AI and Blockchain empowered Health Insurance Fraud Detection

Submitted By

Khyati Kapadiya

20MCEC19

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INSTITUTE OF TECHNOLOGY

NIRMA UNIVERSITY

AHMEDABAD-382481

May 2022

AI and Blockchain empowered Health Insurance Fraud Detection

Major Project

Submitted in partial fulfillment of the requirements

for the degree of

Master of Technology in Computer Science and Engineering

Submitted By

Khyati Kapadiya

(20MCEC19)

Guided By

Prof. Usha Patel

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INSTITUTE OF TECHNOLOGY

NIRMA UNIVERSITY

AHMEDABAD-382481

May 2022

Certificate

This is to certify that the major project entitled **"AI and Blockchain empowered Health Insurance Fraud Detection"** submitted by **Khyati Kapadiya (20MCEC19)**, towards the partial fulfillment of the requirements for the award of degree of Master of Technology in Computer Science and Engineering of Nirma University, Ahmadabad, is the record of work carried out by him under my supervision and guidance. In my opinion, the submitted work has reached a level required for being accepted for examination. The Results embodied in this Major Project Part-II, to the best of my knowledge, haven't been submitted to any other university or institution for award of any degree or diploma.

Prof Usha Patel
Internal Guide & Associate Professor
CSE Department
Institute of Technology
Nirma University, Ahmedabad

Dr Sudeep Tanwar
Professor & PG Coordinator (M.Tech - CSE)
CSE Department
Institute of Technology
Nirma University, Ahmedabad

Dr Madhuri Bhavsar
Professor & Head
CSE Department
Institute of Technology
Nirma University, Ahmedabad

Dr Rajesh Patel
Director
Institute of Technology
Nirma University, Ahmedabad

Statement of Originality

I, **Khyati B. Kapadiya**, **20MCEC19**, give undertaking that the Major Project entitled **"AI and Blockchain empowered Health Insurance Fraud Detection"** submitted by me, towards the partial fulfillment of the requirements for the degree of Master of Technology in **Computer Science & Engineering** of Institute of Technology, Nirma University, Ahmedabad, contains no material that has been awarded for any degree or diploma in any university or school in any territory to the best of my knowledge. It is the original work carried out by me and I give assurance that no attempt of plagiarism has been made.It contains no material that is previously published or written, except where reference has been made. I understand that in the event of any similarity found subsequently with any published work or any dissertation work elsewhere; it will result in severe disciplinary action.

Khyati

Signature of Student

Endorsed by
Prof. Usha Patel
(Signature of Guide)

Acknowledgements

It gives me immense pleasure in expressing thanks and profound gratitude to **Prof.Usha Patel**, Associate Professor, Computer Science and Engineering Department, Institute of Technology, Nirma University, Ahmedabad for his valuable guidance and continual encouragement throughout this work. The appreciation and continual support he has imparted has been a great motivation to me in reaching a higher goal. His guidance has triggered and nourished my intellectual maturity that I will benefit from, for a long time to come. I am also Thankful **Dr. Sudeep Tanwar** ,PG Coordinator and **Rajesh Gupta** , Ph.D.Research scholar for stimulating discussion, guidance, suggestion and providing me with an opportunity to carry out my project with them esteemed guidance is very early stage of my project development.

It gives me an immense pleasure to thank **Dr Madhuri Bhavsar**, Hon'ble Head of Computer Science And Engineering Department, Institute of Technology, Nirma University, Ahmedabad for her kind support and providing basic infrastructure and healthy research environment.

- Khyati B. Kapadiya
20MCEC19

Abstract

Nowadays, health insurance is becoming an important part of people's lives, as the increase in the number of health issues. Healthcare emergencies can be troublesome for the people who can't afford the huge expenses. Health insurance helps people to cover expensive healthcare services in case of a medical emergency and provides financial protection against indebtedness risk. Health insurance along with its several benefits can face many security, privacy, and fraud issues. From the past few years, fraud has been a sensitive issue in the health insurance domain. Health insurance fraud incurs high loss for individuals, private firms, and governments. So, it is essential for national authorities and private firms to develop systems for detecting the fraudulent cases and payments. A high volume of health insurance data in electronic form is generated. This data is highly sensitive, and attracts hackers for cyber attacks. Motivated by these facts, in this research project, we introduce a AI and blockchain-enabled secure health insurance fraud detection system. This project introduces the security issue of health insurance. We proposed an intelligent architecture for health insurance fraud detection and a case study using machine learning and blockchain. We build ML model to predict a healthcare service provider is potentially fraudulent or not to prevent the financial loss. The model build using bagging classifier is able to successfully predict potentially fraudulent classes of healthcare service provider with an F1-score over 0.95. Furthermore, we develop smart contract based on HIC fraud detection system on blockchain technology to validate the transaction of HIC. The proposed smart contract deployed on the Remix Ethereum to protects transaction history to enhance security and HIC data privacy. Finally, the open issues and research challenges of blockchain and an AI-empowered health insurance fraud detection system are collated.

Abbreviations

Abbreviations	Explanation	Abbreviations	Explanation
HI	Health Insurance	SMOTE	Synthetic Minority Over-Sampling Technique
HIC	Health Insurance Claim	LightGBM	Light Gradient Boosting Machine
PHI	Personal Health Information	PCA	Principal Component Analysis
LEIE	List of Excluded Individuals and Entities	DNN	Deep Neural Network
SVM	Support Vector Machine	BILSTM	Bidirectional Long Short-Term Memory
PII	Personally Identifiable Information	PNN	Parallel Neural Network
HIPPA	Health Insurance Portability and Accountability Act	SNN	Simulated Neural Network
EHR	Electronic Health Record	ECM	Evolving Clustering Method
DDOS	Distributed Denial of Service	BERT	Bidirectional Encoder Representations from Transformers
MITM	Man In The Middle	L-SVM	Lagrangian Support Vector Machine
API	Application Programming Interface	GRPC	Google Remote Procedure Call
ML	Machine Learning	NS 3	Network Simulator
DBN	Deep Belief Network	ICD	International Classification of Diseases
RIPPER	Repeated Incremental Pruning to Produce Error Reduction	DPCNN	Deep Pyramid Convolution Neural Network
HAN	Hierarchical Attention Network	TEXT-RNN	Text Recurrent Neural Network
IPFS	InterPlanetary File System	LEAM	Label-Embedding Attentive Model
POS	Proof Of Stake	POW	Proof Of Work
POB	Proof Of Burn		

List of Figures

2.1 Revolutionary HI policy in india 9

3.1 Possible softwere attack on HI and countermeasure 18
3.2 Possible communication network Attack on HI and countermeasure . . . 19

5.1 Architecture of HI Fraud Detection [1],[2] 35
5.2 Case study : HI Fraud Detection from wearable device 38

6.1 Inpatient data 39
6.2 Outpatient data 40
6.3 beneficiary data 40
6.4 Provider Potential fraud 40
6.5 Feature Importance 42
6.6 Bagging Classifier Result 45
6.7 Logistic Regression Result 45
6.8 Support Vector Classifier Result 45
6.9 Random Forest Classifier Result 46
6.10 Area under Curve for four model 47
6.11 Smart contract function 48
6.12 set claim data transaction 49
6.13 Retrieve claim data transaction 49

List of Tables

1.1 Comparative analysis between existing review and proposed review on HIC fraud detection . . . 8

2.1 Types of HI fraud . . . 11

3.1 Possible security attack On HIC . . . 22

4.1 Comparative analysis of existing Unsupervised Learning based HIC fraud detection system . . . 28
4.2 Comparative analysis of existing Supervised Learning based HIC fraud detection system . . . 29
4.3 Comparative analysis of existing Hybrid approach(supervised and unsupervised learning) based HIC fraud detection system . . . 30
4.4 Comparative analysis of existing Blockchain based HIC fraud prevention system . . . 31
4.5 Comparative analysis of existing HIC Fraud detection system based on group of people . . . 32
4.6 Comparative analysis of existing AI and Blockchain based HIC fraud detection system . . . 32

6.1 Parameter of Model . . . 46
6.2 Performance of Model . . . 46

Contents

Certificate iii

Statement of Originality iv

Acknowledgements v

Abstract vi

Abbreviations vii

Abbreviations viii

List of Figures viii

List of Tables ix

1 Introduction 1

1.1 Scope of work 5

1.2 Motivation 6

1.3 Contribution 7

2 Background 9

2.1 Revolution of HI 9

2.2 Types of HI Fraud 11

2.3 Artificial Intelligence (AI) 12

2.4 Blockchain 13

3 Security Issue in HI 16

3.1 Interlinked Structure of Electronic Health Record (EHR) 16

3.2 Weaknesses of HI Portability and Accountability Act (HIPPA) 16

3.3 Cyber Security Attack 17

3.3.1 Software Attacks 17

3.3.2 Communication network attack 18

4 Literature Survey 23

4.1 Supervised Learning 23

4.2 Unsupervised Learning 25

4.3 Hybrid Approach 25

4.4 Blockchain Based HIC Fraud prevention 26

4.5 HIC Fraud detection from Group of people 27

4.6 AI and Blockchain-based HIC fraud detection 27

5 The Proposed System 33

5.1 User layer 33

5.2 Data generation layer 33

5.3 Data analytical layer 34

5.4 Blockchain layer 34

5.5 Case study 35

5.5.1 Real time analytics 37

5.5.2 Security and Privacy 38

6 Experimental Evaluation 39

6.1 dataset description 39

6.2 Experimental Methodology 41

6.2.1 Data Pre-processing 41

6.2.2 Machine Learning 42

6.3 Blockchain Implementation 47

7 Open Issues and Research Challenges 51

7.1 Class imbalance 51

7.2 Standardization of Data Exchange Rule 51

7.3 Real Time Data-set Accessibility 52

7.4 Handling Growing Number of User 52

7.5 Amalgamation with other Technology 52

7.6 Lack of Professional Talent 52

7.6.1 Weakness of Smart Contract Development 53

8 Future Scope and Conclusion 54

Chapter 1

Introduction

Health insurance (HI) is a contract between the insurance provider and insurance subscriber, in which the insurance provider compensates the insurance subscriber's healthcare expenses. The Health insurance association of America stated that healthcare insurance covers losses resulting from the accidents, healthcare expenses, incapacity, accidental injury, and damage [3]. Insurance subscribers have to pay the premium regularly for this compensation. The insurance provider can be either from the commercial world or a government body. Nowadays, HI becomes a necessity of each individual due to the rising hospitalization and treatment costs and getting income tax rebates [4].

Earlier, the health insurance claim (HIC) process was manual and offline that has many shortcomings, which are described as follows. Insurance subscribers need to visit the insurance office during the office hours only for filling out the premium, inquire about the HIC status, and get HI coverage, which wastes time and money in terms of transportation costs for the insurance subscribers. This procedure is completely based on paper, so human resource necessity and the possibility of error are more for auditing HIC. Maintenance and integration of the paper-based health claim data is very tedious and challenging work. Health claim records are easily alterable and accessible. So the chances of fraud occur from the insurance provider, insurance subscriber, and healthcare service provider due to the lesser transparency and privacy. It is less cost-effective due to the involvement of intermediary broker or agent costs [5].

In digital era, every piece of information is gathered in a digital form, which revolutionizes the HIC worlwide. Following are various benefits of digitization: (i) It provides convenience to all the parties involved with HIC. Insurance subscriber does not need to visit the insurance office frequently to purchase the HI and to fill out the premium amount [6].(ii) Communication between insurance subscribers and insurance providers becomes efficient using digital technology. (iii) HIC auditor's complex and tedious work becomes easy through digitization. (iv) Fraudulent behavior can be easily identified in HIC using advanced AI technology. (v) Digitization in the HI sector reduces the human resource cost [6]. (vi) HIC verification becomes fast using web-generated reports, so insurance subscribers get insurance coverage fast and automatically during any medical emergency. There are many other benefits of the digitization of HIC besides those mentioned above.

Despite several benefits of digitized HIC, it faces various challenging issues. Following are some of the issues of digitized HIC.

- *Validation of data and model*: To determine the premium rate and coverage price, insurance providers use a digital business model. This model is developed by a professional who may be unfamiliar with HI rules and with the specific requirements of HI, which may cause professionals to find it difficult to measure the impact of new variables used in the models [7]. In web-based HIC, data are collected from various social media platforms that may be inaccurate, outdated, and incorrect for validating digital business model [7]. The subscriber does not provide the data generated from the social media platform to the insurance provider, and the insurance subscriber does not have a chance to correct this social media platform data, which may be used for determining the premium rate.

- *Lack of talent*: Developing web-based HI claims needs a group of skilled people because it is dependent on complex algorithms and mathematical skills [8]. Lack of talent leads to the insurance sector becoming expensive.

- *Fraud detection system* : Every insurance plan, including HI, is vulnerable to fraud. Each year, HI provider firms lose revenue due to fraudulent claims. Insurance firms hike premiums to maintain profit, which impacts legitimate insurers. It is assessed that fraudsters approximately take fifteen percent of the taxpayer's money, which is utilized to finance government-assisted medicare. So this is necessary for national authorities to develop systems for detecting fraudulent cases and payments. HI fraud is a serious offense that affects people, nations, and a lot of money and time. As a result, a good fraud detection system is necessary for lowering costs and enhancing the security of healthcare [3]. HI fraud is increasing day by day, which is a concern for insurance subscribers and insurance providers of the nation. As there was no option for a HI fraud penalty before, the number of fraudulent cases is increased [8].

- *Connecting to outdated computer system*: Insurance providers need to replace legacy computer systems or customize their systems to interface with new technologies properly, which is costlier. Legacy computer systems was developed around satisfying regulatory requirements rather than enhancing the subscriber's experience [7].

- *Privacy of HI subscriber's data*: The insurance provider's expanded the use of subscriber's data that raises concerns about the privacy of the data. It may not be possible for an insurance subscriber to know exactly what or when data is collected and how data is being used. So, insurance subscriber does not provide explicit consent to the insurance provider, and insurance subscriber loses control over their personal information [7]. So, the identity threat is a big issue in which insiders misuse insurance subscriber's identities for getting insurance coverage.

- *Security of HI system*: Digitization in HI raises various security concerns such as ransomware attacks, phishing attacks, Distributed Denial of Service (DDoS) attacks, replay attacks, etc. Cybercrime affects the HIC industry from both internal and external sources, including the third parties [9]. HIC data is stored in various systems, and it is interlinked between systems which causes authentication and authorization problems. Insurance firms lose lots of revenue, money, and reputation due to the compromised security of the HI system.

To overcome the aforementioned security issues, researchers have given various cryptography-based solutions. Few of the works are: Lou *et al.* [10] presented the access control techniques for patient-related data based on attribute encryption. This data is shared among insurance firms, hospitals, and patients. The encryption technique is one-to-many, and ciphertext can be accessed by a group of people who fulfill certain access policies. Then, Heurix *et al.* [11] discussed pseudonymization and personal metadata encryption techniques for handling and transmission of EHR, personal health record, and billing record of patients. Usage of the encryption technique prevents unauthorized data disclosure and assures data security from internal attackers by decreasing sensitive data leakage. Later, Kumar *et al.* [12] developed an encoding algorithm for storing patient's private data on a server. Admin can access the patient and hospital-related information. Encoding of data is done by implementing the 128 base encoding method, and decoding is done with the help of base64 decoding method. Later, Bellisario *et al.* [13] utilized a X.509 certificate for secure communication between insurance providers / healthcare providers and the authentication server. They encrypted the EHR with a 256-bit AES key according to the hash value. The first-time hash value is stored in the encrypted configuration file for future use, giving centralized access to EHR.

Therefore, even though the solutions mentioned above offer security to the HIC data, it has several disadvantages. The central server used in the aforementioned security solutions could crash due to malicious attack or fault, which affect HIC business operation and work location availability [13]. Another disadvantage is immutability which means data can't be removed or altered after being stored. The traditional security solutions proposed by the authors do not provide immutability and transparency, which can lead to the data tampering in the system [13]. Authentication of user's information depends on the admin, and if the admin's security is compromised, the entire system would be vulnerable to security attack[12]. Data redundancy is also a demerit because multiple copies of the same data of HIC exist in the database. The aforementioned problems of the security-based solution can be resolved with the help of blockchain.

Blockchain distributed ledger allows the group of peers to collaborate to form a single, and decentralized network. Consensus mechanisms are used for communication and sharing data among peers [14]. Blockchains store all transactions on a public ledger in the form of chained blocks without a central server [15]. Blockchain has a decentralized network that enables access to HIC records to all the parties involved in the HI without the involvement of any central authority to oversee the global HI data. The data saved on the blockchain cannot be corrupted, altered, or retrieved to ensure immutability and transparency in the system [15]. It enhances the security of the data utilizing the encryption techniques using cryptographic hash keys and timestamp protocols. The availability of HIC data saved on the blockchain is ensured since the records on a blockchain are duplicated in numerous nodes, and the system can be protected against security attacks on the HIC confidential data availability. It can verify the authenticity of blockchain data without accessing the plaintext of those records [15]. From the discussion of various issues of digitized HI, in this paper,We are mostly focused on HI's fraud, security, and privacy issues.

1.1 Scope of work

In this section, we discuss the scope of the survey and comparative-analysis between the existing survey and the proposed survey on HIC fraud detection.

Many researchers have presented surveys on HI fraud detection. Most of these surveys have given insights into the processing and integration of HIC datasets, data mining, and machine learning (ML) techniques for fraud detection in HI. For example, Duman *et al.* [16] presented the big data analytics aspect in HIC fraud detection. The study provides detailed information about the HIC data source, benefit, and characteristics of the algorithm and model used for HI fraud detection. Then, Thomas *et al.* [4] addressed the supervised learning technique for fraud detection systems that have been optimized in the HI industry to identify the fraudulent claims. Later, Bauder *et al.* [17] analyzed the various CMS medicare and LEIE (List of Excluded Individuals and Entities) datasets, which are being processed for identity fraud in HIC. CMS medicare dataset summarizes the information about healthcare services providers. It also gives information about medical procedures prescribed to HI beneficiaries and prescription drugs managed by the physicians under the Medicare Part D Prescription Drug Program. LEIE dataset contains information about the fraud labels. They also analyzed mapping between the LEIE dataset and CMS dataset to identify the HIC fraud detection. Further, Ekin *et al.* [18] discussed statistical HIC fraud assessment using sampling, overpayment estimation, data mining technique, and the Bayesian approach. They also demonstrate various unsupervised learning techniques for fraud detection in HIC utilizing a real-time dataset.

