electronics

MDPI

Article

Healthcare Big Data Analysis with Artificial Neural Network for Cardiac Disease Prediction

Sulagna Mohapatra [1], Prasan Kumar Sahoo [1,2,*] and Suvendu Kumar Mohapatra [1]

[1] Department of Computer Science and Information Engineering, Chang Gung University, Guishan, Taoyuan 33302, Taiwan; d0521007@cgu.edu.tw (S.M.); d0121007@cgu.edu.tw (S.K.M.)
[2] Department of Neurology, Chang Gung Memorial Hospital, Linkou Medical Center, Guishan, Taoyuan 333423, Taiwan
* Correspondence: pksahoo@mail.cgu.edu.tw; Tel.: +886-3-211-8800 (ext. 3804)

Abstract: The generation of a huge volume of structured, semi-structured and unstructured real-time health monitoring data and its storage in the form of electronic health records (EHRs) need to be processed and analyzed intelligently to provide timely healthcare. A big data analytic platform is an alternative to the traditional warehouse paradigms for the processing, analysis and storage of the tremendous volume of healthcare data. However, the manual analysis of these voluminous, multi-variate patients data is tedious and error-prone. Therefore, an intelligent solution method is highly essential to perform multiple correlation analyses for disease diagnosis and prediction. In this paper, first, a structural framework is proposed to process the huge volume of cardiological big data generated from the hospital and patients. Then, an intelligent analytical model for the cardiological big data analysis is proposed by combining the concept of artificial neural network (ANN) and particle swarm optimization (PSO) to predict the abnormalities in the cardiac health of a person. In the proposed cardiac disease prediction model, an extensive electrocardiogram (ECG) data analysis method is developed to identify the probable normal and abnormal cardiac feature points. Simulation results show the effects of a number of attributes for improving the accuracy of the cardiac disease prediction and data processing time in the cloud with an increase in the number of the cardiac patients.

Keywords: big data; cardiac disease; electrocardiogram (ECG); artificial neural network (ANN)

Citation: Mohapatra, S.; Sahoo, P.K.; Mohapatra, S.K. Healthcare Big Data Analysis with Artificial Neural Network for Cardiac Disease Prediction. *Electronics* **2024**, *13*, 163. https://doi.org/10.3390/electronics13010163

Academic Editor: Tun-Wen Pai

Received: 1 December 2023
Revised: 24 December 2023
Accepted: 28 December 2023
Published: 29 December 2023

1. Introduction

Cardiovascular diseases are very common now due to the changes in lifestyle and food habits [1]. Mostly, an electrocardiogram (ECG) is used to measure the cardiac activity of the heart in the form of a signal that can be beneficial in the care of the chronic heart patients [2,3]. The ECG signal produces a large number of unstructured data sets, which are difficult to process in the traditional approaches. In addition, in-hospital patients generate varieties of clinical data at a tremendous speed, which needs to be processed and stored for further analysis. Based on a survey, approximately 50 petabytes of digital healthcare data are estimated to be generated in 2012, and the trend continues exponentially to reach 2500 petabytes in the future [4]. Hence, a big data analytic framework is the best solution to analyze the gigantic structured, semi-structured and unstructured data in an efficient manner at a single point of time [5]. However, when huge numbers of patients are involved with different health parameters with symptoms of abnormal cardiac health, it becomes a tedious job for the physicians to identify the heart disease. In addition, the complex correlation analysis considering multi-variable health parameters cannot be performed manually [6].

To determine the cardiac complications accurately, it is required to process and analyze large number of co-related parameters together. The statistical analysis of such a huge number of parameters is a highly challenging, time-consuming and tedious task [7]. Moreover,

existing statistical analysis tools are very expensive and offer only a handful number of analysis methods. In the absence of state-of-the-art analysis methods, the in-depth understanding of responsible causes behind cardiac complications remains a challenging issue [8]. To gain in-depth knowledge of available data, intelligent algorithms can be employed as an alternative tool on a case-by-case basis depending on the requirement and scenario on hand. For instance, answer set programming (ASP) is considered one of the practical tools for high-utility data extraction from a large-scale dataset to improve the accuracy and efficiency of big data management [9]. Currently, the artificial intelligence (AI) technique is utilized for healthcare data analysis especially for the cardiogenic issues [10]. The machine mimics the way a human observes, interprets, evaluates, and makes decisions based on the trained data [11]. Powered with multiple supervised and unsupervised learnable algorithms, AI replaces traditional rule-based strategies with data-driven approaches and is capable of learning from the positive and negative experiences. Those inferred clinical diagnoses assist clinicians in faster decision making with a higher reduction in their workload. Considering the importance of AI, a logical framework for studying the evolution of neurological disorders is developed by integrating the concept of artificial neural networks (ANNs) and ASP [12]. The authors have extensively researched the brain's structural connectivity by representing it as a graph network. The proposed protocol achieved impressive results with precision, recall and F-score of 0.89, 0.88 and 0.88, respectively.

The e-health care system can be strengthened using both cloud and big data technology, where the entire healthcare system can be digitized, and any information of a patient can be accessed by the authorized person from anywhere at any time [13]. Considering the platform as a service (PasS) and infrastructure as a service (IaaS), cloud services are used for big data, where the IasS deals with on-demand virtual machines and virtual resources from a large storage pool present in the data centers. PaaS provides a computing platform such as an operating system, database, web service and parallel program execution [14]. Virtualization is one of the key technologies that can act as a backbone of various big data analysis tools such as Hadoop, where chunks of data are processed parallelly in multiple servers [15].

The maintenance of a healthcare system is expensive and difficult, since a bulk amount of emergent data are generated over time [16]. Further, these healthcare data need extensive analysis to make real-time decisions based on the extracted information in the form of knowledge. Therefore, we intend to propose statistical big data analytic models with a machine learning approach for the critical cardiological data analysis in the healthcare environment. Furthermore, an optimization technique is introduced to minimize the storage, transfer, and processing cost in the cloud. In this paper, a dynamic big data analytic framework is designed to handle the high volumes of critical patient data and extend it to analyze the ECG batch data as an application of big data processing. In addition, an intelligent heart disease prediction model is proposed using the approach of an artificial neural network (ANN), which has the potential to establish an implicit relationship between the complex nonlinear patient's health parameters. In addition, the particle swarm optimization (PSO) technique is employed to reduce the computational burden in ANN in terms of accuracy, speed of convergence, and global optimization. Eventually, the resulted data can be stored in the cloud for future analysis by doctors, patients, nurses and researchers.

The rest of the paper is organized as follows. The related works with our contributions are presented in Section 2. A cardiological big data processing and analysis architecture is proposed in Section 3. The delineation of abnormal ECG feature points with heart diseases prediction using an artificial neural network is described in Section 4. Simulation results are given in Section 5, and concluding remarks are made in Section 6.

2. Related Work

Recently, there is a rapid growth of data-intensive applications such as digitized medical records, scientific data analysis, semantic web analysis, sensor data and bio-

informatics data analysis. Hence, big data have drawn attention from industry, academia, scientists and governments. According to [17], the healthcare data analytics are categorized into three different parts, such as descriptive, predictive and prescriptive analytics. The descriptive analytic is described as the report summary of the data sets that are under investigation. It may be used to address the questions like "What is the problem and what precautions are needed?" However, the descriptive analytic is unable to predict the future health condition, which is overcome by the predictive analytic. In the predictive analytic, various statistical models can be utilized to know the future health conditions on the historical data sets, which can answer questions like "What could be the health conditions after one day?". Similarly, the prescriptive analytics can answer the question such as "What is the best scenario?", which is usually used in optimization problems. Hence, our aim is to answer such questions correctly.

A secure medical distributed big data ecosystem is designed using the Hadoop platform in [18], where the personalized healthcare data are stored centrally but analyzed in a distributed fashion through the developed data synchronization module. This Hadoop-based health management service enables the hospital staff to manage patient data efficiently. The authors in [19] have designed a stochastic model to predict future health conditions by correlating multiple health parameters with their current health conditions. The proposed model achieved a prediction accuracy of 98% with a 90% reduction in CPU and bandwidth utilization. A new cloud-based data storage framework is proposed by the authors in [20] for both structured and unstructured data for the heterogeneous IoT devices. The framework is extended and integrated with the Hadoop storage system to handle the diversified collected data types. However, no analysis is made for the healthcare data sets. In [21], the authors have tried to optimize the cost of data migration by choosing the effective data centers for data aggregation and processing, taking routing paths into account.

Although ECG is widely adopted for the cardiac health analysis, the signal pattern is complicated, and manual delineation of the primary feature points—namely P, Q, R, S and T—is really difficult especially when the abnormality is minute [22]. Therefore, different mathematical and AI models are proposed for the automatic identification of those feature points. For instance, the authors in [23] achieved a sensitivity of 99.82% in detecting arrhythmia using Q, R and S feature points. However, other points like P and T waves have greater influence in a generic cardiac health analysis scenario. Similarly, the abnormality in R and Q, R, and S points are identified by others in [24] with sensitivity = 99.8%. Nonetheless, the contamination of Q and S points due to noise is not considered. The classification of QRS-wave morphology is explained in [25,26] considering Q, R and S as the dominant feature points. However, the complexity due to P and T waves are ignored in both of the works.

The concept of CNN is employed to detect cardiovascular disease considering the ECG in [27]. Although the work achieved a detection accuracy with true positive = 99.8%, the precision is very low due to high false negatives. A CNN-based classification model using ECG is developed by the authors [28] to classify cardiac abnormalities in different categories such as normal, atrial premature beat, and premature ventricular contraction. To calculate the risk of atherosclerotic cardiovascular disease (ASCVD), a machine learning (ML) method is proposed in [29] that considers the electronic medical records (EMRs) in the analysis. The proposed work outperforms the conventional pooled cohort equations (PCEs) risk calculator. The authors in [30] have considered 74 independent cardiac-related features such as blood pressure, heart rate, ST depression, etc., to predict the heart disease using different ML models such as decision tree (DT), gradient-boosted tree (GBT), logistic regression (LOG), multilayer perceptron (MPC), naïve Bayes (NB) and random forest (RF). Out of all, the RF performs better with 98.7% accuracy. The comparison of our proposed architecture with the existing works [2,16,19,22,31–41] are summarized in Table 1. In addition, the summary of previous works [40–47] related to intelligent cardiac data analysis is presented in Table 2.

