Computers, Materials & Continua
DOI:10.32604/cmc.2022.019458
Article

Tech Science Press

Applying Apache Spark on Streaming Big Data for Health Status Prediction

Ahmed Ismail Ebada[1], Ibrahim Elhenawy[2], Chang-Won Jeong[3], Yunyoung Nam[4,*], Hazem Elbakry[1] and Samir Abdelrazek[1]

[1]Information Systems Department, Faculty of Computers and Information, Mansoura University, Mansoura, 35516, Egypt
[2]Department of Computer Science, Faculty of Computers and Information, El-Zagazig University, Zagazig, Sharqiyah, 44519, Egypt
[3]Medical Convergence Research Center, Wonkwang University, Iksan, Korea
[4]Department of Computer Science and Engineering, Soonchunhyang University, Asan, Korea
[*]Corresponding Author: Yunyoung Nam. Email: ynam@sch.ac.kr
Received: 14 April 2021; Accepted: 31 May 2021

Abstract: Big data applications in healthcare have provided a variety of solutions to reduce costs, errors, and waste. This work aims to develop a real-time system based on big medical data processing in the cloud for the prediction of health issues. In the proposed scalable system, medical parameters are sent to Apache Spark to extract attributes from data and apply the proposed machine learning algorithm. In this way, healthcare risks can be predicted and sent as alerts and recommendations to users and healthcare providers. The proposed work also aims to provide an effective recommendation system by using streaming medical data, historical data on a user's profile, and a knowledge database to make the most appropriate real-time recommendations and alerts based on the sensor's measurements. This proposed scalable system works by tweeting the health status attributes of users. Their cloud profile receives the streaming healthcare data in real time by extracting the health attributes via a machine learning prediction algorithm to predict the users' health status. Subsequently, their status can be sent on demand to healthcare providers. Therefore, machine learning algorithms can be applied to stream health care data from wearables and provide users with insights into their health status. These algorithms can help healthcare providers and individuals focus on health risks and health status changes and consequently improve the quality of life.

Keywords: Big data; streaming processing; healthcare data; machine learning; IoT data processing; Apache Spark

1 Introduction

Big data in healthcare systems are composed of large amounts of data and can be applied to obtain insights into healthcare data and support healthcare decisions. Healthcare decision-making depends totally on available information on various clinical devices, economic aspects, or social factors. These decisions have no value if they are not made promptly. Poor, late, or wrong

decisions are taken because knowledge is unavailable at the right time. Medical devices based on Internet of Things (IoT) offer service providers with a variety of solutions in the healthcare domain, which is needed to be analyzed to predict unknown data. IoT wearable devices provide medical solutions with data collected from different data sources, which can be collected from various patients. Medical data should be analyzed and further elucidated [1]. However, using streaming medical data is challenging because data are large and incomplete; time is also an important factor. This research is based on the availability of healthcare datasets and advances in machine learning algorithms. Healthcare is equipped with technologies that aids in the diagnosis of many health issues.

Numerous available smart sensors can be used to measure human medical data continuously. For instance, an electrocardiogram (ECG) sensor measures the heart rate (HR) [2]. Heart diseases can be monitored by finding the changes between heartbeats [3]. Two devices are available for calculating HR, ECG, and photo plethysmo gram (PPG), which can be placed on smart wearables [4]. Smartphones and smart wearables help users perform their daily activities. They use wireless connections to provide healthcare systems with recommendations about an optimal healthy life based on available data [5].

Medical wearable devices can be implantable or wearable. They consist of microprocessors, memory storage, sensors, and communication interfaces. They measure physical vital signs, HR, respiration rate, body temperature, and other parameters. Some wearable devices collect data such as population, light, sound, and carbon dioxide density from the surrounding environment [6]. With the commercialization of these devices, more advanced features, such as tracker of heart disease, have been developed [7].

Multiclass neural network and multiclass random forest algorithms represent an efficient inference machine for mobile computing applications. Multiclass classification involves more than two classes, such as classifying a set of images of fruits (e.g., pears, oranges, or apples). The development of mobile smart wearable devices can contribute to the reduction of health prediction problems. As such, this study proposes the use of a multiclass neural network and a multiclass random forest algorithm for breast cancer prediction based on a dataset from electronic health records (EHRs). This model can support health experts in uncertain moments, anywhere and anytime, by using mobile applications, reducing risky situations, and giving a recommendation to users to visit a doctor. This study aims to propose an inteligent big data architecture for healthcare applications based on several components capable of storing, processing, and analyzing a high amount of historical medical data in real time. It also demonstrates the potential of using the proposed system in big data analytics in the healthcare area to find useful information in highly valuable data and thus help healthcare providers.

This paper is organized as follows. Section 2 presents the related works in medical healthcare solutions. Section 3 describes the proposed system based on streaming data analysis by using Apache Spark and its mechanism. Section 4 shows the implementations and datasets, Section 5 discusses the results, and Section 6 presents the conclusion and recommends relevant topics for future work.

1.1 Healthcare Prediction System

The availability of new technologies has led to remarkable achievements not only in disease detection and diagnosis but also in disease prediction. Prediction in healthcare systems aims to determine the occurrence of a potential disease or the early detection of a disease. Prediction is based on different sources, such as electronic medical records from healthcare wearables,

healthcare records, and healthcare reports. Monitoring health status can be informative by collecting healthcare data, such as heart rate, blood sugar, and other health parameters, from wearable devices to monitor changes in health status over time. Wearable biosensors are widely used in applications for hospitals, sports, and fitness by using smartwatches, bracelets, and other devices with heart rate trackers and accelerometers. Being healthy has become a lifestyle because of the awareness that the prevention or early detection of potential diseases, such as heart disease and cancer, can improve the survival rate. Applying machine learning methods to healthcare data provides healthcare providers with solutions that can predict health issues at acceptable accuracy.

1.2 Streaming Health Data to the Cloud

Streaming health data provides an effective solution to support decision support systems in the healthcare area. Many researchers developed new methods and technologies to employ machine learning algorithms on big data and predict diseases before they could even occur. For example, Spark rapidly processes a large amount of data because it can do computations in memory. This characteristic is useful for data processing in runtime and decision-making. Many researchers also used Twitter as a tool to stream different data sources in big data analysis. As a third-party service, Twitter is an efficient and free real-time communication tool. It can support a size of a maximum of 140 characters in a post. It can send messages shown by followers without being visible to others. Therefore, it is an effective communication tool in smartphone environments where resources are insufficient for processing.

2 Related Work

Big medical data can be processed with either of the two methodologies: stream processing or batch processing [8]. In batch processing, data are analyzed data over a certain period. In stream processing, streaming data are examined to obtain quick feedback [9]. Stream processing is very useful because streaming applications can collect and make decisions on runtime, and batching and analyzing data are not time consuming. Streaming applications can be used for a new generation of transportation applications, healthcare, and automation [10]. However, streaming processing is greatly limited by memory capacity. In addition to real-time application usually sends a high amount of data with noise, errors, and missing records because of connection issues.

Feedback about streaming data can be obtained by using some methods, such as frequent pattern mining, clustering, and classification [11]. Real-time feedback on a large amount of patients' data can be provided by big medical data. Streaming data is useful in healthcare systems because it can send alerts when prediction analysis detects possible health issues. This prediction method from streaming data can help decrease medical diagnosis errors and give a recommendation about diseases to healthcare providers [12].

Pongsakornsathien et al. [13] presented a medical system to monitor the health conditions of individuals and collected data such as HR during daily life to recommend a better and healthier lifestyle. They proposed a solution that employs batch data, but they did not provide a solution for streaming data. Ismail et al. [14] developed a wearable device based on wireless connection and multiple sensors for physical measurements to help users receive alerts and recommendations on real-time processing on smartphones. They also provided a solution that saves resources by sending only important alerts, but they did not provide an analysis system for healthcare systems. In another paper [15], different wearable devices based on medical body sensors, such as smart applications, are reviewed in relation to their effectiveness at home and in a hospital. The previous

review indicated that smart wearables are useful for giving more data about patients. However, they are limited and do not achieve the precision and safety required by medical staff.

