

## 1.2 Fraud, Waste, and Abuse

”Fraud, Waste, and Abuse” (often abbreviated as ”FWA”) is a term used in the healthcare industry, including health insurance, to refer to practices that result in unnecessary or excessive healthcare costs, improper payments, or other fraudulent activities. Waste, abuse, and fraud in healthcare can result in substantial financial losses for insurance companies, which can drive up the cost of healthcare for everyone. To combat FWA in healthcare, insurance companies, regulators, and law enforcement agencies work to detect and prevent these activities, investigate potential cases, and prosecute those responsible for engaging in fraudulent activities.

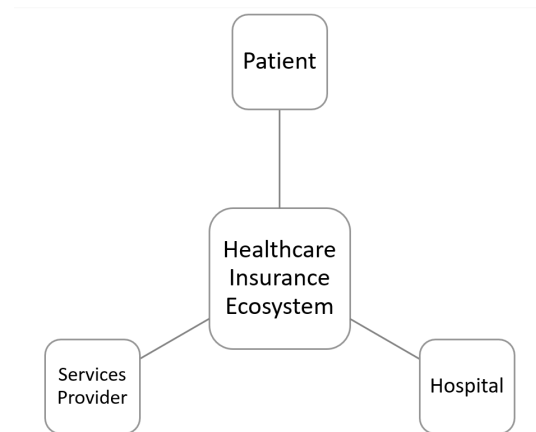


Figure 1.3: Healthcare insurance ecosystem involves patient, hospital, and services providers

### 1.2.1 Fraud

Fraud in Healthcare occurs when individuals or organizations intentionally deceive healthcare providers, insurance companies, or patients in order to gain some type of financial benefit. This can take many forms, including:

- Billing for services that were never performed: This occurs when a healthcare provider submits a claim for services or procedures that were never actually provided to the patient. For example, a provider may bill for a diagnostic test or procedure that was never performed, or bill for a longer visit than was actually spent with the patient.
- Submitting false claims: This involves submitting claims for services or procedures that are not medically necessary, or for which the provider is not qualified to perform. For example, a provider may submit a claim for a specialized surgery that they are not trained or authorized to perform.
- Using someone else insurance information: This involves using another person's insurance information to obtain healthcare services. For example, a person may use a family member's insurance information to obtain

prescription medications or medical procedures without their knowledge or consent.

These actions can result in improper payment or financial gain for the individuals or organizations involved, and can ultimately increase healthcare costs for patients and insurance providers. Healthcare fraud is a serious crime and can result in civil and criminal penalties, including fines, imprisonment, and exclusion from government healthcare programs.

### **1.2.2 Waste**

Waste in healthcare is a significant problem that can lead to unnecessary costs without providing any or way less benefit to patients. It refers to the overuse or misuse of healthcare resources, which can result in inefficient healthcare practices and poor patient outcomes. Examples of waste in healthcare include:

- Ordering unnecessary tests or procedures: This occurs when healthcare providers order tests or procedures that are not medically necessary, such as routine imaging studies or diagnostic tests that do not improve patient outcomes.
- Prescribing expensive brand-name drugs when generic alternatives are available: This involves prescribing more expensive medications when equally effective generic alternatives are available, which can result in higher costs for patients and insurance providers.
- Using higher-cost facilities for routine care: This involves using higher-cost facilities, such as hospitals, for routine care that could be provided in a lower-cost setting, such as a primary care clinic or urgent care center.

Such wastes in healthcare can contribute to rising healthcare costs and reduced access to care for patients. Addressing waste in healthcare is also

important for improving the efficiency and effectiveness of healthcare delivery, while also ensuring that patients receive the appropriate care they need. This can involve implementing strategies to reduce unnecessary testing and procedures, promoting the use of cost-effective medications, and encouraging the use of lower-cost healthcare facilities for routine care.

### **1.2.3 Abuse**

Abuse in healthcare is a significant problem that can lead to unnecessary costs and improper payments. It refers to actions that are inconsistent with accepted business or medical practices, and can result in fraudulent or unethical behavior by healthcare providers or organizations. Some of the many examples of abuse in healthcare are

- **Over-billing for services:** This occurs when healthcare providers bill for services at a higher rate than is allowed, or bill for services that were not actually provided. For example, a provider may bill for a more expensive procedure than was actually performed, or bill for multiple procedures when only one was performed.
- **Billing for services that were not actually provided:** This involves billing for services that were not actually provided to the patient, such as diagnostic tests or procedures that were not performed, or medication that was not actually administered.

These examples of abuse in healthcare can lead to fraudulent or unethical behavior, resulting in unnecessary costs and improper payments. Addressing abuse in healthcare is important for improving the integrity of healthcare delivery, while also ensuring that patients receive the appropriate care they need. This can involve implementing strategies to detect and prevent abuse, such as conducting regular audits of billing practices, implementing fraud detection

software, or establishing clear policies and procedures for billing and reimbursement.

Healthcare insurance fraud is a significant problem in the healthcare industry resulting in billions of dollars in losses each year. Fraudulent activities can occur in various forms, such as billing for the services that were never rendered, over-billing, falsifying medical records, kickbacks, and many others. Such fraudulent activities can cause substantial harm to insurance companies, policyholders, and healthcare providers. We can use healthcare insurance data which is being generated on very large level and is not labelled. It will be difficult to apply well established classification techniques which requires labelled data; and annotation itself requires a lot of time and finances. Therefore, there is a need to design and develop effective unsupervised learning-based technique that can help detect and prevent health insurance fraud, provide actionable insights to relevant stakeholders, such as hospital and insurance providers.

## CHAPTER 2

### RELATED WORK

Healthcare insurance fraud is a pervasive issue that affects countries worldwide, particularly those that have implemented healthcare insurance systems for their citizens. Previous research has predominantly focused on general insurance fraud, employing data mining and machine learning techniques for detection and prevention [12]. Given the scarcity of studies specifically addressing healthcare insurance fraud, this thesis aims to concentrate on the literature that advocates for the application of unsupervised learning techniques in detecting fraudulent activities within the healthcare insurance sector.

Association rule mining is not yet widely researched in the area of healthcare insurance or for any other fraudulent activities. Although it is a widely used data mining technique but still carrying some drawbacks, Yadav et al. discussed some techniques which can help improve the algorithm [13]. Saba et al. share the initial stage of the study, by using the association rule followed by the SVM classification algorithm, they believe their model can address the discrepancies and this reduce fraud in health insurance [14]. Sornalakshmi et al. presented the new technique by combining the MapReduce and Apriori association rule mining. MapReduce makes parallel computing very easy. However, the author believes Apriori algorithm needs to be fully implemented, as there is a lot of improvement needed in Apriori algorithms for parallel and distributed terms [15]. Authors of [16] used the algorithm in medical billing

also believes that Apriori algorithm is good for finding frequent item-sets from billing database.

Data mining helps detect and prevent insurance fraud. Anomaly detection, clustering, and classification can detect fraudulent insurance claims [17]. After finding anomalous claims, further investigation can be required to narrow the focus and identify fraud patterns. Kirlidogab and Asukb [18] used longitudinal data of nine years but also suggest one-year analyses which can be beneficial for detecting "hit and run" frauds that are hard to detect over long periods. Gao [19] proposed the SSIsomap activity clustering method, SimLOF outlier detection method, and the Dempster-Shafer Theory-based evidence aggregation method to detect the unusual categories and frequencies of behaviours simultaneously. Alwan [20] shows how combining machine learning techniques with existing methods for detecting fraud can make it easier to find fraud. Specifically, the paper examines the effectiveness of several data mining techniques, including Decision Tree, Support Vector Machine, K-Nearest Neighbor, and Hidden Markov Model, in detecting credit card fraud. The findings highlight the potential of a hybrid approach that integrates these methods to enhance fraud detection.

Shang [21] suggested the use of One Class Support Vector Machine (OCSVM) for the intrusion detection. Authors describe that OCSVM in anomalies detection fields have advantages, such as fast and strong generalization ability, the less support vector, the simple model, and the great practical value [22]. Maglaras [23] combined the ensemble methods and social networking metrics for the enhancement of the OCSVM, but it needs the improvement in order to decrease false positive and increase detection accuracy. Maglaras [24] developed using an OCSVM classifier and recursive k-means clustering. It is trained offline using network traces, and only severe alerts are communicated to the system. The module is part of an IDS system developed under CoCkpitCI, and its performance is stable and not affected by the selection of

parameters  $\nu$  and  $\sigma$ . However, author believes further evaluation is needed to determine its effectiveness under different anomalies scenarios. Wang [25] proposes an improved particle swarm optimization algorithm to enhance the accuracy of the OCSVM-based power system anomaly detection. The algorithm introduces an adaptive learning factor and splitting and elimination mechanism to improve the population’s diversity and fine searching ability. Amer [26] proposed SVM-based algorithms are effective for unsupervised anomaly detection, outperforming clustering-based methods in most cases. The proposed eta one-class SVM produces the most promising results, with a sparse solution and superior AUC performance. The introduced outlier score calculation method allows for ranking of outliers, providing practical value for anomaly detection applications.

In 2008, Fei Tony Liu and Zhi-Hua Zhou developed an algorithm called the Isolation Forest [27] with the purpose of finding anomalies in data. This particular algorithm makes use of binary trees in order to identify anomalies, and because of its linear time complexity and low memory requirements, it is well suited for the processing of large amounts of data. Isolation Forest algorithm’s low accuracy, execution efficiency, and generalization ability are addressed by Xu’s SAiForest data anomaly detection method [28]. SAiForest optimises the forest by selecting isolation trees with high abnormality detection and difference using simulated annealing and selective integration based on precision and difference value. Cheng [29] proposes the union of Isolation Forest and Local Outlier Factor to detect outliers in multiple datasets. The algorithm calculates each data point’s anomaly score using binary trees and prunes normal samples to find outlier candidates. The proposed method addresses Isolation Forest’s local outlier issues and reduces Local Outlier Factor’s time complexity. Ding [30] proposes an iForest-based anomaly detection framework under the sliding windows framework iForestASD, for streaming data. Four real-world data sets show that proposed method is efficient. Au-



thors believes there is still a lot improvement required in the algorithm, such as defining the threshold and size of sliding window. Lesouple [31] introduced generalized isolation forest for anomaly detection. Although it achieved the less execution time but the false alarm rate is high. In future, authors aim to reduce the false alarm rate and improve anomaly detection.

Cluster Based Local Outlier Factor (CBLOF) was proposed by He et al. in 2022 [32]. It is generally used for outlier detection that considers a combination of local distances to nearby clusters and the size of those clusters. It identifies anomalies as data points that are located in small clusters next to a larger nearby cluster. Such outliers may not be single points but instead, small groups of isolated points. John [33] explained the workings of Local Outlier Factor and Isolation Forest and suggested its use for identification of credit card fraud with the accuracy of 97% and 76% respectively. Kanyama [34] used K-Nearest Neighbor (k-NN), CBLOF, and histogram-based outlier score (HBOS) for anomaly detection in smart water metering networks. After the experimentation, authors believes that CBLOF performs better than KNN in terms of detection rates, but KNN achieved almost zero in terms of False Positive Rate. Irfan [35] performed an experiment for the evaluation of the performance of three unsupervised outlier detection algorithm such as K-Means, LOF, and CBLOF. Authour states that the CBLOF performed better than its competitors, CBLOF was faster in terms of computational complexity. Author recommended to restart the K-Means algorithm multiple times for stable cluster results, but CBLOF may be preferable for applications where processing speed or updating clustered models in streaming data is important. In another experiment Irfan [36] applied the methodology for churn prediction in banking system and came up with the same results in favor of CBLOF.

The main goal is to find the most important features and data sources, such as medical records, billing details, and demographic information, for using unsupervised learning techniques to find health insurance fraud. We also want

to come up with ways to find complicated fraud schemes that involve more than one person, like when healthcare providers, patients, and services work together to commit fraud. We will assess the influence of data preprocessing approaches, like normalization, feature scaling, and missing data imputation, on the accuracy and resilience of fraud detection models. Moreover, we will investigate the potential of ensemble methods, combining multiple unsupervised learning models to enhance accuracy and generalization, and evaluate the performance of various unsupervised learning algorithms in detecting health insurance fraud. One of the main reason of using the unsupervised techniques are the lack of labelled data. Mostly data available in the insurance provider and hospital databases is unlabelled. There are multiple supervised techniques available which can also guarantee better results but data labelling requires huge amount of time and finances. It is of best interest on current circumstances to apply unsupervised techniques.

Overall, less amount of work has been done in the domain of health-care insurance fraud detection. Researchers has mostly focused on one of the stakeholder of insurance triangle either on the frauds done by patients or by hospitals. In this research, we will focus on all stakeholders in insurance triangle for the better identification of frauds commits across the board.