Table 1. Comparison of the proposed work with the existing works.

Related Works	Big Data	Map Reduce	Cloud	Cardiac Healthcare Data	ECG Data
[16]	✓	×	✓	✓	×
[19]	✓	✓	✓	✓	×
[2,22]	×	×	×	×	✓
[31]	✓	×	✓	✓	×
[32,33]	✓	✓	×	×	×
[34]	×	✓	✓	×	×
[35,36]	✓	×	×	✓	×
[37,38]	×	×	✓	✓	×
[39–41]	×	×	×	✓	×
Ours	✓	✓	✓	✓	✓

Table 2. Existing intelligent models used for cardiac data analysis.

Related Works	Dataset	Algorithm Type	Analysis	# of Features	Accuracy
[40]	UCI repository	Learning vector quantization (LVQ)	Classification	10	0.98
[41]	Cleveland heart disease	MMC, Random, Adaptive, QUIRE, and AUDI	Classification	14	0.57
[42]	Cleveland heart disease	Multilayer perceptron (MLP) + PSO	Classification	13	0.84
[43]	Cleveland heart disease	Recurrent neural network (RNN) + Long short term memory (LSTM)	Classification	14	0.95
[44]	Cardiovascular disease (CVD) and Framingham	MLP, Support vector classifier (SVC)	Classification	12 (CVD), 11 (Fram)	0.74 (CVD), 0.71 (Fram)
[45]	Cleveland heart disease	SVM + AdaBoost	Classification	14	0.88
[46]	Cleveland, Hungarian, Switzerland, Long Beach VA, Statlog Data Set	Classification and regression tree (CART)	Classification and Regression	11	0.87
[47]	Department of Cardiology, IGMC	Multinomial logistic regression (MLR)	Classification	26	0.98

2.1. Limitations of Existing Works

Most of the related works on cardiac abnormality detection as discussed above consider only a single modality of data either in the form of physical records, radiological images, or cardiac signal-specific data. There are limited works that consider the analysis of multi-modality data such as the integration of physiological and demographic data with the ECG, which is one of our biggest motivations. In addition, there are a lack of big data analysis strategies for the effective management of both structured and unstructured cardiac healthcare data. In the healthcare industry, the handling and storing of the critical data of the patient is highly essential and is expensive, as the bulk amount of emergency data is generated over time. However, none of the related works on cardiac data analysis have studied the necessity of a distributive cloud-based paradigm for the processing and storage. Although many related works have proposed intelligent models for the cardiac data analysis, the accuracy is not impressive [41,42,44] and has technical limitations [43]. For instance, the proposed work for cardiac abnormality determination has achieved 95% accuracy after applying LSTM. However, LSTM possesses high computational complexity, a longer training time, and a complex procedure of hyperparameter tuning.

2.2. Contributions

Considering the limitations of the previous related works, the main contributions of our work can be summarized as follows.

- An advanced functional cloud-based big data analytic architecture is proposed to process the massive volume of manifold medical data.
- A multi-modal analysis is performed considering patient's demographic, physiological and ECG data as feature attributes.
- ECG signal analysis is performed considering important feature points to identify the normal and abnormal heart condition.
- An intelligent prognosis model is developed by integrating the concept of ANN and PSO to classify the heart disease considering multi-parametric data collected from symptoms, reports and ECG records.
- The proposed intelligent system has achieved a maximum 99% accuracy in determining the cardiac abnormality in critical cases.
- Extensive simulation is performed to validate the proposed cloud-based big data analytic framework and heart disease prediction algorithm.

3. Framework for Cardiological Data Processing and Analysis

In and around the hospitals, a massive volume of data are generated from various body sensors, test records, treatment procedures, and personal information of patients, which leads to the complex storage and analysis issues. For example, the data may come with a greater speed and large volume from various sources of the hospital, such as sensors, instruments, records, pharmacies, research laboratories and reports of the patients in real time, as shown in Figure 1, which is very difficult to manage using traditional methods. As shown in the figure, all users such as doctors, patients, nurses and researchers of different hospitals and laboratories are connected with the cloud to store their respective real-time data. At a single point of time, a huge amount of data are generated and accessed by the users, which motivates us to use the cloud-enabled big data platform to reduce the delay and minimize the average cost of data analysis. In this section, the detailed architecture is presented for the healthcare domain with the layered architecture as explained below.

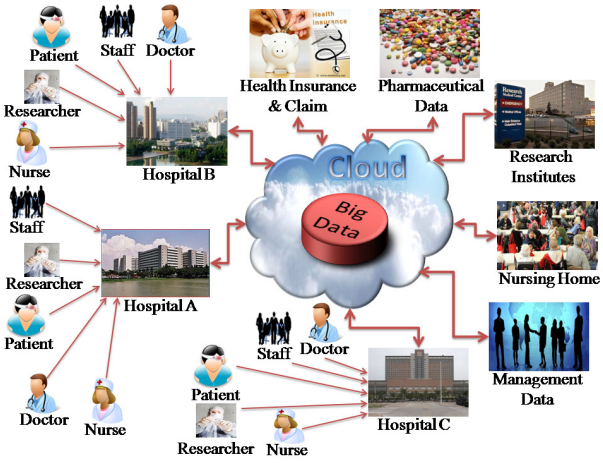

Figure 1. Sources of big data generation in medical environment.

3.1. Structural Architecture

In this section, a cloud-integrated AI-powered big data architecture is introduced for the hospital and healthcare management. A structural view of the big data analytic architecture is presented as shown in Figure 2. In the healthcare application scenario, big data and cloud with machine intelligence are combined together to obtain better results in terms of storage and performance. User layer 1st is the one in which users could be the doctors, patients, nurses, management staffs and researchers. In this layer, the patient's data are generated and the extracted knowledge is used for the purpose of treatment and decision making.

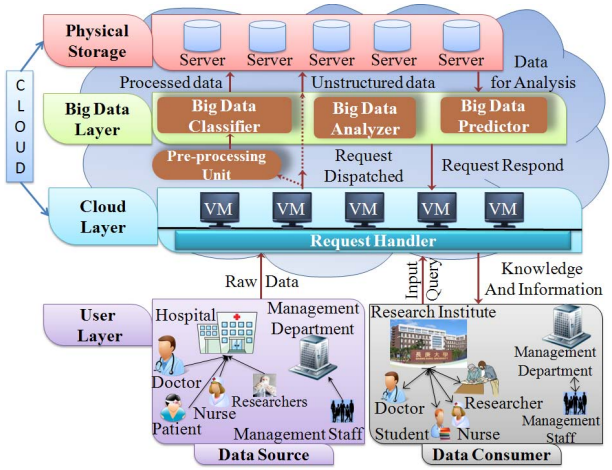

Figure 2. Structural view of proposed big data architecture.

The 2nd layer is known as the Cloud Layer, where a large number of users requests are handled with the help of multiple virtual machines (VMs). Multiple CPU cores are assigned to each VM for CPU-intensive request processing, and a high network bandwidth is allocated for agile data transfer. This layer is also responsible for balancing the workload (CPU and I/O-intensive workload) by dividing the available resources among the jobs. In the proposed architecture, the Big Data Layer is the 3rd layer, where the Big Data Classifier, Big Data Analyzer and Big Data Predictor are present as components for the data classification, processing and prediction, respectively. The AI-based analysis of those high voluminous, multi-variate cardiogenic data is performed in the Big Data Layer. The Cloud Layer is extended to the Physical Storage Layer (4th layer) known as data centers to store the medical data for future access and processing purposes. Finally, the analyzed data can be visualized by the doctors, patients and researchers for future use. The proposed generic

framework for the learnable cardiological big data analysis can be applied effectively in collective healthcare applications.

3.2. Logical Architecture

A logical big data architecture for intelligent analysis is proposed for healthcare applications with the support of a cloud computing framework, as shown in Figure 3. The model is broadly classified into three different layers, i.e., the User Layer, Cloud Layer and Big Data Layer. Each layer not only addresses the data flow but also exhibits an efficient way to handle, store and analyze the healthcare big data.

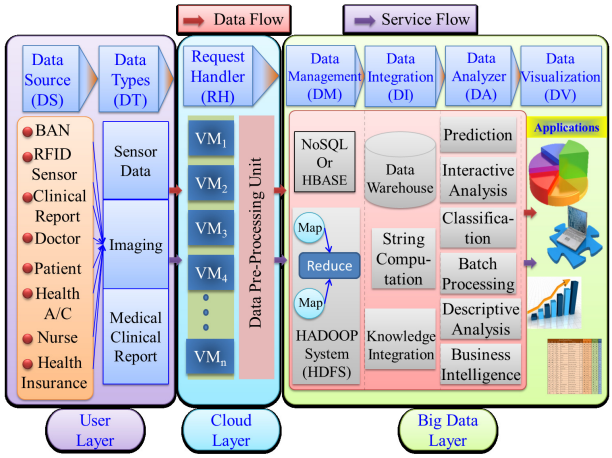

Figure 3. Logical architecture of healthcare big data analytic platform.

3.2.1. User Layer

The user layer is the first layer of the proposed model, which is basically used for the healthcare data input. In this layer, Data Sources (DSs) and Data Types (DTs) are described to explain the big data generation in the healthcare domain. Doctors, patients, nurses and different hospitals are major sources of big data generation. More specifically, cardiac patients' ECG and clinical data are considered as the input data sets. As the ECG data are generated in a continuous manner over time, it becomes huge in volume with clinical records to manage and analyze properly. The collected patients' data are broadly categorized as structured, semi-structured and unstructured data types. Structured data are defined as the data present in any specific format such as text, number or character. Each patient's name, age, gender, etc. can be considered as the structured data, which can be stored in a relational table format. Semi-structured data include the patient's EHRs, physicians prescriptions, email, etc. Unstructured data are generated by various sources such as ECG data collected from the devices, radiology and computerized tomography scan (CT Scan) data, which are difficult to analyze and process in the traditional analytic platform. In addition, integrating those data for analysis and management is challenging. For instance, the analysis and management of clinical data differ from the image analysis. Furthermore, there is a need for parallel processing in a distributed fashion in some cases. Accordingly, the successive cloud, big data, and storage layers fulfill the requirements for efficient processing, management, analysis and storage of the heterogeneous multi-modal data.