The A healthcare system for disease detection based on indoor wearable devices has also been developed [16,17]. It can provide data about health characters over time. The main strength in their work is that data should be analyzed or monitored by a healthcare provider to gain knowledge. However, the implementation of the system is limited to indoor applications. Heryana et al. [18] presented a healthcare service to monitor heart diseases by using wearable devices. They developed a system composed of several phases. First, noninvasive wireless sensors are used. Second, data are stored in a server via a sensor gateway equipped with an alerting system that send an alert only when abnormal conditions are detected. In the third phase, cloud computing is applied to handle the stored data and allow patients to use the service for monitoring their heart conditions. In the fourth phase, an application for tablet devices and smartphones is developed to monitor health issues and measurements. However, such systems are complex, and human experts are needed to analyze health data.

The proposed system differs from the discussed solutions. In particular, the proposed solution is based on a wearable smart device that monitors different health parameters with a multisensor. It is a healthcare system based on the Microsoft Azure cloud system to obtain streaming data from wearable devices and predict breast cancer. It is also a multisensory system that monitors HR, blood pressure, brain activities, and muscle activities. Monitored data are sent to smartphones wirelessly in accordance with the ZigBee [18] radio standard. Smartphones then send data to a user's profile in the cloud, which contains all the medical and historical data of patients. Lastly, a machine learning model is run in the cloud system to inform healthcare providers about breast cancer predictions.

3 Methodology

Many healthcare companies employ medical big data to predict health risks before they could even occur [19]. Apache Spark rapidly deals with the big data of unstructured and structured healthcare datasets because it provides memory computations. It can process data 100 times faster than traditional map-reducing methods. Spark contains a lambda architecture that allows real-time processing and batch processing. It is used to analyze streaming healthcare data and has a library that can handle Twitter data streams. MLlib in Spark, which is a machine learning library, can support decision tree implementation for disease predictions. Spark is an analysis engine for big data during streaming. It can handle streaming better than Hadoop because the former provides memory processing. Data and intermediate results are stored in memory, thereby avoiding the input–output delay of switching data back and forth the hard disk. Spark applies the concept of resilient distributed dataset (RDD) [20], which is an immutable object collection that can be retained in memory and partitioned across nodes for parallel computations.

Though Spark is a batch processing engine, it can process streaming data from different sources, including those from Twitter. Incoming data streams are divided into small batches that are then processed by the Spark engine. In Spark streaming, discretized streams (DStreams) are a sequence of RDDs that represent a continuous data stream. Operations on DStreams are converted to basic RDD transformations, which are then computed by the Spark engine. MLlib consists of popular learning algorithms, such as classification, clustering, and collaborative filtering, which helps resolve machine learning problems.

In this research, wearable medical sensors are used to continuously monitor body measurements, including heart rate, diabetic level, blood pressure, body temperature, physical activities, and blood oxygen levels. Health monitoring has become a lifestyle, and people prefer preventing diseases to curing them [21]. Wearable smart multisensory devices can help predict and detect diseases, such as heart disease and cancer; they can also contribute to a decrease in health risks [22]. The proposed system (Fig. 1) is based on a wearable device connected to a cloud system for transmitting streaming data. The proposed system proposed two methodologies for authentication to allow healthcare providers to check medical data on demand by using a smart ring with radio-frequency identification. In this way, healthcare user's profile in the cloud can be accessed for a certain period depending on a patient's choice. When healthcare providers gain access to a patient user's profile in the cloud, they can check all related medical data, such as laboratory tests, family historical data, previous and current diseases, and data obtained by wearable devices. Thus, users, doctors, nurses, physicians, and healthcare providers save time because data can be securely accessed in the cloud without risks. Furthermore, the proposed system offers a solution for sharing medical data on demand to help doctors make the right decisions.

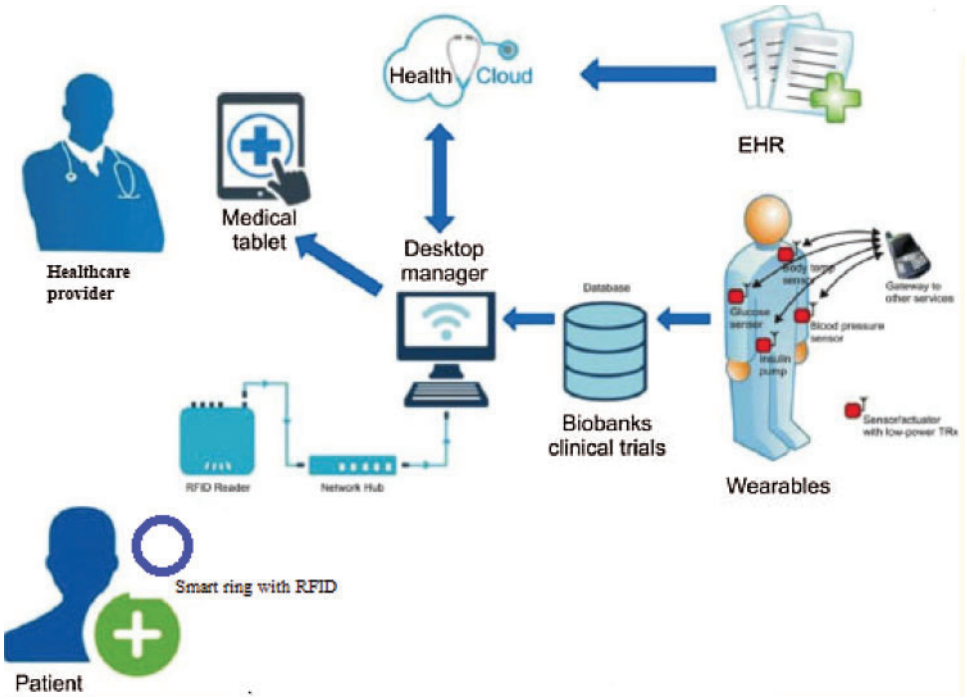

Figure 1: Proposed system for disease monitoring in the cloud

The proposed system is based on 15 characteristics most related to human activities and physical measurements: sex, age, use of psychoactive drugs, number of drugs taken daily, history of previous falls, laboratory reports, ability to stand up and sit down, common diseases in the area, mobility disorders, neuropsychiatric disorders, chronic diseases of the family, and historical medical reports. By applying machine learning algorithms to healthcare data, Spark can analyze and predict health issues with reasonable accuracy [23]. With the proposed method, patients can undergo an initial screening test or access a recommendation system to visit a doctor. Wearable devices collect measurements via a medical sensor. These measurements, along with with available

medical data, are used by the proposed system to give medical recommendations and track health status. Streaming health status data are filtered via abnormality filtration during real-time streaming. Then, the extracted health data are inputted to the machine learning model to predict the health status.

3.1 EHR Applications

EHRs can be accessed by patients and shared as valuable medical records. They are recorded in a private environment or shared securely [3]. They give valuable information from the health data history of users. They can contain measurements from wearable devices or other physical devices at doctors. These measurements include skin temperature, heart rate, blood pressure, respiratory volume, quality of surrounding air, sleep, and food consumed.

Healthcare prediction or a healthcare recommender system, which is the proposed system in this study, is any healthcare system that can help patients have preventive care through diagnosis, treatment, monitoring, and follow-up. EHRs contain all the data needed by clinicians or healthcare provides to give recommendations or prescribe the needed medicines [24]. The proposed system provides a smart solution to manipulate big data coming from different resources and depends on available information. The proposed system relies on information provided by doctors and other experts. The system does not collect just information about the disease for patients; information is about a disease with one choice answer "yes" or "no" and the number that represents the disease level or state. Users also provide system information about normal activities with the number of occurrences.