## CHAPTER 3

### MATERIALS & METHODS

#### 3.1 Dataset

In this study, we made use of the CMS Linkable 2008–2010 Medicare Data Entrepreneurs’ Synthetic Public Use File (DE-synPUF). The claims made by Medicare recipients and a random sample of five percent of those beneficiaries from 2008 to 2010 are included in the dataset. The data have been thoroughly anonymized, which means that none of the beneficiaries included in the DE-SynPUF actually receive Medicare benefits; however, they are all intended to stand in for real beneficiaries. The Centers for Medicare and Medicaid Services (CMS) made twenty random sample files available for us to use in our experiments, and we decided to use the inpatient dataset from subsample 1 of those files. However, there is nothing that restricts us to using only this one sample or using multiple samples at the same time. But in terms of solid reasons, some studies have suggested that inpatient fraud may be more prevalent than outpatient fraud. One of the possible explanation for this is that inpatient care tends to be more expensive than outpatient care, which means that there is a greater potential for fraudulent activity to generate large profits. Additionally, inpatient care may involve more complex procedure and treatments, which can be easier to overbill or manipulate as compare to simpler outpatient services. The selection of this particular method for validating our proposed methodology was completely arbitrary, and in the near future we

want to add more samples to our dataset. Figure 1 presents the database structure, which lists all of the tables and characteristics that are currently accessible.

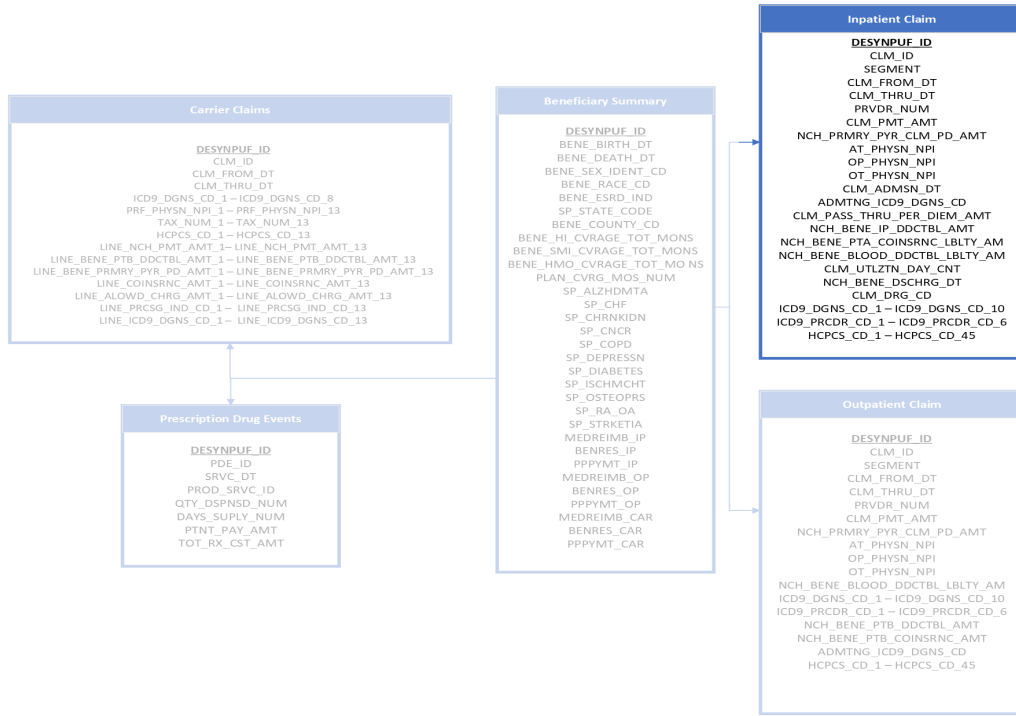


Figure 3.1: All the table and their attributes of DE-SynPUF

Our dataset consists of 81 variables: The beneficiary code (DESYNPUF\_ID) identifies each beneficiary in the dataset, while the claim ID distinguishes claims for the same beneficiary. A record’s claim line section identifies its claim component. The start and end dates indicate the claim period. The provider institution is the medical facility that performed the service, and the claim payment amount is the total amount paid. Attending, operating, and other physician NPI numbers identify service providers. The inpatient admission and discharge dates show when the beneficiary was hospitalised. Diagnosis and procedure codes define illnesses and treatments. Lastly, the revenue centre HCFA common procedure coding system classifies medical services.

### 3.2 Proposed Methodology

The methodology for this research is based on unsupervised learning techniques that aim to discover patterns, anomalies, and relationships within a given dataset without the need for prior labeled data. Unsupervised learning is a powerful approach for analyzing large and complex datasets, where manual labeling is often infeasible or impractical. The key advantage of unsupervised learning is its ability to identify hidden structures and correlations within data, which can provide valuable insights for data-driven decision-making. One area where unsupervised learning techniques can be particularly useful is in healthcare insurance fraud detection. Insurance fraud is a significant problem for the healthcare industry, resulting in significant financial losses and potentially endangering patient health.

The unsupervised learning techniques used in this research include Apriori, Isolation Forest, One-Class SVM (OCSVM), Clustering-based Local Outlier Factor (CBLOF), and Ensemble Correlation-Based Outlier Detection (ECOD). Apriori is a well-known algorithm for mining frequent itemsets and association rules, which is used to identify patterns and relationships between different items in a dataset. Isolation Forest is a tree-based algorithm that partitions the dataset into isolated subspaces, which is used to detect anomalies and outliers. OCSVM is a support vector machine-based algorithm that creates a boundary around the normal data points, which is used to identify anomalous data points that fall outside the boundary. CBLOF is a clustering-based approach that uses k-means clustering to identify local outlier factors, which is used to identify anomalous clusters. ECOD is an ensemble method that combines multiple correlation-based outlier detection methods, which is used to identify anomalous data points that are consistent across multiple methods.

When the model is used for unsupervised learning, it is extremely difficult to evaluate, and when there is no ground truth, it is even more complicated.

There are several common methods for evaluating unsupervised learning models, but there is no universal method. Several factors, such as the type of unsupervised learning model employed, the nature of the data being analyzed, and the nature of the problem being solved, can impact the selection of an evaluation method.

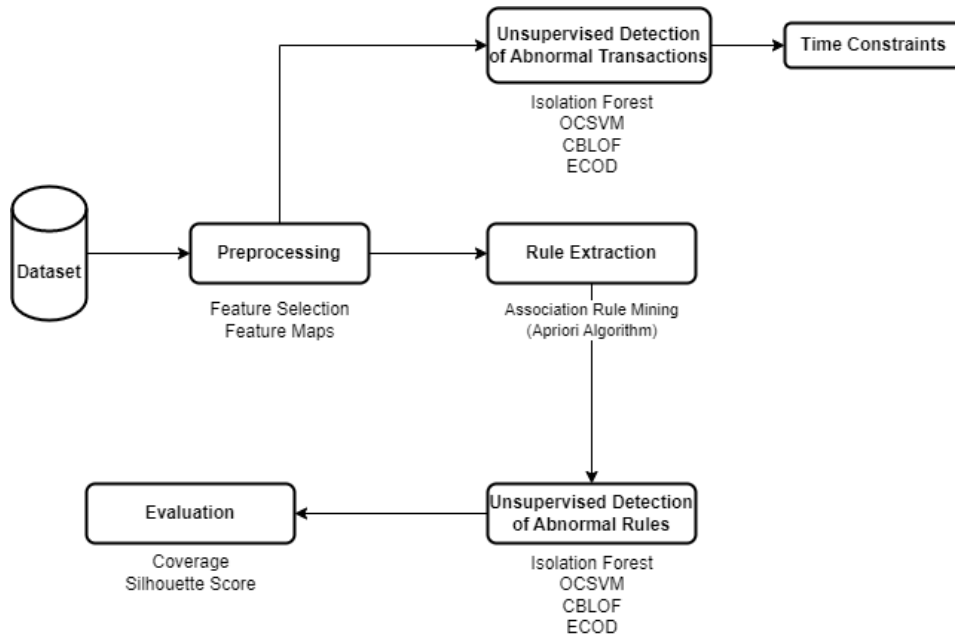


Figure 3.2: Proposed Ensemble Technique

We also conducted another experiment by passing the preprocessed values directly to every anomaly detection classifiers for checking whether the transaction is fraudulent or not. The classifier returns the status of every transactions. The main constraints we faced in this experiment is the time complexity. It was taking more time than the other technique.

### 3.2.1 Apriori Algorithm

Agrawal and Srikant proposed the Apriori algorithm in 1994, which has become a widely used data mining algorithm for identifying frequent item sets in a transaction database [37]. In the field of association rule mining, the

Apriori algorithm is recognized as one of the most well-known algorithms [38]. However, it may not be the optimal choice for detecting anomalies or fraudulent transactions in a database. This is because it is commonly assumed that fraudulent transactions are significantly fewer than normal ones. Therefore, when implementing Apriori, it is expected that the algorithm will generate rules based on normal transactions.

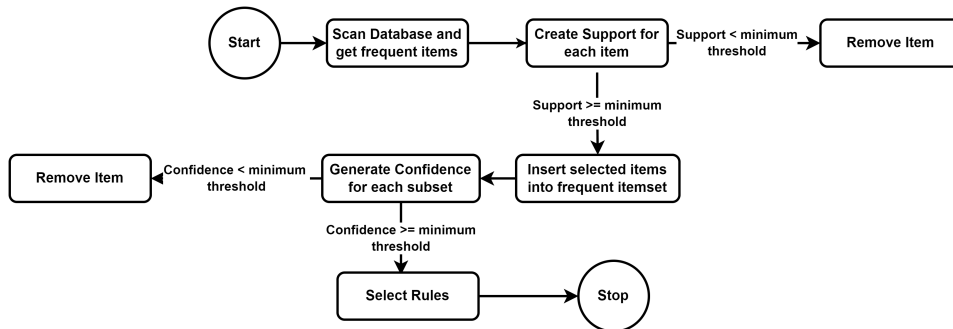


Figure 3.3: Association Rule Mining Process Flow Chart

Apriori algorithm works in two steps for association rule mining. The first step is to find all the frequently occurring item sets from the data. Generating association rules from the set of frequently occurring items is done in the second step. The second step is easy to achieve as compare to the first step [39].

### 3.2.2 Isolation Forest

Isolation Forest was introduced at Lie et al [27] in 2008. Generally, it is designed to detect anomalies from structured data. The iTree, or isolation tree, is a binary tree data structure in which each node corresponds to a subset of data objects. The tree is constructed by randomly sub sampling a subset of  $n$  data objects from the entire dataset and using it as the data pool for the root node. The tree grows by recursively partitioning the data objects in the leaf node into two child nodes, until a single data object remains in the node or the maximum depth limit is reached. The branching criterion for each data object

is determined by comparing a randomly selected feature of the data object to a split value within the range of that feature's values. The path length of a data object in the iTree serves as an indication of the object's abnormal degree. An iForest, or isolation forest, is constructed by creating multiple iTrees, and the anomaly score of a data object is calculated by averaging the path lengths of that object across all iTrees in the forest. The final anomaly score is then normalized using a factor.

Isolation Forest consists of two steps, training and testing phase. In training, the algorithm builds an ensemble of isolation trees, known as iTrees, as shown in fig 3.4. Each tree is built through algorithm. By default 100 iTrees are built in an IForest but changes can be made in experiments for obtaining the best results.

---

**Algorithm 1 Building a Decision Tree**

---

```

1: procedure BUILD ISOLATION TREE
2:   Draw a uniform sample from the data
3:   Select a splitting point  $p$  randomly and an attribute  $q$ 
4:   Divide the data sample
5:   while not at predefined tree height or all data points in sample have
      same value or only one data point remains in sample to split on do
6:     Repeat steps 3 and 4
7:   end while
8: end procedure

```

---

In the next step of IF algorithm, each data point is passed through each built iTree to calculate its corresponding anomaly score  $a(x)$  from 0 to 1. Labels are assigned based on their respective data point's scores. Specifically, those with scores below 0.5 are classified as normal and receive a label of 1. On the other hand, data points with scores that are closer to 1 are deemed as potential anomalies and thus labeled with a value of -1.

Anomalies are detected through

$$a(x, m) = 2 \frac{-E(h(x))}{k(m)} \quad (3.1)$$



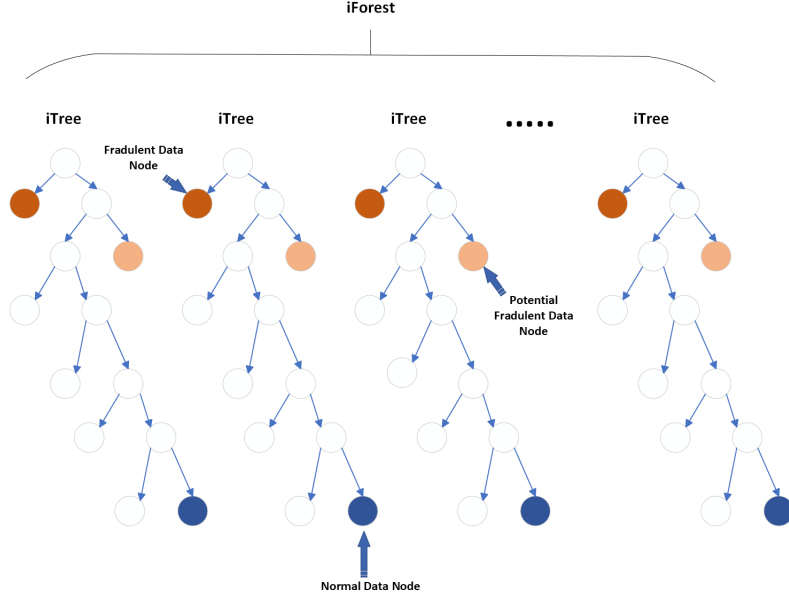


Figure 3.4: Isolation Forest

where  $c(m)$  is a normalization constant for a data set of size  $n$ . The expression  $E(h(x))$  represents the expected or “average” value of this path length across all the Isolation Trees. The expression  $k(m)$  represents the average value of  $h(x)$  given a sample size of  $m$  and is defined using the following equation. Following equation illustrates the formula of the constant  $k(m)$ .

$$c(m) = \begin{cases} 2H(m-1) - \frac{2(m-1)}{m} & : \text{for } m > 2 \\ 1 & : \text{for } m = 2 \\ 0 & : \text{otherwise} \end{cases} \quad (3.2)$$

where  $H$  is the harmonic number, which can be estimated by  $H(i) = \ln(i) + \gamma$ , where  $\gamma = 0.5772156649$  is the Euler-Mascheroni constant.

