

ANKARA YILDIRIM BEYAZIT UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED
SCIENCES



FRAMEWORK MODELING FOR HEALTHCARE
SYSTEM BASED ON MACHINE LEARNING AND
BLOCKCHAIN

Ph.D. Thesis by

AZHAR HASAN NSAIF DREBEE

Department of Computer Engineering

March, 2021

ANKARA

FRAMEWORK MODELING FOR HEALTHCARE SYSTEM BASED ON MACHINE LEARNING AND BLOCKCHAIN

A Thesis Submitted to

The Graduate School of Natural and Applied Sciences of

Ankara Yıldırım Beyazıt University

**In Partial Fulfilment of the Requirements for the Degree of Doctor of
Philosophy in Department of Computer Engineering**

by

AZHAR HASAN NSAIF DREBEE

March, 2021

ANKARA

PhD THESIS EXAMINATION RESULT FORM

We have read the thesis entitled “**FRAMEWORK MODELING FOR HEALTHCARE SYSTEM BASED ON MACHINE LEARNING AND BLOCKCHAIN**”, completed by **AZHAR HASAN NSAIF DREBEE** under the supervision of **ASST. PROF. AHMET ERCAN TOPCU**, and we certify that in our opinion, it is fully adequate, in scope and in quality, as a thesis for the degree of PhD.

Asst. Prof. Ahmet Ercan TOPCU

Supervisor

Prof. Dr. Fatih Vehbi ÇELEBİ

Jury Member

Assoc. Prof. Ömer KARAL

Jury Member

Prof. Dr. Hasan Şakir BİLGE

Jury Member

Assoc. Prof. Mehmet Serdar GÜZEL

Jury Member

Prof. Dr. Ergün ERASLAN

Director

Graduate School of Natural and Applied Science

I hereby declare that, in this thesis which has been prepared in accordance with the Thesis Writing Manual of Graduate School of Natural and Applied Sciences,

- All data, information and documents are obtained in the framework of academic and ethical rules,
- All information, documents and assessments are presented in accordance with scientific ethics and morals,
- All the materials that have been utilized are fully cited and referenced,
- No change has been made on the utilized materials,
- All the works presented are original,

and in any contrary case of above statements, I accept to renounce all my legal rights.

Date:5 March, 2021 Signature :.....

Nam & Surname: AZHAR HASAN NSAIF DREBEE

ACKNOWLEDGMENT

My gratitude is deeply paid to my advisor, Asst. Prof. Ahmet Ercan TOPCU, of the Graduate School of Natural and Applied Sciences, of Ankara Yıldırım Beyazıt University, for his continuous supervision, sharing his experience, encouragement, and guidance throughout my study.

Also, I would like to thank the members of my thesis committee, for their support and valuable suggestions that made great contributions to this work.

Finally, the greatest thanks go to my government and Al-Mustaniriyah University for their infinite support. This thesis is dedicated to them. My gratitude is extended to all of my family and colleagues who supported me with all of the help that they could

2021, 11 March

AZHAR HASAN NSAIF DREBEE

FRAMEWORK MODELING FOR HEALTHCARE SYSTEM BASED ON MACHINE LEARNING AND BLOCKCHAIN

ABSTRACT

Health is a basic need of people and everybody needs good and less costly health facilities. In the current era, there is a great advancement of technology, such system should be provided that ensures easy access and less costly health to everybody. There are a large of research in the medical field, like cure of cancer and uncured diseases, but health cost is still an issue for everybody despite of the availability of much information. The proposed system is developing application can provide real-time information to doctors and health providers about patients, as well as considering the security aspect in the transmission and authentication of the information. This thesis contains two parts: The first part: Provides a better decision support to increase the quality of health and care (using Logistic Regression, Decision Tree, Random Forest, Support Vector Machine and hybrid between Artificial Neural Network and Deep Neural Network), different learning models are trained, without modifying the dataset properties, while their decisions were considered as input to the logical consensus paradigm. The suggested approach abstracts the decision made by different learning models, passing those models behavior to a binary logical decision layer, which produced the results. The training samples, the decision abstraction layer and the final decision were used to update the overall logic of the proposed system. During the training phase, the update logical decision module will keep updating its parameters. After all, the proposed paradigm of combining different learning approaches with a simple binary circuitry achieved almost 100% accuracy. The second part: The next generation blockchain technology can help in reducing the cost of transactions in various government schemes. Healthcare is an industry that requires real-time up gradation and updating in order to meet the intended need for the dispatch of quick and efficient healthcare. A key generator based on logistic map theory can be applied. The key generator is not meant to be regenerated, which means that it is not possible to regenerate the same key. Therefore, the authentication is strengthened.

Keywords: Healthcare, machine learning, blockchain, authentication, logistic map

SAĞLIK SİSTEMİ İÇİN MAKİNE ÖĞRENMESİNE VE BLOKZİNCİRİNE DAYALI ÇERÇEVE MODELLEME

ÖZ

Sağlık, insanların temel bir ihtiyacıdır ve herkesin iyi ve daha az maliyetli sağlık tesislerine ihtiyacı vardır. İçinde bulunduğumuz çağda teknolojiye büyük bir gelişme var, herkese kolay erişim ve daha az maliyetli sağlık sağlayan böyle bir sistem sağlanmalıdır. Tıp alanında kanser ve tedavi edilmemiş hastalıkların tedavisi gibi çok sayıda araştırma vardır, ancak birçok bilginin mevcut olmasına rağmen sağlık maliyeti hala herkes için bir sorundur. Önerilen sistem geliştiriyor uygulaması, doktorlara ve sağlık sağlayıcılarına hastalar hakkında gerçek zamanlı bilgi sağlayabilir ve bilginin iletilmesinde ve doğrulanmasında güvenlik yönünü göz önünde bulundurur. Bu tez iki bölümden oluşmaktadır: Birinci bölüm: Sağlık ve bakım kalitesini artırmak için daha iyi bir karar desteği sağlar (Lojistik Regresyon, Karar Ağacı, Rastgele Orman, Destek Vektör Makinesi ve Yapay Sinir Ağı ile Derin Sinir Ağı arasında hibrit kullanarak), farklı öğrenme modelleri, veri seti özelliklerini değiştirmeden eğitilirken, kararları mantıksal fikir birliği paradigmasına girdi olarak kabul edildi. Önerilen yaklaşım, farklı öğrenme modelleri tarafından alınan kararı özetler, bu modellerin davranışını sonuçları üreten ikili mantıksal karar katmanına aktarır. Eğitim örnekleri, karar soyutlama katmanı ve nihai karar, önerilen sistemin genel mantığını güncellemek için kullanıldı. Eğitim aşamasında, güncelleme mantıksal karar modülü, parametrelerini güncellemeye devam edecektir. Sonuçta, farklı öğrenme yaklaşımlarını basit bir ikili devre ile birleştirmenin önerilen paradigması neredeyse %100 doğruluk elde etmiştir. İkinci bölüm: Yeni nesil blockchain teknolojisi, çeşitli hükümet programlarında işlemlerin maliyetini azaltmaya yardımcı olabilmektedir. Sağlık hizmetleri, hızlı ve verimli sağlık hizmetlerinin gönderilmesi için amaçlanan ihtiyacı karşılamak için gerçek zamanlı yükseltme ve güncelleme gerektiren bir sektördür. Lojistik harita teorisine dayalı bir anahtar üretici uygulanabilmektedir. Anahtar oluşturucunun yeniden üretilmesi amaçlanmamıştır, bu da aynı anahtarı yeniden oluşturmanın mümkün olmadığı anlamına gelir. Bu nedenle kimlik doğrulama güçlendirilmektedir.

Anahtar Kelimeler: Sağlık, makine öğrenme, blockchain, kimlik doğrulama, lojistik harita

CONTENTS

PhD THESIS EXAMINATION RESULT FORM	i
ACKNOWLEDGMENT	ii
ABSTRACT	iv
ÖZ	v
NOMENCLATURE	vii
LIST OF TABLES	v
FIGURES LIST	xii
CHAPTER 1 -INTRODUCTION.....	1
1.1.Machine Learning Techniques.....	2
1.1.1.Logistic Regression.....	2
1.1.2.Decision Tree	3
1.1.3.Random Forest	4
1.1.4.Support Vector Machine	5
1.1.5.Artificial Neural Network	5
1.1.6.Deep Neural Network.....	7
1.2. Materials and Methods.....	8
1.3. Blockchain Technologies.....	9
1.3.1.Decentralization of Consensus	9
1.3.2.Immutability and Transparency	9
1.3.3.Security.....	10
1.3.4.Automation and smart contracts.....	10
1.3.5.Storage	10
1.4 Blockchain Classification	11
1.5. Block Structure	11
1.5.1Block Header	12
1.5.2Block Identifiers – Block Header Hash	12
1.5.3Linking Blocks in the Blockchain.....	13
1.6. Research Objectives and Contributions	13

1.7. Dissertation Organization	15
CHAPTER 2	16
LITERATURE REVIEW	16
2. 1 Performance Tests.....	16
2. 2 Big Data in Healthcare.....	17
2. 3 Health Data Sources.....	18
2. 4 Analytics Techniques in Healthcare	19
2. 5 Application of Big Data Analytics in Healthcare	21
2.5.1 Clinical Decision Support	21
2.5.2 Healthcare Administration	22
2.5.3 Privacy and Fraud Detection.....	23
2.5.4 Mental Health.....	24
2.5.5 Public Health.....	24
2.5.6 Pharmacovigilance	25
2. 6 Blockchain in Healthcare.....	25
2.6.1 Medical Data Sharing	26
2.6.2 Research and Clinical Trials	27
2.6.3 Medical Data Access Control	27
2.6.4 Improved claim auditing and fraud detection	27
2.6.5 Drug Supply Chain Management.....	28
2. 7 Security in Healthcare Data	28
2. 8 Challenges in Healthcare Analytics	28
CHAPTER 3	30
3.1 Dataset	31
3.1.1 Data analysis	31
3.1.2 Dataset Properties.....	34
3.1.3 Data Selection	41
3.2 Machine Learning Models	42
3.2.1 ANN and DNN.....	42
3.3 Performance	43
3.4 Secure HealthCare Framework (blockchain).....	44

3.4.1 Structure of Security System	44
3.4.2 Blockchain Process	46
3.5 Chaotic Maps	49
3.5.1 Chaos-based Cryptography Techniques	49
3.5.2 Benefits and Disadvantages of Chaos Theory Used with Cryptography	50
3.5.3 Logistic Map	50
3.5.4 Create Blocks Stage	54
CHAPTER 4	56
4.1 Section one: System Architecture	56
4.1.1 Individual Model Results	58
4.1.2 SVM Detection of Benign and Malignant.....	59
4.1.3 Logistic Regression Detection of Benign and Malignant	60
4.1.4 Decision Tree Detection of Benign and Malignant.....	61
4.1.5 Random Forest Detection of Benign and Malignant.....	62
4.1.6 ANN Detection of Benign and Malignant	63
4.1.7 DNN Detection of Benign and Malignant	64
4.2 Logical Inference Systems	65
4.3 Section Two: Blockchain cryptography	68
4.3.1 Key Generation	69
4.3.2 Testing and Experimentation	71
4.3.3 Healthcare Services Case Study.....	73
CHAPTER 5	78
5.1 Conclusions.....	78
5.2 Future work.....	79
REFERENCES	80
CURRICULUM VITAE	91

NOMENCLATURE

Acronyms

WDBC	Wisconsin Diagnostics Breast Cancer
LR	Logistic regression
DT	Decision tree
RF	Random forest
SVM	Support vector machine
AI	Artificial intelligence
ANN	Artificial neural network
DNN	Deep neural network
MSL	Mean square loss
P2P	Peer-to-peer
SHA	Secure hash algorithm
NIST	National Institute of Standards and Technology
CPOE	Computerized physician order entry
CDSSs	Clinical decision support systems
EMRs	Electronic medical records
EHRs	Electronic health records
CHD	Coronary heart disease
PSO	Particle swarm optimization
ICUs	Intensive care units
ADRs	Adverse drug reactions
PEM	Prescription event monitoring
MICE	Multiple imputation by chained equations

PDFs	Probability density functions
ANOVA	Analysis of variance
PPV	Positive prevalence value
NPV	Negative prevalence value
POW	Proof of work
APP	Application
HMT	Hash Merkle tree
SOP	Sum-of-product
OSE	Open-system environment

LIST OF TABLES

Table 2.1 Data sources in the healthcare system	20
Table 2.2 Common hashing algorithm	26
Table 3.1 Attributes of the UCI data	34
Table 3.2 ANOVA test for epithelial cell size after adjustment	38
Table 3.3 Nucleoli dataset	40
Table 3.4 Statistical metrics used to assess the models	44
Table 4.1 Confusion matrix or error matrix	58
Table 4.2 Classification results for the benign and malignant values	59
Table 4.3 Classification results for the benign and malignant value models ...	60
Table 4.4 Classification results for the benign and malignant value models ...	61
Table 4.5 Classification results for the benign and malignant value models ...	62
Table 4.6 Classification results for the benign and malignant value models ...	63
Table 4.7 Classification results for the benign and malignant values models...	64
Table 4.8 Function of 2 binary inputs	65
Table 4.9 Matrix of confusion for all 6 of the models individually	67

LIST OF FIGURES

Figure 1.1 Random forest training algorithm	5
Figure 1.2 Artificial neural network (ANN).....	6
Figure 1.3 Blockchain structure.....	13
Figure 2.1 Hashing	26
Figure 3.1 Genral block daigram of the proposed model	31
Figure 3.2 Data analysis flowchart.....	33
Figure 3.3 Boxplots of the UCI breast cancer dataset	35
Figure 3.4 Datasets	36
Figure 3.5 Division of cell mitoses vs. tumor degree	36
Figure 3.6 Mitoses or cell division after adjusting the data	37
Figure 3.7 Epithelial cell size vs. several cancer cases.	38
Figure 3.8 Epithelial cell size data after adjustment	39
Figure 3.9 Number of nucleoli before adjustment	41
Figure 3.10 Number of normal nucleoli after data adjustment.....	41
Figure 3.11 Inter-parameter correlation analysis.....	42
Figure 3.12 Secure healthcare system.....	45
Figure 3.13 Bifurcation diagram of the logistic map	51
Figure 3.14 Lyapunov exponent of the logistic map	51
Figure 3.15 One parameter logistic map	52

Figure 3.16 Two parameter logistic map	52
Figure 3.17 Logistic map with the hybrid parameters.....	53
Figure 3.18 Example of create block stage	55
Figure 4.1 The medical assistant software design.....	57
Figure 4.2 Testing accuracy using SVM	59
Figure 4.3 Testing accuracy using LR.....	60
Figure 4.4 Testing accuracy using DT.....	61
Figure 4.5 Testing accuracy using RF.....	62
Figure 4.6 Testing accuracy using ANN.....	63
Figure 4.7 Testing accuracy using DNN.....	64
Figure 4.8 Multi-model logical inference system.....	66
Figure 4.9 Logic gate representation of the final decision.....	67
Figure 4.10 Confusion matrix chart for all 6 models	68
Figure 4.11 Logistic map with one parameter	69
Figure 4.12 Logistic map with two parameter	70
Figure 4.13 Logistic map with hybrid parameters.....	71
Figure 4.14 Blockchain peer-to-peer network	72
Figure 4.15 Transaction in a blockchain network.....	72
Figure 4.16 Pharmacy and laboratory transaction	73
Figure 4.17 Statistical test one: key space.....	76
Figure 4.18 Statistical test two: dissimilar criteria with dissimilar tests.	77

CHAPTER 1

INTRODUCTION

This chapter contains an overview of machine learning and blockchain technology, and describes some basic terminologies used in blockchain technology.

The term breast cancer has been used to describe a category of malignant growths among the tumors that affect females [1]. Globally, at least 400,000 exclusive deaths of women have been reported as the result of breast cancer annually, comprising 14% of all cancer-related deaths. [2]. The key problem with cancer is the lack of the existence of an ultimate medical solution for its treatment, implying that the best life-saving strategy against it is an early diagnosis. In addition to the above, it is known that medical laboratory tests usually lead to multiple cases, as doctors and laboratory specialists have to analyze and determine the possibility of the diagnosis being cancer or not. The Wisconsin Diagnostics Breast Cancer (WDBC) dataset is an example of a lab test that contains the characteristics of nine patients that require analysis. In this case, the instances of the above-mentioned dataset are not easily used directly due to the high variance of their values. Therefore, to demonstrate a mathematical model with decent accuracy, such data need to be adjusted. In this study, the characteristics within the WDBC dataset were used without an adjustment, i.e. as they came from the source. Then, six different models were presented, and the final decision used the output of those models as inputs into the proposed logical inference system. Online networking and big data analytics have made healthcare-related problems very light. At present, it has become easier to give patient data, while effective algorithms and tools are utilized for analyzing particular data and appropriate treatments are offered to patients, leading to reduced time and cost. Further online networks that make use of big data could also involve other professionals in the field of healthcare to inculcate new materials and firms, which could, in turn, also facilitate peer-to-peer (P2P) learning. It is important, however, that, when using these online networks and the big data of healthcare, privacy-related issues should be mitigated, due to the sensitivity of the healthcare information [2].

Technological advances have strengthened the health-care sector, as in several life sectors. Essentially, advances in technology have participated considerably in improving security, user experience, and various aspects in the field of healthcare. These improvements have been supported by electronic health records (EHRs) and electronic medical records (EMRs). However, they face challenges as a result of certain problems relevant to data integrity, security of the medical records, and user ownership of the data. New technology, such as blockchains, can provide a solution to the aforementioned problems, since they possess the ability to provide a tamper-resistant platform for storing, in a secured way, any information associated with the healthcare field and a record of the medical information of a patient [3].

In the EHR system, each access to the record of a patient is recorded (who accesses the record, where that access occurred, the date and time of the access) for any reason with its history in a log file for post-check-in of the date of arrival. Those log files are then used to reconstruct previous cases of medical records; hence, they can act as legal documents. Therefore, secure protection must be provided to the logfile against any illegal access and should be changed as little as possible, taking into account and big data and online networking [4].

1.1. Machine Learning Techniques

Machine learning is one of the most important topics of study due to the tremendous the influence of AlphaGo and other artificial intelligence (AI) applications. The subregion of AI and computer science is known as machine learning. It represents a region that uses certain unique algorithms to 'learn' computer systems with specific data without programming complexity. In particular, it is a mechanism that helps computer systems or computers to perceive, observe, understand, and predict the world as humans do. Machine learning is the ability for a computer to develop new information and capabilities and reorganize established experiences [5].

1.1.1. Logistic Regression

Logistic regression (LR) is a kind of reverse action analysis, where the categorical results could be prognosticated depending on specific predictors [6]. Probabilities of the potential results are formed, using logistic methods, as a procedure of unattached

variables. LR could be either binomial or multinomial. LR utilizes a linking procedure that transubstantiates the restricted range of a probability, $[0,1]$, into $(-\infty, +\infty)$.

In linear regression patterns with single or repeated independent variables, X , the dependent variable, Y , is an uninterrupted accidental variable in nature. However, in some positions, the dependent variable is qualitative and verbalized by 2 or more classes; in other words, it admits 2 or more values. In this case, the method of least squares does not offer reasonable estimators. A suitable approximation can be attained using LR, which permits the usage of a regression model to estimate or forecast the possibility of a specific event [6,7].

1.1.2. Decision Tree

In data science, the decision tree (DT) algorithm is a very important algorithm. There are many desirable decision techniques, such as classification and regression trees (CART), and Quinlan's C5.0 decision tree algorithm and iterative dichotomiser 3 (ID3) [8]. A DT describes the process formation and flow, in which each essential node refers a test on an element, each transition represents one of the results from the test, and each of the leaves is associated with one of the classes [9,10]. Observations are split into parts to establish trees continuously. In many cases, tree classifiers apply the ranking process in 2 phases, comprising the tree growing phase and tree pruning phase. Tree growing consists of a top-down methodology, wherein division of the tree occurs in a recursive manner. This is determined to have been accomplished 1) at the point when a node has attained identical values to those of the goal variable, or 2) when the prognoses no longer gains value as a result of separation. On the other hand, with tree pruning, absolute evolution of the tree will occur, and this absolutely evolved tree will be cut back so as to prohibit the data from being overfitted. The process of tree pruning enhances the accuracy that the tree attains in a bottom-up manner. DTs are used quite extensively in a great number of fields as a result of their being sufficiently powerful for data deployment.

1.1.3. Random Forest

Random forest (RF) is a combination learning technique that is produced through planting numerous classification trees that contribute collectively to the latest judgment based on the general majority [11,12].

RF consists of a set of classifiers, each of which has a tree structure. Assume that a specific RF has k trees of classifiers that are defined as $h(x, \Theta_n)$ for $n = 1, 2, \dots, k$, where $\{\Theta_n\}_{n=1}^k$ is a set of independent identically distributed (iid) random vectors and x is the input, and each tree votes for the most common class at input x [13].

The RF training goes through the following stages [14,15]:

1. Supposing that the instance number in the training group is N , sample N random states using the original data (bootstrap).
2. The number of characteristics is M . Then, a simple number of m ($< M$) characteristics are selected at random, after which, for those characteristics, optimal division is employed to attain division of the nodes. The value of m stays fixed in the forest development.
3. Each of the trees is fully evolved to attain its maximum potential, because no restrictions exist. Reviewing and establishing a DT a number of, k , times, we get k different DTs, resulting in a arbitrarily produced 'forest', as shown in Figure 1.1 below.

As the RF training ends with k trees, which are unique classifier models, the test phase will use the popular majority vote among those distinct trees [16]:

$$H(x) = \underset{Y}{argmax} \sum_{i=1}^k I(h_i(x) = Y), \quad (1.1)$$

where $H(x)$ represents a combined model of classification, Y represents the variable of the output, $I(\cdot)$ represents the function of indicator, h_i represents a single model of DT. To a certain variable of an input, each tree possesses right for voting in order to choose the optimal result of the classification.

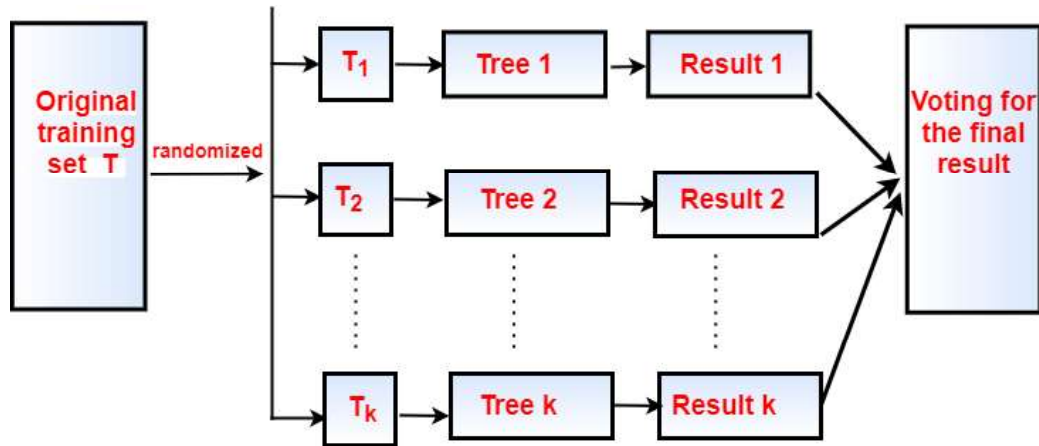


Figure 1.1 Random forest training algorithm

The WDBC breast cancer data, obtained from the University of California Irvine (UCI) Machine Learning Repository, were classified as follows: 80% were placed into the training group and 20% were placed into the test group. Training of the data yielded a RF of 500 trees. The accuracy of the model after testing it was 98.91%.