Later, Ankrah *et al.* [19] presented an exhaustive survey on Ghanaian HIC Dataset. They introduce the dataset feature of this database to the researcher in the context of ghana's healthcare system. This dataset was categorized into six groups, i.e., the information of the client, services offered, diagnosis, investigations, medicines, and client summary. Then, Chen *et al.* [20] described several types of HI fraud and explained recommendations techniques for combat HI fraud. They also explored healthcare fraud cases of pharmacogenetic testing. Later, Mary *et al.* [21] investigated the imbalance classification issue occurred in the method solution for the HI Fraud detection. In an imbalanced classification problem, data distribution across known classes is biased. The comprehensive research study shows that the class imbalance in support vector machine (SVM) classifier is more influenced than decision trees and neural networks classifiers. Further, Magalingam *et al.* [22] presented various data mining methods for fraud detection in finance applications such as insurance fraud, bank fraud, corporate fraud, and cryptocurrency fraud. They found SVM extensively utilized in fraud detection of financial applications from their survey. The research study also shows the list of nations vulnerable to financial fraud.

In this proposed review, we have discussed major security issues and their countermeasures in HIC and proposed blockchain and AI-empowered architecture for HIC fraud detection. Table 1.1 shows a comparative analysis between the existing review and the proposed review on HIC fraud detection.

1.2 Motivation

The motivation of this project is as follows.

- The importance of security, privacy, and fraud detection in HI is key criteria. Without the security and privacy of the HIC system, patient's sensitive personally identifiable information (PII) can be compromised, which can ruin the insurance firm's reputation. Fraud in healthcare insurance causes loss for individuals, private firms, and governments. So secure fraud detection methods for HIC have become a necessity.

- In the existing survey, security issues and HI fraud detection architecture are not discussed. So, there is a need for a comprehensive survey that inspects the secure AI and blockchain empowered HI fraud detection system.

- In the existing paper,bagging classifier(ensemble learning) is not used for HIC fraud detection.

1.3 Contribution

This paper presents a detailed survey of HI fraud detection. We highlight open issues and research challenges in HI fraud detection. The research contributions of the paper can be represented as follows:

- The background and security issues of HI have been discussed. We also discussed possible software attacks on the HI systems and their countermeasure tools.
- We proposed a blockchain and AI-based architecture of HI fraud detection to overcome security issues in HI.
- We also proposed a case study on HIC fraud detection from wearable devices, a futuristic approach for the proposed system.
- We proposed machine learning model for detection of potential fraudulent healthcare service providers.
- HIC fraud detection system-based, Smart Contract on blockchain technology is proposed to validate the transaction of HIC.

Table 1.1: Comparative analysis between existing review and proposed review on HIC fraud detection

Author	Year	Contribution	Pros	Cons
Duman et al. [16]	2017	A review on fraud detection techniques and big data analytics aspects in HI.	Provide significant HIC dataset related information	Only ML based solution discussed
Thomas et al. [4]	2018	Present a review on supervised learning technique's impact on HI fraud	Useful for labeled data	Not given information about tools
Bauder et al. [17]	2018	A survey on medicare dataset integration and processing of fraud detection system	It guides researcher about research gap in medicare dataset processing and integration	Does not analyze Security of dataset
Ekin et al. [18]	2018	A review on statistical healthcare fraud assessment	It provides useful information for the complicated nature of healthcare insurance data and the heterogeneity of healthcare systems.	Not given detailed security aspect in statistical assessment
Ankrah et al. [19]	2018	A Review on HI claim dataset of Ghana	Helpful in drug utilization research aspect	Limited discussion about HI fraud detection from dataset
Chen et al. [20]	2020	A review on recommendations method for healthcare service providers and patients to combat Fraud	Provide awareness information about healthcare fraud	Lack of detailed information about fraud prevention method
Mary et al. [21]	2021	Study present class imbalance issue of ML method in detection of healthcare fraud	The researcher can easily identify affecting factor of class imbalance	Not discussed detail solution of class imbalance issue
Magalingam et al. [22]	2021	A Review on data mining technique for fraud detection in all types of finance application	Cover useful information about the dataset and validation measures for estimating the performance of data mining techniques.	Limited information about HI fraud
Proposed Review	-	A review on blockchain and AI-empowered HI Fraud Detection	Present security issues, solutions, and architecture for HI fraud detection	-

Chapter 2

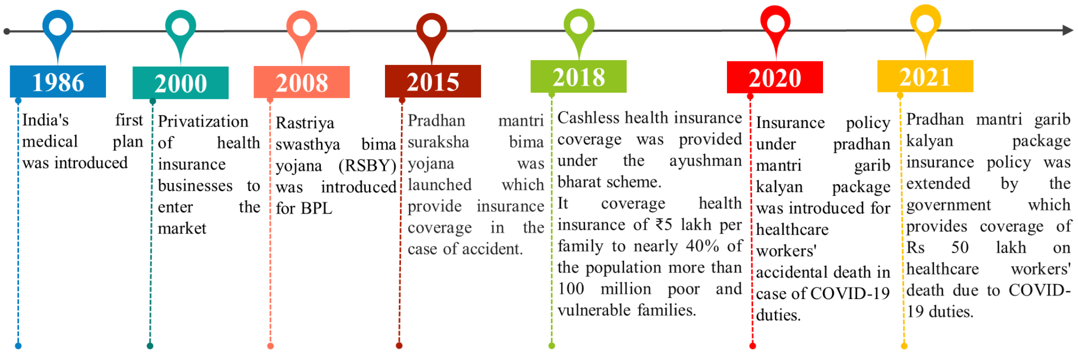
Background

This section provides the background of HI and HIC fraud types and detection of the HIC fraud.

2.1 Revolution of HI

HI is the most effective method of financing healthcare to preserve a person's money and prevent indebtedness. Healthcare insurance helps subscribers to pay for expensive healthcare services in case of a medical emergency [21]. In several countries, healthcare is becoming a considerable expenditure. It is considered to be crucial for the life of several residents in a particular country. Many patients receive inadequate healthcare services due to the unaffordable high medical service costs [23]. HI has already become an important part of people's life. Healthcare spending has been steadily rising, so this has already become a worldwide phenomenon[24].

In India, HI has become the important part of the economy. The availability of medical services in India differs by state. Most states have public medical services;



1986: India's first medical plan was introduced

2000: Privatization of health insurance businesses to enter the market

2008: Rastriya swasthya bima yojana (RSBY) was introduced for BPL

2015: Pradhan mantri suraksha bima yojana was launched which provide insurance coverage in the case of accident.

2018: Cashless health insurance coverage was provided under the ayushman bharat scheme. It coverage health insurance of ₹5 lakh per family to nearly 40% of the population more than 100 million poor and vulnerable families.

2020: Insurance policy under pradhan mantri garib kalyan package was introduced for healthcare workers' accidental death in case of COVID-19 duties.

2021: Pradhan mantri garib kalyan package insurance policy was extended by the government which provides coverage of Rs 50 lakh on healthcare workers' death due to COVID-19 duties.

Figure 2.1: Revolutionary HI policy in india

however, due to a shortage of resources and administration, most citizens prefer private medical services [25]. India consists of global, and single-payer medical system financed by either public or private HI funds and a section of government medical centers that is almost entirely paid by taxes [25]. India has introduced a lot of revolutionary HI policies. Figure 2.1 shows the revolution of HI policies year-wise.

The Constituent Assembly approved the National Health Policy during 1983, which has been further modified during 2002, and 2017 [25]. The update in policy is required due to the growth in diseases and rising incidences of unaffordable expenditure because of medical services costs. India's first medical plan was introduced in 1986 to regulate the terms and conditions of the HI [3]. It covered inpatient costs but excluded existing conditions such as pregnancy, maternity, and HIV. The payments were reimbursed automatically by third-party administrators. India's constitution introduces a new economic policy in which the insurance sector was privatized in 1991[3].

The Indian government liberalized insurance in 2000, which allowed private businesses to enter the market. With the arrival of private health insurers in India, various novel policies and plans such as family floater plans, top-up plans, critical illness plans, and hospital cash were introduced [3]. In the year 2008, the Rashtriya Swasthya Bima Yojana was introduced, in which HI plan aims to ease the progressive rollout of HI projects for BPL workers in all districts of the states [25]. Pradhan Mantri Suraksha Bima Yojana was introduced in 2015. It is a government accident insurance policy in India. Insurance subscribers get coverage of Rs 2 lakh in unfortunate death and Rs 1 lakh in the incident of severe lifelong disability under this policy [26]. In 2018, the Indian government started Ayushman Bharat, a countrywide taxpayer-funded HI scheme. This program intends to address the lower half of the nation's population and provides individuals with the free medical care in both public and private medical centers [3]. In 2020, an insurance policy was launched under Pradhan Mantri Garib Kalyan Package for healthcare workers fighting COVID-19. This policy provides coverage of Rs 50 lakh to private and government healthcare workers who may have accidental death in case of COVID-19 duties [27]. This policy was launched for around 22.12 lakh healthcare providers. In 2021, the government extended the Pradhan Mantri Garib Kalyan Package Insurance policy which provides insurance coverage to healthcare workers who may be suffering from severe disease or death due to the COVID-19 duties.

2.2 Types of HI Fraud

Medical insurance fraud is a serious subject to each country, and the forged behavior patterns vary according to the situation. Various types of HI fraud occur in each country, and various parties are involved with this fraud. There are mainly three parties engaged with this fraud. First, healthcare service providers such as doctors, hospitals, ambulance firms, and laboratories. The second is subscribers such as patients and their employers. The third is HI provider's fraud which includes private insurance firm, and the government sector [16]. Table 2.1 shows the different types of HI fraud.

For decades, fraud has been a significant problem and may be found in any industry.

Table 2.1: Types of HI fraud

Healthcare service provider's fraud	**HI subscriber's fraud**	**HI provider's fraud**
Medical services that have not been performed are being billed [28].	To get a lesser premium rate, misrepresent eligibility documents [16].	Charge of premium taking more from insurance subscriber by falsifying subscriber's claim [29].
Expensive services and medical tests are being billed that are costlier than the original test [28]	Making claims for medical treatments that were never provided [16]	Claims are being blocked without an examination of the claims' authenticity [30].
For the goal of gaining insurance coverage, presenting non-covered therapies as medically not required [28].	Obtaining financial compensation by impersonating another person's policy card [28].	To discourage the insurance subscriber, the HI provider incorrectly dismisses legitimate claims in the expectation that the patient will soon quit [30].
Misrepresentation of a specific diagnosis or history of therapy [16].	Actively participating with scam networks through buying several insurance policies [28].	Useless and forged claims are created by insurance providers for collecting premiums from insurance subscribers [30].

Every time, fraudsters used the new technique to commit fraud. Individuals, businesses, organizations, and government may face some serious losses due to the healthcare fraud. It is also predicted that the $600 to $850 billion per year is lost because of fraud in medicare system of United States and face loss of approximately $125 to $175 billion of the total amount due to the forged activities involved in the system [24]. As a result, combating fraud is becoming a priority. Several experts propose fraud detection methods in HI. Fraud can be attempted by fraud identification method at the time of their occurrence. Manually fraud monitoring and automatic fraud monitoring are the two most common procedures for detecting fraud. Monitoring fraud procedures requires exhausting human labor. This procedure contains complicated transactions which takes more time and involve a wide range of field experience and expertise. As a result, automated techniques have been developed to enhance the effectiveness of fraud detection. In automatic fraud monitoring, various data mining computer-based methods are involved [16].

2.3 Artificial Intelligence (AI)

Artificial Intelligence (AI) techniques have been utilized as a valuable tool for HIC fraud detection. AI automates the HIC fraud detection system. As per the recent studies, AI has been mainly used to solve HIC fraud detection using several ML, deep learning, and data mining models [31]. Behavioral profiling methods based on ML techniques are used to detect anomalies and fraud detection. For this purpose, each individual's behavior pattern is modeled to monitor it for any derivation from norms [16]. ML techniques used in HIC fraud detection are categorized into supervised learning, unsupervised learning, and semisupervised learning [16].

Supervised learning technique in HIC fraud detection uses a dataset of previously known fraudulent and legitimate records. Those records are utilized to capture fraud patterns and build the model. The essential benefit of the supervised learning technique is that the classification results given by this technique are easy to comprehend. Various classification algorithms and supervised learning regression analysis algorithms are used in HIC fraud detection. Neural networks, SVM, decision trees, bayesian approaches, statistical analysis, graph analysis, and rule-based methods are supervised learning techniques used to identify HIC fraud [32]. Apart from its benefits, supervised learning has many drawbacks. In supervised learning techniques, data collection and data generation are challenging tasks. If the dataset of HIC is quite huge, then labeling can be a difficult task to perform. When labels are uncertain and ambiguous, it is difficult to identify between them in fraud data [16]. In some circumstances, supervised learning implementation becomes challenging due to these constraints. These constraints of the supervised learning can be solved with the help of unsupervised learning.

Unsupervised learning techniques identify the fraudulent behaviors of HIC from the unlabeled HIC datasets. The benefit of HIC fraud detection using unsupervised learning is that there is no need for labeled data [16]. Unsupervised learning is used in HIC fraud detection where labeled data is not available. There are various unsupervised learning techniques utilized in the HIC fraud detection, such as association rules mining, data mining, k-means clustering, and k-nearest neighbor. Semi-supervised learning is used to get benefits from both supervised and unsupervised learning. These hybrid semi-supervised learning are used where a limited number of labeled data is available compared to unlabeled data. A predictive model is built using both labeled and unlabeled data in semi-supervised learning [16]. Combined clustering and classification as a semi-supervised learning method is generally used in HIC fraud detection.

2.4 Blockchain

The security issue is one of the key concerns in HI. Existing centralized systems provide security to the certain extent, but they could crash due to malicious attacks or faults. This issue can be solved with the help of a decentralized network, i.e., blockchain. The idea of blockchain was discovered in Satoshi Nakamoto's proposal for virtual currency bitcoin. Blockchain is a digital ledger of transactions stored in a chain of blocks. A cryptographic hash of the previous block, a timestamp, block header, block number, version, nonce, and transaction data are all included in each block. Transaction data is maintained as a Merkle tree [33]. In the blockchain, if any of the transactions in a block are modified or altered, leading to the drastic change in the hash value of the particular block. It further breaks the chain of blocks on the blockchain, which helps to detect the modified transaction. Hence, a transaction can't be modified or altered once it has been added to the blockchain. As a result, data on the blockchain is immutable. Every transaction has been digitally signed with a private key utilizing asymmetric cryptography to assure integrity, security, and immutability [33]. Only the public key owner can authenticate the signing entity. These signatures prevent tampering with the transaction data [34].

The blockchain communication network is decentralized and peers to peer, which solves the problem of a single point of failure. Blockchain utilizes a consensus protocol instead of a central authority to settle disagreements between nodes in a distributed application [33]. A consensus protocol is used in the blockchain networks to ensure reliability and trust between unknown peers in a distributed computing environment. In the consensus protocol, all peers in the blockchain network agree on the distributed ledger's current state. There are four types of blockchain: public, private, consortium, and hybrid. In a public blockchain, everyone with internet connectivity can sign on the blockchain to become an authorized node making public blockchain permissionless [35]. Permissions are required to use a private blockchain. It is only open to those who have the network administrator's permission. A hybrid blockchain combines the benefits of both centralized and decentralized blockchain [35]. Consortium blockchain contains the combined features of a private and public blockchain. There are two types of nodes in the blockchain based on their role, i.e., whether it works as a full node or as a normal node. A full node(mining node) keeps a copy of the entire transaction record and participates in the validation and authentication process, including signature verification, and mining [33]. In a blockchain network, full nodes are the backbone of trust. Because it is in charge of implementing consensus norms and processes, the normal node could create and transfer transactions while maintaining the blockchain's header. It can not participate in the validation process. [35]. The attack that happens on any one node in the blockchain network does not affect the ledger's status due to the data in the blockchain ledger being replicated across all nodes in the network. One of the HIC fraud detection system requirements is the storage of HIC data that are not alterable [35]. The immutability feature of blockchain fulfills aforementioned requirement. Preserving the integrity and authenticity of the insurance subscriber's (patient) records is critical. Insurance subscriber's data on the blockchain is encrypted using cryptographic techniques, which assure that only insurance subscribers with genuine authorization can access and decrypt the data. It further enhances the data security and privacy of the HIC. Insurance subscriber's data can be exchanged among all parties involved with HIC fraud detection without revealing the insurance subscriber's identities because the identities of insurance subscriber's in a blockchain are pseudonymized using cryptographic keys [15]. Blockchain supports smart contract, which is written in the form of an executable code, and designed in solidity language, and deployed on Ethereum Virtual Machine to regulate activities following the agreement [15].

It contains transaction logic that governs the whole HI business process. Smart contract methods are executed once the transaction commits the operation. A smart contract defines rules for the HIC process that allows insurance subscribers to control how their HIC records are shared or used in the network [15].

Chapter 3

Security Issue in HI

The high volume of healthcare data in electronic form is generated due to the advancement in the technology. This data is very sensitive, which attracts the hackers for cyber attacks. So security is a major issue in HIC. Following are major security issues in the HIC.

3.1 Interlinked Structure of Electronic Health Record (EHR)

The EHR contains sensitive data that contain patient's medical information. Physicians, healthcare workers, and insurance firms are sharing crucial healthcare data of patients using EHRs. This will be easier to manage medical services and deal with insurance issues, but this interlinked structure of EHRs creates a security problem. Because of the interlinked structure of EHRs, attackers can access this record which has been gathered for years under patient's identities. Patient records must be shared to provide better treatment, yet this will make networks vulnerable [36].

