An intelligent healthcare monitoring framework using wearable sensors and social networking data

Farman Ali, Shaker El-Sappagh, S.M. Riazul Islam, Amjad Ali, Muhammad Attique, Muhammad Imran, Kyung-Sup Kwak

PII: S0167-739X(19)31605-X
DOI: https://doi.org/10.1016/j.future.2020.07.047
Reference: FUTURE 5769

To appear in: *Future Generation Computer Systems*

Received date : 17 June 2019
Revised date : 24 June 2020
Accepted date : 21 July 2020

An Intelligent Healthcare Monitoring Framework Using Wearable Sensors and Social Networking Data

Farman Ali[1, †], Shaker El-Sappagh[2, 3, †], S.M. Riazul Islam[4], Amjad Ali[5*], Muhammad Attique[1], Muhammad Imran[6], Kyung-Sup Kwak[7*]
[1]Department of Software, Sejong University, South Korea
[2]Centro Singular de Investigación en Tecnoloxías Intelixentes, Universidade de Santiago de Compostela, Santiago, Spain
[3]Department of Information Systems, Benha University, Banha, Egypt
[4]Department of Computer Science and Engineering, Sejong University, Seoul, South Korea
[5]Department of Computer Science, COMSATS University Islamabad, Lahore Campus, Lahore, Pakistan
[6]College of Applied Computer Science, King Saud University, Riyadh, Saudi Arabia
[7]Department of Information and Communication Engineering, Inha University, Incheon, South Korea
E-mail: farmankanju@sejong.ac.kr; shaker_elsapagh@yahoo.com, riaz@sejong.ac.kr, attique@sejong.ac.kr, dr.m.imran@ieee.org
*Corresponding author: K.S. Kwak (kskwak@inha.ac.kr); Amjad Ali (amjad.ali@cuilahore.edu.pk)
†These authors contributed equally to this work and co-first authors.

Abstract: Wearable sensors and social networking platforms play a key role in providing a new method to collect patient data for efficient healthcare monitoring. However, continuous patient monitoring using wearable sensors generates a large amount of healthcare data. In addition, the user-generated healthcare data on social networking sites come in large volumes and are unstructured. The existing healthcare monitoring systems are not efficient at extracting valuable information from sensors and social networking data, and they have difficulty analyzing it effectively. On top of that, the traditional machine learning approaches are not enough to process healthcare big data for abnormality prediction. Therefore, a novel healthcare monitoring framework based on the cloud environment and a big data analytics engine is proposed to precisely store and analyze healthcare data, and to improve the classification accuracy. The proposed big data analytics engine is based on data mining techniques, ontologies, and bidirectional long short-term memory (Bi-LSTM). Data mining techniques efficiently preprocess the healthcare data and reduce the dimensionality of the data. The proposed ontologies provide semantic knowledge about entities and aspects, and their relations in the domains of diabetes and blood pressure (BP). Bi-LSTM correctly classifies the healthcare data to predict drug side effects and abnormal conditions in patients. Also, the proposed system classifies the patients' health condition using their healthcare data related to diabetes, BP, mental health, and drug reviews. This framework is developed employing the Protégé Web Ontology Language tool with Java. The results show that the proposed model precisely handles heterogeneous data and improves the accuracy of health condition classification and drug side effect predictions.

Keywords: Machine learning; Semantic knowledge; Big data analysis, Healthcare monitoring system; Wearable sensors, Social network analysis.

1. Introduction

Diabetes and abnormal blood pressure (BP) are the two most common and extremely dangerous diseases that affect the functionality of the human body, and increase the risk of cardiovascular diseases. Early detection and classification of diabetes and BP into their specific category is therefore essential to treat the patient efficiently. In addition, accurate monitoring of the diabetes and BP can provide an opportunity to secure the patient from cardiovascular diseases.

Due to the advancements in information technology, all the devices in the healthcare industry have been digitalized. These digital devices enable the living better and more comfortable. Therefore, people use various devices, such as smartphones and wearable sensors, in their daily lives. Smartphones can contain sensors that can be used to obtain information about the human body [1]. Wearables can be utilized to collect a huge amount of patient vitals [2]. Both of these devices can be utilized to real-time monitor the patient. To date, numerous healthcare systems have been proposed to monitor the diabetes and BP patients using smartphone and wearable sensors [3–10]. However, these systems are not well-equipped to collect the data in real time. Furthermore, the digital devices generate a huge amount of healthcare data and the existing systems are incapable of storing and processing it efficiently for precise monitoring of patients. Moreover, extracting valuable information from healthcare data and effectively analyzing them has become a new challenge for the existing healthcare monitoring systems. The generated data from these devices are unstructured and are therefore difficult to handle for chronic patient monitoring. In addition, the amount of data is exponentially increasing, which requires a huge storage space. Therefore, a smart methodology and a cloud-based healthcare architecture are required to accurately monitor diabetes and BP patients, store their healthcare data, and perform predictive analysis on the structured and unstructured data.

Recently, the use of social networking in the healthcare industry has been rapidly increasing. The social network data can also be utilized to identify various factors such as emotional status and accrued stress, which might be translated into the status of a patient health. People with diabetes and abnormal BP share their emotions and experiences with others on social networking sites. They share valuable information and motivate each other to fight against diabetes and high/low BP. In addition, diabetes patients publish their opinions about specific drugs. A new patient sees the opinions of others and responds to them about the same drugs. Therefore, the healthcare monitoring systems for diabetes and abnormal BP need social networking data in order to identify emotional disturbances in patients by using their posts, and to monitor drug side effects by using drug reviews. However, the information on social networking sites about patient emotions and drug experiences are unstructured and unexpected, and it would be a challenging task for healthcare monitoring systems to extract the information and analyze it in order to monitor the patient's mental health and to predict drug side effects. Therefore, there is a need of smart approaches that can automatically extract the most relevant features, and reduce the dimensionality of the datasets for better accuracy of healthcare monitoring system.

In recent years, machine learning (ML) techniques, such as the decision tree, the support vector machine (SVM), k-nearest neighbors (KNN), fuzzy logic, and multi-layer perceptron (MLP), are used to monitor diabetes and BP patients and provide competent treatments [11–14]. However, continuous patient monitoring generates a large amount of healthcare data, such as sensor readings, patient profiles, medical records, lab tests, and physician notes. Both healthcare and social networking data have greatly increased in a few years, which is called big data (both unstructured and structured). The traditional approaches and ML techniques may not handle these data very well for the extraction of meaningful information and for abnormality prediction. In addition, these data may not help the healthcare industry until they are processed intelligently in real time. This necessitates a big data cloud platform and an advanced deep learning approach, such as long short-term memory (LSTM).

New advancements in the healthcare monitoring systems are still a big challenge because they need huge multidisciplinary steps. With the new technologies in healthcare sectors and the changes in society, traditional systems are not efficient enough for these new conditions. They need a new framework based on new methods to solve the problems. However, traditional techniques can be used together with a new system.