Data Sources (DS)

Patients related structured, semi-structured and unstructured data are accumulated from various sources such as the body area network (BAN), pathological test reports, and radiological images, which are input to the big data healthcare system.

Data Types (DTs)

Data types are defined as different healthcare data in various formats such as ECG data collected from the ECG sensors, radiological and computerized tomography scan (CT Scan) data in the form of images, and blood pressure data in the form of text.

3.2.2. Cloud Layer

The Cloud Layer is solely used for handling and storing the data across the data centers in a distributed fashion. The query requests of the doctors and nurses are also handled by the request handlers present in this layer. In order to analyze and detect the abnormality of a cardiac patient, huge amount of cardiogenic data are collected from time to time in a hospital. For example, the ECG data of the patients are generated and stored in a fault-tolerant, highly available cloud systems where the data can be accessed anywhere at anytime. It can be achieved only if the data are stored in a highly virtualized distributed cloud so that the scalability, elasticity, fault tolerance, self-manageability, and ability to run on commodity hardware can be achieved. Basically, VMs take care of the requests through a request handler. A preprocessing unit is employed to refine the raw data into a standard format by which data inconsistency and duplicity can be eliminated.

Request Handler (RH)

The Request Handler (RH) is used for interaction and handling the data and computation-intensive requests coming from the user layer. The proposed request handler has four major components: job filter, scheduler or job dispatcher, VM repository and VM auditor. The job filter component is placed at the top of this framework, which is responsible for filtrating the incoming processes. Types of the requests are analyzed efficiently by the job filter and are redirected to their respective queues. For instance, an analysis of heart disease training on the last five years of data is an example of data-intensive job queue. In contrast, medical analysis like radiological images analysis and genome sequencing are the examples of computation-intensive job queues. A dedicated scheduler is designed to utilize the resources optimally by redirecting the large number of small jobs to compute the intensive queue. The VM repository is also known as a reconfigurable VM repository, which acts as the bridge between the server and the user. The VM repository is reconfigured as per the requirements of the computation-intensive and data-intensive jobs. The configuration is varied between the VMs in terms of memory, processing capabilities, network bandwidth and storage. The VM auditor is mainly used for configuring and allocating the VMs as per the user requirements. However, if any shortage of VMs occurs, the VM auditor leases and reconfigures the required VMs. Once the task is performed, all VMs are released and are recomposed by the VM auditor.

The VMs are provisioned to the physical servers present in the data centers for the proper utilization of the resources. The configurations of the VMs are different in terms of memory, processing cores, network bandwidth and storage. After the completion of any task, the allocated VMs are released for future use. The cloud platform for the big data analysis can act as an off-premise computing environment to store the data on the Internet and can be available for analysis irrespective of any locality. Hence, the Cloud Layer is more essential to include in our proposed healthcare data analytic platform for patient data storage and processing by incorporating with the Big Data Layer.

3.2.3. Big Data Layer

The Big Data Layer comprises four sublayers, i.e., Data Management (DM), Data Integration (DI), Data Analyzer (DA) and Data Visualization (DV).

Data Management (DM)

In the Data Management (DM) layer, the incoming patient data are processed into information and are managed through different data-intensive software frameworks like Hadoop MapReduce. A MapReduce programming model is used for the processing of

large data sets in parallel and distributed fashion. Basically, the map function is used to match, filter and sort the processing tasks, whereas the Reduce function is utilized to aggregate and summarize the resultant operations.

Data Integration (DI)

In this phase, the processed data that come from the DM phase are integrated with the existing patient warehouse. In a healthcare environment, the newly generated patient data are consolidated with the existing data to maintain the continuity of the information.

Data Analyzer (DA)

In this phase, different analysis such as prediction, interactive analysis, descriptive analysis, data mining and business intelligence are carried out on the stored data. However, the exact investigation depends on the user's requirement. In the healthcare domain, possible future complications and failure are anticipated using a prediction mechanism for which preventive action can be taken. Therefore, for accurate prediction of the cardiac abnormality within a shorter span of time, a fusion of intelligent ANN and PSO algorithms is incorporated in this analyzer.

Data Visualization (DV)

In Data Visualization (DV), the resultant data are represented in various formats with different granularities. For example, the stored data are presented in graphical formats for the doctors, patients and researchers to take necessary action. In the healthcare system, the output of the data can be shown in different formats such as text, image, audio, video, graphs, charts, and table.

3.2.4. Physical Storage Layer

In the proposed architecture, the Physical Storage layer is exclusively used for storing the analyzed data across different Data Centers (DCs) in a distributed fashion. The Hadoop File System (HDFS) is used by the Hadoop platform to store the patients' data in a distributed fashion. All DCs are networked and distributed geographically. In this layer, a zookeeper is used as a coordinator between Hadoop Executors and the HBase repository for the data storage. HBase distributed databases are used in the cloud which can support the fault tolerance by clustering multiple database and backup nodes, which are also synchronized by the zookeeper. By using this kind of state-of-art architecture, a colossal amount of patient cardiac batch data can efficiently be handled, analyzed and stored for the healthcare automation.

4. Intelligent Model for Cardiac Disease Prediction

The proposed analytical paradigm for coronary heart disease prediction is explained in three substeps. First, a generic hospital environment is elucidated considering how a variety of healthcare big data are generated and how a common relationship can be established among the patients by correlating multiple health parameters. The second step includes an extensive analysis related to the identification of the ECG feature point's abnormalities. In the final phase, the intelligent heart disease prediction algorithm is derived considering 15 attributes related to the cardiac healthcare, as presented in Table 3.

Table 3. Attributes description and normalized values.

Attributes	Descriptions	Original Values	Normalized Values
Age	Age in year	Continuous	0: age $\leq$ 30 1: 30 $\geq$ age $<$ 50 2: 50 $\geq$ age $<$ 70 3: age $\geq$ 70
Sex	Male or Female	(0, 1)	0: Female 1: Male

Table 3. *Cont.*

Attributes	Descriptions	Original Values	Normalized Values
cp	Chest pain	1: Typical angina 2: Atypical pain 3: Non-anginal pain 4: Asymptomatic	1: Typical angina 2: Atypical pain 3: Non-anginal pain 4: Asymptomatic
trestbps	Resting blood pressure	Continuous	Normalization using min–max method
chol	Serum cholesterol in mg/dL	Continuous	0: chol $\leq$ 200 1: 200 $>$ chol $\leq$ 239 2: chol $\geq$ 239
fbs	Fasting blood sugar	Continuous	0: fbs $\leq$ 120 1: fbs $>$ 120
restecg	Resting electro-cardiographic	(0, 1, 2)	(0, 1, 2)
thalach	Maximum heart rate	Continuous	Normalization using min–max method
exang	Exercise -induced angina	Yes or No	0:No 1:Yes
oldpeak	ST depression	(0–4)	(0–4)
slope	Slope of peak exercise ST segment	1: Upsloping 2: Flat 3: Downsloping	1: Upsloping 2: Flat 3: Downsloping
ECG abnormality	Abnormality in any feature points	Yes or No	0:No 1:Yes
ca	Number of major vessels (0–3) colored by fluoroscopy	(0–3)	(0–3)
thal	Normal, Fixed defect, Reversible defect	3: Normal 6: Fixed defect 7: Reversible 7: defect	3: Normal 6: Fixed defect 7: Reversible 7: defect

4.1. Environmental Scenario of Cardiac Healthcare Big Data

In heart disease prediction, the future health condition of a patient is anticipated whether the patient is having any heart disease or not. Based on the proposed environment as shown in Figure 1, patients visit the hospital over time. Let m be the number of patients that visit the hospital at time t. In a hospital, various departments are present, and each department can have several doctors associated with different patients, as shown in Figure 4. The time frame could be represented as hour, day, week or month. However, the general form of time is represented as t_1, $t_2 = t_1 + \Delta t, t_3 = t_2 + \Delta t, \ldots, t_n$ where Δt is a constant time duration. During time t_1 through t_n, all the incoming patients can be represented as $(P_1, t_1), (P_2, t_2), \ldots, (P_m, t_n)$. Similarly, different departments can be represented as $(Dpt_1, Dpt_2, Dpt_3, \ldots, Dpt_k)$, where k is the number of departments present in a hospital. In each department, doctors can be represented as {(Doc_{11}, Doc_{12}, Doc_{13},

$\ldots, Doc_{1z}$), (Doc_{21}, Doc_{22}, Doc_{23}, $\ldots, Doc_{2z}$),$\ldots$,(Doc_{k1}, Doc_{k2}, Doc_{k3}, $\ldots, Doc_{kz}$)}, where z is the total number of doctors present in each department. Out of a total P_m number of patients, each department can have a certain number of patients. For example, Dpt_1 represents having $(P_1, P_2, \ldots, P_a)$, Dpt_2 represents having $(P_{a+1}, P_{a+2}, \ldots, P_b)$, and Dpt_k represents having $(P_{c+1}, P_{c+2}, \ldots, P_m)$ patients, which is expressed in Equation (1).

$$P = \sum_{i=1}^{m} P_i = \sum_{i=1}^{a} P_i + \sum_{i=a+1}^{b} P_i + ... + \sum_{i=c+1}^{m} P_i \tag{1}$$

A patient's heart condition is connected with various health-related parameters such as age, body mass index, sex, type of chest pain, blood pressure, serum cholesterol, fasting blood sugar, electrocardiographic (ECG) data, maximum heart rate, etc. Let $(V_1, V_2, V_3, \ldots, V_l)$ be the maximum health-related parameters correlated with each patient. It is possible that one patient may be associated with all or some of those parameters. Hence, Equation (1) can be rewritten as,

$$P = \sum_{i=1}^{m} P_i \cdot \sum_{i=1}^{l} V_i \tag{2}$$

$$\Rightarrow P = \sum_{i=1}^{a} P_i \cdot \sum_{i=1}^{l} V_i + \sum_{i=a+1}^{b} P_i \cdot \sum_{i=1}^{l} V_i + ... + \sum_{i=c+1}^{m} P_i \cdot \sum_{i=1}^{l} V_i$$

Again, the healthcare parameters are categorized into two different groups named as basic and critical parameters.