### 3.2.3 Cluster Based Local Outlier Factor

Cluster-Based Local Outlier Factor (CBLOF) was proposed by He et al [32] in 2002. The CBLOF definition of anomalies takes into account both the

local distances to neighbouring clusters as well as the sizes of the clusters to which the data point belongs. Algorithm first cluster next to a nearby large cluster are identified as outliers. The Local outliers may not be a singular point, but a small group of isolated points.

In general, the procedure of CBLOF can be describe in the three steps. Initially, a data point is assigned to one and only one cluster. K-means is commonly used as clusteric algorithm for CBLOF. Next, CBLOF ranks clusters according to the cluster size from large to small and get the cumulative data counts. Clusters that holds 90% of the data are considered as "large" clusters rest of them are consider as "small" clusters. The threshold of 0.9 can be fine-tuned as per requirement. Lastly, the outlier detection process involves the calculation of the distance of a data point to the centroid and its corresponding outlier score. For data points belonging to a large cluster, the distance is calculated as the distance from the data point to the centroid of its cluster. The outlier score is then determined as the product of this distance and the number of data points in the cluster. For the smaller clusters the distance is the distance from the data point to the centroid of the nearest large cluster. The outlier score for these data points is determined as the product of this distance and the size of the small cluster to which the data point belongs.

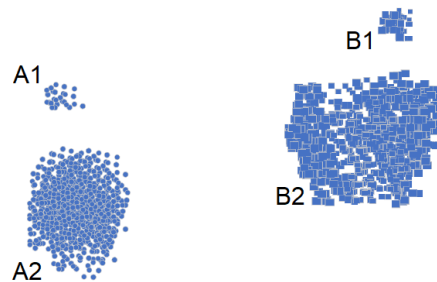


Figure 3.5: Cluster Based Local Outlier Factor

As it can be seen in figure 3, clusters A1 and B1 are the smaller clusters and A2, and B2 are large cluster. A1 and B1 will be considered as outlier as

they do not belong to any of the large clusters A2 and B2. According to the local neighborhood, data in cluster A1 is local outliers to A2, and same with B1 for B2.

### 3.2.4 One-Class Support Vector Machine

An unsupervised learning technique, One-Class Support Vector Machine (OCSVM) is used for outlier detection and constituting an incremental learning process. Its application in Anomaly Detection is widely used around the world such as Outlier Detection, Novelty Detection, and many others. OCSVM is modified to be a single-class learner from SVM that tries to find a hyper-sphere among the instances of the normal classes. This model classifies new data as normal or abnormal, all observations inside the hyper-sphere are normal and those outside the hyper-sphere are abnormal or anomalies.

Let us first examine the conventional two-class support vector machine. Consider a dataset with two-dimensional space  $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ ; points  $x_i \in \mathbf{R}^d$  where  $x_i$  is the  $i$ -th input data point and  $y_i \in \{-1, 1\}$  is the  $i$ -th output pattern, indicating the class membership.

A significant advantage of support vector machines (SVMs) is their capability of generating a non-linear decision boundary by transforming the data through a non-linear mapping  $\phi$  to a higher-dimensional feature space  $F$ . In this feature space, it may be possible to separate the classes with a hyperplane, even if a linear boundary is not feasible in the original input space  $I$ . This process results in a non-linear curve in the input space when the hyperplane is projected back. By utilizing a polynomial kernel for the projection, all the dots are elevated to the third dimension, and a hyperplane can be employed for separation. When the plane's intersection with the space is projected back to the two-dimensional space, it results in a circular boundary.

The hyperplane that separates the classes in an SVM is represented by the equation  $w^T x + b = 0$ , where  $w$  is a vector in the feature space  $F$  and

$b$  is a scalar in  $\mathbf{R}$ . The margin between the classes is determined by this hyperplane, with all data points belonging to class  $-1$  on one side and all data points belonging to class  $1$  on the other. The hyperplane aims to maximize the distance between the closest data points from each class to itself, thus achieving the maximum margin or "separating power."

To address the issue of over fitting in the presence of noisy data, slack variables  $\xi_i$  are introduced to permit some data points to lie within the margin. The trade-off between maximizing the margin and accommodating training errors is controlled by the constant  $C > 0$ . The SVM classifier's objective function is a minimization formulation that balances these factors.

$$\min_{w,b,\xi_i} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i \quad (3.3)$$

$$\begin{aligned} \text{subject to: } & y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i, \quad \text{where } i = 1, \dots, n \\ & \xi_i \geq 0, \quad \text{where } i = 1, \dots, n \end{aligned}$$

According to Scholkopf et al [40], separates all the data points from the origin in the feature space  $F$  and maximizes the distance from hyperplane to the origin. This result in a binary function which returns  $+1$  in a "smaller" region and  $-1$  elsewhere.

$$\min_{w,\xi_i,\rho} \frac{1}{2} \|w\|^2 + \frac{1}{\nu n} \sum_{i=1}^n \xi_i - \rho \quad (3.4)$$

$$\begin{aligned} \text{subject to: } & (w \cdot \phi(x_i)) \geq \rho - \xi_i, \quad \text{where } i = 1, \dots, n \\ & \xi_i \geq 0, \quad \text{where } i = 1, \dots, n \end{aligned}$$

By using Lagrange techniques and using a kernel function for the dot product calculations, the decision function becomes:

$$f(x) = \text{sgn}((w \cdot \phi(x)) - \rho) = \text{sgn}\left(\sum_{i=1}^n \alpha_i K(x, x_i) - \rho\right) \quad (3.5)$$

### 3.2.5 Empirical Cumulative distribution based Outlier Detection

The Empirical Cumulative Distribution-based Outlier Detection (ECOD) method has several advantageous attributes that distinguish it from alternative algorithms. ECOD is unique in its lack of dependence on hyperparameters, its computational efficiency and swiftness, and its ease of interpretation and comprehension. The ECOD approach leverages information regarding the distribution of data to identify points that deviate significantly from the majority, thus indicating their outlier status. The ECOD technique calculates the tail probability of each variable using univariate Empirical Cumulative Distribution Functions ( $\delta$ ) and combines these probabilities through multiplication.

Detection of the anomalies through ECOD is done through the computation of three values. ECDfs are used to generate the left- and right-tail probability values,

1. *O-left = Sum of the negative log of the left-tail probability of every variable*
2. *O-right = Sum of the negative log of the right-tail probability of every variable*
3. *O-auto = Sum of left- or right-tail probability of every variable, depending on whether it is left- or right skewed*

Final outlier score of an observation is obtained through taking the extreme negative log probability score.

$$\text{Outlier Score} = \max(O_{\text{left}}, O_{\text{right}}, O_{\text{auto}}) \quad (3.6)$$

For mathematically-inclined, following are simplified formulations of the three equations describe above

$$O_{left} = - \sum_{j=1}^d \log(\delta_{left}^j(X^j)) \quad (3.7)$$

$$O_{right} = - \sum_{j=1}^d \log(\delta_{right}^j(X^j)) \quad (3.8)$$

$$O_{auto} = - \sum_{j=1}^d \begin{cases} \log(\delta_{left}^j(X^j)) & \text{if } \gamma_j < 0 \\ \log(\delta_{right}^j(X^j)) & \text{if } \gamma_j \geq 0 \end{cases} \quad (3.9)$$

where  $\gamma_j$  is the skewness coefficient

### 3.3 Evaluation

Evaluation of the model becomes extremely challenging when unsupervised learning is involved, and it becomes even more intriguing when there is no ground truth. Although there are some common evaluation techniques for unsupervised learning models, there is no universal method for evaluating these models. Several factors, such as the type of unsupervised learning model being used, the nature of the data being analysed, and the specific problem being solved, can influence the choice of evaluation method.

Selected methodology started with Apriori association rule mining algorithm which is a popular data mining technique used to extract interesting relationships or associations between items in large databases. The goal of the association rule mining to find a set of rules that have high support and confidence values. When performing association rule mining, it is important to filter the mined association rules by using statistical indicators such as support, confidence, and lift ratio. Support is an indicator of how often an association rule appears in the dataset, while confidence is an indicator of how reliable a

rule calculation is. The lift ratio is an indicator of the strength of the dependence between the antecedents and consequences of the association rule. Only association rules whose support, confidence, and lift ratios are greater than the corresponding thresholds are used for further analysis.

$$support(A \rightarrow B) = P(A \cup B) \quad (3.10)$$

$$confidence(A \rightarrow B) = P(B/A) = \frac{P(A \cup B)}{P(A) \times P(B)} \quad (3.11)$$

$$lift(A \rightarrow B) = \frac{confidence(A \rightarrow B)}{support(B)} \quad (3.12)$$

$$leverage(A \rightarrow B) = support(A \rightarrow B) - (support(A) \times support(B)) \quad (3.13)$$

Where  $A$  and  $B$  are the itemset occurring in the database.

Support refers to the frequency of an itemset in the database, while confidence measure the strength of the association between two itemsets. However, existing measures only evaluate the quality of the resulted rules separately, missing the different dependencies between the rules. For the evaluation of the of the rule we calculated the coverage of the rule,

$$Cover(Rule) = \frac{1}{k} \sum_{r_j \in R, i \neq j} Distance(r_i, r_j) \quad (3.14)$$

Here,  $Cover(Rule)$  measures the average distance of every rule with other rules.

In order to evaluate the result of the fraudulent rules, which were obtained through Isolation Forest, CBLOF, ECOD, and OCSVM, there are a variety of validity metrics were proposed in the past but most popular is Silhouette Score [41]. The silhouette coefficient is calculated by taking into account the mean intra-cluster distance  $a$  and the mean nearest-cluster distance

$b$  for each data point i.e.  $(b - a)/\max(a, b)$  [42]. A silhouette score near +1 indicates correct cluster, near 0 suggests possible alternative cluster, and near -1 indicates wrong cluster.



## CHAPTER 4

### RESULTS & ANALYSIS

This section presents the results and discussion of our research on health-care insurance fraud detection using data mining techniques. The study utilizes the open-source CMS 2008-2010 DE-SynPUF dataset, which is preprocessed by removing less important features and encoding the data.

#### 4.1 Descriptive Analysis

To get a clear and concise summary of the data, we analyze, summarize, and interpret it. The descriptive analysis allows us to identify patterns, trends, and relationships in the data, which assists us in drawing important conclusions and making informed decisions. By utilizing descriptive statistics and visualizations, we aim to offer a comprehensive overview of the data and pinpoint key insights that can be used to guide future research or decision-making.

We chose to work with the inpatient dataset for our analysis. This dataset comprises a total of 66,773 insurance claim records. To streamline our analysis, we decided to exclude features related to the Health Care Common Procedure Coding System (HCPCS). These codes represent procedures, supplies, products, and services that may be provided to Medicare beneficiaries and individuals enrolled in private health insurance programs. However, we believe that such information can be obtained through other features in the dataset.

By removing these features, we aim to focus our analysis on the most relevant and informative data points in the inpatient dataset.

As shown in Figure 4.1, the provider institution count is analyzed using a bar chart. The dataset contains 2675 unique provider institutions, with 50% of the total occurring less than 10 times in the complete dataset. We found out that provider institution "23006G" occurs in 772 records. The top 20 most-occurring institution providers share the count of 7524 transactions. 209 provider institutions were only seen once in the complete dataset. Overall, we can infer that the dataset contains a large number of unique provider institutions, but the majority of these institutions occur very few times in the dataset. Additionally, there are a small number of provider institutions that occur frequently, with the top 20 accounting for a significant proportion of the transactions. Finally, a substantial proportion of the provider institutions in the dataset were only seen once. This information could be used to inform further analysis of the dataset, such as identifying outliers or patterns in the data.

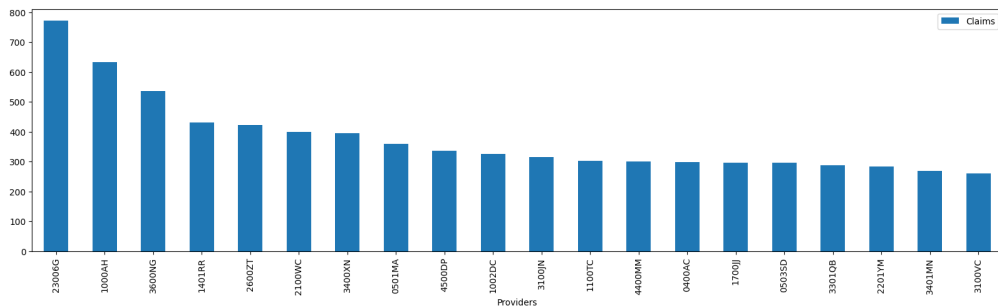


Figure 4.1: Provider institutions occurrence in entire dataset

The Figure 4.2 shows the bar chart of attending physicians. The attending physician is a medical doctor who is responsible for the overall care of a patient in a hospital or clinic setting. An attending physician may also supervise and teach medical students, interns, and residents involved in the patient's care. DE-synPUF dataset contains total number of 16670 unique

attending physicians while 75% of the physicians occurred only once or twice. Attending Physician with id '9011551271' occurred most of the time with the count of 533. Top 20 most occurred attending physician shares the count of 5675 transactions. Overall, this information suggests that there is a large degree of variation in the frequency of attending physicians in the DE-synPUF dataset. While a small number of physicians occur frequently, the majority occur infrequently, which may have implications for analysis of the data.