In this paper, we propose an advanced healthcare monitoring framework based on ontologies and on bidirectional long short-term memory (Bi-LSTM) to precisely analyze healthcare big data in order to improve classification accuracy. The proposed framework integrates different sources of information for efficient healthcare monitoring of chronic patients. It was applied to the classification of patients using their healthcare data related to diabetes, BP, mental health, and drug reviews. The results prove that the proposed model correctly handles heterogeneous data and improves the accuracy of patient health condition classification. The main contributions of our research work are as follows.

- A novel framework is introduced that extracts large amounts of the most useful healthcare data from various sources, such as smartphones, wearable sensors, medical records, and social networks. In addition, a big data cloud repository is utilized to store the extracted data, and MapReduce is applied to intelligently handle and process structured and unstructured data.
- A big data analytics engine is proposed for the analysis of real-world big data. It is used to accurately handle healthcare data containing inconsistencies and that have missing values, noise, different formats, a large size, and high dimensionality. In addition, it is utilized to improve the quality of data processing and to save time. The proposed big data analytics engine uses artificial intelligence (AI) approaches to extract useful features from big data that eventually reduce the dimensionality of data.
- A neural network–based word embedding model called Word2vec is used to represent healthcare textual data with semantic meaning. In addition, specific domain ontologies are integrated with the Word2vec model. These ontologies provide additional information for a neural network model that understands the semantic meaning of unusual words. A novel semantic knowledge with the Bi-LSTM model is used to precisely classify the unstructured and structured healthcare data.
- Several experiments are conducted using principal component analysis (PCA) and information gain (IG) with the ontology and LSTM-based models, and the results are compared with the respective reference models. This comparison helps us to determine the advantages and limitations of the applied approaches and classification models.

The rest of this paper is structured as follows. Section 2 presents discussions of healthcare monitoring using wearable sensors and ML approaches, healthcare monitoring systems based on social networking data and ML approaches, and healthcare big data and ML approaches. Section 3 presents the proposed framework and describes its various modules. Section 4 provides the experimental results. Finally, Section 5 concludes our work.

2. Related Work

Machine learning techniques and big data play a key role in the healthcare monitoring systems of chronic patients. Due to the rapid increase in the usage of wearable sensors and social networking data in the healthcare domain, this section looks at healthcare monitoring of diabetes and BP patients based on wearable sensors and ML approaches, at healthcare monitoring systems based on social networking data and ML approaches, and at healthcare big data and ML approaches.

2.1 ML approaches and healthcare monitoring of diabetes and BP patients using wearable sensors

There have been various studies into wearable sensor–based physiological information extraction and the healthcare monitoring of chronically ill patients. These studies introduced various methods to analyze wearable sensor data [15–17].

However, wearable sensors-based collected data are in large volume and unstructured. These methods do not consider the advanced data analysis approaches to discover useful information in the sensors data. Thus, they may not monitor the patient body effectively. Wearable sensors and an ML-based personalized healthcare monitoring system was presented for diabetes patient monitoring [5]. This system utilizes Bluetooth Low Energy–based sensors to collect data. The authors then applied multi-layer perceptron and LSTM to classify the type of diabetes and to predict the glucose level of the input user. However, the diabetes dataset used in the existing system was limited to Pima Indian woman. Therefore, it may not be possible to use the dataset with robust prediction model for different purposes. A framework based on ML methods and a Hadoop environment was proposed for diabetes prediction [18]. In that work, the authors proposed an IG algorithm for feature extraction. In addition, naive Bayes, the decision tree, and random forest were utilized to predict diabetes. However, the accuracy of these classifiers largely decreased when the amount of data increased. This is because they did not use dimensionality reduction. A 5G smart diabetes system based on wearable sensors, big data, and ML was presented for personalized diabetes diagnosis [12]. The system has five main goals (smartness, personalization, comfortability, effectiveness, and sustainability), which help patients detect diabetes early, and it provides personalized treatment solutions. This model is costly in terms of implementation and lacks a personalized data analysis method. A wearable-technologies and Internet of Things (IoTs)-based recommendation system was proposed for proactive healthcare monitoring [19]. The system solved two main problems in health monitoring: the identification of factors that must be monitored, and the identification of wearable technology that must be used for the measurement of the factors. However, this system recommends tools or technologies using manual approaches, which is time-consuming and may not suit with the ongoing trend of healthcare technology. An LSTM-based model was presented to effectively mine sensor data and lab tests [20]. The authors trained LSTM and MLP techniques by utilizing different attributes of the intensive care unit (ICU) patient, including body temperature, systolic and diastolic blood pressure, blood glucose, and heart rate. The authors then verified the utility of these classifiers and compared their performance. However, this LSTM-based system failed to directly handle irregular sampling and missing values. It needs indicator variables to differentiate actual from missing values. A framework for big data in Mauritius was proposed for diabetes [21]. That work presented the patient's current diabetes status and solutions for diabetes management. In addition, the authors also discussed the challenges of big data in the healthcare domain. A bi-directional LSTM was proposed to predict the future level of blood glucose [22]. In that system, the authors compared the results from simple LSTM with Bi-LSTM using 26 datasets from 20 real patients. However, the system lacks features that decrease the performance of diabetes prediction. In addition, there is no alarm mechanism to predict the up-coming blood-glucose level. The authors presented a review of published work from 2001 to 2017 that summarized the research topics related to wearable sensor technology [23]. In the review paper, the authors discussed frameworks that covered various topics, including data gathering, data processing, and system responses. In addition, the authors looked at the gap between wearable devices and human factors. A knowledge discovery approach was proposed for healthcare assistance [14]. The system simplified the analysis of big data in the cloud environment, and the authors illustrated a new method that determines the classifications to predict unusual conditions in blood pressure patients. However, the system utilized traditional approaches and may solve only specific cases. In addition, the information gathered from sensors is so massive. It is impossible to operate them for knowledge-discovery without cloud environments and big data framework.