3.2 Weaknesses of HI Portability and Accountability Act (HIPPA)

HIPPA was established for the privacy and security of the healthcare data in the USA in 1996. HIPAA's rule covers how to use and how to release healthcare data along with the personal information. This rule ensures the security of the personal information while granting the flow of healthcare data, which is needed for medical choices. HIPPA rule applies to the HIC, and healthcare service providers who share healthcare data with

HI provider [37]. HIPAA does not include Fitbits, Apple watches, fitness, and fertility applications. This rule does not protect the huge amount of data that people generate through their usage of digital devices, apps, ecommerce sites, and social media. HI firms, healthcare service providers, and insurance subscribers have easy access to this data [38].

3.3 Cyber Security Attack

The Healthcare system provides life-saving care to patients using advanced technology. The high volume of healthcare data in electronic form is generated . This data is very sensitive, and it attracts hackers for cyber attacks. This healthcare data is used in HIC. The network of HIC is also vulnerable. It is an open channel for communication, so there is a possibility of network attacks on the data. Various attacks can be possible on HIC to breach security. We categorize possible security attacks on HIC into two categories: 1)software attack 2)communication network attack. The software attacks are a) Phishing b) Ransomware c) SQL Injection d) Malware and e) Bruteforce. Software attacks are performed on the back-end of the HIC software to change its control and operations. The communication network attack are: a) Man in the Middle b) DDoS c) Evesdroping d) Replay and e) Impersonation. Communication network attacks are performed on the HIC network to obtain unauthorized access of HIC data. Table 3.1 describes possible security attacks and consequences on the HIC category-wise.

3.3.1 Software Attacks

In a software attack, attackers design malicious programs to harm HIC system computers and servers. Utilization of the software has completely transformed in the healthcare insurance industry, which undoubtedly helps insurance subscribers and providers. However, there is no adequate security measure to ensure that HI software functions in the authentic manner. As a result, HIC systems are vulnerable to various software and app-related risks, including phishing, ransomware, SQL injection, malware, brute force. The countermeasure tools like netcraft, open SSL, iperf tool, SQL rand, droidMat, specter, iptables, etc are used to protect the confidentiality, integrity, and availability of HIC data and prevent from software attacks. Machine learning and deep learning technology are used to prevent software attacks. Machine learning algorithms such as naive bayes, sequential covering algorithm, random forest, decision

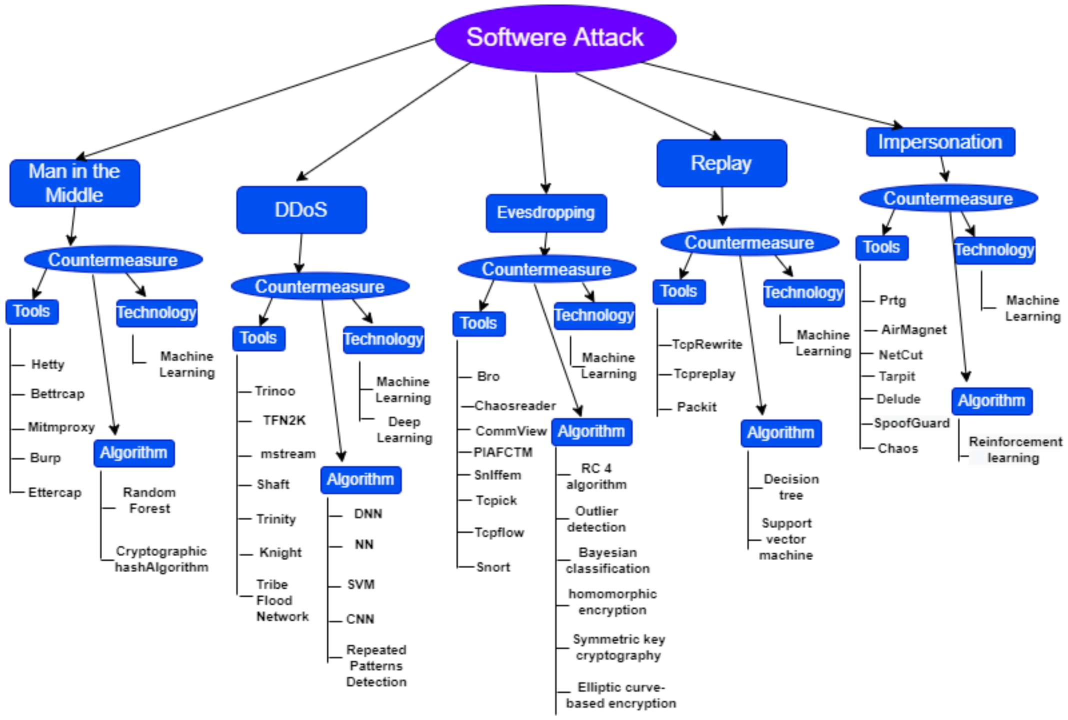


Figure 3.1: Possible softwere attack on HI and countermeasure

tree, clustering, etc are utilized for the countermeasure of software attacks. Figure 3.1 shows possible software attacks on HI and the associated countermeasure tools, technologies, and algorithms.

3.3.2 Communication network attack

Patient's remote monitoring, diagnosis, treatment, and emergency support are all possible with rapid development in the wireless communication. A patient's medical data is generated from wireless communication, which various users share and access, including healthcare professionals, researchers, government organizations, and insurance providers. Attacks on wireless communication seem to be a key concern for the healthcare system. Attackers use eavesdropping, replay, impersonation, DDoS to compromise the system's integrity and authenticity. Tools for the countermeasure of communication network attacks are burp, ettercap, trinity, tcpflow, tcpick, snort, tcprewrite, netcut, etc. Machine learning algorithms such as random forest, DNN, SVM, CNN, decision tree, reinforcement learning, etc are utilized to countermeasure of communication network attacks. Various cryptographic countermeasure algorithms such as symmetric-key cryptography, homomorphic encryption, cryptography hash algorithm, etc are used to prevent from communication network attacks. Figure 3.2

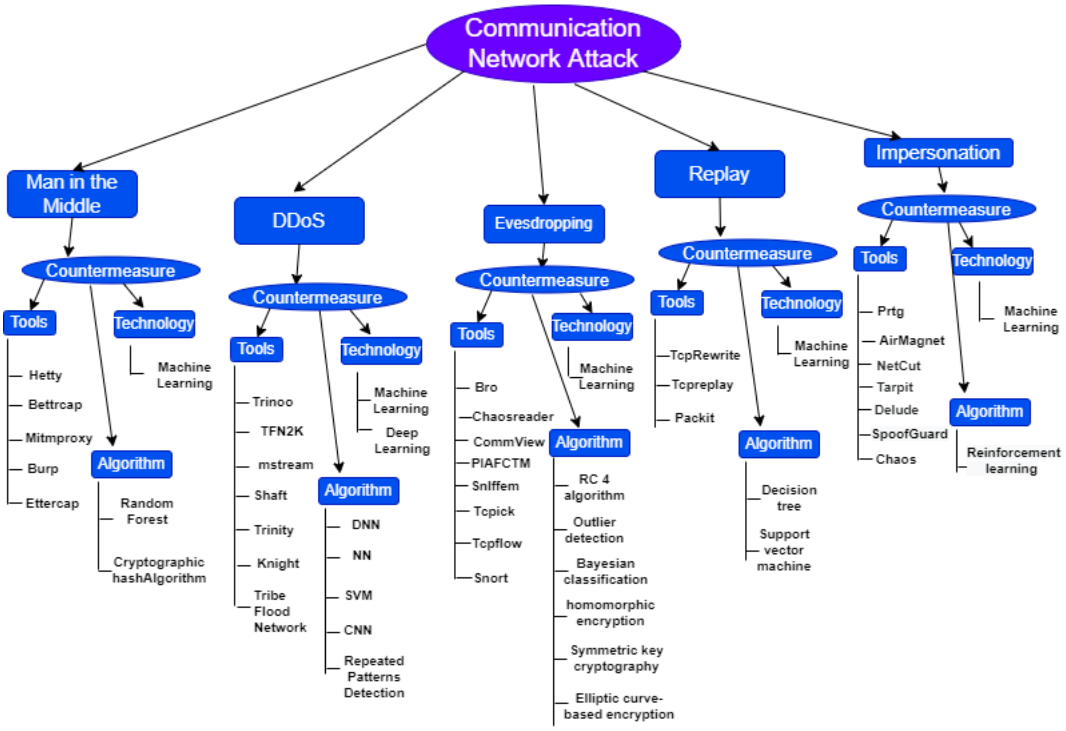


Figure 3.2: Possible communication network Attack on HI and countermeasure

shows the possible communication network attacks on the HI and the countermeasure tools, technologies, and algorithms.

Communication network attacks are classified into active and passive attacks according to the attack's involvement in the HIC system. In active attack, the attacker attempts to modify the content of the communication message. DDoS, man in the middle (MITM), and replay attacks are active attacks.Active attacks on the HIC system compromise integrity and availability of the HIC system. In a passive attack, attackers may eavesdrop to analyze and replicate the communication and use it maliciously. Eavesdropping and impersonation attacks are passive attacks. The passive attack can compromise confidentiality of the HIC system.

We address the several countermeasure tools technologies and algorithms offered by the researchers to protect against the possible attacks described in the table 3.1. For example, Kok *et al.* [39] proposed a two-stage pre-encryption ML algorithm to detect ransomware. In the first phase, they examined the malicious program's application programming interface (API) using ML. This method was employed to guarantee the detection of both existing and undiscovered crypto-ransomware. In phase two, the signature repository is created in which the signature of the predicted crypto-ransomware is generated. The signature repository enables the identification of crypto-ransomware efficiently, using the signature matching method at the pre-encryption stage. Then, Singh *et al.* [40] propose a SQL injection detection technique using a clustering ML algorithm and identify the unauthorized users by maintaining an audit record. An audit record is being used to compare the attack intensity and detection probability to estimate the detection accuracy. Later, Aleroud *et al.* [41] classify the phishing countermeasures based on ML, text mining, human users, profile matching, ontology, honeypot, search engine, and client server-based authentication. Then, Datta *et al.* [42] suggested the malware countermeasure for the android platform. They suggested an android security extension that gives extremely quite security controls over the android apps. It enhances the android authorization model by giving control over which type of contacts access and which sites an app can connect to over the internet. Later, Sadasivam *et al.* [43] proposed honeynet architecture for the detection of distributed brute force attacks. They showed attacker's login attempts determined by the behavioral index. All the single-source attacks are identified according to the behavioral index.

Later, Shyam *et al.* [44] proposed a SaaS framework for DDoS attack mitigation based on a deep belief network (DBN). The weight and activation function of the DBN are fine-tuned using the median fitness sea lion optimization technique. When DBN identifies an attack node, control is passed to a lightweight bait technique that successfully reduces the most frequent attack. Then, Zagrouba *et al.* [45] explained data integrity, routing security, and data privacy countermeasure for MITM attack. For data integrity and routing, security authentication and encryption method are useful. Cryptographic hash functions are the greatest defense from the data integrity attacks.

Later, Yousefpoor *et al.* [46] reviewed several eavesdropping attack's countermeasure encryption methods which ensure the data privacy and data confidentiality across a wireless sensor network. They discussed the homomorphic encryption method, which can enable the sensor nodes to secure their confidential information. An attacker will not decrypt data packets if he obtains the private keys of all sensor nodes. As a result, if an attacker eavesdrops on a communication channel, it can only see the encrypted data packets without accessing the complete information. Then, Waqas *et al.* [47] presented a reinforcement learning-based method to secure the system against impersonation attack. They used channel gain to identify impersonation and generated valid secret keys between authorized users. Later, Bruce *et al.* [48] proposed a security solution for e-healthcare that mitigate the replay and impersonation attacks. They also designed a protocol ensure the security of the system against impersonation attack. If an adversary tries to steal the certificate to access the network, then the designed protocol can be used to verify the authenticity and take action accordingly. After the authentication procedure, the interacting entities generate a session key. This key is unique to each session and cannot be used again once it has ended, preventing replay attacks.

Table 3.1: Possible security attack On HIC

Security attack	Description	Consequence
Phishing	Attackers send mass amounts of email with malicious links to users or employees of HIC for getting credential information of the user [41].	Disclose user's PII information and HI data,After getting credential detail, launch malware
Ransomware	Disrupts a computer's sensitive functionality by encrypting a user's computer's sensitive file until he gets ransom money[49].	Disrupt insurance business operations and harm the reputation of the firm
SQL Injection	Attacker uses SQL to execute malicious code into a server, the server is forced to disclose secured data [50]. HIC website comment section and search bar are used to execute malicious code .	Attacker gets unauthorized access to the HIC database through SQL injection [50]. Database contains sensitive PII information. The attacker can perform malicious activity and fraud in HIC with this information.
Malware	Malware is a malicious software program written to interrupt a computer, server, or communication network [51]. Malware-infected HIC systems might diverge from their intended functions, such as delaying or shutting down.	An attacker can disclose private information of the HIC system and obtain unauthorized access to information or systems, deny users access to information, or inadvertently compromise a user's computer security and privacy.
Bruteforce	Attackers get credentials of the HIC system by guessing all combinations and getting unauthorized access to the system [52].	Attackers steal PII-related data and execute malware to the network through unauthorized access. HI firm's reputation is also ruining
Man-in-the-Middle	Attackers intercept communication between server and client, and the victim is clueless about the attack. This attack can happen between parties involved in HIC [53].	An attacker obtains confidential information from the victim by hiding identity.
DDOS	Attackers overload server with fraudulent service requests containing fraudulent return addresses, distracting the server during authenticating service requests [54].	It causes server down and does not enable service request processing. System goes completely offline. This attack harms the insurance firm's service.
Eavesdropping	Attacker attempts to steal HIC data from a smartphone while the user is sending or receiving data over a communications channel by exploiting the vulnerability of the communication channels [46].	The attackers can be sold important corporate and financial data of HIC for malicious activities.
Replay	A replay attack occurs when an attacker intercepts the HIC network connections and falsely resends them to misdirect the receiver.	If an attacker eavesdrops on HI firm's financial transaction through a replay attack, the firm will lose money.
Impersonation	The attacker masquerades a valid user in the communication network to obtain access to the victim's confidential data.	Email of HI business compromise refers to impersonation frauds in which a victim is misled into completing a cash transfer.

Chapter 4

Literature Survey

In this section, we categorize the taxonomy of HIC fraud detection. In the literature, various possible solution for fraud detection in HI was given, which is discussed in the following section.

4.1 Supervised Learning

Supervised learning is well known for fraud detection. This method describes a comparison of newly arrived claims with the pre-trained models. In this method, the model is trained using data associated with the labels. The trained model can predict the newly arrived HIC data's class. The model can be built using a training dataset. The genuine and fraudulent claims could be the labels of class in the system to detect the medicare fraud [55]. If a claim follows a similar label, it can be classified as genuine. Otherwise, it can be classified as the fraudulent.

Later, Kose *et al.* [28] developed a framework that detects the HI abuse cases independently from the actors and commodities. They defined actors as the insured individuals, physicians, or institutional healthcare service providers and commodities as medication or healthcare services. They used the ZeroR classification supervised learning technique for fraud detection. They also utilized proactive and reactive analysis. The framework includes a visualization tool that significantly reduces the time requirements for the users during the fact-finding process after the RFID Suite alerted the user of risky claims. Then, Richard *et al.* [56] built and tested the anomaly detection model to flag outliers using publicly available 2013 and 2014 insurance data from the Center of Medicare and Medicaid Services. The testing performed in the anomaly detection model to detect potentially fraudulent activities by predicting the medicare provider's specialty according to the number of procedures performed. Suppose a physician is expected to work in a different medical profession; they likely have many one-of-a-kind patients or are engaging in potentially fraudulent activity. The strategies such as feature selection, a healthcare specialty, and grouping are utilized for fraud identification in healthcare insurance sector specialties.

Later, the authors in Cassimiro [57] proposesd few experiments to examine the class imbalance impact of HIC fraud detection systems. This experiment predicts the performance of supervised techniques such as repeated incremental pruning to produce error reduction (RIPPER), SVM, random forest, naïve Bayes due to the imbalanced class. They also used the recovery method synthetic minority over-sampling technique (SMOTE) meta Cost and random oversampling to reduce the HI claims fraud. Later, Herland *et al.* [23] presented various methods for generating one dataset from three medicare CMS (HI claim) datasets after that data processing is performed and a map provider fraud label from LEIE dataset for identifying fraudulent behavior of physician. For medical insurance fraud detection, supervised learning techniques like Random Forest, gradient boosted trees, and logistic regression was utilized for training three CMS medicare dataset and combined dataset.

Then, Pandey *et al.* [58] developed a framework for identifying forged HIC. A scoring model was used to create a fraud indicator to recognize forged claims for this goal. This fraud indicator classified the forged and non-forged claims. Linear regression is used to validate this scoring model. Sowah *et al.* [59] developed a decision-making method for detecting the HI fraud. SVM based on genetic algorithms is utilized for classifying the fraudulent insurance claims from the National HI Scheme dataset. Later, Ilango *et al.* [60] proposed a framework that detects the HI fraud with faster learning from CMS medicare dataset using multi-Layer perceptron, a feed-forward neural network with genetic algorithm and also solves classification imbalance problem. Then, Yang *et al.* [61] proposed a novel light gradient boosting machine (LightGBM) mining algorithm for HI fraud detection. This algorithm eliminates the negative influence of the imbalance of positive and negative samples in the training set by automatically selecting the hard examples and balancing the ratio between fraud samples and nonfraud samples.

4.2 Unsupervised Learning

In an unsupervised learning method, the model was trained using data without any associated label. Unsupervised learning can detect new and existing types of HI frauds as they aren't limited to fraud patterns with predefined class labels[55].