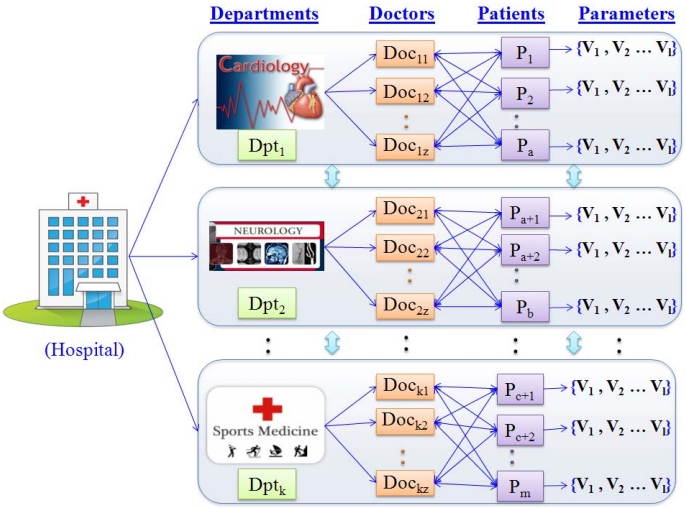

Figure 4. Patient information across various departments of a hospital.

Definition 1. *(Basic parameters) B_i: Basic parameters are defined as the general health variables such as age, sex, body weight, body mass index, etc., which represent having a less importance level of a disease.*

Definition 2. *(Critical parameters) C_i: Critical parameters are expressed as the most substantial health variables like chest pain, blood pressure, cholesterol, ECG data, maximum heart rate, etc. which are having a higher importance level of a disease.*

The basic and critical parameters are expressed in Equation (3) as follows:

$$V_i = \sum_{i=1}^{x} B_i + \sum_{i=x+1}^{y} C_i \tag{3}$$

Equation (2) can be revised as

$$P = \sum_{i=1}^{m} P_i \cdot (\sum_{i=1}^{x} B_i + \sum_{i=x+1}^{y} C_i) \tag{4}$$

where $\sum_{i=1}^{x} B_i$ are the basic parameters and $\sum_{i=x+1}^{y} C_i$ are the critical parameters.

$$\Rightarrow P = (\sum_{i=1}^{m} P_i \cdot \sum_{i=1}^{x} B_i) + (\sum_{i=1}^{m} P_i \cdot \sum_{i=x+1}^{y} C_i) \quad (5)$$

In a healthcare process, let the i^{th} doctor (Doc_i) treat the i^{th} heart patient (P_i), who has different basic and critical parameters. Similarly, the j^{th} patient (P_j) is treated by the j^{th} doctor (Doc_j). However, there must be some common parameters between the patients P_i and P_j. The intersection between the basic parameters set according to Doc_i and Doc_j is $S^b_{Doci \cap Docj}$. For example, if $patient_1$ has a basic parameter set {*age, sex, body weight*} and $patient_2$ has a basic parameter set {*age, body weight, body mass index*} $S^b_{Doc1 \cap Doc2}$ = {*age, body weight*}. The intersection between critical parameters is set according to Doc_i and Doc_j is $S^c_{Doci \cap Docj}$. For example, if $patient_1$ has the critical parameter set {*chest pain type, blood pressure, serum cholesterol, maximum heart rate, smoke*} and $patient_2$ has the critical parameter set {*blood pressure, serum cholesterol, fasting blood sugar, ECG data, maximum heart rate, smoke*}, $S^c_{Doc1 \cap Doc2}$ = {*serum cholesterol, maximum heart rate, smoke*}. Similarly, all possible intersections are calculated for all the patients prescribed by the doctors. The future heart condition is predicted by combining both basic and critical parameters in common as expressed in Equation (6).

$$S^{bc} = S^b_{Doci \cap Docj} \cup S^c_{Doci \cap Docj} \quad (6)$$

4.2. Delineation of ECG Feature Points

In this analysis, ECG data are considered one of the primary inputs for cardiac healthcare prediction. ECG is a time-series record of heart activity to observe the abnormality. In general, normal heart functioning is determined by tracking the orderly progression of the five important feature points P, Q, R, S and T, as shown in Figure 5. The ECG signal consists of many sequential waveforms, which is repeated periodically and described as a P wave (atrial depolarization), QRS complex (ventricular depolarization) and T wave (ventricular repolarization), as shown in Figure 5. Commonly, each wave has a predefined duration. For instance, in case of a person with a healthy heart, the duration of the P wave should be between 0.08 and 0.10 s. Any deviation from that default value could be considered as a sign of a heart problem. The accurate cardiac abnormality from ECG is diagnosed by observing the successive changes in the duration of those feature points for the respective cardiac cycle. However, the correct identification of those primary points is challenging as there is a high chance of signal distortion due to artifacts and variable position of the points in the ECG plot. In this proposed analytic architecture, Hadoop is used as the analysis platform, which is most suitable for the patient batch data. According to the Hadoop platform, the clinical data can be analyzed in parallel by using various Map (*wMap*, *sMap* and *dMap*) and Reduce (*rReduce* and *cReduce*) functions in different stages, as shown in Figure 6.

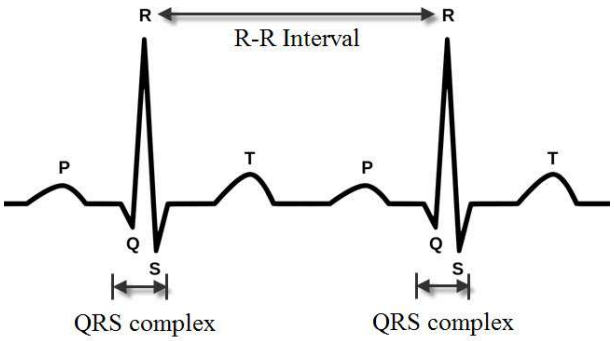

Figure 5. ECG signal with R-R interval and QRS complex.

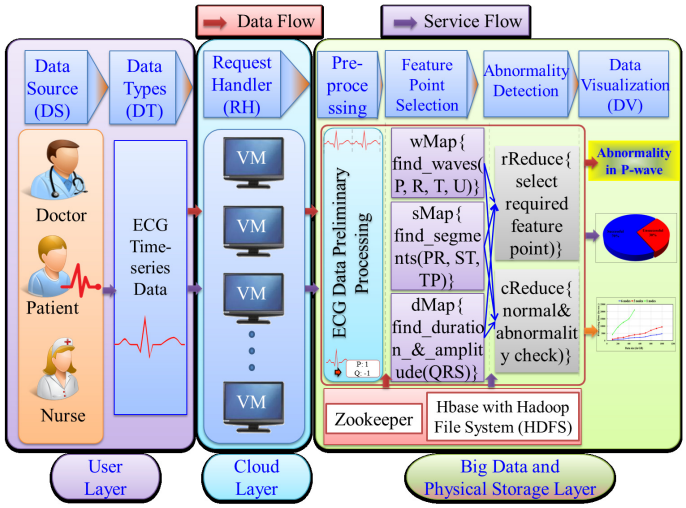

Figure 6. Conceptual big data analytic architecture for cardiac healthcare.

The step-by-step procedure of analysis is described in Algorithm 1. Lines 1–7 present the preprocessing of an unstructured ECG signal. In the data preprocessing phase, the noisy ECG signal is cleaned to extract the imperative features. Nevertheless, various methods are carried out for preprocessing including the noise removal. ECG data deal with the electric waves, and therefore wavelet drift correction and frequency filtering will be performed as preprocessing. Afterward, smoothing and signal enhancement is accomplished for a clear representation of the signals. In the preprocessing stage, the ECG data waves are divided into separate segments by considering the end of the *R-wave* in each *PQRST* cycle and stored locally in *data_ecg*. Essential feature information is the key factor for the identification of abnormalities. Thus, the feature points of ECG data are selected based on the duration and amplitudes of *P*, *R*, and *T* waves (Line 12–25) after the preprocessing phase, which are explained in *wMap* function of Hadoop.

In order to obtain better feature points, various segments such as *RR*, *ST* and *TP* are considered as the *sMap* function. The *QRS* duration is included in the ECG assessment in the *dMap* phase of the feature points selection. Here, our first objective is to find *R, P and T waves*, which are very crucial factors for feature point selection. Sometimes, the P wave is destroyed due to noise in case of low amplitude. In that situation, we simply ignored the noisy signal in the analysis. By including both P and T waves, the analysis is more realistic and provides better feature points. After finding different waveforms, we are highly interested in finding the *QRS* complex, as it is the most distinctive one in the *PQRST* cycle. The QRS duration is included in the ECG assessment (Lines 26–29). Upon the completion of all *Map* functions, the intermediate results are transferred to the *Reduce* phases for final execution. In the *Reduce* phase, two logical reduction functions are defined, where *Reduce* is used to decrease the number of feature points by ignoring the unwanted data sets. Likewise, the *cReduce* function is used to check the abnormality in the processed ECG data sets. Consequently, the cardiac patients are diagnosed with respect to their ECG data values. According to Line 30, if any of the feature points is deviated from the normal range of cardiac functioning, then the corresponding ECG cycle is identified as abnormal. The classified ECG signal from Algorithm 1 is used as one of the primary attributes for heart disease prediction.