2.2 Healthcare monitoring systems based on social networking data and ML approaches

Recently, researchers have been utilizing social networking data and drug information in the field of healthcare monitoring systems, which are extensively discussed in this section [12,24,25]. An emotional healthcare system was presented to detect psychological disturbances in patients [26]. The authors utilized patient messages published on social networking platforms and identified depression and stress levels. They applied the convolution neural network (CNN), the recurrent neural network (RNN), and Bi-LSTM to detect stressful and depressive content. In addition, they proposed an ontology-based recommendation system that sends text to the patient based on results from monitoring. However, the continuous monitoring of patients based on their social networking data is a difficult task. This needs effective statistical analysis, data normalization, feature selection, and ML models to correctly predict the patient mood. Deep learning–based sentiment analysis was presented for perinatal depression [27]. These authors utilized WeChat friends'–circle data for LSTM training, and used emoticons as feature extraction to monitor perinatal depression in patients. This system abbreviated the screening time and decreased the costs of doctor-patient communication. This method does not consider the sentiment analysis of textual data which is potential to improve the performance of perinatal depression monitoring. An artificial intelligence approach was proposed to analyze patient posts to healthcare social networking platforms and to identify critical problems in patients [28]. The system applied the preprocessing methods of text and ML classifiers to automatically analyze patient posts and inform their doctors when needed. The posts on social networks are always unstructured. It is impossible to precisely handle without using deep learning algorithms such as word embeddings and n-grams. A novel system was proposed to identify diabetes risk based on social networking activities on Twitter [29,30]. The authors presented a new approach to data generation, and introduced ML methods to classify the users' diabetes risk based on their Twitter activities. Both the systems have utilized insufficient information for diabetes risk prediction, which may generate inaccurate results and thus perhaps misguide the healthcare professionals. A Bi-LSTM and gated recurrent unit was utilized to intelligently process social media comments of a patient undergoing a medical cure [31]. These authors focused on patient remarks about drugs, and discovered disease-related medical concepts from them. This system considered different types of text to discover disease (e.g., a depressive disorder was discovered in patients using text like "woke up too early"). Feature-based sentiment analysis of drug reviews was presented to detect adverse drugs reactions [32]. This system performs different tasks on drug reviews that identify the effectiveness and side effects of the drugs.

However, deep learning approaches with more sophisticated features can improve the achieved results. A sentiment feature–based system was proposed to detect adverse drug reaction posts on social networks [33]. The authors utilized a large amount of text and conducted experiments in order to develop a real approach to identifying posts related to adverse drug reactions. In this system, there is no data analysis method to efficiently process the data. Also, the system lacks semantic similarity approaches to precisely confirm the identification of drug-related posts.

An ontology-based, feature-level, sentiment analysis system was proposed in the domain of diabetes treatment [34]. The authors used the ontology to provide semantic relationships between features that successfully performed the classification task. The system has two main limitations. First, the general semantic lexicon cannot capture the meanings of diabetes-related text. Second, it requires fuzzy ontology-based semantic knowledge to identify the aspects. An ontology mapping framework was presented to discover the relationships between ontology features and required features in text [35]. In this system, the authors used various ontologies from bio portals, and proposed a novel approach to extracting features from ontologies for semantic word embedding. A deep learning method was proposed to present medical knowledge embedding [36]. In this work, biomedical ontologies related to epilepsy were employed to automatically find medical concepts from clinical text. However, clinical text is very noisy and the generated triples for finding medical concepts may not be correct. Therefore, the integration of semantic attributes is required to address the above issues. An ontology was proposed to boost the deep learning performance for disease name extraction from Twitter text [37]. The authors presented the architecture of an ontology-boosted neural network to automatically extract disease names from Tweets. The system used different types of features for sentiment classification. However, they failed to achieve high accuracy for standard disease names. A System called BO-LSTM was presented to detect biomedical information in text [38]. This approach overcomes the problem of missing semantic entities in word embedding. The authors utilized a domain ontology with word embedding in order to enhance the extraction of biomedical relations using deep-learning LSTM.

2.3 Healthcare Big data and Deep learning approaches

The analysis and representation of healthcare big data with low dimensions is another main issue in healthcare monitoring systems. Recently, various systems have been presented to analyze big data [39–44], to decrease the dimensionality of big data [45], and to manage large amounts of heterogeneous data using ML [46,47]. Jindal et al. [48], collected patient data using remote healthcare applications, and then applied a fuzzy-rule classifier to the data for efficient decision making. They discussed various issues in information clustering, information retrieval, and parallel processing of big data in the cloud environment. However, their rule-based system is limited to specific cases. Therefore, instead of using the traditional fuzzy rule-based system, there is a need of the deep learning model that can learn from the existing data and provide the results for precise decision making. A recommendation system based on numerical reputation and context investigation were proposed to guarantee the veracity of big data [49]. In addition, ML methods and analytic algorithms were used for handling the volume of data and managing the velocity of data, respectively. A disease diagnosis healthcare framework based on the cloud-centric IoT was proposed in order to predict diseases in students [50]. In this system, student healthcare data were generated using medical sensors and a University of California, Irvine (UCI) dataset, and ML algorithms were trained to predict diseases in students. In addition, the system utilized medical measurements for disease diagnoses. However, same measurement algorithms for six different diseases may generate the wrong results. Semantic knowledge with deep learning method is required to handle the issues of disease diagnoses. A big data analytics system was proposed for an effective health recommendation system [51]. This system handles large amounts of structured and unstructured data extracted from patient social networking activities. In addition, ML algorithms were applied to recommend centric treatments for patients. However, this system is possibly incapable to filter the unstructured data properly.

Dwivedi et al. [52], discussed the effects of the IoT on the results of big data sentiment analysis. At first, they introduced the notion of big data sentiment analysis, features, and the value of decisions, and then, they described the framework required to handle big data. However, this framework is complex and needs more management to reduce the task on the Hadoop cluster. A random forest and MapReduce were used to present a big data–analytics-based IoT healthcare system [53]. This system considered patients with different diseases for analysis. In addition, an improved dragonfly algorithm was applied to select optimal attributes for better classification. This random forest classifier is computationally slow because of healthcare big data. This issue can be addressed using deep learning approaches. A new IoT-based framework was proposed to precisely collect the sensor data for healthcare applications [54]. The system utilizes Apache HBase and Apache Pig for the collection and storage of sensor data, and it applies a prediction model of MapReduce to predict heart disease. However, it is not enough to transfer sensor data to the cloud for storage and processing. It needs a smart data analytics strategy with secure architecture for precise heart disease prediction.

Most of the aforementioned systems were based on traditional methods, and they handle the problem of big data to some extent. However, there is no clear framework for using big data analytics to process and analyze big data with high accuracy. In addition, these systems are still not good enough to deal with different types of structured and unstructured data. Due to the limitations of medical semantic knowledge for ML classifiers, these systems can incorrectly classify social networking data.

3. The proposed healthcare monitoring framework for diabetes and abnormal BP patients

The proposed framework is presented in Fig. 1. This framework contains different layers, namely, the data source layer, the data collection layer, the data storage layer, the analytics engine layer, and the data presentation layer. The data source layer deals with heterogeneous data. Sensor devices, medical records, and social networking platforms are the main sources of data. The data collection layer is responsible for gathering data from different domains about diabetes and BP patients. This layer collects physiological information from wearable devices, drug and symptom information from smartphones, and patients' and doctors' discussions on social networks through application programming interfaces (APIs). The third layer is the data storage layer. All the monitored data of patients using wearable devices and the collected data from social networks are offloaded to the cloud server through a wireless communications network. The fourth layer is the analytics engine layer, which is the most important layer of the proposed framework. It is divided into two sub-layers: the data computation layer and the data classification layer. The data computation layer has sub-modules, namely, data preprocessing and analysis, feature extraction, word embedding, and dimensionality reduction. Here, ontology-based semantic knowledge, along with soft computing approaches, are employed to process and analyze the data for the extraction of required information. In the data classification layer, Bi-LSTM with ontologies is utilized for the classification of diabetes, BP, mental health, and drug side effects. This layer intelligently analyzes the multidimensional big data about diabetes and BP to get insights for decision making from the data, and provides a personalized diabetes and BP healthcare system for patients. The proposed system uses Hadoop MapReduce with ML to reduce large-sized data about patient treatments. The final layer is the data representation layer, which presents the analysis results to physicians. This layer combines the generated results with the physician's suggested treatment, and then, the personal diabetes and BP healthcare treatment is recommended to the patient. The proposed system helps warn diabetes and BP patients before their health risk reaches a high level. It supports physicians in offering actual treatments to their patients by monitoring health conditions smartly. The whole idea of the proposed system is displayed in Fig. 2. We briefly discuss the different modules of the proposed system in the following subsections.