1.1.4. Support Vector Machine

Support vector machine (SVM) is what is known as a type of supervised machine learning that is broadly implemented in the domain of cancer identification and forecast. The SVM algorithm chooses significant samples from all of the categories, known as support vectors, and divided these categories by producing a linear method that splits them as widely as possible using these support vectors. Therefore, it could be considered that representation between an input vector to a high dimensionality space is produced with the use of SVM that leads to discovering the most appropriate hyperplane that splits the dataset into categories [17]. This linear classifier leads to an increase in the distances among the decision hyperplanes and the closest data point, widely known as the marginal distance, by discovering the best appropriate hyperplane [18].

1.1.5. Artificial Neural Network

Artificial neural networks (ANNs) present a way to describe artificial neurons for solving complicated issues, similarly, as the human brain does. In past years,

particularly, since in 1950s, there has been increasing attention towards studying the mechanism and structure of the human brain.

This increasing attention to such research has led to the development of novel models of computations, communication systems, or ANNs, depending upon the biological background, to solve complicated issues, such as fast processing, and adapting of information and pattern recognition [19].

This network is typically composed of several layers, which are organized in sequential order, whereas every layer comprises 1 set of neurons that all have the same communication pattern as the neurons that are in the other layer.

The first and last layers are then utilized as variables of the input and output, while the transitional layers are most often considered to be a hidden layer, which can be one or more, based on how complex the problem is. The neuron weights are then adjusted automatically as a result of training the network, in line with the rules of learning until it properly simulates the previous data or conducts the desired assignment. Mathematical functions, well known as functions of neuron transfer, are utilized for converting inputs to outputs, for each neuron [19], as can be seen in Figure 1.2

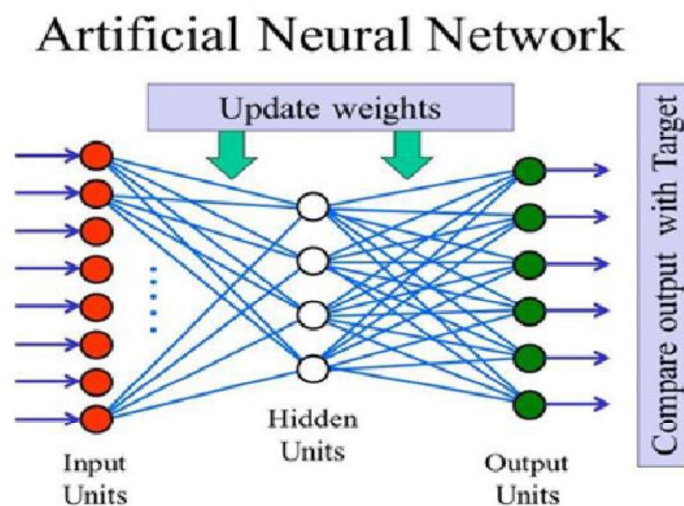


Figure 1.2 Artificial neural network [20]

1.1.6. Deep Neural Network

Recently, to provide patients with suitable medicine-related services, deep neural networks (DNNs) have gained a lot of attention, and many studies have been conducted with regard to the probability distribution and predictability of diseases from younger to older generations in accordance with the definition of precision medicine. Simultaneously, it is necessary that behavioral studies are conducted on the clinical data to obtain predictive results using several effective and AI techniques.

Because the behavioral research that is currently taking place can have a deep impact on the outcomes, obtaining the features that are the most effective is imperative with respect to correlation when attempting to reflect the precision characteristics of both machine learning and a DNN [21].

A DNN is a bio-inspired algorithm, it is graphically introduced as a set of layers, where each layer is a vertical combination of nodes. Each node corresponds to a processing unit that applies a linear function, followed by a non-linear activation function when fed with raw input data. This application will transform the data representation at each layer, leading to the statistical model that helps to thereafter perform the classification/detection task. The output is a series of probabilistic nodes, where each one corresponds to the probability of classifying the input in a certain class. This number of processing layers to which the data was passed and transformed is what inspired the label 'deep'. The strength of neural networks lies in the fact that we could train them for approximating any function, given that there are adequate data and computing time.

Initially, the network is naive, wherein it does not know the function of mapping the inputs to outputs.

We can train the network by updating its parameters according to the loss that the network makes at each step. To find those parameters, it is necessary to know how badly the network predicts the true output. Therefore, the loss function can be computed, which is a forecast error measure. For instance, the mean square loss (MSL) is sometimes utilized in problems related to regression and binary classification.

$$MSL = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \dots \dots (1.2)$$

Here, n represents the number of training examples, Y_i represents the true labels, and \hat{Y}_i represents the predicted labels [22].

1.2. Materials and Methods

It is possible to use the blockchain as a direct database or level record. Customers who use the Bitcoin Center store their blockchain metadata in Google's Level DB database. These blocks are then 'back', meaning that each of them is attached directly to the previous block in the chain. There is a particular visualization in one's mind about a blockchain, which can be visualized as a vertical stack, in which the blocks are placed on top of one another, and the main block always functions as the stack core. Visualizing blocks piled on top of each other leads to use of terms like 'height' to indicate the distance from the first block, and the term 'top' to indicate the block that was the most recently added [23].

One of the unique features of the blockchain is that each of the blocks that is within the blockchain has a distinct hash that is obtained using the secure hash algorithm-256 (SHA-256), which is a cryptographic algorithm, on the header of the block. Additionally, each of the blocks also refers to the previous block, which is called the 'parent block', through a field in the header of the block that is known as the 'hash of the previous block'. This means that the parent hash of each block is comprised in the header of that block. Then, the hash clustering, which connects each of the blocks to its parent, forms a chain that then goes directly back to the primary, or what is commonly known as the 'starting' block, in any point [24].

Any of the blocks can possess a number of children after a while due to the presence of only a single parent. Each of the children indicates a block similar to its parent and encompasses the corresponding parent hash in the field of the 'previous block hash'. Many children come into being in the blockchain 'fork', which is an impermanent condition that occurs when dissimilar blocks are detected, at almost the same time, by different miners.

Ultimately, the fork is resolved once a single child block integrates into the blockchain. Despite this, a block might contain children (more than a child), and each one could have only 1 parent. This is attributed to a block possess that is known as the 'hash of

previous block field', which indicates its only origin (the parent). The major block within the blockchain, which was created in 2009, is commonly known as the 'starting block'. This block is the 'common ancestor' of a great number of the blocks that are in the blockchain, which means that when you start from any block and track a chain in the opposite direction over time, ultimately, you can reach the starting block [25].

1.3. Blockchain Technologies

In broad terms, blockchains can be viewed and described by placing focus on their main driving principles, which include things like the decentralization of data, transparency, immutability, security, and privacy.

1.3.1. Decentralization of Consensus

The distributed nature that the network has means that the untrusted participants have to be able to reach a consensus. In a blockchain, this consensus may be based 1) on the 'rules' that are used to determine, for example, which of the transactions are, or are not, allowed, the number of bitcoins that will be included in the block reward, and the mining difficulty, or 2) on the 'transaction' history, which allows the users to be able to determine who is the owner of what. The idea behind a decentralized agreement on the transactions, is that it controls the process of updating the ledger, through the transmission of these responsibilities to the local nodes, which, in turn, performs a verification of the transactions independently, and then, subsequently, adds the transactions to the computation throughput that is the most cumulative, such as the 'longest chain' rule. Moreover, it is not necessary to have a central authority or integration point to be able to consent to the transactions or stipulated rules. Hence, there is no single point of trust or of failure [24].

1.3.2. Immutability and Transparency

The blockchains are immutable information that can only be added to previous data and once entered, it cannot be changed or lost, which provides an incorruptible past record that will become constant in the system. In addition, transparency is secured while all of the changes are indicated on the ledger and can be audited by any party that is participating in the network [26].

1.3.3. Security

Here, it is important to clarify that the blockchains are tamper-proof, shared, replicated ledgers, comprising irreversible records that cannot be changed as a result of their 1-way cryptographic hash functions. Even though security here is a relative concept, it is important to mention that blockchains are secure to a certain extent. This is because the user is only able to transfer the data if they have possession of a private key. This private key is used to facilitate 1 signature for each of the blockchain transactions that the user must send out. This created signature is then used to determine and ensure that the transaction did in fact come from the user, in addition to preventing anyone else from altering the transaction once the signature has been issued [24].

1.3.4. Automation and smart contracts.

A smart contract is a form of arbitration that is 1) mostly automatable by a computer; however, it may be required that a few select parts may be input and controlled by humans, and 2) enforceable, either through the legal enforcement of rights and responsibilities or through the execution of a tamper-proof computer code. The term 'smart contract' is also a technical meaning, at a low-level, within some technology platforms for distributed ledgers, on which it has been used to describe a replicated code that is synchronously run on more than one node of the distributed ledger. When necessary for to make it clearer, the term 'technical smart contract' can be used to refer to this low-level code [27].

1.3.5. Storage

One of the key features of storage is that storage space, such as that available on a blockchain network, can be used to store and exchange random data structures. The storage of such data may have suffered some size limitations that have been set in place to avoid the problem of 'blockchain bloating'. As an example, metadata can be used to issue meta-coins, which are second-layer systems that use the portability of the underlying coin, which serves as a fuel. Any transactions that occur in the second layer also represent a transaction in the network that underlies it. On the other hand, storage of further data can be handled off-chain, through a private cloud on the infrastructure of the client or public storage (P2P or third-party). Some blockchains, such as

Ethereum, also allow for data to be stored as a variable of a smart contract or as the log event of a smart contract [24].

1.4. Blockchain Classification

Based on the path, blockchains are mainly classified into 3 types, as public, consortium, and private blockchains, and then can be further can be classified into main and side chains, depending on the link of the chains. Moreover, many blockchains are able to form into a network, where the chains in this network are then also interconnected and generate an interchain.

Public Blockchain: This is an 'agreement blockchain', which everyone is able to gain access to. The topology of the blockchain is able to send transactions and be validated. Every person is eligible to compete for the billing rights. Public blockchains are usually considered as completely decentralized, with typical use that is similar to that of the bitcoin blockchain, where the information is fully published.

Private Blockchain: This blockchain is a courtesy to write residue in one community. The subscription to read it may be public or limited to a certain extent. Within a company, there are added options, such as database management, audit, etc. In most instances, public access is not required.

Consortium Blockchain: This is between the public and private blockchains and refers to a blockchain with a consent process that is controlled by nodes that have been preselected. As an example, imagine a system of 20 financial institutions, and each manages 1 node, and 15 of them must approve each of the blocks to be recognized as correct and then added to the chain. The right to be able to read the blockchain may be open to the public or restricted by the participants, or even a hybrid. These chains are partially decentralized [28].

1.5. Block Structure

The block is a structure that contains data that construct the entire exchange for inclusion in an open record, or in the blockchain. Blocks are made up of headers, which comprise metadata, which are then followed by an insignificant list of exchanges that form the major portion of their volume. The block header is 80 bytes, and despite that,

regular exchanges occur in the 250-byte region, and a regular block can comprise over 500 exchanges. Therefore, the total block, with all of the exchanges, is actually several times larger than the block head [29].

1.5.1. Block Header

The header of block header consists of three block metadata arrangements:

- Each block references a hash of previous block which connects a block with a previous block in the blockchain.
- The second part forms difficulty target, timestamp, and nonce, relates to mining challenge.
- The third part is the root of the Merkel tree, a structure of information, which are employed for effectively intensifying each exchange in the block. The difficulty target, timestamp and the nonce are used in the process of mining [29].

1.5.2. Block Identifiers – Block Header Hash

The primary identifier is its encryption hash. This is an impression of computerized distinct fingerprinting that is made twice by the blockhead hash via the SHA-256. The resulting 32-byte hash is called a block hash, or furthermore, it can also be called a block header hash, which is attributed for a reason, as only the block header is employed for computing it. Figure 1.3 illustrates the basic structure of the blockchain [30].

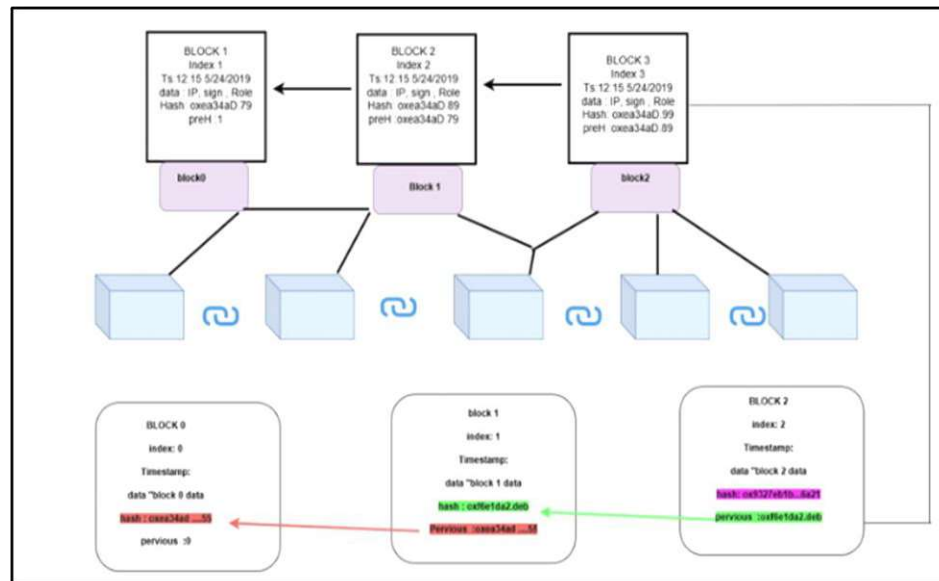


Figure 1.3 Blockchain structure

1.5.3. Linking Blocks in the Blockchain

A local copy of the blockchain is stored in the bitcoin nodes, beginning from the starting block. The local copy of the blockchain, which is continuously being updated as new blocks, is detected and utilized for extending the chain. When a node receives blocks incoming from the network, it validates those blocks and then links them to the present blockchain. For creating a link, the node scans the header of the block that reaches and search for the hash of the previous block [31].

1.6. Research Objectives and Contributions

Healthcare is a very basic need of all people and every person should be able to access good quality and low-cost health facilities. In the current era, in which there has been great advances in technology, such a system should be provided that ensures easy access and low-cost healthcare for every person. The current age has been one of information and big data, which provides many efficient, not to mention the best, types of research and products, the cost of healthcare remains an issue for people, despite the availability of a great deal of information. This data needs to be used to provide reasonable and efficient information that reduces the cost of healthcare, in addition to

securing the privacy of the information of the patients. The work herein was based on the contributions listed below:

A. Due to the ability of blockchains in enabling the safe storage of health information while maintaining a single decentralized version of the truth, different medical organizations and individuals have embraced them for purposes such as analytics to make healthcare problems lighter. Presently, it is easier to provide a patient's data, while efficient algorithms and tools analyze the given data and suitable therapies are given to the patients, resulting in time and cost reductions. Furthermore, online networking which makes use of big data can also be used to engage other healthcare professionals with new companies and materials. Thus, it will also encourage P2P learning. However, privacy issues need to be mitigated while using online networking and big data for healthcare purposes since healthcare information is sensitive [32]. Various methods of key generation and cryptography were proposed for use in blockchains. Cryptography is the study of techniques for secure communications. The various aspects of cryptography include confidentiality, data integrity, and authentication. The idea of using chaos in cryptography is highly appreciated and often researched because it is extremely sensitive to initial condition and control parameters and to pseudo-randomness, and, therefore, it perfectly suits and enhances the requirements of secure algorithms. Any small change in the initial conditions causes a drastic change in the output. The encryption scheme is unique, and the quantitative requirement is defined based on the application [33]. The blockchain receives requests from insurers, hospitals, doctors, and laboratories to access the records of a patient. In this case, each patient will control those who can access his/her data, while healthcare providers can enhance patient care based on more precise data.

B. An end-to-end case study of a blockchain-based healthcare app that we are presently developing by using logistic map equations to generate unique keys that cannot be duplicated is presented here.

C. National Institute of Standards and Technology (NIST) statistical tests are used to build, manage, and track cryptographic keys through a wide range of blockchain network nodes. The NIST statistical tests were chosen to evaluate the features of

correlation and cross-correlation, whereby the key space is sufficient. For that reason, they are appropriate for applications dealing with issues of encryption [34].

1.7. Dissertation Organization

Chapter One: As an introduction, Chapter 1 provides an overview of the main motivations for this work, in addition to the research problem statement, and research objectives and contributions.

Chapter Two: Presents background and related work. It discusses the related algorithms and search algorithms that have focused on the examination of the different analytics techniques used as a key factor in supporting the process of decision-making in the healthcare system, which used a blockchain for the secure application.

Chapter Three: Presents a new approach that was developed, called a multi-model logical inference (MMLI) system. The developed system consisted of a decision abstraction layer (DAL) that was designed to form a logical map using the output of the different initial models. Feedback from the DAL and the final decision were used to update the relationship that led to the best estimation. Without having any adjustment or duplicating the WDBC dataset, almost 100% accuracy was achieved using the proposed logical inference rule and the definition properties of the logistic map. To be able to enlarge the chaotic region of the logistic map, so as to make it better able to generate the key, the mix multiparameter of the logistic map key generation was proposed. A novel key was thus created for every session. The main algorithm showed how a new key could be generated for the hash function of a new block. To generate the unique key, and key generation that is based on the logistic map theory is able to be treated. Key generation should never perform if it is not possible to generate the key same again.

Chapter Four: This chapter on the experimental analysis includes a full description of the algorithms, machine learning, and present privacy issues for better results in the blockchain with the logistic map.

Chapter Five: Presents the conclusion and future work.

CHAPTER 2

LITERATURE REVIEW

A review of the current literature on the analytics of healthcare using big data techniques was conducted and the application of these big data techniques in systems in the field of healthcare was explored. Furthermore, focus was placed on examining the different analytical techniques used as a major factor in supporting the decision-making process in the healthcare system, as well as on big data analytical techniques, types of analytics, data, data sources, and big data applications, all within the healthcare services field. Moreover, a review was then conducted with the aim of exploring the literature on this topic from various sources, with concentration being focused on the objectives, as listed below:

- To identify the various viewpoints in relation to the definition and concepts of both big data and blockchain in the field of healthcare.
- To conduct an examination of the sources of big health data.

2.1 Performance Tests

The inherent purpose of designing a cryptosystem is to secure the ability to transmit medical images related to the medical field, in the environment of an Internet of Healthcare Things. The work herein examined a 2-dimensional triangular map that is designed using a set of familiar sine and cosine logistic maps, which characterized their dynamics. In internet healthcare scenarios, devices are connected to the internet to efficiently deliver the health data of patients, in addition to gaining access to distant health facilities, so as to be able to focus on fast recovery of the health condition suffered by the patient. As an aspect of medical data, medical images are of great significance and need to be transmitted safely. In view of the current situation of medical image security transmission, this paper focused on the prioritization of securing those images. Chaos-based encryption was used to effectively transfer medical images, as this system has the ability to easily block most of the attacks. A triangular map file was also designed, which has an effective system of encryption that can encrypt images of various kinds, particularly in the field of clinical examination.

The main metric for evaluating the performance of an algorithm is how quickly it works. The encryption algorithm should use up the minimum implementation time to be used effectively in encrypting images.

The speed of the encryption algorithm must be ascertained for each image test, so as to determine the running speed that the proposed algorithm has dissimilar volumes of images, constructed in seconds, which would indicate the efficiency of the mechanism proposed. The results that are tested for scaled images will lead to a proposal for the color image mechanism. Some of the pixel information might be blocked out or lost an encrypted image is sent. The transmission channel could also mislead a file that is a part of the message being transmitted. It is imperative that the cryptographic system is powerful when up against block attacks. Moreover, relying upon the analysis conducted herein, the capability to create unique images from blocked code images was tested [35].

2.2 Big Data in Healthcare

Big data has actually changed the very way data, in any field, is now analyzed, managed, and used. One of the most interesting domains where big data can have an impact is on healthcare. In healthcare, Big data comprises large-scale and complex electronic health datasets. The difficulty and impossibility of processing that kind of data with traditional hardware or software can be attributed to their massive number, and in addition, they also cannot be easily processed using common or traditional tools and methods in the field of data management [36]

Massive health data, known as big data, includes a large volume of biological, clinical, environmental, and lifestyle-related information taken from a target sample, ranging from an individual to large groups, on the subject of their wellness status and health, at one or several point points [37]. Big health care data are datasets that are characterized by being complex, and have some uniqueness in their characteristics, beyond their massive size, that both contribute to facilitating the process of knowledge extraction about a specific phenomenon. In addition, large data commonly require the process of data analyzing that automatically programs the analytical model building, which is called the application of automated learning, in order to provide analysis for

the datasets [38].

In the field of medicine, big data comprises miscellaneous, multispectral, uncompleted, and inaccurate observations (such as medical care, demographics, medical care, prevention of disease, diagnosis, morbidity, injury, and mental and physical disabilities), resulting from various sources through the use of non-identical samples [39].

Terminologically, big data is used to give a description for datasets that contain massive amounts or complications that conventionality, in data processing, would not be appropriate to process [40].

Big data can be described as very complex, huge data. To manage and process such a huge volume of data, traditional data tools can be used [41].

Big data can be described as the rapid and sophisticated analysis of a massive volume of very dissimilar data [42].

Big data can generally be determined through the 4 essential V's, which are as follows: volume, velocity, veracity, and variety. Furthermore, an explanation of these 4 can be illustrated separately:

- **Volume** refers to the quantity or scale of the data.
- **Velocity** refers to the analysis of the real-time, near-real-time, and speed of the data.
- **Veracity** refers to the quality assurance of the data.
- **Variety** refers to the dissimilar forms of the data, often from disparate data sources, [43]. The term big data also refers to an ultra-massive amount of data that is not limited with regards to the scope or magnitude when comes to handling definite research questions or conditions related to disease that arise continuously and rapidly [44].

It is possible to define big data as datasets with $\text{Log}(n \times p) \geq 7$. Moreover, it can describe the big data properties, such as a having high velocity and great variety [45].

2.3 Health Data Sources

Big health data sources can be classified into the following different types:

Traditional medical data: These data can be acquired from the computerized physician order entry (CPOE) EMRs, EHRs, medical history of a patient, clinical decision support systems (CDSSs), and (notes written by a physician, and medical images, prescriptions, pharmacy, insurance, laboratory, and other administration-related data), which help to gain a clear view of disease outcomes and reduce the cost of healthcare services [46].

Omics data: These data include genomics, macrobiotics, proteomics, metabolomics, etc., which play a crucial role in understanding the diseases and optimizing the treatment of individuals [47].

Internet of Things (IoT): These data can be obtained from healthcare wearables and trackers, mobile applications, medical devices, and sensors, which can have a key role in providing information regarding the daily health routines and of the lifestyle of an individual.

Social Media: Data are taken on social media sites and their respective applications, such as Facebook, Twitter, LinkedIn, etc., and generally used to analyze the spread\transmission of the disease [48]. These data can also be used to reveal and present evidence of the mood, mental state, health, etc., of an individual.

Insurance claims: These data include pharmacy claims, and private payer and health insurance claims. In summary, healthcare big data originate from 2 major sources:

- a. Internal data sources, such as EMRs, CPOE, biomedical data, and imaging data.
- b. External data sources, such as research and development laboratories, official statistics, and social media [36,49]. Table 1.1 illustrates the diverse kinds of healthcare data sources.

2.4 Analytics Techniques in Healthcare

There are 2 types of analytics that have been identified in the literature [50] show in table 2.1, as follows:

Table 2.1 Data sources in the healthcare system.

