Beyhan ADANUR DEDETÜRK

A Ph.D. Thesis

AGU 2024

DECENTRALIZED ELECTRONIC HEALTH RECORD MANAGEMENT SYSTEM AND DISEASE PREDICTION WITH MACHINE LEARNING METHODS

A THESIS
SUBMITTED TO THE DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING
AND THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF ABDULLAH GUL UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

By
Beyhan ADANUR DEDETÜRK

June 2024

DECENTRALIZED ELECTRONIC HEALTH RECORD MANAGEMENT SYSTEM AND DISEASE PREDICTION WITH MACHINE LEARNING METHODS

A THESIS
SUBMITTED TO THE DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING AND THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE OF ABDULLAH GUL UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Ph. D.

By
Beyhan ADANUR DEDETÜRK
2024

SCIENTIFIC ETHICS COMPLIANCE

I hereby declare that all information in this document has been obtained in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all materials and results that are not original to this work.

Name-Surname: Beyhan ADANUR DEDETÜRK
Signature :

REGULATORY COMPLIANCE

Ph. D. thesis title **"Decentralized Electronic Health Record Management System And Disease Prediction With Machine Learning Methods"** has been prepared in accordance with the Thesis Writing Guidelines of the Abdullah Gül University, Graduate School of Engineering & Science.

Prepared By	Advisor
Beyhan ADANUR DEDETÜRK	Assoc. Prof. Burcu BAKIR GÜNGÖR
Signature	Signature

Head of the Electrical and Computer Engineering Program

Assist. Prof. Samet GÜLER

Signature

ACCEPTANCE AND APPROVAL

Ph. D. thesis title **"Decentralized Electronic Health Record Management System And Disease Prediction With Machine Learning Methods"** and prepared by Beyhan ADANUR DEDETURK has been accepted by the jury in the Electrical and Computer Engineering Graduate Program at Abdullah Gül University, Graduate School of Engineering & Science.

10/ 06 / 2024
(Thesis Defense Exam Date)

JURY: **Signature:**

Advisor : Assoc. Prof. Burcu BAKIR GÜNGÖR

Member : Assist. Prof. Gülay YALÇIN ALKAN

Member : Assoc. Prof. Özkan UFUK NALBANTOĞLU

Member : Assist. Prof. Nazlı TEKİN

Member : Assoc. Prof. Rıfat KURBAN

APPROVAL:

The acceptance of this Ph. D thesis has been approved by the decision of the Abdullah Gül University, Graduate School of Engineering & Science, Executive Board dated ….. /….. / …….. and numbered ……..……. .

…….. /…….. / ……..

Graduate School Dean
Prof. Dr. İrfan ALAN

ABSTRACT

DECENTRALIZED ELECTRONIC HEALTH RECORD MANAGEMENT SYSTEM AND DISEASE PREDICTION WITH MACHINE LEARNING METHODS

Beyhan ADANUR DEDETÜRK
Ph. D. in Electrical and Computer Engineering
Advisor: Assoc. Prof. Burcu BAKIR GÜNGÖR
June 2024

Electronic health records (EHRs) are vital to the advancement of healthcare and can help detect and prevent diseases early. However, EHR sharing faces challenges such as managing large data volumes, ensuring data privacy, security, and interoperability. This thesis aims to develop and analyze a blockchain-based EHR sharing system for disease prediction mechanism integration using SysML. The AguHyper platform, built by merging the InterPlanetary File System (IPFS) with Hyperledger Fabric, ensures the immutability of health records by storing hash values in the blockchain and encrypted records in IPFS. The system architecture and implementation configurations, including CouchDB and the Raft consensus mechanism, are thoroughly examined. The study also presents a novel hybrid approach called CSA-DE-LR, which integrates Differential Evolution (DE) and Clonal Selection Algorithm (CSA) with Logistic Regression (LR) to improve LR weights for precise categorization of cardiovascular diseases. The integration of the AguHyper with the CSA-DE-LR is explained in detail. At the end of our performance evaluations, we concluded that the AguHyper model has the potential to speed up the process of collecting and sharing data, and it offers an efficient platform for the participants.

Keywords: Electronic Health Records, SysML, Blockchain, Hyperledger, IPFS, Disease Prediction, Machine Learning

ÖZET

MERKEZİ OLMAYAN ELEKTRONİK SAĞLIK KAYDI YÖNETİM SİSTEMİ VE MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE HASTALIK TAHMİNİ

Beyhan ADANUR DEDETÜRK
Elektrik ve Bilgisayar Mühendisliği Bölümü Doktora
Tez Yöneticisi: Doç. Dr. Burcu BAKIR GÜNGÖR
Haziran 2024

Elektronik sağlık kayıtları (EHRs), sağlık hizmetlerinin ilerlemesi için hayati öneme sahiptir ve hastalıkların erken tespit edilip önlenmesine yardımcı olabilir. Ancak EHR paylaşımı, büyük veri hacimlerinin yönetilmesi, veri gizliliğinin, güvenliğin ve birlikte çalışabilirliğin sağlanması gibi zorluklarla karşı karşıyadır. Bu tez, SysML kullanarak hastalık tahmin mekanizması entegrasyonu için blokzincir tabanlı bir EHR paylaşım sistemi geliştirmeyi ve analiz etmeyi amaçlamaktadır. Gezegenler Arası Dosya Sisteminin (IPFS) Hyperledger Fabric ile birleştirilmesiyle oluşturulan AguHyper platformu, şifrelenmiş kayıtları IPFS'de ve hash değerlerini ise blokzincirde depolayarak sağlık kayıtlarının tahrip edilemezliğini sağlamaktadır. CouchDB ve Raft konsensüs mekanizması da dahil olmak üzere sistem mimarisi ve uygulama konfigürasyonları kapsamlı bir şekilde incelenmektedir. Çalışma ayrıca, kardiyovasküler hastalıkların kesin kategorizasyonu için LR ağırlıklarını iyileştirmek amacıyla Diferansiyel Evrim (DE) ve Klonal Seçim Algoritmasını (CSA) Lojistik Regresyon (LR) ile birleştiren CSA-DE-LR adı verilen yeni bir hibrit yaklaşım da sunmaktadır. AguHyper'ın CSA-DE-LR ile entegrasyonu ayrıntılı olarak anlatılmaktadır. Performans değerlendirmeleri sonucunda AguHyper modelinin veri toplama ve paylaşma sürecini hızlandırma potansiyeline sahip olduğu ve katılımcılara verimli bir platform sunduğu sonucuna varılmıştır.

Anahtar Kelimeler: Elektronik Sağlık Kayıtları, SysML Blokzincir, Hyperledger, Gezegenler Arası Dosya Sistemi, Hastalık Tahmini, Makine Öğrenmesi

Acknowledgements

I would like to convey my thanks gratefully to my advisor, Assoc. Prof. Burcu BAKIR GÜNGÖR. Her diligence and professional behaviors are always going to be a guide for me in the way of being a scientist. I would like to express my gratitude to Asst. Prof. Gülay YALÇIN ALKAN and Assoc. Prof. Özkan Ufuk NALBANTOĞLU for taking their valuable time for following my progress and helping me with their advice.

I also would like to express my deepest gratitude to my family. I would like to thank my father, Hayrettin, and my mother, Nurhan, for their endless patience, love, and labor. I would like to thank my husband Bilge and my brother Erhan for their guidance in my life, friendship, and support. I couldn't have achieved this without them helping me with my work. Finally, my dear daughter, Tomris, everything I do is for you. Everything I've learned today is thanks to my family. I am grateful for everything.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

EHRs	Electronic Health Records
BC	Blockchain Technology
CSA	Clonal Selection Algorithm
DE	Differential Evolution
LR	Logistic Regression
IPFS	InterPlanetary File System
CAD	Coronary Artery Disease
CVD	Cardiovascular Disease
MCC	Matthew's Correlation Coefficient
MAE	Mean Absolute Error
DSC	Digital Signature Certificate
CA	Central Authority
DoS	Denial-of-Service
tps	Transaction Rate
Tx	Number of Transactions
SUT	System Under Test
WHO	World Health Organization
ML	Machine Learning
FL	Federated Learning
WBCO	Breast Cancer Wisconsin Original
WBCD	Breast Cancer Wisconsin Diagnostic
FS	Feature Selection
ANN	Artificial Neural Network
ACC	Accuracy
FNR	False Negative Rate
FPR	False Positive Rate
ROC	Receiver Operating Characteristic
AUC	Area Under the Curve
std	Standand Deviation
PSO	Particle Swarm Optimization

To my mother and daughter

Chapter 1

Introduction

Technological advances in recent decades have brought about significant developments in our environment, including improvements in healthcare systems. Healthcare systems were limited in the 1970s, 80s, and 90s, and because of a lack of funding, they were not in sync with digital systems. Healthcare providers started creating online platforms between 1990 and 2005 to share information, store data on cloud servers, and provide mobile access to patient records. This allowed for continual access for both the patient and the physician. Electronic Health Records (EHRs) and wearable and implantable technologies were introduced between 2005 and 2016, allowing for the real-time, ubiquitous tracking of a person's medical history. Similar methods of exchanging health data between practitioners and across networked channels have been used by EHR systems. Also improved were the exchanges of information and interactions between patients and providers. There have been high-tech and high-touch solutions used since 2016. Blockchains that provide real-time access to patient clinical data are created using these systems, which combine cloud computing, fog and edge computing, big data analytics, artificial intelligence, and machine learning [1].

EHRs are now essential to the development of the biological sciences. In the field of healthcare, many methods of determining and assessing illness history are currently being researched [2, 3]. For example, vital signs like blood pressure and pulse may be tracked using patient health data from smart watches and other comparable sensors, and a prediction engine can help doctors assess the condition. Because they make it possible to anticipate and prevent diseases before they pose a major threat, these discoveries have a profound effect on people's lives. Additionally, they make it possible to use nutrition, focused therapy, and customized medicine in treatment.

1.1 Motivation and Problem Statement

EHRs provide healthcare professionals with rapid access to patients' historical medical information, helping to make better and faster treatment decisions. Additionally, it contributes to reducing errors, ensuring treatment continuity and improving care coordination. Thanks to EHRs, it becomes easier to store, share and analyze patient information in digital format, increasing effectiveness and efficiency in healthcare. However, EHRs ought to be interchangeable with ease for consistency and efficiency's sake. Regretfully, EHR sharing is still not as widespread as one would like. The following are the main problems that contemporary healthcare systems are facing: i) Managing large volumes of data is challenging, ii) data privacy, security, and interoperability are not always ensured, iii) users do not have control over data access rights, and iv) analysis costs are significant [4]. In traditional systems, EHRs are delivered through an intermediary rather than a direct connection between the data source and the requester. Customers are therefore unable to change their data access rights, even when the charge rises. EHRs are a very sensitive subject since they include personal data about individuals. People want to be sure that their sensitive information is safe and secure as a consequence. Hospitals and clinics may exchange data internally, but because of infrastructure constraints or employee resistance, they are unable to move data between systems. All issues must be resolved, EHRs must be simple to exchange, and data analysis on these records must be routine procedure if efficiency is to be maximized.

Blockchain technology (BC) can be a major and useful instrument for achieving these goals because of its decentralization and capacity to maintain data in an open and irreversible manner. The decentralized and reliable technology known as the Blockchain does away with middlemen and the need for centralized transaction verification [5]. Because transparency in blockchain merges different processing resources from several nodes in the network to produce incredibly rapid computation, it allows speedier access to ledger-based transactions across networks [6]. Because blockchain technology can solve a wide variety of issues in novel ways, it has lately acquired favor in numerous industries [7-9], including EHR systems [10-29].

The main operating concept of blockchain-based EHR sharing systems may be summed up as follows when all research conducted since 2016 are taken into consideration. First, the proper method is used to encrypt EHRs. The data should ideally

be saved off-chain because of the bulk of the encrypted information. Data characteristics and hashes are stored in blockchain. Users who want data as well as those who create it join the blockchain network. The network's data access permissions are configured by the data provider in accordance with their wishes. The individual asks for the pertinent information. The supplier automatically provides the requester with the data address and data key if the necessary requirements are satisfied. Then, using artificial intelligence techniques, illness prediction may be made using this data. Ultimately, a permanent record of every transaction is kept along the chain. Using this method, data providers may take back control of their data access permissions and interact directly with data requesters without the need for a middleman. Costs are lowered as a result, and the data access control issue is fixed. Every user on a blockchain network contributes to the creation of an immutable ledger. As a result, transaction histories become easier to see and follow. Most of the time, data security and privacy are achieved by encrypting the data and keeping it off-chain. Lastly, the integration of off-chain storage with blockchain ensures the interoperability of EHRs. Because the data to be stored in off-chain storage and the quantity, characteristics, and format of the information to be stored in blocks are predetermined by the system.

In addition to their benefits, EHR sharing systems each have unique drawbacks that have yet to be resolved, making it impossible to categorize any of them as the perfect system. The following is a summary of these flaws. There is no discussion of how to guarantee the accuracy and consistency of the data or address the problem of mistrust in authority. Scalability is a problem for which the platform lacks a precise fix. It is unclear how people will be able to access or participate in research utilizing permissionless blockchain. Furthermore, the possibility that a malicious owner might send meaningless data to a recipient has not been considered, even if data is encrypted and exchanged inside the system. Research employing permissioned blockchains sometimes do not provide a detailed description of the circumstances under which users can join the system or the process by which permissions are chosen. Furthermore, it is believed that system administrators are completely reliable, and the impact of hostile agents on the system has not been investigated. Many systems do not specify how data is transmitted during data sharing, even while they provide the recipient to examine the properties of the data prior to data sharing [11–29].

As per our findings, platforms have demonstrated a tendency to address some outstanding issues related to the sharing of EHRs that they deem significant. The whole

data sharing process, including i) access control (data interoperability), ii) permissions, and iii) data verification, recording, and entry processes (privacy and security), as well as user registration and roles, should be carefully examined in order to guarantee optimal efficiency. In this approach, it is required to identify every component that comprises the platform, identify the prerequisites that these components must meet in order for them to function in a safe and healthful manner, and examine the circumstances in which the prerequisites are satisfied. SysML may now be regarded as a useful and efficient tool. A graphical modeling language called SysML is used to define, examine, create, and validate complex systems [30]. Thanks to SysML, it is now possible to anticipate errors and risks that may arise during the system development life cycle and to ensure that the designed systems and their internal structure can be easily understood even by stakeholders from different

1.2 Objectives and Contributions

The purpose of this thesis is to develop and show performance analysis of an ideal EHRs sharing system that is blockchain-based suitable for disease prediction mechanism integration and addresses all EHR sharing problems in detail using SysML. As far as we could find, no research has used SysML to investigate the entire process of the aforementioned blockchain-based EHR sharing platform. To begin in this manner, the relevant platform's constituent parts were identified, and the relationships between them were examined. Furthermore, the necessary conditions for the safe and efficient functioning of the previously described system have been established, and the relationship between these conditions and the elements of the platform has been investigated. Furthermore, an example scenario is used to demonstrate how these criteria might be satisfied and system performance analysis.

In this thesis, the AguHyper framework was proposed and built by merging InterPlanetary File System (IPFS) with Hyperledger Fabric [31]. While IPFS uses decentralized databases to address the issues with centralized storage, using a permissioned BC guarantees safe interactions. The immutability of health records is achieved by storing hash values in the blockchain and encrypted records in IPFS, making the framework impervious to tampering. The system architecture and AguHyper implementation configurations—which include the usage of CouchDB and the Raft consensus mechanism—are thoroughly examined in our study. The CouchDB

database and the Raft consensus mechanism were implemented as part of the experimental setup [32]. System performance was rigorously assessed on datasets of different sizes, with particular attention paid to factors like uploading-downloading time, average transaction latency, and transaction throughput. A comparison examination of the system's performance versus pertinent literature research using various consensus processes and database formats marked the study's conclusion.

Moreover, a novel hybrid approach known as CSA-DE-LR was presented in this thesis. It integrates the Differential Evolution (DE) and Clonal Selection Algorithm (CSA) with Logistic Regression (LR). The purpose of this integration is to effectively improve logistic regression weights for the precise categorization of Cardiovascular Diseases (CVD). Based on the F1 score, the Matthews correlation coefficient (MCC), and the Absolute Error (MAE), the technique uses three optimization methodologies. Comprehensive tests on industry standard datasets, Cleveland and Statlog, show that CSA-DE-LR performs better than the most advanced machine learning techniques. Interestingly, it also shows better efficacy when compared to earlier studies in this field. Finally, the proposed blockchain-based AguHyper was integrated with the new disease prediction mechanism CSA-DE-LR, and these integration stages were explained in detail. We think that this study will be accomplished by bringing this topic to the attention of scientists from other domains and by giving them the opportunity to create fresh strategies to address the challenges these difficulties generate. The following succinctly describes our study's primary contributions:

1) Our study indicates that some of the outstanding issues with EHR sharing that they deem significant have been addressed by the blockchain-based EHR sharing systems that are now in use. The whole data sharing process, including i) access control (data interoperability), ii) permissions, and iii) data verification, recording, and entry processes (privacy and security), as well as user registration and roles, should be carefully examined in order to guarantee optimal efficiency. Thus far, no research has been discovered that employs SysML to investigate the entire set of operations associated with the aforementioned blockchain-based EHR sharing platforms.
2) The gaps in the research have been identified by looking at the current blockchain-based EHR sharing solutions. According to these conclusions, a comprehensive list of the elements of the perfect blockchain-based EHR sharing platform has been established.

3) SysML was used to assess the relevant platform's component parts and the connections among them. Here, the primary goal of using SysML is to facilitate the understanding of the entire present system by academics from many domains. Therefore, we believe that scholars from many fields may offer various strategies for resolving connected issues.
4) In addition, the conditions necessary for the system to function effectively were established, and the relationship between these conditions and the platform's constituent parts was investigated. Through an example situation, it is demonstrated how these needs may be satisfied.
5) This framework has produced a prototype that explores BC technology, fills in the gaps in earlier research, and shows off its possible uses in medical treatments.
6) A thorough description of the BC-based healthcare system's deployment and performance assessment are given. Using the Raft consensus technique, the CouchDB database was deployed as part of the experimental setup. A comparative examination of pertinent literature research using various consensus procedures and database formats was included in the study's conclusion. Our search revealed that there isn't a single study in the literature that evaluates research from this angle.
7) The permissioned BC-based decentralized EHRs sharing architecture and smart contract design that is being proposed outperforms current systems in terms of uploading-downloading time, average transaction latency, and transaction throughput.
8) Using a decentralized file system for off-chain data storage offers superior security against single points of failure, DoS attacks, and degradation of data integrity, all while providing speed that is equivalent to that of current centralized database systems.
9) This thesis also introduces CSA-DE-LR, a novel approach to classification that combines Logistic Regression with DE and CSA. Especially in the context of CVD, this novel hybrid technique is designed to improve LR weights for effective categorization. Without concentrating on ML algorithm training, the majority of investigations in the literature have employed meta-heuristic techniques for feature selection and parameter optimization problems. In contrast to these works, the suggested strategy trains the machine learning

algorithms using metaheuristics. It also gives thorough explanations of the reasoning for the combination of these three particular approaches.

10) The suggested CSA-DE-LR approach provides three different optimization approaches based on the F1 score, MAE, and MCC. In order to get the best classification performance, the model weights must be adjusted using these measures, which also serve as training guidelines.
11) The Cleveland and Statlog datasets, two well-known datasets, are used in the study to thoroughly assess the CSA-DE-LR approach. Accuracy, F1 score, MCC, ROC-AUC, false negative rate, and false positive rate are only a few of the metrics used to evaluate the performance. The outcomes are then compared with other well-known machine learning methods. When contrasting the suggested approach with prior research and presenting the findings, great care is taken to ensure complete transparency and equity. Additionally assessed are the moral ramifications of applying ML models to the medical field.
12) In contrast to previous research, this thesis offers insightful information about the significance of feature selection and model optimization by in-depth examination. It investigates how improving predictive consistency and generalizability might result from removing certain characteristics, emphasizing the significance of dataset-specific tuning and rigorous feature selection using an alternative method.
13) The outcomes demonstrate that, on the Cleveland and Statlog datasets, CSA-DE-LR performs better than earlier techniques in terms of accuracy and precision. This illustrates the method's efficacy and promise for enhancing medical professionals' diagnostic decision-making processes.
14) The proposed blockchain-based AguHyper has been integrated with the new disease prediction mechanism CSA-DE-LR, and how this integration is done is explained in detail.

The organization of thesis is as follows: Chapter 2 explains Background of EHRs; Chapter 3 processes Background of Blockhain; Chapter 4 shows Blockchain Applications in EHRs; Chapter 5 presents Methods used in this study; Chapter 6 explains the Design of an Ideal EHR Sharing Platform based on SysML and Blockchain; Chapter 7 shows Implementation of AguHyper: Results and Discussion; Chapter 8 presents A Novel Classification Algorithm: CSA-DE-LR; Chapter 9 explains Integration of the CSA-DE-LR with AguHyper; finally Chapter 10 shows Conclusions.

Chapter 2

Background of EHRs

2.1 Electronic Health Records

EHRs are digital records of patients that are safely shared, maintained, and accessed by several authorized users to facilitate the efficient and ongoing administration of integrated healthcare [33-35]. EHRs contain information regarding medical histories of patients, including diagnosis, test results, hospital admissions, surgical procedures, and medication histories. They provide a description of the patient's condition, enabling a more thorough diagnosis and course of care [36]. When necessary, EHRs can be shared with other medical professionals. Nevertheless, during transmission, EHRs are vulnerable to a variety of security and privacy threats [36, 37].

The healthcare industry is very interested in creating a safe EHR sharing environment because of its extensive use. The most recent research [38-40] shows that adopting EHR software has several advantages, such as lower costs, better healthcare quality, the development of evidence-based medicine, more extensive data collecting, and flexibility. As a result, the word "EHR" in this study refers to a system that may be used to enforce and maintain data completeness, resilience to failure, high availability, and consistency of security standards, in addition to an electronic database for storing and retrieving health information. Table 2.1 summarizes the features of EHR.

2.2 EHR Privacy

Protecting patients' rights over their data, including both data security and physical privacy, is referred to as EHR privacy. It entails guaranteeing that patients have authority over health-related data that is protected by strict security and privacy guidelines [42]. EHR privacy features also include ways to monitor data transfer and access, protect against social or economic discrimination, and promote confidence in healthcare systems.

Table 2.1 PCS: Privacy, Confidentiality, and Security [41].

Glossary	Description
Privacy	In the context of EHR, privacy is deined as the right of individuals to keep their health information conidential and control over the access and use of this data. It is a fundamental patient right under various health laws and regulations.
Confidentiality	Refers to the ethical and legal duty of healthcare professionals to protect personal health information from unauthorized disclosure. It is critical to maintain trust between patients and healthcare providers.
Security	Encompasses the technical and organizational measures to protect EHR data from unauthorized access, use, disclosure, disruption, modiication, or destruction. Security practices are crucial for maintaining the integrity and availability of health data.
EHR Management	Involves the systematic approach to handling and governing EHR systems, focusing on eicient and secure data handling, storage, and exchange, ensuring compliance with legal and ethical standards.
EHR Systems and Technologies	Refers to the hardware, software, and methodologies used in EHR systems. This includes traditional and emerging technologies like cloud computing, blockchain, and AI-driven analytics used for enhancing EHR functionalities.
PCS Framework	Represents the integrated approach of Privacy, Conidentiality, and Security in EHR systems. It underscores the interrelatedness of these aspects in ensuring holistic protection and governance of health information.

Patients have rights to privacy with regard to both their data and bodily privacy. The foundation of medical practice is the relationship of trust between patients and healthcare providers. A patient must have complete faith in the doctor before disclosing any sensitive, humiliating, or perhaps harmful personal information. A doctor has to have faith that a patient is providing enough details for a diagnosis to be made correctly and that the patient is competent to consent to treatments that carry significant risks [43]. Privacy is a fundamental element of the trust that exists between a physician and a patient. Hippocrates stressed the significance of privacy more than two millennia ago, and medical practice has acknowledged and appreciated privacy ever since [44, 45].

One of the main cybersecurity problems of today is privacy, and both end users and academics often discuss privacy issues in the omnipresent healthcare system. Patients utilizing EHR systems ought to be in charge of their health-related data, and strict national and international EHR privacy and security regulations ought to protect this data. This has to go beyond just compensating people who are at danger and include steps for data breaches that have already happened. Protecting against social or economic discrimination and fostering trust in the healthcare system can be

accomplished in this way. But in order to guarantee that vital health information is still available at the point of service, mechanisms for managing privacy protection must be in place. Appropriate privacy-preserving technologies are necessary for patients to maintain control over their own data; these systems may also aid in monitoring who has accessed and received a record [42]. A patient has the right to know what personal health information has been gathered about them and how it has been used, in accordance with information privacy laws.

EHRs and other software systems that handle sensitive user data are having trouble maintaining a high degree of data privacy [35]. Since health information is sensitive and confidential, it should only be accessed or utilized by those who have been allowed and approved, such as medical professionals. Extensive guidelines and standards have been established to guarantee the security and privacy of users' data. Health data transfer is subject to stringent security regulations, with serious consequences for noncompliance [35].

2.3 EHR Confidentiality

Identification-specific personal health information is protected in EHRs and is only shared with express, informed agreement. It guarantees that private information is protected against unwanted dissemination [46]. Maintaining confidentiality in EHR systems requires taking precautions such data encryption and abiding by privacy legislation [47]. Safeguarding identifiable personal information is a component of confidentiality. An individual's identification and personal data will only be shared with another individual or department with their express informed consent, according to an agreement and informed consent procedure [46, 48, 49]. It should be made clear that maintaining data confidentiality cannot be avoided. Encrypting electronic data is necessary to maintain its confidentiality because granting access to it might jeopardize someone's privacy [46]. One of the fundamental tenets of cybersecurity is confidentiality, which guarantees that personal data is shielded from unlawful exposure [47].

2.4 EHR Security

The main goal of electronic health record security is to prevent unauthorized access, abuse, and breaches of patient data. It includes access control, authorization, and authentication protocols [50]. Technical and administrative methods are used to secure EHRs, and adherence to HITECH and HIPAA regulations is essential [51]. The availability, confidentiality, and integrity of health information are guaranteed by security. Since EHRs are shared by several systems, there is a risk of abuse or unauthorized access due to inadequate security implementation, including authentication, authorization, and access control [50]. This raises concerns regarding patient privacy. This raises concerns regarding patient privacy. Developing policies and procedures for access control is essential to the security of EHR systems. Protecting personal data from unintentional or intentional destruction, loss, change, disclosure, or access is known as data security. In order for the shared care paradigm to be designed, implemented, and managed, it has been crucial to ensure the security of EHR systems. For EHR systems to meet these needs, the security and privacy standards must be established. Healthcare organizations have emphasized the significance of standards in ensuring the security and privacy of EHRs while offering shared and interoperable EHR services [52].

In order to guarantee confidentiality, integrity, and availability of data, EHR interoperability necessitates information security, which includes limiting unauthorized access, use, disclosure, and modification of data [53]. Wireless communication protocols, which are used by EHR connections, have the potential to create enormous amounts of data on a regular basis, providing attackers with an avenue to execute a variety of security assaults. Healthcare records might be compromised by an unsecured Healthcare 4.0 approach [54], giving hackers complete access to patient email addresses, messages, and reports [55]. To prevent unauthorized people from accessing patient data, security techniques are employed to manage access. Operational controls inside a privacy-preserving entity can do this [56].

Chapter 3

Background of Blockchain

3.1 Introduction

Even though the globe is altering and evolving at an astounding rate, technology has unquestionably contributed to this advancement. These days, technology actively influences every facet of our existence. With the advancement of information and communication technology, people's work habits are ever-changing regardless of the type of business they are in. Every modification serves a function and encourages innovation in individuals. It's also critical to stay on top of these advancements and apply the benefits they offer to corporate operations. These days, blockchain technology is receiving a lot of interest from a variety of industries, including EHRs [57–59]. There has been a breakdown in traditional business process models as it provides a fresh approach to challenges. The late 1980s and early 1990s saw the public release of the fundamental blockchain concept. On the other hand, the first cryptocurrency, Bitcoin, debuted in 2008 with the proposal of an anonymous author named Satoshi Nakamoto in his white paper [60]. Data transfer is carried out in various fields (multimedia, communication, web interface, etc.) in today's Internet environment. The new technology known as blockchain enables us to transfer assets that have worth to us as well [61]. It is a chain model-based ledger of transactions. This distributed system is controlled such that all users may decide together without the need for a central management, and it is unbreakable.

3.2 Components

The two main problems with crypto technology are single-point of failure and double-spending. The primary goal of blockchain technology was to avoid double-spending of electronic currency in the absence of a central middleman. In issue resolution, the central middleman is avoided as it may result in a single point of failure.

In order to facilitate distributed verification of operations, every computing node in the blockchain network needs to store every transaction. Additionally, it must implement a distributed timestamp protocol, which uses the actual time of a computer-recording case, to decide which transactions should be approved and which should be denied [62].

Blockchain may seem complicated, but by dissecting each component one by one, it may be made simpler. It uses common procedures from computer science and elementary cryptography. Cryptographic hash functions, digital signatures, transactions, asymmetric-key cryptography, ledgers, blocks, and the arrangement of blocks are the main elements of blockchain. Blocks and transactions are the two fundamental ideas of blockchain technology. A transaction is any instance of content information that takes place within a blockchain network. Depending on the architecture, this data may include variables like client records, fixture input, and money transfers. These records include information on money transfers for virtual currencies. At certain intervals, the records are written into blocks after being merged and processed. Block rewards are given to miners in recognition of their work when they find new blocks. A block is rewarded to the winning miner by being added to the chain as the first transaction. An unchangeable record of transactions is what is meant to be represented by the chain structure that is formed when each newly discovered block is connected to the preceding block. When a block is created, digital signatures and cryptographic hash methods are often employed. Transactions are executed on blocks containing their hash values rather than the underlying data [63]. The block structure is shown in Figure 3.1.

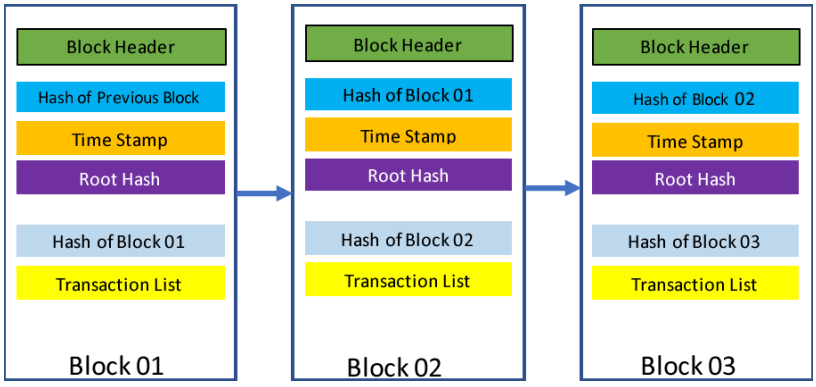

Figure 3.1 Structure of blocks [64].

Using mathematical functions of different data lengths, the hash function generates a unique result of a specified length. As a one-way function, no connection is made between the original text and the summary value when it is examined.

Furthermore, the summary value does not yield the actual data (its effectiveness against quantum computing is explored in [65]). Any modifications made to the source data during the summarizing process will likewise affect the summary value. Because of this, hash functions are frequently employed as comparison and data validation mechanisms. Algorithms like as the RIPEMD, MD, and SHA families are examples of summary functions [66]. The type of digital signature used depends on the document's content. Put another way, because digital signatures are made up of a mix of private keys and message hashes, each one alters in a unique way based on the signed message. The foundation of digital signatures is asymmetric cryptography. An encryption system that uses two distinct keys for the encryption and decryption procedures is known as asymmetric cryptography. It is utilized in two ways: (i) signing with a private key and verifying with a public key; (ii) encryption with a public key and decryption with a private key. While the private key is known solely to the individual, the public key is known to all users on the network. Symmetric cryptography, as contrast to asymmetric cryptography, employs a single key—referred to as the secret key—for both encryption and decryption processes [67].

3.3 Blockchain Categorization

There are two sorts of blockchains: permissioned and permissionless, aside from standard databases and distributed ledger technologies [68]. Anybody may join a permissionless blockchain, also known as a public blockchain, and operate as a user, miner, or developer without requiring permission from any authorities. Since all transactions are visible, anybody can publish blocks and view the specifics of any transaction. Their development aims to achieve a fully decentralized network. Every permissionless blockchain eventually has a token attached to it, most of which are meant to incentivize and reward network members. In contrast to permissionless blockchains, users on permissioned blockchains need authorization from a higher authority in order to publish blocks. This allows for the regulation of which users are able to carry out specific network actions. The chain management organization serves as a crucial oversight body for the participants and governance frameworks. Although networks are more centralized than public blockchains, they may also be created and managed with open source or closed source software. Businesses who wish to work

together and exchange data but do not want their private, confidential information to be shown on a public blockchain are drawn to them. Private and consortium blockchains are the two subclasses of permissioned blockchains [69]. Because consortium blockchains differ in a few ways from private blockchains, they are actually a subset of private blockchains. Consensus blockchains are managed by a group, whereas private blockchains are controlled by a single person. For companies that collaborate but also compete, this collaborative model will be the best choice.