Later, Verma *et al.* [62] proposed an approach for identifying the fraud frequent patterns from HI database using rule mining which analyzed HIC fraudulent pattern according to period and disease. Period-based claim fraud outliers are identified using decision rules based on statistics and k-means clustering. In contrast, disease-based fraud outliers are identified using association rule mining and gaussian distribution. These outliers represent fraud insurance claims in the database. Then, Anbarasi *et al.* [63] proposesd a framework that predicts the suspicious behavior of the HI fraud using probabilistic outlier detection. To reduce the time required for the fact-finding process, proactive and retrospective analysis are combined. Later, the authors in [64] proposed a self-organizing feature map neural network-based HI fraud detection model with principal component analysis (PCA), which considers all the valuable features from the patient's claim data and improves the accuracy of the classification. PCA is utilized for dimension reduction of data. In this study, few influence variables are selected to detect HI fraud.

4.3 Hybrid Approach

Some studies in the reviewed literature suggest a hybrid strategy of supervised and unsupervised learning methods for detecting the HI fraud. This hybrid approach of ML technology provides efficient results in healthcare insurance fraud identification.

For example, Rawate *et al.* [55] developed a supervised and unsupervised learning hybrid strategy for detecting the fraud in the HI sector. This approach utilized the advantage of both classification and clustering techniques. Then, Kareem *et al* [65]. presented a method for recognizing the forged HIC by analyzing the correlations or associations between attributes on the HIC documents. They utilized a support vector machine (SVM) supervised learning technique. This approach uses an unsupervised learning technique evolving clustering and association rule mining due to dynamic HI document data. Later, Jiang *et al.* [66] proposed a parallel framework to deal with the class imbalance and heterogeneous data of HI detection systems. This framework is cost-sensitive to deal with class imbalance. Deep learning algorithms such as deep neural network (DNN), bidirectional long short-term memory (BiLSTM), parallel neural network (PNN), and simulated neural network (SNN) are used to compare this framework's performance. Later, Zhou *et al.* [67] proposed a hybrid strategy for recognized forged repayment of the healthcare insurance sector. Local outlier factor and clustering are combined in this hybrid approach. This approach provides better results for newly arrived insurance claim data.

4.4 Blockchain Based HIC Fraud prevention

Liu *et al.* [68] proposed a blockchain-based healthcare insurance system , i.e., a cloud-based architecture. This system gives medical insurance an anti-fraud service, which determines whether a patient's medical reimbursement request is genuine and fits the policy's requirements. Later, Gera *et al.* [69] developed a framework and consensus mechanism of three peers, which are police, insurance firm staff, and agents for solving the security issues in the HI transactions. Every transaction can be stored as a chain of blocks and cryptographically signed in the IBM blockchain platform to prevent HIC transactions from the fraud. Later, Saldamli *et al.* [70] provides a blockchain-based solution for HI fraud commited by the patient or any malicious entity. In this solution, transactions associated with the patient's health claim are tracked using blockchain. If a transaction of the claim resembles with a past claim transaction, then the patient's insurance application request need to be declined in the network. Then, Ismail *et al.* [71] proposed a taxonomy and HI fraud detection framework using blockchain and examine results according to the time taken by execution and data transmitted when the number of HIC is increased. In this framework, 12 different fraud scenarios are detected, and detection was performed from patients, doctors, pharmaceutical firms, and doctors. Later, Baker *et al.* [72] developed a consortium blockchain-based distributed architecture and consensus mechanism to restrain duplication in the HIC system. They analyze the time required to validate the transactions (validation time), the time needed to upload transactions into the transaction pool (upload time), the time required for inserting blocks to the chain, security of data, and privacy of data.

4.5 HIC Fraud detection from Group of people

As we discussed in section II B Types of HIC fraud, a group of people are involved with HI fraud. Some researchers are analyzing HI fraud detection methods from a group of people. For example, Wanga *et al.* [73] presented an approach for detecting dentist fraudulent claims based on the trustworthiness score of the dentist, which is evaluated from the social network of patient and dentist. In this approach, the suspicious link between two or many dentists is analyzed based on the social network. Later, Figueredo *et al.* [74] developed a HI fraud detection model which detects the fraud from insurance subscriber's claim data. Fraud can identify based on mutual referrals among groups of physicians. Social networking techniques are used to find mutual referrals among groups of physicians. Then, Sun *et al.* [75] developed a method to detect joint fraud based on fraudulent group mining in the HI sector. In this method, mining is evaluated using a similarity adjacency graph and classifies abnormal groups according to the similarity adjacency graph.

4.6 AI and Blockchain-based HIC fraud detection

Integration of blockchain and ML for HIC fraud detection solved many issues of HIC, such as security, privacy, and data interoperability. Blockchain provides secure HIC data storage and tamperproof HIC data transfer in which insurance subscribers consent to share their records with the parties involved with HIC. ML can use this HIC data to identify a fraudulent pattern and generate correct predictions about fraudulent claim. Two studies found in the literature show the integration of ML and blockchain for HIC fraud detection.

Dhieb *et al.* [76] proposed a framework for HIC fraud detection and risk evaluation in which ML techniques are performed on data stored into the blockchain. For better performance results of the fraud detection analysis, data cleaning is performed. They also performed a comparative analysis of fraud indicators using different ML techniques such as SVM, XGBoost, fast decision tree, Stochastic Gradient Descent, Naïve Bayes, Nearest Neighbour. Blockchain network is created using hyper ledger fabric, and for integration between blockchain and ML, they used Rest API. Zhang *et al.* [77] developed a framework for identifying HIC fraud based on bidirectional encoder representations from transformers (BERT-LE) deep learning technique using two real hospital datasets. BeRT-LE technique was used to classify the reasonability of the international classification of diseases-10 (ICD-10) code, which is attached with the E-health report. The reasonability of the ICD code is evaluated from the inpatient's description about their sickness, which they called a chief complaint. They also proposed a health record storage and management method based on the consortium blockchain for data security, reliability, immutability, traceability, and non-repudiation.

Table 4.1: Comparative analysis of existing Unsupervised Learning based HIC fraud detection system

Author	Year	Objective	Pros	Cons	ML Method
Verma et al.[62]	2017	To identify HI fraud patterns periodically and disease wise	It contains a common strategy to generate a prediction of identifying medicare insurance fraud in many healthcare specialties, which is applied flexibly at scale.	This study has no effective strategies for fraud prevention and does not identify newly emerging fraud	K-means clustering, Association rule mining, elbow test, Outlier detection
Anbarasi et al.[63]	2017	To identify abnormal fraud patterns in the HIC system using outlier detection and proactive, reactive analytics integration.	The amount of time taken by the process which finds facts is reduced.	Not given information about a tool for identifying nature of HIC fraud occurrences	Outlier detection, Analytic hierarchical processing's pairwise comparison technique
Cao *et al.* [64]	2019	To extract features from the medical claim database which improve Accuracy of HI fraud classification model using self-organized feature map neural network.	Overall time complexity lowered, The cost of analysis decreased, The overall accuracy of classification has enhanced	The fraud indicator of healthcare insurance data was not precise, and it impacted the model's detection ability.	Self-organizing feature map, PCA

Table 4.2: Comparative analysis of existing Supervised Learning based HIC fraud detection system

Author	Year	Objective	Pros	Cons	ML Method
Kose et al.[28]	2015	To develop a framework for detecting HI abusive claims from the actor and commodities. Actor refers to insurance subscribers, physicians, and commodity refers to healthcare services.	Even with the incomplete data, predicted the risky claim	The network between actors is not considered in the analysis process	Proactive and retrospective analysis, Zero R classification
Richard et al. [78]	2016	To develop a predictive model which detects fraudulent behavior in HIC based on classification of physician specialty	Effective prediction	Similar procedure perform by various type of physicians is not considering	Multinomial naive bayes, Logistic regression
Cassimiro et al.[57]	2017	To analyze the loss of HIC fraud prediction performance in brazillian healthcare due to class imbalance and present recovered method	Give comparison of various ML-based recovery methods for class imbalance issue	Availability of data set is limited, The small training data set	RIPPER, SVM, Random forest, Naive bayes
Herland et al. [23]	2018	To generate one dataset from three Medicare CMS datasets, perform data processing on it and perform mapping of provider fraud label for identifying fraudulent behavior of physician in HIC	Performance of HIC fraud detection from the combined generated dataset is best.	Class imbalance problem is not addressed during the exploratory analysis of medicare fraud	LR, Gradient boosted trees, Random forest
Pandey et al.[58]	2018	To build a scoring model which classifies fraud indicators of HIC and determines the best ML technique for HIC fraud detection	More accurate Result of fraud prediction and scoring for any new HIC data.	The model is unable to handle dynamic data.	LR, Decision tree, Neural networks
Sowah et al. [59]	2019	To develop a model for HI fraud detection using genetic SVM.	Time taken for process claims is reduced	More computation time	Genetic SVM, Naive bayes
Ilango et al. [60]	2020	To develop a framework for discovering the more number of fraud patterns in HIC using by supervised learning techniques	More accurate and time efficient	Not given a detailed analysis of model training time	Multi-layer perceptron with a genetic algorithm,LR
Yang et al.[61]	2021	To develop bootstrapping HIC fraud detection model by choosing fraud samples and removing a legitimate sample for balance between fraud sample and legitimate sample	The model takes less time for training.	Small dataset is taken for model training.	LHEM

Table 4.3: Comparative analysis of existing Hybrid approach(supervised and unsupervised learning) based HIC fraud detection system

Author	Year	Objective	Pros	Cons	ML Method
Rawte et al.[55]	2015	To develop hybrid model using SVM supervised learning (SVM) and unsupervised learning (ECM) for identify duplicate HIC	Scalability	Accuracy information is missing	Evolving Clustering Method (ECM), SVM
Kareem et al.[65]	2017	To identifying correlation or link between features of HIC record to detect the forged HI claims	Scalability	Only the pre-processing forged HIC detection was addressed	ECM, SVM
Jiang et al. [66]	2018	To develop a parallel framework for fraud detection in HIC operation from real-world sequence and descriptive data by comparative analysis of various deep learning approaches	Processed heterogeneous data, Classify imbalance data	-	DNN, BiLSTM, PNN,
Zhou et al. [67]	2020	To propose a method for detecting fraudulent payment in HIC reimbursement the process from real HIC data using local outlier factor and clustering	Scalability	Fraud detection analyzed only from two diseases wise	Outlier detection, clustering

Table 4.4: Comparative analysis of existing Blockchain based HIC fraud prevention system

Author	Year	Objective	Pros	Cons	Protocol
Liu et al.[68]	2019	To develop blockchain-based HIC system to prevent forged data, third-party accidental fraud risk	Provide better data privacy preservation, adaptability	Not given implementation result	GRPC, Gossip protocol
Gera et al.[69]	2020	To developed framework for preventing HI fraud transaction using IBM platform of blockchain and developed consensus involved three peer police, agent and HI firm staff	Security, Non-repudiation, Integrity	Scalability	Proof of Work (POW)
Saldamli et al. [70]	2020	To track all patient's HIC transactions and design a HIC record secure sharing solution using blockchain, which helps to detect fraud in HIC	Easy HIC fraud detection, Efficiency of middleware to scan and communication of HIC record is good, time taken for data processing in blockchain was less	Not stable version of database	-
Ismail et al. [71]	2021	To develop blockchain-based taxonomy of HI fraud detection according to 12 different fraud scenarios.	Maintain performance when healthcare insurance branches and claims transactions grow.	Interoperability issue	NS 3
Baker et al. [72]	2021	To develop blockchain-based distributed architecture and consensus mechanism to prohibit duplication in medicare insurance claims.	Frameworks are run very fast and increase the security and privacy of data	Scalability	First in first out, Proof of work, Proof of stack.

Table 4.5: Comparative analysis of existing HIC Fraud detection system based on group of people

Author	Year	Objective	Pros	Cons	ML method
Wanga et al. [73]	2016	To detect fraudulent dentist claims based on the trustworthiness score of the dentist, which is evaluated from the social network of patient and dentist.	HIC application evaluation is quick, Reduce Claim reviewer's exhaustive workload and improve their review accuracy,	Some fraudulent dentists cannot be identified because their patients' loyalty prohibits them from building social relationships between them.	Zero R classification
Figueredo et al.[74]	2018	To disclose common reference between physicians and identify fraud pattern of physician's practice from HIC data.	Accuracy of the model is excellent.	Clinical test or surgeries was not analyzed in the fraud detection model.	Page rank algorithm, Social network analysis
Sun et al.[75]	2019	TO develop a method for detecting joint fraud in HIC based on fraudulent group mining. This mining is evaluated and classifies abnormal Groups according to a similarity adjacency graph.	More precise, Reduce calculation	Joint fraud detection performed on a limited number of people's record	L-SVM, Abnormal grouping, Density-based incremental local outlier factor

Table 4.6: Comparative analysis of existing AI and Blockchain based HIC fraud detection system

Author	Year	Objective	Pros	Cons	ML method
Dhieb et al. [76]	2020	To develop a framework for HIC fraud detection and risk evaluation using blockchain and machine learning	Easily integrate module into the system, scalable	Authorized party check fraud claim generated from ML module manually	XGBoost, Naïve bayes, Nearest neighbour, SVM.
Zhang et al. [77]	2021	Proposed a framework for detection of HIC fraud using text classification, and HIC data are stored on consortium blockchain.	Reduce HI auditor workload, Secure, traceable, and non-repudiation data storage	Result of prediction was biased, Not consider outpatient scenario during training	DPCNN, Text-RNN, HAN, FastTex, LEAM, BERT, BERT-LE

Chapter 5

The Proposed System

In this chapter, we propose a layered intelligent HI fraud detection system and case study on HI Fraud Detection. The number of layers are four which are named as user layer, data generation layer, data analytical layer, and blockchain layer. Figure 5.1 shows the proposed four-layer architecture. The detailed description of each layer is described in subsequent subsections.

5.1 User layer

There are three primary users involved in the proposed HIC fraud detection architecture, such as insurance provider, healthcare service provider, and insurance subscriber. Insurance providers include private and government insurance firms. Healthcare service providers include hospitals, pharmacies, medical labs, ambulance services, and doctors. The insurance subscriber is a patient, who request an HI claim. This layer is connected to a blockchain network in which each activity of all the users are recorded. Apart from the blockchain, this layer is also connected to the data generation layer.

5.2 Data generation layer

There are two sources for the HIC data collection, which are offline and online processes. Offline data sources include paper-based billing forms, drug prescriptions, Paper based HIC, medical guideline books. Offline data sources such as billing forms, drug prescriptions, and medical guideline books are collected from hospitals (healthcare service provider), ambulance service providers, medical laboratories, pharmacies. Paper based HIC is collected from the HI government sector, or private firms (HI provider). Offline collected HIC data are digitized through internet. Online data sources include publicly available research-based datasets, EHR, HIC data in an electronic form, and social media data which are available through internet. EHR are collected from healthcare service provider and social media data are collected from insurance subscriber (patient). After data collection, the data is generated according to the HIPPA act. Data generated from this layer will share and control through smart contract condition for security purposes because this data contains payment-related sensitive information of insurance subscribers and providers.

5.3 Data analytical layer

HIC data have sparsity, heteroscedasticity, multicollinearity, and missing values. The best way to deal with such data is to use data preprocessing and apply robust ML approaches. In the data analytics layer, data preprocessing is involved with data cleaning, data integration, data transformation, and data reduction. In the data cleaning process, all redundant, incomplete, and unnecessary HIC data are eliminated. HIC data generated from all sources are combined in the data integration process. In the data transformation process, the structure of the HIC data is modified through various techniques such as smoothing, aggregation, and normalization. The data reduction process decreases the HIC data storage, making HIC data analysis easier while producing similar results. After data preprocessing, data is ready for data analytics, which can be performed using machine learning algorithms according to fraud indicators. Various ML algorithms are used to determine if a claim is legitimate or fraudulent. This prediction result is integrated with smart contract through blockchain layer.

5.4 Blockchain layer

The blockchain layer provides a secure environment for all payment and authentication-related transactions. Each layer of architecture can communicate with the blockchain layer through a smart contract. Smart contracts are self-executing predefined rules for authentication and insurance coverage payment. In a smart

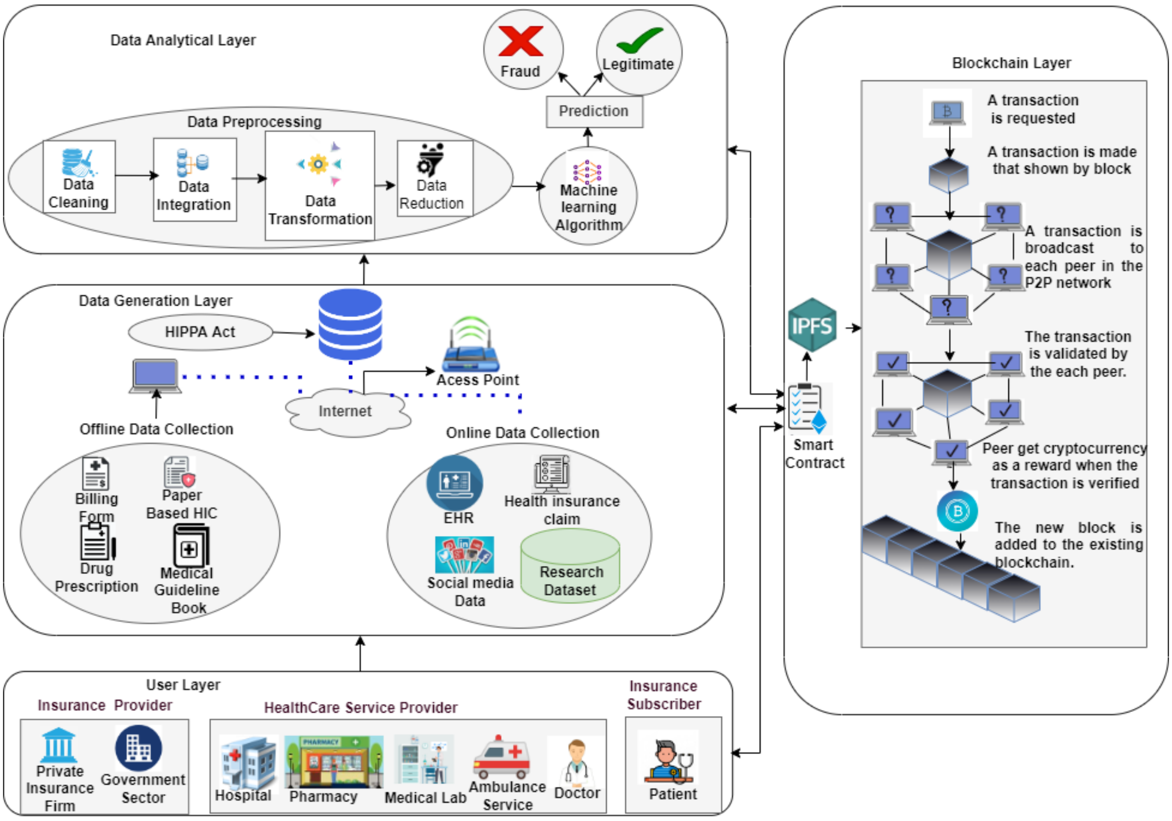


Figure 5.1: Architecture of HI Fraud Detection [1],[2]

contract, consensus algorithms like PoW, proof of burn (PoB), or proof of stake (PoS) are utilized to authenticate every transaction and input. A smart contract is integrated with InterPlanetary File System (IPFS) because it minimizes the overall data storage and cost for fast data retrieval. Blockchain is a decentralized network of peers that communicate using a smart contract that eliminates the need for a third party. In Figure 5.1, blockchain layer shows how all HIC transactions are performed. Cryptography algorithms and the immutable qualities of the blockchain enhance the security of the HIC data. When HIC data is copied into the blockchain, it is impossible to change it.