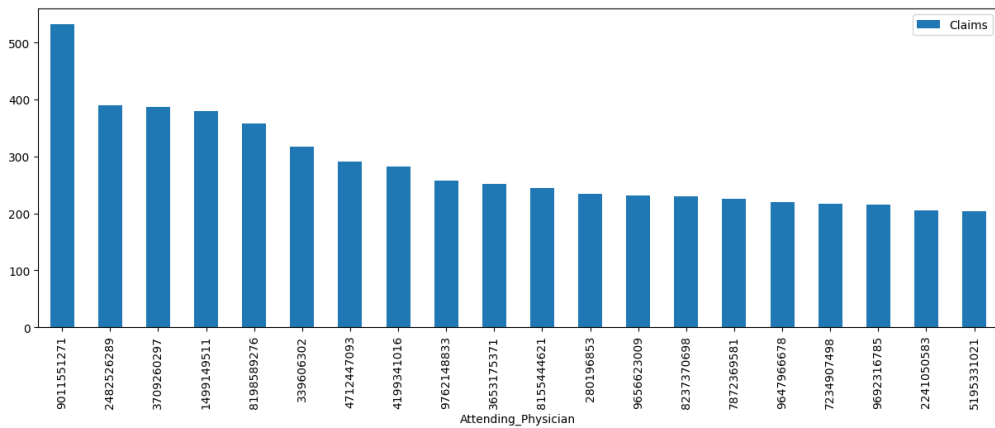


Figure 4.2: Attending physicians occurrences in entire dataset

The following Figure 4.3 shows the occurrence of top 20 operating physicians. The term operating physician refers to a physician (e.g., surgeon) who performs an operative procedure in the medical centre and who has the responsibilities outlined in the medical staff rules and regulations. The dataset contains 12076 unique numbers of operating physicians, while 75% of the physicians occurred only once or twice. Operating Physician with id '9612910514' occurred most of the time with a count of 324. The top 30 most frequent attending physicians shared the count of 4377 transactions. This information suggests that there is a large degree of variation in the frequency of operating physicians in the dataset. While a small number of physicians occur frequently, the majority occur infrequently, which may have implications for analysis of the data.

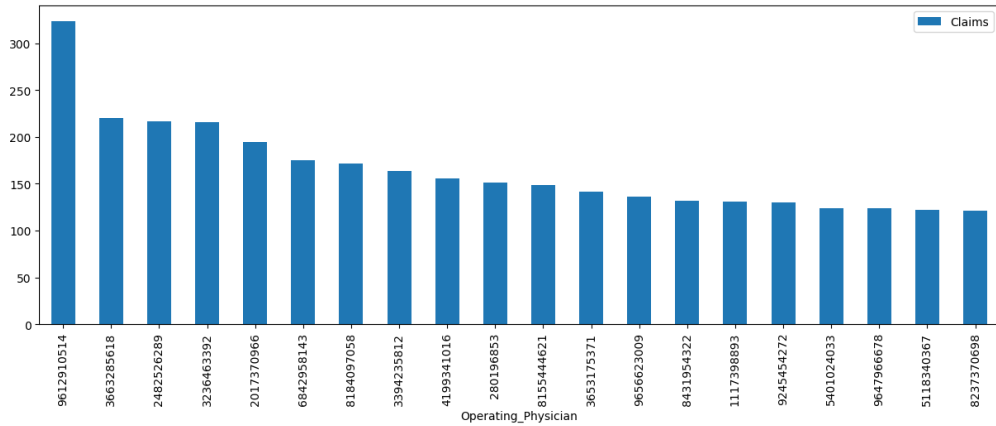


Figure 4.3: Operating physicians occurrences in entire dataset

Upon comparing the features of attending physicians and operating physicians, it can be seen in Figure 4.4 that 26.7% of the physicians were found in both features.

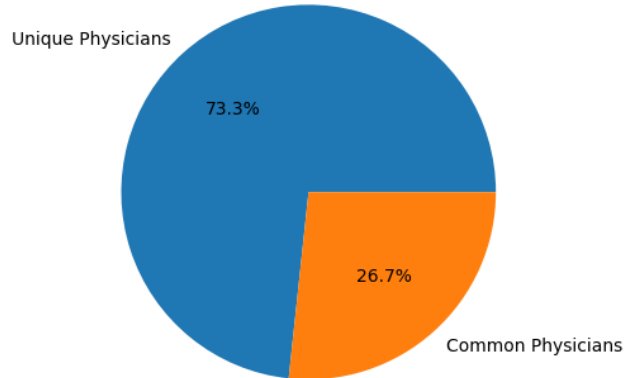


Figure 4.4: Unique & common physicians in attending physicians and operating physicians

In terms of features related to diagnosis codes, the dataset contains 5357 unique diagnosis codes. 50% of the diagnosis codes appeared in fewer than seven transactions. The diagnosis codes are present under the mapping of ICD-9 coding. Diagnosis code '4019' appeared, the most, 23512 times, and

referred to hypertension. Hypertension is also known as high blood pressure. The second most frequently occurring diagnosis code is '25000' which is commonly known as diabetes mellitus without mention of complication, type II or unspecified type, not stated as uncontrolled. Most of them are not much serious diseases, and there are a lot other examples available in the dataset. Figure 4.5 refers to the top 20 most frequent diagnoses in the transactions.

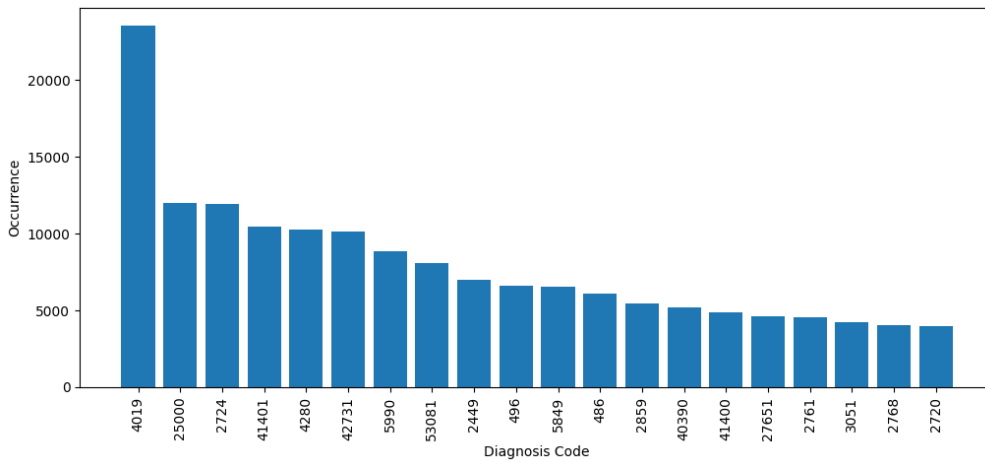


Figure 4.5: All diagnosis codes with their occurrences in the dataset, 4019 occurred the most

As we looked at the procedure codes in the dataset, we noticed a strange pattern: the procedure codes seemed to be the same as the diagnosis codes. Specifically, we noted that certain diagnosis codes, such as 4019 and 2724, were present in the Figure 4.6 we analyzed. In order to learn more about this, we compared the procedure codes in all six features to the diagnosis codes. This gave us some interesting information. A detailed breakdown of the results of this analysis can be found in the table provided 4.1. Except the feature procedure code 1, all of the other features has up to 35% same codes as diagnosis codes.

Figure 4.7 can be referred as overall summary overall summary of finding the common codes between diagnosis and procedure codes. Feature Proce-

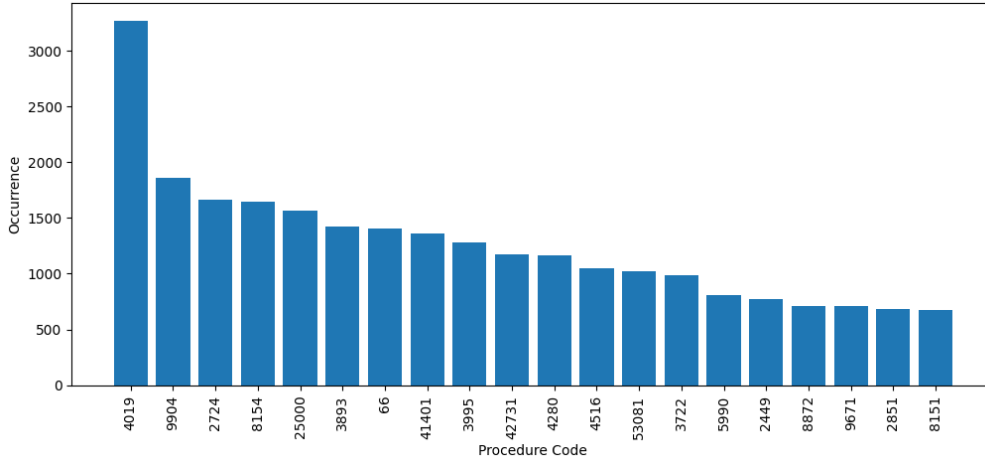


Figure 4.6: All procedure codes with their occurrences in dataset

Table 4.1: Procedure Codes Similarity Check W.R.T. Diagnosis Codes

Feature	Unique %	Common %
Procedure Code 1	95.4%	4.6%
Procedure Code 2	64.2%	35.8%
Procedure Code 3	69.6%	30.4%
Procedure Code 4	75%	25%
Procedure Code 5	79.3%	20.7%
Procedure Code 6	83.0%	17.0%

procedure\_Code\_1 has more than 95% of the procedure codes and rest of the features only have around 50% and also contains diagnosis codes.

## 4.2 Preprocessing

Healthcare insurance fraud is widespread problem and can be perpetrated through various means, including upcoding, misrepresenting procedures

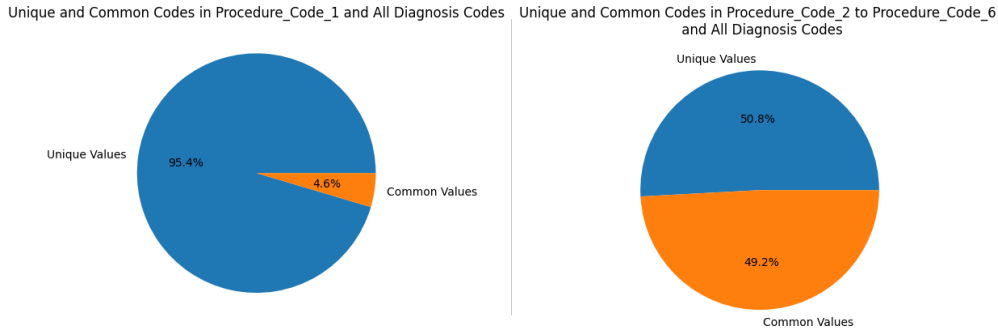


Figure 4.7: Codes Comparison Summary

to obtain payment for non-covered services, over billing, waiving patient co-pays or deductibles, and forging or altering medical bills or receipts. Identity theft is also a common way to commit health insurance fraud [43][44]. Insurance fraud are often performed through the partnership of the services provider, patients, and hospital. Fraudster play through the technicalities of billing, unnecessary treatments and unnecessary procedures in order to get unjust benefits [11]. In an initial part of the preprocessing, we removed all of the less important features from our dataset and kept only those who plays an important part in the fraudulent activities. The reason of the selection of features depends on the specific research question and the goals of the analysis. In this case, we have chosen these nine features based on their relevance to the research question and their potential to provide insight into the relationships or patterns of interest. For example, Provider\_Institution could be relevant to understanding the quality of care provided to beneficiaries, while Claim\_Payment\_Amount could be important for investigating the financial implications of Medicare claims. Attending\_Physician, Operating\_Physician, and Other\_Physician could be useful in identifying patterns of physician involvement in care, while Claim\_Admitting\_Diagnosis\_Code, Claim\_Day\_Spent, Claim\_Diagnosis\_Related\_Group\_Code, and Claim\_Procedure\_Code\_1 could provide insight into the types of medical conditions and procedures that are most common among Medicare beneficiaries. Ultimately, the selection of these fea-

tures was likely based on their potential to answer the research question and provide valuable insights into Medicare claims data.