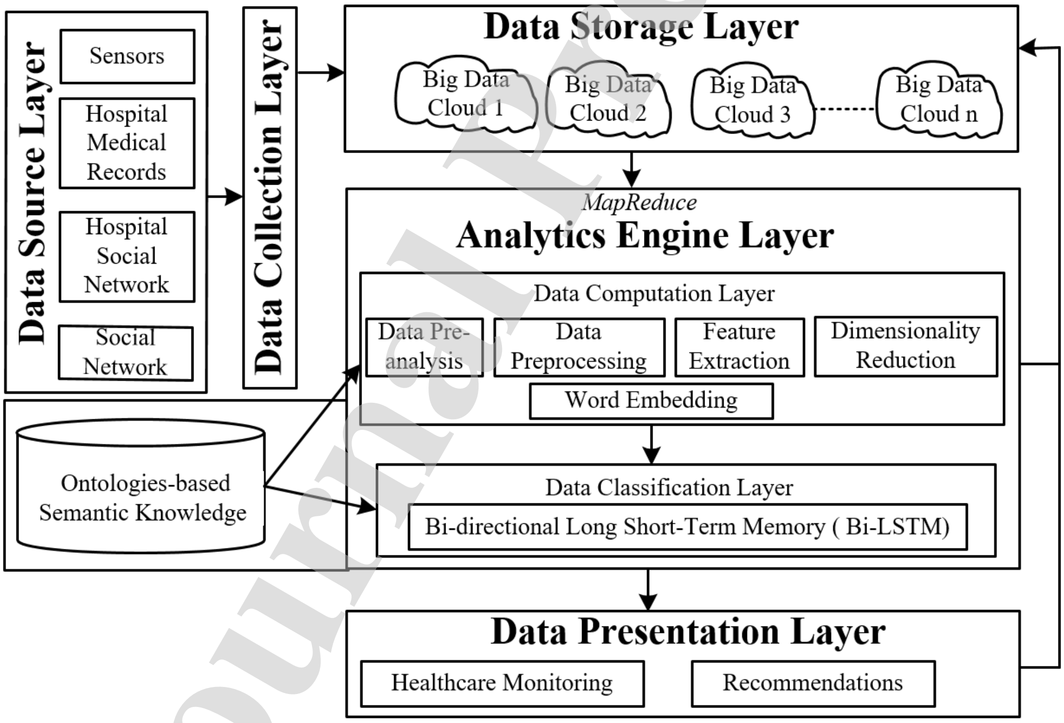

Fig. 1. Different layers in the proposed healthcare monitoring framework. The data source layer provides different sources of data. The data collection layer gathers data from different sources about patients. The data-storage layer is responsible to store the collected data. The analytics engine layer analyzes the healthcare data and predicts the patient health conditions. The data representation layer presents the analysis results to physicians.

3.1 Data collection

This proposed system collects the data from four different sources. In this section, we extensively discuss data collection from all four sources, one by one.

3.1.1 Wearable devices

Smart sensors and wearable devices, such as a blood pressure monitor, a glucometer sensor, the smart watch, pulse oximeters, accelerometers, temperature sensors, and weigh scales, can be utilized to collect the patient's real-time body signals for the monitoring of physiological signs in the patient, such as blood pressure, blood sugar level, heart rate, stress rate, oxygen saturation rate, temperature, and weight. In our system, wearable devices are placed on the patient's body to monitor the bodily functions. In addition, the use of mobile devices has largely increased, and the applications related to health have been well developed in recent years. A mobile device contains sensors that provide the opportunity to collect accurate information regarding different parts of the human body. Therefore, a smartphone is used to collect information on diet, exercise, and other activity information from diabetes and BP patients, along with personal information (age, gender, height, and other information). This information is valuable for healthcare and disease prevention. However, smartphone data are highly unstable and can be damaged easily. Therefore, it is difficult to use smartphone data for accurate healthcare and relevant information extraction. We have utilized data mining techniques and ontology-based semantic knowledge to manage sensor data and extract relevant information for efficient healthcare.

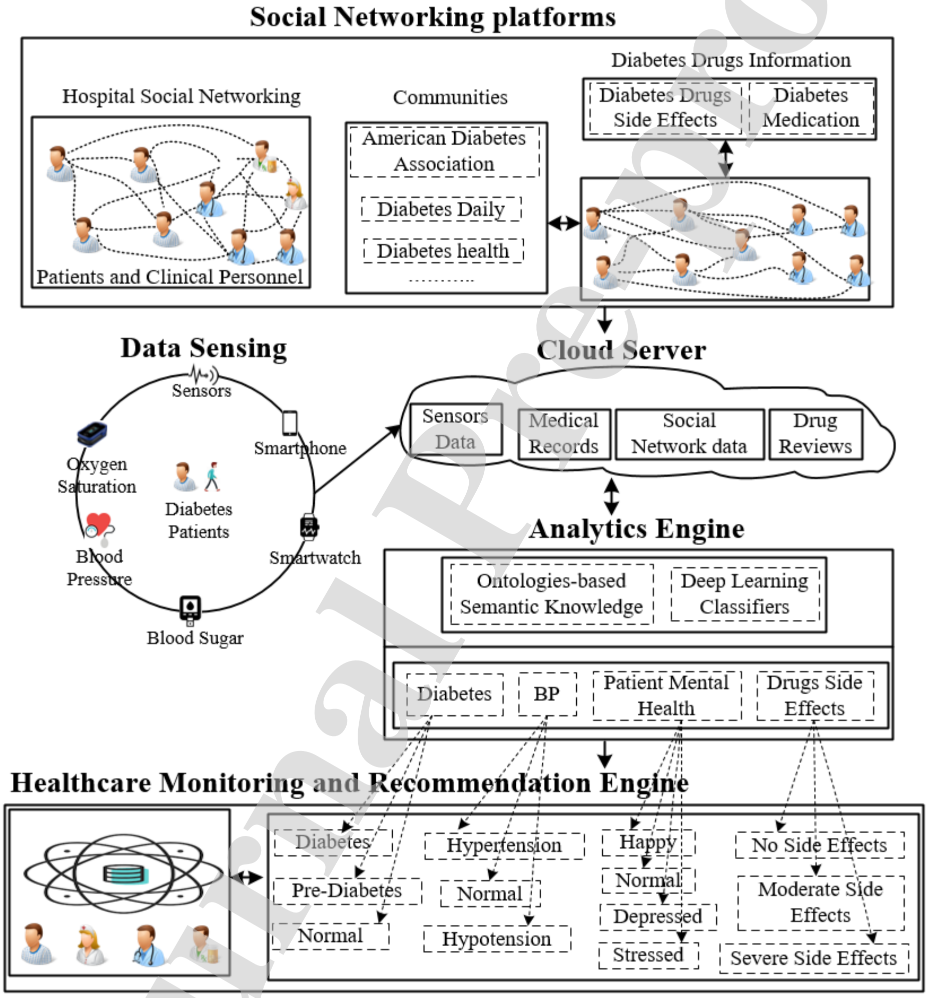

Fig. 2. The workflow of the proposed system. The first step is to collect patient data using wearable sensors and social media platform. The data are then stored in the cloud server. An ontologies-based semantic knowledge is applied to transform the data into a readable form. Next, deep learning models classify the transformed data in order to predict the patient health condition for precise healthcare monitoring.