3.4 Consensus Algorithms

In computer science, consensus algorithms are those that allow for an agreement on certain requests in distributed processes or systems. These algorithms do not require these systems or processes to be dependable in order to compromise them. Therefore, the blockchain's structure—which is not depend on mutual trust—is provided by consensus algorithms. They are essential to maintaining the security and effectiveness of blockchain. Table 3.1 [70] lists the most widely used algorithms in the blockchain sector.

Selecting the appropriate consensus algorithm for a particular issue is essential to enhancing system efficiency, which might lead to a rise in the quantity of blockchain-based apps [71]. Each algorithm has its own advantages and disadvantages, depending on the purpose and requirements of the systems but common goals of blockchain consensus models can be listed as follows:

- To reach an agreement
- To cooperate with the participants
- To offer equal rights to each participant
- To ensure that every member of the group is equally active

3.5 Key Benefits and Open Issues

Prior to discussing the benefits and drawbacks of blockchain technology, it is important to remember that these returns might vary depending on how systems are implemented and used. The following is a broad outline of blockchain's benefits and drawbacks.

Table 3.1 Different types of consensus algorithms

Consensus Algorithms	Explanations
PoW	When a user initiates a transaction, miners attempt to solve a cryptographic problem to test that they have worked a lot
PoS	A user encouraged to spend more on building a block until he becomes a validator
PoWeigth	Similar to PoS but the difference is that it depends on several other variables known as weights
PoB	Based on the amount, users submit the coins back into their wallet that they cannot recover from will receive rewards
PoC	Using this protocol, you can use the user's hard drive functionality
DPoS	As with PoS, but users having more coins will be able to vote and nominate witnesses
DBFT	Focuses on a gamified way of block checking among the qualified node checks
PBFT	Byzantine made use of a specific sequence to keep the rouge users at bay

Key benefits of the blockchain [72];

- Distributed management of the information transmission process is made possible by blockchain, as opposed to central management. This information management procedure is transparent to all parties and is documented in an unbreakable manner.
- Every transaction must be validated in accordance with the method employed by the system's nodes; in this way, the transactions become more reliable and secure.
- Those that participate in the block-level processing of transaction outcomes and provide their processing power to the system get money.
- The system's nodes carry out transactions in an anonymous manner, maintaining independent control over all data and processes.
- Stakeholders may readily trust one another because of digital signatures and verifications.

- Smart contracts enable the automation of some tasks.

Open issues of the blockchain [73];

- The data in the blockchain is held independently in each node, and the consistency of these data is assured as a consequence of each completed operation;
- Blockchain systems based on the Proof of Work algorithm demand a lot of energy. This explains the low performance as compared to standard databases.
- Because every node in the network has visible access to and storage of data, people's privacy may be compromised.
- As the number of apps utilizing blockchain networks has grown, so has the additional burden required by the system. Performance and scalability issues follow as a result. When the demands on a big distributed system grow, the algorithms operating on it will attempt to execute millions of operations per second. As a result, system performance may suffer.

Consequently, it would be wiser to include blockchain technology into a system after determining whether it is necessary for it [74]. Please be aware that not all distributed problems can be solved using blockchain.

Chapter 4

Blockchain Applications in EHRs

Because blockchain technology is decentralized and irreversible, it offers safe solutions for EHRs and has attracted substantial attention in a number of disciplines, including EHR administration. In an effort to solve management issues, researchers have developed a number of blockchain-based EHR sharing systems since 2016 [75]. Research conducted from 2016 to 2018 primarily concentrated on core development to show that blockchain platforms could be implemented in healthcare systems. This included studies on EHR sharing and genomic data [76, 11-13, 15-16]. Studies from 2019 onwards only looked at EHR sharing, progressively lessening the reliance on blockchain and incorporating other strategies. The integration of cloud-based, encryption-based solutions and the assessment of system performance are highlighted by BC technology in the research done between 2019 and 2020 [1, 14, 17, 19-20]. Blockchain developed into a platform between 2021 and the present, with an emphasis on creating blockchain-based healthcare systems with patient monitoring and illness prediction techniques [21-29, 77-83]. This phase signifies the initial stages of building a data ecosystem using blockchain technology.

Using the Ethereum blockchain and IPFS, Jabarulla and Lee [25] present a proof-of-concept for a distributed Patient-Centric Image Management (PCIM) system. The solution provides secure patient data control and decentralized storage in an effort to address issues with medical picture sharing and storage. Distributed access control is implemented using an Ethereum smart contract, and the framework's viability and efficiency are confirmed through testing on an Ethereum testnet. Nonetheless, problems with customer accessibility and data entry procedure clarity still exist. Furthermore, even in an encrypted system, the study does not take precautions against fraudulent data transfer and ignores potential modification of data quality.

The M-DPS architecture was put up by Shah and Rajagopal [26] for decentralized patient data management in the medical field. M-DPS seeks to improve data accessibility, optimize storage, and lower gas costs in comparison to the current DPS

design. The evaluation's findings show notable gains in the efficiency of storage space and the decrease of gas fees, which might be advantageous to customers. A restricted grasp of system operations results from the study's lack of comprehensive information on user registration procedures, roles, permissions, and data sharing protocols. To solve IoT-driven healthcare demands, Azbeg et al. [27] present BlockMedCare, a secure healthcare solution that combines blockchain technology with IoT. The system, which focuses on remote patient monitoring, uses an off-chain IPFS database for scalability, smart contracts for access management, and a re-encryption proxy and blockchain for security.

Kaur et al. [77] offer a permissioned blockchain-based system to solve the challenges of maintaining EHRs dispersed among several healthcare providers. This system makes use of the Identity Based Proxy Re-Encryption (IB-PRE) algorithm for safe data sharing, IPFS for secure off-chain storage of encrypted data, and Hyperledger Fabric for network implementation. Hyperledger Caliper is used for performance testing to evaluate the efficiency of the framework. Evaluations in comparison to current systems demonstrate how well it addresses EHR security and privacy issues. Despite providing a more thorough performance analysis than its peers, the research lacks a diversity of workload viewpoints. The experimental design is similar to that of comparative studies, with the only emphasis being on latency and throughput measurements.

A Hyperledger Fabric-based patient-centric healthcare data management system is presented by Sonkamble et al. [79]. By integrating IPFS and BC, our decentralized architecture prioritizes patient control while guaranteeing the safe storage of EHR data. Smart contracts that employ Secure Password Authentication-Based Key Exchange (SPAKE) provide user control. The system's efficacy in patient-centric access management is demonstrated by the experimental design, which also evaluates performance using important metrics. The study compares performance to previous research and assesses the access control system. But a more thorough examination, especially with regard to blockchain-based performance, may have offered a more accurate comparison with current options. Furthermore, it's yet unknown how users will connect with or interact with these Hyperledger-based platforms.

By delivering powerful computing power near to users, the combination of BC and Mobile Edge Computing (MEC) has improved healthcare efficiency in recent years. Although a number of blockchain-based MEC strategies have been put out to solve

EHR security issues, many of them still encounter difficulties in their effective execution and fall short in addressing issues with automation and scalability. To improve on current schemes, Datta and Namasudra [83] provide a unique blockchain-based EMR sharing architecture that makes use of MEC and consumer electronics. Additional security layers are incorporated into this architecture by using methods like AES. IPFS storage is used to store encrypted EMRs and diagnostic reports, and the blockchain network is used to upload the associated hashes. Different functionalities are managed by smart contracts, and speedier transactions are ensured by the Proof of Authority (PoA) consensus process. Unfortunately, the study does not address scenarios in which malevolent actors might transmit irrelevant data and does not provide specifics on the data input process, which leaves a lack of procedures for data verification. Furthermore, comprehensive details on system permissions are not supplied.

Using the primary parameters shown in Table 4.1, research on EHR sharing from 2016 to the present is methodically compiled and contrasted. Every platform has unique advantages and disadvantages. According to our findings, these systems have successfully solved a number of important EHR sharing concerns. Still, they frequently focus on specific problems instead of covering the whole range of necessary features. The whole data sharing process, including access control (data interoperability), permissions, data verification, data recording, data input (privacy and security), and user registration (roles), must be carefully examined in order to achieve complete efficacy. The suggested system should then be put into action, and its performance should be thoroughly examined after that. Despite the fact that the studies are blockchain-based, our findings show that several of the studies' performance evaluations lack blockchain-specific analysis. While some studies compare their systems with current ones based on throughput and latency measurements, others analyze performance inside their own systems. A novel strategy would, however, compare the suggested system against previous research based on several configurations rather than simple measures. This comparative technique facilitates the discovery of several aspects impacting system performance as well as the development of innovative methods from different perspectives.

Table 4.1 Comparison of features between the proposed work and existing related works.

Research	Year	ACM	P	DV	SP	Roles	DS	S	A	PA	DP
[13]	2012	√	√	X	√	√	√	N/A	√	X	X
[11]	2016	√	√	X	√	√	√	X	X	X	X
[12]	2016	√	X	X	X	X	√	X	X	X	N/A
[15]	2017	√	X	X	X	X	√	√	X	X	X
[16]	2018	√	√	X	√	√	√	X	X	X	X
[14]	2019	X	√	X	X	X	√	X	X	X	X
[17]	2019	X	√	X	X	X	√	X	X	√	X
[19]	2019	√	√	X	X	√	√	√	X	√	X
[20]	2020	√	X	X	X	√	√	√	X	X	X
[1]	2020	√	√	X	X	√	√	√	√	√	X
[21]	2021	N/A	X	X	X	X	√	X	X	X	√
[22]	2021	√	√	X	√	√	X	X	X	X	√
[23]	2021	√	√	X	X	√	X	√	√	X	√
[24]	2021	√	√	X	X	X	√	√	X	X	√
[25]	2021	√	√	X	X	√	√	√	√	√	X
[26]	2022	√	X	√	√	X	X	√	√	√	X
[27]	2022	√	√	X	X	√	√	√	√	√	X
[28]	2022	√	√	X	X	√	√	√	√	√	X
[29]	2022	X	X	X	X	X	X	√	X	X	√
[77]	2022	√	√	X	√	√	X	√	√	√	X
[78]	2023	√	X	√	√	√	X	√	√	X	X
[79]	2023	√	√	X	√	√	X	√	X	√	X
[80]	2023	√	X	X	X	√	√	X	X	X	X
[81]	2023	√	√	X	√	√	√	√	√	X	X
[82]	2023	√	√	X	√	√	√	√	√	X	X
[83]	2024	√	X	X	√	√	√	√	√	√	X
AguHyper	2024	√	√	√	√	√	√	√	√	√	√
-ACM: Access Control Mechanism			-P: Permissions					-DV: Data Verification			
-Security&Privacy (SP)			-DS: Data Sharing					-S: Scability			
-A: Availability			-PA : Performance Analysis based on BC					-DP: Disease Prediction			

The goal of this thesis is to create and demonstrate a performance study of an optimal blockchain-based EHR sharing system that can be integrated with illness prediction mechanisms and thoroughly tackles all EHR sharing issues using SysML. To the best of our knowledge, no study has using SysML to examine every step of the previously stated blockchain-based EHR sharing platform. In this way, the components of the relevant platform were identified, and their interrelationships were analyzed, to start. In addition, the prerequisites for the secure and effective operation of the system previously mentioned have been determined, and the connection between these prerequisites and the platform's components has been examined. Furthermore, an example scenario is used to demonstrate how these criteria might be satisfied and system performance analysis.

In this thesis, IPFS and Hyperledger Fabric were combined to create the AguHyper framework [31]. Using a permissioned BC ensures secure interactions, whereas IPFS employs decentralized databases to overcome the problems with centralized storage. By storing hash values in the blockchain and encrypted information in IPFS, health records may be made immutable, rendering the system immune to manipulation. We do a full analysis of the system architecture and AguHyper implementation configurations, which include the use of CouchDB and the Raft consensus algorithm. The experimental configuration included the implementation of the Raft consensus mechanism and the CouchDB database [32]. Using datasets of varying sizes, the system's performance was carefully evaluated, with special emphasis given to variables like uploading-downloading time, average transaction latency, and transaction throughput. The study concluded with a comparative analysis of the system's performance against relevant literature research utilizing different consensus techniques and database formats.

Furthermore, this thesis introduced CSA-DE-LR, a unique hybrid technique. It combines LR with DE and CSA. This integration is intended to optimize logistic regression weights for accurate CVD classification. The method employs three optimization approaches based on the F1 score, the MCC, and the MAE. Extensive experiments conducted on industry standard datasets, Cleveland and Statlog, demonstrate that CSA-DE-LR outperforms the state-of-the-art machine learning methods. Interestingly, in comparison to previous research in this sector, it also demonstrates superior effectiveness. Lastly, a detailed explanation of the integration

phases between the novel illness prediction mechanism CSA-DE-LR and the planned blockchain-based AguHyper was provided.

Chapter 5

Methods

5.1 SysML

A system may be thoroughly evaluated thanks to SysML, a design language that gives users access to a variety of diagram structures for system analysis and design at different stages of development [84]. These blocks or modules are connected to one another by directional and related arrows, and the strength of the relationship between them varies based on the type and direction of the arrows. In this language, the elements to be used in the structures within the system are called blocks or modules. With the numerous diagrams it offers, SysML enables one to view the system's components from a variety of angles [85].

A block definition diagram is a diagram that shows the hierarchy between the system's subcomponents and the system itself. Internal block diagrams are described as those that depict the sub-components of the system and the contents of these sub-components. Use-case diagrams are schematic representations of the functions and activities that the system's end user will be able to perform [86]. Activity diagrams are those that show how the system and its subcomponents behave and perform. The requirement diagram is the diagram used to ascertain the requirements of the system to be constructed and to select which modules or components would satisfy these requirements [87]. The aforementioned diagram kinds are commonly used to design a system within a general framework, while there are additional sorts of diagrams that are utilized as well. Thanks to SysML, it is now possible to anticipate errors and risks that may arise during the system development life cycle and to ensure that the designed systems and their internal structure can be easily understood even by stakeholders from different disciplines. To create the system components, their relationships, system requirements, the relationships between system components, and a case study for this research, SysML is recommended for these reasons.

5.2 Hyperledger Fabric

Choosing the best blockchain platform to use in the ideation and creation of a project focused on blockchain technology is an important task. There are two primary types of BCs: public and private. Permissionless access and total transparency are features of public blockchains, which enable any member of the network to view the transaction ledger and carry out actions without limitations. On the other hand, private blockchain technology is designed to meet the needs of applications where security and anonymity are of utmost importance [31]. It is simple to set up a closed network and construct several channels that limit usage to certain individuals by changing the network's access rights. In this way, confidential information may be transferred throughout the network without notice and unregistered users are prevented from seeing the ledger [88]. Our study uses Hyperledger Fabric because of the sensitive nature of EHRs and the need for limited access. By selecting this option, reliance on a central authority is removed and secure exchange of healthcare information among pre-specified parties is ensured.

5.3 Consensus Mechanism

A fundamental feature and layer of blockchain is the consensus process that controls transactions. For the purpose of authenticating and modifying transactions in the ledger according to the order in which they occur, this method depends on the smart contracts layer. The consensus mechanism controls the sequence of transactions in the ledger and determines which transactions are rejected if they are not optimum. Three different consensus algorithm implementations are included in Hyperledger Fabric [32, 89]: i) SOLO Ordering Service: This deployable nonproduction ordering service eliminates the requirement for consensus by having a single central authority and a single procedure serving all clients. Although it works well for testing and development, deployment is not advised. ii) An ordering service based on Kafka: This service offers crash-fault tolerance (CFT) and is based on Kafka's publish-subscribe architecture with multiple Kafka brokers and corresponding Zookeeper ensembles. Because it does not have Byzantine fault tolerance, it is defenseless against hostile nodes on the network even if it stores data on other brokers in case of a failure. iii) Raft: Based on the Raft protocol in etcd, Raft is a CFT ordering service. The Raft protocol uses a "leader and

follower" architecture in which each channel has a leader node that is chosen, and its choices are replicated by the followers. Diverse enterprises can contribute nodes to a distributed ordering server using raft ordering services, which are expected to be easier to set up and administer than Kafka-based ordering systems. The Raft mechanism has been selected for our investigation based on the characteristics of these three consensus methods. This choice is consistent with our system needs, and we plan to compare and analyze Raft's performance to alternative consensus algorithms used in other systems.

5.4 State Database

CouchDB and LevelDB are the two peer database formats that Hyperledger Fabric supports. As a key-value store, LevelDB keeps chaincode data in an easy-to-read format that makes it possible to do key, key range, and composite key searches. Conversely, CouchDB makes use of a datastore that is structured in JSON, which offers more flexibility since it permits information to be mapped across various database documents [90]. CouchDB has been carefully selected as the on-chain database for this investigation. Its use improves system compliance in addition to the security and data protection components of the system. According to the needs and goals of the study, the JSON format of CouchDB enables a more flexible and dynamic representation of data inside the blockchain.

5.5 Hyperledger Composer

A collection of cooperative tools called Hyperledger Composer was created specifically for the modeling and building of blockchain business networks. Its goal is to simplify and accelerate the process by which developers and business owners may create blockchain apps and smart contracts. The goal of Hyperledger Composer's design was to make the difficulties involved in working directly with Hyperledger Fabric simpler. It provides a more sophisticated interface that makes it easier for developers to describe their business networks, participants, assets, and transactions. To help and improve the development workflow, this includes offering a modeling language, an API, and a collection of command-line tools [91]. In our investigation, Hyperledger Composer is utilized as a result. Composer is used in this study's setting to produce a business network definition. ACL (.acl) files for access control rules and permissions,

script files (.js) holding smart contracts, model files (.cto) that define assets, and query (.qry) files for creating database queries inside the framework are all included in this description. Then, to make it easier to install the framework's business network onto a distributed ledger, the business network specification is packaged as a. bna file.

5.6 Chaincode

Acting as self-contained chain codes, smart contracts contain the rules governing specific network transactions. These JavaScript-scripted smart contracts in Hyperledger Composer operate on the Hyperledger Fabric blockchain network. The Hyperledger Composer chaincode represents the application logic in charge of managing and defining transactions, supervising asset management, and implementing access control policies in a business network [92]. The AguHyper project has made a conscious choice to employ smart contracts, taking advantage of their built-in advantages, which include the automated fulfillment of contractual duties and the efficient management of relationships and access permissions to data.

5.7 Interplanetary File System

A decentralized, peer-to-peer file system called IPFS has the potential to completely replace HTTP and transform the way the internet is now organized. Using IPFS, one may access a data structure or download a file from the internet by accessing it via network peers by use of the file's identifier, or 'cryptographic hash'—an attribute referred to as IPFS content addressing [93]. IPFS distributes the encrypted data among several nodes to provide secure storage if the data exceeds a certain size threshold. For the purposes of this research, a wide range of medical records and the associated hash kept in the CouchDB database are maintained in IPFS, which acts as an off-chain database [94].

Chapter 6

Design of an Ideal EHR Sharing Platform based on SysML and Blockchain

Every process, subsystem, and component is viewed as a distinct module in our paradigm, which is equivalent to a "block" in SysML settings. The six parts that make up our strategy are System Users, Registration, Data Recording, Asset Layers, Data Entry, and Data Sharing, as seen in Figure 6.1. Furthermore, Figure 6.1's arrows show how the modules communicate with one another. For instance, to submit a registration request, each actor in the system must connect with the registration module. The purpose of each module and how they relate to one another will be covered in detail in the parts that follow.

6.1 System Users

Particularly in the healthcare industry, life-critical systems are involved. Users inside an EHR system have the potential to make decisions that compromise patient safety or worsen the quality of care. As a result, users of the system have to be licensed healthcare professionals with verified IDs. In this sense, in order to be enrolled in the system, users must get a certificate from an appropriate medical institution proving their eligibility for the responsibilities they claim. The seven roles that will be included in the platform are defined in the System Users module and comprise patients, doctors, researchers, nurses, labs, and hospitals.

Patients: are the primary users of the system; they submit data using the Data Entry module and change permissions for data access using the Data Sharing module.

Doctors: among the users who utilize the Data Entry module to enter data into the system. Through the Data Sharing module, they may also use the system to request data and validate patient data.

Researchers: are users who utilize the Data Sharing module to submit requests for data and validate patient data.

Nurses and Laboratories: The Data Entry module is used by laboratories and nurses to enter data into the system.

Hospitals: are system users who sign up new users, gather and store platform-shared transactions, and log them onto the blockchain.

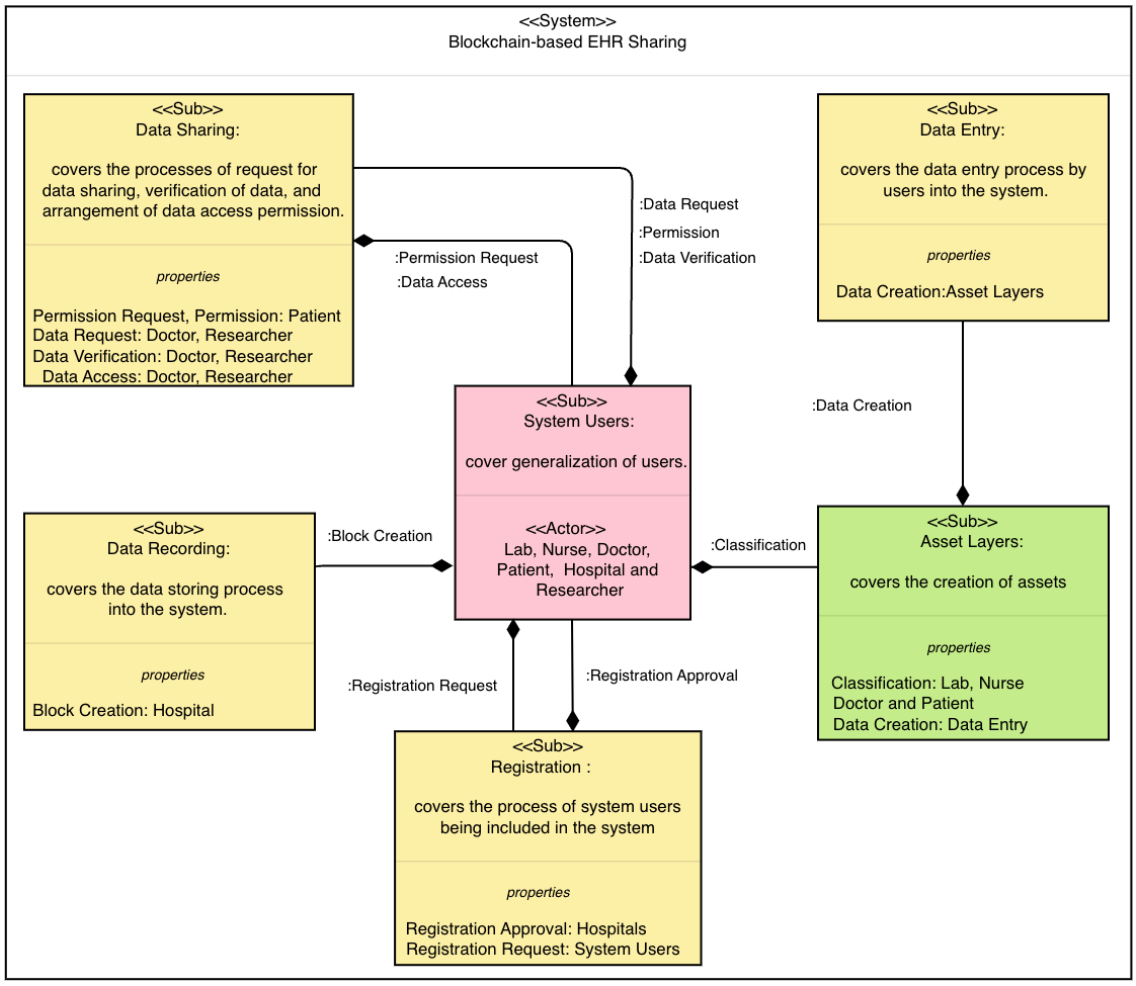

Figure 6.1 General structure of blockchain-based EHR sharing

6.2 Registration

The user registration procedure is implemented by our system as shown in Figure 6.2. A certificate proving qualification for the claimed profession and a registration

request must be sent via the Registration Module to one of the registered hospitals by any new user (other than hospitals) who wants to join the system. The registration is authorized and the applicant is notified through the Registration Module if the certificate is valid.

We should be careful while creating the registration process for a new hospital since hospitals have more power over the decision-making process inside the system. In a similar vein, a hospital that wants to join the system sends a registration request to one of the hospitals that are already enrolled using the Registration Module. In contrast to other users' registration processes, the hospital uses the Registration Module to start a vote scheme that solicits input from other registered hospitals about the applicant's admittance. The appropriate hospital authorizes the registration and notifies the applicant through the Registration Module if the majority of voting hospitals approve it.

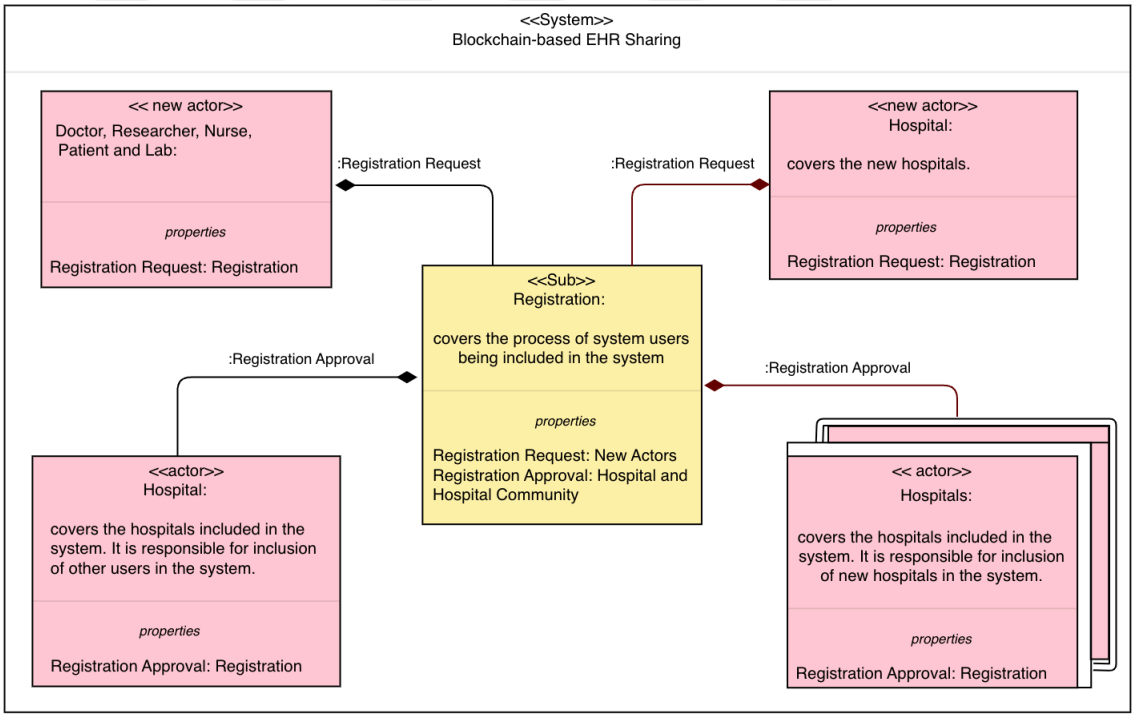

Figure 6.2 Shows the registration module

6.3 Data Entry & Asset Layers

To take part in data sharing, users must first join the network and then submit data into the system. This procedure is the responsibility of the data input module. Nurses, doctors, patients, and labs may all input data into the system. The two steps of data storage in this module are generation and categorization. The act of entering data by

users is referred to as data production, while the process of classifying that data is referred to as classification. Specifically, the users who input the assets are identified by the categorization. This procedure produces four distinct assets: a patient asset, a nurse asset, a doctor asset, and a laboratory asset.

6.4 Data Recording

Hospitals store user data in the IPFS network after the assets are produced. For this procedure, the data recording module is in charge. Similar to how they handle registration module transactions, hospitals handle these data recording transactions in the blockchain. Here, we would want to underline that the owners of the data keep the encrypted version of the data in the IPFS rather than storing it openly. By doing this, unwanted access to the data is avoided. Block formation is a result of data input. With permission from the authorities registered on the platform, the designated authorities (hospitals) will periodically consolidate transactions shared on the platform via the Consensus Module and write them into new blocks that will be added to the chain. The subsequent sections will provide specifics on the consensus and block formation procedures.

6.5 Data Sharing

The processes of data request, data verification, and data access authorization setup are all included in the data sharing module. These mechanisms are detailed in depth in Figure 6.3. Patients, researchers, and doctors all participate in the data sharing process. Firstly, patients control their rights for data access, and these permissions facilitate data interchange. Ensuring the accuracy of the data is crucial, in addition to restricting the rights related to its sharing. The manner by which the system does this is through data verification. Doctors and related researchers validate patient data. The only people who may seek access to data are researchers and doctors. Researchers and physicians send data requests to the system in order to receive data. The data is automatically sent to the researcher or clinician if the system determines that they are authorized to view it and meet the necessary requirements.

6.6 Requirement Analysis

The prerequisites for accurate and trouble-free platform maintenance will be covered in this section. The requirement diagram in Figure 6.4 illustrates these necessary conditions.

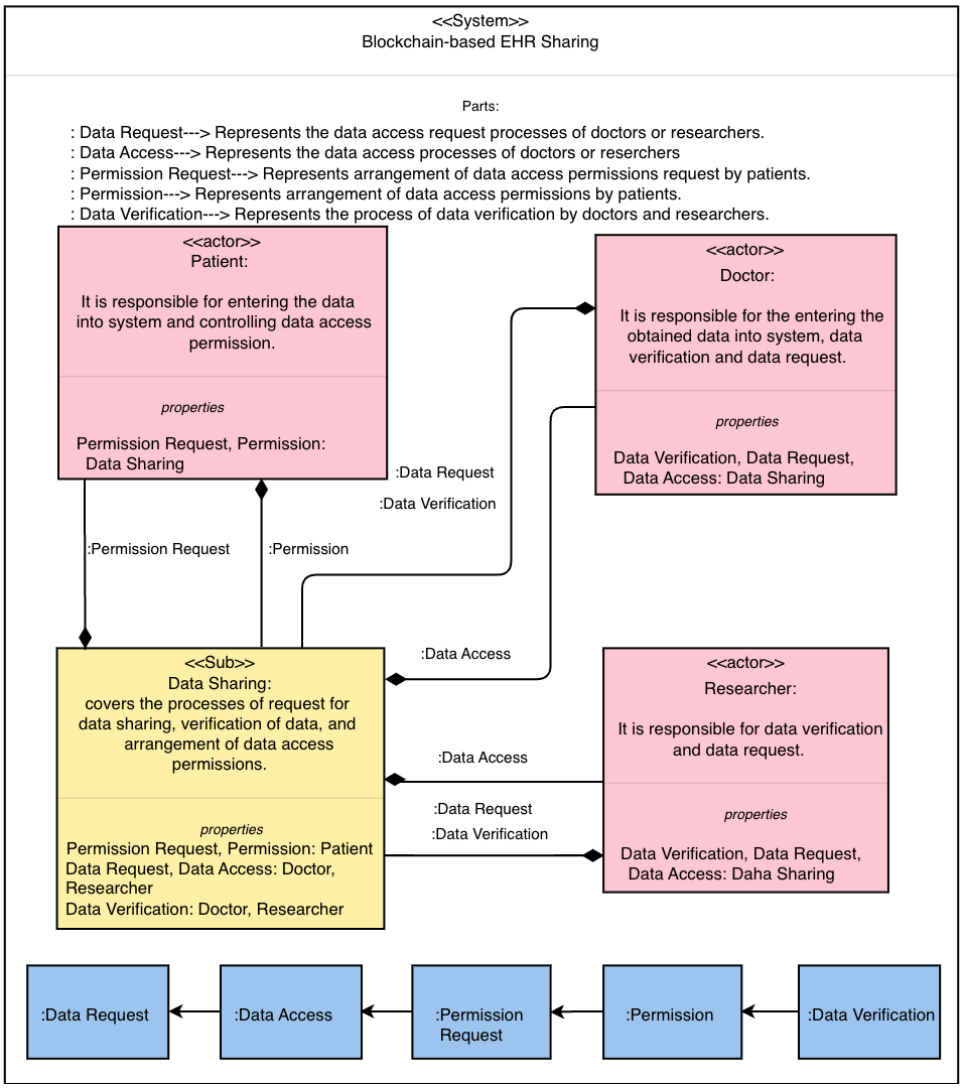

Figure 6.3 The data sharing module

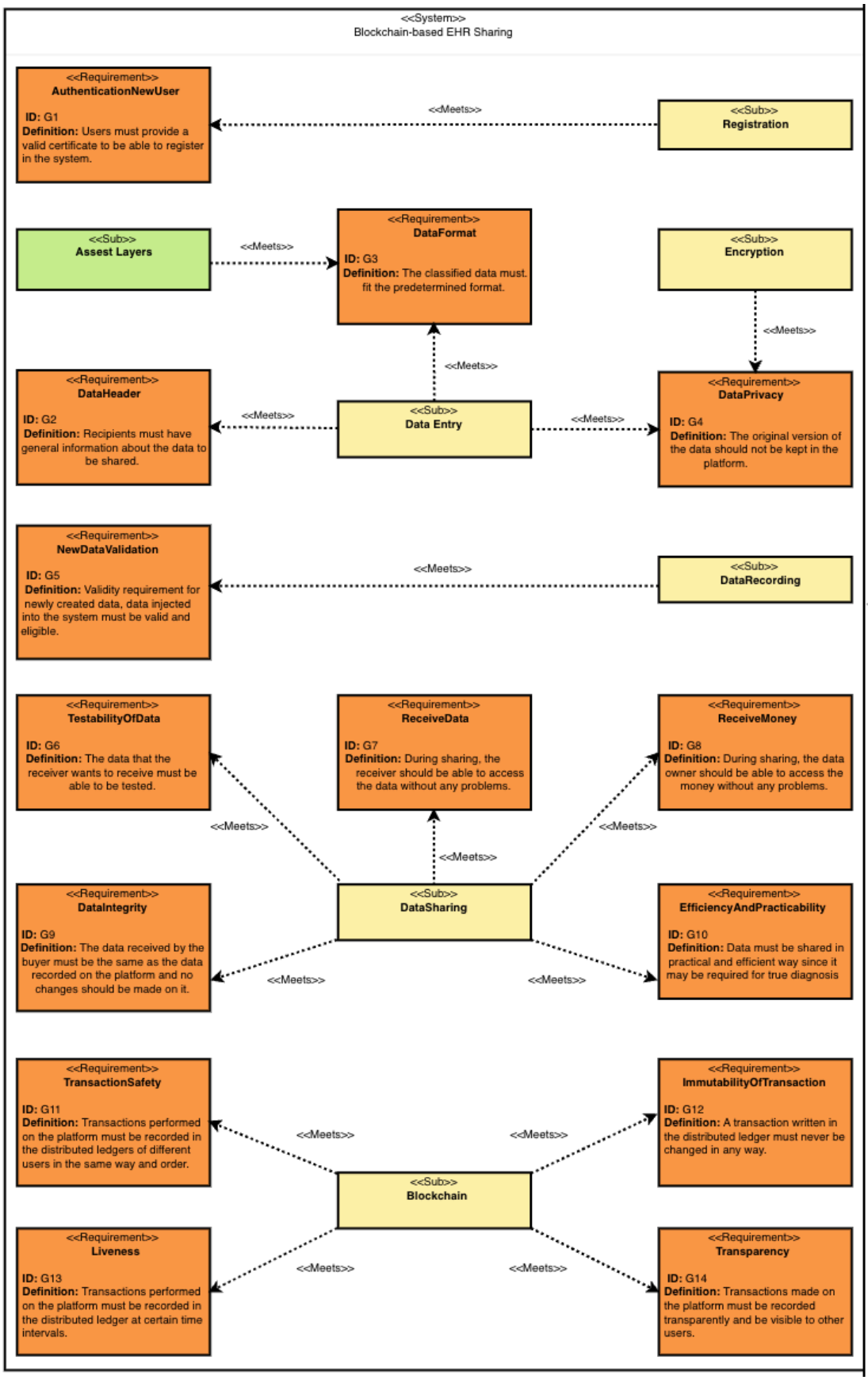

Figure 6.4 Requirement diagram of the system.