Table 4.2: Description of features chosen from the DE-SynPUF

<b>Features</b>	<b>Description</b>	<b>Type</b>
Provider_Institution	Unique Provider Identification Number	Categorical
NCH_PRMRY_PYR _CLM_PD_AMT	NCH Primary Payer Claim Paid Amount	Numerical
AT_PHYSN_NPI	Attending Physician - National Provider Identification - Number	Categorical
OP_PHYSN_NPI	Operating Physician - National Provider Identification - Number	Categorical
OT_PHYSN_NPI	Other Physician - National Provider Identification - Number	Categorical
CLM_UTLZTN _DAY_CNT	Claim Utilization Day Count	Numerical
ADMTNG_IDC9 _DGNS_CD	Claim Admitting Diagnosis Code	Categorical
CLM_DRG_CD	Claim Diagnosis Group Code	Categorical
ICD9_PRCDR_CD.1	Claim Procedure Code 1	Categorical

Inpatient dataset of DE-SynPUF contains the 10 features for Beneficiary Diagnosis, we have dropped those variables as the dataset also contains Claim\_Admitting\_Diagnosis\_Code which indicates the beneficiary's initial di-



agnosis at the time of admission. Mostly the claim is done through this one feature, rest of the diagnosis codes are mainly used for side diseases. Same goes for procedure codes, it has 6 features. Upon our analysis and investigations we found out that the main feature of the dataset is Claim\_Procedure\_Code\_1. Rest of the procedure code features contains the same code as diagnosis code. Lastly, we had 45 features of Revenue Center HCFA Common Procedure which are of no use for us. After the selection of the features, the values within every feature was labeled in such a way that we could distinguish the code after the generation of the rules through association rule mining.

### **4.3 Findings and Interpretations**

Two experiments were conducted in this study. Initially, we provided the preprocessed data to all the chosen unsupervised anomaly detection techniques. However, we discovered that this method was taking an excessive amount of time to complete. A time comparison of the two experiments is presented in tables 4.3 , 4.4. The time delay caused by this method has the potential to result in financial losses, as the experiment would need to be repeated each time a new transaction is added to the database. Therefore, we decided to discontinue this approach and explore alternative methods to improve the efficiency of our analysis.

In second experiment, the Apriori association rule mining algorithm is applied on the preprocessed dataset, resulting in 72 rules that frequently appear together in the CMS 2008-2010 DE-SynPUF dataset (table 4.6). Apriori association rule mining algorithm, a popular data mining method, was used to find interesting relationships between items in large databases. Association rule mining seeks high-confidence rules. Confidence measures the strength of the association between two itemsets, while support measures their frequency in the database. Existing measures only evaluate the quality of the resulted rules separately, missing their dependencies. To evaluate the rules

<b>Dataset / Detector</b>	<b>50%</b>	<b>75%</b>	<b>100%</b>
IF	1.84 Sec	2.67 Sec	3.4 Sec
CBLOF	0.79 Sec	0.80 Sec	2.23 Sec
OCSVM	224.29 Sec	507.96 Sec	897.24 Sec
ECOD	0.6 Sec	1.03 Sec	1.39 Sec
<b>Total Time</b>	227.53 Sec	512.45 Sec	904.24 Sec

Table 4.3: Unsupervised techniques applied independently on the dataset

<b>Algorithm</b>	<b>Time</b>
Apriori Algorithm	867.06 Sec
Apriori > IF	0.08 Sec
Apriori > CBLOF	0.08 Sec
Apriori > OCSVM	0.08 Sec
Apriori > ECOD	0.08 Sec
<b>Total Time</b>	868.18 Sec

Table 4.4: Unsupervised techniques applied on the rules extracted from Apriori algorithm

generated through Apriori association rule mining, we calculated the Coverage score against every rule.

SNo	Antecedents	Consequents	Support	Confidence	Lift	Leverage	Coverage
-----	-------------	-------------	---------	------------	------	----------	----------

1	CADC-7802	A5000	0.0169	0.6712	1.7176	0.0071	0.8545
2	CADC-78650	A5000	0.0232	0.5732	1.4668	0.0074	0.8545
3	PC-3995	A10000	0.0106	0.5498	1.6892	0.0043	0.9484
4	2 Day[s]	A5000	0.0811	0.5482	1.4028	0.0233	0.8545
5	1 Day[s]	A5000	0.0626	0.5416	1.3859	0.0174	0.9202
6	PC-8154	A15000	0.0128	0.5188	3.73	0.0094	0.8545
7	PC-8154	3 Day[s]	0.012	0.4848	2.884	0.0078	0.8545
8	CADC-4280	A10000	0.0122	0.4461	1.3706	0.0033	0.9108
9	3 Day[s]	A5000	0.0747	0.4443	1.1368	0.009	0.8451
10	CADC-486	A10000	0.0155	0.4388	1.3482	0.004	0.9108
11	PC-9904	A10000	0.0122	0.4374	1.3438	0.0031	0.8451
12	CADC-78605	A10000	0.0177	0.4326	1.3292	0.0044	0.9108
13	PC-9904	A5000	0.0117	0.4213	1.078	0.0008	0.8545
14	4 Day[s]	A5000	0.0495	0.4051	1.0366	0.0017	0.8451
15	CADC-78605	A5000	0.0156	0.3797	0.9716	-0.0005	0.8545
16	CADC-4280	A5000	0.0102	0.3712	0.95	-0.0005	0.8638
17	6 Day[s]	A10000	0.0254	0.3708	1.139	0.0031	0.8357
18	5 Day[s]	A10000	0.0339	0.3702	1.1372	0.0041	0.8545
19	5 Day[s]	A5000	0.0336	0.3672	0.9396	-0.0022	0.8451
20	CADC-486	A5000	0.0129	0.3644	0.9324	-0.0009	0.8545
21	4 Day[s]	A10000	0.0444	0.3633	1.1162	0.0046	0.8545
22	8 Day[s]	A10000	0.0144	0.3585	1.1014	0.0013	0.8451
23	7 Day[s]	A10000	0.0191	0.3559	1.0933	0.0016	0.8451
24	3 Day[s]	A10000	0.0557	0.3312	1.0176	0.001	0.8451

25	6 Day[s]	A5000	0.0215	0.3145	0.8048	-0.0052	0.9108
26	2 Day[s]	A10000	0.0446	0.3016	0.9267	-0.0035	0.8545
27	1 Day[s]	A10000	0.0328	0.2838	0.872	-0.0048	0.8545
28	7 Day[s]	A5000	0.0151	0.2818	0.7211	-0.0058	0.8451
29	8 Day[s]	A5000	0.0102	0.2546	0.6514	-0.0055	0.8451
30	CADC-78650	A10000	0.0101	0.2492	0.7655	-0.0031	0.9108
31	A5000	2 Day[s]	0.0811	0.2075	1.4028	0.0233	0.8545
32	A5000	3 Day[s]	0.0747	0.1911	1.1368	0.009	0.8451
33	A15000	3 Day[s]	0.025	0.1797	1.0689	0.0016	0.8545
34	A10000	3 Day[s]	0.0557	0.1711	1.0176	0.001	0.8545
35	A5000	1 Day[s]	0.0626	0.1601	1.3859	0.0174	0.8451
36	6 Day[s]	A15000	0.0108	0.158	1.1361	0.0013	0.9108
37	3 Day[s]	A15000	0.025	0.1487	1.0689	0.0016	0.8545
38	5 Day[s]	A15000	0.0127	0.1389	0.9988	0	0.8545
39	A10000	2 Day[s]	0.0446	0.1371	0.9267	-0.0035	0.9484
40	A10000	4 Day[s]	0.0444	0.1365	1.1162	0.0046	0.8451
41	4 Day[s]	A15000	0.0166	0.1354	0.9737	-0.0004	0.8357
42	A5000	4 Day[s]	0.0495	0.1268	1.0366	0.0017	0.8451
43	A15000	4 Day[s]	0.0166	0.1191	0.9737	-0.0004	0.8545
44	A10000	5 Day[s]	0.0339	0.1041	1.1372	0.0041	0.8451
45	1 Day[s]	A15000	0.0118	0.1021	0.7342	-0.0043	0.8451
46	A10000	1 Day[s]	0.0328	0.1008	0.872	-0.0048	0.8638
47	A15000	2 Day[s]	0.0138	0.0991	0.6697	-0.0068	0.8357
48	2 Day[s]	A15000	0.0138	0.0932	0.6697	-0.0068	0.8545

49	A15000	PC-8154	0.0128	0.0919	3.73	0.0094	0.8545
50	A15000	5 Day[s]	0.0127	0.0914	0.9988	0	0.8545
51	A5000	5 Day[s]	0.0336	0.086	0.9396	-0.0022	0.8357
52	A15000	1 Day[s]	0.0118	0.0848	0.7342	-0.0043	0.8638
53	A10000	6 Day[s]	0.0254	0.0779	1.139	0.0031	0.9014
54	A15000	6 Day[s]	0.0108	0.0777	1.1361	0.0013	0.8451
55	3 Day[s]	PC-8154	0.012	0.0711	2.884	0.0078	0.8451
56	A5000	CADC-78650	0.0232	0.0592	1.4668	0.0074	0.8545
57	A10000	7 Day[s]	0.0191	0.0586	1.0933	0.0016	0.8451
58	A5000	6 Day[s]	0.0215	0.0551	0.8048	-0.0052	0.9014
59	A10000	CADC-78605	0.0177	0.0545	1.3292	0.0044	0.8451
60	A10000	CADC-486	0.0155	0.0477	1.3482	0.004	0.9202
61	A10000	8 Day[s]	0.0144	0.0441	1.1014	0.0013	0.9108
62	A5000	CADC-7802	0.0169	0.0432	1.7176	0.0071	0.8545
63	A5000	CADC-78605	0.0156	0.0399	0.9716	-0.0005	0.8638
64	A5000	7 Day[s]	0.0151	0.0386	0.7211	-0.0058	0.9108
65	A10000	CADC-4280	0.0122	0.0375	1.3706	0.0033	0.9108
66	A10000	PC-9904	0.0122	0.0375	1.3438	0.0031	0.8545
67	A5000	CADC-486	0.0129	0.033	0.9324	-0.0009	0.8545
68	A10000	PC-3995	0.0106	0.0325	1.6892	0.0043	0.8545
69	A10000	CADC-78650	0.0101	0.0309	0.7655	-0.0031	0.8451
70	A5000	PC-9904	0.0117	0.03	1.078	0.0008	0.8545
71	A5000	8 Day[s]	0.0102	0.0261	0.6514	-0.0055	0.9108
72	A5000	CADC-4280	0.0102	0.026	0.95	-0.0005	0.8545

Table 4.6: Results of Apriori Association Rule Mining

The rules only tell us about the itemsets appearing together in the transaction, it doesn't tell us about the nature of the rule; whether it is normal or fraudulent. To overcome the obstacle of identifying the nature of the generated rules, the study then moved towards identifying the fraudulent rules by using Isolation Forest algorithm. The Isolation Forest algorithm is an unsupervised machine learning algorithm used for anomaly detection. It works by creating random decision trees to isolate fraudulent points from normal points in the dataset. The algorithm initially identified 14 fraudulent rules in the DE-SynPUF dataset. However, due to the sensitive nature of healthcare and financial transactions, we applied three additional unsupervised algorithms named CBLOF, ECOD, and OCSVM to obtain more reliable and weighted results.

As a result, these algorithms identified 8, 4, and 8 fraudulent rules, respectively. In total, 52 out of 72 rules were marked as normal by all of the algorithms. However, in combination, 20 rules were marked as fraudulent by one or more algorithms.

The results of our analysis are presented in Table 4.7, which shows the classification of rules according to the number of algorithms that classified them as fraudulent. The table indicates that 52 rules were classified as normal, while 10 were classified as fraudulent by one algorithm, 6 were classified as fraudulent by two algorithms, and 4 were classified as fraudulent by three algorithms. No rules were classified as fraudulent by all four algorithms.

These findings suggest that a combination of unsupervised algorithms help us achieve the most accurate and reliable results in detecting healthcare insurance fraud. Detailed results against the transactions can be seen in Table 4.8. For example, The first rule states that if the Diagnosis Code is 7802 and

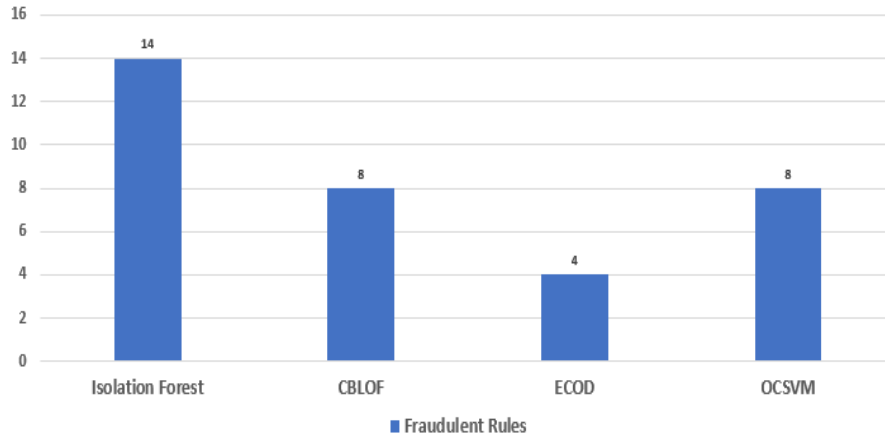


Figure 4.8: Fraudulent Classification Through Classifiers

Rules Classification	Rules
Normal Rules	52
Classified as Fraudulent By 1 Algorithm[s]	10
Classified as Fraudulent By 2 Algorithm[s]	6
Classified as Fraudulent By 3 Algorithm[s]	4
Classified as Fraudulent By 4 Algorithm[s]	0

Table 4.7: Rules Classification Through Applied Techniques

the Patient Pay is less than \$5000 USD, the claim is classified as fraudulent (1) by the One-OCSVM and ECOD, while it is classified as normal (0) by the Isolation Forest and CBLOF detectors. As for the second rule, it states that if the Diagnosis Code is 78650 and the Patient Pay is less than \$5000 USD, the claim is classified as normal (0) by all detectors. The following table shows all transaction along with the fraudulent status by the selected anomaly detection techniques.