3.1.2 Medical records

Medical records are the data about treatments undergone by diabetes and BP patients. We collect the medical records of patients, which contain the patient history (treatments, lab tests, and drug intake). These records can be analyzed to extract valuable information that can help provide insights into improving medical guidelines for diabetes and BP diagnosis. However, the volume of medical records is usually large, and each record comprises data with distributed variables and of high dimensionality. In addition, diabetes and BP patients may face other complications, such as kidney and cardiovascular disease, neuropathy, and skin and eye problems. Therefore, the medical records should be analyzed to identify patients affected by the above complications, and to monitor their status with more specific examinations.

3.1.3 Data from social networking platforms and webpages

The proposed system first extracts patient contents from hospital social networking platforms. However, this task needs extra work, and completely depends on social network privacy settings. The APIs of some specific social networks are not publicly available. In this situation, special software, such as *wrappers*, can be used to extract information (e.g., patient posts) [28]. Generally, diabetes and BP patients regularly contact their physicians, but they also need support, knowledge, and skills for personal monitoring of their healthcare condition. In addition, if patients do not get efficient information from their physicians, social media can play an important role in fulfilling their needs. Therefore, social networking platforms like Facebook and Twitter provide opportunities for patients to get enough knowledge regarding diabetes and BP and to connect with people who have similar health problems and experiences. Social networks provide platforms for both patients and physicians to share their knowledge regarding diabetes treatments. We collect social media data, such as drug reviews and emotional posts of patients, to predict their stress and depression levels, to identify the side effects of diabetes drugs in the context of diet and lifestyle, and to improve patient care and knowledge. The procedure used for data collection from Facebook and Twitter is discussed in the following sections.

Facebook: The popularity of social networks has increased in the healthcare domain. Therefore, patients connect with online communities to meet their needs. Various communities on Facebook, such as the American Diabetes Association, diabetes recipes and food hints, Diabetes Daily, and Diabetes Health, provide platforms for patients to share information with each other. Patients can respond to posts and share them with others. We employed the Graph API along with a Java client, RestFB, to extract data from Facebook pages [55][56]. The Graph API permits us to automatically extract information from Facebook. We first select those communities' pages that contain information about diabetes patients and drugs. Then, we extracted all posts from the communities' pages that were published between January 2017 and January 2019. The responses (reactions and emotions) made in these posts were collected and stored for further processing.

Twitter: The Streaming API and REST API of Twitter were employed to retrieve Tweets holding diabetes data. We retrieved the most recent Tweets using the REST API with various queries [57]. In these queries, the most specific terms related to diabetes should be used to retrieve the most-related Tweets. Thus, we built queries that are based on keywords with the Boolean operators AND and OR, e.g., patient AND (blood pressure OR blood sugar) for diabetes, patient AND (heart rate OR stress rate OR blood sugar) for hypertension, and patient AND (diabetes OR drugs) for diabetes drugs. More than 30 keywords were used to construct queries related to diabetes [58]. We fetched 300,000 Tweets about diabetes patients, treatments, drugs, and symptoms by using the aforementioned keywords-based queries.

Webpages: Diabetes patients are interested in knowing about any adverse side effects, diseases, and symptoms from taking a specific drug. Usually, patients share their experiences about adverse drug reactions in social media websites such as https://www.askapatient.com/, https://www.webmd.com/, http://www.druglib.com/, and https://www.drugs.com/. These websites contain a huge amount of information related to adverse drug reactions from anti-diabetes drugs. Therefore, these websites were chosen as data sources for the proposed system. We used the names of intake drugs for diabetes and BP as queries, retrieved 1600 posts from https://www.askapatient.com/ and https://www.webmd.com/, and used a keyword-based search engine to extract posts that contain only drug-related information. In addition, the proposed system employed an automatic web clawer that gathered 25,000 user comments (reviews) about drugs from http://www.druglib.com/ and https://www.drugs.com/.

3.2 Data store in big data cloud server

In the proposed system, four different sources of data are considered to monitor patient health and provide useful information. Here, wearable devices data are used to extract physiological information about the individual patient. Second, patients' medical records are collected to determine their treatments and medical histories. Third, the patients' contents are extracted from social networking sites to understand their feelings, emotions, and stresses. Fourth, patient reviews about drugs are collected from medical webpages to identify the side effects of the current drug intake. These large volumes of data from four different sources are difficult to store and handle. In addition, the number of patients with various diseases is largely increasing, which generate a large volume of health-related data. Therefore, a smart approach is needed that handles and processes the data competently. Hence, a big data cloud repository is utilized to store the extracted data so it can easily be accessed from any place and at any time. The data must be stored in an intelligent way so the information can be retrieved correctly and rapidly. The data used in our proposed work are in large volumes with versatility and velocity. Therefore, the traditional methods of storage and retrieval may not process the data from different sources. The proposed system is connected with a personal cloud server called Amazon S3 [14], which is highly scalable and secure to store patient information. Amazon S3 stores the data in buckets, and processes them for different purposes. A unique name and URL is assigned to each bucket, and the s3cmd method is utilized to upload the data into Amazon S3. The s3cmd method allows users to upload, retrieve, and manage the data in the cloud database [54]. Amazon S3 provides a facility to upload the data into an HBase cluster. As shown in Fig. 3, wearable sensor and social networking data are transferring from the cloud server to a Hadoop Distributed File System (HDFS). HDFS is a distributed file storage system that provides high bandwidth in order to transfer the data to HBase cluster [52,59]. HBase runs on top of HDFS. It is distributed database management, which stores the data in the form of rows and columns. Flume is used to transfer the data from the cloud server to the Hadoop ecosystem. Flume contains three units (source, decorator, and sink), which present the sources of the data, decorations of the data (e.g., compression and decompression), and targets of the data for specific purposes, respectively.