Type	Source	Description	Source
Internal	EMRs/EHRs Administrative data Diagnostics	Information-related patient, medical history medications, (physician prescriptions) admission, discharge, transfer and financial data diagnosis results (laboratory reports, imaging results)	Hospitals and clinics Laboratories
External	Insurance claims Omics wearables and trackers Medical devices Mobile Applications Social media	Data of medical reimbursement (details of insurance policy, procedures, stay in hospital). Molecular data (metabolomic, transcriptomic proteomic, genomic). Wellness and lifestyle data (smartphones, fitness monitors). Community discussions	Data aggregators Diagnostic, payers, and R&D corporations Device data systems Social media Websites, web health portals

- Descriptive analytics (to turn big data into actionable visions).
- Predictive analytics (i.e. to predict a future event relying on history-related data).
- Prescriptive analytics (i.e. to utilize data to participate in the process of decision support).

Descriptive analytics was applied in the majority of the fields of application in healthcare, except for clinical decision support. However, in pharmacovigilance, which is among the areas of application studies, it has only been employed in descriptive analytics to determine the relationship between the adverse effects of drugs

with medication. Moreover, the application of clinical decision support has made wide use of predictive analyses, as this application incorporates an array of tools that serve as an enhancing element in the decision-making process in the clinical workflow. Although numerous studies have been conducted with regards to morbidity and predicting the danger of heart attack, pain of chest, and various other cases, the use of prescriptive analytics has not been widely used, which can be attributed to the fact that most of these studies have concentrated either on a number of people in a given area or on a particular disease situation. Nevertheless, particular evidence concerning prescriptive analytics have been used in the field of medicine, management, and psychological mental state, which is run by the government. These studies have participated considerably in forming a data repository or a platform for analytical purposes to ease the decision-making process for various conditions.

In addition, big data technologies, such as MapReduce and Hadoop, can also be used in healthcare analytics [50]. Examples of these applications are:

- MapReduce is able to improve the performance of a common signal detection algorithm for pharmacovigilance with nearly linear acceleration rates [51].
- Algorithms, which rely primarily on the Hadoop distributed platform, have the ability to improve alignment of protein structure more accurately than the current algorithms [52].
- MapReduce based algorithms are able to function in improving the performance of neural signal processing [53].
- Algorithms for image reconstruction speed up the process of reconstruction [54].

2.5 Application of Big Data Analytics in Healthcare

Many potential areas where health analytics can be utilized. The most common areas identified in the literature have been mental health, healthcare administration, CDSSs, pharmacovigilance, privacy and fraud detection, and public health.

2.5.1 Clinical Decision Support

In clinical practice, big data analytics can contribute to detecting the disease in its first stage, predict the course of the disease accurately, determine the deviation from the

health condition, and changed courses of the disease. Providing this information would help healthcare facilities to meet the requirement of predictions, targeted therapy and the desired outcome concerning cost-effective care, and reduce resource wasting, by providing a recommended response to individuals, encouraging them to maintain themselves in good health.

Healthcare analytics can be applied in cardiovascular disease to investigate the factors related to coronary heart disease (CHD) [55], diagnosis of CHD [56], categorization of uncertain and high-dimensioned health disease data [57], and prediction of major adverse cardiovascular events [58].

Big data techniques can be applied to predict the types of diabetes, provide early diagnosis of the risk level of a patient and associated complications, determine the essential elements for controlling diabetes, and examine the historical record of the patient to obtain information for appropriate treatment [59].

Another major application of healthcare analytics is cancer. Analytics can be employed to identify patients that have contracted breast cancer through the use of regression and correlation to choose the important features and apply particle swarm optimization to categorize the data within the WDBC dataset [60], classify the breast cancer using genomic data [61], and predict survival for prostate cancer patients using classification models (i.e. DT, ANN, SVM algorithm) [62].

Health analytics can be used in intensive care units (ICUs) to provide a better health plan by predicting the morality rate from the data of the first 24 h of admission [63] and 30 days of admission [64] in the ICU.

2.5.2 Healthcare Administration

Big data techniques can be used for healthcare administration, including data warehouse, reduction of cost, quality enhancement, patient management, and cloud computing, etc.

Analytics can be applied to the data warehouse to improve health services by analyzing the causes of readmission [65]. Post et al. [66] developed a 3-stage framework that was built on cloud computing and big data, in order to combine the data that was gathered from various data sources. In the first stage, heterogeneous data from different sources

were collected, preprocessed, and unified. Data processing and analytics were performed in the second stage. The analysis results and models for health specialists for developing analytical tools were provided in the third stage.

Another key application of analytics in healthcare administration has been cost reduction. In the study of Zhang et al. [67], the classification and clustering algorithm was used to analyze the insurance claim data of 800,000 people over a period of 3 years to predict the cost.

Bertsimas et al. [68] applied the sentiment analysis of 6412 online patient views on a website that belonged to the English National Health Service, in 2010, about their healthcare experience.

Analytics can also be used for the purpose of patient management, which encompasses services concerning the improvement of care and scheduling to the patient through his/her residence in places like a healthcare facility [69].

Analytics can also be applied for analyzing data on the behavior of healthcare professionals and to make an assessment of the practicability of measuring drug-safety alert response through resources of medical information taken from logs available on the Internet [70].

2.5.3 Privacy and Fraud Detection

The privacy of patient data and fraud detection are seen as foremost concerns due to the rise in social media usage and the tendency that people have to keep their information on social media. A novel model was proposed by Callahan et al. [71], where the data of EHR from different health centers were stored online in cloud storage areas.

The model provided a different level of security privileges to the patients, health providers, and practitioners, in order to access and retrieve these data.

For revealing abuse and fraud (i.e., the potential suspicions on care activity, intended misleading of information, and unnecessarily repeated visits), a big data analytics framework was developed based on binary domain experts, that learn about dishonest cases manually from treatment plan data that doctors monitor regularly [72].

2.5.4 Mental Health

Data analytics can be applied in the diagnosis, analysis, or treatment of patients suffering from mental health disorders. According to Yang et al. [73], diagnosing and classifying mental disorders and relevant cases, such as (depression, dementia, alcohol abuse, anxiety, and schizophrenia) can be achieved using classification and aggregation algorithms on large datasets created from facial images, video motion, and imaging data.

To predict mental disorders (e.g., insomnia, dementia) from abnormal physical activity discovered by wearable sensors [74], a reference behavior model was developed based on the data recorded for 1 day or 1 month. The activities of the patient were compared for the reference model to detect abnormalities.

2.5.5 Public Health

For storing and integrating heterogeneous and multidimensional data (for example, nutrients, diabetes, and food) that are applicable to diabetes management, Carús Candás et al. [75] proposed a robust back-end application for internet consultation between the patient, physician, and electronic healthcare administration systems, all the while providing cost-cutting implications.

According to Nimmagadda and Dreher [76], techniques related to data mining and statistics were first used for analyzing data that were gathered for the National Institute of Public Health of a specific region in Slovenia. Then, issues related to the organization aspects of the resources of public health were investigated with the aim of determining the accessibility and availability of public health services for individuals.

Researchers took advantage of data analytics for designing systems of healthcare and systematic response to unexpected health-related occurrences, to determine the effective management of resources, satisfaction level of the patients, as well as improve the automation tool for users who had no specialized knowledge. This ultimately might help to improve patient service and the system of healthcare.

2.5.6 Pharmacovigilance

Data analytics can be involved in pharmacovigilance by identifying and monitoring new additional adverse drug events from physicians or health professionals, where long-term effects are not overlooked and by detecting adverse drug reactions (ADRs) to ensure the safety of the patients [77].

Harpaz et al. [78] applied data mining techniques for detecting adverse drug event signals from prescription event monitoring (PEM) databases, which contained reports from approximately 1 million events that were taken from 78 studies related to PEM.

Basically, the researchers, who argued that muscular and renal adverse events resulted from rosuvastatin, atorvastatin, pravastatin, and simvastatin, thought to apply the techniques of data mining to the 2004–2009 database reports of the FDA Adverse Event Reporting System [79]. The standards that were used in order to determine significant associations were: empirical Bayes geometric mean, proportional reporting ratio, reporting odds ratio, and information component. Additionally, researchers have identified dose dependent ADRs, where they used models that were developed from structured and unstructured EHR data [80], and of those, 4 were found to be related to dose, among the top 5 drugs associated with ADRs.

2.6 Blockchain in Healthcare

The technology of blockchain is a network that enables data sharing among transactional partners, including a sole copy of a distributed, decentralized, public, secure ledger. More specifically, each transactional entity possesses its ledger copy rather than having it held by a centralized party. The main feature of blockchain is that its technical infrastructure is involved in finding solutions to the settlement problem; that is to say, all ledger copies and consensus are synchronized with each other at all times, which removes the need to reconcile any copy of the ledger among transactional entities, and it removes any inefficiencies in the process. The ledger is distributed and secured, which means that it is broadcasted to all of the entities involved in the transaction and cannot be edited by any entity. The main concept in the blockchain is hashing. A hash function is a mathematical formula or algorithm that converts an arbitrary length of data (i.e. text, audio, video, etc.), as an input, into a unique fixed-

size string, called a hash value. A simple process hashing is explained in figure 2.1.

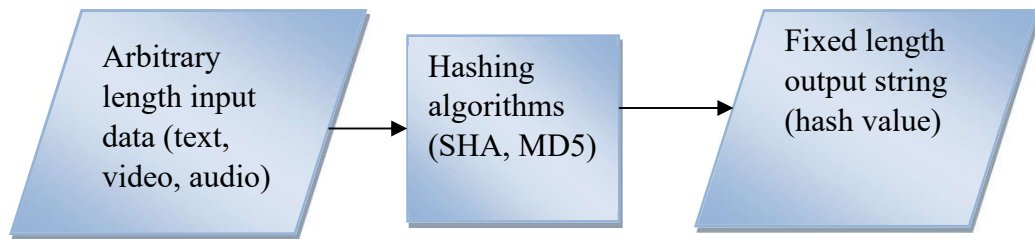


Figure 2.1 Hashing.

Any slight change in the input data results in a different hash value. The hashing process is not reversible; in other words, the data cannot be derived from its generated hash value. The most common hashing algorithms are listed in Table 2.2.

Table 2.2 Common hashing algorithms.

Algorithm	Length variety (bits)
Message Digest	128
SHA	224, 256, 384, 512
BLAKE	224, 256, 384, 512
RACE integrity primitives evaluation message digest	128, 160, 256, 320

Blockchain is a transformational technology for the healthcare services system. There are number of potential applications where the technology of blockchain can be a valuable improvement. The key application areas are briefly described below.

2.6.1 Medical Data Sharing

To bring about the improvement and development in the quality of healthcare services, it is very important to store and share the medical information regularly among all of the entities involved in the healthcare system. A blockchain provides a trustworthy and secure method of data sharing and management among all of the individuals who are involved in the transactions [81, 82].

2.6.2 Research and Clinical Trials

Another useful and fundamental method that is used in the healthcare sector is clinical trials, which require a suitable way to monitor every phase of the pathway. When it comes to design, a blockchain is seen as a decentralized structure (such as, a P2P, non-intermediated). Basically, each organization has the possibility to have full control of its resources, resulting from the computational process, while providing collaboration with other institutions regarding data analysis or data sharing [83]. For example, a blockchain can provide a platform for collecting wanted data, data management of trails, trial mechanism, and monitoring, so that those trails are present for all researchers [84].

2.6.3 Medical Data Access Control

Blockchain technology creates patient-centered healthcare, where the patient serves as the platform. Clearly, this technology both owns and controls getting an access to their data related to health issues. That process plays a key role in eliminating all difficulties for patients to acquire copies of their records related to their medical history or transfer those records to different healthcare facility. As in the case of EHRs and EMRs, several healthcare service facilities have a close association with what the patients need, and they are totally uninformed of the parties that deal with their case with regards getting access, storing, and sharing health-related data that belong to the patients. The technology of a blockchain is not exclusively restricted to enabling the patients to access their health-related information in an extremely secured way, but also ensures that only parties that have authorization can do so, whether they are individuals or institutions, they can have access and change the data [82]. The data are fully encrypted in the technology of the blockchain, and the patient is the only one who can decrypt it using his/her own private key. However, the most important thing about this technology is that if a malicious party was able to infiltrate the data, there would not be an effective way for them to read or reproduce a copy of the patient data [85].

2.6.4 Improved claim auditing and fraud detection

Many individuals can take full advantage and benefit from facilitating audits and detect fraud in a better way based on the remarkable and immutable features that a

blockchain possesses. Those individuals can be private and government insurance payers, regular payers, and individual payers.

2.6.5 Drug Supply Chain Management

Managing medical drug supply can be commonly seen as a crucial factor in the industry of modern medicine, but this concept still faces many challenges, which can include losses and complexities because of pilfering and counterfeiters. Fundamentally, a blockchain can aid in medication verification and the authenticity of its chain of supply to the parties involved. A blockchain can be viewed as an essential part of monitoring the stages that are involved in the management of a pharmaceutical supply chain and providing access to the users [86].

2.7 Security in Healthcare Data

EHRs encompass a considerable volume of sensitive information, such as data records about present and past diagnoses, health-related treatments, results of tests, images of X-rays, and other medical information, that could compromise the privacy and security of the patients. These records are usually shared among healthcare providers for simplifying communication and coordination of care. However, the confidentiality of this information is likely to be misused by many, such as employers, companies working in the field of insurance, or any individual or organization who wants to exploit the data for particular gains. Therefore, it is vital to protect these records from unauthorized access and pernicious attacks, and the theft of these data for profit, to ensure the integrity, confidentiality, and availability of the EHRs [87]. The most common methods employed for providing the security of health-related data are authentication, encryption, data masking, and access control.

2.8 Challenges in Healthcare Analytics

Big data technique application has become increasingly popular in the field of healthcare. However, there are challenges in data analytics in the field of healthcare, which include the computational time, noisy data, high dimensionality, heterogeneity of the data, and its dynamic nature. Future opportunities for research may include the

field of information loss, personal health care, pre-processing, and obtaining healthcare-related data, to conduct the study, multidisciplinary study and specialist knowledge of the field, automation for non-experts, integration into the healthcare system, and a prediction for the application of data mining. The application of blockchain paradigm in the healthcare sector raises various issues and challenges, such as confidentiality/transparency [88], scalability/speed [89], and 51% attack threat [90].

CHAPTER 3

METHODOLOGY

This chapter presents an overview of the proposed model of the healthcare system. The model aimed to put into place a foundation of analysis for the healthcare data via machine learning and provide a unique privacy and security approach through a blockchain framework. First, the proposed model had to analyze the user data, based on the machine learning classification, and then analyze the data and provide therapies that were appropriate for the patients, while at the same time, reducing the time and cost.

Online networks which use big data are also able to engage with other professionals in the field of healthcare to introduce new firms and materials. Therefore, it is also able to encourage P2P learning. However, while using online networks and the big data of healthcare, it is necessary to mitigate privacy issues, due to the sensitivity of the healthcare information.

The secure technique presented in this thesis was based on the blockchain framework with a newly proposed authentication and used logistic map features, where SHA-256 calculated the hash value of the plain text, and the results were then applied to change the initial keys in the logistic map. Figure 3.1 illustrates the structure of the proposed model.

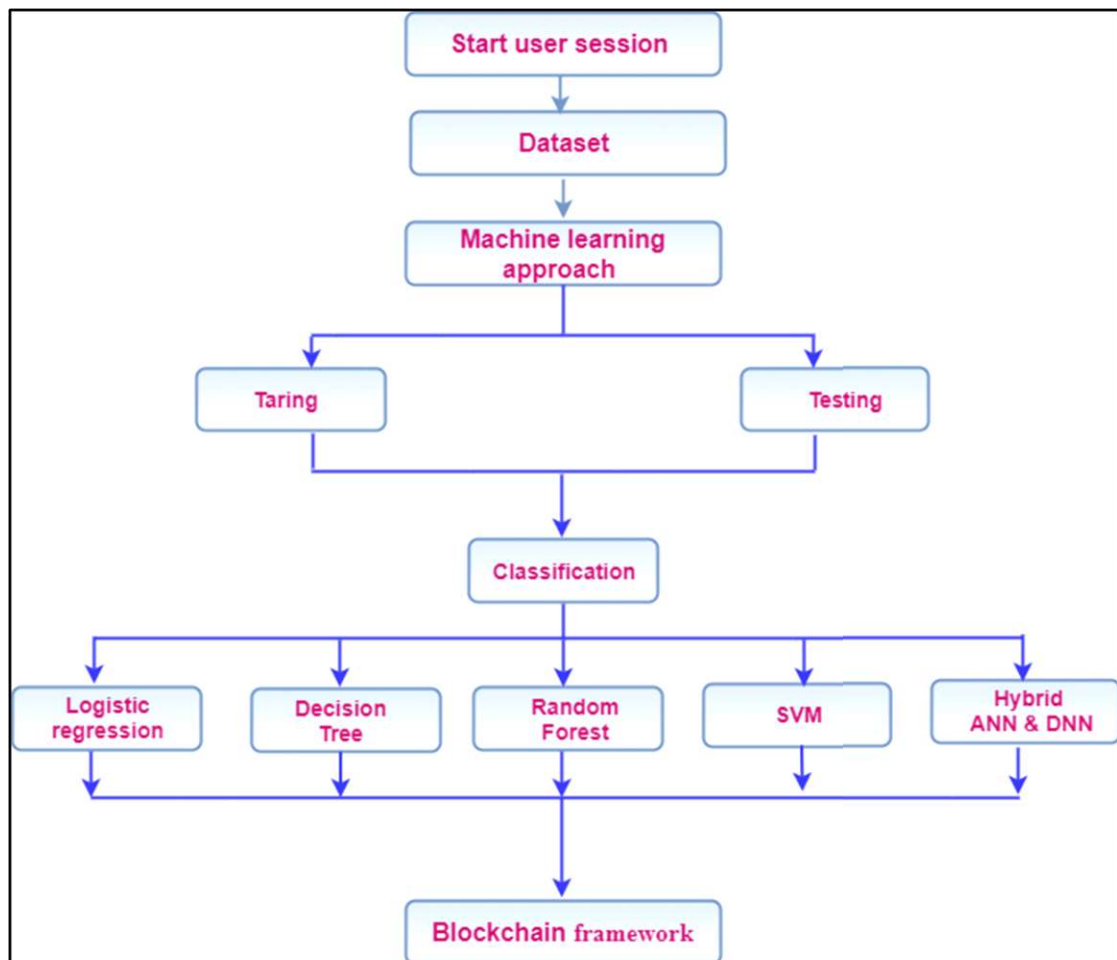


Figure 3.1 Genral block diagram for the proposed model

3.1 Dataset

This section focuses on the data analysis, machine learning, and data mining predictive modeling methods. The common steps used for building a predictive model in this work were as follows:

3.1.1 Data analysis

Analysis steps:

- 1- Load data and remove the mark and ID

- 2- Replace the **NA** values with the most common value for each parameter (column), using the multiple imputation by chained equations (MICE) function from the MICE package in R language.
- 3- Divide the dataset into sets for training and testing, with 80% training and 20% for a test from the whole set. To do this, the R package caTools was used.
- 4- Inspect all of the parameter 'columns' of the dataset using a boxplot and the ggplot2 library in R.
- 5- Depending on the boxplot, some of the parameters needed to be adjusted by grouping the observations of the same trend into a single group. The threshold that was used to decide which observations were similar was chosen by inspecting each parameter individually.
- 6- Verify the selected threshold in step 5 by doing an F-test (this was a verification step).
- 7- Modeling: there are many analytical methods available in the R programming environment. For instance, DTs, LR, RF regression/classification, and SVM.
- 8- RF was chosen as it had the best evaluation accuracy among the other methods.
- 9- Testing: the evaluation step uses a portion of training data, whereas that portion was already used to build the model in the first place. The test, however, was about entering a new input, which was not seen by the model in the training phase. The test accuracy was 98.57% with a confidence interval of 95%. Figure 3.2 shows the steps for the data analysis flowchart.

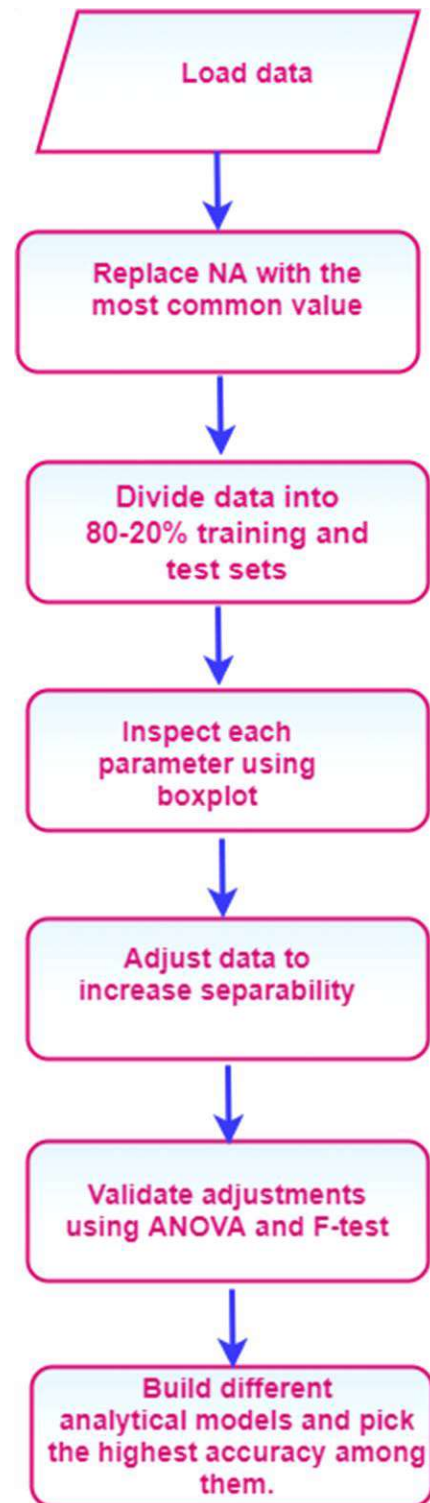


Figure 3.2 Data analysis flowchart.

3.1.2 Dataset Properties

The WDBC dataset obtained from UCI was originally collected from the University of Wisconsin Hospital, Madison, Wisconsin, USA. The dataset reflects real clinical cases that are grouped in a chronological order. There are 699 instances with the following attributes show in table 3.1 [91]. :

Table 3.1 Attributes of the UCI data [91]

Attribute	Domain
1. Sample code number	ID Number
2. Clump Thickness	1 – 10
3. Uniformity of Cell Size	1 – 10
4. Uniformity of Cell Shape	1 – 10
5. Marginal Adhesion	1 – 10
6. Single Epithelial Cell Size	1 – 10
7. Bare Nuclei	1 – 10
8. Bland Chromatin	1 – 10
9. Normal Nucleoli	1 – 10
10. Mitoses	1 – 10
11. Class	2 for benign, 4 for malignant

The original dataset measured 9 cases, each possessing 1 of 2 possible categories: malignant or benign. The values in the domain were the medical lab measures that were scaled from 1 to 10, as provided by the UCI [92]. For instance, the clump thickness was measured by the degree of layer alignment of the cell. On the other hand, marginal adhesion was a scale of how the cells stuck together among each other. Normally, the cells were tightly stuck together, but they lost contact in the cancer cells.

Benign tumor cells usually have a nuclei that is not bounded by cytoplasm. Therefore, bare nucleoli are associated with the degree of malignancy. Bland chromatin measures the degree of the chromatin pattern in the cell. Cancer cells have uneven chromatin distribution, while it is evenly distributed in normal cells. Finally, the mitoses factor measures the degree of cell-division, which is in an uncontrollable situation in cancer

cells [93]. There are many outliers that will certainly affect the statistical model that would be used later in the classification. The most critical instance is mitosis, as shown in the boxplots of Figure 3.3 below.