SNo	Rules	IF	CBLOF	OCSVM	ECOD
1	Diagnosis Code = 7802 $\wedge$ Patient Pay Less Than 5000 USD	0	0	1	1
2	Diagnosis Code = 78650 $\wedge$ Patient Pay Less Than 5000 USD	0	0	0	0
3	Procedure Code = 3995 $\wedge$ Patient Pay More Than 5000 USD and Less Than 10000 USD	0	0	0	1
4	Patient Stay = 2 Day[s] $\wedge$ Patient Pay Less Than 5000 USD	0	0	0	0
5	Patient Stay = 1 Day[s] $\wedge$ Patient Pay Less Than 5000 USD	0	0	0	0
6	Procedure Code = 8154 $\wedge$ Patient Pay More Than 10000 USD and Less Than 15000 USD	1	1	0	1
7	Procedure Code = 8154 $\wedge$ Patient Stay = 3 Day[s]	0	1	1	1
8	Diagnosis Code = 4280 $\wedge$ Patient Pay More Than 5000 USD and Less Than 10000 USD	0	0	0	0
9	Patient Stay = 3 Day[s] $\wedge$ Patient Pay Less Than 5000 USD	0	0	0	0
10	Diagnosis Code = 486 $\wedge$ Patient Pay More Than 5000 USD and Less Than 10000 USD	0	0	0	0
11	Procedure Code = 9904 $\wedge$ Patient Pay More Than 5000 USD and Less Than 10000 USD	0	0	0	0



12	Diagnosis Code = 78605 $\wedge$ Patient Pay More Than 5000 USD and Less Than 10000 USD	0	0	0	0
13	Procedure Code = 9904 $\wedge$ Patient Pay Less Than 5000 USD	0	0	0	0
14	Patient Stay = 4 Day[s] $\wedge$ Patient Pay Less Than 5000 USD	0	0	0	0
15	Diagnosis Code = 78605 $\wedge$ Patient Pay Less Than 5000 USD	0	0	0	0
16	Diagnosis Code = 4280 $\wedge$ Patient Pay Less Than 5000 USD	0	0	0	0
17	Patient Stay = 6 Day[s] $\wedge$ Patient Pay More Than 5000 USD and Less Than 10000 USD	0	0	0	0
18	Patient Stay = 5 Day[s] $\wedge$ Patient Pay More Than 5000 USD and Less Than 10000 USD	0	0	0	0
19	Patient Stay = 5 Day[s] $\wedge$ Patient Pay Less Than 5000 USD	0	0	0	0
20	Diagnosis Code = 486 $\wedge$ Patient Pay Less Than 5000 USD	0	0	0	0
21	Patient Stay = 4 Day[s] $\wedge$ Patient Pay More Than 5000 USD and Less Than 10000 USD	0	0	0	0
22	Patient Stay = 8 Day[s] $\wedge$ Patient Pay More Than 5000 USD and Less Than 10000 USD	0	0	0	0

23	Patient Stay = 7 Day[s] $\wedge$ Patient Pay More Than 5000 USD and Less Than 10000 USD	0	0	0	0
24	Patient Stay = 3 Day[s] $\wedge$ Patient Pay More Than 5000 USD and Less Than 10000 USD	0	0	0	0
25	Patient Stay = 6 Day[s] $\wedge$ Patient Pay Less Than 5000 USD	0	0	0	0
26	Patient Stay = 2 Day[s] $\wedge$ Patient Pay More Than 5000 USD and Less Than 10000 USD	0	0	0	0
27	Patient Stay = 1 Day[s] $\wedge$ Patient Pay More Than 5000 USD and Less Than 10000 USD	1	0	0	0
28	Patient Stay = 7 Day[s] $\wedge$ Patient Pay Less Than 5000 USD	0	0	0	0
29	Patient Stay = 8 Day[s] $\wedge$ Patient Pay Less Than 5000 USD	0	0	0	0
30	Diagnosis Code = 78650 $\wedge$ Patient Pay More Than 5000 USD and Less Than 10000 USD	0	0	0	0
31	Patient Less Than 5000 USD $\wedge$ Patient Stay = 2 Day[s]	0	0	0	0
32	Patient Less Than 5000 USD $\wedge$ Patient Stay = 3 Day[s]	0	0	0	0
33	Patient Pay More Than 10000 and Less Than 15000 USD $\wedge$ Patient Stay = 3 Day[s]	0	0	0	0

34	Patient Pay More Than 5000 and Less Than 10000 USD $\wedge$ Patient Stay = 3 Day[s]	0	0	0	0
35	Patient Less Than 5000 USD $\wedge$ Patient Stay = 1 Day[s]	0	0	0	0
36	Patient Stay = 6 Day[s] $\wedge$ Patient Pay More Than 10000 USD and Less Than 15000 USD	1	0	0	0
37	Patient Stay = 3 Day[s] $\wedge$ Patient Pay More Than 10000 USD and Less Than 15000 USD	0	0	0	0
38	Patient Stay = 5 Day[s] $\wedge$ Patient Pay More Than 10000 USD and Less Than 15000 USD	1	0	0	0
39	Patient Pay More Than 5000 and Less Than 10000 USD $\wedge$ Patient Stay = 2 Day[s]	0	0	0	0
40	Patient Pay More Than 5000 and Less Than 10000 USD $\wedge$ Patient Stay = 4 Day[s]	0	0	0	0
41	Patient Stay = 4 Day[s] $\wedge$ Patient Pay More Than 10000 USD and Less Than 15000 USD	1	1	0	0
42	Patient Less Than 5000 USD $\wedge$ Patient Stay = 4 Day[s]	0	0	0	0
43	Patient Pay More Than 10000 and Less Than 15000 USD $\wedge$ Patient Stay = 4 Day[s]	1	1	0	0

44	Patient Pay More Than 5000 and Less Than 10000 USD $\wedge$ Patient Stay = 5 Day[s]	0	0	0	0
45	Patient Stay = 1 Day[s] $\wedge$ Patient Pay More Than 10000 USD and Less Than 15000 USD	1	0	0	0
46	Patient Pay More Than 5000 and Less Than 10000 USD $\wedge$ Patient Stay = 1 Day[s]	1	0	0	0
47	Patient Pay More Than 10000 and Less Than 15000 USD $\wedge$ Patient Stay = 2 Day[s]	1	1	0	0
48	Patient Stay = 2 Day[s] $\wedge$ Patient Pay More Than 10000 USD and Less Than 15000 USD	1	1	0	0
49	Patient Pay More Than 10000 and Less Than 15000 USD $\wedge$ Procedure Code = 8154	1	1	0	1
50	Patient Pay More Than 10000 and Less Than 15000 USD $\wedge$ Patient Stay = 5 Day[s]	1	0	0	0
51	Patient Less Than 5000 USD $\wedge$ Patient Stay = 5 Day[s]	0	0	0	0
52	Patient Pay More Than 10000 and Less Than 15000 USD $\wedge$ Patient Stay = 1 Day[s]	1	0	0	0
53	Patient Pay More Than 5000 and Less Than 10000 USD $\wedge$ Patient Stay = 6 Day[s]	0	0	0	0

54	Patient Pay More Than 10000 and Less Than 15000 USD $\wedge$ Patient Stay = 6 Day[s]	1	0	0	0
55	Patient Stay = 3 Day[s] $\wedge$ Procedure Code = 8154	0	1	1	1
56	Patient Less Than 5000 USD $\wedge$ Diagnosis Code = 78650	0	0	0	0
57	Patient Pay More Than 5000 and Less Than 10000 USD $\wedge$ Patient Stay = 7 Day[s]	0	0	0	0
58	Patient Less Than 5000 USD $\wedge$ Patient Stay = 6 Day[s]	0	0	0	0
59	Patient Pay More Than 5000 and Less Than 10000 USD $\wedge$ Diagnosis Code = 78605	0	0	0	0
60	Patient Pay More Than 5000 and Less Than 10000 USD $\wedge$ Diagnosis Code = 486	0	0	0	0
61	Patient Pay More Than 5000 and Less Than 10000 USD $\wedge$ Patient Stay = 8 Day [s]	0	0	0	0
62	Patient Less Than 5000 USD $\wedge$ Diagnosis Code = 7802	0	0	1	1
63	Patient Less Than 5000 USD $\wedge$ Diagnosis Code = 78605	0	0	0	0
64	Patient Less Than 5000 USD $\wedge$ Patient Stay = 7 Day[s]	0	0	0	0

65	Patient Pay More Than 5000 and Less Than 10000 USD $\wedge$ Diagnosis Code = 4280	0	0	0	0
66	Patient Pay More Than 5000 and Less Than 10000 USD $\wedge$ Procedure Code = 9904	0	0	0	0
67	Patient Less Than 5000 USD $\wedge$ Diagnosis Code = 486	0	0	0	0
68	Patient Pay More Than 5000 and Less Than 10000 USD $\wedge$ Procedure Code = 3995	0	0	0	1
69	Patient Pay More Than 5000 and Less Than 10000 USD $\wedge$ Diagnosis Code = 78650	0	0	0	0
70	Patient Less Than 5000 USD $\wedge$ Procedure Code = 9904	0	0	0	0
71	Patient Less Than 5000 USD $\wedge$ Patient Stay = 8 Day [s]	0	0	0	0
72	Patient Less Than 5000 USD $\wedge$ Diagnosis Code = 4280	0	0	0	0

Table 4.8: Identification of fraudulent rules through applied techniques

While the following table carries the transactions which are classified as fraudulent by one or more identifiers.

SNo	Rules	IF	CBLOF	OCSVM	ECOD
-----	-------	----	-------	-------	------

1	Diagnosis Code = 7802 $\wedge$ Patient Pay Less Than 5000 USD	0	0	1	1
2	Procedure Code = 3995 $\wedge$ Patient Pay More Than 5000 USD and Less Than 10000 USD	0	0	0	1
3	Procedure Code = 8154 $\wedge$ Patient Pay More Than 10000 USD and Less Than 15000 USD	1	1	0	1
4	Procedure Code = 8154 $\wedge$ Patient Stay = 3 Day[s]	0	1	1	1
5	Patient Stay = 1 Day[s] $\wedge$ Patient Pay More Than 5000 USD and Less Than 10000 USD	1	0	0	0
6	Patient Stay = 6 Day[s] $\wedge$ Patient Pay More Than 10000 USD and Less Than 15000 USD	1	0	0	0
7	Patient Stay = 5 Day[s] $\wedge$ Patient Pay More Than 10000 USD and Less Than 15000 USD	1	0	0	0
8	Patient Stay = 4 Day[s] $\wedge$ Patient Pay More Than 10000 USD and Less Than 15000 USD	1	1	0	0
9	Patient Pay More Than 10000 and Less Than 15000 USD $\wedge$ Patient Stay = 4 Day[s]	1	1	0	0
10	Patient Stay = 1 Day[s] $\wedge$ Patient Pay More Than 10000 USD and Less Than 15000 USD	1	0	0	0

11	Patient Pay More Than 5000 and Less Than 10000 USD $\wedge$ Patient Stay = 1 Day[s]	1	0	0	0
12	Patient Pay More Than 10000 and Less Than 15000 USD $\wedge$ Patient Stay = 2 Day[s]	1	1	0	0
13	Patient Stay = 2 Day[s] $\wedge$ Patient Pay More Than 10000 USD and Less Than 15000 USD	1	1	0	0
14	Patient Pay More Than 10000 and Less Than 15000 USD $\wedge$ Procedure Code = 8154	1	1	0	1
15	Patient Pay More Than 10000 and Less Than 15000 USD $\wedge$ Patient Stay = 5 Day[s]	1	0	0	0
16	Patient Pay More Than 10000 and Less Than 15000 USD $\wedge$ Patient Stay = 1 Day[s]	1	0	0	0
17	Patient Pay More Than 10000 and Less Than 15000 USD $\wedge$ Patient Stay = 6 Day[s]	1	0	0	0
18	Patient Stay = 3 Day[s] $\wedge$ Procedure Code = 8154	0	1	1	1
19	Patient Less Than 5000 USD $\wedge$ Diagnosis Code = 7802	0	0	1	1
20	Patient Pay More Than 5000 and Less Than 10000 USD $\wedge$ Procedure Code = 3995	0	0	0	1

Table 4.9: Identification of fraudulent rules through applied techniques



We apply the Silhouette Scores method to evaluate the effectiveness of four different anomaly detection techniques: Isolation Forest, CBLOF, ECOD, and OCSVM. The results of this evaluation are shown in Table 4.10. The Silhouette Scores for each technique are listed in the "Scores" column, while the "Classifier" column specifies the name of the anomaly detection technique used. As we can see from the table, the CBLOF technique has the highest Silhouette Score of 0.114, followed by Isolation Forest with a score of 0.103. The ECOD and OCSVM techniques had lower scores of 0.063 and 0.060, respectively.