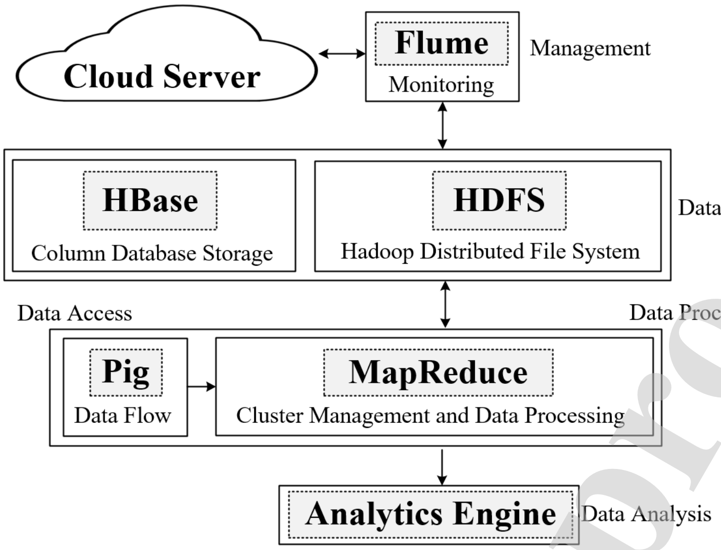

Fig. 3. Big data cloud server and the Hadoop Distributed File System. This system is based on HDFS and MapReduce. The HDFS provides high bandwidth in order to transfer the data to HBase cluster, which stores the data in the form of rows and columns. The MapReduce utilizes Apache Pig to handle structured and unstructured data and represents them for data analysis.

We use Apache Pig to transfer the data extracted by wearable devices into MapReduce. Apache Pig handles structured and unstructured data and represents them for data analysis. MapReduce is a parallel processing system in Hadoop for large datasets, which works in a distributed environment. It has two main functions, namely, Map and Reduce. The map task collects values from the big data in key-value pairs. The reduce function stores the value set for a specific key. MapReduce intelligently handles structured and unstructured data.

3.3 Data analysis using the big data analytics engine

The data of the proposed system consists of sensing data, medical records, and social networking data. However, it is extremely difficult to handle real-world big data due to its inconsistencies, missing values, noise, different formats, large size, and high dimensionality. Low-quality and noisy data produce low-quality results. The data preprocessing step is applied before actual processing, which improves the quality of the data processing and saves time. Our system includes pre-analysis of sensor data, preprocessing and filtering of sensor data, preprocessing of medical records, and preprocessing of social network content, as shown in Fig. 4.

3.3.1 Pre-analysis of sensor data

The physiological information of the patient is extracted using wearable biomedical and behavioral sensors. Different parameters are sensed from diabetes and BP patients using sensors and a smartphone, as shown in Table 1. The sensed parameters cover most of the symptoms of diabetes, abnormal BP, and other diseases. In addition, other parameters are extracted from the patient's body. However, for simplicity, only these parameters are mentioned in the proposed work. In pre-analysis, various steps are performed. First, the dataset is converted into comma separated value (CSV) files for easy parsing. Names are assigned to each attribute (parameter), and they are presented in the form of columns along with numerical values. Identities (IDs) of sensor data are then plotted to the actual sensor name (e.g., S1 is the ID for blood sugar). We use various attributes for different purposes. For example, we utilize blood sugar, blood pressure, heart rate, body mass index (BMI), age, gender, and activities attributes to classify the health condition of diabetes and BP patients. However, instead of using an explicit year and time for age and activities, respectively, we divide the attribute age into day, months, and years, and divide activities time into morning, afternoon, and evening. These generated features provide valuable insights into patient health, such as any pattern of normal and abnormal conditions based on whether it is the start or the end of the month, or if there is any serious condition during the starting and ending months of the year. Afterward, the final datasets are uploaded into the Hadoop cloud environment for further processing.

3.3.2 Preprocessing and filtering of sensor data

Data collected using wearable sensors have various limitations. They contain a lot of incorrect and useless information. In addition, sensor data are corrupted by signal artifacts, such as noise and missing values, which highly decrease classification performance. Therefore, the data are preprocessed and filtered before analysis. We filter the data to remove inconsistences and noise. The data are cleaned by removing ASCII characters. A well-known filtering approach called Kalman filtering is utilized to remove such noise from the data [60]. Furthermore, the unsupervised filter called *ReplaceMissingValues* is utilized to replace all missing numeric values in the dataset with means and modes from the available data. Useless attributes are removed with a maximum variance of 90% using an unsupervised filter called the *RemoveUseless* filter. The numerical values are then normalized using a *Normalize* filter to limit them to between 0 and 1 for any classification. After these steps, the *EmEditor* is applied to divide the dataset into *n* data files, and they are uploaded into the Hadoop cloud environment for further processing.

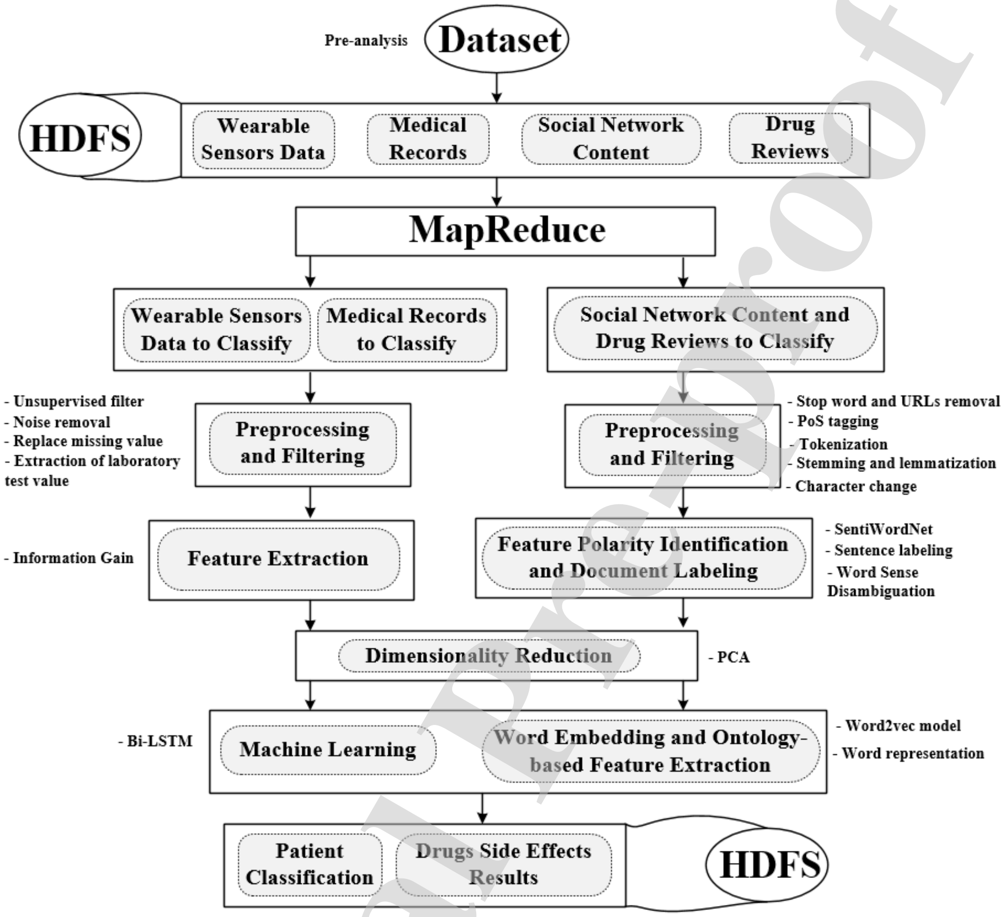
Pre-analysis
Dataset
HDFS
Wearable Sensors Data
Medical Records
Social Network Content
Drug Reviews
MapReduce
Wearable Sensors Data to Classify
Medical Records to Classify
Social Network Content and Drug Reviews to Classify
- Unsupervised filter
- Noise removal
- Replace missing value
- Extraction of laboratory test value
Preprocessing and Filtering
Preprocessing and Filtering
- Stop word and URLs removal
- PoS tagging
- Tokenization
- Stemming and lemmatization
- Character change
- Information Gain
Feature Extraction
Feature Polarity Identification and Document Labeling
- SentiWordNet
- Sentence labeling
- Word Sense Disambiguation
Dimensionality Reduction
- PCA
- Bi-LSTM
Machine Learning
Word Embedding and Ontology-based Feature Extraction
- Word2vec model
- Word representation
Patient Classification
Drugs Side Effects Results
HDFS