Standard directional dashed lines in the diagram show the link between the requirements and the pertinent modules. An example scenario will be used to illustrate how these requirements may be accomplished in the next section. In order to register with the system, new users need to present a valid certificate [95]. This is because it is necessary to verify that the roles they are claiming to be legitimate. A digital signature system can be used to achieve this goal [96]. The ID:G1 in Figure 6.4 serves as an illustration of this need. Utilizing the "Registration" Module satisfies the prerequisite. The following procedures will be followed, in turn, by the new user here:

1. Choose the actor role for which you wish to receive the Digital Signature Certificate (DSC) and submit a registration request to the Certifying Authority.
2. Complete the required fields.
3. Identity verification from the certifying body.
4. If all the details are accurate
5. Obtain clearance for both registration and a DSC.

In order to facilitate data scanning on the platform and enable recipients to assess the data they wish to see, a data-related header should be supplied on the platform when constructing the data record [97]. The ID: G2 is used in Figure 6.4 to demonstrate this need. This need is satisfied by the "Data Entry" Module, which demands that the data-header be entered during the data input procedure. The system will check each data point to make sure its format matches the data type before categorizing it. The ID: G3 is used in Figure 6.4 to demonstrate this need. To comply with the standards, the "Data Entry" and "Asset Layers" modules are utilized.

1. Depending on the kind of data, the system determines several forms for data entry.
2. Choose a data type.
3. Complete the required fields.
4. Data format proof.

There is one thing to keep in mind. The Data Entry Module oversees Steps 1 through 3 while the Asset Layers Module verifies Step 4.

Because the system is visible, users may easily access data and data privacy is compromised if the shared data is maintained in its original form within the platform. As a result, the information needs to be kept private inside the system. With the ID: G4, this need is demonstrated in Figure 6.4. To comply with the standards, the "Data Entry" and "Encryption" Modules are utilized. In this case, data can be stored encrypted on the

blockchain; however, this will be a scalability issue since it would significantly expand the blockchain's size and, therefore, the volume of data it must process. Alternatively, IPFS allows data to be stored encrypted [98].

In this instance, maintaining the data's proof information within the system is essential to guaranteeing safe commerce and verification. All of this may be summed up in the following order:

1. The platform does not retain original data. Rather, the data's address and any associated basic information ought to be retained on the platform. This module "Data Entry" satisfies this demand.
2. The data address shouldn't be preserved on the platform in its original format. Data addresses have to be maintained on the platform as passwords inside the data's secrecy. We have the "Encryption" module to meet this purpose.

The newly input data into the system has to be eligible and legitimate. The ID:G5 in Figure 6.4 serves as an illustration of this need. Utilizing the "Data Recording" Module satisfies the prerequisite. Because of this, the data must first be validated by certain system actors based on its level of criticality before it can be recorded in blocks. The data is processed into chunks after approval. The sample usage scenario will go into depth on how to fulfill this need.

Malicious actors could submit meaningless material to the platform with the intention of tricking recipients. Receivers should thus have the option to test data before sharing it with the system. Furthermore, throughout the sharing process, a malicious actor may alter data posted to the platform, which would be detrimental to the recipient. Thus, at this stage, data integrity needs to be guaranteed. Stated differently, it needs to be guaranteed that the data a recipient wants to see is identical to the data that was initially posted to the platform and that no modifications have been done. Furthermore, it should be guaranteed that the owner may consistently get all payments after providing all data to the recipient, even though the recipient should be able to obtain all data without any issues following payment of the necessary cost. The IDs G6, G9, G7, and G8 are used in Figure 6.4 to show these criteria. To comply, the "Data Recording" Module is utilized. In order to compare the hash values of recorded data with sharing phase data, integrity issues require that, in addition to the address of the data being saved on a block, the relevant data's hash value be preserved as well. Hash values are then contrasted. The hash value needs to match the primary hash value that was recorded before sharing [99]. The permissions of the data and the state in which the data

will be shared may be readily controlled by the system through the use of smart contracts.

Since accurate diagnosis and prompt clinical treatment may depend on data sharing, it is imperative that this be done in a practical and effective manner. The ID: G10 is used in Figure 6.4 to demonstrate this need. Utilizing the "Data Sharing" Module satisfies the need. The platform must maintain the confidentiality of the data, as we said while elucidating the data privacy criterion. In this scenario, the data can remain encrypted within the blockchain; however, this will reveal the scalability problem as it would greatly expand the blockchain's size and require a greater volume of data to function [100]. Alternatively, data can be transmitted using a hashed version of the data address after being encrypted in IPFS. In this way, data sharing is efficient and practicability is realized.

A shared database system dispersed over a specific network or collection of networks is called a distributed ledger. Each user on the network is able to have a local copy of this ledger, and using consensus methods, updates to the ledger are mirrored in all of these copies on a regular basis. Two fundamental needs must be satisfied by a strong distributed ledger: liveness and safety. The IDs G11 and G13 are used in Figure 6.4 to show these requirements. The first condition says that network users will mutually agree on transactions to be put to the ledger; the second says that a successfully performed transaction will eventually be accepted and posted to the ledger by network users. It is assumed in this study that blockchain technology is used to construct the distributed ledger system. These prerequisites are satisfied by using the "Blockchain" Module.

A malevolent user has the ability to alter a record within the distributed ledger to benefit themselves. For instance, it has the ability to alter data sharing records to appear as though they have never been completed. In a similar vein, by tampering with the entries in the distributed ledger, a sharing record that has never been produced might appear to have been made. An unchangeable storage of a transaction published to the distributed ledger is necessary to increase the system's resilience to such situations. This will stop malevolent users from altering system records in order to benefit from them. Similar to this, a malevolent user may lead to an incomplete transaction record by cooperating with the authorities, which might harm the reputation of the vendor or buyer. Because of this, all platform transactions must to be transparently documented [101]. Other users will be able to examine the recorded transactions in this way,

boosting platform confidence. The IDs G12 and G14 in Figure 6.4 serve as illustrations of these criteria, and the "Blockchain" Module is utilized to satisfy them.

6.7 A Case Study

This section will utilize the patient's position to demonstrate how the previously described requirements may be met, as seen in Figure 6.5. A patient first makes a request to be added to the database.

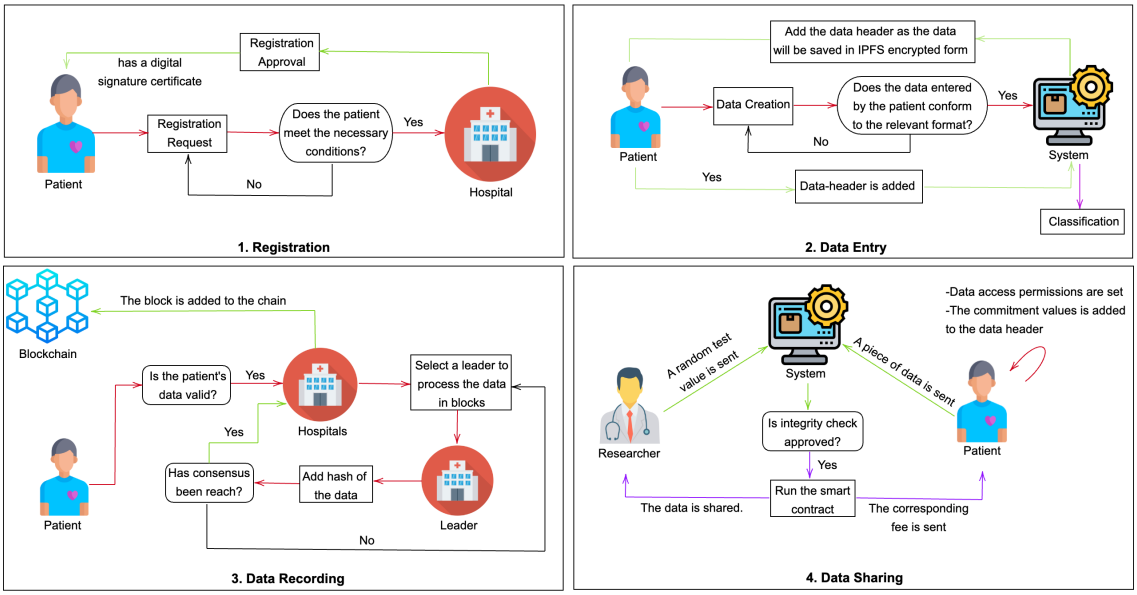

Figure 6.5 Demonstration of how requirements can be met by the proposed system using the patient role.

A patient first makes a request to be added to the database. In accordance with this request, the approved hospital confirms it provided the patient satisfies the requirements. The patient can now utilize a digital signature certificate in the system. The patient's second want is to input information into the system. For every data entry, the system has predetermined a specific format. Patients must submit data into the system in forms appropriate for the data type they are entering. Users of the system can immediately access shared data if it is maintained in its original format within the platform, jeopardizing data protection. As a result, the information needs to be kept private inside the system. In this case, data on the blockchain can be encrypted, although doing so will cause scalability issues. Rather, IPFS encrypts data and shares its address with recipients. This solves the issues of both scalability and availability.

Recipients must get comprehensive details on the information. Patients must input data header information during the data entry procedure in order to do this. Following the patient's entry of data header information in accordance with these boards, the system classifies these assets.

The data has to be saved in the system after the data entering procedures are finished. Blocks are created by these procedures. The process of capturing data involves verifying its legitimacy and integrity. Once these are established, the data is processed into blocks. Concurrently, during the data recording phase, decisions are made on who will handle the data in the blocks and what type of consensus procedure will be employed in this regard. The detailed process of the complete data recording module is depicted in Figure 6.6. Patients' data that they submit into the system has to be accurate. This requires the approval of users selected from among hospitals. The process of processing blocks cannot begin without passing the data approval stage.

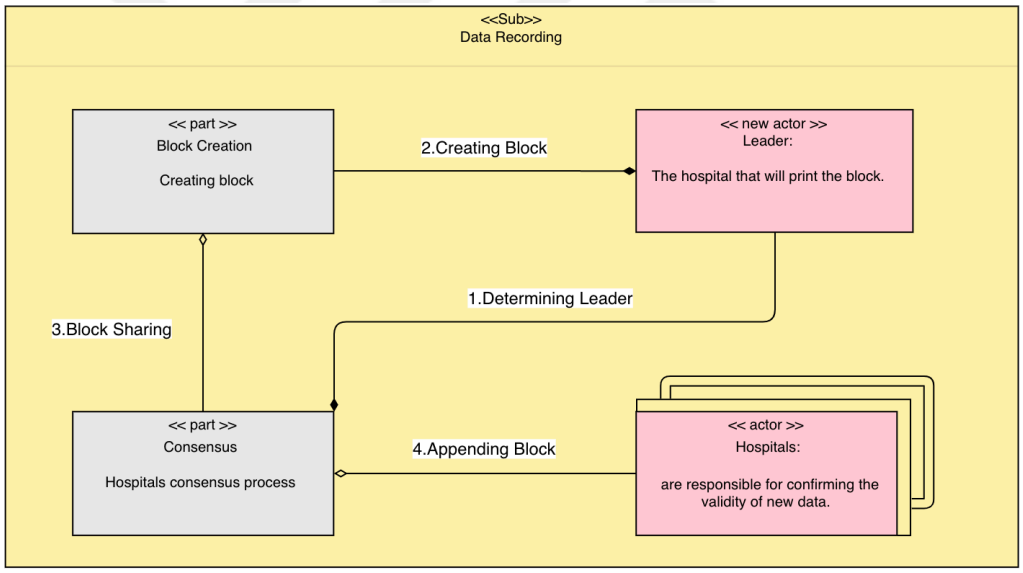

Figure 6.6 Data Recording Process

The hash of the data is written to the blocks during the data writing procedure to ensure that there hasn't been any degradation in data integrity throughout the data sharing phase. Blockchain technology leaders that print blocks are rewarded. Additionally, deciding which leader will print a block is required. Figure 6.6 part 1 illustrates this procedure, whereas Figure 6.6 part 2 demonstrates putting the data into blocks. The data must be submitted to consensus by certain nodes before it is blocked,

as shown in Figure 6.6 parts 3 and 4. Consequently, the transaction must be finished. Blocks representing the patient's data are created during the data recording phase.

Researchers and medical professionals can get data summary information on the chain. Patients and physicians share data in different ways than patients and researchers do. The data is intended for use by researchers in their own analyses. As a result, it could be easier for researchers to get data and profitable for patients to share their data. Figure 6.5, Part 4, provides an illustration of this circumstance. Researchers use the data sharing module to get in touch with the appropriate patient when they wish to review one of these data. The technology can allow the researcher to test the data before sharing.

We refer to this need as testability of data, and it may be satisfied if the sender and recipient share a modest amount of data. It is now required to demonstrate that this shared element is a part of the data. To ensure that the data is supplied with integrity, the sender must demonstrate to the recipient that the data is exactly as it was originally recorded and has not been altered. By integrating the Proof of Space module into the system, these two needs may be satisfied. In addition to the aforementioned, the Smart Contract module that may be implemented into the system can satisfy the needs linked to the data sharing module, which we refer to as the sender receiving the fee and the recipient receiving the data. It would be helpful to give a quick overview of the Proof of Space protocol and the idea of a smart contract before going on to discuss how the Proof of Space and Smart Contract modules operate.

A proof of space protocol is based on the memory space of the device being used and is utilized between two users, referred to as the prover and the confirmer. The protocol enables the prover to demonstrate that, within the allotted time frame, the region in his local memory designated for a particular purpose remains unaltered. There are two steps to the procedure. The prover creates a commitment value, which is essentially a summary of the N-bit data, in the first stage and stores it in its local memory. It then shares this value with the validator and uses it in the second step. To determine if the prover retains the pertinent data in its original format and communicates this value with the prover, the confirmatory creates a random test value in the second phase. At this point, the prover generates a proof value that matches the test value to demonstrate that it has preserved the data in its original format. It then provides the validator with this proof value. Ultimately, the approver decides whether to accept or reject the proof value by determining if the proof value produced is consistent

with the commitment value that the proofreader supplied at the first step. Dziembowski et al. were the first to provide the Proof of Space approach, which necessitates communication between the prover and the validator. It is a consensus algorithm that was put forth by [102]. The proof value generated in the second stage was rather little in comparison to the original file since the Merkle tree structure was employed in the suggested method's proof production process.

A smart contract is a computer code that runs on the blockchain and has a certain number of rules and actions that must be followed. It can only function when these rules are given [103]. A smart contract cannot be altered or terminated once it is in operation. It is mostly used to trade valuable assets, like money or data, and because it removes intermediaries, it lowers transaction costs. Furthermore, a smart contract's output is dependable as it is simple for other system users to verify and approve it [104]. The Proof of Space and Smart Contract modules incorporated into the system, as seen in Figure 6.7, can assist meet the requirements for the data sharing module that were established in the preceding section.

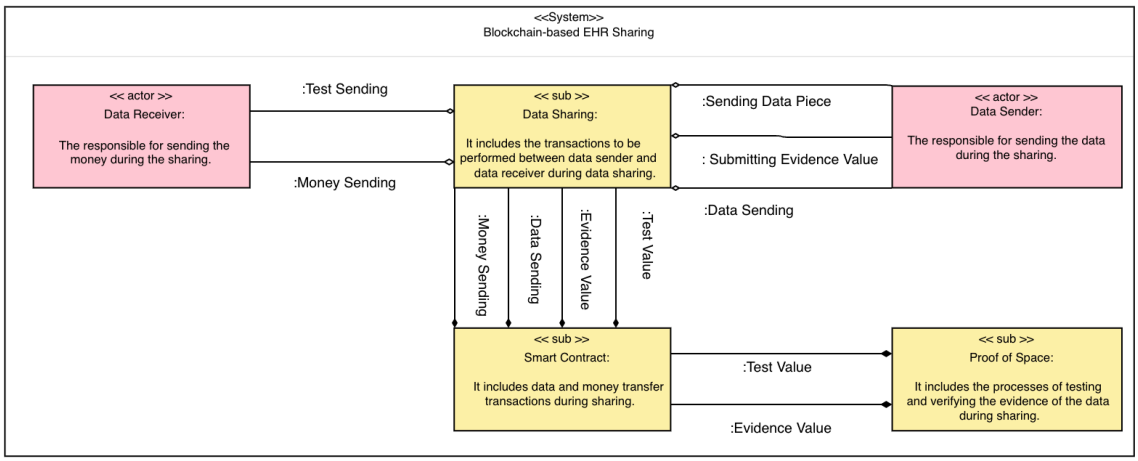

Figure 6.7 Data Sharing module for example scenario

The Dziembowski [102] proof of space protocol, which is applied by using the Merkle tree structure, may be utilized in the Proof of Space module. Senders must create a record about the data they wish to transmit and add the commitment value of the data to the data header for usage in the protocol in order for the system to function correctly with this protocol. At this stage, senders construct a Merkle tree over the n bits of data they currently have in hand. Then, they append to the data header the number n, which serves as the commit value, and the root of the related Merkle tree. The recipient selects a random number from the set {1,...,n} as a test value and sends it to the

appropriate sender and the Smart Contract module via the data sharing module when they wish to receive any of the data on the platform. This allows them to test the data beforehand and use it in the proof of space protocol to be applied for the proof of data integrity. The proof value is then constructed to be utilized in the proof of space protocol and sent to the Smart Contract module by the sender along with the data piece that corresponds to the test value via the Data Sharing module to the recipient. The recipient, who assesses the data piece, submits this request to the sender together with the payment needed for the sharing transaction to the Smart Contract module through the data sharing module. In response to this request, the sender provides the encrypted data to the Smart Contract module, or if the data is stored in IPFS, the appropriate secret key together with the recipient's identity and the applicable password. On the other side, the smart contract module uses the proof of space protocol on the test value from the sender, the proof value from the receiver, and the associated commitment value added to the chain earlier through the Proof of Space module to determine if the data integrity is given. The Smart Contract module will carry out the data-fee exchange if the proof of space protocol confirms the accuracy of the data.

The consensus module determines the leader of each platform transaction, who then writes it to the blocks. New transactions are added to the chain by getting authorized over the consensus protocol, which is likewise managed by the consensus module [105]. The leader who will form the next block is chosen ahead of the consensus module, as shown in Figure 6.6. The leading authority then builds a new block based on the transactions that were shared on the platform within the pertinent time frame. To create a strong chain, leaders further provide a hash value by applying a collision-resistant cryptographic hash function to the most recent block added to the chain. They then put this value at the start of the subsequent block. By using a collision-resistant cryptographic hash function to join one block to the next, an unbreakable chain of blocks is created. This makes it possible to keep the records that are processed in blocks in an unchangeable format. The leaders then distribute the blocks they make to other platform users. All users who have registered on the site will be able to view the transactions that are made there clearly thanks to its structure. After confirming the transactions in the block and the hash value entered at the beginning of the block, other authorities that receive the freshly formed block add the pertinent block to the chain. Verifying the accuracy of the block is not the responsibility of other users on the platform.

Chapter 7

Implementation of AguHyper: Results and Discussion

7.1 Layers of AguHyper

The Storage Layer, the Blockchain Layer, and the User Layer are the three distinct levels that make up the AguHyper framework as shown in Figure 7.1.

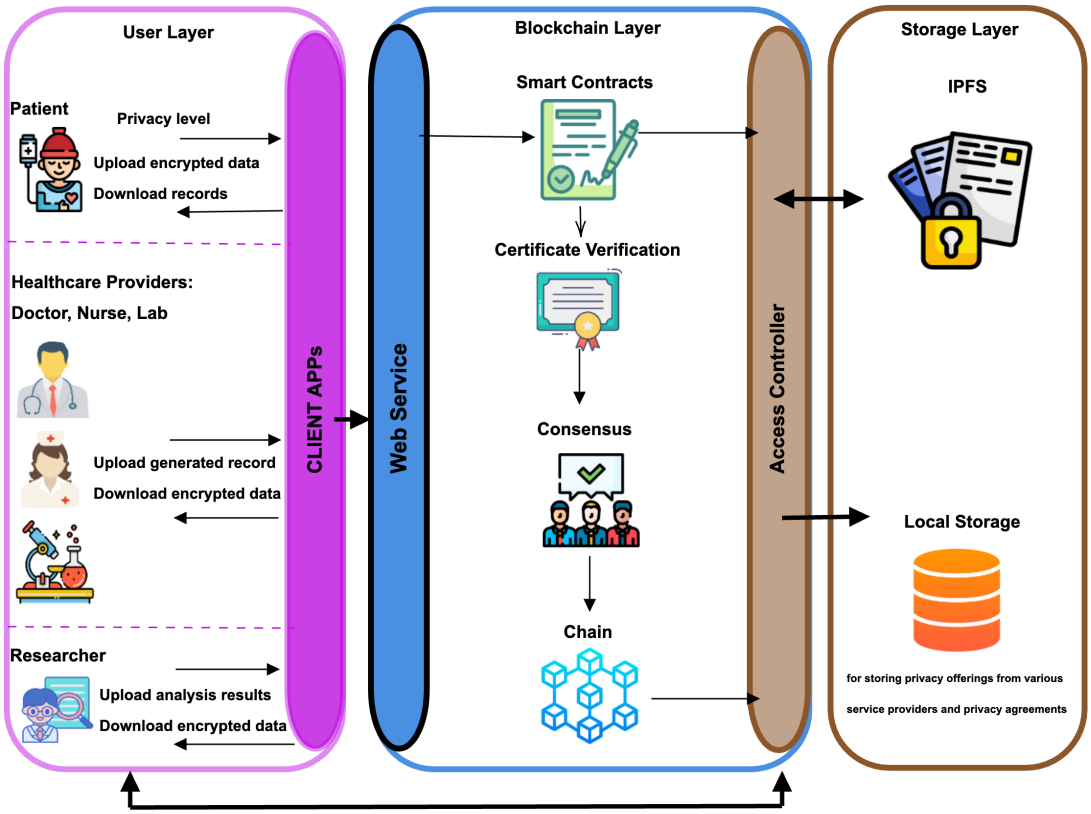

Figure 7.1. Architecture of AguHyper [115].

Potential users of the BC network are included in the User Layer. People utilize Client APP to submit the required data into the system before they may become system users. The Blockchain Layer receives these registration requests via the API. The Blockchain Layer generates a digital signature by having the certificate authority (CA) provide the individual a public-private key. Upon obtaining a digital signature, the user's digital identity and permissions based on their function within the system are stored by the MSP (AguHyper, the system organization). Following these phases, the individual assumes the position of a system user. Using the client APP, users may share, request, and input data to the Blockchain Layer based on their responsibilities and permissions. An off-chain distributed file system specifically designed to store users' encrypted data is included into the Storage Layer. These hashes serve as a means of referencing and organizing the material in a methodical manner. One of the authorized users submits a request to the Blokchain Layer. Using the IPFS Layer APIs and user data, the approved organization in the Blokchain Layer records the pertinent encrypted data in IPFS. Following this procedure, the pertinent data's hash is noted on the BC. Maintaining ownership information and metadata for files kept in the decentralized file storage is the responsibility of the Blockchain Layer. Additionally, it facilitates safe data transfer between companies by offering permission management services. The Blockchain Layer is the fundamental layer in communication between the User tier and the Storage Layer, even if communication between each tier is facilitated by APIs.

7.1.1 Storage Layer

We have decided to use the IPFS to store encrypted data blocks rather than BC for the storing of medical records. The fact that IPFS functions without a single point of failure and can effectively distribute large volumes of data without redundancy is notable [106]. EHRs are created by users and distributed over IPFS storage nodes for storage. Every file that is submitted to the IPFS system is given a distinct hash string, which makes it easier to retrieve it later. Data integrity is guaranteed by the IPFS system when it is connected with the BC network. The storage node sends the data's hash to the BC network after storing the data. Any unlawful adjustments may be easily detected thanks to this approach.

7.1.2 User Layer

Each user launches a decentralized application created especially to make communication with the distributed file system and the BC easier. A CA uses the user's private key, which also contains the user's public key, to establish public-private key pairs for each user in the system and to construct a digital signature for each user. The MSP saves the user's digital identity and permissions according to their position in the system when they get a digital signature. Concurrently, the MSP system keeps track of a folder that has a list of users' digital signatures. A transaction is signed using the client's private key at the time it is completed. This transaction is processed onto the blockchain by orderer nodes. The transaction undergoes verification with the client's public key according to the relevant consensus mechanism before being processed onto the blockchain.

Only patients, lab personnel, nurses, and physicians are permitted to enter data into our system. Researchers and physicians are permitted to seek data in the interim.

- **Patients:** Every patient node is in charge of overseeing one or more EHRs. They provide the IPFS storage node their encrypted data. These nodes show signs of being able to produce and distribute transactions. Patients have complete control over their EHR access rights.
- **Hospitals:** Nodes operate as system users, taking care of member registration, gathering and entering platform-shared transactions into the blockchain.
- **Doctors:** they are able to ask the system for data. Additionally, they demonstrate the ability to safely transfer encrypted EHRs to the assigned storage node.
- **Researchers:** they are those who make requests for data and then make the findings of their analysis available to others.
- **Nurses and Laboratories:** these users are capable of safely transferring encrypted EHRs to the assigned storage node.

7.1.3 Blockchain Layer

Based on a permissioned blockchain architecture, the suggested solution uses pre-specified nodes as miners. As a result of their reputation for dependability, these nodes are in charge of verifying transactions and generating new blocks for the network. In our case, reputable hospitals are the establishments entrusted with this responsibility. These trusted authorities carry out many functions, such as adding new data to the

decentralized file system, uploading relevant transactions to the Blockchain, and validating additional transactions, such as grants and permission requests, that are initiated by external users.

7.2 Smart Contracts

Three different kinds of smart contracts are included in the Blockchain layer: participantCreation contract, assetCreation contract, and dataSharing contract.

participantCreation Contract: All users register anonymously within the participantCreation contract in order to protect the system from malevolent users that try to add false data or exploit information. Public keys for each user are registered, along with the responsibilities that go along with them. The pseudocode explaining the phases involved in participant formation is included in Algorithm 1.

Algorithm 1: participantCreation Contract

Input: userPublicKey, userRole
Output: success of Registration

```
// The MSP register the user in the system with the necessary permissions and roles
after approving the digital signature by the CA.
if protectSystemFromMaliciousUsers() == True then
anonymouslyStoreUserDetails(userPublicKey, userRole);
return "SUCCESS";
else
return "USER CREATION ERROR";
end
```

assetCreation Contract: The assetCreation contract maintains a record list that outlines the association between users and their respective data. Each entry in this list includes the public key of the data owner and the hash of the encrypted data, referencing the raw data stored off-chain. To streamline this process, the data contract offers functional interfaces for the addition of data. Algorithm 2 presents the pseudocode, elucidating the steps of the asset creation process.

dataSharing Contract: A dataSharing agreement carefully records access rights, outlining the various rights that users possess over the data that is contained in the data-sharing agreement. Three tuples make up each access permission: the hash and ID of the

data, the public key of the person requesting the permission, and the public key of the person granting the permission. The pseudocode that explains the phases involved in the data sharing process is presented in Algorithm 3.

Algorithm 2: assetCreation Contract

Input: userPublicKey, encryptedDataHash
Output: success of Data Addition

```
1: //Allow data entry if the digital signature of the person who wants to upload data
   is matched with the digital signature registered in MSP.
2: if SystemUsersVerify() == True then
3: record = createRecord(userPublicKey, encryptedDataHash);
4: addRecordToUserList(record);
5: return "SUCCESS";
6: else
7: return "DATA ADDITION ERROR";
8: end
```

Algorithm 3: Sharing Contract

Input: userPublicKey, requesterPublicKey, EncryptedDataHash, dataID
Output: success of Permission Granting

```
1: //Share the relevant information with the requester, İf the integrity of the data is verified
   and the data owner accepts the request
2: if DataIntegrityVerify() == True && PermissionAcceptedbyUser()== True then
3: permission = createPermission(userPublicKey, requesterPublicKey, EncryptedDataHash,
   dataID);
4: addPermissionToDataSharing(permission);
5: return "SUCCESS";
6: else
7: return " PERMISSION GRANTING ERROR";
8: end
```

7.3 System Operation Details

7.3.1 Add Records

Users conduct two primary operations to add records to the system. These operations include i) uploading data to the IPFS and ii) sending metadata to the BC. During the data uploading process to IPFS, the data undergoes encryption, and the hash value is derived from the encrypted data. The upload procedure is finalized by saving the encrypted data. In the metadata sending process to BC, the transaction content is initially generated. This content encompasses pertinent information, including the

encrypted key. Subsequently, the transaction is authenticated through the user's key and transmitted. In the supplementary EHRs add-on process, the data entry procedures for healthcare providers, who exclusively input patient data, differ from those performed by the patients themselves. Notably, there is an absence of an encrypted key in the content of the patient transaction.

7.3.2 Data Sharing Request

Medical professionals or academics interested in certain data can submit a permission request over the Blockchain network once metadata about the BC becomes available. To achieve this, submit a transaction that causes the dataSharing contract to activate. Upon the transmission of a permission request to the dataSharing contract for accessing specific data, the data owner receives a notification and is afforded the option to either grant or deny the request. If authorization is granted, a transaction is created that has the following elements: the ID of the requested data, the requester's public key, and the key that will be used to decode the requested data once it has been encrypted using the requester's public key. After approving the permission, the user gets the data from an IPFS node in the area. The data that has been obtained is then decrypted.

7.3.3 Analysis Result Share

Data sharing was taken into account for two distinct users when creating AguHyper. Data sharing with researchers is the second, and sharing with the doctor is the first. Users that want data for study, such illness prediction, are called researchers. These users communicate the findings of their study, such as illness prediction, with the appropriate patient if their request for data is granted.

7.4 Security and Functional Analysis

As shown in Figure 7.2, the patient's role is used in this part to clarify the system's working mechanism. First, a patient uses the system to submit a registration request. Once the request has been carefully reviewed and all requirements have been completed, the approved hospital approves it. As a result, the patient receives a certificate with a digital signature to use in the system. The patient then attempts to enter information into the system. The patient must follow format requirements

pertaining to the data type while entering data since the system has specified forms for each kind of entry.