Algorithm 1 ECG feature point analysis in Hadoop cluster

Input: data_ecg: ECG data sets.
Output: R, P, T, QRS interval.

```
Preprocessing(){
String[] tokens = value.toString().split("\t");
for i = 10 to tokens.length do
    time[i/2] = tokens[i];
    data_ecg[i/2] = Float.parseFloat(tokens[i + 1]);
end for
}
Select the R-waves;
for i = 1 to range − 1 do
    data_dif[i] = data_ecg [i + 1] − data_ecg [i];
end for
Select the range for R-Wave;
for j = 1 to range − 1 do
    if data_dif[j] < −0.5 && data_dif[j] > 0.5 then
        Return R value;
    end if
end for
Select the P-waves;
if 0.08 < data_ecg < 0.10 then
    Return P value;
end if
Select the T-waves;
if 0.10 < data_ecg < 0.25 then
    Return T value;
end if
Select the QRS-waves;
if 0.06 < data_ecg < 0.10 then
    Return QRS value;
end if
Otherwise, the patient has abnormal value;
}
```

4.3. Cardiac Disease Prediction Using ANN and PSO

In this section, the formation of artificial neural networks (ANNs) [48] is narrated with back-propagation supervised learning using historical patient data. However, the decision to include or exclude health parameters is analyzed for each individual patient. Particle swarm optimization (PSO) [49] is used to optimize the synaptic weights used in ANNs for better results, which is also computationally inexpensive with respect to the processing and speed of healthcare big data.

4.3.1. Artificial Neural Network

An artificial neural network is incorporated with a distributed network, connection strengths and processing units. Knowledge is acquired from the environment by the network through a learning process, which is associated with synaptic weights of inter-neurons' connection strengths [48]. A neural network is trained through the training medical data sets, i.e., information are stored by associating it with other information in the memory (neurons). The required information is invoked by the neural network based on partially incorrect inputs, which is more suitable for the unstructured medical data.

An artificial neural network is a computational model, which is combined with three working layers as shown in Figure 7. The first layer is called the input layer through which l number of patient data can be pushed into the network denoted as $(V_1, V_2, V_3, \ldots, V_l)$. The next layer is known as the hidden layer, where the inputs are taken from the input

layer and are passed to the next output layer via a Sigmoid activation function (Φ), which is given in Equation (7).

$$\Phi = \sum_{i=1}^{l} V_l \cdot W_l + \Delta t_h \tag{7}$$

where W_l represents the synaptic weights associated with distinct neurons present in different layers, which can be adjusted dynamically based on the behavior of the patient's input data with respect to the required output. In between the neurons, the weights are associated and represented as $(W_1, W_2, W_3, \ldots, W_l)$, respectively. The threshold in artificial neurons is usually represented by Δt_h.

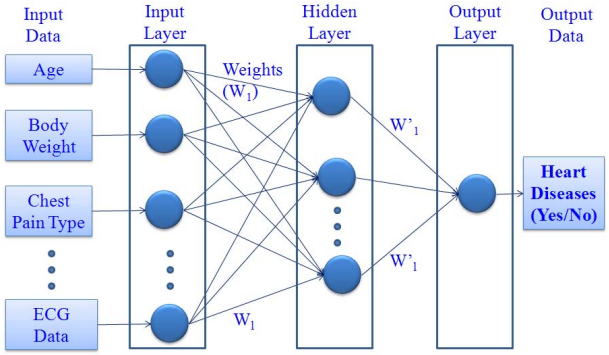

Figure 7. ANN working layers.

The output layer is the third layer, where the output results come out from the network. In our case, the heath condition of a patient results as an output, and the value of the output neuron is a function of its activation as represented in the following equation.

$$\ddot{O} = f(\Phi) \tag{8}$$

4.3.2. Particle Swarm Optimization

Particle swarm optimization (PSO) provides a global optimized solution based on the population on a d-dimensional space without any prior knowledge of the issues. By taking the advantages of PSO [49], the synaptic weights are optimized in ANN, which result in a time-efficient prediction in the healthcare domain. In PSO, each particle is evaluated by the objective function at its current location. A swarm particle i is associated with the velocity vector (α_i^d) and position vector (β_i^d). In the network, the velocity vector set is represented as $\alpha_i = \{\alpha_i^1, \alpha_i^2, \ldots \alpha_i^d\}$ and the position vector set is symbolized as $\beta_i = \{\beta_i^1, \beta_i^2, \ldots \beta_i^d\}$, where d is the dimension of the solution space. Initially, random solutions are kept by the network and are eventually used to update the generations to find the best solution. The movement of each swarm particle is decided by combining the random perturbations, current and best (best fitness) locations. During the evolution process, the velocity and position of the particle i on dimension d are updated as follows,

$$\alpha_i^d = \omega \alpha_i^d + \zeta_1 \, R_1^d (P_{Best} - \beta_i^d) + \zeta_2 \, R_2^d (G_{Best} - \beta_i^d) \tag{9}$$

$$\beta_i^d = \beta_i^d + \alpha_i^d \tag{10}$$

where ω is the inertia weight, and ζ_1 and ζ_2 are the acceleration coefficient constants. R_1 and R_2 are two random numbers and are uniformly distributed within [0, 1]. P_{Best} is the best fitness position found so far known as *LocalOptima* for the i^{th} particle, and G_{Best} is the best position in the network, which is called *GlobalOptima*. Once the network is satisfied with the termination condition, the process stops and returns the *GlobalOptima*, which is used by the ANN. The detailed APSO (ANN-PSO) algorithm mechanism is described in Algorithm 2.

Algorithm 2 APSO algorithm

Initialization: Initialize swarm particles (# of neurons (input and hidden), population size, maximum iterations, momentum weight and random variables)
The positions and velocities of all particles are randomized in the search space
for $i = 1$ **to** $i \leq n$ **do**
  $P_{Besti} = \beta_i$
end for
Set $G_{Best} = Min\{f(\beta_i)\}$, where $f(i)$ evaluates the fitness value
Check for termination
if count == maxIter **then**
  Output = G_{Best}
else
  Go to Step 3
end if
Velocity and position update
for $i = 1$ **to** $i \leq n$ **do**
  Update the velocity vector $(\alpha_i) = \omega\alpha_i + \zeta_1 R_1(P_{Best} - \beta_i) + \zeta_2 R_2(G_{Best} - \beta_i)$
  Update the position vector $(\beta_i) = \beta_i + \alpha_i$
  Evaluate the fitness function$(f(\beta_i))$ of i^{th} particle
  if $f(\beta_i)$ is better than $f(P_{Besti})$ **then**
    $P_{Besti} = \beta_i$
  end if
end for
Update $G_{Best} = Min\{f(\beta_i)\}$
Go to Step 2

The patients' data are collected from the Cleveland database [50], where 13 attributes are considered as described in Table 3. Another attribute corresponding to the outcome of the ECG data analysis named "ECG abnormality" is added in the analysis. The duplicate data are extracted and the missing values are filled by preprocessing the modules. Furthermore, the preprocessed data are normalized according to our heart disease scenario. Now, the MapReduce technique is applied to enhance the performance by executing the tasks in parallel within the same or different stages. A two-stage MapReduce function is applied on those historical data sets for training and testing purposes, as shown in Figure 8.

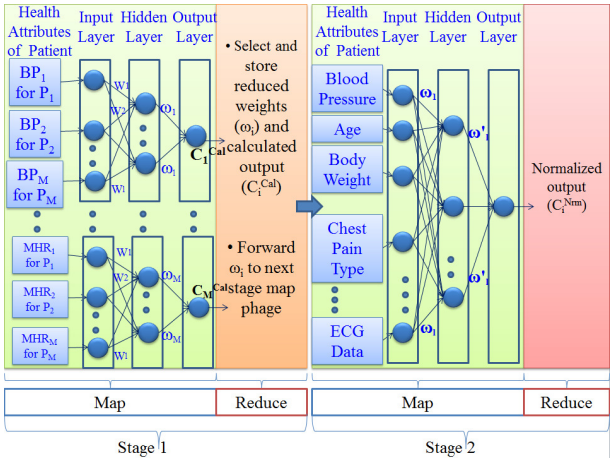

Figure 8. Use of ANN model in MapReduce.

In the first MapReduce stage, our previously defined ANN model is taken as the map function to determine the optimum weight. In this first Map stage, the common parameters S^{bc} such as Blood Pressure (BP), Maximum Heart Rate (MHR), etc. along with some other critical parameters ($\sum_{i=x+1}^{y} C_i$) are considered as the input to the model. Initially, the random

weights are selected for the input to the hidden layer and the hidden to the output layer in an ANN model. Error (Er) is calculated by comparing the calculated output denoted as C_i^{Cal} and the average of the input parameters termed as C_i^{Avg} for each parameter, as shown in Equation (11). By adjusting the weights itself, the error is minimized in each iteration. A threshold error (E_{Th}) is set to $\pm 1\%$ as stopping criteria for the ANN model. This map process is carried out until the error is less than or equal to E_{Th}. After this map process, C_i^{Cal} and final weights involved during stopping iterations are sent to the first reduction phase. In the first reduction phase, we reduce the number of weights between the hidden and the output layer in an ANN instead of all weights. The stored optimum weights are used as the initial weights for the second stage map function in place of random weights.

$$\left.\begin{aligned} Er_i = C_i^{Cal} - C_i^{Avg} \\ s.t.\ Er_i \leq E_{Th} \end{aligned}\right\} \tag{11}$$

In the second stage of MapReduce, the ANN model is used as the map function in which all parameters ($\sum_{i=1}^{l} V_i$) of a patient (P_i) are taken as input. The reduced synaptic weights are assigned to the links between the input and the hidden layer in the second stage. During this supervised learning, the patient's value is provided as 0: $NoHeartDisease$ or ≥ 1: $YesHeartDisease$. By using these historical data, our model is trained to minimize the difference between the computed value and historic value. Similar to the first stage, the E_{Th} value is set to be $\pm 1\%$ as stopping criteria for the ANN model. The output value of the second map function is passed to the second reduction phase for normalization. The normalized output (C_i^{Nrm}) ranges between 0 and 1. If C_i^{Nrm} is greater than 0.5, the patient has the heart disease; else, there is no heart disease. After these two stages of MapReduce, our trained model is ready to predict the heart disease in an efficient way for the newly visiting patients with the given health parameters. The complete disease prediction process is presented in Figure 9.