5.5 Case study

Nowadays, people have become more conscious of their healthy lifestyle choices due to the Covid-19 pandemic. They've begun to live healthier lifestyles using advanced technology such as wearable technology. Wearable devices and smartphone healthcare apps have become more popular due to this. The wearable device is a smart electronic device that can be worn as jewelry, embedded in clothing or body [79]. Smartphone healthcare apps provide healthcare-related services. Data can be captured from this healthcare app and wearable device, used for labs and clinical research.

Some Insurance provider firms use wearable devices, and healthcare apps data containing blood sugar tracker data, fitness tracker data, pedometer data, and telehealth data [80]. They use wearable devices as part of their marketing scheme to connect with insurance subscribers. Under certain conditions, insurance provider firms offer rewards or discounts on the premium price and free wearables as part of their marketing scheme. Conditions are insurance subscribers use a wearable, give consent about wearable tracker data sharing, and show healthy behaviors such as walking a certain number of steps each week, count of calories, maintaining a healthy heart rate through the wearable [79].

Humana Vitality insurance firm launched a platform in 2016 in the United state for motivating and rewarding the insurance subscriber to participate in healthy activities through wearable devices. According to the firm tracked data from wearable devices, Humana provides challenges and competitions to insurance subscribers. Insurance subscribers are rewarded with fitness equipment, gift cards, and reduced HI premium rates [81]. Oscar insurance firm of US partnered with the Misfit to distribute Misfit fitness trackers to all the insurance subscribers free. This firm is one of the first insurance provider firms who evaluate the insurance premium rate of insurance subscribers through wearables data. They are linked insurance subscribers' biometric data to with their HI account [81]. AXA healthcare insurance firm of France encourage insurance subscribers to share their health-related personal data from a fitness device. Insurance subscribers of AXA can earn fit points by sharing their fitness device data. AXA uses this data for its insurance underwriting process [81]. Medibank, Australia's leading insurance provider firm, formed a partnership with flybys to integrate insurance subscriber's data from various wearable fitness trackers to their medibank accounts. This data is used for insurance risk assessment, and pricing of premiums [81].

As we discussed above, wearable device data is used by the insurance firm only to determine the pricing of premiums and insurance risk assessment. They are not used wearable device data for fraud detection in HIC. The HI fraudsters become more smart day by day and find new ways for fraud, so there is a need for the insurance firm to find a new, and improved futuristic way to combat fraud using wearable device data. Hence we present a novel case study to fulfill the above requirement of the insurance firm. The Figure 5.2 shows the flow of the case study. This case study also solves security and privacy issues in HI.

In the figure 5.2, there are various wearable devices worn by insurance subscribers such as smart glasses, blood pressure sensors, fitness bands, smart shoes, etc. The insurance firm may provide wearable for marketing and fraud detection in HIC. Wearable wear by insurance subscribers is connected with insurance subscribers' smartphones through wifi or Bluetooth. The data are generated from wearable are stored in IPFS, which are in the encryption form and digitally signed using blockchain. Insurance providers and healthcare service provider data are encrypted, digitally signed, and uploaded in the fraud detection system. After data generation, data preprocessing is performed that includes data cleaning, integration, transformation, and reduction method. Before data preprocessing, insurance subscribers distribute consent about sharing personal healthcare data with a predefined ethereum smart contract. The insurance provider firm provides the public key to insurance subscribers and healthcare providers. The insurance firm will reward cryptocurrency to insurance subscribers if insurance subscribers give consent about sharing personal healthcare data. The ML algorithm is applied to preprocessed data that predict that HIC is fraud or legitimate. If a claim is legitimate, blockchain network consensus enables automated claim coverage payment to insurance subscribers. But if the claim is fraudulent, the subscriber would pay the penalty.

5.5.1 Real time analytics

Earlier, the HIC management system relies on insurance subscriber reporting. Analysis and auditing of the HIC are done after the hospitalization emergency occurs. At the same time, HIC management uses wearable analysis real-time events such as cardiovascular events. It also increases insurance subscribers' engagement with the insurance provider in real-time and frequently. The location of the insurance subscribers will be traced using the wearable device. Location data and the real-time

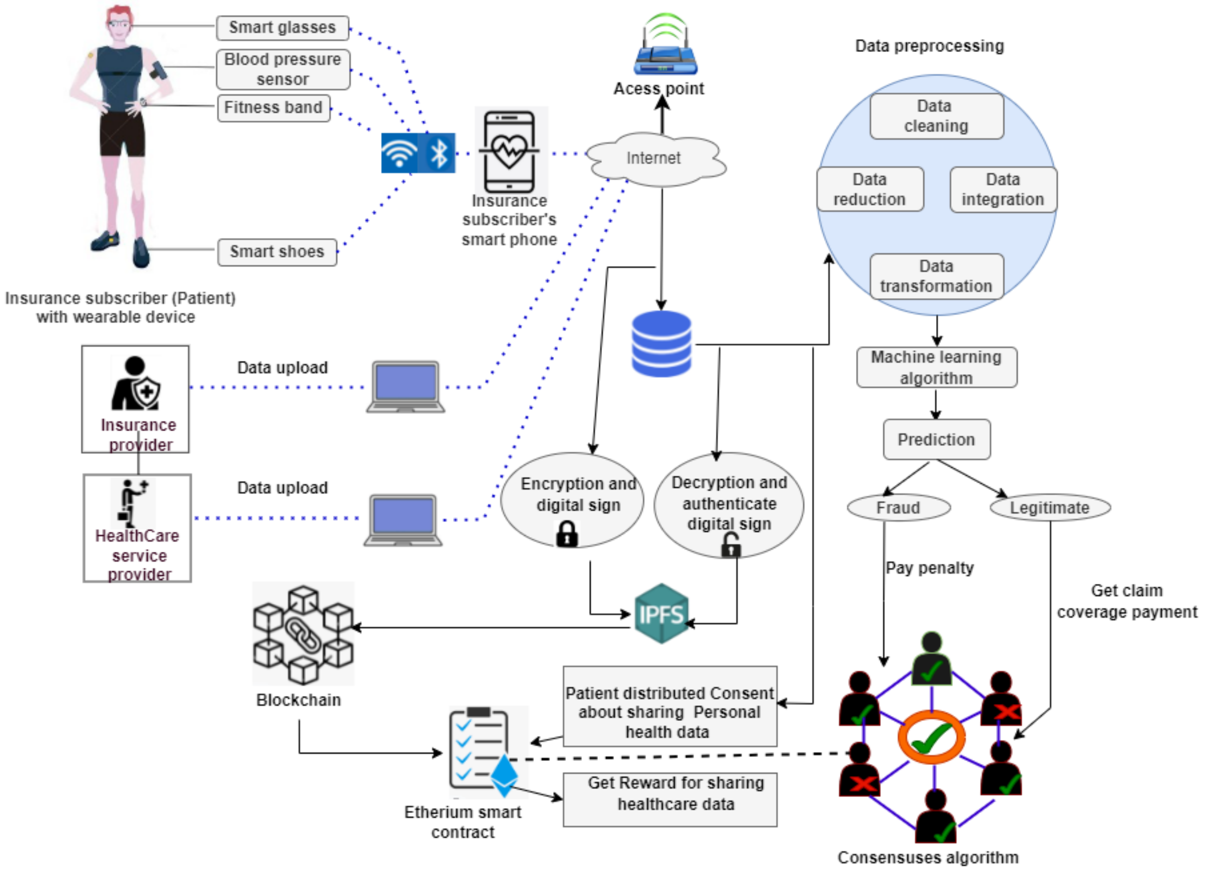

Smart glasses
Blood pressure sensor
Fitness band
Smart shoes
Insurance subscriber (Patient) with wearable device
Insurance subscriber's smart phone
Acess point
Internet
Data preprocessing
Data cleaning
Data reduction
Data integration
Data transformation
Machine learning algorithm
Prediction
Fraud
Legitimate
Pay penalty
Get claim coverage payment
Insurance provider
Data upload
HealthCare service provider
Data upload
Encryption and digital sign
Decryption and authenticate digital sign
IPFS
Blockchain
Patient distributed Consent about sharing Personal health data
Get Reward for sharing healthcare data
Etherium smart contract
Consensuses algorithm

Figure 5.2: Case study : HI Fraud Detection from wearable device

health status of insurance subscribers can observe what happened before, during, and after a hospitalization emergency. Insurance firms can use this data to detect pain levels, counteract medication-seeking behavior, and discover inconsistent behavior with a HIC. Earlier, real-time data is not available to the HIC fraud detection system.

5.5.2 Security and Privacy

Data use in the HIC is sensitive because it may cause the breach of the payment related and insurance subscriber's personal health-related sensitive information. Insurance subscriber's data can be exchanged among all parties involved with HIC fraud detection without revealing the insurance subscriber's identities because the identities of insurance subscriber's in a blockchain are pseudonymized using cryptographic keys. The proposed system use encryption, decryption, and digital sign to secure the HIC system from various attack and privacy breach.

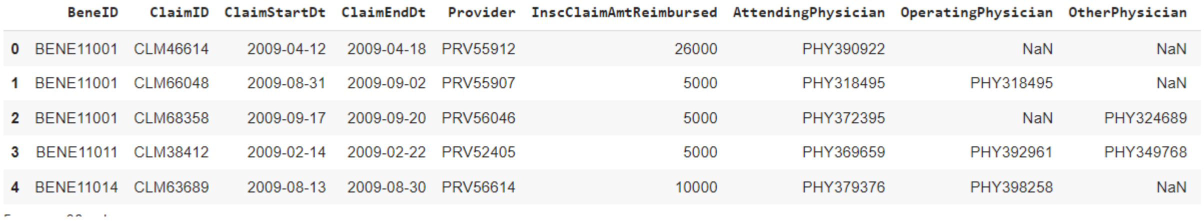
Chapter 6

Experimental Evaluation

In this chapter, we have discussed the selection of the dataset, experimental methodology, and result for the proposed parking pricing approach along with the experimental setup.

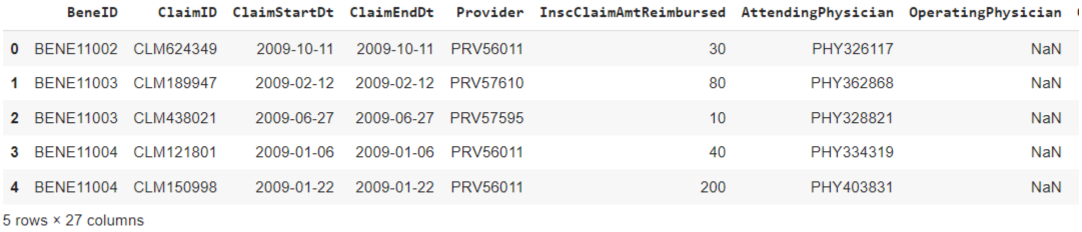
6.1 dataset description

In this project, the healthcare provider fraud detection analysis dataset was used, which can be downloaded from Kaggle [82]. We take 4 CSV data files from this dataset named inpatient,outpatient, beneficiary , and provider fraud. The inpatient data file having 30 feature and 40474 row which consists HIC information of patients who have been admitted to the hospital.Figure 6.1 describe inpatient feature and its value. The outpatient data file having 27 feature and 517737 row which consists HIC information of patients who have been not admitted to the hospital.The feature of outpatient dat file are shown in 6.2 The beneficiary data file having 25 feature shown in fig 6.3 and 138556 row which consists KYC information of patients. The provider fraud file having 2 feature named potential fraud, and provider and 5410 row. The feature of provider fraud file are shown in fig 6.4a



	BeneID	ClaimID	ClaimStartDt	ClaimEndDt	Provider	InscClaimAmtReimbursed	AttendingPhysician	OperatingPhysician	OtherPhysician
0	BENE11001	CLM46614	2009-04-12	2009-04-18	PRV55912	26000	PHY390922	NaN	NaN
1	BENE11001	CLM66048	2009-08-31	2009-09-02	PRV55907	5000	PHY318495	PHY318495	NaN
2	BENE11001	CLM68358	2009-09-17	2009-09-20	PRV56046	5000	PHY372395	NaN	PHY324689
3	BENE11011	CLM38412	2009-02-14	2009-02-22	PRV52405	5000	PHY369659	PHY392961	PHY349768
4	BENE11014	CLM63689	2009-08-13	2009-08-30	PRV56614	10000	PHY379376	PHY398258	NaN

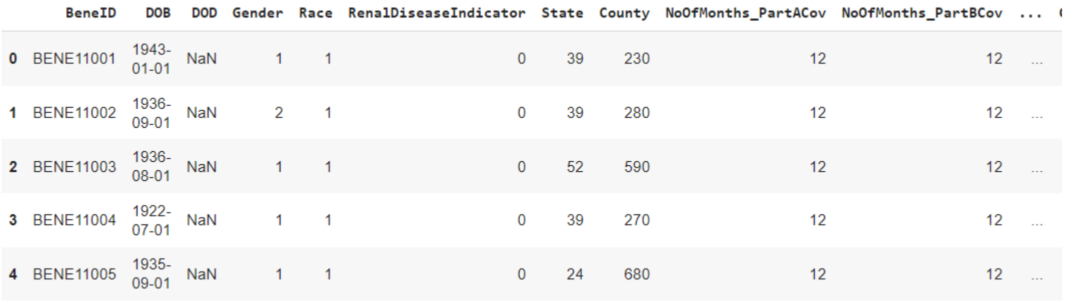
Figure 6.1: Inpatient data



	BeneID	ClaimID	ClaimStartDt	ClaimEndDt	Provider	InscClaimAmtReimbursed	AttendingPhysician	OperatingPhysician
0	BENE11002	CLM624349	2009-10-11	2009-10-11	PRV56011	30	PHY326117	NaN
1	BENE11003	CLM189947	2009-02-12	2009-02-12	PRV57610	80	PHY362868	NaN
2	BENE11003	CLM438021	2009-06-27	2009-06-27	PRV57595	10	PHY328821	NaN
3	BENE11004	CLM121801	2009-01-06	2009-01-06	PRV56011	40	PHY334319	NaN
4	BENE11004	CLM150998	2009-01-22	2009-01-22	PRV56011	200	PHY403831	NaN

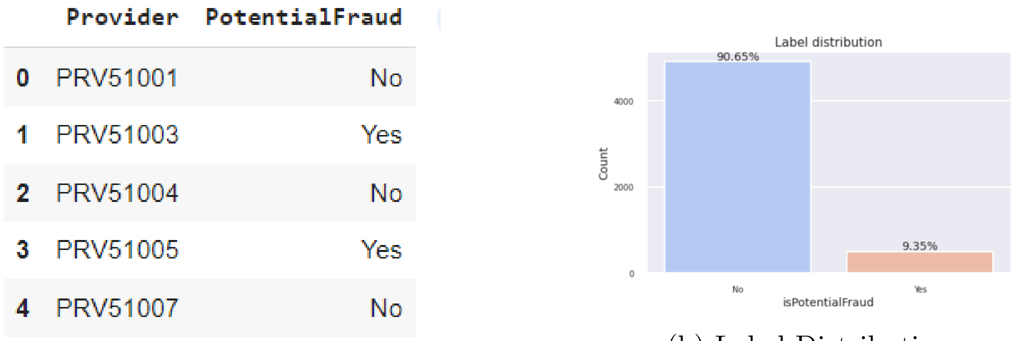
5 rows × 27 columns

Figure 6.2: Outpatient data



	BeneID	DOB	DOD	Gender	Race	RenalDiseaseIndicator	State	County	NoOfMonths_PartACov	NoOfMonths_PartBCov	...
0	BENE11001	1943-01-01	NaN	1	1	0	39	230	12	12	...
1	BENE11002	1936-09-01	NaN	2	1	0	39	280	12	12	...
2	BENE11003	1936-08-01	NaN	1	1	0	52	590	12	12	...
3	BENE11004	1922-07-01	NaN	1	1	0	39	270	12	12	...
4	BENE11005	1935-09-01	NaN	1	1	0	24	680	12	12	...