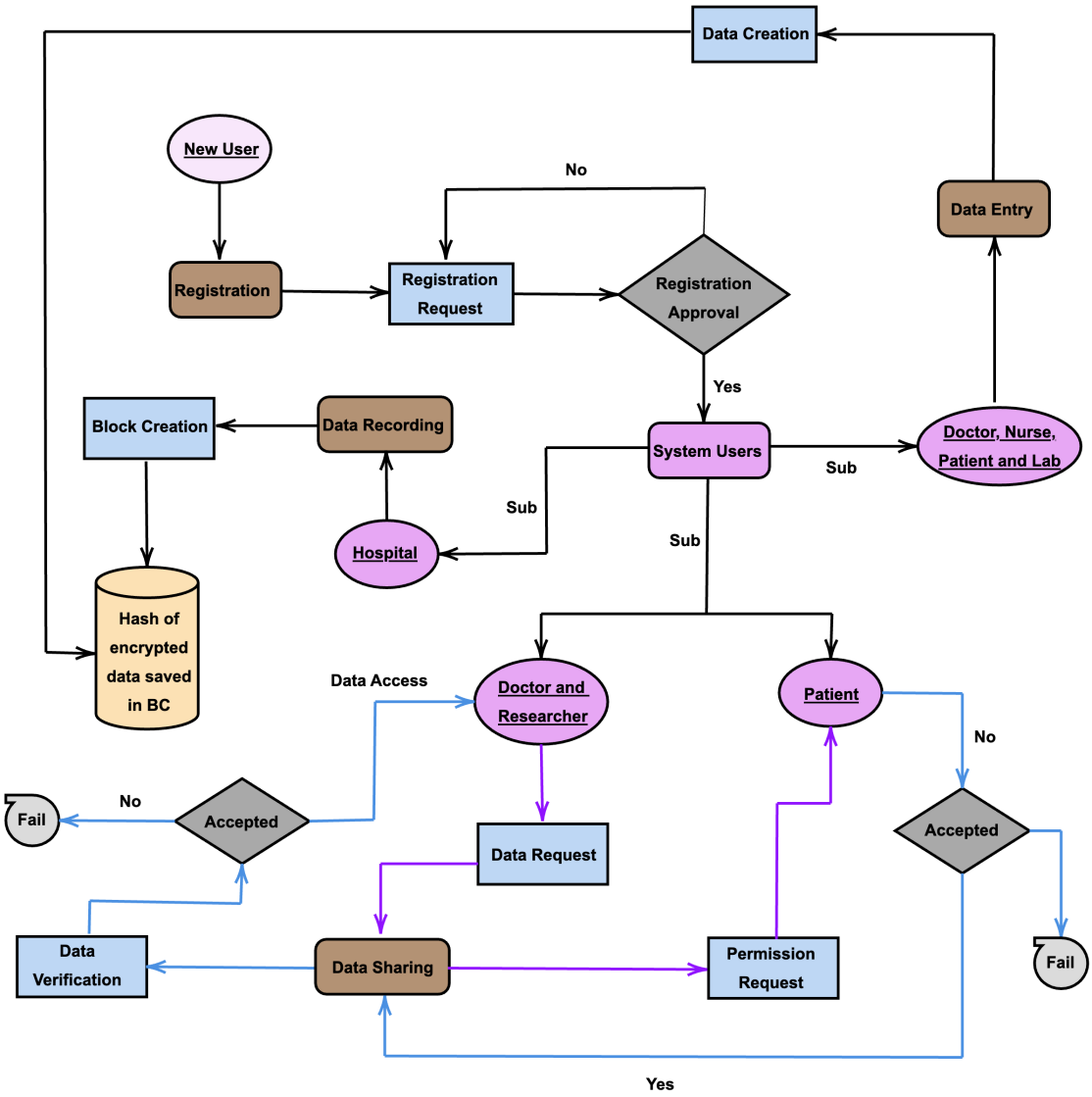

Figure 7.2 Activity Diagram of AguHyper [115].

Incorrect data entering may be avoided in this method. Because EHRs include personal information about specific persons, they are extremely sensitive. It makes sense that people would prioritize keeping their privacy protected in this way. Direct user access is made easier if shared data is kept on the platform in its original format; nevertheless, data privacy is jeopardized. Thus, it is essential to keep data private inside the system. EHRs are not preserved originally, but rather encrypted to IPFS to protect data privacy. The bulk of the data would be a scalability issue if encrypted data were kept directly in the blockchain as opposed to IPFS. It was recommended to maintain

hash values safely kept inside the BC and encrypted data in IPFS in order to address issues with both scalability and availability. Due to the fact that IPFS uses a dispersed network of nodes to store material. Because material may be fetched from numerous nodes, even if some nodes are down or having problems, this design improves availability. Because there is no one server or central authority that, in the event of a compromise, might bring down the entire system, the decentralized architecture of IPFS and BC also reduces the possibility of single points of failure. Both recipients of the data in the system need comprehensive details about it, and classification of the data is necessary after data entry is complete. Patients must put basic and general facts about the data into the data header during the data submission procedure in order to comply with these criteria and make it simpler for recipients to browse the data on the platform. Following the entry of data header information in accordance with specified criteria, the system categorizes the data.

After data entry is finished, it becomes necessary to store the data in the system in an orderly manner, which initiates the building of blocks. After data is carefully examined for authenticity and integrity, it is converted into blocks as part of the data recording process. Simultaneously, the phase of data recording plays a crucial role in defining the entities that are accountable for processing the data included in the blocks and choosing the particular consensus method that will be used during the data recording process. The immutability feature of IPFS and BC technology [107] is exploited to protect the integrity of healthcare data. The verification procedure in our system comprises a careful comparison between the hash applied to the encrypted data that is retrieved from storage and the hash of the encrypted data that is recorded on the ledger. Maintaining consistency among these hashes confirms the data's integrity and speeds up its delivery to the requester. On the other hand, a difference in the hashes indicates possible data corruption and notifies users. The patient's data is segmented into blocks during the data recording phase, and then the patient moves on to the data sharing phase.

Through the BC, recipients can access the records of the data produced on the platform. Through the dataSharing contract, a physician or researcher can get in touch with the appropriate patient to study one of these data. If a physician or researcher wants to obtain particular information from a patient, the patient is informed of this request. The requester is lawfully permitted to access the hash and key associated with the data, along with their own identity, upon receipt of this request and if the patient agrees to

provide access to the encrypted data stored in IPFS. Every transaction carried out on the platform is logged on a regular basis in blocks and added to the chain after being approved by the consensus procedure.

Authorized users actively carry out external checks to confirm the legitimacy of medical records [108]. Because transactions in our system need the user's signature, patients and healthcare providers share accountability for their data. This guarantees the undeniable source of user-generated data. By precisely identifying the accountable parties in situations of probable breaches, smart contract implementations of data access control and data usage audits facilitate the resolution of medical disputes and ensure responsibility. In addition to these initiatives, BC uses a number of defenses against denial-of-service (DoS) attacks, which try to stop a network or service from operating as usual. First off, because BC is decentralized, control and data are dispersed throughout a network of nodes, removing single points of failure and lessening the impact of conventional DoS assaults directed at centralized systems. Second, before a transaction is added to the BC, consensus mechanisms demand that nodes confirm and concur on its legitimacy. Because a majority of nodes must agree for a transaction to be deemed authentic, this agreement procedure stops malevolent actors from flooding the network with fake transactions. Finally, authorized hospital grants are used to complete the system registration procedure. The system's capabilities are restricted based on the responsibilities of its users.

7.5 Implementation

The Hyperledger Composer Business Network [109] and IPFS were set up, tested, and the network's functionality under varied workloads was shown in order to prepare it for practical implementation. The implementation details of the framework will be provided in this section. To make the process of developing apps on the Hyperledger Fabric blockchain more efficient, a development framework called Hyperledger Composer was created. Its main goal is to let users create blockchain apps on Hyperledger Fabric without requiring them to have a deep comprehension of the complex complexities related to BC networks. Furthermore, it comes with an online platform called the Hyperledger Composer Playground [110], which makes it easier to configure, install, and test a corporate network using a web browser rather than requiring a local network setup.

Composer defines four categories of resources using its in-house object-modeling language: i) Assets: Symbols for objects that the application is monitoring; ii) Participants: Identifiers of entities that are interacting with the network and have individual permissions; iii) Transactions: Messages that are sent to update an asset or a participant; they can also be used to carry out custom logic; iv) Events: Resulting from transaction logic, to which participants can subscribe. This research created the AguHyper Hyperledger Composer Business Network in order to take use of the previously described benefits. To configure, deploy, and test AguHyper, the Hyperledger Composer Playground was used. Three separate files make up the AguHyper Business Network: the model, the script, and the access control. Definitions of the assets, participants, transactions, and events are included in the model file. The access control file outlines the rights given to participants, assets, and transactions; the script file includes transaction logic in the form of functions.

Participants in the AguHyper include a patient, physician, researcher, nurse, and lab. Hospitals are the actual system administrators; ii) Patient data is an asset; iii) ParticipantCreation, assetCreation, DataSharingDoctor, and DataSharingResearcher are transactions. In the "ParticipantCreation" transaction, a participant is created by obtaining from users the data required for system registration. The "assetCreation" transaction includes the process of creating an asset, in which users are gathered to provide the encrypted data hash and other necessary information. Data sharing was taken into account for two different user situations during the AguHyper design process. In the first, data sharing with physicians is included, while in the second, data sharing with researchers is. Researchers declare their need for and request data in the "DataSharingResearcher" transaction in order to use it for analysis—such as illness prediction. Relevant data is supplied with the researchers when the request for data is approved. Researchers communicate analytic results, including illness forecasts, with the related patient after data exchange. Among other things, physicians do illness diagnosis in the "DataSharingDoctor" transaction. Doctors request data, and relevant information about the data is provided with them upon permission, in a manner akin to data sharing with researches. The following are the permissions on the system's participants, assets, and transactions:

- Information from doctors and researchers is readable by patients.
- Patients can access their belongings in full.
- Data request transactions are readable by patients.

- The meta data of assets may be accessed by physicians and researchers.
- Only the owners of the encrypted material have access to the hash.
- Transactions for Data Requests can be submitted by doctors and researchers.
- The researcher and physicians are granted authorization to view the encrypted data hash and the relevant data's pertinent details, provided that the right criteria are met.

We used a variety of API requests to access the Hyperledger Composer REST server [111] to evaluate the system's performance. With the help of the Hyperledger Composer Rest Server, an established Hyperledger Fabric business network may be transformed into a REST API that is simple for HTTP or REST clients to use. The Create, Read, Update, and Delete (CRUD) functions of the Hyperledger Composer REST server allow for the modification of asset and participant statuses as well as the query-based submission and retrieval of transactions. We used bespoke Node.js scripts for API requests.

7.6 Performance Analysis and Discussion

Using many API calls on the Hyperledger Composer REST Server [111], this section assesses the efficacy of the proposed architecture based on a variety of trials. In order to evaluate the effectiveness of the suggested framework, a patient-provider data exchange scenario was implemented. Transaction throughput, expressed in transactions per second (tps), average transaction latency, expressed in seconds, and upload and download times are the main performance measures used [112]. The System Under Test (SUT) blockchain finalizes legitimate transactions at a specific frequency within a defined timeframe, known as transaction throughput. It's important to note that this metric encompasses the aggregate performance across all nodes within the SUT rather than focusing solely on individual node activity. On the other hand, Transaction Latency provides a holistic assessment of the duration required for a transaction's impact to become functional throughout the network. This evaluation encompasses the time interval from when the transaction is initially submitted to when its outcome achieves widespread accessibility across the network. Such assessment incorporates factors like propagation duration and any settlement periods influenced by the prevailing consensus mechanism.

7.6.1 Experimental Setup

The research presented "AguHyper," a Hyperledger Composer Business Network. Using the Hyperledger Composer Playground, AguHyper was configured, deployed, and tested. In order to trigger the EHRs-Data-Creation and Data-Sharing chaincodes, we had to write special Node.js code. AguHyper was designed twice to provide a comparison study with SOLO-based research in the literature, hence facilitating a thorough review. AguHyper used three peer nodes in one organization in the first arrangement, and one peer node for each of the two organizations in the second setup, for a total of two organizations. An Intel Core-i9-9900K-16 CPU, 32 GB of RAM, and a server with 500 GB of storage capacity powered the complete system. Because Ubuntu 18.04 is compatible with Hyperledger Fabric 1.4, it was selected as the operating system. Hyperledger Fabric chose CouchDB as the world-state database, and the fabric block size was set to 256 MB. Furthermore, IPFS v0.4.22 was used for research using IPFS. For performance benchmarking, the study uses several network settings to implement different use cases.

7.6.2 Scenario 1

The first stage is to understand how changing the transaction rate (TPS) and number of transactions (Tx) affects throughput and average latency. Table 7.1 provides specifics about the first phase's network configuration.

Table 7.1 System configuration and simulation parameters for phase 1 [115].

	Phase 1
Processor	Intel Core-i9-9900K-16 CPU
Memory	32 GB
OS	Ubuntu 18.04
Hyperledger Fabric	v1.4
Rounds	10
Transactions	100, 250 and 500
Transaction send rate (tps)	5, 25, 50, 75, 100
State DB	CouchDB
Orderer and size	Raft and 2 Org-1peer each

The transaction rates for each individual transaction category were modified during the measuring period. Figure 7.3 shows how system throughput improves as the rate of transactions per second (tps) rises. However, as the number of transactions rises while

the present TPS rates are maintained, system throughput decreases. As the transaction rate and transaction volume grow, Figure 7.4 shows that the average latency increases as well. Additionally, it is interesting that even when the transaction rate increases, the delay does not show a substantial increase when examining equal transaction number groups. It is clear that more parameter tuning or the creation of better Smart Contracts might improve the system's throughput and latency.

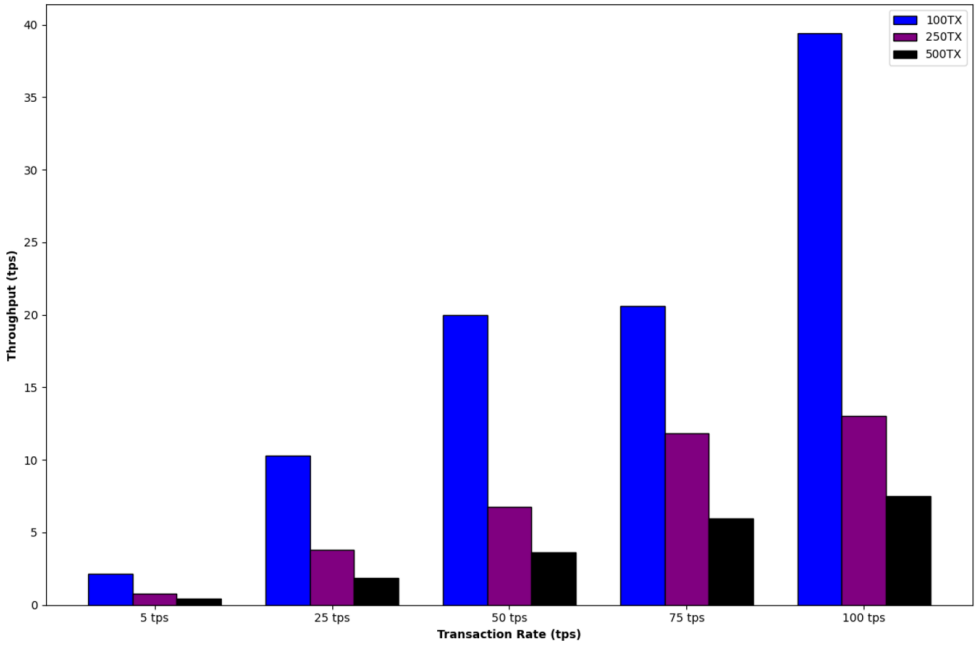

Figure 7.3 The influence of altering the number of transactions (Tx) and rate (TPS) on throughput [115].

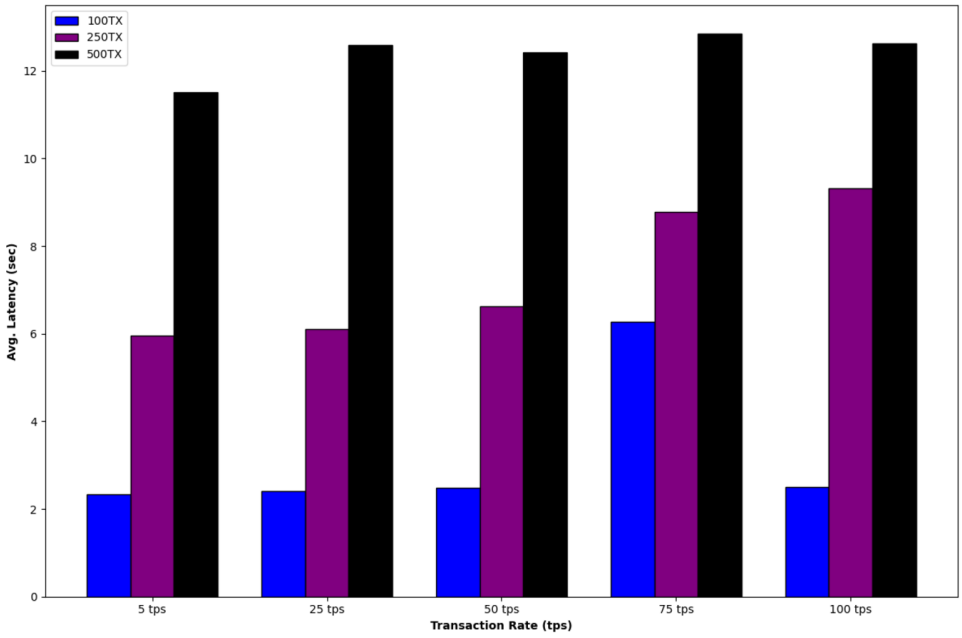

Figure 7.4 The influence of altering the number of transactions (Tx) and rate (TPS) on latency [115].

7.6.3 Scenario 2

The purpose of the second phase is to assess how scalable the IPFS-stored medical data is. It is made up of the amount of the data and the seconds needed for uploading and downloading it. Public text files that are randomly created serve as the data sets for study. The data size ranges from 0.003 MB to 100 MB, as shown in Figure 7.4.3.1 Notably, the chart shows that the uploading and downloading times for the data both rise as the data amount increases.

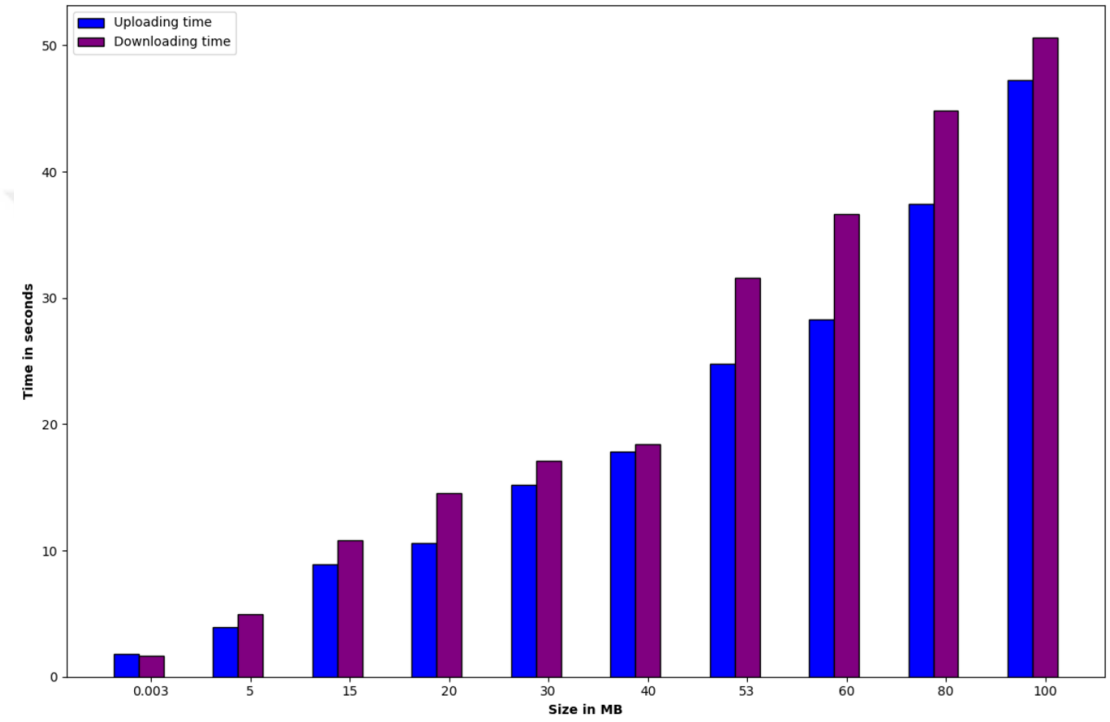

Figure 7.5 The process of uploading and downloading EHR data using IPFS [115].

7.6.4 Scenario 3

The third phase involves a comparative analysis between the performance indicators of AguHyper and the experiment data presented in [77, 113, 114]. The comparison is conducted based on the settings outlined in Table 7.2. The primary objective of this phase is to assess the impact of different consensus protocols on system performance by measuring throughput in transactions per second (tps) and average latency in seconds.

According to system configuration and simulation parameters for phase 3, the performance comparison of existing related works [77, 113, 114] and the proposed

work is demonstrated in Tables 7.3 and 7.4 based on throughput and average transaction latency.

Table 7.2 System configuration and simulation parameters for phase 3 [115].

Phase 3:	**Configuration**
Processor	Intel Core-i9-9900K-16 CPU
Memory	32 GB
OS	Ubuntu 18.04
Hyperledger Fabric	v1.4
Transactions	100, 200, 300, 400 and 500
State DB	CouchDB
Orderer and size	**AguHyper:** Raft and 2 Org-1peer each **Compared works (Kaur et al. [77]; Chelladurai and Pandian [113]; Chelladurai et al. [114]):** SOLO and 2 Org-1peer each

Table 7.3 Phase 3, a performance comparison between the proposed work and existing related works Kaur et al. [77], Chelladurai and Pandian [113] and Chelladurai et al. [114] are conducted based on throughput [115].

Transaction Groups (Throughput)	**AguHyper**	**Kaur et al.**	**Chelladurai and Pandian**	**Chelladurai et al.**
100	37.6217	36.1	4.2	5.82
200	39.67	39.5	10	10.54
300	34.8397	40.9	12	14.57
400	37.0006	40.1	16	17.89
500	38.1500	37	20.73	21.73

Table 7.4 Phase 3, a performance comparison between the proposed work and existing related works Kaur et al. [77], Chelladurai and Pandian [113] and Chelladurai et al. [114] are conducted based on throughput [115].

Transaction Groups (Average Latency)	**AguHyper**	**Kaur et al.**	**Chelladurai and Pandian**	**Chelladurai et al.**
100	2.625	1.74	2.1	2.12
200	4.9	3.14	2.8	2.74
300	6.84	4.57	3.4	3.46
400	9.04	5.32	4.2	4.28
500	11.23	5.9	4.85	4.81

Table 7.3 shows that the proposed system performs better than the studies by [113, 114]. for all transaction groups. It also outperforms the study by [77] in the 100, 200, and 500 transaction groups. As a result of Table 7.4 it is observed that the average transactional latency of the proposed system is marginally higher than the existing works. The systems under comparison utilize the SOLO consensus mechanism, whereas the proposed system employs the Raft consensus mechanism. Phase 3 experiments were

conducted under identical conditions to the existing systems. Therefore, it can be inferred that the utilization of Raft instead of SOLO contributes to an increase in both system throughput and latency.

7.6.5 Scenario 4

In the fourth phase, our objective is to evaluate the correlation between various performance indicators of AguHyper and the experiment data presented in [79]. This assessment is carried out in accordance with the settings specified in Table 7.5. The fourth phase aims to evaluate the impact of different consensus protocols and state databases on overall system performance by measuring uploading and downloading times.

Table 7.5 System configuration and simulation parameters for phase 4 [115].

Phase 4:	**Configuration**
Processor	Intel Core-i9-9900K-16 CPU
Memory	32 GB
OS	Ubuntu 18.04
Hyperledger Fabric	v1.4
Data Size	0.003 MB, 5 MB, 15MB, 20MB, 30 MB, 40 MB, 53 MB, 60 MB, 80 MB, 100 MB.
State DB	**AguHyper:** CouchDB **Compared work [79]:** LevelDB
Orderer and size	**AguHyper:** Raft and 1 Org-3peer **Compared work (Sonkamble et al. [79]):** SOLO and 1 Org-3peer

As per the system configuration and simulation parameters for Phase 4, the performance comparison between the proposed work and existing related work Sonkamble et al. [79] is depicted in Figures 7.6 and 7.7, focusing on uploading and downloading time. The uploading time encompasses the duration required for uploading data of a fixed size, including its encryption time. On the other hand, downloading time encompasses the total time for downloading the fixed data and the time required for its decryption. Figures 7.6 and 7.7 illustrate that the data size ranges from 0.003 MB to 100 MB. Specifically, as the data size increases, both uploading and downloading times increase. However, it is observed that the rate of increase in downloading time is higher than that of the uploading time with the increase in block size. The system under comparison utilizes the SOLO consensus mechanism and LevelDB, whereas the proposed system employs the Raft consensus mechanism and CouchDB. Phase 4

experiments were conducted under identical conditions to the existing system. Therefore, it can be inferred that the utilization of Raft and CouchDB instead of SOLO and LevelDB contributes to a decrease in both uploading and downloading time.

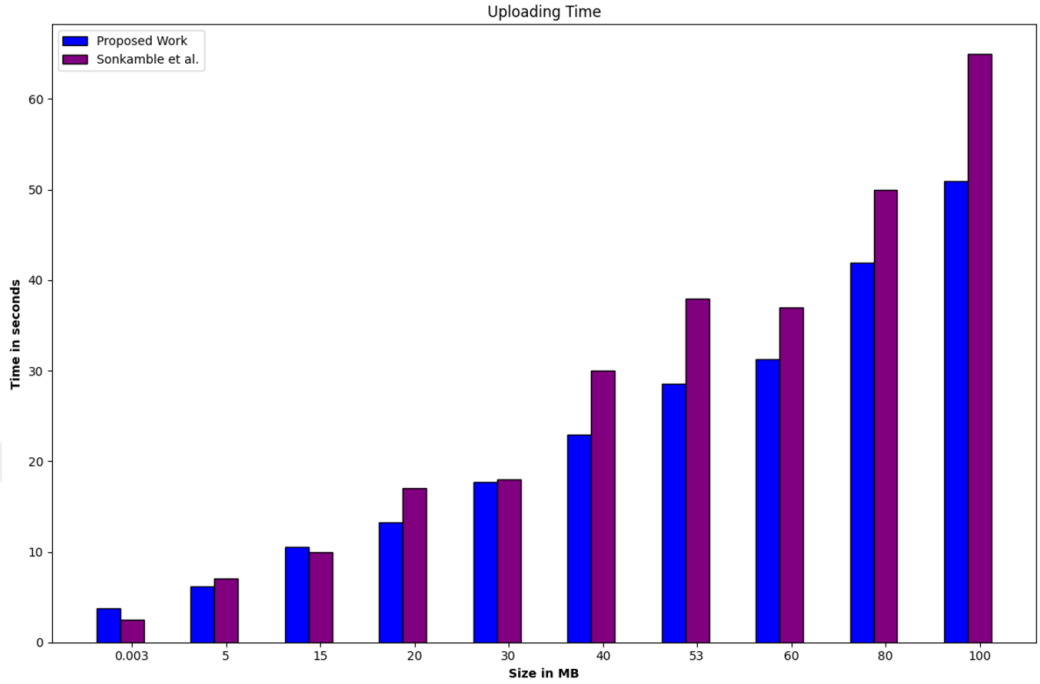

Figure 7.6 Phase 4, a performance comparison between the proposed work and existing related work Sonkamble et al. [79] is conducted based on uploading time [115].

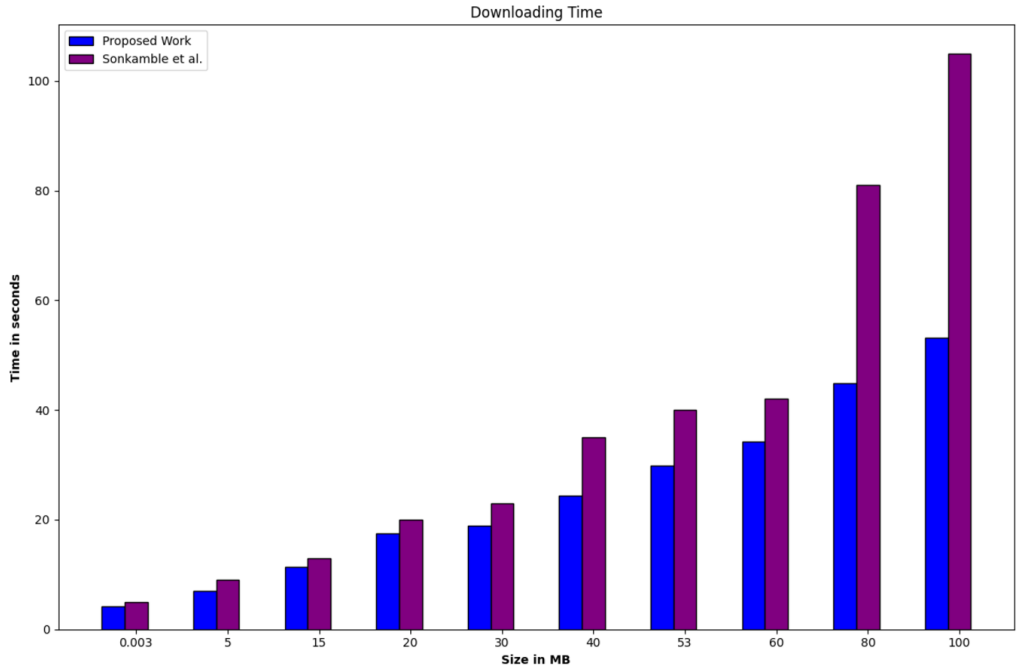

Figure 7.7 Phase 4, a performance comparison between the proposed work and existing related work Sonkamble et al. [79] is conducted based on downloading time [115].

7.6.6 Scenario 5

In the final phase, we conduct a feature-based comparison between AguHyper and existing works based on the ten different questions: Do the studies: i) use access control mechanisms?, ii) explain system permissions?, iii) use data verification mechanisms?, iv) solve security and privacy issues?, v) explain user roles in detail?, vı) use data sharing mechanism?, vıı) solve scability issue?, vııı) provide the availability?, ıx) show performance analysis based on BC?, and x) provide the appropriate basis for disease prediction?

A thorough comparison of features between the proposed work and existing related works is provided in Section 4, and a summary is presented in Table 4.1. In contrast to prior research, our proposed solution primarily enables the utilization of EHRs and ensures the secure sharing of these data. We assure information confidentiality, integrity, and optimal data transmission rates across all aspects.

Chapter 8

A Novel Classification Algorithm: CSA-DE-LR

8.1 Introduction

People worry and experience stress because of their busy schedules and daily routines. Furthermore, there has been a discernible surge in the prevalence of obesity and tobacco addiction, resulting in an escalation of illnesses such as cancer, heart issues, and other maladies [116]. Because these illnesses are so predictable, they pose a special risk. The onset date of various ailments is unpredictable. World Health Organization (WHO) estimates underscore a striking fact: Cardiovascular diseases cause 17.9 million fatalities globally annually, underscoring the alarming statistic that heart-related issues account for more than 32% of all mortality [117]. Of the several cardiovascular conditions, coronary artery disease (CAD) is a particularly challenging and prevalent illness.

CAD is the result of blocked coronary arteries, which supply the heart with essential blood. The accumulation of cholesterol and other substances causes these arteries to eventually fill with plaque, which obstructs blood flow. In its early stages, chest pain could be a sign of this artery constriction. Sadly, it can be challenging to diagnose CAD, thus patients usually encounter severe symptoms like heart attacks or heart failure as their primary warning signals.

The first step in figuring out whether CAD is present in a patient is figuring out if they are high-risk. A patient is subjected to a battery of tests, including electrocardiograms, echocardiograms, coronary angiography, blood tests, and chest X-rays, if they are deemed to be high-risk [118]. The complexity and cost of these diagnostic tests make them necessary, but they also make detecting and treating CAD more challenging and expensive. Thus, research and development of new techniques are

still needed in the medical field to increase the precision and accessibility of cardiovascular disease diagnosis.

The utilization of machine learning (ML) is highly recommended for the prediction of cardiovascular illness due to its ability to extract highly accurate and efficient data from large datasets, hence expediting the prediction process [119, 120]. As the cornerstone of machine learning, it is highly effective at handling large amounts of data, digesting information quickly, and providing predictions in the early phases of development. Applications of machine learning (ML) play a critical role in minimizing hospital mistakes, driving progress in illness prevention, early diagnosis, health policy, and lowering preventable hospital deaths. Similar goals have been pursued by a number of research, which have identified ML techniques that are effective in identifying CAD [121].