Fig. 4. The flow chart of the big data analytics engine. This analytics engine handles three different types of data: wearable sensors data, medical records, and social network contents. First, the preprocessing and filtering techniques are applied to preprocess the data. Second, the information gain approach is used to extract features from sensor's data. In addition, the method of feature polarity identification is applied to label the textual data. After this, the data dimensionality is reduced using PCA. Finally, word embedding model is trained to represent textual data, and machine learning classifiers are applied to predict patient health conditions and drug side effects.

3.3.3 Preprocessing of Medical Records

The Medical Records (MR) contain complete patient history in the digital format. It consists of different medical data describing the patient's health, such as laboratory tests, self-examination answers, and medications taken. Laboratory tests are data from medical devices that can be utilized to judge the health status of the patient in terms of reference values. In addition, the evaluation of the patient's health status can change depending on the patient's disease history and family disease history. Self-examination data contain information, which is collected from patients through questions. The questions are regarding periods of indigestion, drinking and smoking habits. The medication data include information about prescribed drugs for diagnoses. Some of the attributes of the MR can be used for patient classification. Therefore, the ID and reference value (0, 1, or 2) are assigned to each attribute of the MR for data analysis. The reference values 0, 1, and 2 indicate the patient's health condition (normal, pre-diabetes, and diabetes). In addition, we also use the MR data to overcome the limitations of sensors-based generated data; for example, we replace a missing value with the current MR attribute value in the dataset.

3.3.4 Preprocessing of social network content

Preprocessing of social network content is an important task before text classification. It reduces the noise that occurs in the corpus data, and transforms the data into a set of words that are useful representations for ML classifiers. The proposed system collects social network content and drug reviews, and stores them in the HDFS. The collection procedures for this content are explained in Section 3.1. The following steps are applied to transform the data into structured form. In addition, these steps remove useless data, which helps to easily extract features and opinion words.

Table 1. Different health parameters collected from sensors, smartphones, and medical records.

Resource	ID	Parameter	Explanation
Sensors	S1	Blood sugar	Blood glucose (mg/dL)
	S2	Body temperature	Patient's current body temperature
	S3	Blood pressure	Systolic blood pressure (mmHg)
			Diastolic blood pressure (mmHg)
	S4	Oxygen saturation	Patient's SpO_2 consumption (mmHg)
	S5	Heart rate	Patient's heart rate (bpm)
	S6	ECG	Patient's electrocardiography using ECG sensor
	S7	EEG	Patient's electroencephalography using EEG sensor
	S8	Stress	Patient's stress calculation using ECG+EEG pattern
Smartphone	SP1	Age	Age of the patient in days, months, and years
	SP2	Height	Stadiometer to find the patient's height
	SP3	BMI	Body mass index (kg/m^2)
	SP4	Gender	Patient's gender (0/1)
	SP6	Activities	Lifestyle: sedentary, slightly active, moderately active, active, or very active
Hospital medical records	MR1	Lipoprotein level	Low-density lipoprotein level (LDL cholesterol)
			High-density lipoprotein level (HDL cholesterol)
	MR 2	Hemoglobin	Glycated hemoglobin (A1c) of patient (%)
	MR 3	Blood sugar	Patient's blood test
	MR 4	Serum creatinine	Patient's blood test
	MR5	Triglycerides	Patient's blood test
	MR6	Cholesterol	Patient's blood test
	MR7	AST (SGOT)	Blood test to check for liver damage
	MR8	ALT (SGPT)	Blood test to check for liver damage
	MR9	Drugs intake	Extracted from prescription list
	MR10	Smoking	Yes/No (self-examination)
	MR11	Drinking	Yes/No (self-examination)
	MR12	Indigestion	Yes/No (self-examination)
	MR13	Family's disease history	Yes/No (self-examination)
	MR14	Patient's disease history	Yes/No (self-examination)

Stop word removal: Words such as prepositions (to, in, and of), all articles (a, an, and the), symbols (#, @, etc.), and URLs in the corpus data do not disturb the meaning of the document. However, they reduce the accuracy of text classification. Therefore, it is essential to remove them to decrease noise in the text. We utilize a well-known method called Rainbow to delete this content [58].

Tokenization: Tokenization separates a complex text in the corpus into small terms or tokens by removing white space and delimiters. Generally, white space and delimiters occur in a complex text. Therefore, the n-gram tokenizer is applied to delete white space and delimiters. The output is then saved for further analysis, such as extracted words' part of speech (PoS) tagging, and lemmas [61].

PoS tagging: PoS tagging defines words in the text. We split the corpus text into sentences, and then use Stanford Core Natural Language Processing (CoreNLP) for POS tagging. After tagging, it is confirmed that every sentence has a complete clause with a noun and a verb [55].

Stemming and lemmatization: Stemming converts the words in the corpus text into their own basic forms. The system applies a suffix-dropping algorithm for stemming. Lemmatization expresses the lemma of words used in the text. After lemmatization, the system easily obtains the lexical information of each word. For example, *blood sugar* is related to *blood glucose*. Therefore, the stem and lemma words are utilized for further processing.

Character conversion: Patients employ unusual words (e.g., *depresssssssed*) on social networks that affect the results of classifiers. Therefore, we convert a series of characters that appear more than twice into a general word (e.g., *depressssed* becomes *depressed*).