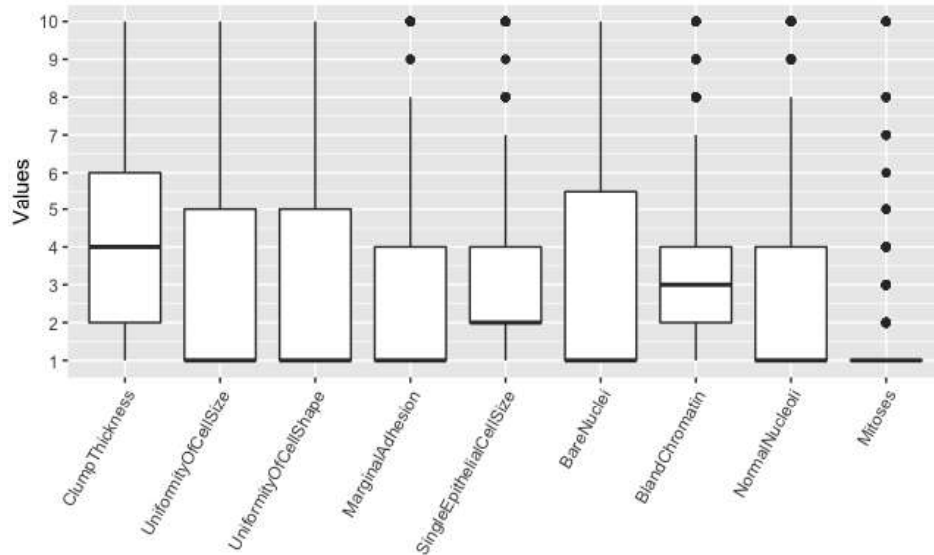


Figure 3.3 Boxplots of the UCI breast cancer dataset.

From the boxplot in Figure 3.3, it can be seen that some instances have plenty of extreme samples. It shows how each attribute is distributed around its average value.

It can be seen that there are many flavors of probability density functions (PDFs). Moreover, the WDCB dataset has a relatively limited number of samples, specifically 699 samples, which is not expected to be sufficient for training a highly accurate model. The diseases were limited to 2 classes, benign or malignant, and this class information was divided into 458 benign instances and 241 malignant instances.

Therefore, adjusting the data or doubling it would be the first step suggested by any data scientist, so that a better model can be obtained. Adjusting the dataset, or artificially adding values to it, or even using valid statistical methods, like the nearest k-means average, may raise an ethical issue for such a life-threatening decision. Figure 3.4 shows the dataset for the values determined as malignant or benign.

Uniformity of cell size	Uniformity of cell Shape	Normal Nucleoli	Bare Nuclei	Single Epithelial cell size	Clump Thickness	Marginal Adhesion	Bland Chromatin	Mitoses	Class
5	1	1	1	2	1	3	1	1	2
5	4	4	5	7	10	3	2	1	2
3	1	1	1	2	2	3	1	1	2
6	8	8	1	3	4	3	7	1	2
4	1	1	3	2	1	3	1	1	2
8	10	10	8	7	10	9	7	1	4
.
.
.
.
.
3	1	1	1	3	2	1	1	1	2
2	1	1	1	2	1	1	1	1	2
5	10	10	3	7	3	8	10	2	4
4	8	6	4	3	4	10	6	1	4
4	8	8	5	4	5	10	4	1	4

Figure 3.4 Datasets.

From the deep investigation of mitoses in Figure 3.5, it can be seen that the benign and malignant samples have similar patterns. However, for mitoses of less than 2, the number of samples is higher. This type of data malfunction comes either from a data collection procedure or due to the nature of the disease. In order to reduce the outlier effect on the classification model, it is more appropriate to consider 2 categories of mitoses: below and above the 2 divisions.

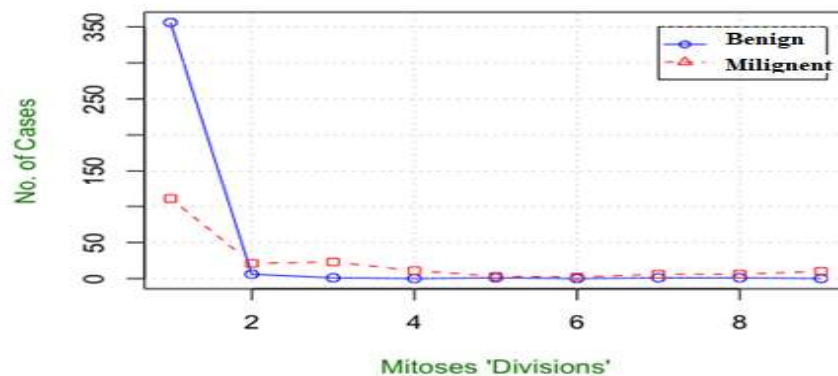


Figure 3.5 Divisions of cell mitoses vs. tumor degree.

By doing so, the correlation coefficient will be increased from 0.407 to 0.51, which will increase the classification model efficiently. Figure 3.6 shows how the classification will be facilitated after the suggested adjustment.

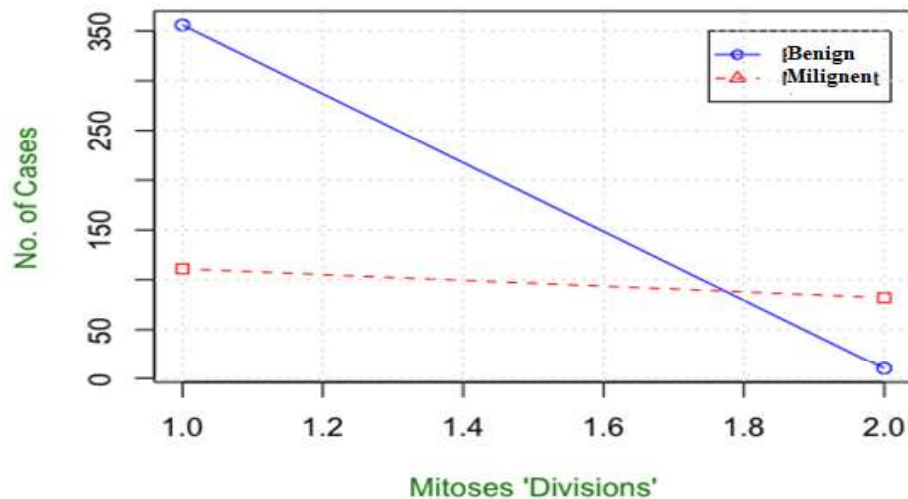


Figure 3.6 Mitoses or cell division after adjusting the data.

For confirming the validity and usefulness of the suggested data adjustment, the following hypotheses were assumed:

Null Hypothesis: for mitoses, for value 2, the possibility of a malignant tumor occurring is less than or equal to value 1.

Alternate Hypothesis: for mitoses, for value 2, the possibility of a malignant tumor occurring is greater than value 1.

By testing both hypotheses using a t-test, the outcome P-value for 95% was 2.2e-16, which had 2 implications:

1. The mitoses data after the proposed adjustment had a relationship with the tumor being malignant, and it did not happen by chance.
2. The small P-value suggested that the null hypothesis should be rejected, which meant that with a tumor that has a value of 2 or higher, the possibility of the occurrence of a malignant state of breast cancer would be great. So that:

- Null hypothesis: For the mitoses column, the possibility of the development of a malignant tumor for value 2 is less than or equal to value 1.
- Alternative hypothesis: For the mitoses column, the possibility the development of a malignant tumor for value 1. The Welch 2-sample t-test of $t = 17.144$, degree of freedom = 165.3 and $P < 2.2e-16$.

Based on these results, the alternative hypothesis is a real difference in means is greater than 0 with 95% confidence.

For the epithelial cell size data, the proportion of being a malignant or benign tumor seemed to be close for sizes 1 and 2. For sizes above 4, the number of cases was very similar, as shown in Figure 3.7.

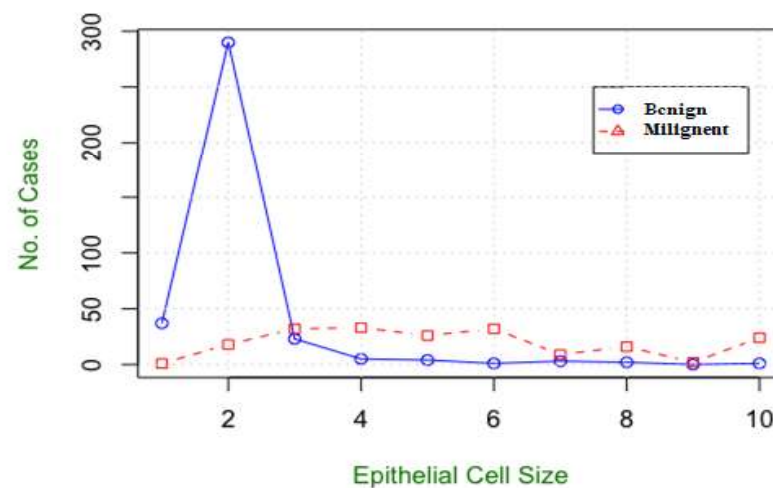


Figure 3.7 Epithelial cell size vs. several cancer cases.

To make sure that the adjustment procedure was valid, analysis of variance (ANOVA) conducted on the modified data. Table 3.2 shows the test results for the ANOVA.

Table 3.2 ANOVA test for epithelial cell size after adjustment

	Df	Sum Sq	Mean Sq	F-value	Pr(>F)
Epithelial cell size	2	80.65	40.32	490.4	<2e-16
Residuals	556	45.72	0.08		

The very small P-value that is shown in Table 3.2, as seen in the $\text{Pr}(>F)$ field, indicates that certain differences must exist with regards to the possibility that a tumor is malignant, due to the different values given for the adjusted epithelial cell size data. This means that even after adjustment, the information is still included in the modified data. Figure 3.8 shows the modified epithelial cell size data.

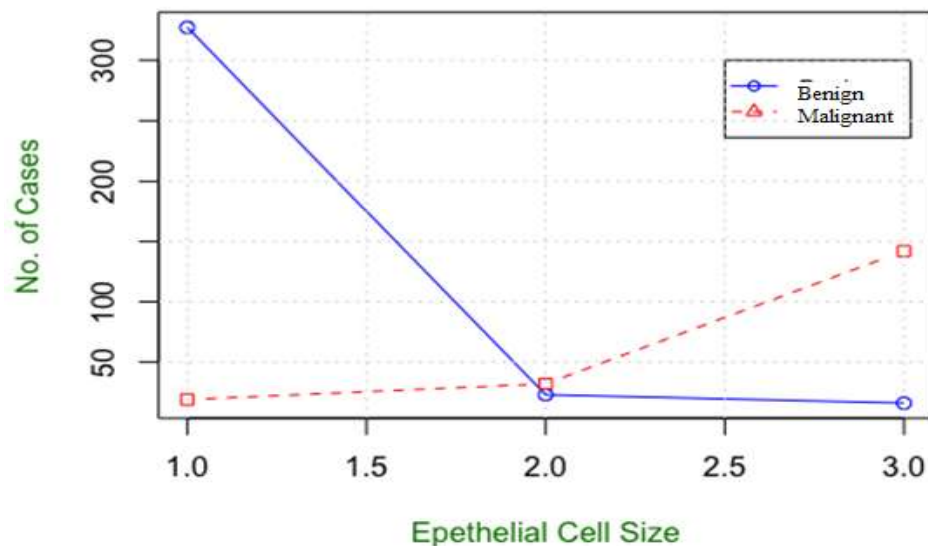


Figure 3.8 Epithelial cell size data after adjustment.

The nucleoli part of the data could also be adjusted by grouping samples of the same behavior. By inspecting Table 3.3, it is clear that the sampled data have a similar trend after 4 normal nucleoli. Before 3 nucleoli, the behavior is also similar for the data. When the normal nucleoli number is 3, the number of patients has begun, and the malignant tumors are somehow unique. Therefore, 3 nucleoli samples will be kept as a specific group so that no important information will be lost.

After grouping the nucleoli data into 3 groups, classification is now more feasible than before. Figures 3.9 and 3.10 show the number of normal nucleoli, vs. the number of cases of benign and malignant tumors, before and after data adjustment.

Table 3.3 Nucleoli dataset

No. of normal nucleoli	Benign	Malignant
1	324	33
2	22	5
3	9	25
4	1	16
5	1	16
6	4	16
7	1	13
8	3	17
9	1	13
10	0	39

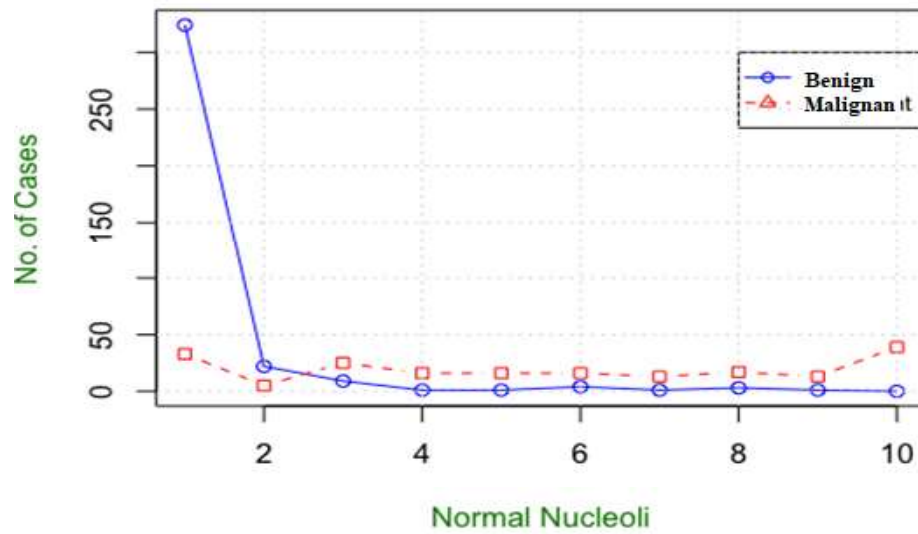


Figure 3.9 Number of nucleoli before adjustment.

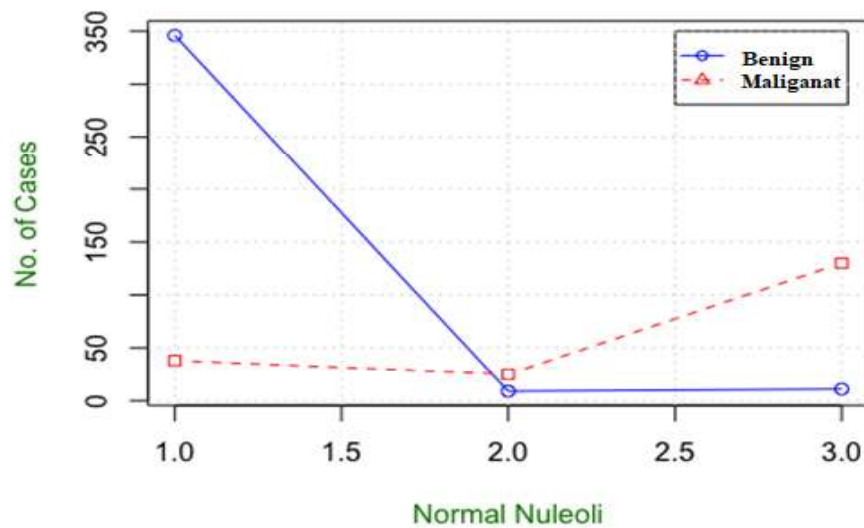


Figure 3.10 Number of normal nucleoli after data adjustment.

3.1.3 Data Selection

The numeric fields of the UCI breast cancer data comprise marginal adhesion, thickness of the clump, cell size and shape uniformity, bare nuclei, and bland chromatin. The autocorrelation of each parameter has to be higher than the cross-correlation within each other. Figure 3.11 shows the inter-parameter correlation.

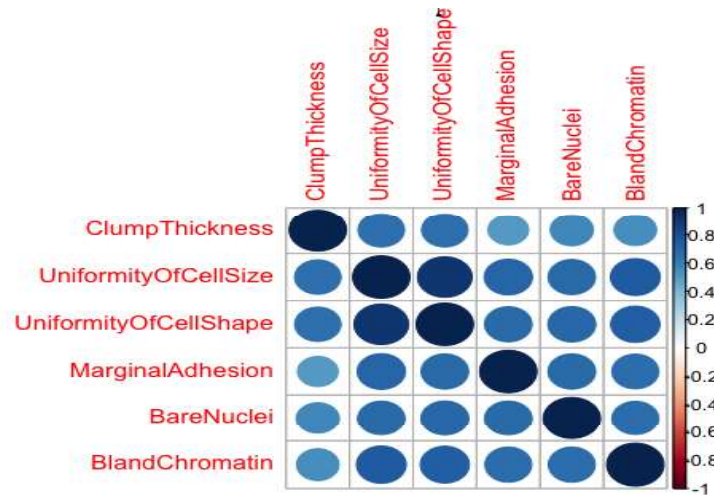


Figure 3.11 Inter-parameter correlation analysis

3.2 Machine Learning Models

There are 6 models that were trained using the WDCB dataset, as they were, without duplicating or modifications. This will increase the credibility of the decision that will be based on an unobserved sample in the future. The following sub-sections will briefly introduce each one of them. While ANNs and deep learning use the specific library in R, there are many analytical methods available in the R programming environment. For instance, SVM-LR, DT, RF, SVM, and a hybrid comprising an ANN and DNN. Herein, the new algorithm was a hybrid of the ANN and DNN.

3.2.1 ANN and DNN

The ANN is well-known for its superior efficiency in performing pattern recognition. An ANN consists of a set of perceptron states that are arranged in 3 different layers, comprising first an input layer, a hidden layer, and an output layer. Each state is trained collaboratively with other states on each layer, so that their parameters are adaptively altered for each trial. The state-of-the-art approach for using the ANN in speech recognition is the DNN, which is a multilayer perceptron structure with many hidden layers [94]. In order to build an ANN model, the (nnet) library is required in the R environment. The following pseudo-code was used to train the ANN model:

```

initialize network weights = 0.5

set hidden layer node = 5

enable soft max // normalize output layer PDFs

do
    for each sample of training called sa
        prediction = neural-net-output (network, sa) // forward pass
        actual = actual-output(sa)
        calculate error (prediction - actual) at unit output
        calculate  $\Delta\omega$  for every weights starting from the layer of input to hidden
layer // backward pass continued
        update network weights
    return the network

```

While deep learning required the (Deepnet) library, the deep ANN training was similar to the conventional ANN. However, the ANN had a single layer (with 5 nodes in the current case). While the deep ANN had more than 1 layer, in this research, 3 layers were used.

3.3 Performance

The statistical concepts, which are a set of concepts that were used as tools in designing and implementing the proposed system, will be reviewed in the section titled Statistical Analysis of the Data. The statistical metrics that were used herein to differentiate between the modeling approaches are presented in Table 3.4.

Table 3.4 Statistical metrics used to assess the models

NO.	Metric	Formula
1	Accuracy	$Acc = \frac{A + D}{A + B + C + D} \times 100\%$
2	Sensitivity	$S_v = \frac{A}{A + C}$
3	Specificity	$S_p = \frac{D}{B + D}$
4	Prevalence	$P_v = \frac{A + C}{A + B + C + D}$
5	Positive prevalence value (PPV)	$PPV = \frac{S_v \times P_v}{(S_v \times P_v) + ((1 - S_p)(1 - P_v))}$
6	Negative prevalence value (NPV)	$NPV = \frac{S_p \times (1 - P_v)}{((1 - S_v)P_v) + ((S_p)(1 - P_v))}$
7	Detection rate	$D_r = \frac{A}{A + B + C + D}$
8	Detection of prevalence	$D_P = \frac{A + B}{A + B + C + D}$
9	Balanced accuracy	$B_{ACC} = \frac{S_v + S_p}{2}$

3.4 Secure HealthCare Framework (blockchain)

The basic idea of the proposed system was to maintain the level of privacy in the authentication in the blockchain network through the following points:

3.4.1 Structure of Security System

First, the system proposed herein acted on the application of the blockchain technique to provide more security for the healthcare system. The blockchain verified any activity that deviated from the normal standard in the activities of the system. Primarily, the blockchain was designed for the digital currency of Bitcoin, but, at this time, the tech community has reached several various utilizations of technology like that. Thus far, the total value of Bitcoin is approaching \$150 billion. However, just as

with the internet, it is necessary to have knowledge regarding how a blockchain works to be able to use it. The traditional blockchain technologies have not been able to independently support the original users of the different input data; hence, new forms of attacks are now included. The main challenges faced when using blockchains are as given below:

- **Computing power:** to generate, add, receive, or transmit a block, and integrate necessary security features, which all increase the complexity involved in its computing power.
- **Storage Memory:** As a result of storing all of the received or generated blocks from the other nodes, each of the nodes has requirements with regards to high memory for storage of the data. As it is not possible to conduct authentication using proof of work (POW), a new hash function had to be created that depended on the logistics map and then a new key had to be created. Figure 3.12 presents a depiction of the overall description of the proposed system for application of the blockchain technology in healthcare systems.

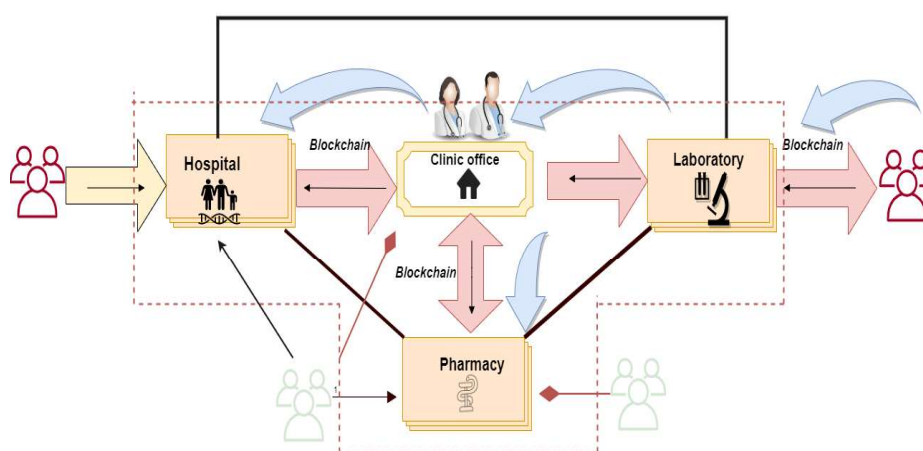


Figure 3.12 Secure healthcare system

A system for the management of access and identity was proposed by leveraging the blockchain technology, so as to support the authentication and authorization of users/institutions in a digital system. A preliminary version can be used for illustrating the process of implementing a blockchain for purposes relevant to management of identity and access. A blockchain that has an preliminary size of roughly 2 parameters

will be able to fulfill the basic processes of authentication and authorization in 2 to 3 seconds.

The proposed identity and access management architecture was composed of 2 servers, where the first sever was used as an authorization server (Auth-server), while the second sever was used as an application server (APP-server), in addition to a client APP as well as a database. The process of transaction validation that occurred in the blockchain network resulted in the system authenticating and authorizing the users. For clarifying that, it can be said that the application formed connections with the application server, and the application server conducted authentication of the users via the credentials that were previously stored within the database. Next, after successful user authentication, it performed essential processes, which were, for instance, represented by the update or retrieve data within the database.

At present, blockchains are able to provide secure and real-time data, although in the past, blockchains were only used to store bitcoins and currency information. To put it another way, the blockchain network is now able to be used in healthcare systems, which promises the authentication of information.

At the same time, the ledger contains blocks of information, that are interconnected in chronological order, in order to form the blockchain, which represents the current set of transactions at any given time.

Ledger integrity is supported by a single source of truth via the use of cryptography. Other users/institutions that wish to perform transactions can do so using a digital signature or a hash, which is a challenging mathematical problem, in which a miner must solve the problem to be able to find a block. The transaction is therefore authenticated, which allows the network to be able to recognize the transaction as valid. However, no other party will have the ability of using it without the cryptographic information.