Medical big data include images (e.g., X-rays, MRIs, and photos), audio files, medical reports, and analysis of waves such as EEG and ECG [25]. Apache Spark collects medical EHR data from different sources, such as wearable devices and users' medical data profiles stored in Hadoop Distributed File System. The collected heterogeneous data can be filtered via Spark filter transformation to remove noisy data. The proposed system is based on applying a real-time connection to Twitter Streaming API to access real-time tweet streams, which require authorization. Twitter is using OAuth, which is the authentication protocol that authorizes applications to access services on users' behalf without sharing their password. As such, a Twitter application needs to be created and registered with a reading/write access. Consumer secret keys and access tokens must be generated because they are required so that the health care application can have authorized access to tweet streams.

3.2 Proposed System

Studies on healthcare have widely explored the application of machine learning algorithms in available large medical data rather than traditional methods. In this study, an efficient architecture is presented to work with streaming data from different wearable devices in healthcare systems. Microsoft Azure instead of a standalone server is utilized with streaming data. With the proposed healthcare system, a data-driven approach can be used to provide recommendations for breast cancer. An available dataset collected from the University of Wisconsin Hospital in Madison, Wisconsin, USA, is used. The proposed system employed a confusion matrix for multiclass neural network accuracy and a multiclass random forest to classify the dataset. The results reveal that the proposed system has a high accuracy in predicting breast cancer with low false alarms.

The proposed healthcare prediction system provides healthcare providers with patient-centric information on a patient's profile in the cloud to enhance the quality of healthcare services. The proposed prediction system provides patients with integrated data from doctors, medication

history, clinic reports, laboratory reports, medical images, wearable device data, and health forum data into a prediction diagnosis system to provide recommendations. A recommendation is a group of services such as preventative care reminders, diagnosis prediction, drug interaction alerts, health insurance suggestions, follow-up reminders, diet recommendations, and rebuying medicines. Through the proposed system, a user's medical data can be shared with an expert system to predict diseases or offer recommendations based on machine learning algorithms.

Traditional prediction systems are used mainly in marketing and recommending an approach regarding earlier interactions between customers and products, predicting the user's ratings on items, and determining whether or not a user liked a product [26]. The proposed system (Fig. 2) collects and models EHRs based on medical data, family medical data, streaming data from wearable devices, and users' profile data, such as location and common diseases. Data are categorized on the basis of users' current medical disorders, family chronic diseases, and healthcare providers' recommendations. The system gives a prediction based on the high rates of risks regarding data insights.

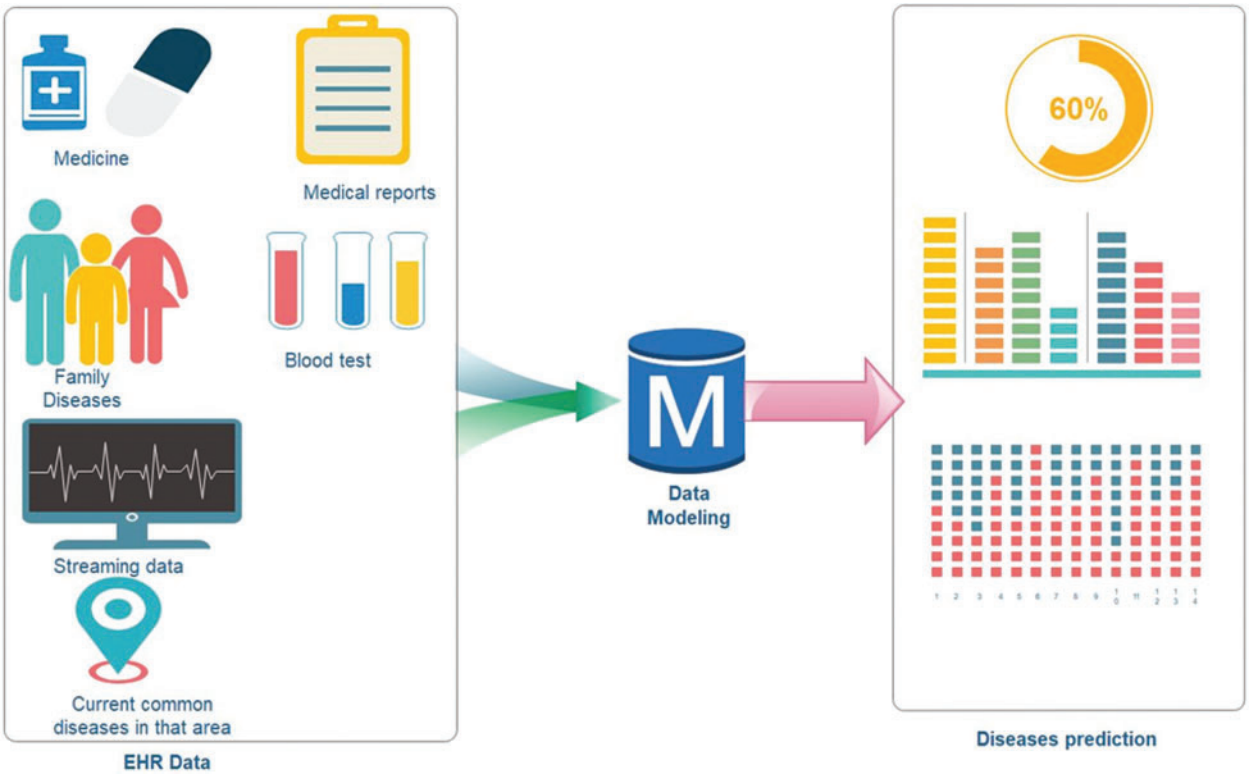

Figure 2: Proposed data nodeling

The proposed system has a telemedical system that helps patients share their profiles in the cloud with healthcare providers who offer recommendations. In this way, the costs of medical visits or medical checkups can be reduced. This series is called the Medical Internet of Things (MIoT). In the proposed system, MapReduce from Hadoop is employed to break down data from different medical sources into smaller datasets to be processed separately. The results of these

smaller datasets are sum reduced and mapped back to provide the outcome of processed data. The map-reduced framework can handle a large dataset to be analyzed by Apache Spark.

In the proposed system, a general model is developed to handle different diseases. It can be used as a general recommendation system by healthcare providers. Different medical history profiles from patients are created and correlated to present representations based on disease conditions and symptoms and determine the optimal recommendations for the same diseases. In Fig. 3, the proposed system is based on a medical ring that sends medical data to a mobile phone through a Bluetooth connection. Medical data are transmitted through the Internet to a user's profile in the cloud and saved on Cosmos DB as a hybrid database. Medical data from healthcare providers are recorded on a Datawarehouse. The Datawarehouse data are presented with Microsoft power BI in the cloud to show predictions.