Initially, ML algorithms like Logistic Regression (LR), XGBoost, Support Vector Machine (SVM), and Naive Bayes (NB) were used for CVD prediction [122-125], but they struggle with handling complex and multidimensional data, leading to lower success rates [126]. A significant barrier to conventional methods, the local minima problem influences convergence to optimal solutions. Gradient-based optimization techniques such as gradient descent are used to iteratively update the model parameters in order to minimize the objective function. Algorithms might become trapped in local minima, though, if initialization or parameter updates force them into less-than-ideal solutions [127]. Researchers have used metaheuristics to pick features, adjust parameters, and train traditional machine learning algorithms to improve classification accuracy by avoiding local minima in the literature.

Without concentrating on ML algorithm training, the majority of research in the literature have employed metaheuristic techniques for feature selection (to lower dimensionality and accelerate computation time) and hyperparameter optimization (to identify almost ideal configurations for ML models) problems [128–133]. These works differ from the issues with the suggested technique since metaheuristics are employed in this study to train the LR while preserving its desirable aspects.

Researchers have used metaheuristics in a few studies to train standard machine learning algorithms such that they can improve classification accuracy by avoiding local minima [134–138]. In comparison to existing methods, these hybrid algorithms—which fuse machine learning with metaheuristics—have demonstrated better performance in the diagnosis of cardiovascular disease (CVD). To increase detection rates and

classification performance, these approaches need a substantial investment of time and energy. The machine learning approach to be employed must first be computationally efficient for huge datasets and have simplicity-interpretability in illness detection issues in order to eliminate these drawbacks of the present hybrid methods. Afterwards, it is necessary to choose the metaheuristic algorithm that will best overcome the disadvantages of the ML method to be used and will be successful for the train and suitable for the relevant problem [121].

Clonal Selection (CSA) is a crucial Artificial Immune System (AIS) technique that enhances the immune system's response to antigens by gradually producing antibodies with higher affinity. Because CSA may leverage receptor editing and hyper-mutation processes to explore the solution space for both local and global solutions, it is frequently employed in optimization issues [139]. Furthermore, studies have demonstrated that CSA-based tactics outperform alternative bio-inspired and optimization techniques in a range of settings [139–141]. However, it's possible that current CSA techniques need to be refined because they're not offering adequate search power [142–144].

In an effort to solve the shortcomings of earlier research and enhance the low detection rate in CVD prediction, a novel hybrid classifier known as CSA-DE-LR is developed in this thesis [121]. In this work, LR is employed as a classification model for the diagnosis of CVD since it is computationally affordable for large data sets and easily interpretable in sickness diagnostic concerns. Then, to increase LR's classification accuracy by avoiding local minima, the CSA-DE optimization approach is used for model training rather than the gradient descent algorithm [148–149]. The method employs three optimization algorithms, which are based on the F1 score, the Matthews correlation coefficient (MCC), and the Mean Absolute Error (MAE). Comprehensive tests on benchmark datasets like Cleveland and Statlog show that CSA-DE-LR performs better than state-of-the-art ML techniques. Furthermore, the Breast Cancer Wisconsin Original (WBCO) and Breast Cancer Wisconsin Diagnostic (WBCD) datasets are used to assess generalization tests. Interestingly, the suggested model shows better efficacy than other studies conducted in this field. The findings of this work mark a significant advancement in the fields of medical data analysis and CVD diagnosis by demonstrating how hybrid ML approaches may enhance diagnostic precision.

The main contributions of proposed model are as follows:

- In this study, a novel classification methodology called CSA-DE-LR is introduced. It combines a CSA and DE with LR. Especially in the context of CVD, this novel hybrid technique is designed to improve LR weights for effective categorization. Without concentrating on ML algorithm training, the majority of research in the literature have employed meta-heuristic techniques for feature selection and hyperparameter optimization challenges. In contrast to these works, the suggested strategy trains the machine learning algorithms using metaheuristics. It also gives thorough explanations of the reasoning for the combination of these three particular approaches.
- Three different optimization techniques based on the F1 score, MCC, and MAE are provided by the suggested CSA-DE-LR approach. In order to get the best classification performance, the model weights must be adjusted using these measures, which also serve as training guidelines.
- The paper uses two popular datasets, the Cleveland and Statlog datasets, to thoroughly assess the CSA-DE-LR approach. Accuracy, F1 score, MCC, ROC-AUC, false negative rate, and false positive rate are only a few of the metrics used to evaluate the performance. The outcomes are then compared with other well-known machine learning methods. When contrasting the suggested approach with prior research and presenting the findings, great care is taken to ensure complete transparency and equity. Furthermore, WBCO and WBCD datasets are used to assess generalization tests. Additionally assessed are the moral ramifications of applying ML models to the medical field.
- In contrast to previous research, the report offers insightful information on the significance of feature selection and model optimization by in-depth analysis. It investigates how improving predictive consistency and generalizability might result from removing certain characteristics, emphasizing the significance of dataset-specific tuning and rigorous feature selection using an alternative method.
- The study demonstrates that, when applied to the Cleveland and Statlog datasets, CSA-DE-LR performs more accurately and precisely than earlier techniques. This illustrates the method's efficacy and promise for enhancing medical professionals' diagnostic decision-making processes.

8.2 Related Works

Numerous studies are presently being conducted in the area of diagnosing cardiovascular illness, each with a different goal in mind [121]. The goal of these studies is to improve the efficiency and accuracy of diagnosis. The aforementioned goals encompass the identification of ideal characteristics that are essential for precise diagnosis, the development of inventive classification models customized to the intricacies of cardiac ailments, and the creation of exceptionally effective classification techniques that may expedite the diagnostic procedure. A number of strategies, including machine learning and metaheuristic-based procedures, have been put forth in an effort to produce thorough and efficient diagnostic results [150, 151].

8.2.1 Machine Learning Techniques

For CVD prediction, ML algorithms such LR, XGBoost, SVM, NB, and others were employed in the early research projects [122–125].

To improve performance, Kolukisa et al. (2019) [123] broadened the scope of feature selection techniques. Additionally, by reducing the amount of characteristics in the diagnosis of coronary artery disease, Fisher linear discriminant analysis was used to save computing time and provide well-performing models for each dataset. On the Cleveland dataset, they obtained an accuracy of 82.5% and an F-measure of 83.80% by using the Multi-Layer Perceptron (MLP) classifier. An innovative self-optimized and adaptable ensemble machine learning method was presented in research [124]. Using a variety of coronary artery disease datasets, this approach finds the best machine learning models on its own and guarantees excellent accuracy. 83.43% accuracy on the Cleveland dataset was attained. Moreover, Kolukisa and Bakir-Gungor's study [122], which included the Cleveland, Statlog, and Z-Alizadehsani datasets, suggested a probabilistic ensemble FS technique as well as an exhaustive ensemble feature selection (FS) method. Six different categorization algorithms and four voting algorithm variations were evaluated. For the Cleveland dataset, the accuracy values that were obtained were 85.47%, whereas for the Statlog dataset, they were 85.55%.

LR and XGBoost are thoroughly examined using the Statlog heart disease dataset as a benchmark by Dhanka et al. [125]. Random SearchCV hyperparameter tweaking is used to optimize the model parameters. Analysis of both non-optimized and optimized

models is included in the paper. The 10-fold cross-validation results show that the accuracy of LR and XGBoost are 85.2% and 81.5%, respectively.

While typical machine learning algorithms have demonstrated significant effectiveness in the prediction of CVD, their management of complex and multidimensional data is generally problematic, resulting in reduced success rates [126]. Moreover, the local minima problem poses a significant challenge to conventional machine learning techniques, impacting the convergence of algorithms towards optimal solutions. The fundamental issue with this topic is the complexity and non-convexity of objective functions, which might have several local minima in addition to a global minimum. Local minima are places in the parameter space where the objective function approaches a local low; these are not always the lowest locations overall, though. Model parameters are regularly modified iteratively utilizing gradient-based optimization techniques, including gradient descent, in order to minimize the objective function. However, these algorithms may become trapped in local minima if initialization or parameter updates force them into regions of the parameter space that correspond to these less-than-ideal solutions [127]. Breaking out of such local minima is necessary to achieve optimal model performance. Researchers using ML algorithms have used a variety of metaheuristic approaches to tackle these issues.

8.2.2 Hybrid Approaches using Metaheuristics and ML Algorithms

Researchers have used metaheuristics in the literature to choose features, adjust parameters, and train the common machine learning algorithms to increase classification accuracy by avoiding local minima.

In order to choose the best features across various CVD dataset types, the majority of research in the literature have employed metaheuristic techniques for feature selection issues without concentrating on training ML algorithms [130–133]. In contrast to the suggested work, these works train machine learning algorithms using metaheuristic techniques while maintaining their desirable characteristics. Metaheuristics aid in reducing dimensionality and expediting computation in feature selection challenges. One of these research is [130], in which the authors combined three bioinspired algorithms with artificial neural networks (ANN) to produce a super learner. Three feature sets were selected using the Bacterial Foraging Optimization (BFO), Krill Herd (KH), and Cat Swarm Optimization (CSO) approaches. Using the

characteristics chosen by each method, a Backpropagation Neural Network (BPNN) was trained. The Statlog dataset yielded an accuracy of 86.36%, whereas the Cleveland dataset produced an accuracy of 84%.

In order to increase the accuracy of heart disease prediction, the study [131] suggested BAPSO-RF, a Bat Algorithm and Particle Swarm Optimization-based Random Forest. The suggested BAPSO-RF is evaluated using 270 records and 14 variables from the UCI heart disease dataset. Using criteria like accuracy, precision, recall, and f1-score values of around 98.71%, 98.67%, 98.23%, and 98.45%, respectively, the model beat competing techniques including GAPSO-RF, GA, and GA-RBF.

The work by [132] makes use of a publicly available dataset from the UCI machine learning repository to predict cardiac disease using ensemble classifiers with parameter tuning. Thirteen factors that affect heart disease are included in the dataset. Principal Component Analysis (PCA) was utilized for feature extraction and Particle Swarm Optimization (PSO) for feature selection. Machine learning techniques like SVM, Deep learning, and Ensemble Classifier have all been used with parameter optimization. The findings revealed that SVM parameters and Deep Learning had the highest accuracy, with 83.51% accuracy being achieved using SVM bagging.

The paper [133] suggested a way to increase prediction accuracy by extracting important characteristics from a cardiac dataset using Binary Particle Swarm Optimization and Attention-based Deep Network (BPSO-ADN). The method finds the best appropriate subset for heart disease prediction using a fitness evaluation method that guides BPSO feature selection. ADN is used for in-depth pattern analysis.

Researchers have used metaheuristics in certain studies to improve parameters in an effort to increase the accuracy of their classifications [128, 129]. In contrast to the suggested work, these works train machine learning algorithms using metaheuristic techniques while maintaining their desirable characteristics. When it comes to efficiently investigating complex search spaces in hyperparameter optimization problems, metaheuristics are a valuable resource. This facilitates the process of determining the best or almost best settings for machine learning models. These methods offer flexible and dependable solutions by breaking out from local optima and improving model performance across a variety of tasks and datasets through the use of parallelization, stochastic search, and adaptive exploration. One of these studies is [128], in which the hybrid system was used by the authors to diagnose cardiac disease.

SVM and MLP classifiers were used for the classification. Three evolutionary algorithms were used to optimize the parameters: PSO, FA, and GSA (Gravity Search Algorithm). In MLP, learning rate and momentum were maximized. Margin was SVM-optimized. The approach was validated using five datasets pertaining to cardiovascular illness. The system obtained 94.1% accuracy on the Cleveland dataset, 90.74% on the Statlog dataset, 89.5% on the SPECT dataset, 90.6% on the SPECTF dataset, and 91.4% on the Eric dataset.

Two intelligent models were created by the study [129]: the HyOPTRF (Model 1) and the HyOPTXGBoost Classifier (Model 2), which were used with both hyper-tuned and default parameters on the Statlog HD dataset. The HyOPTXGBoost Classifier recorded the greatest Accuracy of 96.30% and F1 Score 96.77% on Trial No. 33, while the HyOPTRF recorded the highest Accuracy of 92.59% and F1 Score 93.75% on Trial No. 2. The recommended models were validated using the Stratify Kfold Cross-Validation method and contrasted with the other models that were currently in use.

Researchers have used metaheuristics in a few studies to train standard machine learning algorithms such that they can improve classification accuracy by avoiding local minima [134–138]. In the literature, the resulting algorithms that fuse machine learning with metaheuristics are called hybrid algorithms. For instance, in [134], the authors used a hybrid optimization technique in an attempt to improve the performance of ANNs. Three benchmark datasets were used to test this strategy: Cleveland, Pima Indian Diabetes, and Wisconsin Breast Cancer. The backpropagation (BP) algorithm was used for local search, while a mixture of DE and PSO for global search were used in the ANN training. Min-max normalization was applied to the datasets before building the model. A 10-fold cross-validation was conducted, and the proposed approach, termed Differential Evolution with Global Information and Back Propagation (DEGI-BP), was compared with DE-BP and PSO-BP. The experiments conducted on the Cleveland dataset demonstrated that the proposed approach outperformed other hybrid optimization algorithms, achieving an accuracy of 86.66%.

The study [135] suggested a neural network-based method for CAD diagnosis. The neural network's weights were tuned with the Genetic Algorithm (GA). The backpropagation method was used to train the ANN. One hundred chromosomes made up the first population of GA. The fitness value of the chromosomes was calculated using the untrained ANN's root mean square error, or RMSE. The roulette wheel algorithm was used by GA for selection. Two-point crossover was used, with a

crossover probability of 1. The mutation was carried out using Gaussian mutation. Each chromosome had all of the neural network's weights, and each gene on a chromosome carried one neural network weight. The features were chosen using SVM. Z-Alizadeh Sani was used as the dataset for system evaluation. The accuracy of the system was 93.85%.

To forecast cardiac illness, [136] introduced a hybrid classifier. Features were selected using the Orthogonal Local Preserving Projection (OLPP). The categorization was done using the ANN. Four neurons made up the input layer of the neural network, one hundred neurons made up the buried layer, and five neurons made up the output layer. The weights assigned to the connections across neurons ranged from -10 to 10. By figuring out the weights, Levenberg-Marquardt (LM) and Group Search Optimization (GSO) were utilized to optimize the network. The two sets of weights that LM and GSO had generated were used to choose the network's ideal weights. The results were validated using three datasets: Switzerland, Hungarian, and Cleveland. Using the Cleveland dataset, the accuracy rate of the method was 94%.

In a groundbreaking study [137], an Emotional Neural Network (EmNN) and PSO were combined. This new method's performance was contrasted with that of a hybrid model called PSO-ANFIS, which combines fuzzy logic and an ANN. The research focused on utilizing brain-based emotional learning in EmNNs, which are known for their high accuracy. To optimize the suggested neural network, PSO was used. Three datasets were included in the evaluation: Cleveland, Statlog, and Z-Alizadeh Sani. Eight features were chosen for the Statlog dataset and seven features for the Cleveland dataset after feature selection was carried out without any data preparation. For the Cleveland dataset, the accuracy was 84%, while for the Statlog dataset, it was 85.2%.

A hybrid technique called MLP-PSO was presented in a paper by [138] using the Cleveland dataset with 13 features and 303 samples. This approach used feature scaling and categorical data encoding techniques, substituting feature-specific mean values for missing values. Weights and biases that were optimized using the PSO technique were used to train the MLP model. Grid search was used for hyperparameter tweaking, and 5-fold cross-validation was used for performance measurement. When 10 distinct machine learning algorithms were used to compare the findings, the suggested MLP-PSO technique showed the best accuracy, coming in at 84.60%. To enhance classification performance, they refined the feature extraction process and training step of a neural

network (NN) following the methodology described in [152] Statistical and higher-order statistical features were extracted from the dataset, and PCA was performed.

The results show that, when it comes to several performance parameters, metaheuristic-based machine learning approaches perform better than other ML methods when it comes to diagnosing CVD. However, the application and usefulness of these approaches are severely limited since they often take a long time to develop in terms of low detection rates and a great deal of work to attain good classification performance. The machine learning approach to be employed must first be computationally efficient for huge datasets and have simplicity-interpretability in illness detection issues in order to eliminate these drawbacks of the present hybrid methods. The metaheuristic algorithm that will best address the drawbacks of the ML approach to be employed, be successful for the train, and be appropriate for the pertinent problem must then be selected [121].

In order to address the limitations of previous research and improve the poor detection rate in CVD prediction, this study introduces a novel hybrid classifier known as CSA-DE-LR [121]. Because LR is easily interpretable in illness diagnostic issues and computationally economical for big data sets, it is used as a classification model in this work for the diagnosis of CVD. Then, by avoiding local minima, the CSA-DE optimization approach is used for model training rather than the gradient descent strategy, increasing the classification accuracy of LR.

The following justifications support the selection of CSA-DE as an optimization technique. An important AIS algorithm called the CSA helps the immune system respond better to antigens by gradually producing more affinities in the antibodies it produces. Because it may employ receptor editing and hyper-mutation techniques to seek for both local and global solutions in the solution space, CSA is well-liked for optimization tasks [139]. Additionally, research has shown that, in a variety of settings, CSA-based strategies perform better than other bio-inspired and optimization techniques [139–141]. Traditional CSA methods, however, do not provide enough search power and might want some enhancement [142–144].

In conclusion, the hybrid classifier CSA-DE-LR that has been suggested combines two optimization techniques and applies them to LR during training to get improved performance. Notably, it provides three optimization choices based on F1 score, MCC, and MAE, which increases its flexibility. Furthermore, applying feature selection improved the results of this investigation considerably. The CSA-DE-LR technique

demonstrated the importance of careful feature selection in enhancing diagnostic models for CVD by achieving amazing accuracy and efficiency by finding and leveraging the most essential characteristics. Using ten-fold cross-validation and Bayesian optimization to fine-tune hyperparameters, CSA-DE-LR presents a significant breakthrough in medical diagnostics by improving diagnosis accuracy significantly across datasets.

8.3 Methods

8.3.1 Logistic Regression

Selecting LR is a form of statistical modeling that's especially suitable for situations where the outcome variable is binary, meaning it takes on two possible outcomes. The primary concept behind LR is to model the probability that a given input point belongs to a particular category. This probability estimation is achieved by fitting the data to a logistic curve, hence the name "Logistic Regression."

$$y_i' = \begin{matrix} 0, & p_i < 0.5 \\ 1, & p_i \geq 0.5 \end{matrix} \tag{8.1}$$

The training set of features $\{(\vec{x}_1, y_1), \ldots, (\vec{x}_M, y_M)\}$ comprises M instances. Each instance has a feature vector $\vec{x}_i \in R^D$ and the corresponding y_i is the label for each feature vector, which, in this binary classification scenario, can either be 0 or 1.

In mathematical terms, the prediction for a given $\vec{x}_i$ is determined by inputting the weighted sum of its features into the sigmoid function, and formula (8.1) is used to describe the class of formula (8.1).

where p_i is determined by formula (8.2) LR uses the sigmoid function, which outputs a value between *0* and *1* for any input as outlined in formula (8.3). This value can be interpreted as the probability that the input instance belongs to the class labeled as *1*.

$$p_i = \sigma(\vec{w}\vec{x}_i) \tag{8.2}$$

$$\sigma(a) = \frac{1}{1+e^{-a}} \tag{8.3}$$

The goal of the learning algorithm is to adjust its internal parameters (typically weights associated with each feature and a bias term) to minimize the difference

between its predicted probabilities and the actual outcomes in the training set. This difference is often captured by a cross-entropy cost function given in formula (8.4).

$$J\left(\vec{w}\right) = -\sum_{i=1}^{m} y_i \log(p_i) + (1 - y_i)\log(1 - p_i) \tag{8.4}$$

8.3.2 Clonal Selection Algorithm

The CSA [153] emulates the adaptive immune system's response to antigenic stimuli. An algorithm, CLONALG, was developed based on the principles of clonal selection and affinity maturation inherent in immune responses. Different adaptations of CLONALG were employed to handle pattern recognition and optimization problems [140, 146, 153]. For the application of CSA, a version of CLONALG optimized for such tasks was employed. The primary objective of CSA is to identify an antibody with peak affinity [139, 153]. The key steps of CSA are outlined in Algorithm 1.

Algorithm 1: The Clonal Selection Algorithm

```
Set up the control parameters: total antibodies (P), the highest iteration count,
receptor editing rate (B), and clonal multiplication coefficient (α).
Create an initial set of P antibodies.
Determine the affinity score for every antibody Ab.
for each iteration do
    Clone the antibodies α times and calculate the affinity scores for these clones.
    for each clone Ci do
        Apply reverse mutation to Ci, resulting in the mutated clone σi.
        if f(σi) > f(Ci) then
            Ci := σi
        else
            Carry out pair-wise mutation on Ci to produce σi.
            Compute the affinity value of σi.
            if f(σi) > f(Ci) then
                Ci := σi
            else
                Ci := Ci
            end if
        end if
    end for
    for each antibody Abi do
        Choose the clone Cj from Abi clones with the topmost affinity
        Abi := Cj
    end for
    Substitute the lowest performing B% of antibodies with the newly formed ones.
end for
```

The set of P antibodies $Ab = \{Ab_1, Ab_2, \ldots, Ab_P\}$ is initially formed randomly as indicated in formula (8.5). This set undergoes enhancement in every cycle through

processes like selection, cloning, hyper-mutation, re-selection, and receptor-editing until the highest iteration count is reached [153]. Each antibody $Ab_i = [Ab_{i,1}, Ab_{i,2}, \ldots, Ab_{i,D}]$ within the group represents a potential solution. The ultimate goal is to identify an antibody with the highest affinity score once all cycles have concluded.

$$Ab_{i,j} = lb_j + rand(0,1) \times \left(ub_j - lb_j\right) \tag{8.5}$$

where $rand(0,1)$ is a function producing random numbers uniformly spread between 0 and 1. The terms lb_j and ub_j denote the minimum and maximum limits for the j^{th} parameter, respectively. After establishing a population of P antibodies, the fitness score for every antibody in the Ab set can be determined, as illustrated in formula (8.6).

$$f(Ab_i) = \frac{1}{1+J(Ab_i)} \tag{8.6}$$

$f(Ab_i)$ represents the fitness function determining the fitness score of Ab_i. Meanwhile, $J(Ab_i)$ serves as the cost function, as depicted in formula (8.4) providing the cost measure of Ab_i. The clone count (α_i) for every chosen antibody Ab may remain consistent [153]. The aggregate count of clones within the clone group C is derived from formula (8.7).

$$|C| = \sum_{i=1}^{n} \alpha_i = \alpha \times n \tag{8.7}$$

where α represents the clonal multiplication coefficient and is a positive whole number. In this study, every antibody in the group is chosen for cloning $(n = P)$ and the quantity of clones for each selected antibody remains consistent $(\alpha_i = \alpha)$ to aid in identifying multiple optimal solutions [153].

Following the formation of the clone group, the antibodies undergo enhancement via hyper-mutation processes, namely inverse mutation and pair-wise mutation. In the inverse mutation method, for each clone $C_i = [C_1, C_2, \ldots, C_D]$, parameters j and l are chosen at random, ensuring $|j - l| > 2$, and the parameters between j and l within C_i are inverted to produce the mutated clone σ_i. If σ_i affinity surpasses that of C_i, it replaces C_i. If not, pair-wise mutation is applied to C_i. Here, parameters j and l of C_i are randomly selected and swapped. The affinity of σ_i, resulting from pair-wise mutation, is assessed. If $\sigma_i's$ affinity is superior to that of C_i, it takes the place of C_i; if not, C_i remains unaltered.

Post hyper-mutation, a re-selection step ensures the antibody population size stays consistent. For each antibody, Ab_i, the highest affinity clone among Ab_i clones is chosen and allocated to Ab_i. Concludingly, receptor editing is executed, substituting the least efficient B% of antibodies with new ones. The procedures of the CSA optimization approach are detailed in Algorithm 1.

8.3.3 Differential Evolution

The DE algorithm [154] operates as a collective method that encompasses processes like crossover, mutation, and selection. Its core mechanism hinges on mutation, which derives from the distinctions between randomly chosen solution pairs within the collective. This algorithm harnesses mutation as an exploration tool and the selection process to guide the exploration towards favorable areas in the solution environment. The DE algorithm also employs a distinctive crossover that may favor parameters from one parent over another. By leveraging attributes from current collective members to formulate trial solutions, the crossover operator adeptly redistributes insights about potent combinations, facilitating a more effective search for optimal solutions. Initially, DE establishes a random set of solution vectors. This set undergoes enhancements through the application of mutation, crossover, and selection processes. In the DE method, every newly generated solution is compared against a mutated one, and the superior of the two emerges victorious. The DE algorithm has captured the attention of scholars in diverse fields and has proven valuable in solving numerous real-world challenges [154, 155, 145-147]. The essential steps of the DE algorithm are described as follows:

Algorithm 2: Differential Evolution Algorithm

```
1: Initialize Population
2: Evaluation
3: repeat
4:     Mutation
5:     Recombination
6:     Evaluation
7:     Selection
8: until requirements are met
```

During the mutation process, all of the M parameter vectors are subjected to mutation. This mutation step broadens the exploration area. A mutated solution vector, denoted as w_i, is formulated by formula (8.8):

$$\vec{w}_i{}' = \vec{w}_{i_1} + sf \times (\vec{w}_{i_3} - \vec{w}_{i_2}), 1 \leq i \leq M \tag{8.8}$$

where sf represents the scaling factor with values from [0,1], and the solution vectors i_1, i_2, i_3 are selected randomly and are required to conform to $i_1 \neq i_2 \neq i_3 \neq i$, where i represents the current solution's index. During the crossover procedure, the parent vector merges with the mutated vector, generating a trial vector as given in Eq. 9.

$$\vec{w}_{i,j} = \begin{cases} \vec{w}_{i,j}{}', & r_j \leq cr \\ \vec{w}_{i,j}, & r_j > cr \end{cases} \tag{8.9}$$

where cr represents the crossover constant, r_j is a real number picked randomly from the range [0,1], and j indicates the j^{th} element of the related array.

Every solution within the population possesses an equal probability of being chosen as a parent, regardless of its fitness value. After undergoing mutation and crossover processes, the offspring's performance is assessed. Subsequently, a comparison between the offspring and their parent takes place, with the superior entity prevailing. If the parent remains superior, it is preserved in the population.

8.3.4 Proposed Method (CSA-DE-LR)

This study introduces a novel classification methodology, leveraging a hybrid approach that combines CSA and DE to optimize the LR weights for classification tasks. The proposed method, henceforth referred to as CSA-DE-LR, offers three optimization techniques based on different performance metrics: F1 score, MCC, and MAE. These metrics guide the training process to fine-tune the model weights to achieve an optimal balance between precision and generalizability.

The CSA-DE-LR method begins by initializing a population of P antibodies, each representing a potential solution to the classification problem. The antibodies undergo cloning and local search procedures via DE, followed by a selection phase that favors the most promising candidates. Receptor editing is applied to introduce diversity by replacing a portion of the least-fit antibodies with new candidates. This iterative process continues until the maximum evaluation number (*MEN*) is reached, at which point it returns the best-performing antibody, indicative of the optimal model weights.

The detailed pseudocode for the CSA-DE-LR classification method is presented from Algorithm 3 to Algorithm 11. Algorithm 3 outlines the primary procedure, which utilizes sub-procedures defined in Algorithms 4 to 11.

Algorithm 3: Proposed CSA-DE-LR Classification Method

1: Determine the input parameters: Input matrix $XM \times N$, target $\vec{y}M$, number of antibodies P, population of P antibodies $WP \times D$, percentage of receptor editing B, maximum evaluation number MEN, lower bound lb, upper bound ub, number of clones for each antibody α, scaling factor sf, crossover rate cr

Output:

1: $D \leftarrow N + 1$
2: $W \leftarrow CreateAntibodies(P, D)$
3: $W \leftarrow W$
4: $\vec{f}it \leftarrow CalculateFitness(W)$
5: $evaluation\ number \leftarrow 0$
6: **while** $evaluation\ number < MEN$ **do**
7: *Cloning*()
8: *LocalSearchViaDE*()
9: *Selection*()
10: *ReceptorEditing*()
11: *FindBestAntibody*()
12: **end while**
13: *return* $g\vec{p}ar$

Algorithm 4: Create Population of *P* Antibodies

1: **procedure** *CreateAntibodies*(P, D)
2: **for** $i \leftarrow 1 : P$ **do**
3: **for** $j \leftarrow 1 : D$ **do**
4: $W[i, j] \leftarrow lb + rand(0, 1) \times (ub - lb)$
5: **end for**
6: **end for**
7: *return* W
8: **end procedure**

Algorithm 5: Clone Each Antibody α Times

1: **procedure** *Cloning*()
2: **for** $i \leftarrow 1 : P$ **do**
3: $C[i \times \alpha : (i+1) \times \alpha, :] \leftarrow W[i, :]$
4: **end for**
5: **end procedure**

Each sub-procedure is dedicated to a specific task within the optimization process, including the initialization of the antibody population (Algorithm 4), cloning (Algorithm 5), local search via DE (Algorithm 6), selection (Algorithm 7), receptor editing (Algorithm 8), and the identification of the best antibody (Algorithm 9).

A critical phase within this process is detailed in Algorithm 7, the Selection Phase, where the efficacy of each antibody's clone is rigorously evaluated. Every antibody's clones are compared, and the clone with the highest fitness value is identified.

Algorithm 6: Local Search via DE

procedure *LocalSearchViaDE*()
  for $i \leftarrow 1 : P \times \alpha$ **do**
    $j \leftarrow i // \alpha$
    $\vec{ind}s \leftarrow \{x | x \in Z, 0 \leq x < P, x \neq j\}$
    $pars = rand\ \ choice(\vec{ind}s, 3) \times \alpha + randint(0, \alpha, 3)$
    $\vec{arr} \leftarrow C[pars[0],:] + sf \times (C[pars[2],:] - C[pars[1],:])$
    $\vec{ar} \leftarrow rand(low = 0, high = 1, size = (D))$
    $\vec{\rho} \leftarrow \vec{ar} \leq cr$
    $C[i, \vec{\rho}] \leftarrow \vec{arr}[\vec{\rho}]$
    $\vec{vec} \leftarrow C[i, \vec{\rho}]$
    $\vec{vec}[\vec{vec} < lb] \leftarrow lb$
    $\vec{vec}[\vec{vec} > ub] \leftarrow ub$
    $C[i, \vec{\rho}] \leftarrow \vec{vec}$
  end for
  $\vec{cfit} \leftarrow CalculateFitness(C)$
end procedure

Algorithm 7: Selection Phase

procedure *Selection*()
  $cfit' \leftarrow reshape(\vec{cfit}, size = (P, \alpha))$
  $\vec{max\_idxs} \leftarrow argmax(cfit', axis = 1)$
  $\vec{ind}s \leftarrow [0 : 1 : P] \times \alpha$
  $\vec{idxs} \leftarrow \vec{ind}s + \vec{max\_idxs}$
  $\vec{bestIdxs} = \vec{cfit}[\vec{idxs}] > \vec{fit}$
  $W[\vec{bestIdxs}, :] = C[\vec{idxs}, :][\vec{bestIdxs}, :]$
  $\vec{fit}[\vec{bestIdxs}] = \vec{cfit}[\vec{idxs}][\vec{bestIdxs}]$
end procedure

Algorithm 8: Receptor Editing Phase

procedure *ReceptorEditing*()
  $\vec{fIndex} \leftarrow argsort(\vec{fit})$
  $n \leftarrow round(P \times B)$
  $worstNindex \leftarrow \vec{fIndex}[0 : n]$
  $newNantibodies \leftarrow CreateAntibodies(n, D)$
  $newNfitness \leftarrow CalculateFitness(newNantibodies)$
  $W[worstNindex,:] = newNantibodies$
  $\vec{fit}[worstNindex] \leftarrow newNfitness$
end procedure

Every antibody's clones are compared, and the clone with the highest fitness value is identified. If this clone surpasses the original antibody in terms of fitness, it is preferentially selected as a superior solution. This approach ensures that the proposed model continuously evolves towards higher accuracy by adopting the most advantageous traits of each generation.

Following this, Algorithm 8, the Receptor Editing Phase, comes into play. Here, antibodies are ranked based on their fitness values, and the bottom percentile, amounting to PxB antibodies, is identified for replacement. This mechanism introduces strategic diversity to the population by substituting the least-fit antibodies with newly created ones, thus preventing premature convergence and maintaining a robust search within the solution space.

Prediction calculations are performed per Algorithm 10, where the sigmoid function (σ) is applied to the weighted sum of inputs plus the bias term. The fitness of each antibody is evaluated using one of the three fitness functions (Algorithms 11 to 13), each corresponding to one of the selected optimization metrics.

Algorithms 14 to 16 encapsulate the calculation of the MCC, F1 score, and MAE, respectively. The MCC computation follows the standard formula involving true positives, true negatives, false positives, and false negatives. Similarly, the F1 score is computed using precision and recall derived from the confusion matrix. The MAE, on the other hand, is inverted to ensure that a lower error results in higher fitness.