3.4.2 Blockchain Process

In fact, there are various difficulties that should be overcome in order to provide more security to the healthcare system, especially with regards to malware code sparsity, network modeling, and hacking policies. To handle those difficulties, the framework

herein was proposed based on a blockchain, which comprises the key processes that lead to creation of the blockchain and thus, improvement of the healthcare system. A user login process was planned via the blockchain that can occur through a smart contract, which the blockchain creates. Because the idea of a blockchain is widely distributed, any computer or virtual machine can invoke a smart contract based on this algorithm 1:

Algorithm 1: Algorithm to initialize blockchain framework for healthcare services	
Input:	User Request Login Process
Result:	Valid or Invalid
<pre> // Access blockchain While Not Null do if Create new block for patient, then if previous block null, then Create genesis block calculate HashFunction, Create Smart Contract for valid user, else Generate block, Get last block as a previous, Create new time stamp, Create Smart Contract for valid user, Calculate HashFunction base on algorithm 2, end else return end end </pre>	

The initialized blockchain advantages can be listed as follows:

- Detection entirely about the attacker (intruder) access.
- Preventing the attacker from transmitting bogus data.

- Capturing (DoS), man in the middle, Internet protocol spoofing, and types of threat injection attacks.

This process comprises a pool of function calls, as below:

Process = {PO₁, PO₂, . . . , PO_n} The main features of the processes can be listed as follows:

- **PO1:** The user makes a request to gain access to the healthcare system.
- **PO2:** A block is generated for each new request.
- **PO3:** Broadcast blocks in the domain.
- **PO4:** Validation is performed for requests using the authentication key.
- **PO5:** A new key is made for the validation for each session.
- **PO6:** The mathematical formula is used, which is based on algorithm 2.

The initialization process for the blockchain is presented in algorithm 1. Moreover, the development a hidden key can be seen as one of the most significant stages that occurs in this process. The assumption that all of the users are in possession of a health account in the blockchain is based on the platform approach used herein. The scenario stages for that are given in the first step of the algorithm was followed in order to build the concept of the blockchain. Algorithm 2 provides the process that is used in the creation of a new key for the most recent of the hash functions.

Algorithm 2: Algorithm to generate secret key.
Input: Rate, Most recent numbers of blocks and latest blocks number in blockchain
Result: Generate Hash function
<pre> // calculation While Not Null do if user login then if Create new block for patient then Rate= LatestBlock/lastBlock, Alpha from node get the second from datetime, n=LatesBlock-lastBlock, $g[\sum \text{rate} * a(X_n) * (1 - X_n) + g[\sum \text{rate} * a(Y_n^\alpha (1 - Y_n^\alpha))]]$ end else return end end end </pre>

3.5 Chaotic Maps

Chaotic maps are the most common patterns used to enrich the power of image encryption schemes. The dynamic systems' chaotic properties, such as ergodicity, high sensitivity to initial conditions and topological transience affect the appearance of different schemes of encryption, which are dependent on chaotic maps. It is widely known that an algorithm of high-quality encryption should have a greater sensitivity to the secret key and should also have a large key space, resisting brute force attacks and make them unworkable [95].

3.5.1 Chaos-based Cryptography Techniques

This term refers to asymmetric and symmetric key algorithms. Symmetric algorithms consist of the stream cipher and block cipher approaches, while asymmetric algorithms

consist of public-key cryptography and cryptographic hash functions. An overview of the techniques that are related to chaos-based cryptography will be presented in this study.

3.5.2 Benefits and Disadvantages of Chaos Theory Used with Cryptography

It is possible to define the algorithms of classical cryptography via integer number fields, whereas it is possible to define chaos-based cryptography via continuous number fields. Hence, it refers to the continuous wavelength encryption of the signal with no sampling or quantization, which is its benefit. The drawbacks in relation to the classical processes of cryptography include short-cycle lengths and excess data. In spite of the flaws of chaos-based cryptography, tremendous effort has been made to overcoming these challenges. Cryptography with many algorithms and with deterministic chaos has been described. Thus, the main advantage that the chaos provides is the ability to produce multialgorithm solutions, owing to the potential development of the unlimited numbers of an algorithm. The approach of the multialgorithm results from the use of dissimilar algorithms for the encryption processes of the different blocks of the data, which also participate considerably in solving the problem of a short-cycle length, which is related to chaotic iterations [96].

3.5.3 Logistic Map

A logistic map is a 1-dimensional map, which can be used to model simplistic nonlinear discrete systems. It is possible to explain the logistic map through the recursive function, as shown below:

$$X_{n+1} = L (r - X_n) X_n = (1 - X_n) \dots\dots (3.1)$$

r represents the parameter and X_n $[0, 1]$.

Looking at the logistics map, $L: [0, 1] \rightarrow [0, 1]$ based on equation (3.1), the parameter r is located in the interval $(0, 4]$.

Figure 3.13 presents a diagram of logistic map bifurcation, while Figure 3.14 presents a diagram of a Lyapunov exponent that can be presented by equation (3.2):

$$LE = \frac{1}{n} \sum_{i=1}^n \ln |f'(x_i)| \quad (3.2)$$

Accordingly, it is now possible to confirm the dynamic behavior of the various values of parameter r that are mentioned above [97].

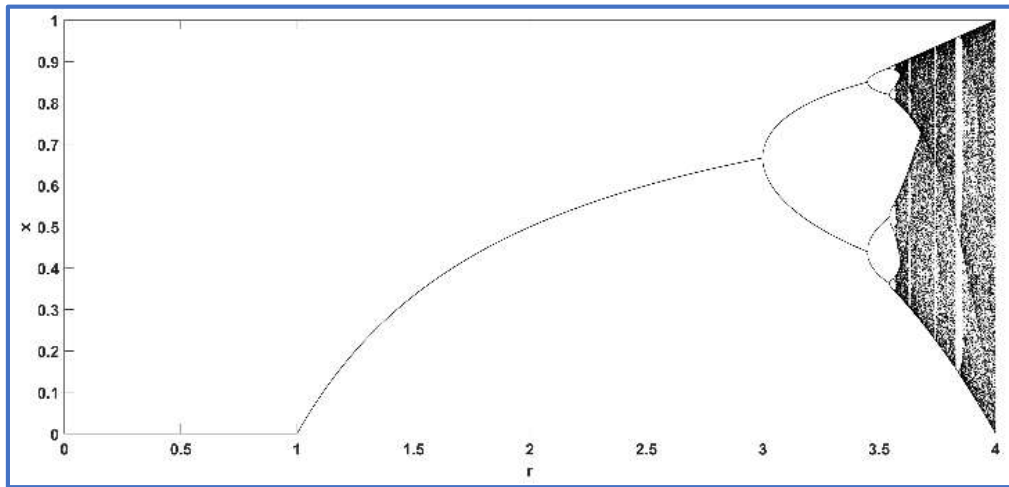


Figure 3.13 Logistic map bifurcation [97]

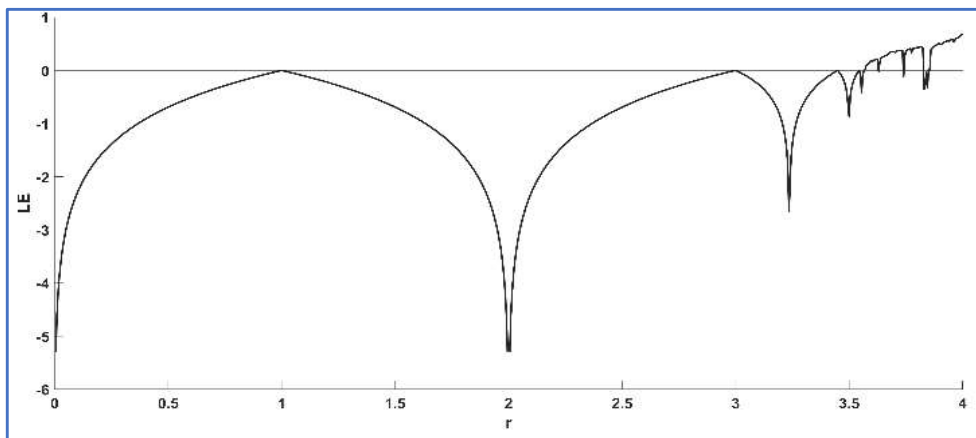


Figure 3.14 Lyapunov exponent of the logistic map [97]

A description of the behavior of chaotic systems is presented in the deterministic nonlinear settings, in addition to the underlining complexity. Applying chaotic maps in the development of an encryption system can have its own basis in the essential point, which are distinguished by these features:

- Encryption rate is high, which is attributed to the simple implementation of the software and hardware.
- The orbital evolution cannot be predicted.

- Parameters of control and initial condition with high sensitivity [98].

a. Logistic Maps with One and Multiple Parameters

Recently, many techniques that are related to information encryption have used chaotic maps with 1 dimension due to it being highly efficient and simple. However, they contain various shortcomings, including small key spaces and poor security. Therefore, logistics maps with one or many parameters have been employed to overcome those shortcomings. The map presented here was based on parameter r , from $3.57 = 4$, and the chaotic behavior is displayed through the map, as presented in Figures 3.15 and 3.16.

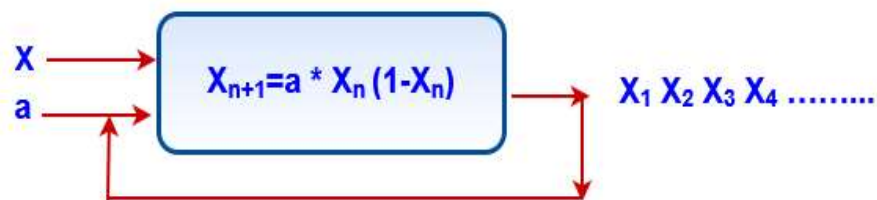


Figure 3.15 One-parameter logistic map.

Here, $a = (0.4]$ and $0 < X_n < 1$.

Following extraction of the values from the model, as can be seen in the formula in Figure 5, where $0 < X < 1$, all of these values will be transferred, which are then extracted, to a binary system, as follows:

if $X_{n+1} < 0.5$ **0**

if $X_{n+1} \geq 0.5$ **1**

b. Two-parameter logistic map

Figure 3.16, below, represents a logistic map with 2 parameters.

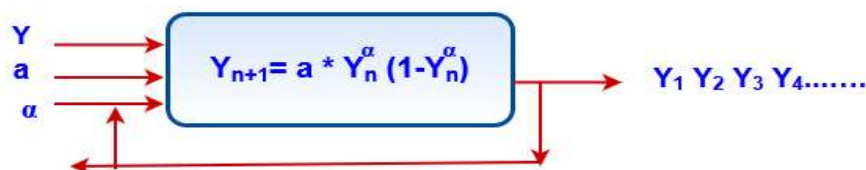


Figure 3.16 Two-parameters logistic map

Here, $a = (0.4]$, $\alpha = 0.5$, and $0 < Y_n < (0.5)^2$.

As in the first formula, an extraction process will be conducted to the values from the second formula and it will transfer all of those values, which are then extracted to a binary system, as follows:

if $Y_{n+1} < 0.5$ 0

if $Y_{n+1} \geq 0.5$ 1

c. Logistic Map with Hybrid Parameters

Here, formulas 1 and 2 had to be merged, resulting in the creation of a hybrid logistic function. The values of **X** and **Y** were then added to this hybrid function, and these values, which were extracted from the hybrid formula, were then converted to a binary system, as is shown in Figure 3.17. By making comparison to the results from both formulas 1 and 2, as well as from the hybrid, it can be seen that the results indicated that the most valuable formula was the hybrid, which contained a number of random properties.

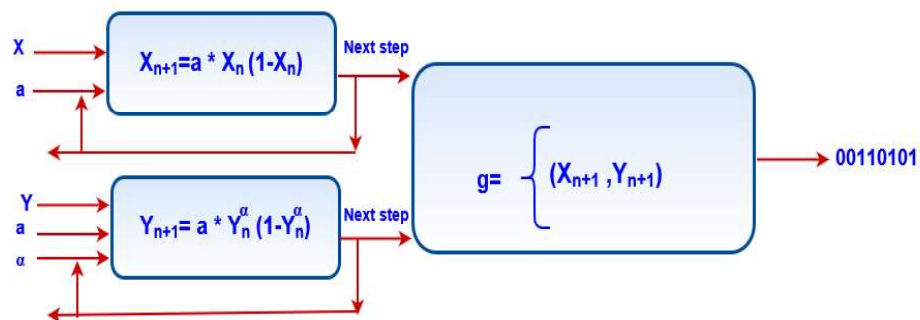


Figure 3.17 Logistic map with hybrid parameters

The conditions of the logistic map with hybrid parameters are as follows:

$$\begin{cases} 0 & \text{if } X_i < Y_i \\ 1 & \text{if } X_i \geq Y_i \end{cases}$$

Plotting successive iterations can allow the representation of dissimilar dynamic behaviors, which can be seen in the system of the logistic map, where in the static system, $a = 3.57$ and $a = 4$ were successful and can be utilized to generate keys due to the number of values from $a = 0.1, 0.2, 0.3, \dots, 3.57$ generate keys is nearly 0.

3.5.4 Create Blocks Stage

Each block in the blockchain network comprises **block header** and a **block body**.

a) The block header contains:

- **Hash Merkle Tree (HMT):** This comprises a hash value that was formed of all of the transactions in the block, which were calculated and presented in the previous section.
- **Timestamp:** This represents the time and date for the transaction registration in the block.
- **Previous-hash:** This represents the cryptography SHA-256 for the previous block. More specifically, the first block in the blockchain and the previous hash, so the value of this is equal to zero.
- **Current hash:** This represents the hash value for the current block header.

b) Block body includes:

- The body section of the block can be constructed of a set of transactions.
- The block header hash (80 bytes) of the N block, resulting in 32-bytes (256 bits using SHA-256) is stored as a 'hash of the former block header', which is part of the block header for the N + 1 block.

In the creating blocks stage, 2 values are calculated, which are:

1- Compute the value of current-hash by copying the content of the block header (timestamp value, previous-hash value, HMT value), then, apply the SHA-256 on the block header contents to generate 1 hash value, as shown in Figure 3.18.

2- Compute the timestamp value that represents the time to create the current block. The block timestamp is different than the time stamp it is located inside of in the block.

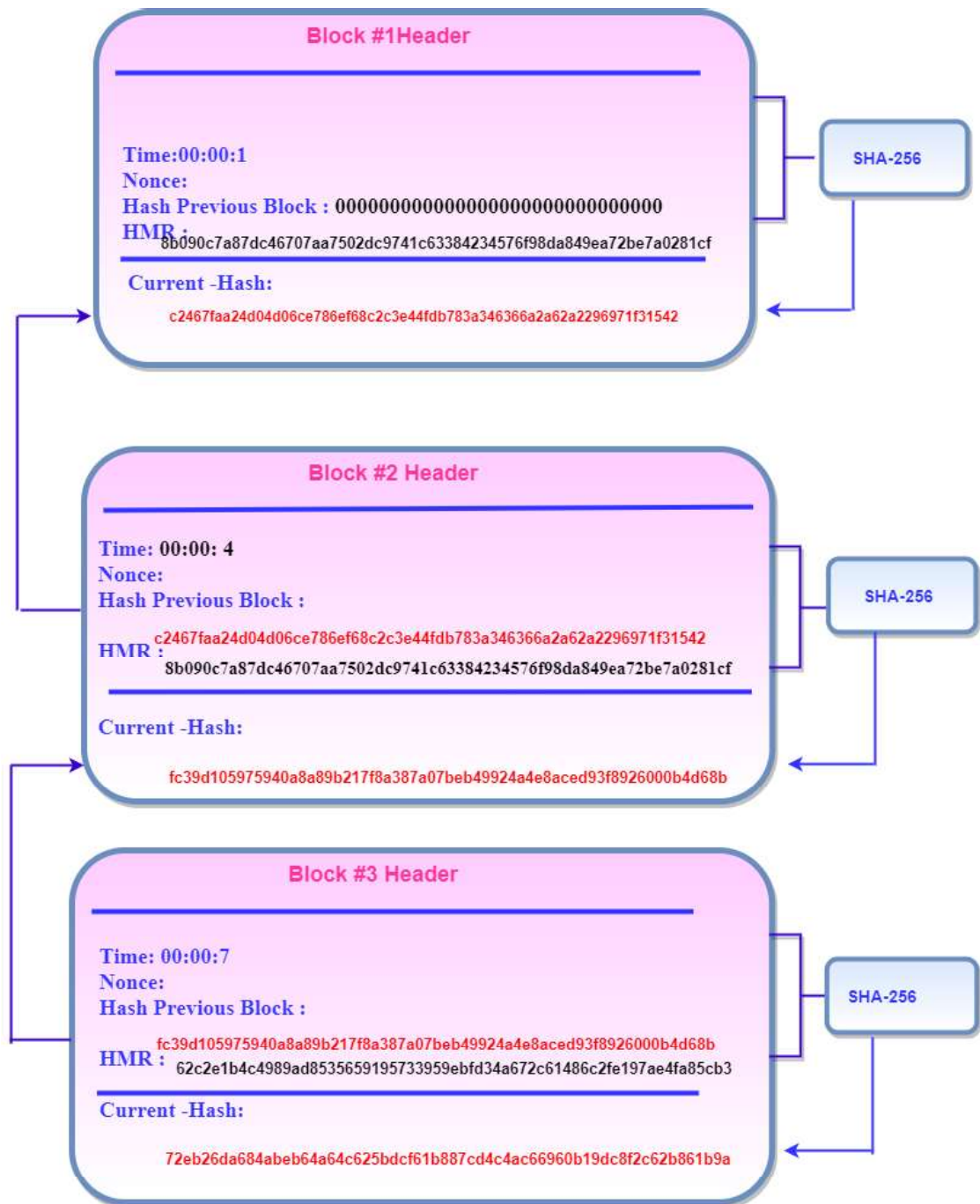


Figure 3.18 Example of the create block stage

CHAPTER 4

EXPERIMENTAL RESULTS

This chapter presents the results that were obtained using the proposed system. The method to the suggested approach abstracted the decisions made by the different learning models, passing the behavior of these models on to a binary logical decision layer, which produced the results. The training samples, the decision abstraction layer, and the final decision were all then used to perform an update of the overall logic of the newly proposed system. The experiments and results were divided into 2 sections:

Section 1: The use of machine learning to obtain a high-accuracy model can be accomplished by duplicating the training dataset or changing the training test segmentation of the data. This model used the output of the symptoms model in order to pick the disease. The lab test results information will be used as an input into the different models, where each model is related to some disease. Only the authorized doctor will be able to access the analysis results of a certain patient. He/she will also be allowed to review the original symptoms and lab test results using the cloud service.

Section 2: Blockchain cryptography: to secure the data and information flow between the agents and the users. Such as credential security, as remote authentication is required by all of the agents and users of the proposed system.

4.1 Section one: System Architecture

The general block diagram of the proposed system for the medical system software blueprint. Figure 4.1 shows the architecture

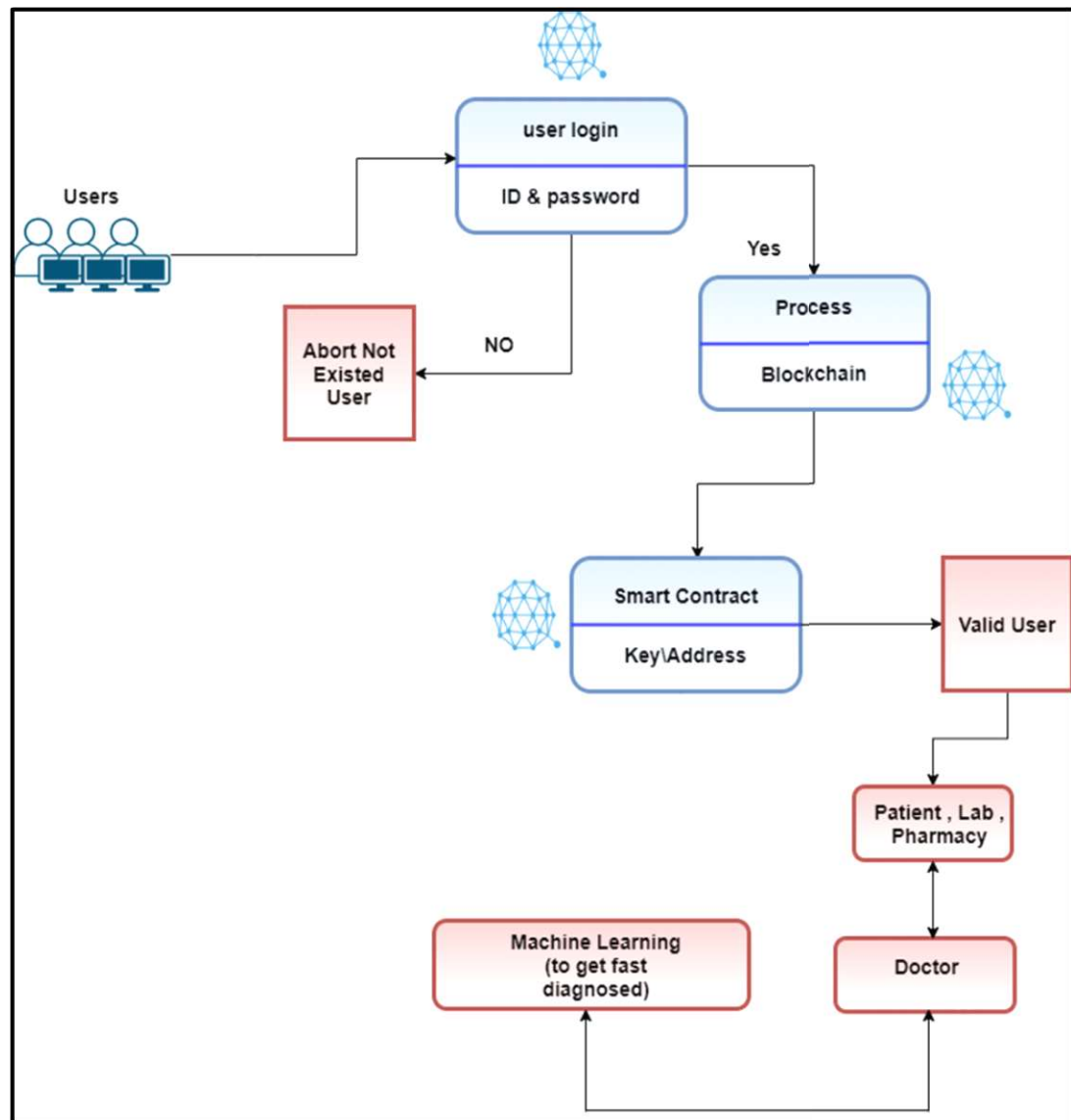


Figure 4.1 Medical assistant software design

There are 2 important categories of inputs:

- A. Symptoms: The patient, or any authorized person who is allowed to access the patient medical information, is able to access the cloud service. This privilege is given to them so that they can update the record of the patient with their symptoms.
- B. Lab information: Any authorized staff from the lab can enter the lab test results of the patient. Both of those entry categories will be stored in the cloud database and then

analyzed by matching them to some trained models. The matching process will indicate whether the symptoms and lab tests confirm the suspected disease or not.

Two separate models will be used, one is for the symptoms and the other is for the lab test results. The first symptom model can also be used to direct the lab staff and the doctor to the most suitable test. The symptom models are based on textual information and they are built using language models and syntax inference models. For example, when patient tells the system that she/he feels tired and loses weight, the symptom model will suggest a diabetes test as the first step.