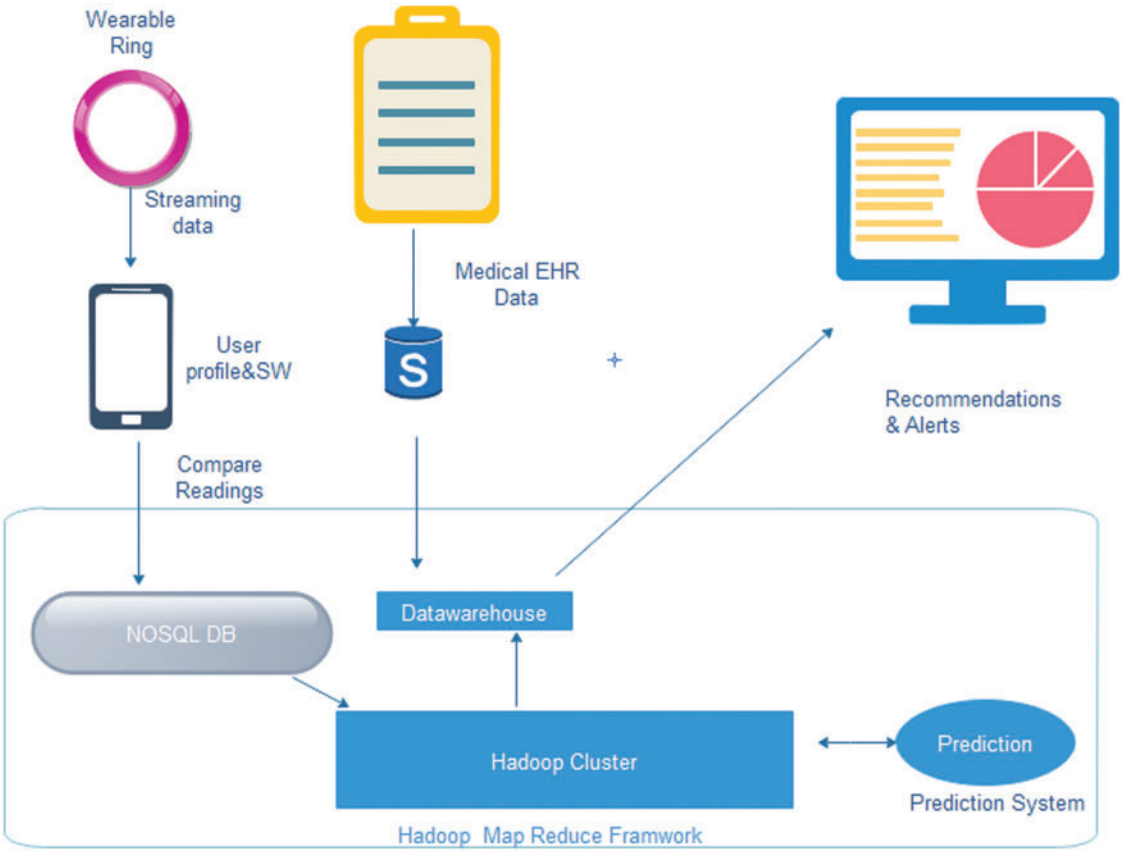

Figure 3: Proposed recommendation system

3.3 Disease Prediction Methodology

In the proposed healthcare system, the system provides decisions as a recommendation or alerts to healthcare providers. The proposed recommendation algorithm collects data from a user's profile and medical devices to give insights into how the indicators of a disease are more risky than completely safe (Tab. 1).

Table 1: Algorithm 1: Disease diagnosis recommendation methodology

```
Input 1: Sensor data (i.e., blood pressure measurements)
Input 2: Set of historical data of a user's diseases (i.e., previous diseases)
Input 3: Set of historical data of diseases of the user's family (i.e., the user's father has diabetes)
Input 4: Set of common diseases in the user's region (i.e., measurements based on GPS location)
Input 5: Set of the user's profile data (i.e., habits, weight, habits, food, activities)
Output records: Disease Name, Disease Level, Disease Value, Disease Probability)
Start
{
// scale analysis for measurements.
Scale Result = Scale Analytics for the inputs.
The result of Irregular Scale = Scale Result. Size ( ).
// Start pattern analysis for measurements.
Pattern Result = execute Pattern Analytics (measurements);
Irregular Pattern Result = Pattern Result. Size ( );
// frequency analysis for the measurements.
Frequency = Frequency Analysis execution for the inputs.
Irregular Frequency = Frequency Result. Size ( );
If (The result value in Irregular Range)
UDA. add (Disease, Level, Probability);
Return UDA;
}
                                End
```

User disease alert is the input record that shows alerts or disease prediction. Disease prediction is presented on the basis of the current health state of users, and a prediction value is generated for a given disease. Every disease has an attribute for a certain disease level, and disease a threshold-based probability. When the UDA of probability is lower than the disease threshold value, then the health state of users is registered to be safe. If the probabilistic value is greater than the disease threshold value, then the person is marked unsafe. The alert-based threshold of the proposed system (α) applies the following conditions:

(1) If USERi HEALTH = Unsafe) AND (P(user) $< \alpha$), then a warning alert is shown in the healthcare system. This signal provides healthcare providers with timely information about personal health to avoid future health issues.
(2) If (USERi HEALTH = Unsafe) AND (P(user) $> \alpha$), then a warning alert is sent to a healthcare emergency department nearest the user or relatives. Alert messages are also delivered to users on their wearable devices. This algorithm describes the alert generation methodology. The disease name with its probability gives specific knowledge about a user's current health status to healthcare providers. Thus, the proposed healthcare system provides a diagnosis method based on IoT wearable devices. It can help healthcare providers diagnose diseases at early stages to increase the possibility of curing chronic diseases, such as breast cancer.

The proposed system is based on streaming healthcare data to continue the live monitoring of healthcare status. In Fig. 4, users' health status are sent through smartphones via internet

connections. Then, the streaming API keeps sending the health status data with push technology and provides a subscription mechanism so that users can choose to receive alerts in real time. Once the Apache Spark streaming starts, it continuously receives streaming data every 30 s and then filters data attributes for data analysis. The proposed system is built on the Microsoft Azure cloud to handle data and apply the machine learning model on the breast cancer dataset. In this way, it sends alerts based on users' health status.