Algorithm 9: Find Best Antibody

```
procedure FindBestAntibody()
    index ← argmax(f⃗it)
    gmax ← f⃗it[index]
    gpar⃗ ← W[index,:]
end procedure
```

Algorithm 10: Calculate Prediction

```
procedure CalculatePrediction(φ)
    w ← φ[:, 1:]
    b ← φ[:, 0]
    prediction ← σ(X.dot(w^T) + b)
    return prediction
end procedure
```

By integrating CSA and DE with LR, the proposed CSA-DE-LR method aims to effectively navigate the search space and converge to an optimal set of weights for the logistic regression classifier, thereby enhancing classification accuracy and model robustness.

Algorithm 11: Calculate Fitness Function using Matthews Correlation Coefficient

procedure $CalculateFitnessMCC(\varphi)$
  $a \leftarrow CalculatePrediction(\varphi)$
  $p \leftarrow round(a)$
  $f \leftarrow MCC(\vec{y}M, p)$
  $evaluation\ number \leftarrow evaluation\ number + len(f)$_
  $return\ \ f$
end procedure

Algorithm 12: Calculate Fitness Function using F1 Score

procedure $CalculateFitnessF1(\varphi)$
  $a \leftarrow CalculatePrediction(\varphi)$
  $p \leftarrow round(a)$
  $f \leftarrow F1(\vec{y}M, p)$
  $evaluation\ number \leftarrow evaluation\ number + len(f)$_
  $return\ \ f$
end procedure

Algorithm 13: Calculate Fitness Function using Mean Absolute Error (MAE)

procedure $CalculateFitnessMAE(\varphi)$
  $p \leftarrow CalculatePrediction(\varphi)$
  $f \leftarrow MAE(\vec{y}M, p)$
  $f \leftarrow 1/(f+1)$
  $evaluation\ number \leftarrow evaluation\ number + len(f)$_
  $return\ \ f$
end procedure

Algorithm 14: Calculate Matthews Correlation Coefficient

procedure $MCC(actual,predicted)$
  $tp \leftarrow sum(predicted * actual, axis = 0)$
  $tn \leftarrow sum((1 - predicted) * (1 - actual)), axis = 0)$
  $fp \leftarrow sum(predicted, axis = 0) - tp$
  $fn \leftarrow sum(actual, axis = 0) - tp$
  $mcc \leftarrow (tp * tn - fp * fn)/(tp + fn) * (tp + fp) * (tn + fn) * (tn + fp)$
  $return\ \ mcc$
end procedure

Algorithm 15: Calculate F1 score

```
procedure F1(actual,predicted)
    tp ← sum(predicted ∗actual, axis = 0)
    fp ← sum(predicted, axis = 0) − tp
    fn ← sum(actual, axis = 0),tp
    precision ← tp/(tp + fp)
    recall ← tp/(tp + fn)
    F1 ← 2 ∗precision ∗recall/(precision + recall)
    return F1
end procedure
```

Algorithm 16: Calculate MAE

```
procedure MSE(actual,predicted)
    mae = mean((actual − predicted)^2, axis = 0)
    return mae
end procedure
```

8.4 Experiments

8.4.1 Datasets and Preprocessing

Four popular datasets—the Cleveland, Statlog, Breast Cancer Wisconsin Original (WBCO), and Breast Cancer Wisconsin Diagnostic (WBCD) datasets—were utilized for empirical analysis in this work. The UCI Machine Learning Repository makes all four of these datasets publically accessible, and they are often utilized in studies on medical categorization.

- The Cleveland dataset contains 303 instances, each described by 13 attributes. It is used to classify instances as indicative of CAD or representing a healthy state.
- The Statlog dataset consists of 270 instances, also described by 13 attributes. It is similar in structure to the Cleveland dataset and is used to classify the presence of CAD.
- The WBCO dataset comprises 699 instances, each described by 9 attributes based on biopsy data. This dataset is used to classify tumors as either benign or malignant.
- The WBCD dataset contains 569 instances, each described by 30 attributes. This dataset is used to classify tumors as benign or malignant.

Preprocessing the data was essential to ensuring accurate findings and uniform scaling. To ensure data integrity, any missing values were first found and fixed. Six

missing value cases were eliminated from the Cleveland dataset, while 16 missing value cases were eliminated from the WBCO dataset. There were no missing values in the Statlog or WBCD databases.

After addressing the missing data, the scaling process was carried out. For the Cleveland, Statlog, and WBCO datasets, the training data for each fold was normalized using the MinMaxScaler, which scales data values to the [0, 1] range. For the WBCD dataset, the StandardScaler was applied to normalize the data by centering and scaling based on the mean and standard deviation. The scalers were first fit and applied to the training data, and subsequently, the transformation was applied to the test data to ensure consistent scaling.

Following the scaling procedures, a 10-fold cross-validation process was employed to evaluate the model's performance. This involved dividing the data into 10 equally sized folds. Each fold was used as a test set while the remaining nine folds formed a training set, allowing the model to be trained and evaluated 10 separate times. The individual results from each fold were then averaged to provide a more comprehensive and reliable assessment of the model's overall performance.

8.4.2 Evaluation Metrics

Several important criteria were used in this study to assess and contrast the effectiveness of the categorization procedures. Among them are:

- Accuracy (ACC): This metric measures the ratio of correctly predicted observations to the total observations and the formula for ACC is provided by formula (8.10).

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \tag{8.10}$$

  where TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives.

- F1 Score: The F1 Score is a metric that balances precision (the quality of the positives identified) and recall (the ability to find all relevant instances). The F1 Score is calculated as:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{8.11}$$

where Precision = TP / (TP + FP), Recall = TP / (TP + FN). This metric is critical for understanding the model's accuracy in classifying positive cases.

- Matthews Correlation Coefficient (MCC): This coefficient is a reliable statistical rate which yields a value between -1 and +1. It is especially useful for imbalanced datasets. The formula for MCC is:

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (8.12)$$

- False Negative Rate (FNR) and False Positive Rate (FPR): Understanding the many kinds of mistakes the model makes depends on these rates. Whereas FPR gauges the percentage of negatively categorized instances that are mistakenly identified as positive, FNR gauges the rate at which positive cases are wrongly classified as negative. Their equations are:

$$FNR = \frac{FN}{TP+FN} \quad (8.13)$$

$$FPR = \frac{FP}{TN+FP} \quad (8.14)$$

- ROC-AUC Score: The model's capacity for class distinction is shown by the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC). A score of around 0.5 is little better than guesswork, but a score of close to 1 denotes excellent categorization.

8.4.3 Hyper-parameter optimization

Thoroughly optimizing the hyperparameters of every classification method—CSA-DE-LR, CSA-LR, DE-LR, and other well-known machine learning techniques like Decision Tree (DT), Linear Discriminant Analysis (LDA), Logistic Regression (LR), Multi-Layer Perceptron (MLP), Random Forest (RF), XGBoost, and Support Vector Machine (SVM)—was an essential part of the experimental design. Table 8.1 lists the hyperparameter ranges and optimal values for each classifier on the Cleveland, WBCD, Statlog, and WBCO datasets, which were attained after 300 tuning iterations using the Hyperopt [156] technique.

For the CSA-LR approach, the hyperparameters tuned included the lower and upper bounds (*lb* and *ub*), population size (*P*), number of clones (α), and receptor editing rate (*B*). The DE-LR method involved optimizing the lower and upper bounds (*lb* and *ub*), population size (*P*), scaling factor (*sf*), and crossover rate (*cr*). These adjustments ensured that both methods could operate efficiently within their designed optimization frameworks.

In the proposed CSA-DE-LR approach, a combination of hyperparameters was tuned, including thel ower and upper bounds (*lb* and *ub*), population size (*P*), number of clones (α), receptor editing rate (*B*), scaling factor (*sf*), and crossover rate (*cr*). These adjustments ensured that both methods could operate efficiently within their designed optimization frameworks.

Table 8.1 Hyperparameter ranges and the best values attained after 300 iterations for different classifiers using Cleveland and Statlog datasets.

Classifier	Parameter	Low	High	Statlog (Best)	Cleveland (Best)	WBCD (Best)	WBCO (Best)
DT	Min Samples Split	2	100	83	92	3	54
	Min Samples Leaf	1	100	65	76	4	12
LDA	Shrinkage	0	1	0.747	0.875	0.422	0.698
MLP	Learning Rate	10^{-8}	10^{-1}	0.028	0.290	0.261	0.331
	Number of Hidden Units	2	40	6	16	18	25
	Batch Size	1	1024	182	247	501	209
	Number of Epochs	1	50	22	31	5	33
RF	Number of Tress	1	200	75	122	172	168
SVM	C	0.001	1	0.014	0.577	0.953	0.772
XGBoost	Eta	0.1	1	0.222	0.948	0.997	0.935
	Depth	1	40	14	15	27	16
LR	C	10^{-4}	10^{4}	63576.24	19600.49	35937.61	13519.16
CSA-LR	*lb*	-64	-16	-60.839	-47.767	-44.305	-61.344
	ub	16	64	53.250	43.126	45.543	27.328
	P	10	80	74	61	73	20
	α	2	6	4	3	4	5
	B	0.05	0.2	0.050	0.162	0.104	0.194
DE-LR	*lb*	-64	-16	-55.968	-34.463	-53.381	-59.298
	ub	16	64	22.412	22.809	38.460	53.334
	P	10	80	53	29	70	50
	sf	0.01	2	0.940	1.790	0.116	0.257
	cr	0.01	1	0.755	0.432	0.651	0.884
CSA-DE-LR	*lb*	-64	-16	-48.353	-27.010	-36.035	-63.765
	ub	16	64	19.487	24.245	29.282	32.241
	P	10	80	78	27	37	10
	α	2	6	4	4	3	5
	B	0.05	0.2	0.198	0.165	0.132	0.053
	sf	0.01	2	0.670	0.070	0.167	1.610
	cr	0.01	1	0.577	0.554	0.333	0.958

In the proposed CSA-DE-LR approach, a combination of hyperparameters was tuned, including thel ower and upper bounds (*lb* and *ub*), population size (*P*), number of clones (α), receptor editing rate (*B*), scaling factor (*sf*), and crossover rate (*cr*). The optimization process balanced exploration and exploitation, finding the best parameter settings to maximize performance.

To manage tree depth and avoid overfitting, the DT concentrated on "Min Samples Split" and "Min Samples Leaf" for the other classifiers. The "Shrinkage" parameter was adjusted by LDA to increase generalization. LR optimized the regularization strength-controlling 'C' parameter.

The learning rate, batch size, number of hidden units, and number of epochs were among the hyperparameters in the MLP that were adjusted to find the optimal values. The main focus of optimization for RF was the quantity of trees in the forest.

The "C" parameter, which specifies the trade-off between accurately categorizing training points and smooth decision boundaries, has to be tuned for the SVM. Finally, in order to enhance predictive performance, XGBoost concentrated on "Eta" (learning rate) and "Depth" (tree depth).

8.5 Performance Results and Discussion

Different performance patterns that represent the differences between the Statlog and Cleveland datasets are revealed through a comparative examination of various optimization algorithms used to both datasets. As shown in Table 8.2, the F1-Optimization technique performs better for the Statlog dataset, especially in the areas of accuracy, F1 score, and ROC-AUC metrics. This indicates a well-balanced approach in terms of precision and recall, which is essential for achieving a harmonious balance in classification tasks where both aspects are equally critical. This result highlights the F1-Optimization's applicability for comparable data types by indicating that it can handle the Statlog dataset's unique distribution and nature.

However, the Cleveland dataset offers an alternative situation in which the MCC-Optimization approach performs better than the others, especially when looking at the MCC and ROC-AUC metrics. This highlights its effectiveness in dealing with potentially imbalanced data structures and its superior capability in distinguishing between classes with higher precision. The better results of MCC optimization in this

situation emphasize how crucial it is to select an optimization approach that is in line with the intrinsic properties of the dataset. This difference in how well optimization techniques performed across the two datasets highlights the need for a customized strategy in machine learning applications that takes into account the particular characteristics of each dataset.

Table 8.2 A Comparative Study of the Proposed Method's Optimization Strategies (F1-Opt, MAE-Opt, and MCC-Opt) Using 10-Fold Cross Validation on the Statlog and Cleveland Datasets. Performance Metrics: ACC, F1 Score, MCC, ROC-AUC Score, FNR, and FPR with Stand Deviations (Std)

	Statlog			**Cleveland**		
Criteria	F1-Opt	MAE-Opt	MCC-Opt	F1-Opt	MAE-Opt	MCC-Opt
ACC±Std	**88.15±0.039**	87.78±0.040	87.04±0.041	86.00±0.053	85.67±0.073	**86.67±0.059**
F1±Std	**86.73±0.049**	84.76±0.062	83.98±0.062	84.44±0.061	82.87±0.101	**84.64±0.066**
MCC±Std	**76.74±0.075**	75.83±0.081	74.46±0.082	71.79±0.105	71.76±0.145	**74.32±0.115**
ROC-AUC±Std	**88.42±0.039**	87.20±0.048	86.50±0.048	85.67±0.053	85.26±0.076	**86.52±0.056**
FNR±Std	**0.099±0.071**	0.199±0.109	0.199±0.118	**0.161±0.083**	0.209±0.144	0.191±0.105
FPR±Std	**0.137±0.046**	0.057±0.043	0.070±0.054	0.125±0.065	0.085±0.055	**0.077±0.079**

The importance of context-dependent strategy selection in machine learning efforts is highlighted by these findings. The differences in performance between the two datasets show that there is no one-size-fits-all method and that, in order to get the best results, each dataset's unique characteristics must be carefully taken into account. Building strong and efficient machine-learning models requires a sophisticated grasp of the interactions between dataset properties and optimization techniques.

The classification methods assessment on the Cleveland and Statlog datasets, as shown in Tables 8.3 and 8.4, provides an extensive overview of the performance of several widely used classification strategies. The suggested CSA-DE-LR technique sticks out in this context as a cutting-edge strategy, piqueing interest. Metrics including ACC, F1-score, MCC, ROC-AUC score, FNR, and FPR were used to evaluate the performance. The results of 10-fold cross-validation were collected, and each metric's mean and standard deviation (Std) were provided. The results indicate that the CSA-DE-LR technique achieved the greatest ACC, F1, MCC, and ROC-AUC scores in the Statlog dataset (Table 8.3), significantly outperforming the other methods across all parameters. This clear superiority of CSA-DE-LR, particularly in significantly lowering the FNR to 0.099, instills confidence in its effectiveness, making it an efficient choice for the Statlog dataset.

The CSA-DE-LR approach also performed better on the Cleveland dataset (Table 8.4), producing higher ACC, F1, MCC, and ROC-AUC scores. For this dataset, CSA-DE-LR proved to be the most successful approach despite having a slightly higher FNR than XGBoost. This was due to its exceptional performance in lowering the FPR. The fact that CSA-DE-LR constantly performs well on both datasets highlights how well it can handle the variety of features included in data on heart disease. Other categorization techniques, however, demonstrated differing degrees of efficacy.

Table 8.3 The comparative performance of other machine learning techniques and the suggested approach (CSA-DE-LR) on the Statlog Dataset was assessed, using metrics like ACC, F1 Score, MCC, ROC-AUC Score, FNR, FPR, and training time in seconds (Time). The fin findings were obtained using 10-fold cross-validation.

Method	ACC±Std	F1±Std	MCC±Std	ROC-AUC±Std	FNR±Std	FPR±Std	Time±Std
DT	80.37 ± 0.057	76.67 ± 0.078	61.20 ± 0.125	80.24 ± 0.062	0.254 ± 0.111	0.140 ± 0.094	0.001 ± 0.000
LDA	85.18 ± 0.043	82.52 ± 0.058	70.82 ± 0.084	84.95 ± 0.044	0.190 ± 0.102	0.110 ± 0.079	0.001 ± 0.000
MLP	85.55 ± 0.034	82.99 ± 0.049	71.17 ± 0.069	85.08 ± 0.039	0.183 ± 0.107	0.114 ± 0.064	0.004 ± 0.000
RF	84.81 ± 0.074	81.72 ± 0.091	70.00 ± 0.147	84.48 ± 0.076	0.213 ± 0.122	0.096 ± 0.086	0.086 ± 0.001
XGBoost	83.70 ± 0.052	81.05 ± 0.068	67.87 ± 0.109	83.78 ± 0.054	0.197 ± 0.093	0.127 ± 0.086	0.048 ± 0.021
SVM	83.70 ± 0.041	79.71 ± 0.058	68.55 ± 0.075	83.21 ± 0.042	0.260 ± 0.109	**0.075 ± 0.074**	0.005 ± 0.000
LR	83.33 ± 0.041	80.50 ± 0.054	66.88 ± 0.087	83.03 ± 0.044	0.206 ± 0.103	0.132 ± 0.075	0.002 ± 0.001
CSA-LR	86.30 ± 0.060	83.22 ± 0.067	72.78 ± 0.108	86.29 ± 0.063	0.205 ± 0.116	0.069 ± 0.061	10.36 ± 0.116
DE-LR	86.30 ± 0.055	84.27 ± 0.064	72.94 ± 0.099	86.84 ± 0.055	0.118 ± 0.106	0.145 ± 0.098	9.175 ± 0.596
CSA-DE-LR	**88.15 ± 0.039**	**86.73 ± 0.049**	**76.74 ± 0.075**	**88.42 ± 0.039**	**0.099 ± 0.071**	0.137 ± 0.046	0.494 ± 0.025

But they fell short of CSA-DE-LR's balanced performance, especially when it came to obtaining low false negative rates without appreciably raising false positive rates. These results, which are presented in Tables 8.3 and 8.4, emphasize how crucial it is to include advanced optimization methods in logistic regression models, such as CSA and DE. Because of its accuracy and versatility, CSA-DE-LR might potentially improve diagnostic decision-making processes in medical diagnostic applications.

The performance of the suggested technique was assessed on the Breast Cancer datasets (WBCO and WBCD) to show its generalizability and offer further validation, even though the research's main focus was on CAD utilizing the Statlog and Cleveland datasets. As shown in Table 8.5, the various optimization procedures (F1-Opt, MAE-Opt, and MCC-Opt) yielded diverse outcomes when assessing the WBCO and WBCD datasets. The MCC-Opt method performed the best for the WBCD dataset in terms of ACC (98.93%), F1 (98.57%), MCC (97.76%), and ROC-AUC (98.68%), among other parameters. Additionally, it demonstrated high classification skills by minimizing the False Positive Rate (FPR) at 0.003 and the False Negative Rate (FNR) at 0.024.

Table 8.4 Using metrics such as ACC, F1 Score, MCC, ROC-AUC Score, FNR, FPR, and training time in seconds (Time), based on 10-fold cross-validation results, a comparison is made between the suggested approach and a number of well-known classifiers on the Cleveland dataset.

Method	ACC±Std	F1±Std	MCC±Std	ROC-AUC±Std	FNR±Std	FPR±Std	Time±Std
DT	81.66 ± 0.060	78.19 ± 0.094	63.14 ± 0.130	80.96 ± 0.067	0.238 ± 0.149	0.142 ± 0.063	0.001 ± 0.000
LDA	85.33 ± 0.068	83.31 ± 0.079	71.05 ± 0.132	85.04 ± 0.067	0.189 ± 0.105	0.109 ± 0.092	0.001 ± 0.000
MLP	85.66 ± 0.068	83.35 ± 0.088	71.78 ± 0.130	85.35 ± 0.069	0.187 ± 0.136	0.104 ± 0.064	0.005 ± 0.002
RF	85.33 ± 0.061	83.34 ± 0.067	71.05 ± 0.123	84.72 ± 0.058	0.197 ± 0.100	0.107 ± 0.091	0.027 ± 0.001
XGBoost	86.00 ± 0.051	84.14 ± 0.055	72.43 ± 0.099	85.74 ± 0.047	**0.180** ± **0.087**	0.104 ± 0.079	0.013 ± 0.001
SVM	83.33 ± 0.066	80.40 ± 0.092	66.90 ± 0.132	82.78 ± 0.068	0.228 ± 0.123	0.115 ± 0.091	0.005 ± 0.000
LR	83.00 ± 0.078	80.78 ± 0.087	66.50 ± 0.145	82.75 ± 0.075	0.205 ± 0.115	0.139 ± 0.119	0.002 ± 0.001
CSA-LR	84.00 ± 0.047	81.98 ± 0.055	68.40 ± 0.088	83.86 ± 0.045	0.184 ± 0.094	0.139 ± 0.082	9.107 ± 0.124
DE-LR	83.33 ± 0.061	82.08 ± 0.065	67.14 ± 0.123	83.31 ± 0.058	0.156 ± 0.079	0.178 ± 0.099	8.755 ± 0.207
CSA-DE-LR	**86.67** ± **0.059**	**84.64** ± **0.066**	**74.32** ± **0.115**	**86.52** ± **0.056**	0.191 ± 0.105	**0.077** ± **0.079**	0.720 ± 0.033

Table 8.5 Comparative Analysis of Optimization Strategies (F1-Opt, MAE-Opt, and MCC-Opt) of the Proposed Method on WBCD and WBCO Datasets Using 10-Fold Cross Validation. Performance metrics include training time (Time) in

seconds with standard deviations (Std), ROC-AUC, FNR, FPR, ACC, F1, MCC, and FNR.

	WBCD			WBCO		
Criteria	F1-Opt	MAE-Opt	MCC-Opt	F1-Opt	MAE-Opt	MCC-Opt
ACC±Std	98.21 ± 0.016	98.39 ± 0.015	**98.93 ± 0.010**	97.65 ± 0.016	**97.94 ± 0.015**	97.65 ± 0.013
F1±Std	97.84 ± 0.017	97.74 ± 0.022	**98.57 ± 0.013**	96.54 ± 0.028	**96.93 ± 0.026**	96.64 ± 0.020
MCC±Std	95.95 ± 0.035	96.52 ± 0.033	**97.76 ± 0.019**	94.88 ± 0.037	**95.49 ± 0.035**	94.92 ± 0.029
ROC-AUC±Std	98.08 ± 0.016	97.98 ± 0.019	**98.68 ± 0.015**	98.05 ± 0.013	**98.28 ± 0.012**	97.90 ± 0.011
FNR±Std	0.036 ± 0.034	0.035 ± 0.033	**0.024 ± 0.027**	**0.008 ± 0.016**	**0.008 ± 0.016**	0.015 ± 0.019
FPR±Std	0.003 ± 0.008	0.006 ± 0.011	**0.003 ± 0.007**	0.031 ± 0.024	**0.027 ± 0.023**	0.027 ± 0.024
Time±Std	1.208 ± 0.055	**1.028 ± 0.035**	1.260 ± 0.056	**1.242 ± 0.042**	1.366 ± 0.054	1.674 ± 0.062

The MAE-Opt method was the top performer for the WBCO dataset, with excellent scores in ROC-AUC (98.28%), F1 (96.93%), MCC (95.49%), and ACC (97.94%). It also had the lowest FNR (0.008) and FPR (0.027), demonstrating how well it could identify breast cancer data. Additionally, the MAE-Opt strategy's computational efficiency was highlighted by the fact that it took the least amount of time to train on the WBCD dataset.

These results highlight how crucial it is to use customized optimization techniques to take into account the unique features of every dataset. While MCC-Opt has the greatest classification performance for the WBCD dataset, MAE-Opt offers a solid balance between accuracy and processing efficiency for the WBCO dataset. In the end, this comparative research highlights the importance of a context-driven approach by showing how the right optimization method may have a substantial influence on a machine learning model's performance.

The performance of the suggested CSA-DE-LR model on the WBCD dataset is obviously better than that of other classifiers, as shown in Table 8.6. With an accuracy of 98.93% and an F1 score of 98.57%, the model demonstrates its remarkable performance across key measures, indicating its ability to effectively categorize the benign and malignant classifications. With scores of 97.76% and 98.68% for MCC and ROC-AUC, respectively, these metrics demonstrate the model's strong predictive capacity and capacity to distinguish between classes. Furthermore, the model minimizes errors and lowers the chance of inaccurate classification with a FNR of 0.024 and an FPR of 0.003. Its effectiveness is demonstrated by the 1.26 second training period, which it maintains despite the great accuracy.

Table 8.6 Comparison of the proposed method CSA-DE-LR with LR, CSA-LR, DE-LR, and several popular classifiers on the WBCD dataset, measured using metrics like ACC, F1 Score, MCC, ROC-AUC Score, FNR, FPR, and training time in seconds (Time), based on 10-fold c ross-validation results.

Method	ACC±Std	F1±Std	MCC±Std	ROC-AUC±Std	FNR±Std	FPR±Std	Time±Std
DT	93.92 ± 0.026	91.81 ± 0.032	87.40 ± 0.051	93.17 ± 0.031	0.105 ± 0.068	0.032 ± 0.034	0.004 ± 0.001
LDA	96.07 ± 0.021	94.46 ± 0.028	91.79 ± 0.040	94.97 ± 0.025	0.098 ± 0.051	**0.003 ± 0.007**	**0.001 ± 0.000**
MLP	98.04 ± 0.017	97.33 ± 0.023	95.88 ± 0.035	97.82 ± 0.021	0.032 ± 0.043	0.011 ± 0.019	0.011 ± 0.005
RF	96.79 ± 0.016	95.64 ± 0.023	93.30 ± 0.033	96.49 ± 0.021	0.047 ± 0.053	0.023 ± 0.029	0.150 ± 0.005
XGBoost	97.50 ± 0.016	96.51 ± 0.024	94.68 ± 0.035	96.92 ± 0.022	0.053 ± 0.046	0.009 ± 0.013	0.026 ± 0.003
SVM	97.32 ± 0.022	96.27 ± 0.029	94.21 ± 0.045	96.78 ± 0.024	0.053 ± 0.036	0.011 ± 0.014	0.011 ± 0.000
LR	96.25 ± 0.036	95.03 ± 0.044	92.05 ± 0.074	95.94 ± 0.036	0.052 ± 0.042	0.029 ± 0.039	0.016 ± 0.011
CSA-LR	97.67 ± 0.022	96.82 ± 0.033	95.03 ± 0.050	97.33 ± 0.027	0.038 ± 0.045	0.014 ± 0.019	16.89 ± 0.073
DE-LR	98.03 ± 0.014	97.23 ± 0.022	95.77 ± 0.032	97.50 ± 0.018	0.044 ± 0.034	0.005 ± 0.011	19.59 ± 0.320
CSA-DE-LR	**98.93 ± 0.010**	**98.57 ± 0.013**	**97.76 ± 0.019**	**98.68 ± 0.015**	**0.024 ± 0.027**	**0.003 ± 0.007**	1.260 ± 0.056

Table 8.7 Comparison of the proposed method CSA-DE-LR with LR, CSA-LR, DE-LR, and several popular classifiers on the WBCO dataset, measured using metrics like ACC, F1 Score, MCC, ROC-AUC Score, FNR, FPR, and training time in seconds (Time), based on 10-fold cross-validation results.

Method	ACC±Std	F1±Std	MCC±Std	ROC-AUC±Std	FNR±Std	FPR±Std	Time±Std
DT	95.00 ± 0.029	92.79 ± 0.042	89.12 ± 0.061	94.82 ± 0.026	0.059 ± 0.033	0.045 ± 0.044	**0.001 ± 0.000**
LDA	96.18 ± 0.027	94.62 ± 0.035	91.76 ± 0.057	95.57 ± 0.030	0.070 ± 0.049	**0.018 ± 0.023**	0.001 ± 0.000
MLP	97.65 ± 0.015	96.60 ± 0.022	94.86 ± 0.033	97.75 ± 0.015	0.020 ± 0.027	0.025 ± 0.021	0.007 ± 0.004
RF	97.50 ± 0.015	96.32 ± 0.025	94.53 ± 0.034	97.72 ± 0.014	0.019 ± 0.026	0.026 ± 0.023	0.017 ± 0.000
XGBoost	96.91 ± 0.019	95.54 ± 0.028	93.25 ± 0.041	97.03 ± 0.020	0.031 ± 0.033	0.029 ± 0.022	0.036 ± 0.004
SVM	97.21 ± 0.021	95.96 ± 0.032	93.87 ± 0.047	97.37 ± 0.021	0.024 ± 0.026	0.029 ± 0.024	0.006 ± 0.000
LR	96.62 ± 0.023	95.26 ± 0.030	92.66 ± 0.048	96.36 ± 0.025	0.048 ± 0.039	0.025 ± 0.022	0.002 ± 0.002
CSA-LR	97.35 ± 0.021	96.10 ± 0.033	94.24 ± 0.047	97.40 ± 0.022	0.027 ± 0.040	0.024 ± 0.028	11.60 ± 0.098
DE-LR	97.35 ± 0.021	96.15 ± 0.033	94.27 ± 0.046	97.75 ± 0.018	0.011 ± 0.017	0.033 ± 0.029	9.600 ± 0.085
CSA-DE-LR	**97.94 ± 0.015**	**96.93 ± 0.026**	**95.49 ± 0.034**	**98.28 ± 0.011**	**0.007 ± 0.015**	0.026 ± 0.023	1.366 ± 0.054

Comparably, Table 8.7 offers information about how well CSA-DE-LR and other classifiers performed using the WBCO dataset. With an accuracy of 97.94% and an F1 score of 96.93%, the suggested model demonstrates its strength in categorization. The ROC-AUC of 98.28% and MCC of 95.49% demonstrate how well the model can distinguish between the two groups. Furthermore, its dependability in reducing classification mistakes is demonstrated by a low FNR of 0.007 and an FPR of 0.026. Even with its extensive functionality, CSA-DE-LR still manages to retain an effective training time of 1.366 seconds.

Overall, Tables 8.3, 8.4, 8.6, and 8.7 data show that the CSA-DE-LR model combines the advantages of DE and CSA optimization techniques with logistic regression to provide a significant improvement over previous approaches. It is a useful tool for medical diagnostics because of its continuously strong performance over a wide range of measures, which guarantees accurate and dependable categorization.

To assess the statistical significance of performance differences between CSA-DE-LR and the other classifiers, Wilcoxon signed-rank tests were used in this investigation. Thirty tests were conducted on four datasets for each classifier, employing the hyperparameters that yielded the greatest results in the past.

Table 8.8 Wilcoxon test results indicating p-values for comparisons between CSA-DE-LR and other classifiers across multiple datasets.

	Cleveland			Statlog			WBCD			WBCO		
Classifier	ACC	F1	MCC	ACC	F1	MCC	ACC	F1	MCC	ACC	F1	MCC
LR	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09
CSA-LR	1.86e-09	1.30e-08	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09
DE-LR	1.86e-09	3.73e-09	1.86e-09	2.51e-06	1.86e-09	3.73e-09	9.76e-06	8.01e-08	8.01e-08	2.53e-06	3.73e-09	3.73e-09
MLP	9.31e-09	1.30e-08	9.31e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09
RF	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09
XGBoost	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09
DT	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09
SVC	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09
LDA	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09	1.86e-09

The final findings are provided as averages for Accuracy (ACC), F1 score (F1), and Matthews Correlation Coefficient (MCC). The performance results were computed using 10-fold cross-validation. Table 8.8 shows that the suggested CSA-DE-LR technique outperforms other classifiers in classification across four datasets in a consistent and statistically significant way. The Wilcoxon signed-rank tests, which consistently produced extremely low p-values—generally regarded as significant at less than 0.05—across all three assessment measures are the source of this finding. The statistical significance of the performance differences between CSA-DE-LR and the

other classifiers is indicated by the low p-values, which suggest that the disparities cannot be the result of chance.

CSA-DE-LR outperformed other classifiers in the Cleveland dataset with statistically significant gains; p-values indicated a substantial difference. CSA-DE-LR had a large performance advantage over other approaches in the Statlog dataset, as confirmed by the Wilcoxon tests, which also revealed significant variations in performance. Comparable patterns were seen in the WBCD dataset, where CSA-DE-LR routinely outperformed other classifiers and proved its efficacy in all three parameters. Lastly, CSA-DE-LR continued to outperform in the WBCO dataset, showing notable gains in all assessment parameters.

These findings show that, in a variety of data contexts, the suggested CSA-DE-LR technique consistently outperforms other classifiers in terms of classification performance. The statistically significant differences indicated by the low p-values suggest that CSA-DE-LR is a dependable and efficient classification technique that performs better than other available classifiers in most cases.

With its many optimization choices, the suggested technique shows strong and dependable classification performance; yet, there is still a great deal of room for improvement when it comes to improving its classification performance on the Cleveland and Statlog datasets. Regarding this, a thorough investigation of the models that yield the best outcomes using F1-Opt and MCC-Opt on the Statlog and Cleveland datasets, together with an analysis of their corresponding weights, may provide priceless information for feature selection and model improvement. The average feature weights from a 10-fold cross-validation of the top-performing models for both datasets are shown in Figure 8.1. In the context of cardiac disease, the comparison analysis shown in Figure 8.1 emphasizes the predictive ability of specific clinical factors. Key characteristics that show significant positive weights in both datasets are 'ca' (number of major vessels colored by fluoroscopy), 'thal' (thalassemia), and 'cp' (chest pain type). This indicates the essential role these parameters play in predicting cardiovascular events. These traits, independent of patient cohort or dataset characteristics, have been repeatedly recognized as critical elements in the diagnosis of heart disease. On the other hand, 'thalach' displays a statistically significant negative weight, indicating an inverse relationship with the target variable.