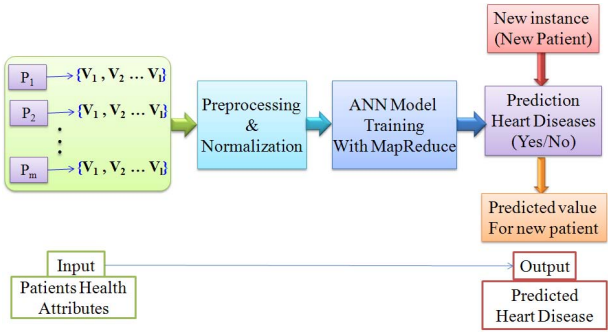

Figure 9. Complete steps of heart disease prediction.

5. Results and Discussions

In this section, simulation results of heart disease prediction are illustrated. In our simulation, heart disease-related medical data are collected from the publicly available Cleveland Clinic Foundation database [50]. The dataset contains the demographic and physiological features of 300 patients (135 normal and 165 heart diseases). The entire Cleveland dataset provides 14 feature attributes where 13 attributes are used as features for heart disease prediction, whereas the attribute named "diagnosis of heart disease" is employed for validating the outcome of heart disease prediction for a patient. The respective dataset contains both categorical and numerical feature values, as shown in Table 3. The numerical data values such as the age of a patient ranging from 29 to 77 years, resting blood pressure varies within 94–200 mmHG, and serum cholesterol (chol) are included with a range of values 126–564 mg/dL. Furthermore, the primary feature predictors such as heart rate (thalach) of values within 71–202 BPM and ST depression (old peak) are studied. Similarly, there are seven attributes such as Sex, Chest pain (cp), Fasting blood sugar (fbs), Resting

electrocardiographic (restecg), Exercise-included angina (exang), Slope of peak exercise ST segment (slope), Presence of thalassemia (thal) and Diagnosis of heart disease (outcome) represent the categorial features. These categorical features are described in terms of two, three, or four categories of values. According to [50], some attributes are normalized and are kept in a range of 0 through 4, as shown in Table 3. Here, 0 is represented as no heart disease, and greater than 0 is symbolized as the presence of heart disease. For instance, the value of age is normalized by assigning 0 if the age is less than 30 years. Likewise, 1 is assigned to the patients who have an age greater than 30 years and less than 50 years.

To validate the proposed ECG abnormality detection model, the popular open MIT-BIH Arrhythmia Database is used [51]. The respective database consists of 48 half-hour excerpts of two-channel ambulatory ECG recordings generated from 47 patients recruited by the BIH Arrhythmia Laboratory between 1975 and 1979. The collected ECG data points are given as input to the derived mathematical model for the determination of the abnormality in the corresponding ECG signal. The prediction from the ECG data analysis named "ECG abnormality" is included as one of the feature sets for the heart disease prediction (Table 3). Hence, the prediction model is trained with 14 features in total by considering "diagnosis of heart disease" as the outcome feature.

For the model derivation, two-thirds of the entire data set is used (200 patients) for training, and one-third of the data set (100 patients) is used for testing purposes. In addition, to establish an efficient predictive model with diversified data values, all 300 patients' data are evaluated through the cross-validation strategy. The efficiency of the proposed solutions are simulated by using the MapReduce framework in Cloudsim simulator and Matlab.

5.1. Simulation Setup

The impact of the proposed cardiac big data analysis framework is evaluated using performance metrics such as accuracy, processing time, total incurred cost and CPU utilization based on data size, and the number of patients. The primary reasons for considering those performance metrics are as follows. Accuracy is one of the influential metrics used to evaluate the prediction model's performance in diagnosing normal and cardiac patients. In case of heterogeneous multi-modal data analysis, the processing time is considered as a crucial factor to measure the efficiency of the classification algorithm. It can justify whether multiple parallel servers are required or whether the processing can be finished using a single server. Furthermore, the CPU utilization rate in a big data center varies based on the data size and the number of patients. The size of one patient's data might be more voluminous than that of multiple patients. In such cases, the CPU utilization varies. Also, this performance can show how effectively data should be managed before processing. Furthermore, the cost analysis should be performed to balance the inflow and outflow of the money. Generally, the incurred cost will be increased with more patients. However, such analysis can assist the regulatory committee in prioritizing their resource investment.

5.2. Simulation Result

In Figure 10, both original and predicted values are plotted by calculating their respective-average normalized values. Instead of all the patients, 100 patients are considered to visualize the data with more clarity in the testing phase. It is observed from the graph that the average predicted values are almost close to the average original values. Hence, our proposed architecture is an efficient predictive model to predict the heart disease for the patient correctly.

The role of important attributes in predicting the heart disease is shown in Figure 11. It is clearly noticed that the accuracy of the prediction is increased as the number of attributes are increased. When we have only one attribute, the accuracy changes to 33% with the help of the training data sets, whereas it boosts up to 38% when two attributes are considered. The accuracy is crossed to 50% with seven attributes, and eventually, a maximum accuracy of 99% is achieved when the number of attributes is increased to 14.

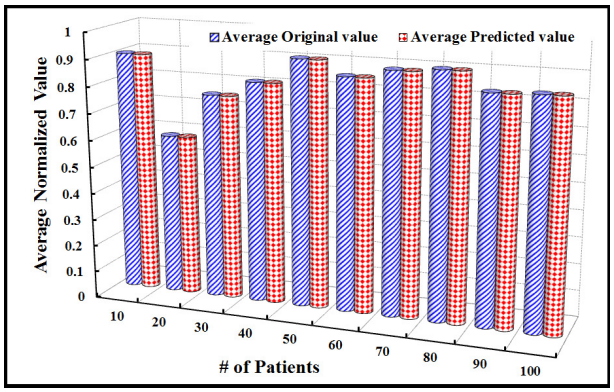

Figure 10. Average predicted values verses average original values.

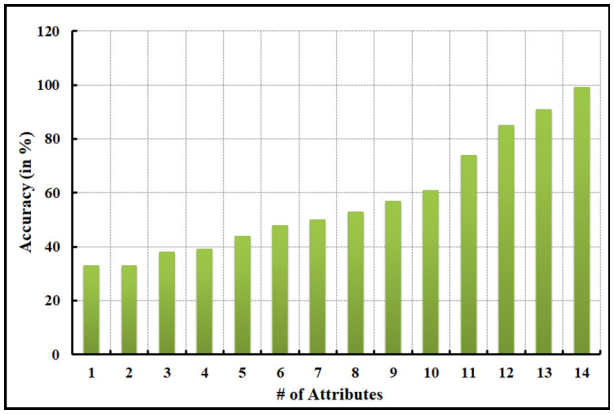

Figure 11. Prediction accuracy verses number of attributes.

To show the correctness, another graph is outlined as shown in Figure 12 by calculating the root mean square (RMS) error. The RMS value is fluctuating due to the varying nature of the predicted and original values. In this sketch, both the average original and predicted value are represented along with a minimum RMS error value for our method.

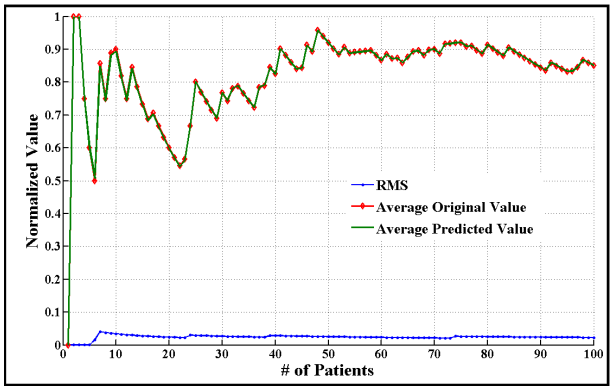

Figure 12. Root mean square error taking different number of patients.

In the CloudSim simulator, the task processing, utilization of the servers, optimum data storage and total cost are analyzed. It is assumed that the data centers are networked and geographically distributed with many servers. The number of users, the current time and the flag are required for tracing all the events. Following the initialization process, all other components like the broker must be initialized, where the broker acts as a bridge in between the user and the cloud provider. An optimum resource utilization with parallel execution is the main focus in data centers; as a result, cost incurred by those data centers are minimized.

Processing time is another crucial factor to measure the efficiency of the prediction algorithm. From Figure 13, it is observed that the parallel servers are beneficial only when a large volume of patient's data are processed and predicted for the heart disease. Initially, more time is taken by the parallel servers to predict the heart condition of fewer patients as compared to the single server as the data are distributed over different data centers. However, when the number of patients increases significantly, a single server is unable to accommodate all the prediction tasks. Hence, more waiting time is required to process in a single server as compared to the parallel servers. Therefore, the processing time is defined as the summation of both execution time as well as the data transfer time coming from the servers located at different places. For example, initially, the time taken by the parallel server (>7 ms) is more than that of the single one up to 60 patients. However, the time consumption is reduced more in parallel servers than in the single one afterwards (<7 ms) and the trend continues to decrease for the rest of the patients. Simultaneously, the average processing time of the single server for the # of patients > 70 is increased to 8 ms.

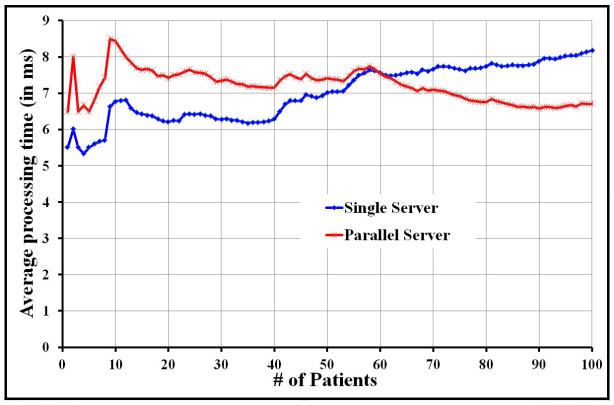

Figure 13. Processing time with number of patients.