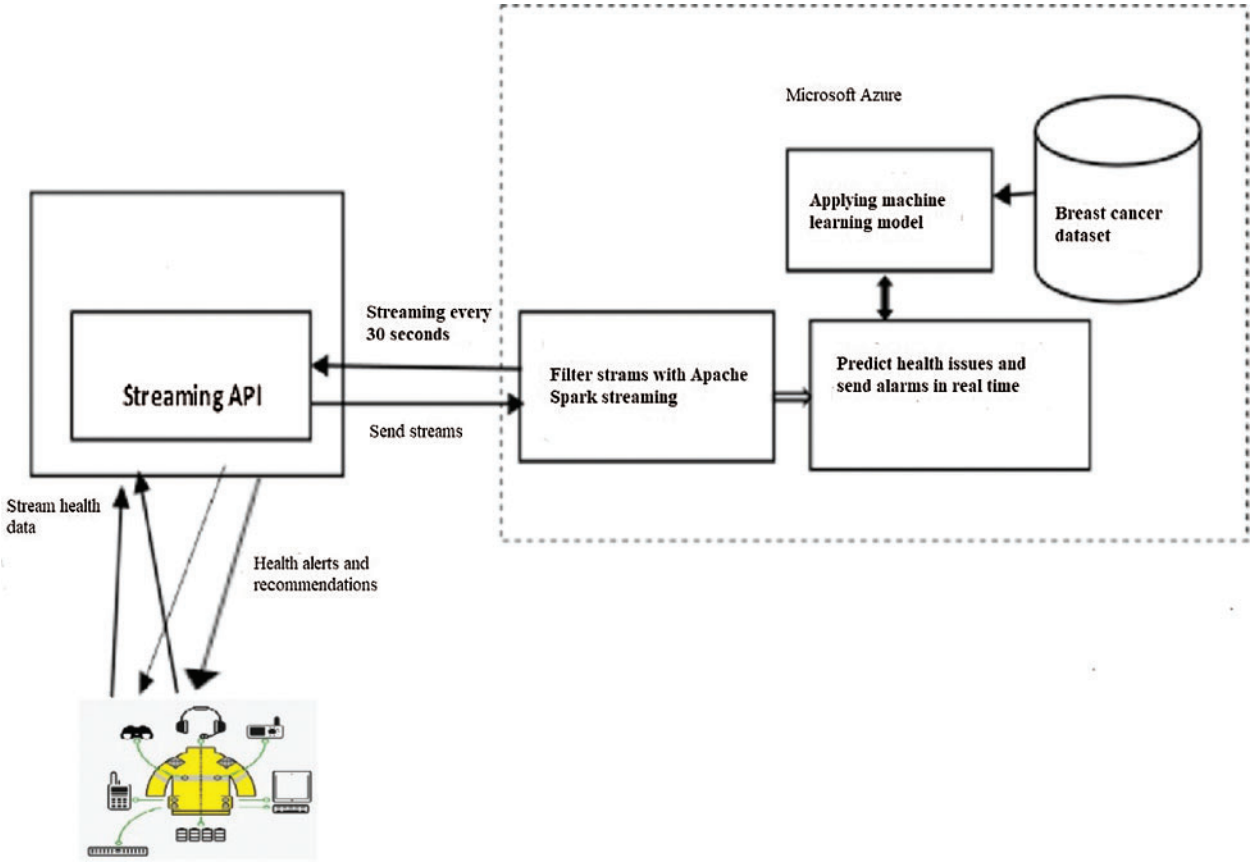

Figure 4: Healthcare recommender system block diagram

3.4 On-Demand Technique

The proposed system presents an on-demand technique to receive data from healthcare wearables and only sends the data to servers on demand based on the threshold value of a patient's medical condition (Fig. 5). Thus, this system provides an effective solution to save the needed records only when abnormal readings are detected by sensors.

In the proposed system, Apache Spark is used to stream medical data and Cosmos DB to serve as a NoSQL database; that is, Apache Spark obtains data from Cosmos DB in real time every 30 s as a default interval time, which can be changed by a user based on healthcare issues. In Spark streaming, data from wearable devices are received to analyze abnormal healthcare issues based on a user's thresholds. Then, alerts are sent to Cosmos DB, which is used to notify doctors

about emergencies. Spark streaming is applied to analyze health data continuously and use data attributes for health risk prediction. The output from the data is presented using Microsoft Power Bi to present health issues as visualizations that can help healthcare providers act as promptly as possible.

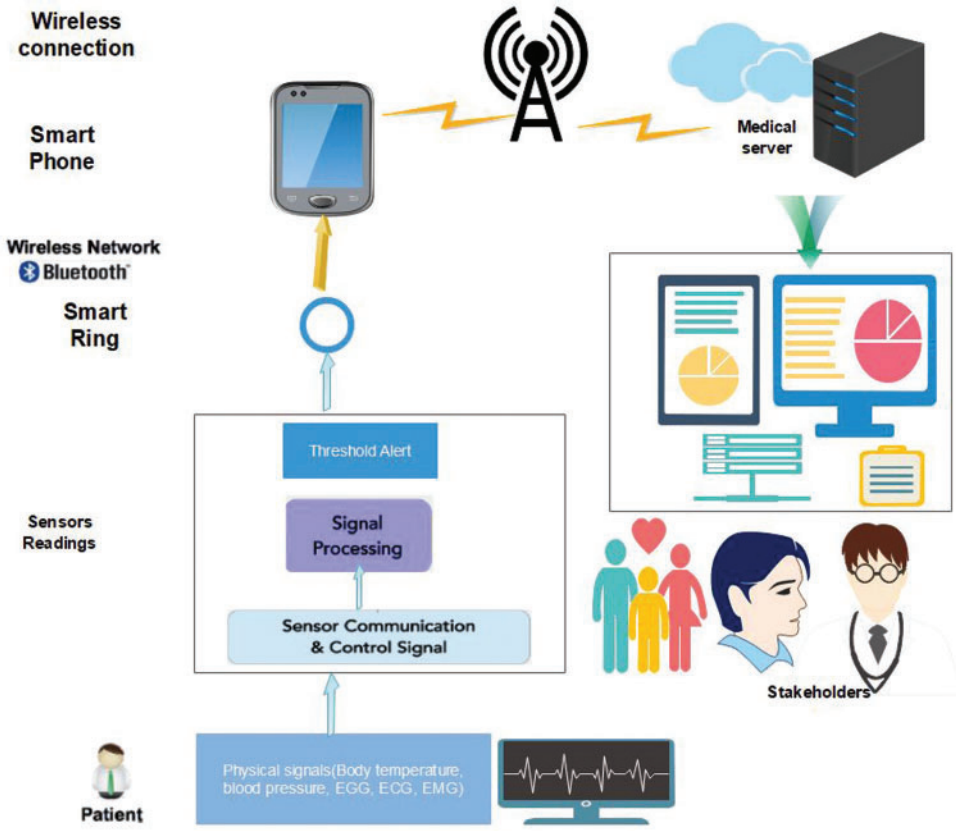

Figure 5: Healthcare system architecture

Spark streaming components, such as MLib, help perform predictive analytics on healthcare data by using a machine learning algorithm [27]. They also contribute to real-time analytics on data generated by wearable health devices. They generate data such as weight, BP, respiratory rate ECG, and blood glucose levels. Analysis can be performed on these data by using k-clustering algorithms. It will intimate any critical health condition before it could happen. Apache Spark's RDD-based computations are extremely fast in processing a large amount of data. Real-time streaming data from social networking sites can be processed effectively. Mila–Spark's built-in library supports machine learning, which is essential for designing health recommender systems. The prediction and recommendation components are built using a machine learning algorithm [28].

The proposed system (Fig. 6) is based on input data to make decisions and predictions. Input data are part of patients' data such as medical reports as measurements and streaming data. They are cleaned by removing any unneeded data.

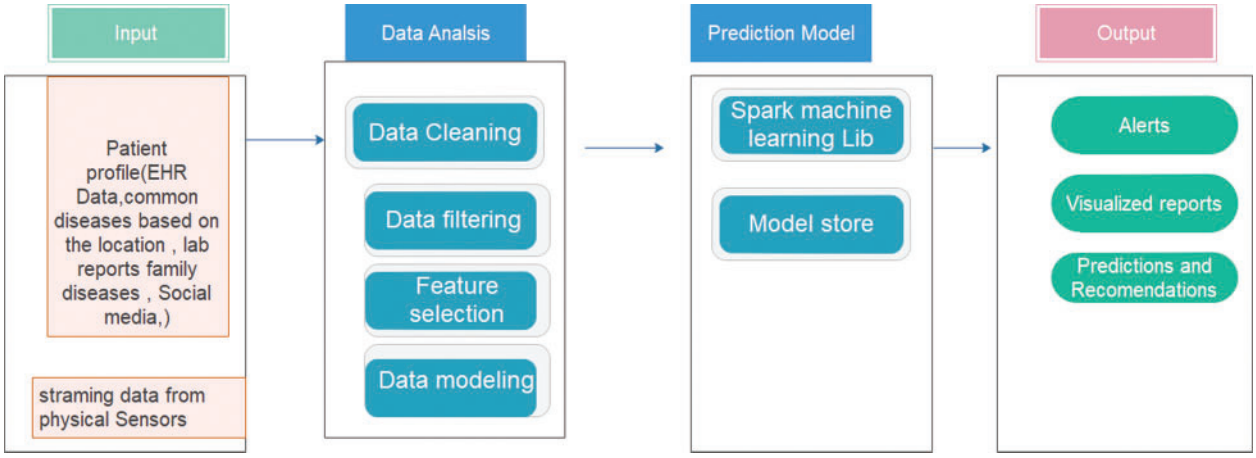

Figure 6: Spark data interaction structure

Spark Streaming can handle streaming data from different sources in real time. Data streams are sent as small batches called DStreams to be processed in the Spark engine. Apache provides MLib that contains various algorithms, such as clustering, classification, and collaborative filtering, to filter attributes in data [29].