Nonetheless, it's interesting to see that some traits show different weights in the two datasets. Statlog's 'age' variable has a positive weight, indicating a direct association

with heart disease; in Cleveland, however, it has a negative weight, showing a less direct or even inverse relationship in that particular patient population. Likewise,'sex' has a higher weight in Statlog than it does in Cleveland.

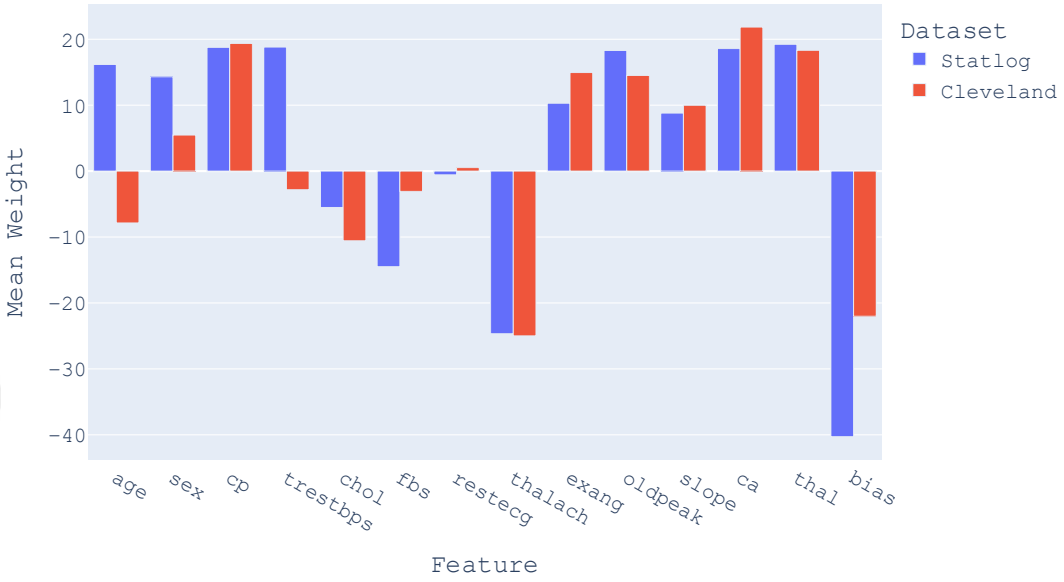

Figure 8.1 Mean weight of each feature for the Statlog and Cleveland datasets

These disparities may result from different patterns of illness presentation across the groups or from demographic variations in the datasets.

Additionally, 'trestbps' (resting blood pressure) has a significant variation in Cleveland, with a negative weight and a large positive weight in Statlog. This implies that, depending on the dataset, the same clinical measurement may have a varied predictive value. This might be due to factors such as the feature's underlying distribution, how it interacts with other variables, or health patterns that are unique to a certain community.

Additionally, 'restecg' (resting electrocardiographic results) exhibits a weight close to zero in both datasets, implying its limited predictive value in the context of these datasets. This discovery is consistent with the principle of parsimony, indicating that eliminating this variable may lower the model's complexity without a major loss of information. This kind of simplification might enhance the model's interpretability and generalizability, facilitating its application in clinical settings.

The negative weights for features such as 'chol' (serum cholesterol) and 'fbs' (fasting blood sugar) in both datasets challenge common assumptions about the role of these factors in heart disease, prompting a re-evaluation of their predictive significance. This

could reflect the multifactorial nature of heart disease, where the relevance of certain risk factors may be diminished or outweighed by others in specific populations.

In conclusion, this analysis underlines the necessity of dataset-specific model tuning and the careful consideration of feature selection based on their differential impact across datasets. It calls for a detailed knowledge of the contribution of each clinical condition and emphasizes the significance of context in the development of prediction models for heart disease.

Nonetheless, inconsistencies between equivalent features across the Statlog and Cleveland datasets necessitate a more nuanced examination. Accordingly, Figures 8.2 and 8.3 illustrate the fold-specific weights for each feature within the Statlog and Cleveland datasets, respectively. Under normal circumstances, one would anticipate fold weights for the same attribute to exhibit similar directionality and closely clustered values. As depicted in Figure 8.2, such consistency is largely maintained for the Statlog dataset, with the notable exception of the 'restecg' attribute.

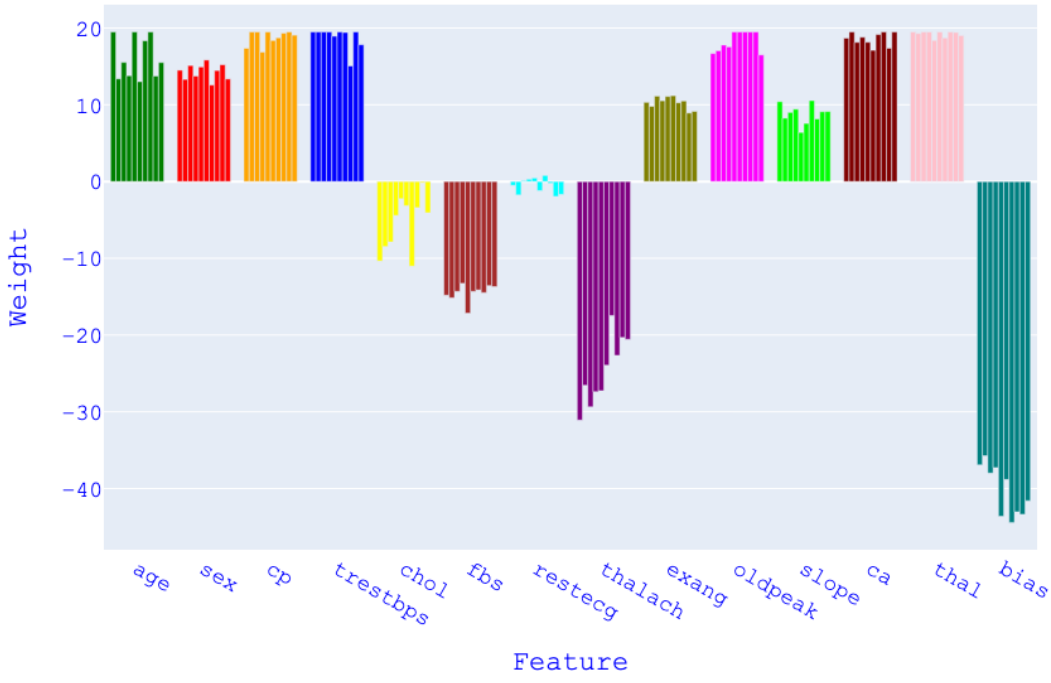

Figure 8.2 Fold-specific weights for each feature of the best-performing model on the Statlog dataset.

These discoveries prompted a rigorous reevaluation that involved removing the 'trestbps', 'fbs', and 'restecg' features from the Cleveland dataset and the 'restecg' feature from the Statlog dataset. This was done to see what affects it would have on

classification performance and to see whether removing these factors would result in a more predictively consistent and broadly applicable model. The objective of this study is to improve the robustness of the model by removing variables that show a high degree of fluctuation and do not consistently add to the accuracy of predictions.

Figure 8.3 Fold-specific weights for each feature of the best-performing model on the Cleveland dataset.

This approach is especially useful in the field of medical diagnostics, where a predictive model's interpretability and dependability are critical.

Following this strategic reevaluation and feature omission, the performance of the optimized models was tested: CSA-DE-LR with F1-Opt for the Statlog dataset and CSA-DE-LR with MCC-Opt for the Cleveland dataset. These models were executed using the same hyperparameter settings that previously yielded optimal results. Table 8.9 provides specifics on the results of this procedure. By using these models with improved feature sets, researchers hoped to learn more about how successful the feature selection strategy was. This stage, which primarily focused on predicted accuracy and consistency, was essential in figuring out how feature removal affected the overall performance of the model.

The best outcomes from using feature selection are shown in Table 8.9, which compares the original models directly. This comparison not only demonstrates the demonstrable advantages of feature selection in improving CSA-DE-LR performance, but it also emphasizes the usefulness of feature selection.

Table 8.9 Enhancing Diagnostic Performance: A Comparative Analysis of the CSA-DE-LR Method with Feature Selection on Cleveland and Statlog Datasets.

Criteria	Statlog (F1-Opt)	Cleveland (MCC-Opt)
ACC±Std	88.15±0.027	88.00±0.049
F1±Std	86.86±0.037	86.08±0.059
MCC±Std	76.63±0.054	76.17±0.100
ROC-AUC±Std	88.44±0.027	87.54±0.049
FNR±Std	0.099±0.071	0.175±0.089
FPR±Std	0.132±0.051	0.074±0.050

Accuracy, F1 score, MCC, and ROC-AUC increases for both datasets demonstrate how much better the technique can identify and categorize instances after feature selection. These improvements highlight the practical significance of our study, especially in medical diagnostic applications where accuracy is critical. Moreover, the consistency of FNR and FPR rates shown after feature selection in both datasets validates the CSA-DE-LR method's dependability and adds to its practicality and use.

To summarize, Table 8.9's findings confirm that the suggested method's feature selection is effective and emphasize how important it is to strike a balance between model complexity and performance. In the realm of medical diagnostics, maintaining the model's accuracy and interpretability is crucial, and this balance is necessary to make sure of that.

Lastly, the suggested approach is contrasted with findings from earlier research. The Statlog and Cleveland datasets have been used in many research, but the 10-fold cross-validation method has been applied in very few of them. To ensure a fair comparison, studies that used similar preprocessing techniques and 10-fold cross-validation were chosen. These studies included the following examples: i) using the highest value obtained after cross-validation rather than the mean value of all results [128], ii) displaying resampled results [129], and iii) placing the obtained results on the train set rather than the test set. As shown in Table 8.10, a comparison analysis with these studies not only highlights the efficacy of the CSA-DE-LR approach but also makes a strong case for its uniqueness and superiority.

Slightly improved ACC and F1 scores (88.0% and 86.08%, respectively) are obtained for the Cleveland dataset using 10-fold cross-validation-optimized CSA-DE-LR. This outperforms the outcomes of earlier techniques such as MGOHBO-KELM

(2022), NN-DEGI-BP (2016), several MLP implementations (2019–2023), Ensemble (2020), and PSO-EmNN (2020).

Table 8.10 A Review of CSA-DE-LR Performance Using Cleveland and Statlog Heart Disease Datasets and Historical Comparison with Other Research Studies.

Dataset	Method	ACC (%)	F1 (%)	K-Fold CV	Paper
Cleveland	NN-DEGI-BP	86.66	-	10	Leema et al. (2016) [134]
	MLP	82.50	83.80	10	Kolukısa et al. (2019) [123]
	Ensemble	83.43	81.10	10	Kolukisa et al. (2020) [124]
	PSO-EmNN	84	82.29	10	Shahid and Singh (2020) [137]
	MLP-PSO	84.60	84.40	5	Al Bataineh and Manacek (2022) [138]
	MGOHBO-KELM	82.22	-	10	Shan et al. (2022) [157]
	MLP	85.47	83.90	10	Kolukisa and Bakir-Gungor (2023) [122]
	CSA-DE-LR	**88.00**	**86.08**	10	**Proposed Method (2024)**
Statlog	PSO-EmNN	85.20	84	10	Shahid and Singh (2020) [137]
	MGOHBO-KELM	81.85	-	10	Shan et al. (2022) [157]
	MLP	85.55	85.30	10	Kolukisa and Bakir-Gungor (2023) [122]
	LR	85.2	-	10	Dhanka et al. (2023) [125]
	XGBoost	81.5	-	10	Dhanka et al. (2023) [125]
	CSA-DE-LR	**88.15**	**86.86**	10	**Proposed Method (2024)**

The improvement is especially noticeable when compared to the most current MLP approach from 2023, highlighting the improvements in classification accuracy and precision produced by CSA-DE-LR.

Similarly, CSA-DE-LR performs better on the Statlog dataset than all other specified algorithms, such as XGBoost (2023), MLP (2023), MGOHBO-KELM (2022), and PSO-EmNN (2020). The effectiveness of the suggested strategy is further validated by the CSA-DE-LR method, which shows its superiority in this dataset with an ACC of 88.15% and an F1 score of 86.86%.

With these 2024 results, CSA-DE-LR is positioned as a front-runner in the categorization of cardiac disease. This method's high accuracy, F1-score, and ROC-AUC suggest that it may be able to identify heart disease cases with lower mistake rates. CSA-DE-LR achieves low False Negative and False Positive rates, improving patient outcomes by lowering unnecessary medications and misdiagnoses. More targeted and efficient diagnostic methods are made possible by the feature weights analysis, which offers insights into significant clinical factors. These components suggest that CSA-DE-LR might be a useful tool in medical contexts by accelerating diagnostic processes and improving resource allocation, thereby saving healthcare costs, coupled with the method's durability and adaptability.

Chapter 9

Integration of the CSA-DE-LR with AguHyper

9.1 Machine and Deep Learning-Enabled Blockchain Technologies for EHR sharing and Disease Prediction

Blockchain technology, together with ML and deep learning (DL), is revolutionizing illness prediction and EHR exchange. A strong ecosystem for safe data transmission and advanced predictive analytics is created by this junction [158-160]. An outline of how these technologies affect illness prediction and EHR sharing is provided below:

9.1.1 Blockchain for EHR Sharing

Blockchain provides a secure, decentralized platform for EHR sharing. It guarantees data security and privacy while facilitating data sharing between patients, healthcare professionals, and researchers [161]. Among the essential components are:

- **Secure Decentralized Ledger:** The immutable ledger of blockchain technology guarantees that data cannot be changed once it is saved, lowering the possibility of manipulation and unauthorized access.
- **Smart Contracts for Access Control:** A customizable and automated method for controlling permissions is made possible by smart contracts on the blockchain, which may provide guidelines for who is allowed access to EHRs and under what circumstances.

- **Interoperability and Standardization:** Blockchain can serve as a common framework for sharing EHRs across different healthcare systems, promoting interoperability and reducing data silos.

9.1.2 Machine Learning for Disease Prediction

Large-scale health data may be analyzed using machine learning and deep learning algorithms to find patterns and forecast the course of diseases. These technologies, when coupled with blockchain, provide a scalable and safe platform for sophisticated analytics. Principal advantages consist of [162]:

- **Predictive analytics:** Using past electronic health records, genetic data, lifestyle characteristics, and other pertinent information, machine learning algorithms can forecast the chance of developing certain diseases. Healthcare professionals may now proactively manage patient care because to this skill.
- **Deep Learning for complicated Patterns:** Deep learning is a branch of machine learning that handles high-dimensional, complicated data, such genome sequences or medical imaging. This makes illness prediction more precise and thorough.
- **Personalized Healthcare:** By utilizing machine learning, medical professionals may give patients with individualized treatment regimens and preventative measures. This strategy lowers healthcare expenses while improving patient outcomes.
- **Anomaly Detection and Early Warning Systems:** By identifying odd patterns in EHRs, deep learning can assist in the early detection of health hazards. Early interventions and increased patient safety may result from this.

9.1.3 Blockchain and Machine/Deep Learning Integration

Blockchain technology combined with deep learning and machine intelligence enables a safe space for exchanging electronic health records and advanced illness prediction. This is the operation of this integration [163]:

- **Secure Data Sharing for ML/DL Models:** Blockchain ensures that EHRs are shared securely, providing a reliable source of data for machine and deep learning models without compromising patient privacy.

- **Model Provenance and Auditability:** Machine and deep learning models have an audit trail thanks to blockchain's immutability. This guarantees accountability in healthcare decision-making and guarantees transparency in model development.
- **Real-Time Disease Prediction:** By combining blockchain technology with machine learning, healthcare professionals will be able to immediately respond to new health hazards by receiving forecasts and insights in real-time.

9.1.4 Challenges and Considerations

The integration of blockchain technology with machine and deep learning capabilities for illness prediction and EHR sharing poses several obstacles, including the requirement for strong data governance, computational resource needs, and regulatory compliance. Additionally, maintaining patient confidence and adhering to privacy legislation depends on assuring the ethical use of AI in healthcare.

9.2 Federated Learning-Enabled Blockchain Technologies for EHR sharing and Disease Prediction

Blockchain technology with federated learning (FL) capabilities provide a novel way to share EHRs and forecast illness. Distributed machine learning is made possible by blockchain technology and federated learning, all while maintaining data confidentiality and privacy. Healthcare providers can work together on machine learning models with this combination without exchanging raw patient data [158-160]. The influence of blockchain technology provided by federated learning on EHR sharing and illness prediction is as follows:

9.2.1 Federated Learning for Secure Collaborative Training

Without transferring their data, many parties can work together to train machine learning models thanks to FL. Using their EHRs, each party trains a local model; only the model updates (weights and biases, for example) are sent to a central server. This

method permits collaborative learning across decentralized data sources while protecting the privacy of personal information [164].

9.2.2 Blockchain for Data Security and Model Provenance

Blockchain provides a secure, decentralized platform for FL. It ensures the integrity of the model updates and allows for a transparent audit trail. Here's how blockchain enhances federated learning:

- **Immutable Ledger:** Blockchain technology offers a tamper-proof audit trail for collaborative learning by guaranteeing that model alterations recorded are unchangeable.
- **Smart Contracts for Model Governance:** FL rules, including who may take part, how updates are verified, and how models are combined, can be specified by smart contracts running on the blockchain. This encourages a governance structure that is automated and transparent.
- **Data Security and Privacy:** By keeping raw data local and leveraging blockchain technology, federated learning addresses privacy issues by preventing patient electronic health records from being accessed by unauthorized parties**.**

9.2.3 Federated Learning-Enabled Disease Prediction

Federated learning, with its focus on privacy, can be used to build collaborative models for disease prediction without compromising patient data. How it helps with illness prediction is as follows [165]:

- **Distributed Learning for Robust Models:** By using FL, healthcare practitioners may make contributions to a common model while maintaining the privacy of sensitive EHRs. Better illness prediction is the outcome, as more robust models trained on a variety of datasets are produced.
- **Personalized Healthcare:** Without centralizing patient data, FL can assist in the development of models for tailored healthcare. By using local EHRs to train models, healthcare professionals may get personalized forecasts for each patient.
- **Anomaly Detection and Early Warning:** Models enabled by FL are able to identify abnormalities and peculiar patterns across decentralized electronic

health records, therefore offering early warning indicators of possible health hazards.

9.2.4 Blockchain and Federated Learning Integration

FL and BC combine to provide a safe space for exchanging collaborative EHRs and illness prediction. This is the operation of this integration [166]:

- **Secure Model Aggregation:** Blockchain technology guarantees a safe and impenetrable federated learning model update aggregation. This offers a solid foundation for group model training.
- **Interoperability and Data Sharing:** Blockchain enables data sharing and interoperability across various healthcare providers, enabling them to exchange model changes without disclosing raw data. This preserves patient privacy while fostering collaboration.
- **Real-Time Predictions:** Models with federated learning enabled by them are able to offer real-time insights and predictions about diseases, enabling medical professionals to act promptly in the face of new health hazards.

9.2.5 Challenges and Considerations

The obstacles of using blockchain technology with FL for illness prediction and EHR sharing include guaranteeing ethical AI use, adhering to legal requirements, and managing computing resource needs. Furthermore, careful consideration of data quality and diversity is necessary to guarantee model correctness and mitigate biases in FL.

9.3 Working principle of machine learning training on blockchain-based EHR

When using a blockchain-based health data sharing system, access to data is crucial to training the machine learning model. With data security and privacy in mind, several different approaches can be used to train the machine learning model [158-160]:

1. **Collaboration with Data Holders:**A blockchain-based health data sharing system enables data holders (hospitals, clinics, research institutions, etc.) to

safely share data. For the training of the ML model, it is possible to collaborate with these data holders. Data holders can provide the data set needed for the model's training, and this data can be securely shared across the blockchain. However, it should not be forgotten that this approach may raise some privacy concerns and that data holders may not be willing to share their data.

2. **Advanced Data Sharing Methods:**Blockchain-based systems ensure that data is shared securely, but not all of this data needs to be available to everyone. Data holders can share the data set needed for training the machine learning model with specific parties in a private and secure manner. This can be done with advanced data sharing methods, and blockchain technology can facilitate this kind of secure data sharing.
3. **Federated Learning:** FL is a non-centric learning approach. In this approach, model training takes place directly on the data holders' devices or on local systems, and the training results are integrated into a central server. A blockchain-based system supports the federated learning approach, allowing data holders to train the model without sharing their data.

Each approach has its advantages and disadvantages. Which approach is preferred should be determined by taking into account security, privacy, data ownership and collaboration issues. However, blockchain-based systems allow this process to be managed more securely and transparently. Table 9.1 tcompares the advantages and disadvantages of different methods. When comparing, we must consider how each method impacts factors such as security, privacy, collaboration, and practicality. Depending on the application and usage scenarios, one or a combination of these methods may be the ideal solution.

9.5 How to integrate the proposed blockchain-based AguHyper with the novel disease prediction mechanism CSA-DE-LR?

Users perform four key operations to add records to the AguHyper. These operations involve i) uploading data to IPFS and ii) sending metadata to the Blockchain (BC).

Table 9.1: Comparison of the advantages and disadvantages of the three main methods used for machine learning training on blockchain-based EHR

Approaches	Advantages	Disadvantages
Collaboration with Data Holders	- Provides a reliable data source. -Data owners can share data needed for model training. - Provides centralized control and supervision.	- Data owners may have privacy concerns. - Sharing data requires additional security measures. - Data owners may not be willing to share their data.
Advanced Data Sharing Methods	- Data can be shared securely on the blockchain. - Private data can be shared with certain parties. - Privacy and security are prioritized.	- Advanced technology and protocols may be required for data sharing. - Data owners may still not be willing to share their data.
Federated Learning	- Data owners can train the machine learning model without sharing their data. - Privacy and security are protected at the highest level. - Training is performed on data owner devices or local systems, reducing dependence on central servers.	- Combined training of the model can be complex. - There are central control and audit difficulties. - Requires higher computing resources and hardware.

During the data upload to IPFS, the data is encrypted, and the hash value is derived from the encrypted data, concluding with the storage of the encrypted data. In the metadata sending process to BC, the transaction content, including the encrypted key, is generated and authenticated through the user's key before transmission. The data entry procedures for healthcare providers, who exclusively input patient data, differ from those performed by patients themselves in the supplementary EHRs add-on process. Notably, there is an absence of an encrypted key in the content of the patient transaction.

Upon the availability of metadata on the Blockchain, medical practitioners or researchers interested in specific data can initiate a permission request within the Blockchain network. This is accomplished by submitting a transaction that triggers the activation of the dataSharing contract. The data owner, upon receiving a permission request, has the option to grant or deny it. In the event of authorization, a transaction is generated, encapsulating components such as the ID of the requested data, the public key of the requester, and the key for decrypting the requested data, encrypted with the

requester's public key. Post permission approval, the user retrieves the data from a nearby IPFS node, followed by decryption.

In designing AguHyper, data sharing considerations were extended to two user types: sharing with doctors and sharing with researchers. Researchers need data to predict disease with CSA-DE-LR. Researchers first make a request for data sharing to the data owner according to the relevant data characteristics. After the data owner approves data sharing, the researcher automatically obtains the data from the nearby IPFS node. Then, it runs the CSA-DE-LR method with this data. As a result of the analysis, it shares the relevant result with the blockchain network. So AguHyper uses *"Advanced Data Sharing Methods"* used for CSA-DE-LR training on blockchain-based EHR. Our primary reason for using the Advanced Data Sharing Method is that we want to leave every decision regarding the data to the data owner. The reason for this is that the main reason that prevents sharing of EHRs is data privacy and security. If we used federated learning for machine learning training on blockchain-based EHRs, data sharing would be out of the question. But this time we will encounter different disadvantages in our system. We can summarize them as follows [167].

Federated learning provides the advantage of developing machine learning models without collecting data centrally. In doing so, model training takes place directly on the data holders' devices or on local systems. It is not possible today to perform EHR data model training on the device of the data owners. For this, each patient must have information about the relevant model in the researcher setting. Hospitals have the biggest responsibility in performing EHR data model training in the local system. For this, hospitals must have advanced knowledge about the relevant model and can perform model training blindly with the permission of the data owner. These training results can then be shared with the relevant person via blockchain. The main disadvantage of this is that local operation centralizes the system and requires very advanced hardware resources for complex trainings.

Because data consistency, heterogeneity, training time and bandwidth problems may be encountered between different nodes. Reverse engineering through model updates can compromise data confidentiality and attacks against the central server can cause security issues. The fact that federated learning requires coordination with a central server increases security risks, while connection problems or operation of nodes at different speeds may reduce the efficiency of model training. Issues of fair participation and representation are also important; Some nodes may have more data, which may

overfit the model by nodes. Additionally, sufficient computing power is required for local training, but not all nodes may have these resources. To overcome these challenges, it is necessary to focus on strong security measures, centralized coordination, network infrastructure and inter-node data consistency.

Chapter 10

Conclusion

EHRs can aid in early illness detection and prevention and are essential to the progress of healthcare. However, EHR sharing faces challenges such as managing large data volumes, ensuring data privacy, security, and interoperability. The purpose of this thesis is to develop and show performance analysis of an ideal EHRs sharing system that is blockchain-based suitable for disease prediction mechanism integration and addresses all EHR sharing problems in detail using SysML. To begin in this manner, the relevant platform's constituent parts were identified, and the relationships between them were examined using SysML. Furthermore, the necessary conditions for the safe and efficient functioning of the previously described system have been established, and the relationship between these conditions and the elements of the platform has been investigated. Furthermore, an example scenario is used to demonstrate how these criteria might be satisfied and system performance analysis.

Then, a permissioned architecture is designed to enable the safe exchange and privacy protection of EHRs, based on the Hyperledger blockchain. To prevent malicious assaults and illegal access, the proposed system incorporates IPFS as a distributed storage solution for EHRs. This guarantees that encrypted patient informations are maintained securely. The distributed ledger of the blockchain then contains hash values associated with these records. Using a variety of datasets, the research carefully examines the system architecture, configures AguHyper for implementation, and carefully assesses performance. CouchDB and the Raft consensus method are included in the experimental configuration, and the system's throughput and latency are monitored closely. This evaluation is made more complete and in-depth by comparing it to previous researches. Crucially, this study adds a unique viewpoint to the body of knowledge already available in the area.

In order to improve CVD diagnosis accuracy, this study also presented CSA-DE-LR, a unique hybrid classification technique that combines the CSA and DE. The Cleveland and Statlog datasets were empirically analyzed, and the results demonstrated

that CSA-DE-LR beats the most advanced machine learning techniques available today in terms of balanced performance and classification accuracy. Numerous optimization strategies demonstrate the method's versatility in a variety of settings, such as the F1 score, MCC, and MAE. Finally, CSA-DE-LR is integrated with AguHyper. and this process is explained in detail.

CSA-DE-LR is a hybrid approach that offers excellent accuracy and balanced performance over various parameters. Its flexibility allows it to be tailored to specific datasets through feature selection, improving its generalizability and usefulness in various scenarios. However, the study acknowledges limitations, such as higher computational complexity and interpretability challenges in clinical settings. Future research should test CSA-DE-LR on various medical datasets to evaluate its generalizability and usefulness. Optimizing the hybrid model's interpretability and computational efficiency could provide more comprehensive understanding. The viability of CSA-DE-LR implementation in clinical settings could be a significant step forward for real-time diagnostic tools. Integrating CSA-DE-LR with deep learning techniques could further improve its performance and application range. In conclusion, CSA-DE-LR shows promise as a technique for diagnosing CVD while addressing the drawbacks of conventional machine learning and metaheuristic techniques.

The analysis's conclusions of AguHyper show that the recommended approach is workable and skillfully satisfies a number of security requirements. It shows a great deal of potential for protecting health data's security, privacy, confidentiality, integrity, and scalability. In the future, the framework's functionality may be improved to respond to requests more quickly, which would lower response times, latency, and total expenses. Moreover, an aim is to broaden the scope of the framework to include more data exchange situations. Prospective research endeavors may delve into sophisticated encryption methodologies to enhance data security. These endeavors will contribute to the continued evolution and refinement of AguHyper, fostering its adoption and relevance in the dynamic landscape of healthcare data management.

BIBLIOGRAPHY

[1] Tanwar, S., Parekh, K., & Evans, R. (2020, February). Blockchain-based electronic healthcare record system for healthcare 4.0 applications. Journal of Information Security and Applications, 50, 102407. https://doi.org/10.1016/j.jisa.2019.102407

[2] Elangovan, D., Long, C. S., Bakrin, F. S., Tan, C. S., Goh, K. W., Yeoh, S. F., Loy, M. J., Hussain, Z., Lee, K. S., Idris, A. C., & Ming, L. C. (2022, January 20). The Use of Blockchain Technology in the Health Care Sector: Systematic Review. JMIR Medical Informatics, 10(1), e17278. https://doi.org/10.2196/17278

[3] Al Mamun, A., Azam, S., & Gritti, C. (2022). Blockchain-Based Electronic Health Records Management: A Comprehensive Review and Future Research Direction. IEEE Access.

[4] Adanur, B. (2020). Blockchain based data sharing platform for bioinformatics field. (Master's Thesis), Abdullah Gul University, Turkey.

[5] Kabra N , Bhattacharya P , Tanwar S , Tyagi S . Mudrachain: blockchain-based framework for automated cheque clearance in financial institutions. Fut Gener Comput Syst 2020;102:574–87.

[6] Vora J , Nayyar A , Tanwar S , Tyagi S , Kumar N , ObaidatM , et al. Bheem: a blockchain-based framework for securing electronic health records. In: 2018 IEEE globecom workshops (GC Wkshps); 2018. p. 1–6 .

[7] Yli-Huumo, J., Ko, D., Choi, S., Park, S., & Smolander, K. (2016). Where is current research on blockchain technology?—a systematic review. PloS one, 11(10), e0163477.

[8] Dedeturk, B. A., Soran, A., & Bakir-Gungor, B. (2021). Blockchain for genomics and healthcare: a literature review, current status, classification and open issues. PeerJ, 9, e12130.

[9] Attaran, M. (2022). Blockchain technology in healthcare: Challenges and opportunities. International Journal of Healthcare Management, 15(1), 70-83.

[10] Sookhak M, Jabbarpour MR, Safa NS and et al. 2021. Blockchain and smart contract for access control in healthcare: A survey, issues and challenges, and open issues. Journal of Network and Computer Applications, Volume 178.

[11] Azaria A, Ekblaw A, Vieira T , and Lippman A, "Medrec: Using blockchain for medical data access and permission management," in Open and Big Data (OBD), International Conference on. IEEE, 2016, pp. 25–30.

[12] Kannan S, Smith M. 2016. GemOS platform whitepaper. 1-12. Available at https://enterprise.gem.co/wp-content/ uploads/ 2016/ 10/ GemOSPlatformWhitepaper.pdf .

[13] e Estonia. 2012. F. A. questions; estonian blockchain technology. Available at https:// e-estonia.com/wp-content/ uploads/ faq-a4-v02-blockchain.pdf .

[14] Abul-Husn NS, Kenny EE. 2019. Personalized medicine and the power of electronic health records. Cell 177(1):58-69 DOI 10.1016/j.cell.2019.02.039.

[15] Mcfarlane C, Beer M, Brown J, Prendergast N. 2017. Patientory: a healthcare peer-to-peer emr storage network. v1.1. 1-19.

[16] Medicalchain. 2018. Whitepaper: Medicalchain 2.1. 1–42. https://Medicalchain.com/Medicalchain-Whitepaper-EN.pdf.

[17] Liu X, Wang Z, Jin C, Li F and Li G. (2019). A blockchain-based medical data sharing and protection scheme. IEEE Access, 7, 118943-118953.

[18] IBM's Medical Blockchain. https://github.com/IBM/Medical-Blockchain/blob/master/README.md .

[19] Al Omar A, Bhuiyan MZA., Basu A, Kiyomoto S and Rahman, MS. (2019). Privacy-friendly platform for healthcare data in cloud based on blockchain environment. Future generation computer systems, 95, 511-521.

[20] Niu S, Chen L, Wang J and Yu F. (2020). Electronic health record sharing scheme with searchable attribute-based encryption on blockchain. IEEE Access, 8, 7195-7204.

[21] Veeramakali T., Siva R, Sivakumar B, Mahesh PS and Krishnaraj N (2021). An intelligent internet of things-based secure healthcare framework using blockchain technology with an optimal deep learning model. The Journal of Supercomputing, 1-21.

[22] Połap D, Srivastava G and Yu K. (2021). Agent architecture of an intelligent medical system based on federated learning and blockchain technology. Journal of Information Security and Applications, 58, 102748.

[23] Chen M, Malook T, Rehman AU, Muhammad Y, Alshehri MD and et al. (2021). Blockchain-Enabled healthcare system for detection of diabetes. Journal of Information Security and Applications, 58, 102771.

[24] Arul R, Al-Otaibi YD, Alnumay WS, Tariq U, Shoaib U and Piran, MJ. (2021). Multi-modal secure healthcare data dissemination framework using blockchain in IoMT. Personal and Ubiquitous Computing, 1-13.

[25] Jabarulla, M. Y., & Lee, H.-N. (2021). Blockchain-based distributed patient-centric image management system. Applied Sciences, 11 . doi:https://doi.org/10.3390/app11010196.