Figure 6.3: beneficiary data



	Provider	PotentialFraud
0	PRV51001	No
1	PRV51003	Yes
2	PRV51004	No
3	PRV51005	Yes
4	PRV51007	No

(b) Label Distribution

(a) PotentialFraud data

Figure 6.4: Provider Potential fraud

6.2 Experimental Methodology

6.2.1 Data Pre-processing

In this phase, the dataset is pre-processed to obtained preferable feature for model implementation, which is discussed over here. Some column of dataset contain categorical value such as RenalDiseaseIndicator, ChronicCond and remaining column contain numerical value. We convert categorical value into numeric value so the model is able to interpret and extract useful data easily. We add some new features from the existing data such as Hospital_stay_period, claim_duration, additional_Claim_Days, age, and ChronicCond_risk_score. Hospital_stay_period is measured from DischargeDt and AdmissionDt. The value of claim_duration was calculated from ClaimStartDt and ClaimEndDt. The additional_Claim_Days feature value was calculated from Hospital_stay_period and claim_duration. If the claim_duration value is greater than Hospital_stay_period then additional_Claim_Days value becomes a subtraction of claim_duration and Hospital_stay_period otherwise its value is zero. The value of age is measured from the date of birth(DOB) and date of death(DOD). If DOD is not given then we calculate age from the current date and DOB. The value of ChronicCond_risk_score can be calculated by summation of all chronic conditions of benefier. In this experiment, we create a new dataset by joining inpatient data, outpatient data, Beneficiary data, and provider fraud data to get a single data frame. We found that AttendingPhysician, OperatingPhysician, OtherPhysician, diagnosis code, procedure codes column having some missing value so we fill these missing values with 0. We perform group operation by provider id and take the mean of an above discussed new feature, reimbursed amount, and deducted amount because the main objective of the experiment is to identify suspicious providers. The provider is linked with physicians, beneficiaries, diagnoses, and procedures. So we add another feature by performing group operations by provider id and taking the count of physicians, beneficiaries,diagnosis, and procedures. After that, we also perform grouping by provider id and take the sum for creating a feature corresponding to every provider.

Feature selection is a key strategy for identifying the most important feature in a dataset. We choose the most important feature and discarded less important ones for reducing overfitting and training time using a decision tree classifier and with the help

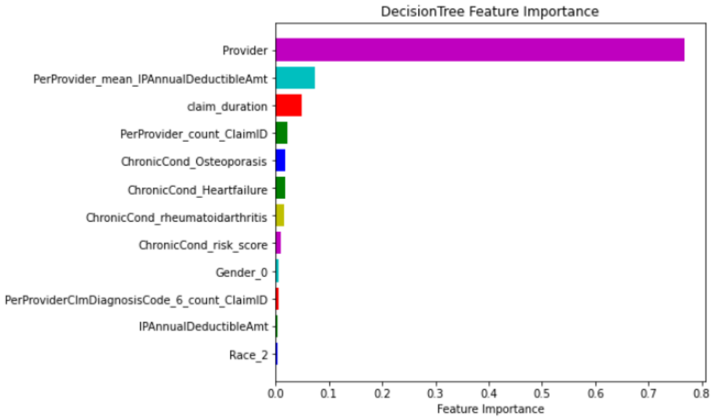


Figure 6.5: Feature Importance

of a visualization technique. The figure 6.5 shows the importance of features included in the ML decision tree classifier algorithm for the prediction of fraudulent healthcare providers.

When class labels are not equally distributed, ML technique face struggle to learn. Figure 6.4b shows potential fraud label distribution ratio is 90.65% : 9.35%. From the figure 6.4b, we can see that the given data is highly imbalance. Due to imbalance training data, ML model will over classify the 'no' label. As a result, sample corresponding to the 'yes' class are more frequently falsely labeled than those corresponding to the 'no' class . To solve this problem, we use SMOTE oversampling method, so that model can be efficiently trained. Before apply ML technique on data, scaling is performed to get every features to the similar range of magnitudes. Scaling the data becomes easier for a model to train and learn.The experiment system environment is Windows 11 operating system with configuration Intel(R) Core(TM) i5 .The software operating environment is google collaboratory notebook which provide scientific python development environment.

6.2.2 Machine Learning

Pre-processed data is fed to model and trained using ML technique where data is split into 80% for training and 20% for testing.The ML algorithms used in our experiment and result are described in the following subsection.

Logistic regression

Logistic regression is a well-known statistical approach for classification in ML. The logistic regression describes the predictive analysis that can be binomial, multinomial, or continuous. We evaluate if something might occur or not based on particular predictors using logistic regression. In our experiment, we apply logistic regression to predict potentially fraudulent HI providers.

Support Vector Classifier

Support vector machine is a linear classification approach that maps the feature representation in the vector form of an object in space. For new objects, the border can also give a good classification result. This classification approach plot a border to classify pattern into two types. So the support vector machine is well suited for our experiment to classify potentially fraudulent HI providers or not.

Random Forest

The random forest approach is utilized for classification and regression. It is applied for classification by building several decision trees. When there are more trees in the forest, this approach produces better results and prevents the model from overfitting. Since forests consist of more trees, this approach produces accurate results and prevents the model from overfitting. In the forest, every decision tree produces a result. To produce a more accurate and consistent prediction, the result of every decision tree are combined.

Bagging Classifier

Ensemble learning(EL) method is combination of more than one classifier to build a particular learning method for improving the prediction of single classifier. EL method is utilizing to improve the classifier's performance for prediction. There are several EL method like bagging, boosting, stacking are utilizing to improve performance for prediction. We used bagging EL method in our experiment for classification of potentially fraudulent HI providers.

Bagging classifiers were utilized to fit base classifiers on each randomly generated with replacement subset of the training dataset for improving the accuracy of prediction. Bagging classifiers is also referred to as smoothing process to enhance the accuracy of regression and classification trees. In the bagging EL method, several weak learners might join to build a strong learner. The weak learner is referred as individual decision tree.

When newly arrived sample is classified, every tree gives vote for a class prediction. The final class prediction of newly arrived sample is achieved through maximum vote of class or taking. In this project, we used bagging classifier with base estimators decision tree. We choose decision tree as a base estimator because our training dataset is highly imbalanced. The decision tree performs exceptionally well by weighting the output of the trees and minimizing the variance of the training dataset and the overfitting. Ensemble classifier of bagging are fast and train model with hundreds of feature easily.

Result

In this project, we update some parameters to enhance the model's performance. The table 6.1 shows the related parameters of the four models. The model's F1 Score, precision, recall, accuracy, specificity, and sensitivity are used to assess the performance of the models. All of the above metrics can be measured from a confusion matrix. Confusion matric compares real data value with ML model's prediction and describes classification performance. The following equations 6.1,6.2,6.3,6.4,6.5 are expressed to measure the value of the F1 Score, precision, recall, accuracy, specificity, and sensitivity.

$$Precision = \frac{TP}{TP + FP} \tag{6.1}$$

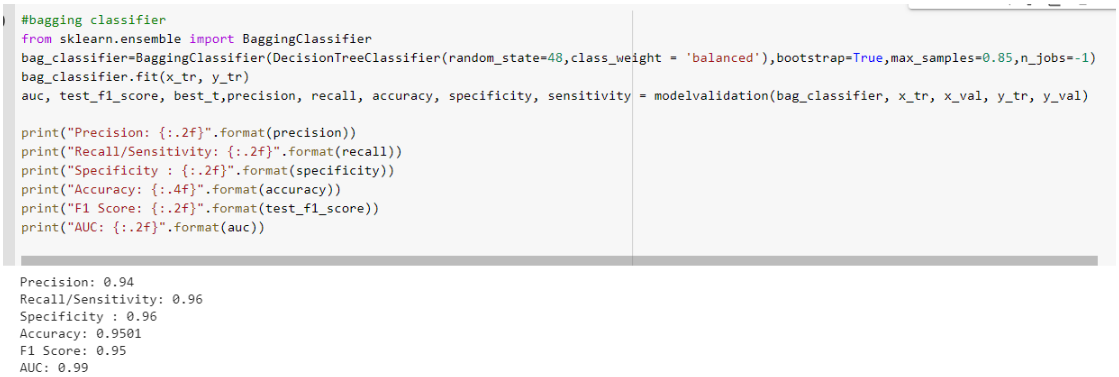
$$Recall/Sensitivity = \frac{TP}{FN + TP} \tag{6.2}$$

$$Specificity = \frac{TN}{FP + TN} \tag{6.3}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6.4}$$

$$F1score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{6.5}$$

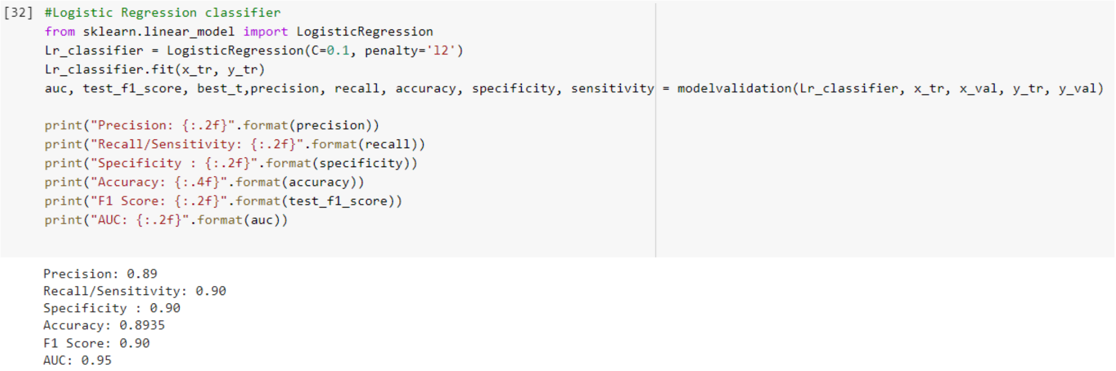
We also draw the ROC curve for comparative analysis of all four models which is shown in the figure 6.10a,6.10b,6.10c,6.10d. The ROC curve of the bagging classifier stays farthest and it has the largest area u nder the curve compared to other models which demonstrate the performance of the bagging classifier is best to classify the positive class in the dataset. From the figure 6.10d, we can also analyze that the random forest classifier is also good because its ROC score value is nearby the bagging classifier's ROC score value.



```
#bagging classifier
from sklearn.ensemble import BaggingClassifier
bag_classifier=BaggingClassifier(DecisionTreeClassifier(random_state=48,class_weight = 'balanced'),bootstrap=True,max_samples=0.85,n_jobs=-1)
bag_classifier.fit(x_tr, y_tr)
auc, test_f1_score, best_t,precision, recall, accuracy, specificity, sensitivity = modelvalidation(bag_classifier, x_tr, x_val, y_tr, y_val)

print("Precision: {:.2f}".format(precision))
print("Recall/Sensitivity: {:.2f}".format(recall))
print("Specificity : {:.2f}".format(specificity))
print("Accuracy: {:.4f}".format(accuracy))
print("F1 Score: {:.2f}".format(test_f1_score))
print("AUC: {:.2f}".format(auc))
```

```
Precision: 0.94
Recall/Sensitivity: 0.96
Specificity : 0.96
Accuracy: 0.9501
F1 Score: 0.95
AUC: 0.99
```

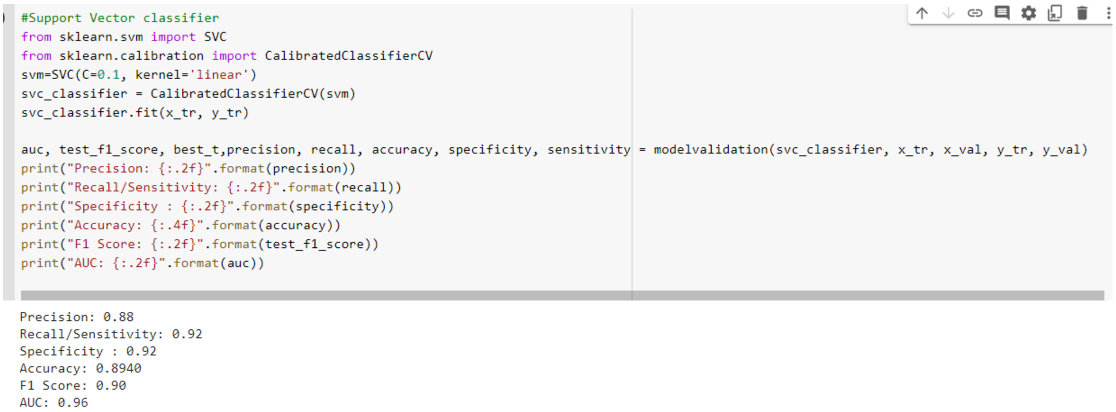
Figure 6.6: Bagging Classifier Result



```
[32] #Logistic Regression classifier
     from sklearn.linear_model import LogisticRegression
     Lr_classifier = LogisticRegression(C=0.1, penalty='l2')
     Lr_classifier.fit(x_tr, y_tr)
     auc, test_f1_score, best_t,precision, recall, accuracy, specificity, sensitivity = modelvalidation(Lr_classifier, x_tr, x_val, y_tr, y_val)

     print("Precision: {:.2f}".format(precision))
     print("Recall/Sensitivity: {:.2f}".format(recall))
     print("Specificity : {:.2f}".format(specificity))
     print("Accuracy: {:.4f}".format(accuracy))
     print("F1 Score: {:.2f}".format(test_f1_score))
     print("AUC: {:.2f}".format(auc))
```

```
Precision: 0.89
Recall/Sensitivity: 0.90
Specificity : 0.90
Accuracy: 0.8935
F1 Score: 0.90
AUC: 0.95
```

Figure 6.7: Logistic Regression Result



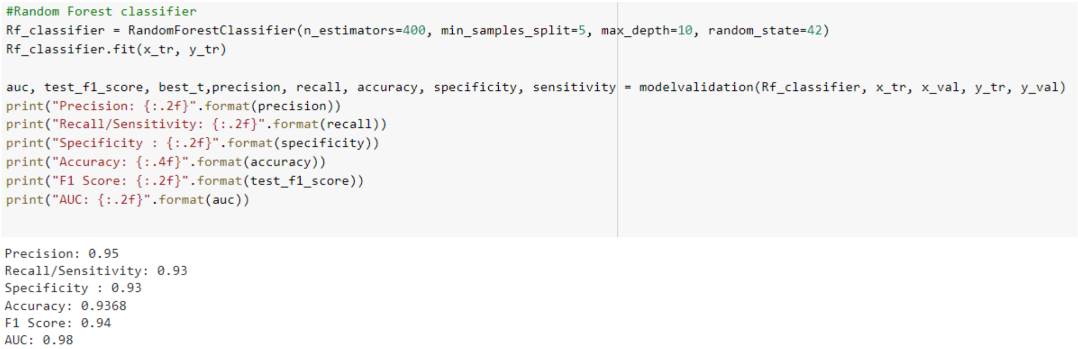
```
#Support Vector classifier
from sklearn.svm import SVC
from sklearn.calibration import CalibratedClassifierCV
svm=SVC(C=0.1, kernel='linear')
svc_classifier = CalibratedClassifierCV(svm)
svc_classifier.fit(x_tr, y_tr)

auc, test_f1_score, best_t,precision, recall, accuracy, specificity, sensitivity = modelvalidation(svc_classifier, x_tr, x_val, y_tr, y_val)
print("Precision: {:.2f}".format(precision))
print("Recall/Sensitivity: {:.2f}".format(recall))
print("Specificity : {:.2f}".format(specificity))
print("Accuracy: {:.4f}".format(accuracy))
print("F1 Score: {:.2f}".format(test_f1_score))
print("AUC: {:.2f}".format(auc))
```

```
Precision: 0.88
Recall/Sensitivity: 0.92
Specificity : 0.92
Accuracy: 0.8940
F1 Score: 0.90
AUC: 0.96
```

Figure 6.8: Support Vector Classifier Result

Table 6.1: Parameter of Model

Model	**Parameter**
Bagging Classifier	DecisionTreeClassifier(random_state,class_weight), bootstrap, max_samples, n_jobs
Logistic Regression	C, penalty
SVC	C, kernel
Randomforest	n_estimators, min_samples_split, max_depth, random_state



```
#Random Forest classifier
Rf_classifier = RandomForestClassifier(n_estimators=400, min_samples_split=5, max_depth=10, random_state=42)
Rf_classifier.fit(x_tr, y_tr)

auc, test_f1_score, best_t,precision, recall, accuracy, specificity, sensitivity = modelvalidation(Rf_classifier, x_tr, x_val, y_tr, y_val)
print("Precision: {:.2f}".format(precision))
print("Recall/Sensitivity: {:.2f}".format(recall))
print("Specificity : {:.2f}".format(specificity))
print("Accuracy: {:.4f}".format(accuracy))
print("F1 Score: {:.2f}".format(test_f1_score))
print("AUC: {:.2f}".format(auc))
```

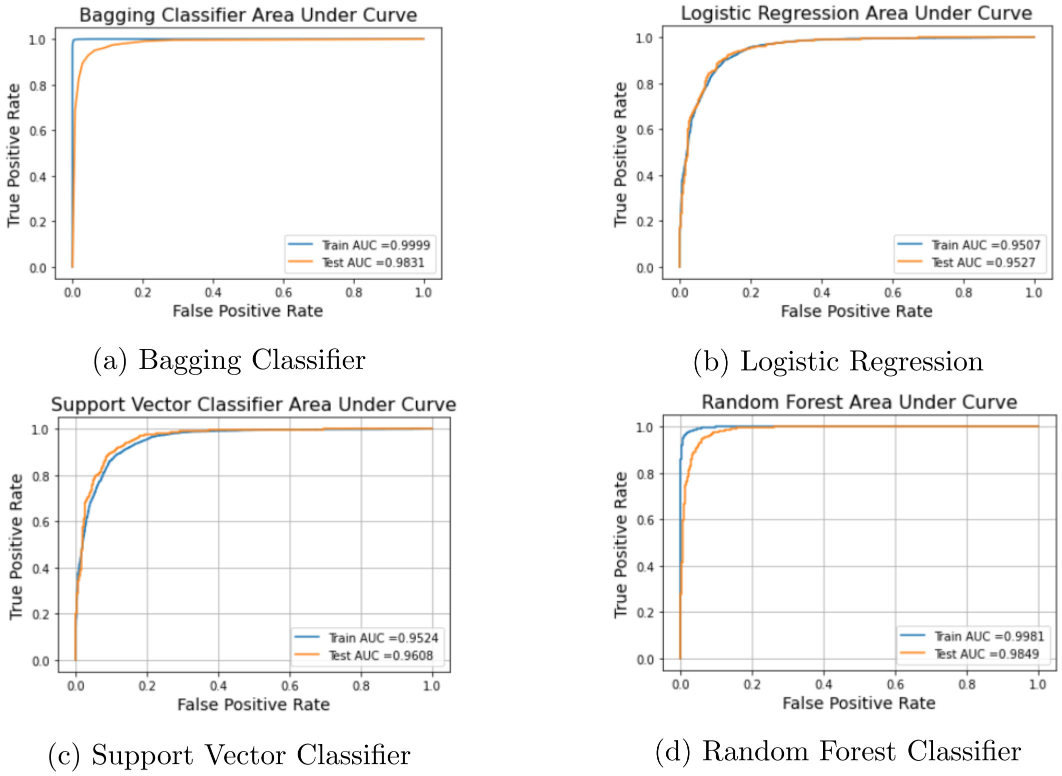
```
Precision: 0.95
Recall/Sensitivity: 0.93
Specificity : 0.93
Accuracy: 0.9368
F1 Score: 0.94
AUC: 0.98
```

Figure 6.9: Random Forest Classifier Result

Table 6.2: Performance of Model

Evaluation Measure	Bagging Classifier	Logistic regression	SVC	Randomforest
Precision	0.94	0.89	0.90	**0.95**
Recall/Sensitivity	**0.95**	0.90	0.90	0.93
Specificity	**0.95**	0.90	0.9	0.93
Accuracy	**0.9439**	0.8935	0.8986	0.9368
F1 Score	**0.95**	**0.90**	**0.9**	**0.94**
AUC	**0.98**	0.95	0.96	0.98

In this experiment,dataset have very few sample of fraud label. The training dataset is highly imbalance. The accuracy metrics are always misleading for imbalance training data so we cannot use accuracy metric for performance evaluation of model. Precision score indicates percentage of the non-fraudulent classes that are labelled as fraudulent, whereas recall score indicates the percentage of fraudulent classes that are labelled as non-fraudulent. We choose F1 score as performance metric because it is harmonic mean of precision and recall. The table 6.2 shows a comparison of four classification algorithms with the evaluation metrics. Compared with all the four algorithms, the bagging classifier produced better results in terms of recall, specificity, accuracy, F1 score, and AUC.