In Figure 14, the CPU utilization of the data centers is shown due to processing the huge number of patient records. In our simulation, the input data size is set in gigabytes ranging from 5 to 50 GB by keeping numbers of patients constant at 5000. Heterogeneous servers for each data center are kept with an unequal number of servers ranging from 50 through 70. From the plotted graph, it is concluded that with the increase in the size of data coming to the data center, the utilization of servers increases. For instance, the CPU utilization of the data center DC 1 is increased to 90% for the data size 50 GB. From Figure 14, the harmonic growth of the utilization can be visualized, where for each 5 GB data, the % of CPU utilization is increased by 10%. However, when the amount of data exceeds the capacity of a single data center (that increases the processing time), a new data center needs to be deployed to balance the utilization. Our goal is to maximize the resource utilization without compromising the processing deadline.

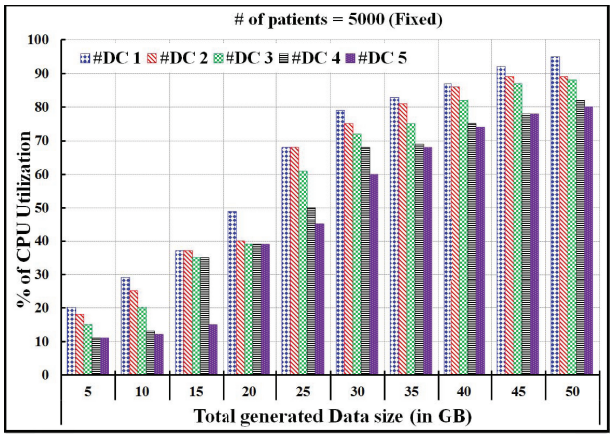

Figure 14. Percentage of CPU utilization with increase in data size.

In Figure 15, the behaviors of different data centers are examined with varying numbers of patients, where each data center is equipped with an unequal number of servers. It is observed that the average percentage of CPU utilization depends on both number of patients and different data centers with a fixed data size. For instance, initially, the CPU utilization of DC_3 is 69%, though the CPU utilization eventually boosts up afterward to 92% with the same 50 GB data size. Furthermore, when the number of patients increased from 1000 to 5000 and the number of data centers rose from 1 to 5, the CPU utilization also increased from 75% to 85% and the trend is continued for the other data centers. But the utilization is decreased from the data centers due to the same amount of workload distribution among the data centers.

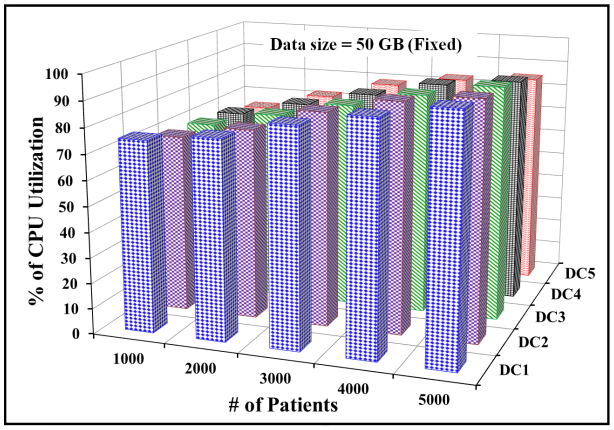

Figure 15. Percentage of CPU utilization in different data centers with varying numbers of patients.

From the revenue point of view, cost is another major factor that cannot be ignored. The incurred cost associated with the increasing number of patients is depicted in Figure 16. In the simulation, bandwidth, storage, computation and data migration costs are taken into consideration. For the cost calculation, the Amazon Web Service (AWS) pricing model is considered, and it is observed that the cost per patient comes out to be approximately 55$. The growth rate of the cost per patient does not follow the same trend irrespective of the location of the data centers. For instance, the total cost is increased linearly for processing the data up to 500 patients though it becomes steady between the patients 600 to 1000.

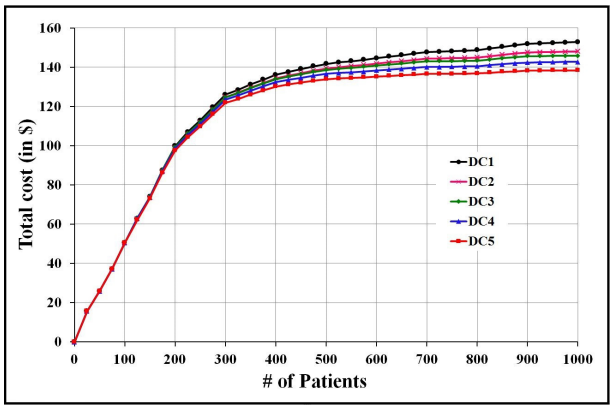

Figure 16. Incurred cost with increase in number of patients.

6. Conclusions

In the current work, a statistical big data analytic framework with the machine learning concept for critical cardiological data analysis in a healthcare environment is proposed. An optimization technique is introduced to minimize the storage, transfer, and processing cost in the cloud. A dynamic big data analytic framework is designed to handle the high volumes of critical patient data and extend it to analyze the ECG batch data as an application for big data processing. In addition, an intelligent heart disease prediction

model is proposed using the approach of ANN, which has the potential to establish an implicit relationship between the complex nonlinear patient's health parameters. The impact of computational and processing complexity on the number of patients and data sizes is experimentally presented. The developed ANN-based big data analytical model enables faster and accurate cardiac disease prediction compared to the manual diagnosis method. The proposed algorithm can efficiently conclude the presence or absence of cardiac abnormality by considering numerous physiological features with the ECG outcome. The proposed framework has various technical and computational implications as it is necessary to normalize the attributes before analysis; otherwise, the accuracy becomes low due to dispersed values, resulting in under-fitting of the model. Furthermore, the number and importance of feature attributes significantly improve the accuracy as some features are essential for the prediction. The chances of over-fitting can be reduced using a large population with a maximum variation of patient data. The ANN is the best solution for faster categorical and numerical data analysis when the number of attributes is few, exhibiting lower correlation among each other. However, LSTM is the optimal choice for the temporal or sequencing medical data analysis, although it possess higher computational complexity during training.

The ECG and physiological health data analysis can be simultaneously performed by integrating the concept of machine learning with deep learning by adopting the feature-fusion methodology. The proposed architecture for processing and analysis could be applied to neonatal health prediction, neurological speculation, especially EHR plus electroencephalogram (EEG) data, and disease classification concerning the patients, diseases, and associated risks. However, the hyperparameter tuning and the analysis time vary based on the feature complexity. Furthermore, the computational and time complexity of batch data processing is lower than that of the real-streaming data. The necessity of a cloud environment can only be realized in case of higher data volume and smaller data sizes. The current study has several limitations in terms of small study population, the varieties of attributes being limited to numerical and categorical data, consideration of batch data and limited number of attributes. Finally, the interdependencies among the attributes for the disease diagnosis are not considered. In our future work, we plan to perform extensive research for cardiac disease prediction considering other data types and modalities such as daily living data, sensor data from wearable devices, social behavior data, and ultrasound image data. Designing an integrated model considering both ML and DL methodologies for analyzing physiological, temporal, cardiac signaling, and image data would be our primary future work. The applicability of the prediction algorithm will be validated in an actual hospital scenario considering the real-time data.

Author Contributions: Conceptualization, S.M., P.K.S. and S.K.M.; methodology, P.K.S.; formal analysis, S.M. and S.K.M.; investigation, P.K.S.; writing—original draft preparation, S.M. and S.K.M.; writing—review and editing, P.K.S.; visualization, S.K.M.; supervision, P.K.S.; project administration, P.K.S. and S.M.; funding acquisition, P.K.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported in part by the National Science and Technology Council (NSTC), Taiwan under grant number 110-2221-E-182-008-MY3.

Data Availability Statement: Data available in a publicly accessible repository. The data presented in this study are openly available in [Heart Disease] at [https://doi.org/10.24432/C52P4X], reference number [50] and [MIT-BIH Arrhythmia Database] at [https://doi.org/10.13026/C2F305], reference number [51].