[26] Shah, R., & Rajagopal, S. (2022). M-dps: a blockchain-based efficient and cost-effective architecture for medical applications. International Journal of Information Technology, 14 , 1909–1921. doi:http://dx.doi.org/10.1007/s41870-022-00912-1.

[27] Azbeg, K., Ouchetto, O., & Jai Andaloussi, S. (2022). Blockmedcare: A healthcare system based on iot, blockchain and ipfs for data management security. Egyptian Informatics Journal , 23 , 329–343. doi:https://doi.org/10.1016/j.eij.2022.02.004.

[28] Jayabalan, J., & Jeyanthi, N. (2022). Scalable blockchain model using off-chain IPFS storage for healthcare data security and privacy. Journal of Parallel and Distributed Computing, 164, 152-167.

[29] Mantey, E. A., Zhou, C., Srividhya, S. R., Jain, S. K., & Sundaravadivazhagan, B. (2022). Integrated Blockchain-Deep Learning Approach for Analyzing the Electronic Health Records Recommender System. Frontiers in Public Health, 10, 905265.

[30] Kotronis, C., Nikolaidou, M., Dimitrakopoulos, G., Anagnostopoulos, D., Amira, A., & Bensaali, F. (2018, June). A model-based approach for managing criticality requirements in e-health iot systems. In 2018 13th annual conference on system of systems engineering (SoSE) (pp. 60-67). IEEE.

[31] Androulaki E, Barger A, Bortnikov V, Cachin C, Christidis K, De Caro A, Enyeart D, Ferris C, Laventman G, Yacov M, Muralidharan S, Murthy C, N Nguyen B, Sethi M, Singh G, Smith K, Sorniotti A, Stathakopoulou C, Vukolic M, Cocco S, & Yellick J. Hyperledger fabric: a distributed operating system for permissioned blockchains. In: Proceedings of the Thirteenth EuroSys Conference. ACM; 2018. p. 30.

[32] Hyperledger-Fabric. (Accessed by 2023, 23 November). The Ordering Service. https://hyperledger-fabric.readthedocs.io/en/release-2.5/orderer/ordering_service.html .

[33] Muhammad Anshari. 2019. Redeining electronic health records (EHR) and electronic medical records (EMR) to promote patient empowerment. IJID (International Journal on Informatics for Development) 8, 1 (2019), 35ś39.

[34] Hsuan-Yu Chen, Zhen-Yu Wu, Tzer-Long Chen, Yao-Min Huang, and Chia-Hui Liu. 2021. Security privacy and policy for cryptographic based electronic medical information system. Sensors 21, 3 (2021), 713.

[35] Mohammad Shahid Husain, Muhamad Hariz Bin Muhamad Adnan, Mohammad Zunnun Khan, Saurabh Shukla, and Fahad U Khan. 2021. Pervasive Healthcare: A Compendium of Critical Factors for Success. Springer.

[36] Maithilee Joshi, Karuna Joshi, and Tim Finin. 2018. Attribute based encryption for secure access to cloud based EHR systems. In 2018 IEEE 11th International Conference on Cloud Computing (CLOUD). IEEE, 932ś935.

[37] Raza Nowrozy, Khandakar Ahmed, Hua Wang, and Timothy Mcintosh. 2023. Towards a universal privacy model for electronic health record systems: an ontology and machine learning approach. In Informatics, Vol. 10. MDPI, 60.

[38] Bassim Al Bahrani, Itrat Medhi, and ITRAT MEHDI. 2023. Copy-Pasting in Patients' Electronic Medical Records (EMRs): Use Judiciously and With Caution. Cureus 15, 6 (2023).

[39] Roberto Cerchione, Piera Centobelli, Emanuela Riccio, Stefano Abbate, and Eugenio Oropallo. 2023. Blockchain's coming to hospital to digitalize healthcare services: Designing a distributed electronic health record ecosystem. Technovation 120 (2023), 102480.

[40] Vivek Subbiah. 2023. The next generation of evidence-based medicine. Nature medicine 29, 1 (2023), 49ś58.

[41] Nowrozy, R., Ahmed, K., Kayes, A. S. M., Wang, H., & McIntosh, T. R. (2024). Privacy preservation of electronic health records in the modern era: A systematic survey. ACM Computing Surveys.

[42] Maryam Farhadi, Hisham Haddad, and Hossain Shahriar. 2018. Static analysis of hippa security requirements in electronic health record applications. In 2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC), Vol. 2. IEEE, 474ś479.

[43] Nilüfer Demirsoy and Nurdan Kirimlioglu. 2016. Protection of privacy and conidentiality as a patient right: physicians' and nurses' viewpoints. Biomedical Research 27, 4 (2016), 1437ś1448.

[44] Tasha Glenn and Scott Monteith. 2014. Privacy in the digital world: medical and health data outside of HIPAA protections. Current psychiatry reports 16, 11 (2014), 1ś11.
[45] John D Halamka, Andrew Lippman, and Ariel Ekblaw. 2017. The potential for blockchain to transform electronic health records. Harvard Business Review 3, 3 (2017), 2ś5.
[46] Aisling R Cafrey and Austin R Horn. 2021. Considerations for protecting research participants. In Pragmatic Randomized Clinical Trials. Elsevier, 273ś292.
[47] Mike Chapple and David Seidl. 2021. Cyberwarfare: Information operations in a connected world. Jones & Bartlett Learning.
[48] W Bani Issa, I Al Akour, A Ibrahim, A Almarzouqi, S Abbas, F Hisham, and J Griiths. 2020. Privacy, conidentiality, security and patient safety concerns about electronic health records. International Nursing Review 67, 2 (2020), 218ś230.
[49] Gabriela Kato Lettieri, Aline Hung Tai, Aline Rodrigues Hütter, André Luiz Torres Raszl, Mariana Moura, and Raquel Barbosa Cintra. 2022. Medical conidentiality in the digital era: an analysis of physician-patient relations. Revista Bioetica 29 (2022), 814ś824.
[50] Fei Zhu, Xun Yi, Alsharif Abuadbba, Ibrahim Khalil, Xinyi Huang, and Feihong Xu. 2023. A Security-Enhanced Certiicateless Conditional Privacy-Preserving Authentication Scheme for Vehicular Ad Hoc Networks. IEEE Transactions on Intelligent Transportation Systems (2023).
[51] BD Deebak and Fadi Al-Turjman. 2023. Secure-user sign-in authentication for IoT-based eHealth systems. Complex & Intelligent Systems 9, 3 (2023), 2629ś2649.
[52] Nancy D Harada, Laural Traylor, Kathryn Wirtz Rugen, Judith L Bowen, C Scott Smith, Bradford Felker, Deborah Ludke, Ivy TonnuMihara, Joshua L Ruberg, Jayson Adler, et al. 2023. Interprofessional transformation of clinical education: the irst six years of the Veterans Afairs Centers of Excellence in Primary Care Education. Journal of interprofessional care 37, sup1 (2023), S86śS94.
[53] Fatemeh Rezaeibagha, Khin Than Win, and Willy Susilo. 2015. A systematic literature review on security and privacy of electronic health record systems: technical perspectives. Health Information Management Journal 44, 3 (2015), 23ś38.
[54] Sreelakshmi Krishnamoorthy, Amit Dua, and Shashank Gupta. 2023. Role of emerging technologies in future IoT-driven Healthcare 4.0 technologies: A survey, current challenges and future directions. Journal of Ambient Intelligence and Humanized Computing 14, 1 (2023), 361ś407.
[55] Sarah Qahtan, Khaironi Yatim, Hazura Zulzalil, Mohd Hafeez Osman, AA Zaidan, and HA Alsattar. 2023. Review of healthcare industry 4.0 application-based blockchain in terms of security and privacy development attributes: Comprehensive taxonomy, open issues and challenges and recommended solution. Journal of Network and Computer Applications 209 (2023), 103529.
[56] Gurvirender PS Tejay and Zareef A Mohammed. 2023. Cultivating security culture for information security success: A mixed-methods study based on anthropological perspective. Information & Management 60, 3 (2023), 103751.
[57] A. Maxmen, 'AI researchers embrace Bitcoin technology to share medical data', Nature, 555(7696), (2018).
[58] P. Mamoshina, L. Ojomoko, Y. Yanovich, A. Ostrovski, A. Botezatu et al., 'Converging blockchain and next-generation artificial intelligence technologies to decentralize and accelerate biomedical research and healthcare', Oncotarget, 9(5), 5665, (2018).

[59] S. Angraal, HM. Krumholz, WL. Schulz, 'Blockchain technology: applications in healthcare', Circulation: Cardiovascular quality and outcomes, 10, 1–3, (2017).
[60] S. Nakamoto, 'Bitcoin: A Peer-to-Peer Electronic Cash System', 1-9, (2008), https://bitcoin.org/bitcoin.pdf.
[61] M. Di Pierro, 'What is the blockchain?', Computing in Science & Engineering, 19 (5), 92-95, (2017).
[62] TT. Kuo, HE. Kim, L. Ohno-Machado, 'Blockchain distributed ledger technologies for biomedical and health care applications', J Am Med Informatics Assoc 24**,** 1211–1220, (2017).
[63] D. Yaga, P. Mell, N. Roby, K.Scarfone,' Blockchain technology overview', arXiv preprint arXiv:1906.11078, (2019).
[64] JI. Zahid, A. Ferworn, F. Hussain, 'Blockchain: A technical overview', IEEE Internet Policy Newsl, 1-3, (2018).
[65] V. Mavroeidis, K. Vishi, MD. Zych, A. Jøsang, 'The impact of quantum computing on present cryptography', arXiv preprint arXiv:1804.00200, (2018).
[66] IF. Kaderali, 'Foundations and Applications of Cryptology', (2007).
[67] Y. Kumar, R. Munjal, R, H. Sharma. 'Comparison of symmetric and asymmetric cryptography with existing vulnerabilities and countermeasures', International Journal of Computer Science and Management Studies, 11(03), 60-63, (2011).
[68] K. Sultan, U. Ruhi, R. Lakhani, 'Conceptualizing Blockchains: Characteristics & Applications', arXiv preprint arXiv:1806.03693, (2018).
[69] O. Dib, KL. Brousmiche, A. Durand, E. Thea, EB. Hamida, 'Consortium blockchains: Overview, applications and challenges' International Journal On Advances in Telecommunications, 11(1&2), 2018.
[70] H. Anwar, 'Consensus Algorithms: The Root Of The Blockchain Technology', 101 Blockchains Website, (2018).
[71] D. Mingxiao, M. Xiaofeng, Z. Zhe, W. Xiangwei, C. Qijun, 'A review on consensus algorithm of blockchain', IEEE Int Conf Syst Man, Cybern SMC 2017, 2567–2572, (2017).
[72] V. Gatteschi, F. Lamberti, C. Demartini, C. Pranteda, V. Santamaria. 'To blockchain or not to blockchain: That is the question', IT Professional, 20(2), 62-74, (2018).
[73] F. Casino, TK. Dasaklis, C. Patsakis, 'A systematic literature review of blockchain-based applications: current status, classification and open issues', Telematics and Informatics, 36, 55-81, (2019).
[74] K. Wüst, A. Gervais, 'Do you need a blockchain?', In 2018 Crypto Valley Conference on Blockchain Technology (CVCBT), 45-54, (2018).
[75] Junaid, S. B., Imam, A. A., Balogun, A. O., De Silva, L. C., Surakat, Y. A., Kumar, G., ... & Mahamad, S. (2022, October). Recent Advancements in Emerging Technologies for Healthcare Management Systems: A Survey. In Healthcare (Vol. 10, No. 10, p. 1940). MDPI.
[76] Dedeturk, B. A., Soran, A., & Bakir-Gungor, B. (2021). Blockchain for genomics and healthcare: a literature review, current status, classification and open issues. PeerJ, 9, e12130.
[77] Kaur, J., Rani, R., & Kalra, N. (2022). A Blockchain-based Framework for Privacy Preservation of Electronic Health Records (EHRs). Transactions on Emerging Telecommunications Technologies, 33(9), e4507.

[78] Sharma, P., Namasudra, S., & Lorenz, P. (2023, May). Blockchain-Based Cloud Storage System with Enhanced Optimization and Integrity Preservation. In ICC 2023-IEEE International Conference on Communications (pp. 3744-3749). IEEE.
[79] Sonkamble, R. G., Bongale, A. M., Phansalkar, S., Sharma, A., & Rajput, S. (2023). Secure Data Transmission of Electronic Health Records Using Blockchain Technology. Electronics, 12(4), 1015.
[80] Yang, X., Li, W., & Fan, K. (2023). A revocable attribute-based encryption EHR sharing scheme with multiple authorities in blockchain. Peer-to-peer Networking and Applications, 16(1), 107-125.
[81] Rai, B. K. (2023). PcBEHR: patient-controlled blockchain enabled electronic health records for healthcare 4.0. Health Services and Outcomes Research Methodology, 23(1), 80-102.
[82] Abdelgalil, L., & Mejri, M. (2023). HealthBlock: A Framework for a Collaborative Sharing of Electronic Health Records Based on Blockchain. Future Internet, 15(3), 87.
[83] Datta, S., & Namasudra, S. (2024). Blockchain-Based Smart Contract Model for Securing Healthcare Transactions by Using Consumer Electronics and Mobile Edge Computing. IEEE Transactions on Consumer Electronics.
[84] Tsuji, H., Shii, M., Yokoyama, S., Takamido, Y., Murase, Y., Masaki, S., & Ohara, K. (2020). Reusable robot system for display and disposal tasks at convenience stores based on a SysML model and RT Middleware. Advanced Robotics, 34(3-4), 250-264.
[85] Wolny, S., Mazak, A., Carpella, C., Geist, V., & Wimmer, M. (2020). Thirteen years of SysML: a systematic mapping study. Software and Systems Modeling, 19(1), 111-169.
[86] J. Huckaby and H. I. Christensen, "A case for SysML in robotics," 2014 IEEE International Conference on Automation Science and Engineering (CASE), 2014, pp. 333-338, doi: 10.1109/CoASE.2014.6899347.
[87] Wrycza, S., & Marcinkowski, B. (2011, September). SysML requirement diagrams: Banking transactional platform case study. In EuroSymposium on Systems Analysis and Design (pp. 15-22). Springer, Berlin, Heidelberg.
[88] Iftekhar, A., Cui, X., Tao, Q., & Zheng, C. (2021). Hyperledger fabric access control system for internet of things layer in blockchain-based applications. Entropy, 23(8), 1054.
[89] Zheng, Z., Xie, S., Dai, H., Chen, X., & Wang, H. (2017, June). An overview of blockchain technology: Architecture, consensus, and future trends. In 2017 IEEE international congress on big data (BigData congress) (pp. 557-564). Ieee.
[90] Ndzimakhwe, M., Telukdarie, A., Munien, I., Vermeulen, A., Chude-Okonkwo, U. K., & Philbin, S. P. (2023). A Framework for User-Focused Electronic Health Record System Leveraging Hyperledger Fabric. Information, 14(1), 51.
[91] Mali, A. S., Jagtap, A. M., Katekar, S., Shinde, S., & Ashtekar, K. (2023, April). Food Supply Chain Management Using Hyperledger. In 2023 7th International Conference on Trends in Electronics and Informatics (ICOEI) (pp. 82-89). IEEE.
[92] Sasikumar, R., & Karthikeyan, P. (2023). Heart disease severity level identification system on Hyperledger consortium network. PeerJ Computer Science, 9, e1626.
[93] Benet J. Ipfs-content addressed, versioned, p2p file system. arXiv preprint arXiv:14073561. 2014.

[94] Liu, F., & Wang, P. (2023). A novel privacy protection method of residents' travel trajectories based on federated blockchain and InterPlanetary file systems in smart cities. PeerJ Computer Science, 9, e1495.

[95] Forouzan, B. A., & Mukhopadhyay, D. (2015). Cryptography and network security (Vol. 12). New York, NY, USA:: Mc Graw Hill Education (India) Private Limited.

[96] Ramesh, K., Ganeshkumar, P., Gokul, B., & Sharma, K. Y. (2022). Digital Certificate-Based User Authentication to Access Networks and Hosts. In International Conference on Computing, Communication, Electrical and Biomedical Systems (pp. 437-445). Springer, Cham.

[97] Kulemin N, Popov S, Gorbachev A. 2017. The Zenome Project : Whitepaper blockchain-based genomic ecosystem.

[98] Sadat MN, Aziz MM Al, Mohammed N, Chen F, Jiang X, Wang S. 2018. SAFETY: Secure GWAS in Federated Environment Through a hYbrid solution. IEEE/ACM Trans Comput Biol Bioinforma 1–17.

[99] Kairaldeen, A. R., Abdullah, N. F., Abu-Samah, A., & Nordin, R. (2021). Data Integrity Time Optimization of a Blockchain IoT Smart Home Network Using Different Consensus and Hash Algorithms. Wireless Communications and Mobile Computing, 2021.

[100] Chauhan, B. K., & Patel, D. B. (2022). A Systematic Review of Blockchain Technology to Find Current Scalability Issues and Solutions. In Proceedings of Second Doctoral Symposium on Computational Intelligence (pp. 15-29). Springer, Singapore.

[101] Rajasekaran, A. S., Azees, M., & Al-Turjman, F. (2022). A comprehensive survey on blockchain technology. Sustainable Energy Technologies and Assessments, 52, 102039.

[102] Dziembowski, S., Faust, S., Kolmogorov, V., & Pietrzak, K. (2015, August). Proofs of space. In Annual Cryptology Conference (pp. 585-605). Springer, Berlin, Heidelberg.

[103] Sharma, P., Jindal, R., & Borah, M. D. (2022). A review of smart contract-based platforms, applications, and challenges. Cluster Computing, 1-27.

[104] Vacca, A., Di Sorbo, A., Visaggio, C. A., & Canfora, G. (2021). A systematic literature review of blockchain and smart contract development: Techniques, tools, and open challenges. Journal of Systems and Software, 174, 110891.

[105] George, J. T. (2022). Consensus Algorithms for Blockchains. In Introducing Blockchain Applications (pp. 149-161). Apress, Berkeley, CA.

[106] Shuaib, K., Abdella, J., Sallabi, F., & Serhani, M. A. (2022). Secure decentralized electronic health records sharing system based on blockchains. Journal of King Saud University-Computer and Information Sciences, 34(8), 5045-5058.

[107] Conti, M., Kumar, E. S., Lal, C., & Ruj, S. (2018). A survey on security and privacy issues of bitcoin. IEEE communications surveys & tutorials, 20(4), 3416-3452.

[108] Xu, J., Xue, K., Li, S., Tian, H., Hong, J., Hong, P., & Yu, N. (2019). Healthchain: A blockchain-based privacy preserving scheme for large-scale health data. IEEE Internet of Things Journal, 6(5), 8770-8781.

[109] Dhillon V, Metcalf D, Hooper M. The hyperledger project. In: Blockchain enabled applications. Springer; 2017. p. 139–149.
[110] Hyperledger Foundation, Playground Tutorial, https://hyperledger.github.io/composer/v0.19/tutorials/playground-tutorial.html
[111] Hyperledger Foundation. (Accessed by 2023, 23 November). Hyperledger Composer. https://hyperledger.github.io/composer/v0.19/reference/rest-server .
[112] Hyperledger Foundation. (Accessed by 2023, 23 November). Hyperledger: Blockchain Performance Metrics. https://www.hyperledger.org/learn/publications/blockchain-performance-metrics .
[113] Chelladurai U, Pandian S. A novel blockchain based electronic health record automation systemfor healthcare. J Ambient Intell Humaniz Comput. 2021;1-11. doi:10.1007/s12652-021-03163-3 .
[114] Chelladurai MU, Pandian DS, Ramasamy DK. A blockchain based patient centric electronic health record storage and integrity management for e-Health systems. Health Policy Technol. 2021. doi:10.1016/j.hlpt.2021.100513 .
[115] Adanur Dedeturk B, Bakir-Gungor B. 2024. Aguhyper: a hyperledger-based electronic health record management framework. PeerJ Comput. Sci. 10:e2060 DOI 10.7717/peerj-cs.2060
[116] Pouriyeh, S., Vahid, S., Sannino, G., De Pietro, G., Arabnia, H., & Gutierrez, J. (2017, July). A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease. In 2017 IEEE symposium on computers and communications (ISCC) (pp. 204-207). IEEE.
[117] World Health Organization (WHO): Cardiovascular diseases (CVDs), 2011. Available at https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) .
[118] CDC: Coronary Artery Diseases (CAD), 2021. Available at https://www.cdc.gov/heartdisease/coronary_ad.htm .
[119] Alkayyali, Z. K., Idris, S., and Abu-Naser, S. S. (2023). A systematic literature review of deep and machine learning algorithms in cardiovascular diseases diagnosis. Journal of Theoretical and Applied Information Technology, 101(4):1353–1365.
[120] Azmi, J., Arif, M., Nafis, M. T., Alam, M. A., Tanweer, S., and Wang, G. (2022). A systematic review on machine learning approaches for cardiovascular disease prediction using medical big data. Medical Engineering & Physics, 105:103825.
[121] Naser, M. A., Majeed, A. A., Alsabah, M., Al-Shaikhli, T. R., & Kaky, K. M. (2024). A Review of Machine Learning's Role in Cardiovascular Disease Prediction: Recent Advances and Future Challenges. Algorithms, 17(2), 78.
[122] Kolukisa, B. and Bakir-Gungor, B. (2023). Ensemble feature selection and classification methods for machine learning-based coronary artery disease diagnosis. Computer Standards & Interfaces, 84:103706.
[123] Kolukisa, B., Hacilar, H., Kus¸, M., Bakır-Gungor, B., Aral, A., and Gungor, V. C. (2019). Diagnosis of coronary heart disease via classification algorithms and a new feature selection methodology. International Journal of Data Mining Science, 1(1):8–15.
[124] Kolukisa, B., Yavuz, L., Soran, A., Bakir-Gungor, B., Tuncer, D., Onen, A., and Gungor, V. C. (2020).Coronary artery disease diagnosis using optimized adaptive ensemble machine learning algorithm. International Journal of Bioscience, Biochemistry and Bioinformatics, 10(1):58–65.

[125] Dhanka, S., Bhardwaj, V. K., and Maini, S. (2023). Comprehensive analysis of supervised algorithms for coronary artery heart disease detection. Expert Systems, page e13300.

[126] Ramudu, K., Mohan, V. M., Jyothirmai, D., Prasad, D. V. S. S. S. V., Agrawal, R., & Boopathi, S. (2023). Machine learning and artificial intelligence in disease prediction: Applications, challenges, limitations, case studies, and future directions. In Contemporary Applications of Data Fusion for Advanced Healthcare Informatics (pp. 297-318). IGI Global.

[127] Ghassemi, M., Naumann, T., Schulam, P., Beam, A. L., Chen, I. Y., & Ranganath, R. (2020). A review of challenges and opportunities in machine learning for health. AMIA Summits on Translational Science Proceedings, 2020, 191.

[128] Nalluri, M. R., Kannan, K., Manisha, M., & Roy, D. S. (2017). Hybrid disease diagnosis using multiobjective optimization with evolutionary parameter optimization. Journal of healthcare engineering, 2017.

[129] Dhanka, S., & Maini, S. (2024). HyOPTXGBoost and HyOPTRF: Hybridized Intelligent Systems using Optuna Optimization Framework for Heart Disease Prediction with Clinical Interpretations. Multimedia Tools and Applications, 1-49.

[130] Murugesan, S., Bhuvaneswaran, R. S., Khanna Nehemiah, H., Keerthana Sankari, S., & Nancy Jane, Y. (2021). Feature selection and classification of clinical datasets using bioinspired algorithms and super learner. Computational and mathematical methods in medicine, 2021, 1-18.

[131] Torthi, R., Marapatla, A. D. K., Mande, S., Gadiraju, H. K. V., & Kanumuri, C. (2024). Heart Disease Prediction Using Random Forest Based Hybrid Optimization Algorithms. International Journal of Intelligent Engineering & Systems, 17(2).

[132] Muliawan, A., Rizal, A., & Hadiyoso, S. (2023). Heart Disease Prediction based on Physiological Parameters Using Ensemble Classifier and Parameter Optimization. Journal of Applied Engineering and Technological Science (JAETS), 5(1), 258-267.

[133] Sampathkumar, K., & Periyasamy, S. (2024). A Deep Learning Approach with Binary Particle Swarm Optimization for Optimizing Prediction of Heart Disease. Nature Inspired Optimization Theories (NIOT).

[134] Leema, N., Nehemiah, H. K., and Kannan, A. (2016). Neural network classifier optimization using differential evolution with global information and back propagation algorithm for clinical datasets.Applied Soft Computing, 49:834–844.

[135] Arabasadi, Z., Alizadehsani, R., Roshanzamir, M., Moosaei, H., & Yarifard, A. A. (2017). Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm. Computer methods and programs in biomedicine, 141, 19-26.

[136] Poornima, V., & Gladis, D. (2018). A novel approach for diagnosing heart disease with hybrid classifier. Biomed Res, 29(11), 2274-2280.

[137] Shahid, A. H., & Singh, M. P. (2020). A novel approach for coronary artery disease diagnosis using hybrid particle swarm optimization based emotional neural network. Biocybernetics and Biomedical Engineering, 40(4), 1568-1585.

[138] Al Bataineh, A., & Manacek, S. (2022). MLP-PSO hybrid algorithm for heart disease prediction. Journal of Personalized Medicine, 12(8), 1208.

[139] Rahman, W., Aneek, R. H., Moinuddin, M., Sakib, M. S., Iqbal, M. S., & Rahman, M. M. (2023, December). Cardiovascular Disease Prediction Utilizing Machine Learning and Feature Selection with Clonal Selection Algorithm. In 2023 5th International Conference on Sustainable Technologies for Industry 5.0 (STI) (pp. 1-6). IEEE.

[140] Duru, C., Ladeji–Osias, J., Wandji, K., Otily, T., & Kone, R. (2022, June). A review of human immune inspired algorithms for intrusion detection systems. In 2022 IEEE World AI IoT Congress (AIIoT) (pp. 364-371). IEEE.

[141] Haktanirlar Ulutas, B. and Kulturel-Konak, S. (2011). A review of clonal selection algorithm and its applications. Artificial Intelligence Review, 36:117–138.

[142] Gong, M., Jiao, L., and Zhang, L. (2010). Baldwinian learning in clonal selection algorithm for optimization. Information Sciences, 180(8):1218–1236.

[143] Zhang, L., Gong, M., Jiao, L., and Yang, J. (2008). Optimal approximation of linear systems by an improved clonal selection algorithm. In 2008 IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence), pages 527–534. IEEE.

[144] Xu, N., Ding, Y., Ren, L., and Hao, K. (2017). Degeneration recognizing clonal selection algorithm for multimodal optimization. IEEE transactions on cybernetics, 48(3):848–861.

[145] Mostafa, R. R., Khedr, A. M., Al Aghbari, Z., Afyouni, I., Kamel, I., & Ahmed, N. (2024). An adaptive hybrid mutated differential evolution feature selection method for low and high-dimensional medical datasets. Knowledge-Based Systems, 283, 111218.

[146] Azevedo, B. F., Rocha, A. M. A., & Pereira, A. I. (2024). Hybrid approaches to optimization and machine learning methods: a systematic literature review. Machine Learning, 1-43.

[147] Song, H., Bei, J., Zhang, H., Wang, J., & Zhang, P. (2024). Hybrid algorithm of differential evolution and flower pollination for global optimization problems. Expert Systems with Applications, 237, 121402.

[148] Dedeturk, B. K. and Akay, B. (2020). Spam filtering using a logistic regression model trained by an artificial bee colony algorithm. Applied Soft Computing, 91:106229.

[149] Dedeturk, B. K., Akay, B., and Karaboga, D. (2021). Artificial bee colony algorithm and its application to content filtering in digital communication. Nature-Inspired Metaheuristic Algorithms for Engineering Optimization Applications, pages 337–355.

[150] Cai, Y., Cai, Y. Q., Tang, L. Y., Wang, Y. H., Gong, M., Jing, T. C., ... & Zhang, G. W. (2024). Artificial intelligence in the risk prediction models of cardiovascular disease and development of an independent validation screening tool: a systematic review. BMC medicine, 22(1), 56.

[151] Rani, P., Kumar, R., Jain, A., Lamba, R., Sachdeva, R. K., Kumar, K., & Kumar, M. (2024). An Extensive Review of Machine Learning and Deep Learning Techniques on Heart Disease Classification and Prediction. Archives of Computational Methods in Engineering, 1-19.

[152] Cherian, R. P., Thomas, N., & Venkitachalam, S. (2020). Weight optimized neural network for heart disease prediction using hybrid lion plus particle swarm algorithm. Journal of Biomedical Informatics, 110, 103543.

[153] de Castro, L. N. and Von Zuben, F. J. (2002). Learning and optimization using the clonal selection principle. IEEE Transactions on Evolutionary Computation, 6(3):239–251.

[154] Storn, R. and Price, K. (1997). Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. Journal of global optimization, 11:341–359.

[155] Corne, D., Dorigo, M., Glover, F., Dasgupta, D., Moscato, P., Poli, R., and Price, K. V. (1999). New ideas in optimization. McGraw-Hill Ltd., UK.

[156] Bergstra, J., Yamins, D., and Cox, D. (2013). Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In Dasgupta, S. and McAllester,D., editors, Proceedings of the 30th International Conference on Machine Learning, volume 28 of Proceedings of Machine Learning Research, pages 115–123, Atlanta, Georgia, USA. PMLR.

[157] Shan, W., Qiao, Z., Heidari, A. A., Gui, W., Chen, H., Teng, Y., Liang, Y., and Lv, T. (2022). An efficient rotational direction heap-based optimization with orthogonal structure for medical diagnosis. Computers in Biology and Medicine, 146:105563.

[158] Kumar, R., Arjunaditya, S. D., Srinivasan, K., & Hu, Y. C. (2022). AI-powered blockchain technology for public health: a contemporary review, open challenges, and future research directions. Healthcare 11 (1), 81 (2023).

[159] Zukaib, U., Cui, X., Hassan, M., Harris, S., Hadi, H. J., & Zheng, C. (2023). Blockchain and Machine Learning in EHR Security: A Systematic Review. IEEE Access, 11, 130230-130256.

[160] Khalid, N., Qayyum, A., Bilal, M., Al-Fuqaha, A., & Qadir, J. (2023). Privacy-preserving artificial intelligence in healthcare: Techniques and applications. Computers in Biology and Medicine, 106848.

[161] Jayabalan, J., & Jeyanthi, N. (2024). A Review on State-of-Art Blockchain Schemes for Electronic Health Records Management. Cybernetics and Information Technologies, 24(1).

[162] Das, A., Choudhury, D., & Sen, A. (2024). A collaborative empirical analysis on machine learning based disease prediction in health care system. International Journal of Information Technology, 16(1), 261-270.

[163] Fazel, E., Nezhad, M. Z., Rezazadeh, J., Moradi, M., & Ayoade, J. (2024). IoT convergence with machine learning & blockchain: A review. Internet of Things, 101187.

[164] Zhu, M., Yuan, J., Wang, G., Xu, Z., & Wei, K. (2024). Enhancing Collaborative Machine Learning for Security and Privacy in Federated Learning. Journal of Theory and Practice of Engineering Science, 4(02), 74-82.

[165] Tripathy, S. S., Bebortta, S., Chowdhary, C. L., Mukherjee, T., Kim, S., Shafi, J., & Ijaz, M. F. (2024). FedHealthFog: A federated learning-enabled approach towards healthcare analytics over fog computing platform. Heliyon.

[166] Issa, W., Moustafa, N., Turnbull, B., Sohrabi, N., & Tari, Z. (2023). Blockchain-based federated learning for securing internet of things: A comprehensive survey. ACM Computing Surveys, 55(9), 1-43.

[167] Gupta, M., Sharma, P., & Kalra, R. (2024). Federated Learning and Artificial Intelligence in E-Healthcare. In Federated Learning and AI for Healthcare 5.0 (pp. 104-118). IGI Global.

CURRICULUM VITAE

2013– 2017	Bachelor, Computer Engineering, Erciyes University, Kayseri, TURKEY
2018 – 2020	Master, Electrical and Computer Engineering, Abdullah Gul University, Kayseri, TURKEY
2020 – 2024	Doctoral Candidate, Electrical and Computer Engineering, Abdullah Gul University, Kayseri, TURKEY
2020-	Research Assistant, Computer Engineering, Abdullah Gul University, Kayseri, TURKEY

SELECTED PUBLICATIONS AND PRESENTATIONS

J1) Dedeturk, B. A., Soran, A., & Bakir-Gungor, B. (2021). Blockchain for genomics and healthcare: a literature review, current status, classification and open issues. PeerJ, 9, e12130.

J2) Adanur Dedeturk B, Bakir-Gungor B. 2024. Aguhyper: a hyperledger-based electronic health record management framework. PeerJ Comput. Sci. 10:e2060 DOI 10.7717/peerj-cs.2060.

C1) Adanur, B., Bakir-Güngör, B., & Soran, A. (2020, October). Blockchain-based fog computing applications in healthcare. In 2020 28th Signal processing and communications applications conference (SIU) (pp. 1-4). IEEE.