A. For data acquisition:

1. Text mining: This is required to extract information from the symptoms and lab inputs.
2. Optical character recognition: This might be necessary if the input is an image of a prescription or lab test.

B. For decision models: The results of the datasets and configuration of the algorithms are used as input for the classification algorithms in machine learning and implemented.

4.1.1 Individual Model Results

Confusion matrix is a useful tool for evaluating the prediction accuracy of classifiers. In order to separate the 2 classes, the specificity, NPV, PPV, and sensitivity are all first computed by the use of the confusion matrix. The detection rate is the rate of true classifications and the prevalence of detection is the occurrence of expected accidents. Assume a table of 2×2 , with reference notation for an anticipated event, No event. As is shown in Table 4.1

Table 4.1 Confusion matrix or error matrix

	References	
Anticipated	Event	No event
Event	A	B
No event	C	D

Here, B is the number of times the model accepted false samples, C is the false rejected ones, while A and D are the number of correctly accepted and rejected iterations, respectively. The WDBC dataset obtained from the UCI [97] were randomly divided into 80%–20% for the training and testing procedure.

4.1.2 SVM Detection of Benign and Malignant

The classification results of the training set by SVM are given in Table 4.2. The SVM classifier had a nearly perfect accuracy value of 96.4% with the original features, without selecting the best feature or reducing the number of features the performance parameters of the SVM were accuracy: 0.964, specificity: 0.967, sensitivity: 0.958, PPV: 0.939, NPV: 0.978, prevalence: 0.343, rate of detection: 0.329, prevalence of detection: 0.350, and balanced accuracy (BA): 0.963. As is shown in Figure 4.2.

Table 4.2 Classification results for the benign and malignant values

Model	Prediction	Reference	
		Benign	Malignant
SVM	Benign	89	2
	Malignant	3	46

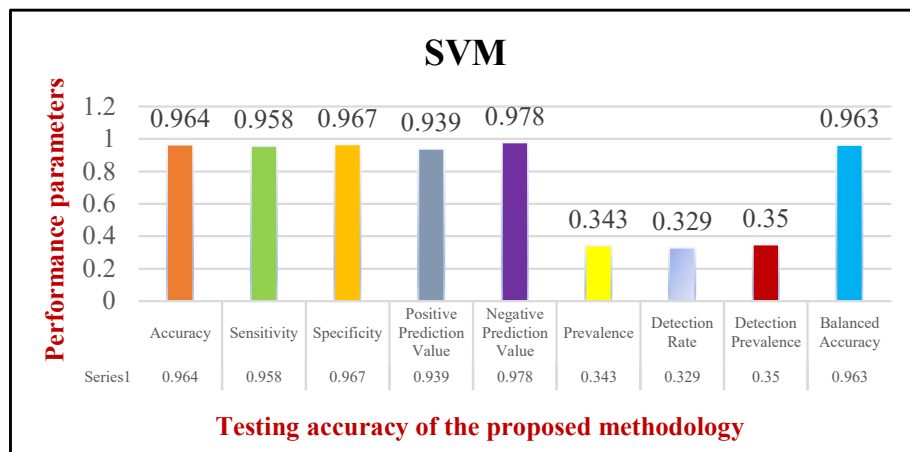


Figure 4.2 Testing accuracy using SVM.

4.1.3 Logistic Regression Detection of Benign and Malignant

After training LR for the breast cancer data, the results of the classification were displayed, which shows a detailed classification of the test samples are given in Table 4.3. The benign and reference columns represent the true situation, while the row values are the predicted ones. For a 100% accurate model, the LR classifier had a nearly perfect accuracy value of 96.4% with the original features, without selecting the best feature or reducing the number of features. The performance parameters of LR were accuracy: 0.964, specificity: 0.978, sensitivity: 0.938, PPV: 0.957, NPV: 0.968, prevalence: 0.343, rate of detection: 0.321, prevalence of detection: 0.336, and BA: 0.958. As shown in Figure 4.3.

Table 4.3 Classification results for the benign and malignant values models

Model	Prediction	Reference	
		Benign	Malignant
LR	Benign	90	3
	Malignant	2	45

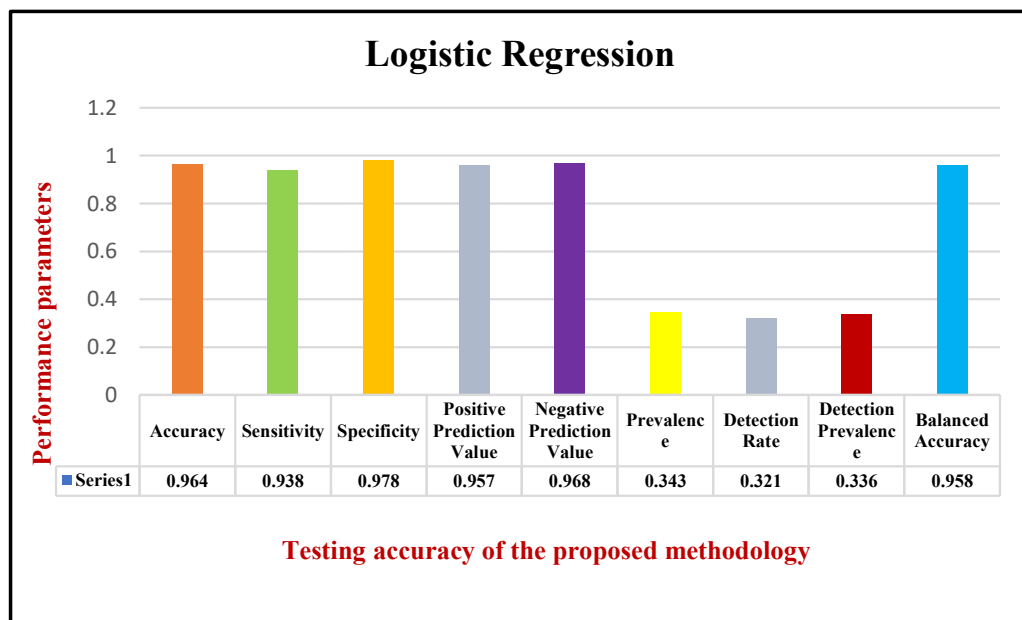


Figure 4.3 Testing accuracy using LR.

4.1.4 Decision Tree Detection of Benign and Malignant

In the experiment, 2 classes were used, and thus a 2×2 confusion matrix was applied, which achieved 95.7% accuracy as is shown in Table 4.4. The experiments were conducted on the model. The dataset was well trained for every classifier and from that, a model we obtain and then validation was performed using the test data, and outcomes were attained. After that, the obtained outcomes were computed and assessed, especially with regards to measurements such as accuracy. When the original features were used, without selecting the best feature or reducing the number of features, the performance parameters of DT were accuracy: 0.957, specificity: 0.957, sensitivity: 0.958, PPV: 0.920, NPV: 0.978, prevalence: 0.343, rate of detection: 0.329, prevalence of detection: 0.357, and BA: 0.957. As shown in Figure 4.4.

Table 4.4 Classification results for the benign and malignant values models

Model	Prediction	Reference	
		Benign	Malignant
DT	Benign	88	2
	Malignant	4	46

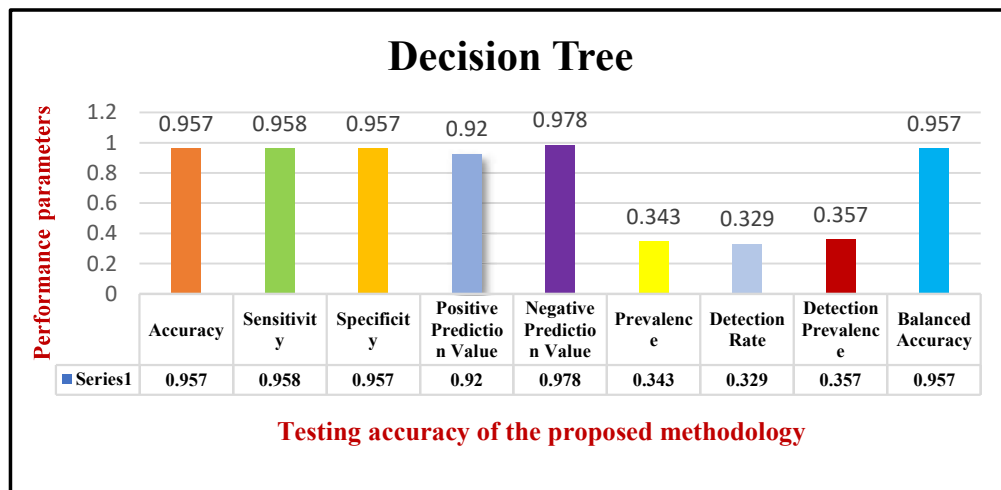


Figure 4.4: Testing accuracy using DT.

4.1.5 Random Forest Detection of Benign and Malignant

The results of the training classification set by the RF are given in Table 4.5. The identification of the most significant features within the dataset, conducting an analysis of the significance that these features had for many of the selected features, and testing the different sizes of the RFs, a final model was obtained with an accuracy of 97.9%. However, when building a base model for comparison before performing any analysis on the dataset, it is generally best to select a basic classification model, which can then be used as a point of reference to determine if a machine learning model was effective. When the original features were used, without selecting the best feature or reducing the amount of features, the performance parameters of the RF were accuracy: 0.979, specificity: 0.978, sensitivity: 0.979, PPV: 0.959, NPV: 0.989, prevalence: 0.343, rate of detection: 0.336, prevalence of detection: 0.350, and BA: 0.979. As show in Figure 4.5.

Table 4.5 Classification results for the benign and malignant values models

Model	Prediction	Reference	
		Benign	Malignant
RF	Benign	90	1
	Malignant	2	47

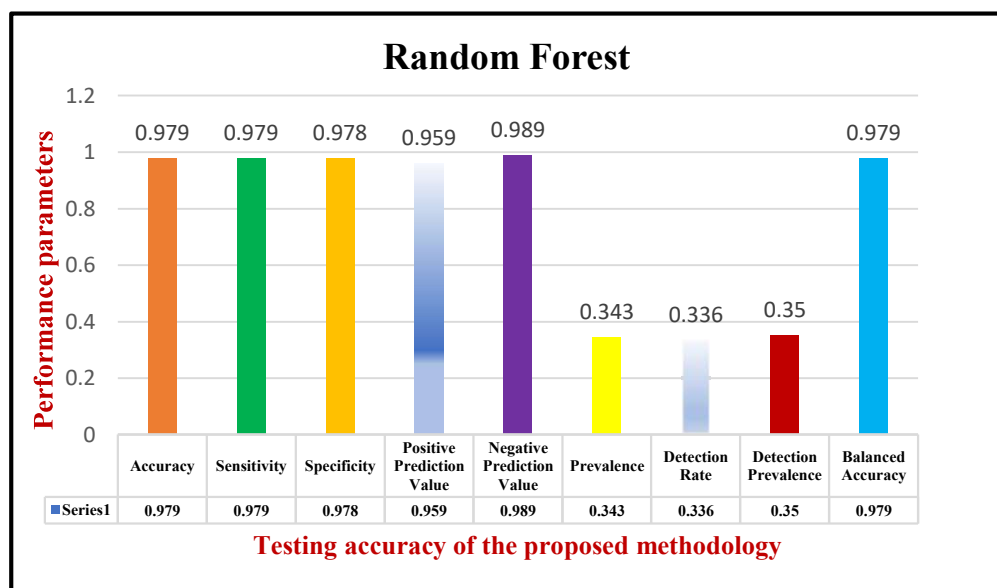


Figure 4.5 Testing accuracy using RF.

4.1.6 ANN Detection of Benign and Malignant

The classification results of the training set using an ANN are given in the Table 4.6. Due to features concerning the mechanism of information processing, the ANN was a significant and effective application. Moreover, ANNs can be successfully applied to a vast range of applications that use a large amount of data. ANNs have been used in many different fields as a result of their ability solve very complex problems such as classification, detection, prediction classification, and prediction, and has been described as highly accurate, reaching 95%. The classification results of the training set using an ANN are given in Table 4.5. When the original features were used, without selecting the best feature or reducing the number of features, the performance parameters of the ANN were accuracy: 0.95, specificity: 0.958, sensitivity: 0.946, PPV: 0.977, NPV: 0.902, prevalence: 0.657, rate of detection: 0.621, prevalence of detection: 0.636, and BA: 0.952, as show in Figure 4.6.

Table 4.6 Classification results for the benign and malignant values models

Model	Prediction	Reference	
		Benign	Malignant
ANN	Benign	87	2
	Malignant	5	46

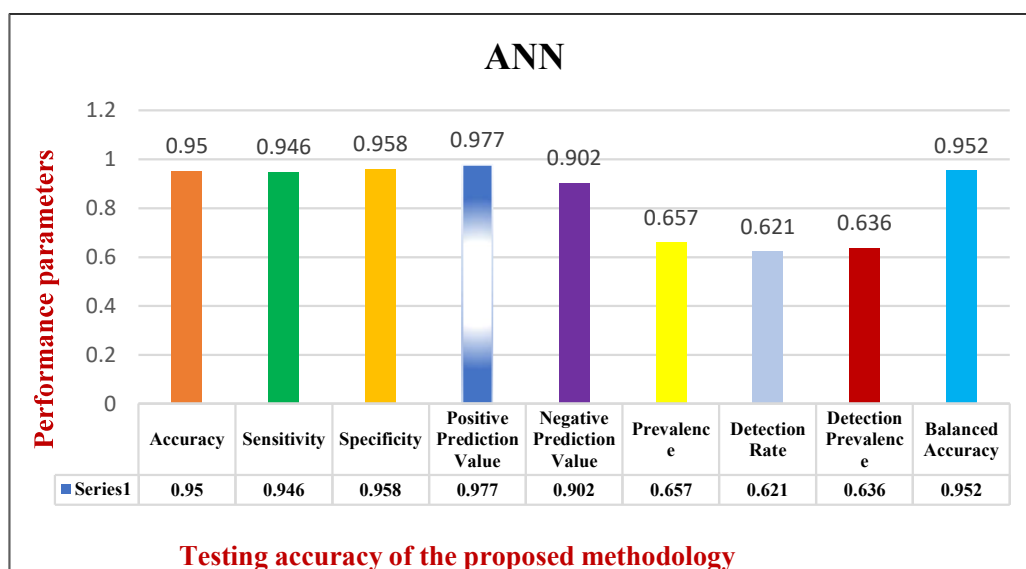


Figure 4.6 Testing accuracy using an ANN.

4.1.7 DNN Detection of Benign and Malignant

The classification results of the training set using a DNN are given in Table 4.7. The model was trained by feeding the inputs first and creating an output (predict) to be compared with the input label. When the original features were used, without selecting the best feature or reducing the number of features, the performance parameters of the DNN were accuracy: 0.97, specificity: 0.958, sensitivity: 0.978, PPV: 0.978, NPV: 0.958, prevalence: 0.657, rate of detection: 0.643, prevalence of detection: 0.657, and BA: 0.968, as show in Figure 4.7.

Table 4.7 Classification results for the benign and malignant values models

Model	Prediction	Reference	
		Benign	Malignant
DNN	Benign	90	2
	Malignant	2	46

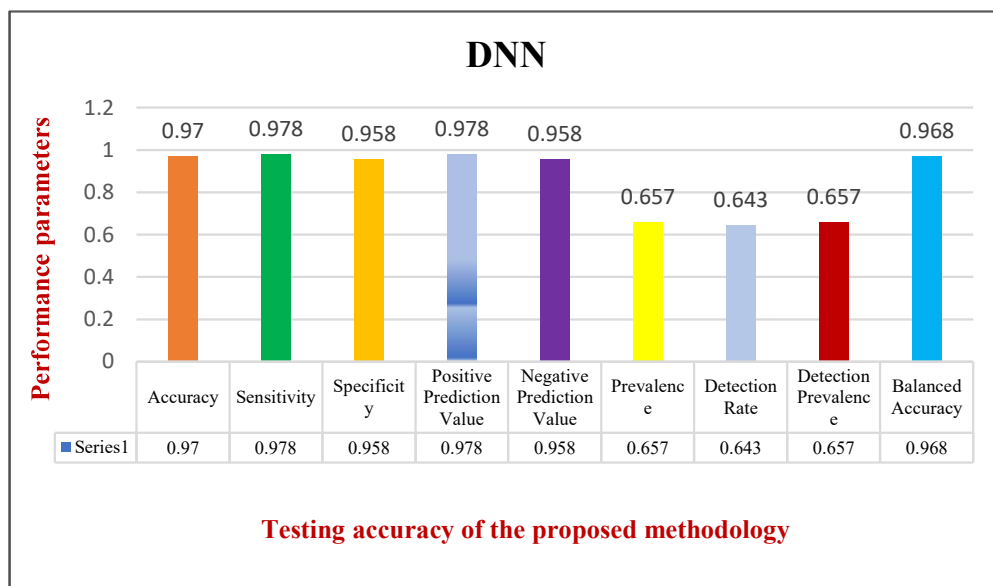


Figure 4.7 Testing accuracy using a DNN.

4.2 Logical Inference Systems

In this study, different models were used to give their decision in the first stage, and then, a logical inference layer was introduced. Figure 4.8 depicts the proposed logical inference system. The training set was used to train 6 different models, from A to F.

The decision abstraction layer converted the output of the models into 2 logical states: 1 for malignant and 0 for benign. Currently, the decision abstraction uses hard decision. The logical inference uses the training feeds to determine the final decision logical expression.

The sum-of-product (SOP) is a logical expression that is used to define the relationship between the inputs and the output when they have binary form. For instance, let the following truth in Table 4.8 show a function of 2 binary inputs, A and B.

Table 4.8 Function of 2 binary inputs

A	B	Output
0	0	0
0	1	1
1	0	1
1	1	0

Then, the SOP canonical expression is:

$$output = \underline{A}B + A\underline{B}, \quad (4.1)$$

which is the sum of the product of A and B, and means that the output is logic '1', only if A = '0' and B = '1' or vice versa.

According to the used WDBC dataset, the outputs of the 6 models ended up with the following SOP logical expression (after reduction using Boolean algebra using an online Karnaugh map solution):

$$Final\ Decision = ADEF' + AB'CD'E + AB'C'D'E'F' \quad (4.2)$$

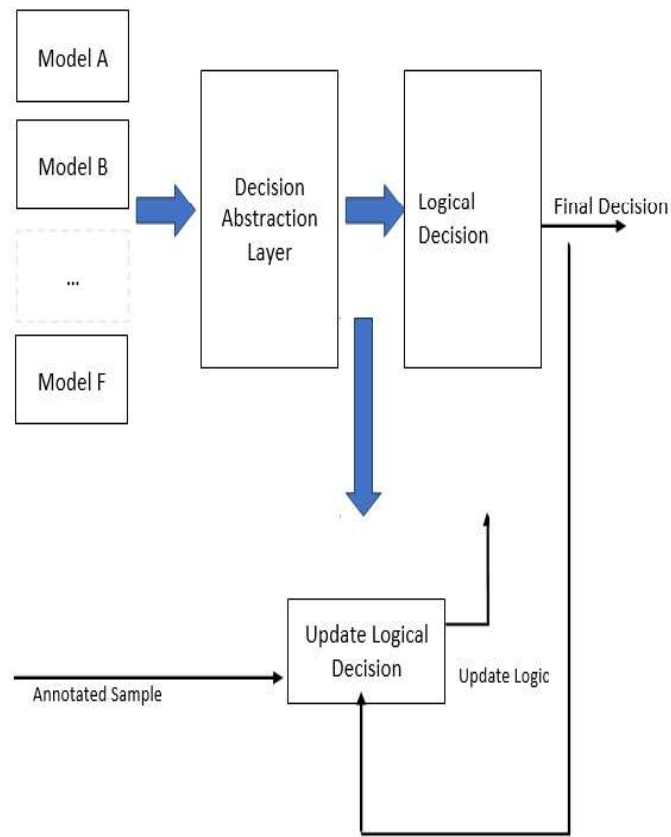


Figure 4.8 Multimodal logical inference system.

Here, A–F are the hard decision outputs of the models, where the multiplication among them is a logical AND, ‘+’ is a logical OR operation. Therefore, A–F is the hard decision of the prediction of the following models:

A: RF, B: LR, C: DT, D: SVM, E: ANN, and F: DNN. Figure 4.9 below depicts the final decision using logic gates.

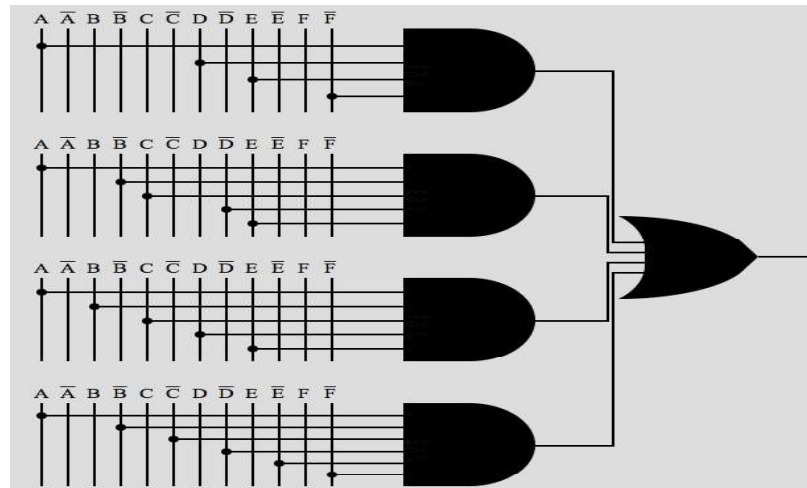


Figure 4.9 Logic gate representation of the final decision

Moreover, the obtained results of the machine learning techniques were compared, as displayed in Table 4.9.

Table 4.9 Matrix of confusion for all 6 of the models individually

Statistic	SVM	Logistic Regress ion	Decision Tree	Random Forest	ANN	DNN
Accuracy	0.964	0.964	0.957	0.979	0.950	0.971
Sensitivity	0.958	0.938	0.958	0.979	0.946	0.978
Specificity	0.967	0.978	0.957	0.978	0.958	0.958
Positive Prediction Value	0.939	0.957	0.920	0.959	0.977	0.978
Negative Prediction Value	0.978	0.968	0.978	0.989	0.902	0.958
Prevalence	0.343	0.343	0.343	0.343	0.657	0.657
Detection Rate	0.329	0.321	0.329	0.336	0.621	0.643
Detection Prevalence	0.350	0.336	0.357	0.350	0.636	0.657
Balanced Accuracy	0.963	0.958	0.957	0.979	0.952	0.968

These equations provide the accuracy ratio to the classified instances as illustrated in Figure 4.10 for the dataset.

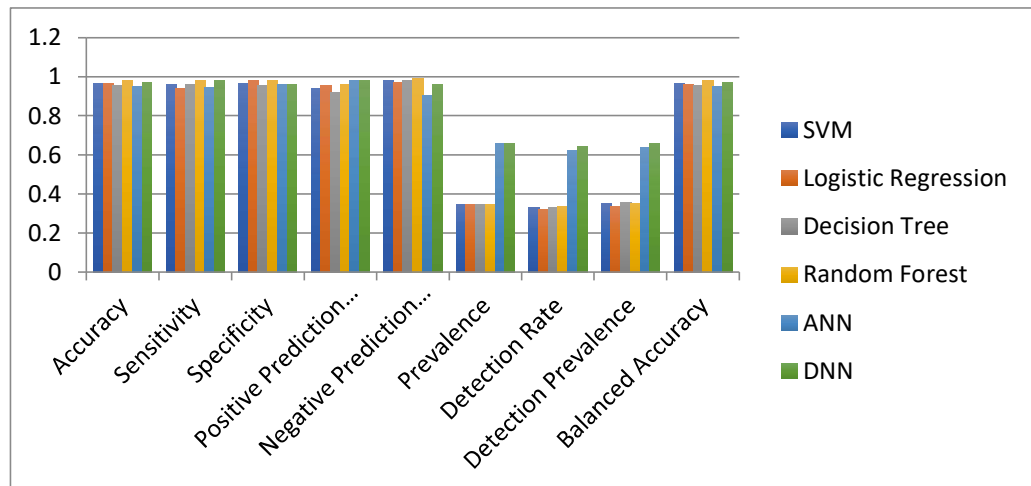


Figure 4.10 Confusion matrix chart for all 6 of the models.