The proposed system has an architecture (Fig. 7) that deals with streaming data applications for the healthcare system. Input data can be obtained from different input sources such as Kafka and Flume. This cloud-based system employs Microsoft Azure Databricks to focus on streaming medical data and use the power of Spark streaming for the analysis of streaming data in the cloud. Then, streaming data are made into microbatches so that they can be processed through the Spark core. After data are processed, they are sent to Cosmos DB connected to Microsoft Power BI to present data on dashboards and assist in healthcare decision-making.

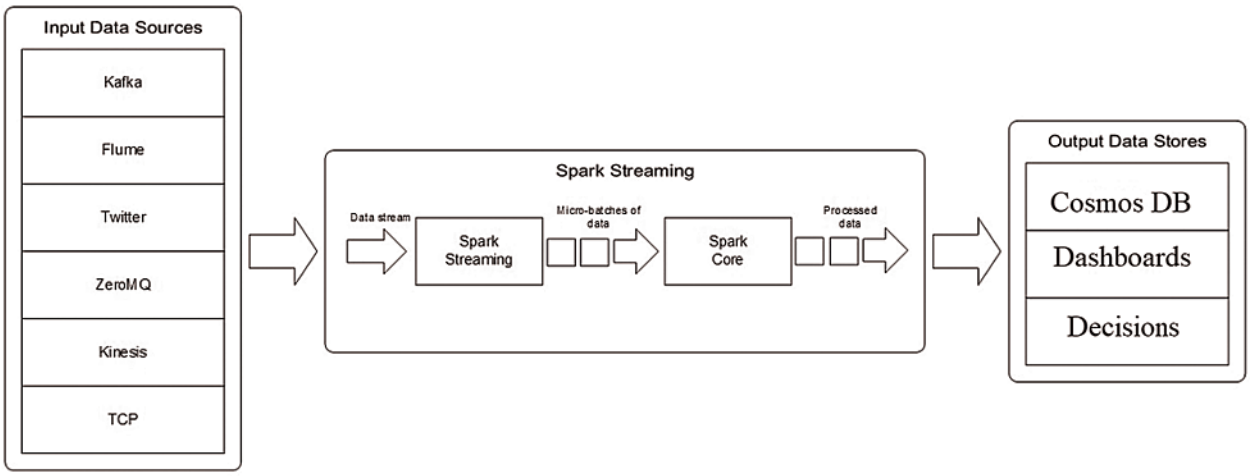

Figure 7: Streaming data module

The proposed system has five layers (Fig. 8). In the first layer, medical data detected by sensors are collected as event sources and sent to the second layer through smartphones. Users can access private medical data through smartphones, tablets, laptops, and computers. In the second layer, data are transmitted to different paths on demand. The first path is connected to streaming data in the third layer, which sends only the JSON data about measurements by medical sensors from smartphones. The second path is connected to a slow track that collects all other medical data, such as medical images (e.g., X-ray and CRT), medical reports, prescriptions, laboratory tests, historical medical data, family historical data, and environmental parameters (e.g., location), and common diseases, from healthcare providers. In the fast-track third layer, recommendations and alerts are provided in real time. In the slow-track third layer, data are batched into Cosmos DB to be analyzed and shown in the last layer. In the last layer, or the decision layer, healthcare providers are given indicators about the users' health status.

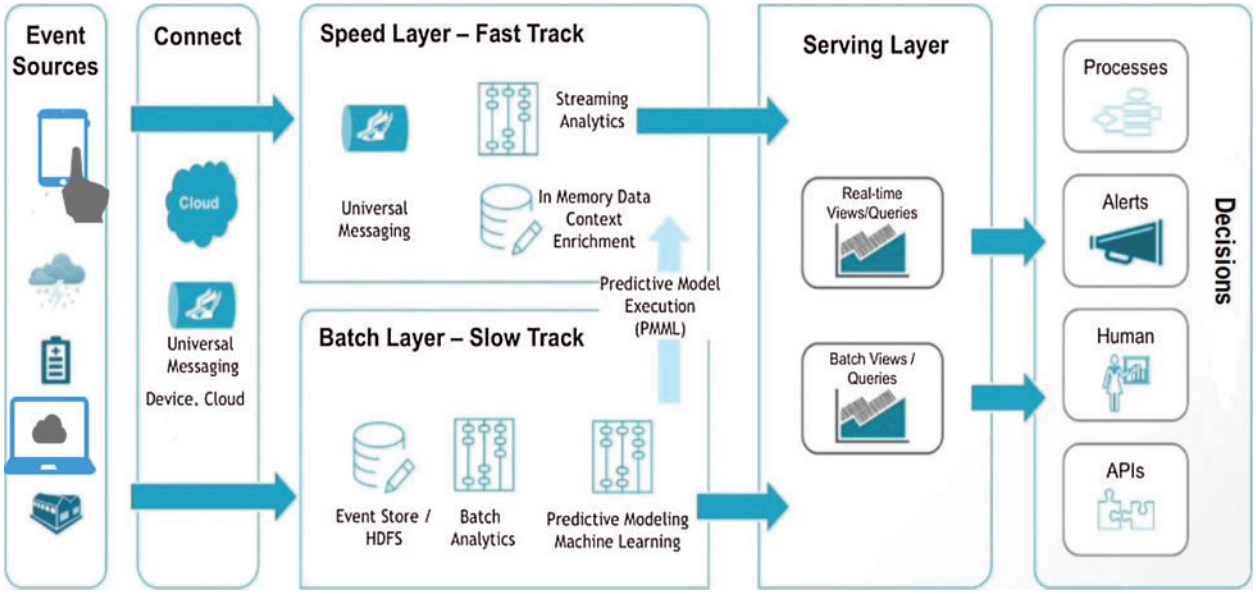

Figure 8: Proposed powerful streaming architecture for medical applications

The proposed system sends streaming data through an event hub, which is a cloud streaming platform on Microsoft Azure. After the analysis process on the Azure Databricks, the hub transmits data to give recommendations and present the anomalous measurements from wearable devices. Then, Cosmos DB receives the data. The Azure Data Factory is used to integrate the data from the Spark streaming data and the data from the batched data for machine learning classification [30]. Classification results are categorized as safe or unsafe and sent to Cosmos DB. They can be monitored remotely using visual studio application insights. The resulting data from machine learning are sent through Data Factory to Microsoft Azure Cosmos DB. They are presented using Microsoft Power Bi to give healthcare providers indicators about the health status of patients (Fig. 9).