No.	Classifier	Scores
1	Isolation Forest	0.103
2	CBLOF	0.114
3	ECOD	0.063
4	OCSVM	0.060

Table 4.10: Silhouette Score of applied techniques

Another popular technique for evaluating the results is through the visual representation. Visual representation of the classifiers are as follow

The study was performed on a Windows 10 PC with 11th Gen Intel (R) Core (TM) i7 processor, 32 GB of RAM, and a Nvidia GeForce GTX 1650 with 4GB dedicated memory. Apriori was implemented by using mlxtend python library as it is easy to implement. For the implementation of Isolation Forest, we used sklearn library. OCSVM, ECOD, and CBLOF were implemented using Python Outlier Detection (PyOD) library. PyOD is considered as the best anomaly detection library containing most of the unsupervised algorithm.

These findings demonstrate the effectiveness of data mining techniques

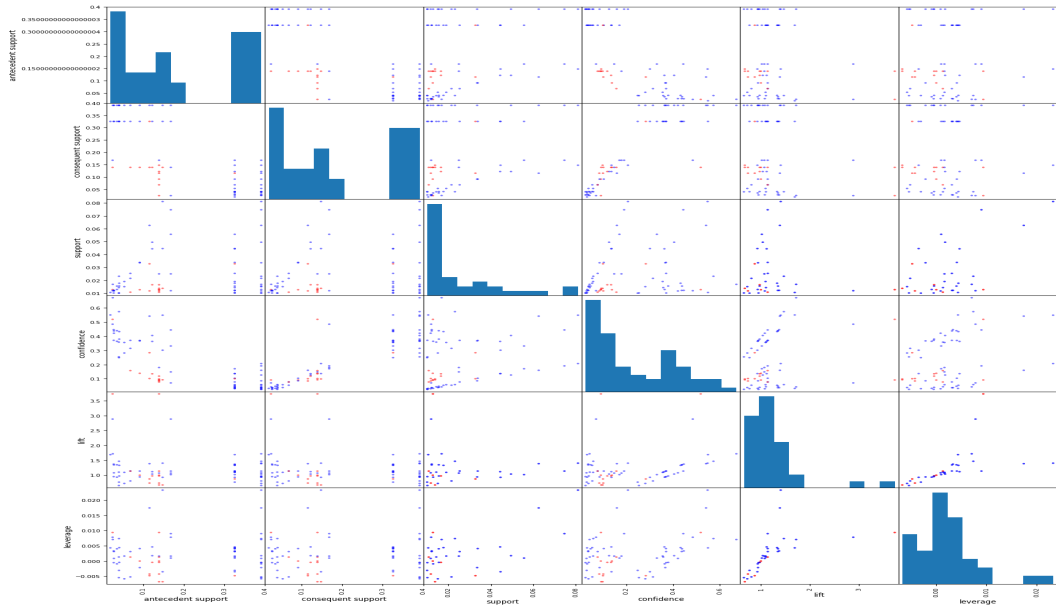


Figure 4.9: Fraudulent Classification Through Isolation Forest

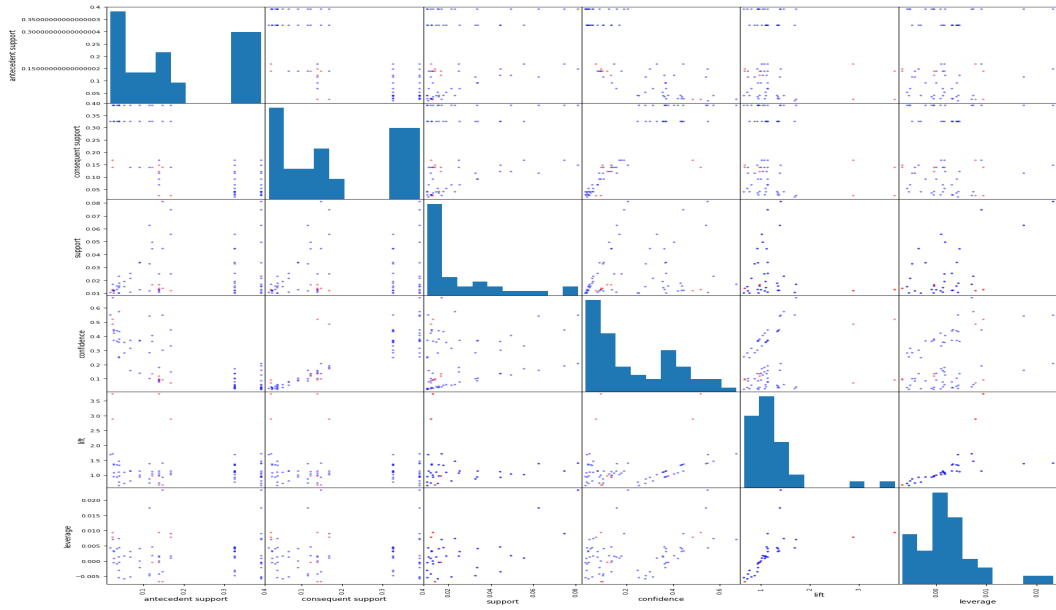


Figure 4.10: Fraudulent Classification Through CBLOF

for healthcare insurance fraud detection and can have important implications for fraud prevention efforts in the healthcare industry.

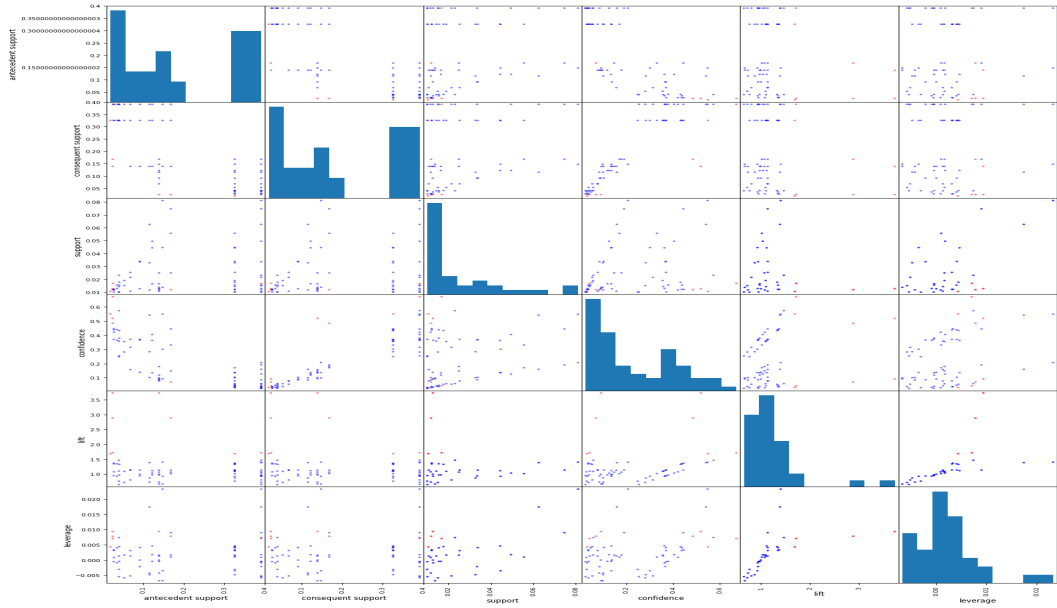


Figure 4.11: Fraudulent Classification Through ECOD

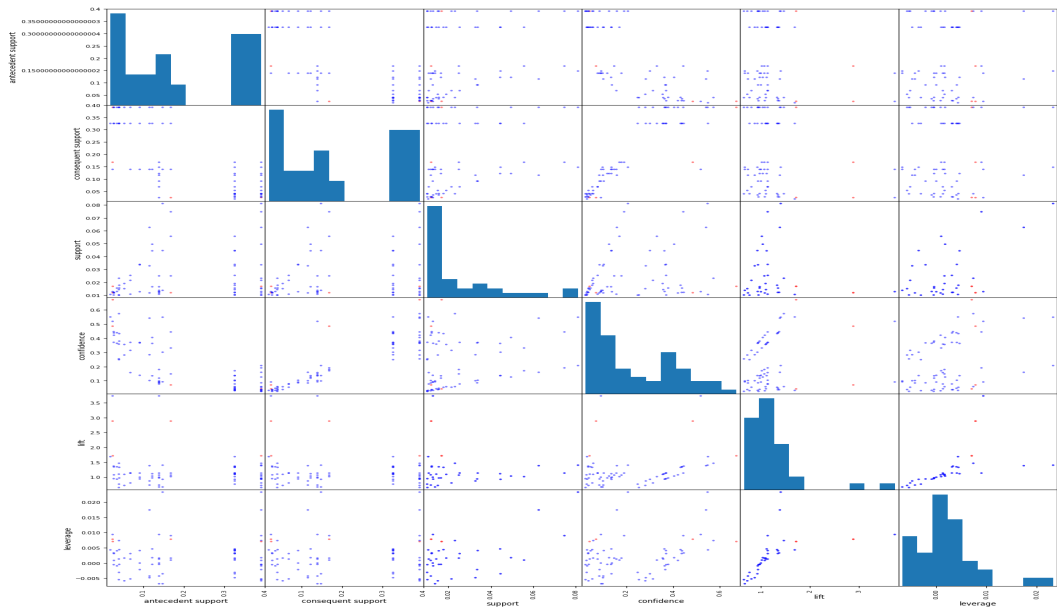


Figure 4.12: Fraudulent Classification Through OCSVM

## CHAPTER 5

### CONCLUSION & FUTURE WORK

The complexity and substantial monetary value of the healthcare industry make it a desirable target for fraudulent activity. Due to the growing older population, healthcare insurance has been a consistent focus. The Centers for Medicare & Medicaid Services (CMS) and other organizations work ceaselessly to reduce fraudulent operations. The use of publicly accessible healthcare insurance data to identify and prevent potential fraudulent actions is a recent development, despite the issue's longevity. Effective machine learning solutions can drastically minimize fraudulent occurrences and the resources necessary to investigate probable fraud cases.

In our study, we present a model based on five unsupervised learning techniques for detecting healthcare insurance fraud. We use Apriori association rule mining technique, which was not previously used on CMS 2008-2010 DE-SynPUF dataset. We obtain 72 rules, which are further provided to the anomaly detection algorithms such as Isolation Forest, OCSVM, ECOD, and CBLOF. After combining all results, we find out that 20 rules are classified as fraudulent by one or more than one algorithm, and 52 are marked as normal.

While our study shows promising results in detecting healthcare insurance fraud using unsupervised learning techniques, there is still room for improvement and further research. One area for future work is to test the methodology on other healthcare insurance datasets with labels. This will allow for a

more thorough evaluation of the algorithm's effectiveness and accuracy.

Moreover, we plan to explore and experiment with additional unsupervised learning techniques and algorithms to further enhance the model's performance. This could include the integration of deep learning algorithms and techniques to improve the overall accuracy and robustness of the model.

Overall, the results of this study provide a strong foundation for future research in the detection of healthcare insurance fraud using unsupervised learning techniques. We will continue to work towards improving the model's performance and developing a more comprehensive and effective approach for detecting fraudulent activities in healthcare insurance datasets.

## REFERENCES

- [1] G. of Pakistan, Introduction — sehat sahulat program (2019).  
URL <https://sehatinsafcard.com/introduction.php>
- [2] G. of Pakistan, Benefits package (2019).  
URL <https://sehatinsafcard.com/benefits.php>
- [3] (1965). [link].  
URL <https://www.medicare.gov/>
- [4] J. Gee, M. Button, G. Brooks, The financial cost of healthcare fraud: what data from around the world shows (2010).
- [5] D. M. Berwick, A. D. Hackbarth, Eliminating waste in us health care, *Jama* 307 (14) (2012) 1513–1516.
- [6] K. M. King, Progress made, but more action needed to address medicare fraud, waste, and abuse (Apr 2014).  
URL <https://www.gao.gov/assets/gao-14-560t.pdf>
- [7] P. Barrett, Global claims fraud survey (2017).
- [8] A. Miller, Health and hard time (2013).
- [9] A. Hansson, H. Cedervall, Insurance fraud detection using unsupervised sequential anomaly detection (2022).
- [10] C. Gomes, Z. Jin, H. Yang, Insurance fraud detection with unsupervised deep learning, *Journal of Risk and Insurance* 88 (3) (2021) 591–624.