In order to maintain 100% accuracy, any unobserved annotated sample could help to update the system. If a new observation for a known case state is available, the final decision is compared against its reference class, and then all of the logic decisions are updated. The proposed system in Figure 4.8 maintained 100% accuracy for the test part of the WBDC dataset. Without extracting any features or processes in the breast cancer data before training, the proposed inference system maximized the true acceptance and rejection ratio.

In addition to improving accuracy, the proposed logical decision can be easily updated if annotated samples exist. As it was based on a SOP canonical expression, updating the final decision rule can be as easy as adding another AND expression to the representation in Figure 4.9. The update procedure happens only if the proposed system misclassifies the annotated new sample.

4.3 Section Two: Blockchain cryptography

This thesis introduced a notion in term of providing protection for data used in the healthcare field, particularly, in an open-system environment (OSE) using a blockchain. Additionally, this work implemented proof of principle (proof-of-concept prototypes) to allow a user to have access to the healthcare system.

4.3.1 Key Generation

Below are the results of generating public keys using the logistic map for a 1-parameter function, when the input values for $x_1 = 0.1$, $a = (0.4]$, as shown in Figure 4.11, which were declared for the whole system.

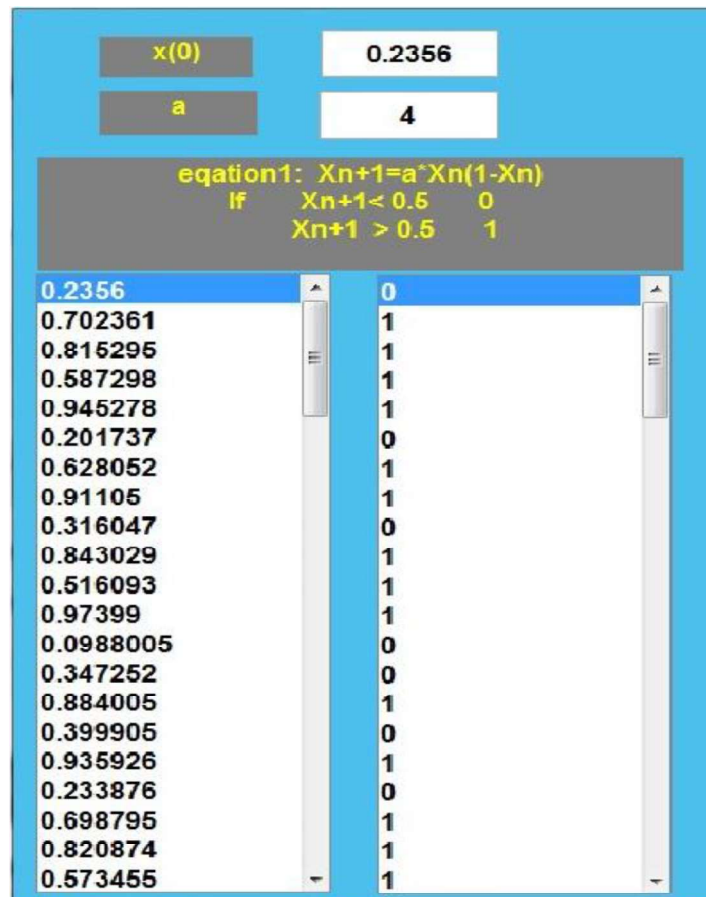


Figure 4.11 Logistic map with one parameter

These results for 2 of the parameters, and according to equation (3.16), were observed as in the table shown in Figure 4.12.

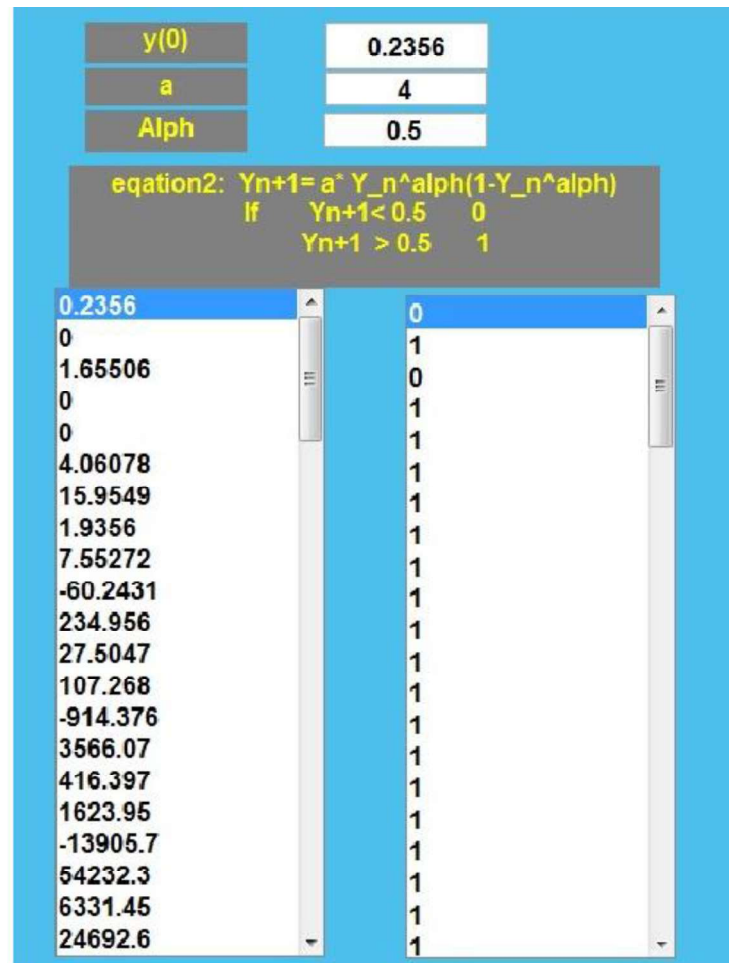


Figure 4.12 Logistic map with two parameters

Now, when making a hybrid between the 2 equations, i.e., equation (3.15) and (3.16), the results are as shown in Figure 4.13.

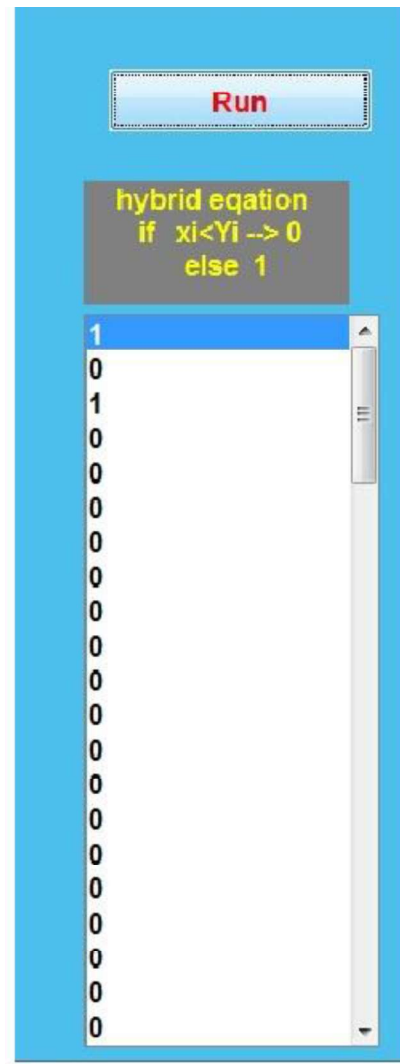


Figure 4.13 Logistic map with hybrid parameters.

4.3.2 Testing and Experimentation

This thesis introduced a notion in terms of providing protection for data used in the healthcare field, particularly, in an OSE using a blockchain. Additionally, this work implemented proof of principle (proof-of-concept prototypes) to allow a user to have access to the healthcare system. At the beginning, communicating must begin with the healthcare system. Here, the tool of OMNeT ++ was utilized for simulating that connection. The Figures 4.14 to 4.15 below illustrate the method used. Each node in the network is a peer, and as a result, each user within the blockchain can be regarded as a node and each user can be capable of performing the operation required for the transaction.

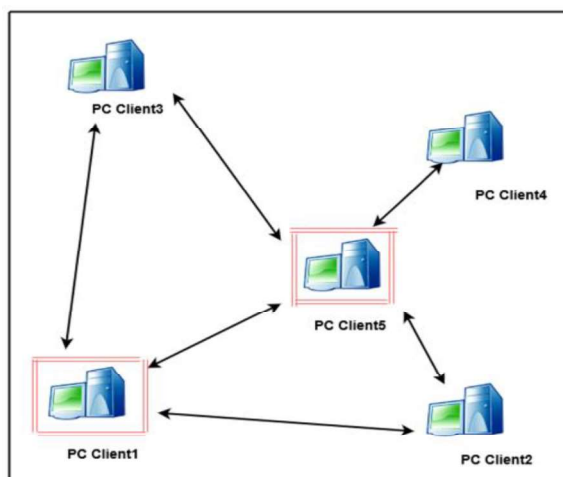


Figure 4.14 Blockchain peer- to -peer network.

Figure 4.14 illustrates the (P2P) blockchain network, which displays in what way communication is flowing among the nodes and thus how communication with a system occurs in a P2P manner in blockchain. Here, suppose that a PC client is a health center, and the other nodes are a lab, drugstore, office management, or hospital financial officer. Figure 4.15 illustrates the process of a transaction within a blockchain, and the red square indicates the transaction occurring between this node and another, just like the transaction occurring between a physician and a lab.

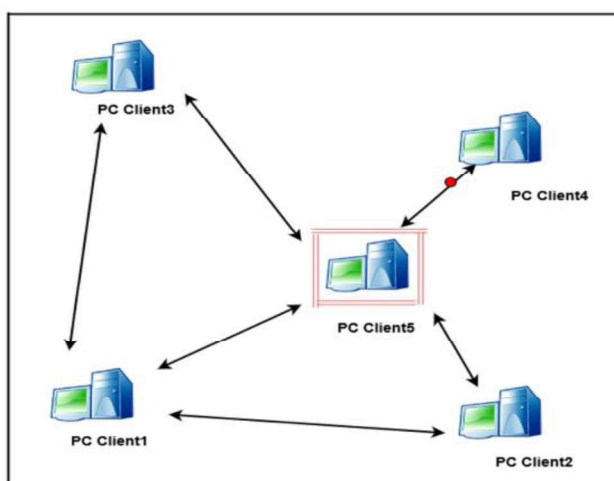


Figure 4.15 Transaction in a blockchain network.

When a user attempts to enter the healthcare system using a blockchain, he/she logs in using a password and username, and the blockchain will then conduct a validation of

this process to confirm that that particular user possesses valid access to the system. If the blockchain grants a user access to the network, the system can be used. If not, the user needs to register as a ‘new user’ in the block chain to verify the access to the system. Figure 4.16 illustrates a transacting process between a drugstore and a lab.

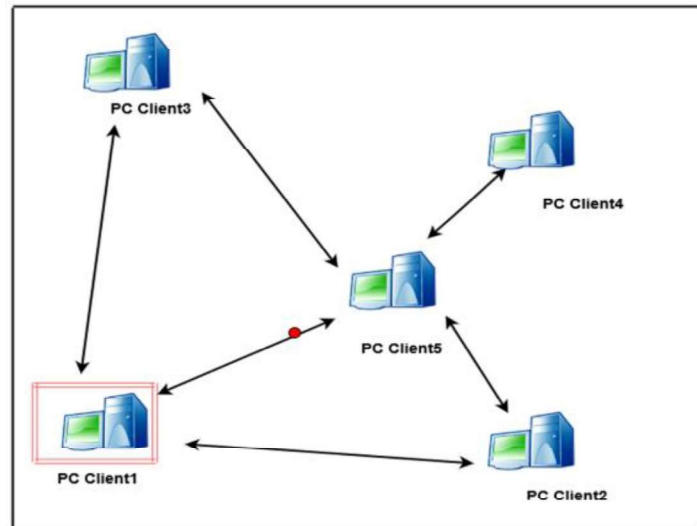


Figure 4.16 Pharmacy and laboratory transaction.

4.3.3 Healthcare Services Case Study

When a client report needs to be viewed by a physician, it is imperative that he/she is granted access to the system through authentication. For authentication, the methodology of the hash function should be used for granting access to the user, rather than using POW consensus. The solution that was proposed herein to create the hash function was used for verification, and this was due to its speed, where it was characterized by being quicker for the system than using POW consensus, since it might require 10 to 60 min. Under these circumstances, the methodology herein of a logistic map was used for the generation of a unique key, which relied on the assumption given below:

Initial. condition = (rnd.next double) * π X = time

Key = $g[\sum \text{rate} * a(X_n) * (1 - X_n) + \text{initial.condition}], g[\sum \text{rate} * a(Y_n^\alpha (1 - Y_n^\alpha)) + \text{initial.condition}]$

Here, $g[\sum \text{rate} * a(X_n) * (1 - X_n) + \text{initial.condition}], g[\sum \text{rate} * a(Y_n^\alpha (1 - Y_n^\alpha)) + \text{initial.condition}] =$

$$\begin{cases} 0 & \text{if } x_i < Y_i \\ 1 & \text{if } x_i \geq Y_i \end{cases}$$

As illustrated in Table 4.10, MATLAB code was written for generating a random number. Furthermore, a detailed explanation of the steps required for generating a random number were given. To determine the algorithm for the key generation, the sensitivity of the key was analyzed, since any changes to the key values, of any kind, leading to the creation of a new encryption, must be highly sensitive to the original key that was used in the algorithm. The modification of a single bit in the key will result in the keys being completely dissimilar. This proposed methodology was tested by utilizing a kit for statistical research from the National Institute of Technology and Standards (NITS).

The NITS has 15 tests to verify the randomness of the series and P-values were identified for each test.

Table 4.10 Random number generator.

N	X_i, Y_i	Generation	SHA256	Key
1	X_1, Y_1	$\sum \text{rate} * a(X_1) * (1 - X_1)$	9200e36f3c0494da7fd0cd 629f7b3dbd03b2f050c8e6 e2835d188ca7439d8564	0.21202418382945 665
		$\sum \text{rate} * (Y_1^\alpha (1 - Y_1^\alpha))$	2fbfa8c3fc6192aa41ae26c b0bb06c5856fdc9986ac1f 9485c1b4a3797fdf151	0.03292494192480 34
2	X_2, Y_2	$\sum \text{rate} * a(X_2) * (1 - X_2)$	57116edc290d051e4f7fd0 3abcd76e8baa7d369118a2 f5a16764fda95cf64cff	0.03827843716565 0991
		$\sum \text{rate} * (Y_2^\alpha (1 - Y_2^\alpha))$	fef2857b352d73232a7923 c9cbc087ff92b544612003 b2e234043551a6830dac	1.46577682166859 15E-07

3	X_3, Y_3	$\sum \text{rate} * a(X_3) * (1 - X_3)$	2ae368c1c8462f47f63e056b6a3cf27424dbb35ff21811d014f77f63c903a09c	0.01840659920690315
		$\sum \text{rate} * (Y_3^\alpha (1 - Y_3^\alpha))$	74da64115599e8354885556d8cd291cfddef5715cd da66a8a319d5feecdca90	1.0742508454704157E-14
.
.
∞	X_n, Y_n	$\sum \text{rate} * a(X_n) * (1 - X_n)$.	
		$\sum \text{rate} * (Y_n^\alpha (1 - Y_n^\alpha))$		

In the proposed technique, 2 adjusted logistic maps were used, each with dissimilar initial criteria and conditions, and 1 logistic map with respect to the key space show in the Figure 4.17. Thus, double precision of the floating-point data model was achieved dissimilar criteria with dissimilar tests show in Figure 4.18.

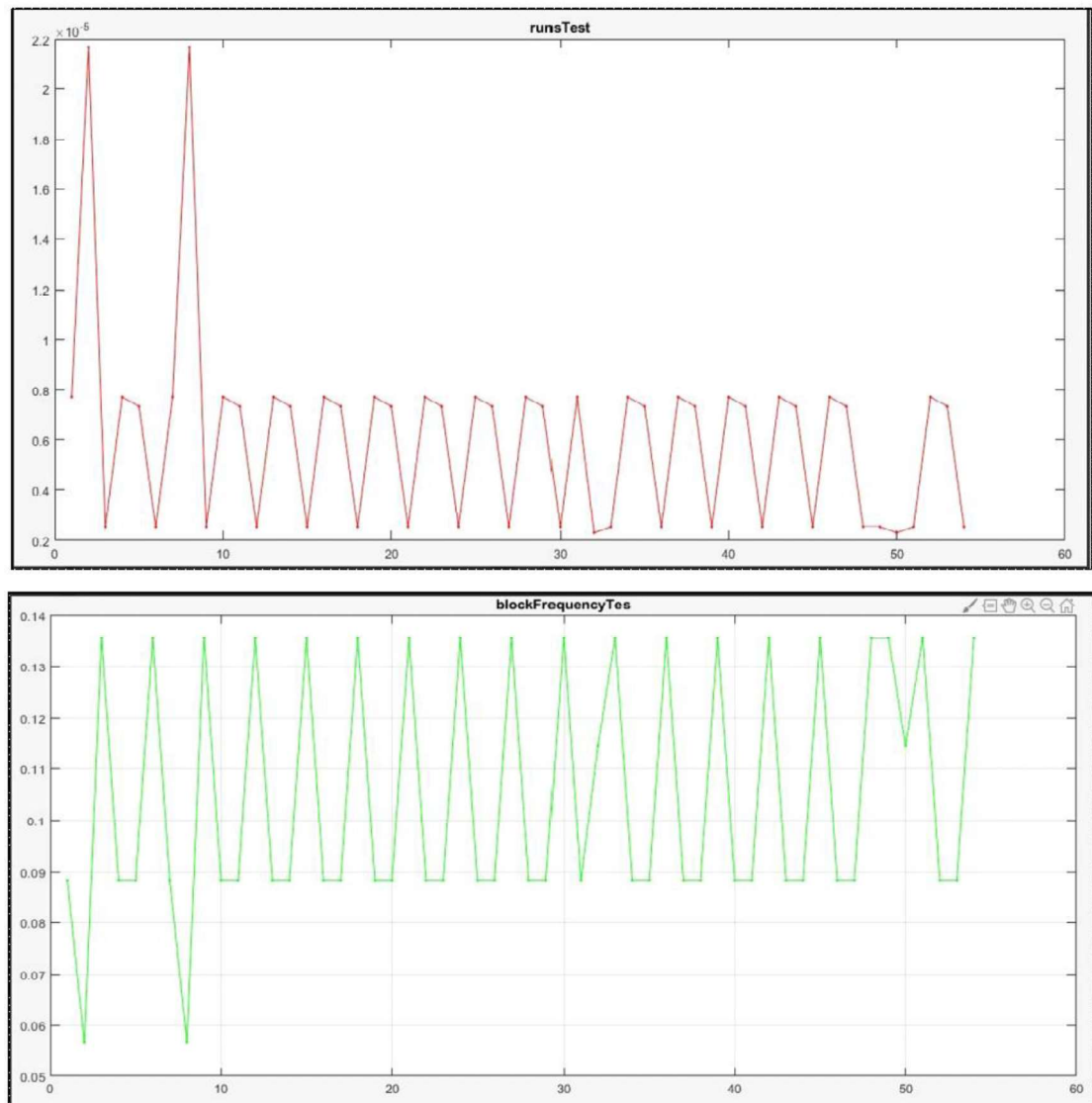


Figure 4.17 Statistical test 2: 'key space'

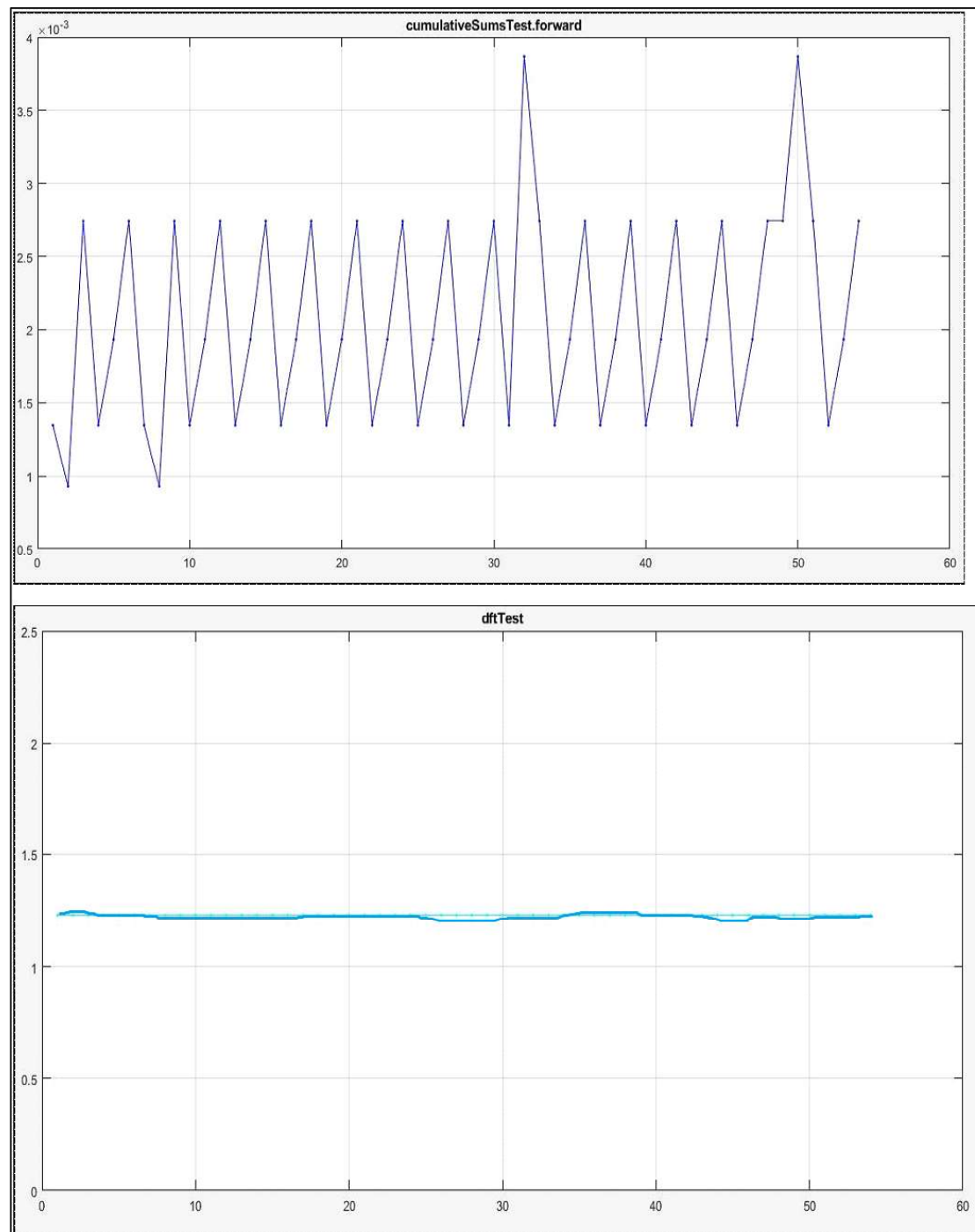


Figure 4.18 Statistical test two: ‘dissimilar criteria with dissimilar tests

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

5.1 Conclusions

The proposed system was summarized, and the following conclusions were taken from the collection of the test results. Some of these conclusions are listed in the following:

- A successfully built system and algorithm can reduce the cost of healthcare by providing secure and adaptive information. The algorithm can process the patient data in a reasonable way and produce effective decision-making criteria for doctors and other healthcare providers in real-time and should be secured, so that no one can temper with it, and every individual would be responsible for his/her own duty and initialization.
- The doctor's reports, patient information, hospital information, and previous health-related data will be used to extract useful results and decisions that help doctors with treatment and easily visualize the results of reports using the centralized and secure source, a blockchain. The system would provide real-time information to health providers and can result in real-time monitoring.
- The accuracy is expected to increase. In the current case, all of the models could be seen as if they had a 2-state single output, with the state '0' for benign and state '1' for malignant. Therefore, those 6 outputs were considered as digital inputs into a digital system. The training data were fed into the 6 models and their outputs were labeled using a binary code. Then, the best way to derive the final logical vote was to simplify the logical relationship using the SOP realization. In the test phase, the WDBC dataset attributes flowed through each model as usual. The individual decisions out of each model were labeled using the binary code: logic '0' for benign and '1' for malignant. Then, those binary encoded states were fed into the logical SOP realization to estimate the final decision. Adding a logical inference layer, that combined the results of the different machine learning models, improved the accuracy.