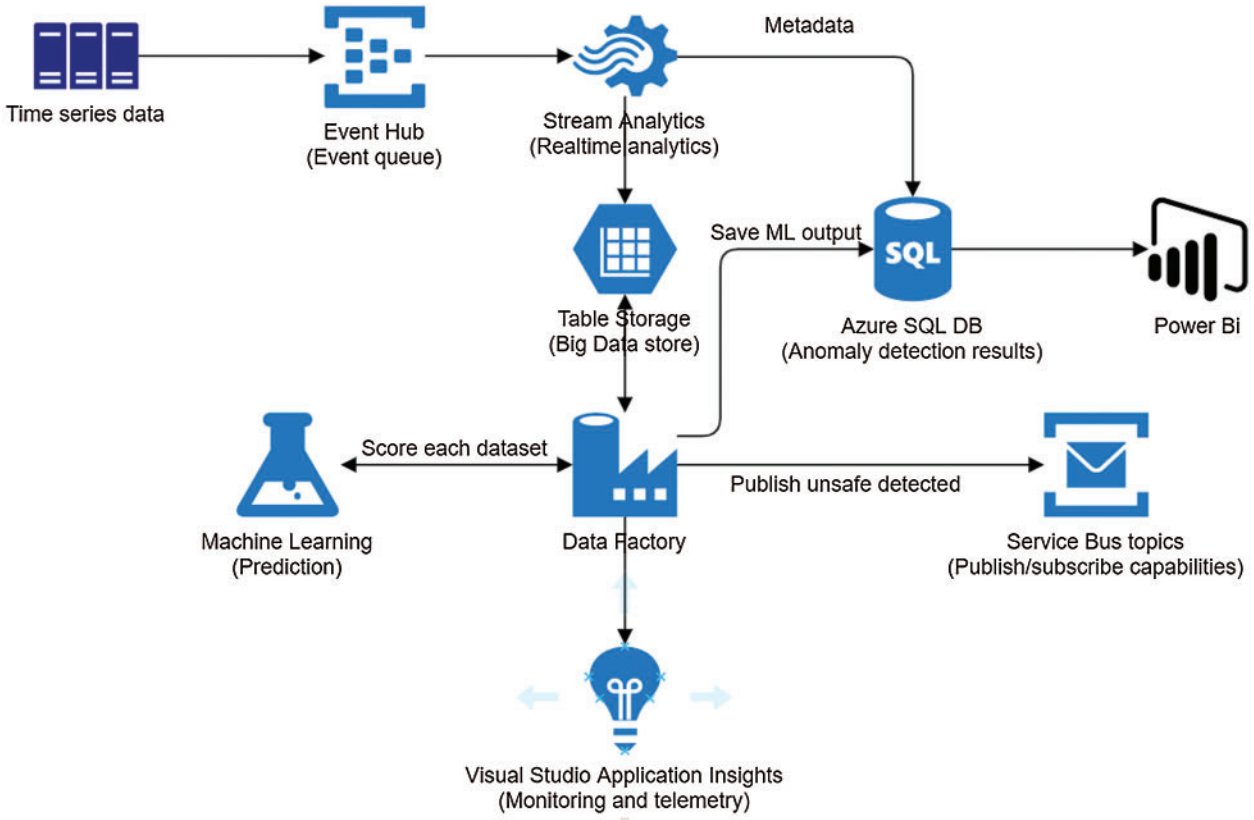

Figure 9: Proposed machine learning algorithm on Microsoft Azure

4 Results

The proposed system used Apache Spark for medical streaming data analysis. It can handle real-time query processing with Spark SQL and DataFrames and real-time streaming. The proposed system has been implemented and tested using Microsoft Azure. It has also been trained in the cloud platform on Microsoft Azure machine learning studio to use the efficiency of cloud platform capabilities. The prediction system is utilized to predict malignant or benign breast cancer from the breast cancer dataset. Multiclass neural networks are employed to predict the type of breast cancer based on other parameters. The Breast Cancer Wisconsin dataset is described in detail in another study [31]. Different algorithms are applied to evaluate the model. The Breast Cancer Wisconsin dataset is used to train a model for the prediction of breast cancer. After the dataset is verified, data should be pre-processed and discretized for cleaning. Data are cleaned to remove duplicate and missing data and resolve inconsistencies. In the next step, columns are selected, missing data are cleaned, and data are split. Then, multiclass neural network and multiclass random forest are applied to determine the algorithm most suitable for the prediction process. The results are evaluated to examine the model (Fig. 10).

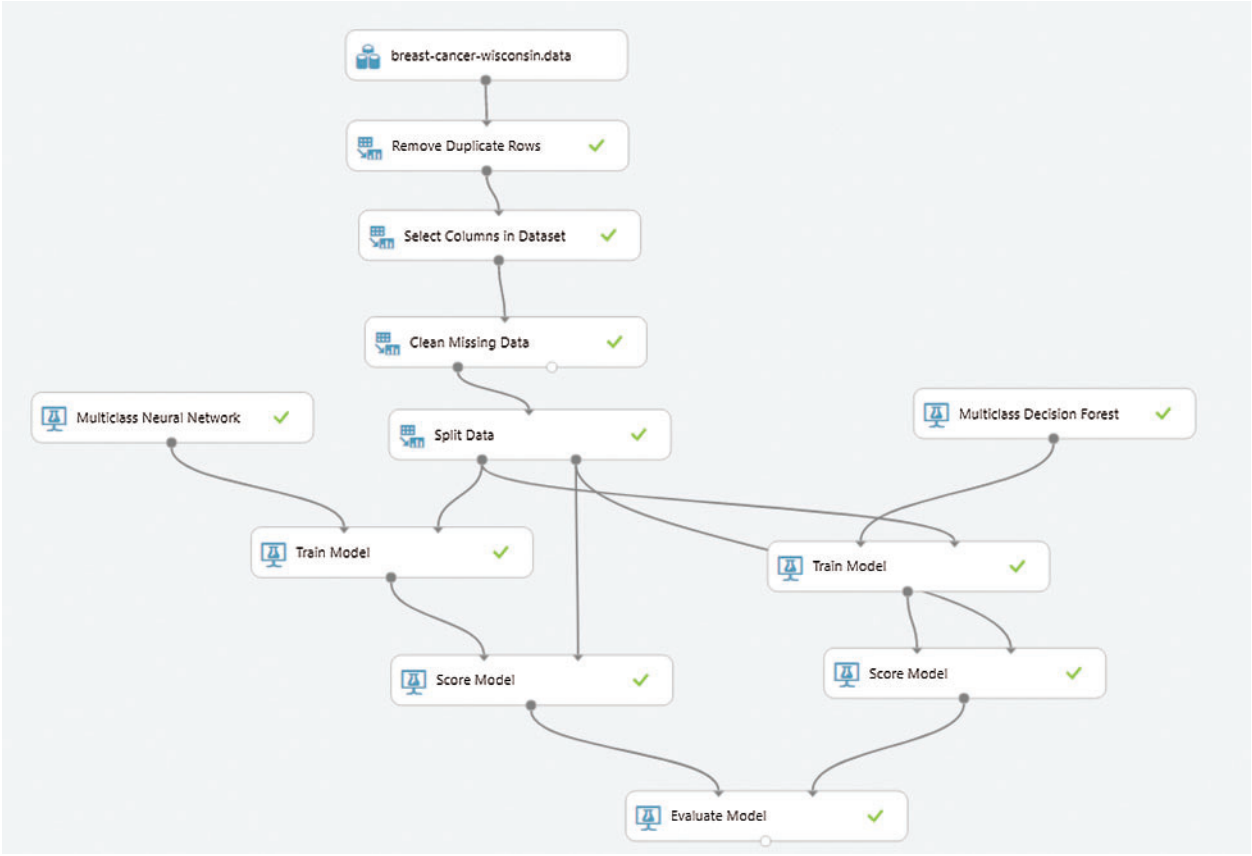

Figure 10: Training module

5 Conclusion

Medical big data analysis should be applied to help healthcare organizations and individuals detect health risks and predict them. This study proposes a system based on Apache Spark to handle streaming data in the cloud. With this system, the analysis features of Spark can be used to predict breast cancer by utilizing machine learning algorithms. The most recent medical big data streaming technologies are presented from different scientific journals, and the proposed healthcare system architecture is described. The application of a multiclass neural network is more efficient than that of a multiclass random forest in terms of predicting breast cancer. This system provides an effective solution to deal with streaming data and determine the processing time to give recommendations. A decision level fusion algorithm is used to define the appropriate threshold for each sensor so that it sends data when readings show abnormal levels. This prediction system helps users identify the potential risk based on the given data; that is, the system can present potential diseases with a probability based on historical data, family data, and users' profile based on EHR and streaming data. Although this system does not provide a solution without the assistance of doctors, it refers specific doctors to be visited and recommends the best time to consult them. This research develops advanced machine learning models to analyze time-series health data collected from wearable devices and predict several vital signs and health disorders. Health-related big data

from wearables and users' profiles can enhance the performance of a prediction model. Thus, this model can be integrated into real clinical systems.

Future studies should focus on using deep learning with streaming data to provide more precise decisions based on patients' scans upon checking and offer an additional basis for making medical decisions. Studies should also develop prediction systems with medical images to increase the probability of healthcare predictions.