(a) Bagging Classifier

(b) Logistic Regression

(c) Support Vector Classifier

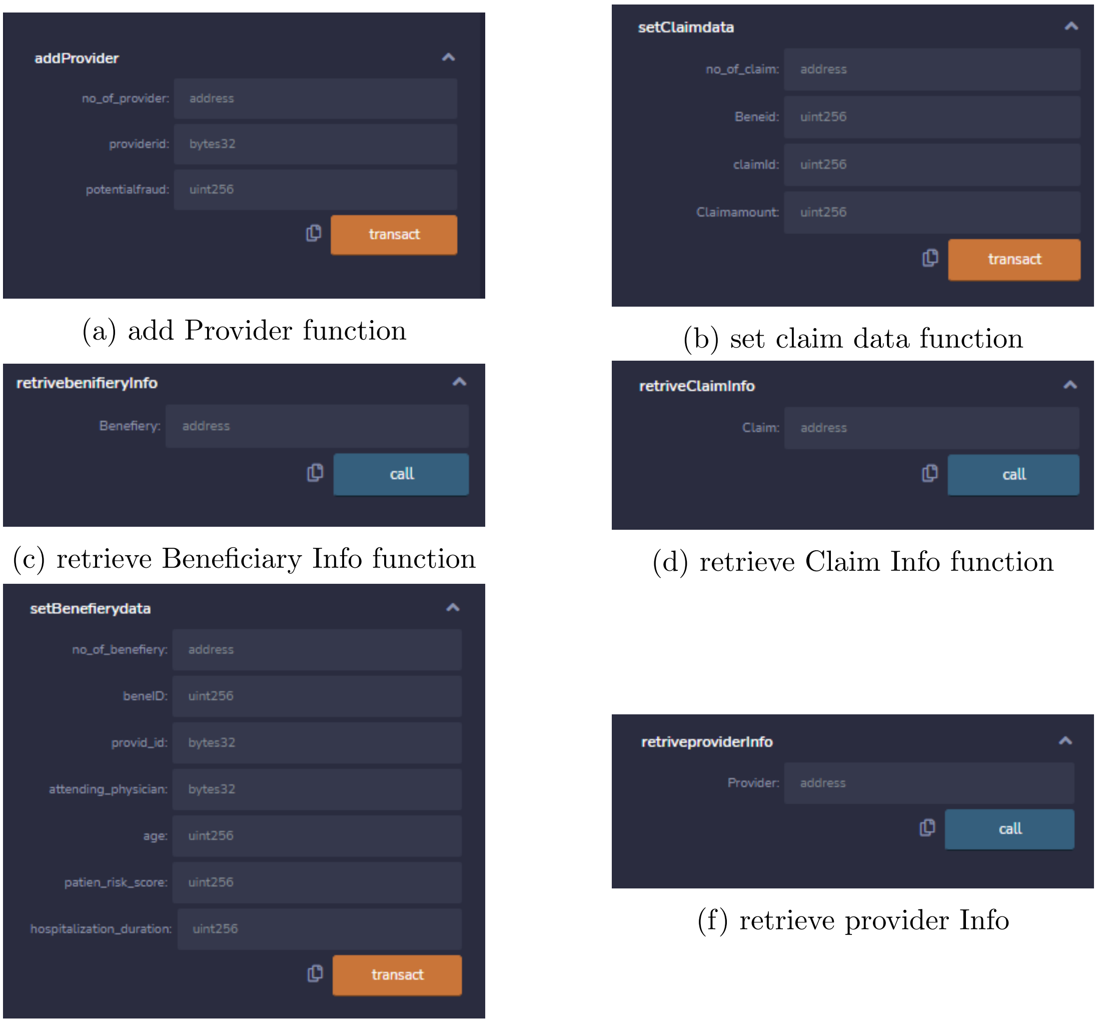
(d) Random Forest Classifier

Figure 6.10: Area under Curve for four model

6.3 Blockchain Implementation

Blockchain as a decentralized framework is beneficial to HIC applications which allow operations of distributed HIC applications without relying on a centralized server. Replication of blockchain ledger data across all the nodes generate an environment of transparency among all the parties involved in the HIC process. For decentralized blockchain platform, Ethereum is used. Ethereum creates peer-to-peer network and it also executes and verifies application code. This application code is known as smart contract and it written in solidity language. We can deploy the HIC application transaction without third party agent using smart contract. All HIC related transaction can be traced and can not be manipulated. In this implementation, we develop a smart contract for HI which contain some condition which execute automatically.

In this project, we used the Ethereum remix IDE to creating a Ethereum platform for smart contract development in solidity language. Remix IDE provide different types of user account which contain fake hundred ether. In this implementation, we used three account. First account used by healthcare service provider, second account used by HI beneficiary and third account used by HI provider firm.In smart contract we developed



(a) add Provider function

(b) set claim data function

(c) retrieve Beneficiary Info function

(d) retrieve Claim Info function

(e) set Beneficiary data function

(f) retrieve provider Info

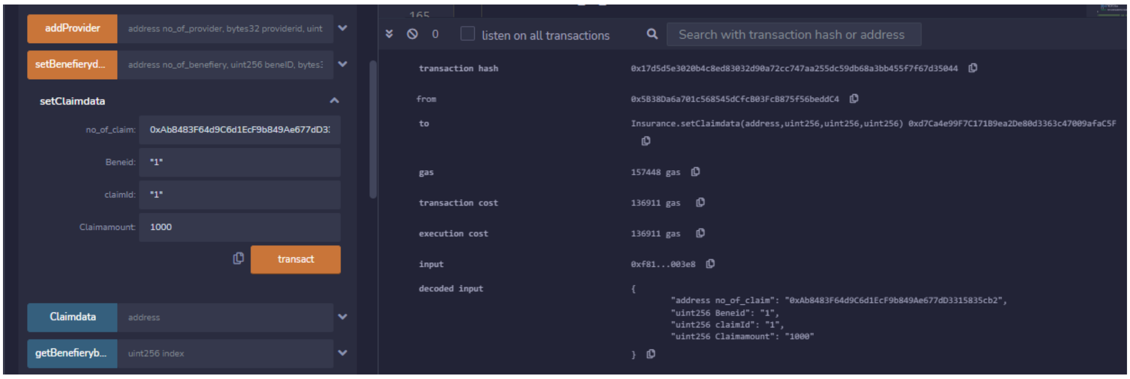
Figure 6.11: Smart contract function

function for validation of data. The GUI of smart contract function are shown in figure 6.11 The function of developed HIC smart contract are as follows:

- *addProvider*: This function can only executed by HI provider firm to add provider information. This function help to add information of fraudulent provider and non fraudulent information.

- *setBeneficiarydata*: This function can only executed by healthcare service provider to add beneficiary related information. This function set information of beneficiary such as beneID, providerid, attending physician, age, patient risk score, hospitalization duration.

- *retriveBenifieryInfo*: This function can executed by healthcare service provider and Benifiery to view the beneficiary related information.


addProvider
setBenefieryd...
Claimdata
getBenefieryb...
transact
listen on all transactions
Search with transaction hash or address

setClaimdata

no_of_claim: 0xAb8483F64d9C6d1EcF9b849Ae677dD3:

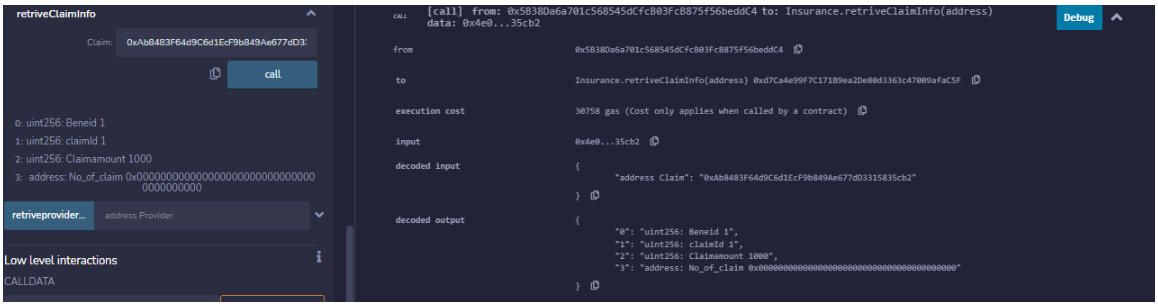
Beneid: "1"

claimId: "1"

Claimamount: 1000

```
transaction hash     0x17d5d5e3020b4c8ed83032d90a72cc747aa255dc59db68a3bb455f7f67d35044
from                 0x5B38Da6a701c568545dCfcB03FcB875f56beddC4
to                   Insurance.setClaimdata(address,uint256,uint256,uint256) 0xd7Ca4e99F7C171B9ea2De80d3363c47009afaC5F
gas                  157448 gas
transaction cost     136911 gas
execution cost       136911 gas
input                0xf81...003e8
decoded input        {
                             "address no_of_claim": "0xAb8483F64d9C6d1EcF9b849Ae677dD3315835cb2",
                             "uint256 Beneid": "1",
                             "uint256 claimId": "1",
                             "uint256 Claimamount": "1000"
                     }
```

Figure 6.12: set claim data transaction



retriveClaimInfo

Claim: 0xAb8483F64d9C6d1EcF9b849Ae677dD3:

```
0: uint256: Beneid 1
1: uint256: claimId 1
2: uint256: Claimamount 1000
3:  address: No_of_claim 0x000000000000000000000000000000
    0000000000
```

```
CALL  [call] from: 0x5B38Da6a701c568545dCfcB03FcB875f56beddC4 to: Insurance.retriveClaimInfo(address)
      data: 0x4e0...35cb2

from               0x5B38Da6a701c568545dCfcB03FcB875f56beddC4
to                 Insurance.retriveClaimInfo(address) 0xd7Ca4e99F7C171B9ea2De80d3363c47009afaC5F
execution cost     30758 gas (Cost only applies when called by a contract)
input              0x4e0...35cb2
decoded input      {
                           "address Claim": "0xAb8483F64d9C6d1EcF9b849Ae677dD3315835cb2"
                   }
decoded output     {
                           "0": "uint256: Beneid 1",
                           "1": "uint256: claimId 1",
                           "2": "uint256: Claimamount 1000",
                           "3": "address: No_of_claim 0x0000000000000000000000000000000000000000"
                   }
```

Figure 6.13: Retrieve claim data transaction

- *setClaimdata*: The function is created for setting claim information data. This function is only executed by beneficiary(insurance subscriber) who want to subscribe for HIC. The setClaimdata function's transaction cost and execution cost in terms of gas amount value are shown in figure 6.12.

- *retriveClaimInfo*: This function can executed by healthcare service provider and Benifiery to view the beneficiary related information. The retriveClaimInfo function's transaction cost and execution cost in terms of gas amount value are shown in figure 6.13.

- *getBeneficiaryCount*: This function return count of the beneficiary which are set by healthcare service provider.

- *signAgrementdata*: This function can only executed if user is beneficiary . It add the information of agreements. This agreement is decide between beneficiary and HI provider firm. This function only executed when some condition are fulfilled such as beneficiary age must be above 18, patient risk score is not more then threshold. If beneficiary take service from fraudulent provider then this agreement is not happen.

- *retriveProviderInfo*: This function can executed by HI provider firm and Benifiery to view the information of fraudulent and check that provider is fraudulent or not.

Chapter 7

Open Issues and Research Challenges

This section highlights research challenges and open issues of blockchain and AI-empowered healthcare insurance fraud detection.

7.1 Class imbalance

Class imbalance is the vital challenge that reduces the performance of the HIC fraud classifier built based on the ML models. The majority class in the imbalance classification predictive modeling problem has more records than the minority (fraud) class [21]. The minority class is more challenging to predict because it contains fewer HIC records. There is no detailed and advanced solution for the class imbalance challenge of HIC fraud detection in the existing research study.

7.2 Standardization of Data Exchange Rule

There is no universal rule or regulation for the cross-border exchange of HIC data. The government privacy rule for HIC data exchange differs from one country to another [15]. The advantage of blockchain and AI-enabled HIC fraud detection system data sharing may be restricted by the cross-border exchange of HIC data where different jurisdictions are involved. So, there is a need of the further research work into the standardization of cross-border HIC data retrieval, storage rules, and regulations.

7.3 Real Time Data-set Accessibility

The public and private healthcare sectors should provide more opportunities for researchers to access real-world HIC data. The availability of the real-world HIC data sets has been restricted due to the legal, privacy, and competitive concerns [83]. Future research with open source HIC real-world datasets is essential, allowing other researchers to validate their results. Researchers should collaborate with the healthcare insurance sector to access the proposed solutions using real-world healthcare insurance data.

7.4 Handling Growing Number of User

One of the challenging tasks is to handle the growing number of users in AI and blockchain-enabled HIC fraud detection framework. Due to the population growth, the need for the healthcare insurance sector is increasing for the wellbeing of patients. As the number of insurance subscribers on the framework grows, AI and blockchain-enabled frameworks become exponentially more difficult to implement in real time scenario.

7.5 Amalgamation with other Technology

HI firms have a scope to amalgamate with wearables and sensor technologies used in the healthcare for the additional functionality of the business operations [84]. HI provider can utilize the wearable devices to collect real data of insurance subscriber, to analyze premium rate and, HIC fraud detection. So, future research work should focus on the wearable and IoT-based healthcare devices data availability and data analytics for blockchain and AI-enabled HIC fraud detection system.

7.6 Lack of Professional Talent

Blockchain and AI have become one of the most rapidly emerging technologies in the present era.Developing blockchain and AI-enabled HIC fraud detection applications require professionally skilled people who know the complex AI, cryptography, and advanced mathematics algorithms. Currently, only some of the researchers and professionals are working on ensuring blockchain and AI-based scheme robust and adaptive. So, adaptation of the AI and blockchain in HIC fraud detection will be a challenging task due to the lack of professional talent.

7.6.1 Weakness of Smart Contract Development

Smart contracts are executable codes designed in the solidity language and deployed on Ethereum Virtual Machines to regulate action according to the agreement. In the development phase of the smart contract, there is a possibility of a software bug. There is a requirement to handle the software bug before the smart contracts are deployed in the blockchain network as they can't be altered after the deployment. So, research solutions should be improved to deploy the smart contract efficiently and reliably.

Chapter 8

Future Scope and Conclusion

In this research project, we proposed a four-layer intelligent HI fraud detection system architecture. Further, project work has introduced a futuristic approach for the proposed HIC fraud detection system using a wearable device. Later, we implement a machine learning model which detects potential fraudulent healthcare service providers. We apply four ML algorithms to the training dataset. The dataset used in the experiment is highly imbalanced. To solve the class imbalance issue we used SMOTE method. Four classifier learners were compared and tested to find their efficiency in building the ML model. Among all Bagging classifiers produced better F1 scores. After that HIC fraud detection system-based, Smart Contract on blockchain technology is developed to validate the transaction of HIC. The proposed smart contract was created and deployed on the Remix Ethereum. Also, we have listed various research challenges and the open issue associated with the blockchain and AI-based proposed system during its real-time deployment. In the future, we will integrate the ML module with the smart contract and develop a decentralized organizational HIC application.

Bibliography

[1] J. J. Hathaliya, S. Tanwar, S. Tyagi, and N. Kumar, “Securing electronics healthcare records in healthcare 4.0: a biometric-based approach,” *Computers & Electrical Engineering*, vol. 76, pp. 398–410, 2019.

[2] A. A. Monrat, O. Schelén, and K. Andersson, “A survey of blockchain from the perspectives of applications, challenges, and opportunities,” *IEEE Access*, vol. 7, pp. 117134–117151, 2019.

[3] H. insurance, “Health insurance in india.” https://en.wikipedia.org/wiki/Health_insurance_in_India. Accessed: 2021.