- Different algorithms were run in the first stage, and a logical voting scheme was applied to reduce the false rejection and acceptance ratios. The logical decision module had the ability of dynamic adaptation if a new annotated observation was fed into the system at any operating point.

The secure case, herein, a newly proposed framework that would protect healthcare systems, using a blockchain and P2P networks, was proposed. This work comprised both the design and the implementation of a secure healthcare system, making use of a novel blockchain method. The current work reduced the cost effectively.

An approach was developed that generated a unique key via implementation of the logistic map method. The key was then encrypted twice using SHA-256. Hence, the proof of the algorithm was also given.

5.2 Future work

During this work, the possible future work for machine learning and blockchain took several directions, as follows:

- A soft decision can be tested and compared to the current hard decision. Moreover, the effect of reducing the number of machines learning models in the first stage can be studied and compared.
- New approaches for the health services framework (e.g., Cloud, online apps, and Mashups) and new problems in decreasing costs at this stage will be implemented. In investigating the existence of threats in said networks, semantic and latent semantic spaces play an important role.
- Using various machine learning techniques (2 or more) for the proposed system to obtain a better accuracy rate.
- Developing a new different method for key generation in logistic map using blockchain security. A hybrid of 3 or 4 key generation equations can be created, which would generate a very unique key using the logistic map method.
- This work is based on a blockchain framework, developing a payment service within the full blockchain platform, including all payment transactions bitcoins.

REFERENCES

- [1] Anderson, B. O., Yip, C. H., Ramsey, S. D., et al. “Breast cancer in limited-resource countries: health care systems and public policy”. *The Breast Journal*, 12(1), S54–S69. 2006
- [2] Parkin, D. M., Pisani, P. & Ferlay, J., “Global cancer statistics, 2002”. *CA: A Cancer Journal for Clinicians*, 55(2), 74–108. 2005
- [3] Shahnaz, A., Qamar, U. & Khalid, A., “Using blockchain for electronic health records”. *IEEE Access*, 7, 147782–147795. 2019
- [4] Tith, D., Lee, J. S., Suzuki, H., Wijesundara, W. M. A. B., Taira, N., Obi, T., & Ohyama, N. “Application of blockchain to maintaining patient records in electronic health record for enhanced privacy, scalability, and availability.” *Healthcare informatics research*, 26(1), 3. 2020
- [5] Achi, I. I., Inyama, H. C., Bakpo, F. F. & Agwu, C. O., “Machine learning based on intelligent tutoring system”, *International Journal of Engineering Trends and Technology*, 26(4), 2015
- [6] Curtes, A. A., “Plots, transformations and regression; an introduction to graphical methods of diagnostic regression analysis”. No. 04; QA278. 2, A8. 1985
- [7] Hosmer Jr., D. W., Lemeshow, S. & Sturdivant, R. X., “Applied logistic regression”, Vol. 398. John Wiley & Sons, New York, 2013
- [8] Polat, K. & Günes, S., “Breast cancer diagnosis using least square support vector machine”. *Digital Signal Processing*, 11, 694–701. 2007
- [9] Keles, A., Keles, A. & Yavuz, U., “Expert system based on neuro-fuzzy rules for diagnosis breast cancer”. *Expert Systems with Applications*, 38, 5719–5726. 2011

- [10] Robert, D., “What is the right organization structure? Decision tree analysis provides the answer”. *Organizational Dynamics* 7(3), 59–80. 1979
- [11] Zhang, W., “A comparative study of ensemble learning approaches in the classification of breast cancer metastasis”, 2009 International Joint Conference on Bioinformatics Systems Biology and Intelligent Computing, 242-245. 2009
- [12] Lakshmanaprabu, S. K., Shankar, K., Ilayaraja, M., Nasir, A.W., Vijayakumar, V. & Chilamkurti, N., “Random forest for big data classification in the internet of things using optimal features”. *International Journal of Machine Learning and Cybernetics* 1–10. 2019
- [13] Breiman, L., “Using adaptive bagging to debias regressions”, Technical Report 547, Statistics Dept. UCB, 1999
- [14] Jiang, R., Yang, H., Sun, F. & Chen, T., “Searching for interpretable rules for disease mutations: simulated annealing strategy”, *BMC Bioinformatics*, 7, 417. 2006
- [15] Jiang, R., Tang, W., Wu, X. & Fu, W., “A random forest approach to the detection of epistatic interactions in case-control studies”, *BMC Bioinformatics*, 10(1): S65. 2009
- [16] Yanli, L., Wang, Y. & Zhang, J., “New machine learning algorithm: random forest”. *International Conference on Information Computing and Applications*, (246–252). Springer, Berlin, Heidelberg, 2012
- [17] Mu, T. T. & Nandi, A. K., “Breast cancer detection from FNA using SVM with different parameter tuning systems and SOM-RBF classifier”, *Journal of the Franklin Institute*, 344, 285–311. 2007
- [18] Krishnan, M. M. R., Banerjee, S., Chakraborty, C., Chakraborty, C. & Ray, A. K., “Statistical analysis of mammographic features and its classification using support vector machine”. *Expert Systems with Applications*, 37, 470–478. 2010

- [19] Gholami, V. & Khaleghi, M. R., “A comparative study of the performance of artificial neural network and multivariate regression in simulating springs discharge in the Caspian Southern Watersheds”, *Applied Water Science*, 9(1),1-10. 2019
- [20] Lee, S. ,” Artificial Neural Networks”,
<https://www.google.com/url?sa=i&url=https%3A%2F%2Fgroup.March>
8,2019
- [21] Jeong, T., “Deep neural network algorithm feedback model with behavioral intelligence and forecast Accuracy”, September 2020
- [22] Benmeziane, H., “Comparison of deep learning frameworks and compilers”, Thesis, June 2020
- [23] Esposito, C., Santis, A., Tortora, G., Chang, H. & Choo, K., “Blockchain: a panacea for healthcare cloud-based data security and privacy”, *IEEE Cloud Computing*, 5(1), 31–37. 2018
- [24] Hanke, T. & Lerner, S. D., “Block mining methods and apparatus”, U.S. Patent Application No. 15/141,063. 2017
- [25] Crosby, M., Pattanayak, P., Verma, S. & Kalyanaraman, V., “Blockchain technology: beyond bitcoin”, *Applied Innovation*, 2(6–10), 71. 2016
- [26] Tasca, P. & Tessone, C. J., “A taxonomy of blockchain technologies: principles of identification and classification”, *Ledger*, 4, 1–39. 2019
- [27] Makridakis, S. & Christodoulou, K., “Blockchain: current challenges and future prospects/applications”, *Institute for the Future, University of Nicosia*, 11(12), 258. 2019
- [28] Eyal, I., Gencer, A., Sirer, E. & Renesse, R., “Bitcoin-ng: a scalable blockchain protocol”. In: *The 13thth USENI X Symposium on Networked Systems Design and Implementation*, pp. 45–59. 2016

- [29] Mann, O. & Shteingart, Z., “System and method for providing shared hash engine architecture for a bitcoin blockchain”, U.S. Patent Application No. 15/513,175. 2017
- [30] Shiong, P., Kupwade-Patil, H., Seshadri, R. & Witchey, N. J., “Homomorphic Encryption in a healthcare network environment, system, and methods”, 2019
- [31] Pareek, N., Patidar, V., & Sud, K., “Image encryption using a chaotic logistic map”. *Image and Vision Computing*, 24(9), 926–934. 2006
- [32] Darwish, A., Hassanien, A., Elhoseny, M., Sangaiah, A. & Muhammad, K., “The impact of the hybrid platform of the internet of things and cloud computing on healthcare systems: opportunities, challenges, and open problems”, *Journal of Ambient Intelligence and Humanized Computing*, 10(10), 4151–4166. 2019
- [33] Sekar, J. & Arun, C., “Comparative performance analysis of chaos-based image encryption techniques”, *Journal of Critical Reviews*, 7(9), 2020
- [34] Moysis, L., Tutueva, A., Volos, C., Butusov, D., Munoz-Pacheco, J. M., & Nistazakis, H., “A two-parameter modified logistic map and its application to random bit generation”, 12(5), 829. 2020
- [35] Nestor, N., De Dieu, D., Jacques, E., Iliyasu, Y. A. & El-Latif, A.A., “A multidimensional hyperjerk oscillator: dynamics analysis, analogue and embedded systems implementation, and its application as a cryptosystem”, *Sensors*, 20(1), 83. 2019
- [36] Raghupathi, W. & Raghupathi, V., “Big data analytics in healthcare: promise and potential”, *Health Information Science and Systems*, 2(1), 3. 2014
- [37] Auffray, C., Balling, R., Barroso, I., Bencze, L., Benson, M., Bergeron, J., et al., “Making sense of big data in health research: towards an EU action plan. *Genome Medicine*, 8(1), 71. 2016
- [38] Ward, M. J., Marsolo, K. A. & Froehle, C. M., “Applications of business analytics in healthcare”. *Business Horizons*, 57(5), 571–582. 2014

- [39] Dinov, I. D., “Volume and value of big healthcare data”. *Journal of Medical Statistics and Informatics*, 4. 2016
- [40] Lee, C. H. & Yoon, H. J., “Medical big data: promise and challenges”. *Kidney Research and Clinical Practice*, 36(1), 3. 2017
- [41] Tan, S. S. L., Gao, G. & Koch, S., “Big data and analytics in healthcare”. *Methods of Information in Medicine*, 54(6), 546–547. 2015
- [42] Liyanage, H., De Lusignan, S., Liaw, S.T., Kuziemy, C., Mold, F., Krause, P., et al., “Big data usage patterns in the health care domain: A use case driven approach applied to the assessment of vaccination benefits and risks: contribution of the times primary healthcare working group”. *Yearbook of Medical Informatics*, 9(1), 27. 2014
- [43] Roski, J., Bo-Linn, G. W. & Andrews, T. A., “Creating value in health care through big data: opportunities and policy implications”, *Health Affairs*, 33(7), 1115–1122. 2014
- [44] Kruse, C. S., Goswamy, R., Raval, Y. & Marawi, S., “Challenges and opportunities of big data in health care: a systematic review”, *JMIR Medical Informatics*, 4(4), e38. 2016
- [45] Ghani, K. R., Zheng, K., Wei, J. T. & Friedman, C. P., “Harnessing big data for health care and research: are urologists ready?”, *European Urology*, 66(6), 975–977. 2014
- [46] Baro, E., Degoul, S., Beuscart, R. & Charizard, E., “Toward a literature-driven definition of big data in healthcare”, *BioMed Research International*, 2015
- [47] Swan, M., “The quantified self: fundamental disruption in big data science and biological discovery”, *Big Data*, 1(2), 85–99. 2013
- [48] Mehta, N. & Pandit, A., “The concurrence of big data analytics and healthcare: a systematic review”, *International Journal of Medical Informatics*, 114, 57–65. 2018

- [49] Huang, T., Lan, L., Fang, X., An, P., Min, J. & Wang, F., “Promises and challenges of big data computing in health sciences”, *Big Data Research*, 2(1), 2–11. 2015
- [50] Islam, S., Hasan, M., Wang, X., Germack, H. D., et al., “A systematic review on healthcare analytics: application and theoretical perspective of data mining”, In: *Healthcare*, Vol. 6, pp. 54. Multidisciplinary Digital Publishing Institute, 2018
- [51] Mohammed, E. A., Far, B. H. & Naugler, C., “Applications of the MapReduce programming framework to clinical big data analysis: current landscape and future trends”, *BioData Mining*, 7(1), 22. 2014
- [52] Wang, W., Hadrian, K., Salmasian, H., Harpaz, R., Chase, H. & Friedman, C., “A drug-adverse event extraction algorithm to support pharmacovigilance knowledge mining from Pubmed citations”, In: *AMIA Annual Symposium Proceedings*, Vol. 2011, pp. 1464. American Medical Informatics Association, 2011
- [53] Hung, C. L. & Lin, Y.L., “Implementation of a parallel protein structure alignment service on the cloud”, *International Journal of Genomics*, 2013.
- [54] Wang, L., Chen, D., Ranjan, R., Khan, S. U., Kolodziej, J. & Wang, J., “Parallel processing of massive EEG data with MapReduce”, In: *Proceedings of the IEEE 18th International Conference on Parallel and Distributed Systems*, pp. 164–171. 2012
- [55] Meng, B., Prax, G. & Xing, L., “Ultrafast and scalable cone-beam CT reconstruction using MapReduce in a cloud computing environment”, *Medical Physics*, 38(12), 6603–6609. 2011
- [56] Karaolis, M. A., Moutiris, J. A., Hadjipanayi, D. & Pattichis, C. S., “Assessment of the risk factors of coronary heart events based on data mining with decision trees”, *IEEE Transactions on Information Technology in Biomedicine*, 14(3), 559–566. 2010

- [57] Tsipouras, M. G., Exarchos, T. P., Fotiadis, D. I., Kotsia, A. P., Vakalis, K. V., Naka, K. K. & Michalis, L. K., “Automated diagnosis of coronary artery disease based on data mining and fuzzy modeling”, *IEEE Transactions on Information Technology in Biomedicine*, 12(4), 447–458. 2008
- [58] Nguyen, T., Khosravi, A., Creighton, D. & Nahavandi, S., “Classification of healthcare data using the genetic fuzzy logic system and wavelets”, *Expert Systems with Applications*, 42(4), 2184–2197. 2015
- [59] Vock, D. M., Wolfson, J., Bandyopadhyay, S., Adomavicius, G., Johnson, P. E., Vazquez-Benitez, G. & O’Connor, P. J., “Adapting machine learning techniques to censored time-to-event health record data: a general-purpose approach using the inverse probability of censoring weighting”, *Journal of Biomedical Informatics*, 61, 119–131. 2016
- [60] Razavian, N., Blecker, S., Schmidt, A. M., Smith-McLallen, A., Nigam, S. & Sontag, D., “Population-level prediction of type 2 diabetes from claims data and analysis of risk factors”, *Big Data*, 3(4), 277–287. 2015
- [61] Yeh, W. C., Chang, W. W. & Chung, Y. Y., “A new hybrid approach for mining breast cancer patterns using discrete particle swarm optimization and statistical method”, *Expert Systems with Applications*, 36(4), 8204–8211. 2009
- [62] Chauhan, R. & Kumar, A., “Cloud computing for improved healthcare: techniques, potential, and challenges”, In: *Proceedings of the IEEE International Conference on E-Health and Bioengineering*, pp. 1–4. 2013
- [63] Delen, D., “Analysis of cancer data: a data mining approach”, *Expert Systems*, 26(1), 100–112. 2009
- [64] Kim, S., Kim, W. & Park, R.W., “A comparison of intensive care unit mortality prediction models through the use of data mining techniques”, *Healthcare Informatics Research*, 17(4), 232–243. 2011

- [65] Lee, J., Maslov, D. M. & Dubin, J.A., “Personalized mortality prediction is driven by electronic medical data and a patient similarity metric”, *PloS One*, 10(5), e0127428, 2015.
- [66] Post, A. R., Kurc, T., Cholleti, S., Gao, J., Lin, X., Bornstein, W., et al., “The Analytic Information Warehouse (AIW): a platform for analytics using electronic health record data”, *Journal of Biomedical Informatics*, 46(3), 410–424. 2013
- [67] Zhang, Y., Qiu, M., Tsai, C. W., Hassan, M. M. & Alamri, A., “Health-CPS: healthcare cyber-physical system assisted by cloud and big data”, *IEEE Systems Journal*, 11(1), 88–95. 2017
- [68] Bertsimas, D., Bjarnadóttir, M. V., Kane, M. A., Kryder, J. C., Pandey, R., Vempala, S. & Wang, G., “Algorithmic prediction of health-care costs”, *Operations Research*, 56(6), 1382–1392. 2008
- [69] Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A. & Donaldson, L., “Use of sentiment analysis for capturing patient experience from free-text comments posted online”, *Journal of Medical Internet Research*, 15(11), 2013
- [70] Glowacka, K. J., Henry, R. M. & May, J. H., “A hybrid data mining/simulation approach for modeling outpatient no-shows in clinic scheduling”, *Journal of the Operational Research Society*, 60(8), 1056–1068. 2009
- [71] Callahan, A., Pernik, I., Stiglic, G., Leskovec, J., Strasberg, H. R. & Shah, N. H., “Analyzing information seeking and drug-safety alert response by health care professionals as new methods for surveillance”, *Journal of Medical Internet Research*, 17(8), 2015
- [72] Youssef, A. E., “A framework for secure healthcare systems based on big data analytics in mobile cloud computing environments. *International Journal of Ambient Systems and Applications*, 2(2), 1–11. 2014

- [73] Yang, W. S. & Hwang, S. Y., “A process-mining framework for the detection of healthcare fraud and abuse”, *Expert Systems with Applications*, 31(1), 56–68. 2006
- [74] Diederich, J., Al-Ajmi, A. & Yellowlees, P., “Ex-ray: data mining and mental health”, *Applied Soft Computing*, 7(3), 923–928. 2007
- [75] Carús Candás, J. L., Peláez, V., López, G., Fernández, M. Á., Álvarez, E. & Díaz, G., “An automatic data mining method to detect abnormal human behavior using physical activity measurements”, *Pervasive and Mobile Computing*, 15, 228–241. 2014
- [76] Nimmagadda, S. L. & Dreher, H. V., “On robust methodologies for managing public health care systems”, *International Journal of Environmental Research and Public Health*, 11(1), 1106–1140. 2014
- [77] Pur, A., Bohanec, M., Cestnik, B., Lavrač, N., Debeljak, M. & Kopač, T., “Data mining for decision support: an application in public health care”, In: *International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems*, pp. 459–469. Springer, 2005
- [78] Harpaz, R., Vilar, S., DuMouchel, W., Salmasian, H., Hadrian, K., Shah, N. H., Chase, H. S. & Friedman, C., “Combing signals from spontaneous reports and electronic health records for detection of adverse drug reactions”, *Journal of the American Medical Informatics Association*, 20(3), 413–419. 2012
- [79] Heeley, E., Wilton, L. V. & Shakir, S. A., “Automated signal generation in prescription-event monitoring”, *Drug Safety*, 25(6), 423–432. 2002
- [80] Sakaeda, T., Kadoyama, K. & Okuno, Y., “Statin-associated muscular and renal adverse events: data mining of the public version of the FDA adverse event reporting system”, *PloS One*, 6(12), e28124. 2011
- [81] Eriksson, R., Werge, T., Jensen, L. J. & Brunak, S., “Dose-specific adverse drug reaction identification in electronic patient records: temporal data mining in an inpatient psychiatric population”, *Drug Safety*, 37(4), 237–247. 2014

- [82] Kuo, S. T., Kim, H. E. & Ohno-Machado, L., “Blockchain distributed ledger technologies for biomedical and health care applications”, *Journal of the American Medical Informatics Association*, 24(6), 1211–1220. 2017
- [83] Ivan, D., “Moving toward a blockchain-based method for the secure storage of patient records”, In: *ONC/NIST Use of Blockchain for Healthcare and Research Workshop*. Gaithersburg, Maryland, United States: ONC/NIST, 1-11. 2016
- [84] Kuo, T. T. & Ohno-Machado, L., “Modelchain: decentralized privacy-preserving healthcare predictive modeling framework on private blockchain networks”, *CoRR*, abs/1802.01746, 2018
- [85] Benchoufi, M., Porcher, R. & Ravaud, P., “Blockchain protocols in clinical trials: transparency and traceability of consent. *F1000Research*, 6, 2017.
- [86] Yang, H. & Yang, B., “A blockchain-based approach to the secure sharing of healthcare data”, In: *Proceedings of the Norwegian Information Security Conference*, 100-111. 2017
- [87] Clauson, K. A., Breeden, E. A., Davidson, C. & Mackey, T. K., “Leveraging blockchain technology to enhance supply chain management in healthcare”, *Blockchain in Healthcare Today*, 1(3), 1-12. 2018
- [88] Abouelmehdi, K., Beni-Hssane, A., Khaloufi, H. & Saadi, M., “Big data security and privacy in healthcare: a review”, *Procedia Computer Science*, 113, 73–80. 2017
- [89] De Filippi, P., “The interplay between decentralization and privacy: the case of blockchain technologies”, *Journal of Peer Production*, 7, 2016
- [90] Brodersen, C., Kalis, B., Leong, C., Mitchell, E., Pupo, E., Truscott, A. & Accenture, L. L. P., “Blockchain: securing a new health interoperability experience”, Accenture, L. L. P. (ed.), 1-11. 2016
- [91] Ohno-Machado, L., Bafna, V., Boxwala, A. A., Chapman, B. E., Chapman, W. W., Chaudhuri, K., et al., “Dash: integrating data for analysis, anonymization,

- and sharing”, *Journal of the American Medical Informatics Association*, 19(2), 196–201. 2011
- [92] Mangasarian, O. L., Setiono, R. & Wolberg, W. H., “Pattern recognition via linear programming: theory and application to medical diagnosis”, In: *Large-scale Numerical Optimization*, Philadelphia, 1990.
- [93] Ronak, S., Vishnusri, N. & Jeyalatha, S., “Diagnosis of breast cancer using decision tree data mining technique”, *International Journal of Computer Applications*, 98(10), 2014
- [94] Yu, D. & Le, D., “Automatic speech recognition, a deep learning approach”, London: Springer-Verlag, 2015, ISBN: 978-1-4471-5778-6. DOI: 10.1007/ 978-1-4471-5779-3.
- [95] Sadkhan, S. B., Al-Sherbaz, A., & Mohammed, R. S. Chaos based cryptography for voice encryption in wireless communication. In 2013 International Conference on Electrical Communication, Computer, Power, and Control Engineering (ICECCPCE) (pp. 191-197) IEEE. December, 2013
- [96] Hussain, A., “A clean novel watermarking technique using a chaotic logistic map and multiple embedding”, In: *The 10th International Conference on Graphics and Image Processing*, Vol. 11069, pp. 1106920. 2019
- [97] Moysis, L., Volos, C., Jafari, S., Munoz-Pacheco, J. M., Kengne, J., Rajagopal, K. & Stouboulos, I., “Modification of the logistic map using fuzzy numbers with application to pseudorandom number generation and image encryption”, *Entropy*, 22(4), 474. 2020
- [98] Alipour, M., Gerardo, B. & Medina, R., “A secure image encryption architecture based on pseudorandom number generator and chaotic logistic map”, In: *Proceedings of the 2nd International Conference on Data Science and Information Technology*, pp. 154–159. 2019