- [11] I. Matloob, S. Khan, H. ur Rahman, F. Hussain, Medical health benefit management system for real-time notification of fraud using historical medical records, *Applied Sciences* 10 (15) (2020) 5144.
- [12] B. Benedek, C. Ciumas, B. Z. Nagy, Automobile insurance fraud detection in the age of big data—a systematic and comprehensive literature review, *Journal of Financial Regulation and Compliance* (2022).
- [13] C. Yadav, S. Wang, M. Kumar, An approach to improve apriori algorithm based on association rule mining, in: 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), IEEE, 2013, pp. 1–9.
- [14] S. Kareem, R. B. Ahmad, A. B. Sarlan, Framework for the identification of fraudulent health insurance claims using association rule mining, in: 2017 IEEE Conference on Big Data and Analytics (ICBDA), IEEE, 2017, pp. 99–104.
- [15] M. Sornalakshmi, S. Balamurali, M. Venkatesulu, M. N. Krishnan, L. K. Ramasamy, S. Kadry, S. Lim, An efficient apriori algorithm for frequent pattern mining using mapreduce in healthcare data, *Bulletin of Electrical Engineering and Informatics* 10 (1) (2021) 390–403.
- [16] U. Abdullah, J. Ahmad, A. Ahmed, Analysis of effectiveness of apriori algorithm in medical billing data mining, in: 2008 4th International Conference on Emerging Technologies, IEEE, 2008, pp. 327–331.
- [17] D. Thornton, G. van Capelleveen, M. Poel, J. van Hillegersberg, R. M. Mueller, Outlier-based health insurance fraud detection for us medicaid data., in: *ICEIS* (2), 2014, pp. 684–694.
- [18] M. Kirlidog, C. Asuk, A fraud detection approach with data mining in health insurance, *Procedia - Social and Behavioral Sci-*

- ences 62 (2012) 989–994, world Conference on Business, Economics and Management (BEM-2012), May 4–6 2012, Antalya, Turkey. doi:<https://doi.org/10.1016/j.sbspro.2012.09.168>.
- [19] Y. Gao, C. Sun, R. Li, Q. Li, L. Cui, B. Gong, An efficient fraud identification method combining manifold learning and outliers detection in mobile healthcare services, *IEEE Access* 6 (2018) 60059–60068. doi:[10.1109/ACCESS.2018.2875516](https://doi.org/10.1109/ACCESS.2018.2875516).
- [20] R. H. Alwan, M. M. Hamad, O. A. Dawood, A comprehensive survey of fraud detection methods in credit card based on data mining techniques, in: *AIP Conference Proceedings*, Vol. 2400, AIP Publishing LLC, 2022, p. 020006.
- [21] W. Shang, P. Zeng, M. Wan, L. Li, P. An, Intrusion detection algorithm based on ocsvm in industrial control system, *Security and Communication Networks* 9 (10) (2016) 1040–1049.
- [22] L. A. Maglaras, J. Jiang, T. Cruz, Integrated ocsvm mechanism for intrusion detection in scada systems, *Electronics Letters* 50 (25) (2014) 1935–1936.
- [23] L. A. Maglaras, J. Jiang, T. J. Cruz, Combining ensemble methods and social network metrics for improving accuracy of ocsvm on intrusion detection in scada systems, *Journal of Information Security and Applications* 30 (2016) 15–26. doi:<https://doi.org/10.1016/j.jisa.2016.04.002>.
- [24] L. A. Maglaras, J. Jiang, Ocsvm model combined with k-means recursive clustering for intrusion detection in scada systems, in: *10th International conference on heterogeneous networking for quality, reliability, security and robustness*, IEEE, 2014, pp. 133–134.



- [25] Z. Wang, Y. Fu, C. Song, P. Zeng, L. Qiao, Power system anomaly detection based on ocsvm optimized by improved particle swarm optimization, *IEEE Access* 7 (2019) 181580–181588.
- [26] M. Amer, M. Goldstein, S. Abdennadher, Enhancing one-class support vector machines for unsupervised anomaly detection, in: *Proceedings of the ACM SIGKDD workshop on outlier detection and description*, 2013, pp. 8–15.
- [27] F. T. Liu, K. M. Ting, Z.-H. Zhou, Isolation forest, in: *2008 Eighth IEEE International Conference on Data Mining*, 2008, pp. 413–422. doi:10.1109/ICDM.2008.17.
- [28] D. Xu, Y. Wang, Y. Meng, Z. Zhang, An improved data anomaly detection method based on isolation forest, in: *2017 10th international symposium on computational intelligence and design (ISCID)*, Vol. 2, IEEE, 2017, pp. 287–291.
- [29] Z. Cheng, C. Zou, J. Dong, Outlier detection using isolation forest and local outlier factor, in: *Proceedings of the conference on research in adaptive and convergent systems*, 2019, pp. 161–168.
- [30] Z. Ding, M. Fei, An anomaly detection approach based on isolation forest algorithm for streaming data using sliding window, *IFAC Proceedings Volumes* 46 (20) (2013) 12–17.
- [31] J. Lesouple, C. Baudoin, M. Spigai, J.-Y. Tourneret, Generalized isolation forest for anomaly detection, *Pattern Recognition Letters* 149 (2021) 109–119.
- [32] Z. He, X. Xu, S. Deng, Discovering cluster-based local outliers, *Pattern Recognition Letters* 24 (9) (2003) 1641–1650. doi:[https://doi.org/10.1016/S0167-8655\(03\)00003-5](https://doi.org/10.1016/S0167-8655(03)00003-5).

- [33] H. John, S. Naaz, Credit card fraud detection using local outlier factor and isolation forest, *International Journal of Computer Sciences and Engineering* 7 (2019) 1060–1064. doi:10.26438/ijcse/v7i4.10601064.
- [34] M. N. Kanyama, C. Nyirenda, N. Clement-Temaneh, Anomaly detection in smart water metering networks, in: *The 5th International Workshop on Advanced Computational Intelligence and Intelligent Informatics (IWACIII2017)*, 2017, pp. 1–10.
- [35] I. Ullah, H. Hussain, S. Rahman, A. Rahman, M. Shabir, N. Ullah, K. Ullah, Using k-means, lof, and cblof as prediction tools.
- [36] I. Ullah, H. Hussain, I. Ali, A. Liaquat, Churn prediction in banking system using k-means, lof, and cblof, in: *2019 International conference on electrical, communication, and computer engineering (ICECCE)*, IEEE, 2019, pp. 1–6.
- [37] R. Agrawal, R. Srikant, et al., Fast algorithms for mining association rules, in: *Proc. 20th int. conf. very large data bases, VLDB*, Vol. 1215, Santiago, Chile, 1994, pp. 487–499.
- [38] X. Liu, Y. Zhao, M. Sun, An improved apriori algorithm based on an evolution-communication tissue-like p system with promoters and inhibitors, *Discrete Dynamics in Nature and Society* 2017 (2017).
- [39] M. H. Santoso, Application of association rule method using apriori algorithm to find sales patterns case study of indomaret tanjung anom, *Brilliance: Research of Artificial Intelligence* 1 (2) (2021) 54–66.
- [40] B. Schölkopf, R. C. Williamson, A. Smola, J. Shawe-Taylor, J. Platt, Support vector method for novelty detection, *Advances in neural information processing systems* 12 (1999).

- [41] P. J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, *Journal of Computational and Applied Mathematics* 20 (1987) 53–65. doi:[https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7).
- [42] K. R. Shahapure, C. Nicholas, Cluster quality analysis using silhouette score, in: *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*, IEEE, 2020, pp. 747–748.
- [43] Health care fraud (Jun 2016).
- [44] What is health insurance fraud?

th

---

ORIGINALITY REPORT

---

18%

SIMILARITY INDEX

14%

INTERNET SOURCES

13%

PUBLICATIONS

7%

STUDENT PAPERS

---

PRIMARY SOURCES

---

1 [jurnal.itscience.org](http://jurnal.itscience.org) 1%  
Internet Source

---

2 [link.springer.com](http://link.springer.com) 1%  
Internet Source

---

3 [www.cnblogs.com](http://www.cnblogs.com) 1%  
Internet Source

---

4 [www.researchgate.net](http://www.researchgate.net) 1%  
Internet Source

---

5 [www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov) 1%  
Internet Source

---

6 [e-theses.imtlucca.it](http://e-theses.imtlucca.it) 1%  
Internet Source

---

7 Submitted to American University of Culture and Education 1%  
Student Paper

---

8 [lib.dr.iastate.edu](http://lib.dr.iastate.edu) 1%  
Internet Source

---

9 [medium.com](http://medium.com) 1%  
Internet Source

---

10 Leandros A., Jianmin Jiang. "A real time OCSVM Intrusion Detection module with low overhead for SCADA systems", International Journal of Advanced Research in Artificial Intelligence, 2014  
Publication <1 %

---

11 Shang, Wenli, Peng Zeng, Ming Wan, Lin Li, and Panfeng An. "Intrusion detection algorithm based on OCSVM in industrial control system : Intrusion detection algorithm based on OCSVM", Security and Communication Networks, 2015.  
Publication <1 %

---

12 [www.udemy.com](http://www.udemy.com)  
Internet Source <1 %

---

13 Submitted to University College London  
Student Paper <1 %

---

14 [repository.tudelft.nl](http://repository.tudelft.nl)  
Internet Source <1 %

---

15 Submitted to University of Warwick  
Student Paper <1 %

---

16 Andreas Bayerstadler, Linda van Dijk, Fabian Winter. "Bayesian multinomial latent variable modeling for fraud and abuse detection in health insurance", Insurance: Mathematics and Economics, 2016  
Publication <1 %

---

17	docksci.com Internet Source	<1 %
18	Djenouri, Youcef, Youcef Gheraibia, Malika Mehdi, Ahcene Bendjoudi, and Nadia Nouali-Taboudjemat. "An efficient measure for evaluating association rules", 2014 6th International Conference of Soft Computing and Pattern Recognition (SoCPaR), 2014. Publication	<1 %
19	essay.utwente.nl Internet Source	<1 %
20	Submitted to Indian Institute of Technology, Madras Student Paper	<1 %
21	Richard Bauder, Raquel da Rosa, Taghi Khoshgoftaar. "Identifying Medicare Provider Fraud with Unsupervised Machine Learning", 2018 IEEE International Conference on Information Reuse and Integration (IRI), 2018 Publication	<1 %
22	Submitted to University of Huddersfield Student Paper	<1 %
23	Charu C. Aggarwal. "Outlier Analysis", Springer Science and Business Media LLC, 2017 Publication	<1 %

24

Internet Source

&lt;1 %

25

"Advances in Databases: Concepts, Systems and Applications", Springer Science and Business Media LLC, 2007

Publication

&lt;1 %

26

Jette Henderson, Joyce Ho, Joydeep Ghosh. "gamAID: Greedy CP tensor decomposition for supervised EHR-based disease trajectory differentiation", 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2017

Publication

&lt;1 %

27

[vdoc.pub](http://vdoc.pub)

Internet Source

&lt;1 %

28

[www.atmseminar.org](http://www.atmseminar.org)

Internet Source

&lt;1 %

29

Submitted to Gitam University

Student Paper

&lt;1 %

30

Submitted to The University of Manchester

Student Paper

&lt;1 %

31

Yan Hu, Dah-Ming Chiu, John C. S. Lui. "Application Identification Based on Network Behavioral Profiles", 2008 16th International Workshop on Quality of Service, 2008

Publication

&lt;1 %

---

32 rdr.io  
Internet Source <1 %

---

33 "Advances in Knowledge Discovery and Data Mining", Springer Science and Business Media LLC, 2006  
Publication <1 %

---

34 Richard Bauder, Taghi M. Khoshgoftaar, Naeem Seliya. "A survey on the state of healthcare upcoding fraud analysis and detection", Health Services and Outcomes Research Methodology, 2016  
Publication <1 %

---

35 hdl.handle.net  
Internet Source <1 %

---

36 Submitted to University of Durham  
Student Paper <1 %

---

37 wwwp.uniriotec.br  
Internet Source <1 %

---

38 Ali A. Ghorbani, Wei Lu, Mahbod Tavallaee. "Chapter 4 Theoretical Foundation of Detection", Springer Science and Business Media LLC, 2010  
Publication <1 %

---

39 hal-utt.archives-ouvertes.fr  
Internet Source <1 %

---

hrsonline.isr.umich.edu



40

Internet Source

<1 %

41

[mdsoar.org](https://mdsoar.org)

Internet Source

<1 %

42

Honggang Yang, Shaowen Li, Lijing Tu, Rongrong Ma, Yin Chen. "Unsupervised Outlier Detection Mechanism for Tea Traceability Data", IEEE Access, 2022

Publication

<1 %

43

[bisite.usal.es](https://bisite.usal.es)

Internet Source

<1 %

44

[dokumen.pub](https://dokumen.pub)

Internet Source

<1 %

45

[www.cdc.gov](https://www.cdc.gov)

Internet Source

<1 %

46

[www.slideshare.net](https://www.slideshare.net)

Internet Source

<1 %

47

Amina elmahalawy, Hayam Mousa, Khalid Amin. "A Comparative Study for Outlier Detection Strategies Based On Traditional Machine Learning For IoT Data Analysis.", IJCI. International Journal of Computers and Information, 2021

Publication

<1 %

48

Heiko Paulheim, Robert Meusel. "A decomposition of the outlier detection

<1 %

problem into a set of supervised learning problems", Machine Learning, 2015

Publication

---

49 Leandros Maglaras, Helge Janicke, Jianmin Jiang, Andrew Crampton. "chapter 9 Novel Intrusion Detection Mechanism with Low Overhead for SCADA Systems", IGI Global, 2017 <1 %

Publication

---

50 Mei-Ling Shyu, Shu-Ching Chen, R. L. Kashyap. "Generalized Affinity-Based Association Rule Mining for Multimedia Database Queries", Knowledge and Information Systems, 2001 <1 %

Publication

---

51 ruor.uottawa.ca <1 %

Internet Source

---

52 towardsdatascience.com <1 %

Internet Source

---

53 www.mdpi.com <1 %

Internet Source

---

54 Rhodes, B.J.. "Taxonomic knowledge structure discovery from imagery-based data using the neural associative incremental learning (NAIL) algorithm", Information Fusion, 200707 <1 %

Publication

---

55 irl.umsl.edu <1 %

Internet Source

---