[4] A. Sheshasaayee and S. S. Thomas, “A purview of the impact of supervised learning methodologies on health insurance fraud detection,” in *Information Systems Design and Intelligent Applications*, pp. 978–984, Springer, 2018.

[5] H. K. Patil and R. Seshadri, “Big data security and privacy issues in healthcare,” in *2014 IEEE international congress on big data*, pp. 762–765, IEEE, 2014.

[6] M. Ojha and K. Mathur, “Proposed application of big data analytics in healthcare at maharaja yeshwantrao hospital,” in *2016 3rd MEC International Conference on Big Data and Smart City (ICBDSC)*, pp. 1–7, IEEE, 2016.

[7] M. Eling and M. Lehmann, “The impact of digitalization on the insurance value chain and the insurability of risks,” *The Geneva papers on risk and insurance-issues and practice*, vol. 43, no. 3, pp. 359–396, 2018.

[8] R. Dutt, “The impact of artificial intelligence on healthcare insurances,” in *Artificial Intelligence in Healthcare*, pp. 271–293, Elsevier, 2020.

[9] P. Kumar and H.-J. Lee, “Security issues in healthcare applications using wireless medical sensor networks: A survey,” *sensors*, vol. 12, no. 1, pp. 55–91, 2012.

[10] M. Li, W. Lou, and K. Ren, “Data security and privacy in wireless body area networks,” *IEEE Wireless communications*, vol. 17, no. 1, pp. 51–58, 2010.

[11] J. Heurix, M. Karlinger, and T. Neubauer, “Pseudonymization with metadata encryption for privacy-preserving searchable documents,” in *2012 45th Hawaii International Conference on System Sciences*, pp. 3011–3020, IEEE, 2012.

[12] K. M. Kumar, T. S, and S. Swarnalatha, “Effective implementation of data segregation amp; extraction using big data in e - health insurance as a service,” in *2016 3rd International Conference on Advanced Computing and Communication Systems (ICACCS)*, vol. 01, pp. 1–5, 2016.

[13] D. Ulybyshev, C. Bare, K. Bellisario, V. Kholodilo, B. Northern, A. Solanki, and T. O’Donnell, “Protecting electronic health records in transit and at rest,” in *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 449–452, IEEE, 2020.

[14] “The benefits and threats of blockchain technology in healthcare: A scoping review,” *International Journal of Medical Informatics*, vol. 142, p. 104246, 2020.

[15] C. C. Agbo, Q. H. Mahmoud, and J. M. Eklund, “Blockchain technology in healthcare: A systematic review,” *Healthcare*, vol. 7, no. 2, 2019.

[16] E. A. Duman and Ş. Sağıroğlu, “Heath care fraud detection methods and new approaches,” in *2017 International Conference on Computer Science and Engineering (UBMK)*, pp. 839–844, IEEE, 2017.

[17] R. Bauder and T. Khoshgoftaar, “A survey of medicare data processing and integration for fraud detection,” in *2018 IEEE international conference on information reuse and integration (IRI)*, pp. 9–14, IEEE, 2018.

[18] T. Ekin, F. Ieva, F. Ruggeri, and R. Soyer, “Statistical medical fraud assessment: exposition to an emerging field,” *International Statistical Review*, vol. 86, no. 3, pp. 379–402, 2018.

[19] D. Ankrah, J. Hallas, J. Odei, F. Asenso-Boadi, L. Dsane-Selby, and M. Donneyong, "A review of the ghana national health insurance scheme claims database: possibilities and limits for drug utilization research," *Basic & clinical pharmacology & toxicology*, vol. 124, no. 1, pp. 18–27, 2019.

[20] Z. X. Chen, L. Hohmann, B. Banjara, Y. Zhao, K. Diggs, and S. C. Westrick, "Recommendations to protect patients and health care practices from medicare and medicaid fraud," *Journal of the American Pharmacists Association*, vol. 60, no. 6, pp. e60–e65, 2020.

[21] A. J. Mary and S. A. Claret, "Imbalanced classification problems: Systematic study and challenges in healthcare insurance fraud detection," in *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, pp. 1049–1055, IEEE, 2021.

[22] K. G. Al-Hashedi and P. Magalingam, "Financial fraud detection applying data mining techniques: a comprehensive review from 2009 to 2019," *Computer Science Review*, vol. 40, p. 100402, 2021.

[23] M. Herland, T. M. Khoshgoftaar, and R. A. Bauder, "Big data fraud detection using multiple medicare data sources," *Journal of Big Data*, vol. 5, no. 1, pp. 1–21, 2018.

[24] M. Ahmed and M. Ahamad, "Combating abuse of health data in the age of ehealth exchange," in *2014 IEEE International Conference on Healthcare Informatics*, pp. 109–118, 2014.

[25] Healthcare, "Healthcare in india." https://en.wikipedia.org/wiki/Healthcare_in_India. Accessed: 2021.

[26] P. M. Suraksha, "Pradhan mantri suraksha bima yojana." https://financialservices.gov.in/insurance-divisions/Government-Sponsored-Socially-Oriented-Insurance-Schemes/Pradhan-Mantri-Suraksha-Bima-Yojana(PMSBY). Accessed: 2021.

[27] P. M. garib kalyan, "pradhan mantri garib kalyan-package." https://www.india.gov.in/spotlight/pradhan-mantri-garib-kalyan-package-pmgkp. Accessed: 2021.

[28] I. Kose, M. Gokturk, and K. Kilic, “An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance,” *Applied Soft Computing*, vol. 36, p. 283–299, 08 2015.

[29] N. Rayan, “Framework for analysis and detection of fraud in health insurance,” in *2019 IEEE 6th International Conference on Cloud Computing and Intelligence Systems (CCIS)*, pp. 47–56, 2019.

[30] N. Rayan, “Framework for analysis and detection of fraud in health insurance,” in *2019 IEEE 6th International Conference on Cloud Computing and Intelligence Systems (CCIS)*, pp. 47–56, 2019.

[31] S. Khezr, M. Moniruzzaman, A. Yassine, and R. Benlamri, “Blockchain technology in healthcare: A comprehensive review and directions for future research,” *Applied sciences*, vol. 9, no. 9, p. 1736, 2019.

[32] L. Settipalli and G. Gangadharan, “Healthcare fraud detection using primitive sub peer group analysis,” *Concurrency and Computation: Practice and Experience*, p. e6275, 2021.

[33] M. H. Nasir, J. Arshad, M. M. Khan, M. Fatima, K. Salah, and R. Jayaraman, “Scalable blockchains—a systematic review,” *Future Generation Computer Systems*, vol. 126, pp. 136–162, 2022.

[34] J. Shah, S. Agarwal, A. Shukla, S. Tanwar, S. Tyagi, and N. Kumar, “Blockchain-based scheme for the mobile number portability,” *Journal of Information Security and Applications*, vol. 58, p. 102764, 2021.

[35] R. Saranya and A. Murugan, “A systematic review of enabling blockchain in healthcare system: Analysis, current status, challenges and future direction,” *Materials Today: Proceedings*, 2021.

[36] issues facing patient privacy, “top 3 issues facing patient privacy.” https://www.healthcareitnews.com/news/top-3-issues-facing-patient-privacy. Accessed: 2021.

[37] C. Thapa and S. Camtepe, "Precision health data: Requirements, challenges and existing techniques for data security and privacy," *Computers in biology and medicine*, vol. 129, p. 104130, 2021.

[38] atlantatech, "addressing the inadequacies of hipaa law and politics." https://www.atlantatech.news/analysis/addressing-the-inadequacies-of-hipaa-law-and-politics-involving-healthcare-it/. Accessed: 2021.

[39] S. H. Kok, A. Abdullah, N. Jhanjhi, and M. Supramaniam, "Prevention of crypto-ransomware using a pre-encryption detection algorithm," *Computers*, vol. 8, no. 4, 2019.

[40] G. Singh, D. Kant, U. Gangwar, and A. P. Singh, "Sql injection detection and correction using machine learning techniques," in *Emerging ICT for Bridging the Future-Proceedings of the 49th Annual Convention of the Computer Society of India (CSI) Volume 1*, pp. 435–442, Springer, 2015.

[41] A. Aleroud and L. Zhou, "Phishing environments, techniques, and countermeasures: A survey," *Computers Security*, vol. 68, pp. 160–196, 2017.

[42] R. Raveendranath, V. Rajamani, A. J. Babu, and S. K. Datta, "Android malware attacks and countermeasures: Current and future directions," in *2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)*, pp. 137–143, 2014.

[43] G. K. Sadasivam, C. Hota, and B. Anand, *Honeynet Data Analysis and Distributed SSH Brute-Force Attacks*, pp. 107–118. Singapore: Springer Singapore, 2018.

[44] S. Reddy and G. K. Shyam, "A machine learning based attack detection and mitigation using a secure saas framework," *Journal of King Saud University - Computer and Information Sciences*, 2020.

[45] R. Zagrouba and R. AlHajri, "Machine learning based attacks detection and countermeasures in iot," *International Journal of Communication Networks and Information Security (IJCNIS)*, vol. 13, no. 2, 2021.

[46] M. S. Yousefpoor, E. Yousefpoor, H. Barati, A. Barati, A. Movaghar, and M. Hosseinzadeh, “Secure data aggregation methods and countermeasures against various attacks in wireless sensor networks: A comprehensive review,” *Journal of Network and Computer Applications*, vol. 190, p. 103118, 2021.

[47] S. Tu, M. Waqas, S. U. Rehman, T. Mir, G. Abbas, Z. H. Abbas, Z. Halim, and I. Ahmad, “Reinforcement learning assisted impersonation attack detection in device-to-device communications,” *IEEE Transactions on Vehicular Technology*, vol. 70, no. 2, pp. 1474–1479, 2021.

[48] N. Bruce, M. Sain, and H. J. Lee, “A support middleware solution for e-healthcare system security,” in *16th International Conference on Advanced Communication Technology*, pp. 44–47, 2014.

[49] “Ransomware: Recent advances, analysis, challenges and future research directions,” *Computers Security*, vol. 111, p. 102490, 2021.

[50] V. Akashe, R. L. Neupane, M. L. Alarcon, S. Wang, and P. Calyam, “Network-based active defense for securing cloud-based healthcare data processing pipelines,” in *2021 International Conference on Computer Communications and Networks (ICCCN)*, pp. 1–9, 2021.

[51] S. A. Roseline and S. Geetha, “A comprehensive survey of tools and techniques mitigating computer and mobile malware attacks,” *Computers & Electrical Engineering*, vol. 92, p. 107143, 2021.

[52] M. M. Najafabadi, T. M. Khoshgoftaar, C. Kemp, N. Seliya, and R. Zuech, “Machine learning for detecting brute force attacks at the network level,” in *2014 IEEE International Conference on Bioinformatics and Bioengineering*, pp. 379–385, IEEE, 2014.

[53] A. M. Amin and M. S. Mahamud, “An alternative approach of mitigating arp based man-in-the-middle attack using client site bash script,” in *2019 6th International Conference on Electrical and Electronics Engineering (ICEEE)*, pp. 112–115, 2019.

[54] K. Xylogiannopoulos, P. Karampelas, and R. Alhajj, "Early ddos detection based on data mining techniques," in *IFIP International Workshop on Information Security Theory and Practice*, pp. 190–199, Springer, 2014.

[55] V. Rawte and G. Anuradha, "Fraud detection in health insurance using data mining techniques," in *2015 International Conference on Communication, Information & Computing Technology (ICCICT)*, pp. 1–5, IEEE, 2015.

[56] M. Herland, R. A. Bauder, and T. M. Khoshgoftaar, "Medical provider specialty predictions for the detection of anomalous medicare insurance claims," in *2017 IEEE International Conference on Information Reuse and Integration (IRI)*, pp. 579–588, 2017.

[57] J. C. Cassimiro, A. M. Santana, P. S. Neto, and R. L. Rabelo, "Investigating the effects of class imbalance in learning the claim authorization process in the brazilian health care market," in *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 3265–3272, 2017.

[58] P. Pandey, A. Saroliya, and R. Kumar, "Analyses and detection of health insurance fraud using data mining and predictive modeling techniques," in *Soft Computing: Theories and Applications* (M. Pant, K. Ray, T. K. Sharma, S. Rawat, and A. Bandyopadhyay, eds.), (Singapore), pp. 41–49, Springer Singapore, 2018.

[59] R. A. Sowah, M. Kuuboore, A. Ofoli, S. Kwofie, L. Asiedu, K. M. Koumadi, and K. O. Apeadu, "Decision support system (dss) for fraud detection in health insurance claims using genetic support vector machines (gsvms)," *Journal of Engineering*, vol. 2019, 2019.

[60] S. S.K. and V. Ilango, "A time-efficient model for detecting fraudulent health insurance claims using artificial neural networks," in *2020 International Conference on System, Computation, Automation and Networking (ICSCAN)*, pp. 1–6, 2020.

[61] W. Yang, W. Hu, Y. Liu, Y. Huang, X. Liu, and S. Zhang, "Research on bootstrapping algorithm for health insurance data fraud detection based on decision tree," in *2021 7th IEEE Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart*

Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS), pp. 57–62, 2021.

[62] A. Verma, A. Taneja, and A. Arora, "Fraud detection and frequent pattern matching in insurance claims using data mining techniques," in *2017 Tenth International Conference on Contemporary Computing (IC3)*, pp. 1–7, 2017.

[63] M. S. Anbarasi and S. Dhivya, "Fraud detection using outlier predictor in health insurance data," in *2017 International Conference on Information Communication and Embedded Systems (ICICES)*, pp. 1–6, 2017.

[64] H. Cao and R. Zhang, "Using pca to improve the detection of medical insurance fraud in sofm neural networks," pp. 117–122, 01 2019.

[65] S. Kareem, R. Binti Ahmad, and 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)*, pp. 99–104, 2017.

[66] X. Jiang, S. Pan, G. Long, F. Xiong, J. Jiang, and C. Zhang, "Cost-sensitive parallel learning framework for insurance intelligence operation," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 12, pp. 9713–9723, 2019.

[67] S. Zhou and R. Zhang, "A novel method for mining abnormal expenses in social medical insurance," in *2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, pp. 1–5, 2020.

[68] W. Liu, Q. Yu, Z. Li, Z. Li, Y. Su, and J. Zhou, "A blockchain-based system for anti-fraud of healthcare insurance," in *2019 IEEE 5th International Conference on Computer and Communications (ICCC)*, pp. 1264–1268, 2019.

[69] J. Gera, A. R. Palakayala, V. K. K. Rejeti, and T. Anusha, "Blockchain technology for fraudulent practices in insurance claim process," in *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, pp. 1068–1075, 2020.

[70] G. Saldamli, V. Reddy, K. S. Bojja, M. K. Gururaja, Y. Doddaveerappa, and L. Tawalbeh, "Health care insurance fraud detection using blockchain," in *2020*

Seventh International Conference on Software Defined Systems (SDS), pp. 145–152, IEEE, 2020.

[71] L. Ismail and S. Zeadally, "Healthcare insurance frauds: Taxonomy and blockchain-based detection framework (block-hi)," *IT Professional*, 07 2021.

[72] B. Alhasan, M. Qatawneh, and W. Almobaideen, "Blockchain technology for preventing counterfeit in health insurance," in *2021 International Conference on Information Technology (ICIT)*, pp. 935–941, 2021.

[73] S.-L. Wang, H.-T. Pai, M.-F. Wu, F. Wu, and C.-L. Li, "The evaluation of trustworthiness to identify health insurance fraud in dentistry," *Artificial intelligence in medicine*, vol. 75, pp. 40–50, 2017.

[74] V. F. de Santana, A. P. Appel, L. G. Moyano, M. Ito, and C. S. Pinhanez, "Revealing physicians referrals from health insurance claims data," *Big data research*, vol. 13, pp. 3–10, 2018.

[75] C. Sun, Z. Yan, Q. Li, Y. Zheng, X. Lu, and L. Cui, "Abnormal group-based joint medical fraud detection," *IEEE Access*, vol. 7, pp. 13589–13596, 2018.

[76] N. Dhieb, H. Ghazzai, H. Besbes, and Y. Massoud, "A secure ai-driven architecture for automated insurance systems: Fraud detection and risk measurement," *IEEE Access*, vol. 8, pp. 58546–58558, 2020.

[77] G. Zhang, X. Zhang, M. Bilal, W. Dou, X. Xu, and J. J. Rodrigues, "Identifying fraud in medical insurance based on blockchain and deep learning," *Future Generation Computer Systems*, vol. 130, pp. 140–154, 2022.

[78] R. A. Bauder, T. M. Khoshgoftaar, A. Richter, and M. Herland, "Predicting medical provider specialties to detect anomalous insurance claims," in *2016 IEEE 28th international conference on tools with artificial intelligence (ICTAI)*, pp. 784–790, IEEE, 2016.

[79] M. McCrea and M. Farrell, "A conceptual model for pricing health and life insurance using wearable technology," *Risk Management and Insurance Review*, vol. 21, no. 3, pp. 389–411, 2018.

[80] B. Nayak and S. S. Bhattacharyya, "The changing narrative in the health insurance industry: Wearables technology in health insurance products and services for the covid-19 world," *Journal of Health Management*, vol. 22, no. 4, pp. 550–558, 2020.

[81] wearable devices, "wearable devices will they really catch on in the health insurance industry." https://fdocuments.in/document/wearable-devices-will-they-really-catch-on-in-the-health-insurance-industry.html. Accessed: 2021.

[82] kaggle, "kaggle dataset." https://www.kaggle.com/datasets/rohitrox/healthcare-provider-fraud-detection-analysis. Accessed: 2021.

[83] T. McGhin, K.-K. R. Choo, C. Z. Liu, and D. He, "Blockchain in healthcare applications: Research challenges and opportunities," *Journal of Network and Computer Applications*, vol. 135, pp. 62–75, 2019.

[84] A. Tandon, A. Dhir, N. Islam, and M. Mäntymäki, "Blockchain in healthcare: A systematic literature review, synthesizing framework and future research agenda," *Computers in Industry*, vol. 122, p. 103290, 2020.