Funding Statement: This study was financially supported by the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), the Ministry of Health and Welfare (HI18C1216), and the Soonchunhyang University Research Fund.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] A. S. Abiodun, M. H. Anisi and M. K. Khan, "Cloud-based wireless body area networks: Managing data for better health care," *IEEE Consumer Electronics Magazine*, vol. 8, no. 3, pp. 55–59, 2019.

[2] N. A. Salleh, N. A. A. Ghani, N. Hasannudin and A. A. Shafi, "Heart rate variability derived from wrist photoplethysmography sensor for mental stress assessment," *Journal of Tomography System and Sensor Application*, vol. 2, no. 2, pp. 7–11, 2019.

[3] A. Ismail, S. Abdelrazek and I. M. Elhenawy, "Big data analytics in heart diseases prediction," *Journal of Theoretical and Applied Information Technology*, vol. 98, no. 11, pp. 15–19, 2020.

[4] Z. Zhang, "Photoplethysmography-based heart rate monitoring in physical activities via joint sparse spectrum reconstruction," *IEEE Transactions on Biomedical Engineering*, vol. 8, no. 5, pp. 1902–1910, 2015.

[5] A. Ismail, S. Abdlerazek and I. El-Henawy, "Development of smart healthcare system based on speech recognition using support vector machine and dynamic time warping," *Sustainability*, vol. 12, no. 6, pp. 2403, 2020.

[6] G. Manogaran and D. Lopez, "Health data analytics using scalable logistic regression with stochastic gradient descent," *International Journal of Advanced Intelligence Paradigms*, vol. 98, no. 6, pp. 118–132, 2018.

[7] L. Nair, S. Shetty and S. D. Shetty, "Applying spark-based machine learning model on streaming big data for health status prediction," *Computers & Electrical Engineering*, vol. 65, pp. 393–399, 2019.

[8] S. Gallego, B. Krawczyk, S. Garcia, M. Woezniak and F. Herrera, "A survey on data preprocessing for data stream mining: Current status and future directions," *Neurocomputing*, vol. 239, no. 1, pp. 39–57, 2017.

[9] S. P. Singh, A. Nayyar, R. Kumar and A. Sharma, "Fog computing: From architecture to edge computing and big data processing," *Journal of Supercomputing*, vol. 7, no. 5, pp. 2070–2105, 2019.

[10] W. N. Price, I. G. Nicholson and C. Glenn, "Privacy in the age of medical big data," *Nature Medicine*, vol. 98, no. 3, pp. 37–43, 2019.

[11] A. Yassine, S. Singh, M. S. Hossain and G. Muhammad, "IoT big data analytics for smart homes with fog and cloud computing," *Future Generation Computer Systems*, vol. 108, no. 9, pp. 563–573, 2019.

[12] S. R. Kumar, N. Gayathri, S. Muthuramalingam, B. Balamurugan, C. Ramesh *et al.,* "Medical big data mining and processing in e-healthcare," *Internet of Things in Biomedical Engineering in Academic Press*, vol. 102, no. 66, pp. 323–339, 2019.

[13] N. Pongsakornsathien, A. Gardi, Y. Lim, R. Sabatini and T. Kistan, "Performance characterization of wearable cardiac monitoring devices for aerospace applications," *The Fifth International Workshop on Metrology for Aerospace, IEEE*, vol. 44, no. 9, pp. 76–81, 2019.

[14] A. Ismail, L. Osman, M. Elhoseny and M. El-Henawy, "Quantified self-using IoT wearable devices," in *Conf. on Advanced Intelligent Systems and Informatics*, Cairo, Egypt, vol. 101, pp. 820–831, 2017.
[15] J. M. Vaquero, A. H. Encinas, A. Q. Dios, J. J. Bullon and A. M. Nova, "Review on wearables to monitor foot temperature in diabetic patients," *Sensors*, vol. 80, no. 5, pp. 112–117, 2019.
[16] H. Chen, S. Bao, C. Lu, I. Wang and J. Ma, "Design of an integrated wearable multi-sensor platform based on flexible materials for neonatal monitoring," *IEEE Access*, vol. 98, no. 88, pp. 14–19, 2020.
[17] J. Klemets, J. Määttälä and I. Hakala, "Integration of an in home monitoring system into home care nurses workflow: A case study," *International Journal of Medical Informatics*, vol. 3, no. 2, pp. 29–36, 2019.
[18] A. Heryana and Suhardi, "Smart personal health care monitoring services design using UML," *IEEE International Conf.*, vol. 44, no. 3, pp. 124–130, 2014.
[19] N. El-Rashidy, S. El-Sappagh, S. M. Islam, H. M. E. Bakry and S. Abdelrazek, "End-to-end deep learning framework for coronavirus (covid-19) detection and monitoring," *Electronics*, vol. 9, no. 9, pp. 1439, 2020.
[20] L. R. Nair and S. D. Shetty, "Applying spark-based machine learning model on streaming big data for health status prediction," *Computers & Electrical Engineering*, vol. 65, no. 1, pp. 393–399, 2018.
[21] N. El-Rashidy, S. El-Sappagh, T. Abuhmed, S. Abdelrazek and H. M. El-Bakry, "Intensive care unit mortality prediction: An improved patient-specific stacking ensemble model," *IEEE Access*, vol. 8, no. 1, pp. 133541–133564, 2020.
[22] H. Elhoseny, M. Elhoseny, S. Abdelrazek, A. M. Riad and A. E. Hassanien, "Ubiquitous smart learning system for smart cities," in *Conf. ICICIS*, Cairo, Egypt, pp. 329–334, 2017.
[23] A. Ismail, S. M. Abdelrazek and I. M. Elhenawy, "Development of smart healthcare system based on speech recognition using support vector machine and dynamic time warping," *Sustainability*, vol. 12, no. 6, pp. 2403–2418, 2020.
[24] A. Ismail, A. Shehab and I. M. El-Henawy, "Healthcare analysis in smart big data analytics: Reviews, challenges, and recommendations," in *Security in Smart Cities: Models, Applications, and Challenges*, Cairo, Egypt: Springer, pp. 27–45, 2019.
[25] V. Jeyabalaraja, P. Abirami, K. Janapriya and N. Raghavi, "Analysis of disease from multiple healthcare data using Hl7 message on HDFS," *International Journal of Computer & Mathematical Sciences IJCMS*, vol. 98, no. 6, pp. 12–18, 2018.
[26] R. Logesh, V. Subramaniyaswamy, V. Vijayakumar and X. Li, "Efficient user profiling based intelligent travel recommender system for individual and group of users," *Mobile Networks and Applications*, vol. 108, no. 9, pp. 1018–1033, 2019.
[27] J. Archenaa and M. Anita, "Health recommender system using big data analytics," *Journal of Management Science and Business Intelligence*, vol. 86, no. 5, pp. 17–24, 2017.
[28] A. I. Ebada, S. Abdelrazek and I. Elhenawy, "Applying cloud-based machine learning on biosensors streaming data for health status prediction," in *Conf. IISA50023*, Piraeus, Greece, pp. 1–8, 2020.
[29] H. Elhoseny, M. Elhoseny, S. Abdelrazek and A. M. Riad, "Evaluating learners' progress in smart learning environment," in *Int. Conf. on Advanced Intelligent Systems and Informatics*, Cairo, Egypt, Cham, Springer, pp. 734–744, 2017.
[30] S. M. Abdelrazek, H. M. El-Bakry, W. F. A. Elwahed and N. Mastorakis, "Collaborative virtual environment model for medical e-learning," in *Proc. ACACOS'10*, pp. 191–195, 2010.
[31] UCI, "Breast Cancer Wisconsin (Original) Data Set," 2021. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+%28original%29.