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhang, L.; Liu, J. Research Progress of ECG Monitoring Equipment and Algorithms Based on Polymer Materials. *Micromachines* **2021**, *12*, 1282. [CrossRef] [PubMed]
2. Sahoo, P.K.; Thakkar, H.K.; Lee, M.Y. A cardiac early warning system with multi channel SCG and ECG monitoring for mobile health. *Sensors* **2017**, *17*, 711. [CrossRef] [PubMed]
3. Vodička, S.; Susič, A.P.; Zelko, E. Implementation of a savvy mobile ECG sensor for heart rhythm disorder screening at the primary healthcare level: An observational prospective study. *Micromachines* **2021**, *12*, 55. [CrossRef] [PubMed]
4. Kaur, K.; Rani, R. Managing data in healthcare information systems: Many models, one solution. *Computer* **2015**, *48*, 52–59. [CrossRef]
5. Rehman, A.; Naz, S.; Razzak, I. Leveraging big data analytics in healthcare enhancement: Trends, challenges and opportunities. In *Multimedia Systems*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 1–33.
6. Hong, L.; Luo, M.; Wang, R.; Lu, P.; Lu, W.; Lu, L. Big data in health care: Applications and challenges. *Data Inf. Manag.* **2018**, *2*, 175–197. [CrossRef]
7. Mehta, N.; Pandit, A.; Shukla, S. Transforming healthcare with big data analytics and artificial intelligence: A systematic mapping study. *J. Biomed. Inform.* **2019**, *100*, 103311. [CrossRef] [PubMed]
8. Thakkar, H.K.; Sahoo, P.K. Towards automatic and fast annotation of seismocardiogram signals using machine learning. *IEEE Sens. J.* **2019**, *20*, 2578–2589. [CrossRef]
9. Cauteruccio, F.; Terracina, G. Extended High-Utility Pattern Mining: An Answer Set Programming-Based Framework and Applications. In *Theory and Practice of Logic Programming*; Cambridge University Press: Cambridge, UK, 2023; pp. 1–31.
10. Sermesant, M.; Delingette, H.; Cochet, H.; Jaïs, P.; Ayache, N. Applications of artificial intelligence in cardiovascular imaging. *Nat. Rev. Cardiol.* **2021**, *18*, 600–609. [CrossRef]
11. Chang, Z.; Zhang, C.; Li, C. Motor Imagery EEG Classification Based on Transfer Learning and Multi-Scale Convolution Network. *Micromachines* **2022**, *13*, 927. [CrossRef]
12. Calimeri, F.; Cauteruccio, F.; Cinelli, L.; Marzullo, A.; Stamile, C.; Terracina, G.; Durand-Dubief, F.; Sappey-Marinier, D. A logic-based framework leveraging neural networks for studying the evolution of neurological disorders. *Theory Pract. Log. Program.* **2021**, *21*, 80–124. [CrossRef]
13. Awotunde, J.B.; Jimoh, R.G.; Ogundokun, R.O.; Misra, S.; Abikoye, O.C. Big data analytics of iot-based cloud system framework: Smart healthcare monitoring systems. In *Artificial Intelligence for Cloud and Edge Computing*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 181–208.
14. Sellami, M.; Mezni, H.; Hacid, M.S. On the use of big data frameworks for big service composition. *J. Netw. Comput. Appl.* **2020**, *166*, 102732. [CrossRef]
15. Hussain, I.; Park, S.J. Big-ECG: Cardiographic predictive cyber-physical system for stroke management. *IEEE Access* **2021**, *9*, 123146–123164. [CrossRef]
16. Sahoo, P.K.; Mohapatra, S.K.; Wu, S.L. SLA based healthcare big data analysis and computing in cloud network. *J. Parallel Distrib. Comput.* **2018**, *119*, 121–135. [CrossRef]
17. Muneeswaran, V.; Nagaraj, P.; Dhannushree, U.; Ishwarya Lakshmi, S.; Aishwarya, R.; Sunethra, B. A Framework for Data Analytics-Based Healthcare Systems. In *Innovative Data Communication Technologies and Application*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 83–96.
18. Allam, S. Research on intelligent medical big data system based on Hadoop and blockchain. *Int. J. Emerg. Technol. Innov. Res.* **2021**, *8*, 1393–1398.
19. Sahoo, P.K.; Mohapatra, S.K.; Wu, S.L. Analyzing Healthcare Big Data With Prediction for Future Health Condition. *IEEE Access* **2016**, *4*, 9786–9799. [CrossRef]
20. Jiang, L.; Da Xu, L.; Cai, H.; Jiang, Z.; Bu, F.; Xu, B. An IoT-oriented data storage framework in cloud computing platform. *IEEE Trans. Ind. Inform.* **2014**, *10*, 1443–1451. [CrossRef]
21. Rahman, L.A.; Rana, M.E. The Convergence Between Big Data and the Cloud: A Review. In Proceedings of the 2021 International Conference on Data Analytics for Business and Industry (ICDABI), Sakheer, Bahrain, 25–26 October 2021; pp. 592–598.
22. Sahoo, P.K.; Thakkar, H.K.; Lin, W.Y.; Chang, P.C.; Lee, M.Y. On the design of an efficient cardiac health monitoring system through combined analysis of ECG and SCG signals. *Sensors* **2018**, *18*, 379. [CrossRef]
23. Rodríguez-Jorge, R.; De León-Damas, I.; Bila, J.; Škvor, J. Internet of things-assisted architecture for QRS complex detection in real time. *Internet Things* **2021**, *14*, 100395. [CrossRef]
24. Bae, T.W.; Kwon, K.K. Efficient real-time R and QRS detection method using a pair of derivative filters and max filter for portable ECG device. *Appl. Sci.* **2019**, *9*, 4128. [CrossRef]
25. do Vale Madeiro, J.P.; Marques, J.A.L.; Han, T.; Pedrosa, R.C. Evaluation of mathematical models for QRS feature extraction and QRS morphology classification in ECG signals. *Measurement* **2020**, *156*, 107580. [CrossRef]
26. Bae, T.W.; Lee, S.H.; Kwon, K.K. An adaptive median filter based on sampling rate for R-peak detection and major-arrhythmia analysis. *Sensors* **2020**, *20*, 6144. [CrossRef] [PubMed]
27. Zhang, X.; Gu, K.; Miao, S.; Zhang, X.; Yin, Y.; Wan, C.; Yu, Y.; Hu, J.; Wang, Z.; Shan, T.; et al. Automated detection of cardiovascular disease by electrocardiogram signal analysis: A deep learning system. *Cardiovasc. Diagn. Ther.* **2020**, *10*, 227. [CrossRef] [PubMed]

28. Avanzato, R.; Beritelli, F. Automatic ECG diagnosis using convolutional neural network. *Electronics* **2020**, *9*, 951. [CrossRef]
29. Li, Q.; Campan, A.; Ren, A.; Eid, W.E. Automating and improving cardiovascular disease prediction using Machine learning and EMR data features from a regional healthcare system. *Int. J. Med. Inform.* **2022**, *163*, 104786. [CrossRef] [PubMed]
30. Gárate-Escamila, A.K.; El Hassani, A.H.; Andrès, E. Classification models for heart disease prediction using feature selection and PCA. *Inform. Med. Unlocked* **2020**, *19*, 100330. [CrossRef]
31. Manimurugan, S.; Almutairi, S.; Aborokbah, M.M.; Narmatha, C.; Ganesan, S.; Chilamkurti, N.; Alzaheb, R.A.; Almoamari, H. Two-stage classification model for the prediction of heart disease using IoMT and artificial intelligence. *Sensors* **2022**, *22*, 476. [CrossRef] [PubMed]
32. Choi, S.Y.; Chung, K. Knowledge process of health big data using MapReduce-based associative mining. *Pers. Ubiquitous Comput.* **2020**, *24*, 571–581. [CrossRef]
33. Demirbaga, U.; Aujla, G.S. MapChain: A blockchain-based verifiable healthcare service management in IoT-based big data ecosystem. *IEEE Trans. Netw. Serv. Manag.* **2022**, *19*, 3896–3907. [CrossRef]
34. Babar, M.; Jan, M.A.; He, X.; Tariq, M.U.; Mastorakis, S.; Alturki, R. An Optimized IoT-Enabled Big Data Analytics Architecture for Edge–Cloud Computing. *IEEE Internet Things J.* **2022**, *10*, 3995–4005. [CrossRef]
35. Safa, M.; Pandian, A.; Gururaj, H.; Ravi, V.; Krichen, M. Real time health care big data analytics model for improved QoS in cardiac disease prediction with IoT devices. *Health Technol.* **2023**, *13*, 473–483. [CrossRef]
36. Shaik, K.; Ramesh, J.V.N.; Mahdal, M.; Rahman, M.Z.U.; Khasim, S.; Kalita, K. Big Data Analytics Framework Using Squirrel Search Optimized Gradient Boosted Decision Tree for Heart Disease Diagnosis. *Appl. Sci.* **2023**, *13*, 5236. [CrossRef]
37. Kim, J. Energy-efficient dynamic packet downloading for medical IoT platforms. *IEEE Trans. Ind. Inform.* **2015**, *11*, 1653–1659. [CrossRef]
38. Chakraborty, C.; Kishor, A. Real-time cloud-based patient-centric monitoring using computational health systems. *IEEE Trans. Comput. Soc. Syst.* **2022**, *9*, 1613–1623. [CrossRef]
39. Gupta, C.; Saha, A.; Reddy, N.S.; Acharya, U.D. Cardiac Disease Prediction using Supervised Machine Learning Techniques. *J. Phys. Conf. Ser.* **2022**, *2161*, 012013. [CrossRef]
40. Srinivasan, S.; Gunasekaran, S.; Mathivanan, S.K.; Benjula Anbu Malar, M.B.; Jayagopal, P.; Dalu, G.T. An active learning machine technique based prediction of cardiovascular heart disease from UCI-repository database. *Sci. Rep.* **2023**, *13*, 13588. [CrossRef]
41. El-Hasnony, I.M.; Elzeki, O.M.; Alshehri, A.; Salem, H. Multi-label active learning-based machine learning model for heart disease prediction. *Sensors* **2022**, *22*, 1184. [CrossRef]
42. Al Bataineh, A.; Manacek, S. MLP-PSO hybrid algorithm for heart disease prediction. *J. Pers. Med.* **2022**, *12*, 1208. [CrossRef]
43. Bhavekar, G.S.; Goswami, A.D. A hybrid model for heart disease prediction using recurrent neural network and long short term memory. *Int. J. Inf. Technol.* **2022**, *14*, 1781–1789. [CrossRef]
44. Pathan, M.S.; Nag, A.; Pathan, M.M.; Dev, S. Analyzing the impact of feature selection on the accuracy of heart disease prediction. *Healthc. Anal.* **2022**, *2*, 100060. [CrossRef]
45. Pan, C.; Poddar, A.; Mukherjee, R.; Ray, A.K. Impact of categorical and numerical features in ensemble machine learning frameworks for heart disease prediction. *Biomed. Signal Process. Control* **2022**, *76*, 103666. [CrossRef]
46. Ozcan, M.; Peker, S. A classification and regression tree algorithm for heart disease modeling and prediction. *Healthc. Anal.* **2023**, *3*, 100130. [CrossRef]
47. Verma, L.; Srivastava, S.; Negi, P. A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data. *J. Med. Syst.* **2016**, *40*, 1–7. [CrossRef] [PubMed]
48. Auger, S.D.; Jacobs, B.M.; Dobson, R.; Marshall, C.R.; Noyce, A.J. Big data, machine learning and artificial intelligence: A neurologist's guide. *Pract. Neurol.* **2021**, *21*, 4–11. [CrossRef] [PubMed]
49. Chou, F.I.; Huang, T.H.; Yang, P.Y.; Lin, C.H.; Lin, T.C.; Ho, W.H.; Chou, J.H. Controllability of Fractional-Order Particle Swarm Optimizer and Its Application in the Classification of Heart Disease. *Appl. Sci.* **2021**, *11*, 11517. [CrossRef]
50. Clevenland Database. Available online: https://archive.ics.uci.edu/dataset/45/heart+disease (accessed on 12 March 2021).
51. MIT-BIH Arrhythmia Database. Available online: https://www.physionet.org/content/mitdb/1.0.0/ (accessed on 30 March 2021).