3.4 Feature polarity identification and document labeling

The proposed system detects the patient's stress and depression using their published contents on social networks. In addition, the system performs multiple tasks on drug reviews to identify their opinions on the efficiency and side effects of diabetes drugs. We use a sentiment analysis approach for the aforementioned two tasks. Therefore, it is important to find the feature polarity and document labeling for sentiment classification. After preprocessing of social networking data and webpage contents, we use SentiWordNet (SWN) to identify the polarity of the opinion words of the features [62,63]. The results of the feature polarity are then accumulated to find the polarity of the whole document. SWN is a resource of lexicons, which links each synset of a WordNet with three numeric values: positive, negative, and objective. However, SWN contains senses for each word to indicate if it is a noun, verb, adjective, or adverb. Therefore, Word Sense Disambiguation (WSD) is used to handle this issue by extracting the needed category sense for each word. In addition, it assigns a zero value to the input word if SWN does not hold any sense for it. After WSD, we extract the SWN scores for the identical senses of the opinion words, and then calculate the polarity of each feature using the following equations [64]:

$$Pos_{score}\ (F_i) = \sum_{w \in wf_i}^{n} Pos_{score}SWN_w \qquad (1)$$

$$Neg_{score}\ (F_i) = \sum_{w \in wf_i}^{n} Neg_{score}SWN_w \qquad (2)$$

$$Nue_{score}\ (F_i) = \sum_{w \in wf_i}^{n} Neu_{score}SWN_w \qquad (3)$$

where $Pos_{score}SWN_w$, $Neg_{score}SWN_w$, and $Neu_{score}SWN_w$ indicate positive, negative, and neutral scores of the feature, respectively. This score is computed by arithmetic means of SWN for individual word w. A detailed example is provided in [64]. If $Pos_{score}\ (F_i) > Neg_{score}\ (F_i)\ and\ Neu_{score}\ (F_i)$, then the system considers the feature polarity as positive. In contrast, the feature polarity is negative if $Neg_{score}\ (F_i) > Pos_{score}\ (F_i)\ and\ Neu_{score}\ (F_i)$. Lastly, it is considered neutral if $Neu_{score}\ (F_i) > Pos_{score}\ (F_i)\ and\ Neg_{score}\ (F_i).W_{1,1}$

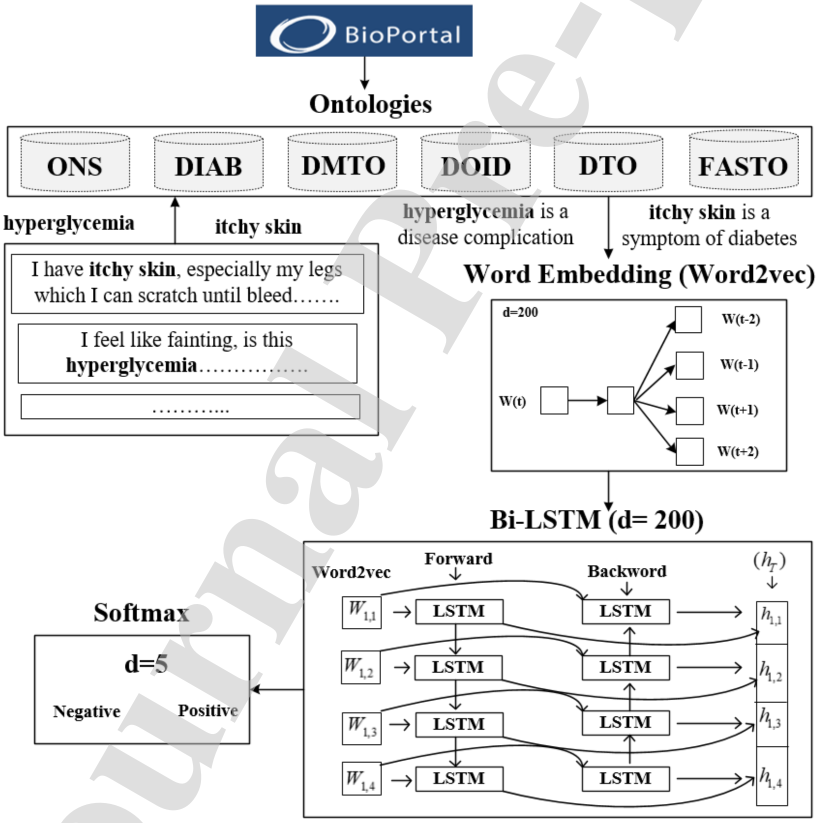

Fig. 5. Ontologies and Word2vec-based word representation. Ontologies provide additional information for Word2vec model to understand the semantic meanings of unusual words.

3.5 Word embedding and ontology-based features extraction from textual data

Word embedding is a word representation approach where words in the document are encoded with numeric values for advanced analysis. Two well-known approaches can be applied for word representation: one-hot word encoding and word embedding [55,65]. In this work, the word embedding approach is applied to represent words as numeric values. It sets the dimensionality, d, and then uses the d-dimensional values to express the word. For example, when d is set to 3, “metformin” is expressed as {0.3, 0.6, and 0.8}. This method drastically decreases the dimensions. However, the embedding approach avoids the relationships between words in a matrix. Therefore, we use a neural network–based word embedding model called

Word2vec for word representation. Word2vec contains two architectures: continuous bag-of-words (CBoW) and the skip-gram model. The CBoW model uses the surrounding context of the word for word prediction. The skip-gram model predicts the surrounding context using the current word, as shown in Fig. 5. We trained the skip-gram model of Word2vec with 200-dimensional vectors on the dataset. It identifies the relationships between words, and helps predict the neighboring words of the input word. This Word2vec model is then utilized to train ML classifiers to detect the association between features. Generally, ML classifiers miss the basic semantics of the features according to their particular domain. An ontology can represent the semantics of a given domain in some cases.

An ontology aims to provide semantic knowledge of concepts and their relations in a particular domain. Our online available biomedical ontologies cover various topics related to diabetes, depression, hypertension, drugs, and foods. In this paper, we use particular domain ontologies, where each class of the ontology is a concept or feature of the domain, and their properties are relations between features. This is the basic representation of the online available biomedical ontologies in the National Center for Biomedical Ontology BioPortal, which is currently the normal method to formalize knowledge about features. We use the latest releases of the Ontology for Nutritional Studies (ONS), the BioMedBridges Diabetes Ontology (DIAB), the Diabetes Mellitus Treatment Ontology (DMTO), the Human Disease Ontology (DOID), the Drug Target Ontology (DTO), and the Fast Healthcare Interoperability Resources (FHIR) and semantic sensor network (SSN)-based Type 1 Diabetes Ontology (FASTO), which contain lots of concepts and relationships, as shown in Table 2. These ontologies provide additional information for a neural network model that understands the semantic meanings of unusual words. In addition, they affect the word-level semantics in word embedding and text classification.

Table 2. BioPortal ontologies of different domains as datasets.

Ontology and Acronym	Definition	Number of Classes	Number of Properties	Number of Individuals
Ontology for Nutritional Studies (ONS) [66]	It provides a description of composite nutritional studies.	3442	66	104
BioMedBridges Diabetes Ontology (DIAB) [67]	It represents the relations between diabetes phenotypes for text mining.	375	4	0
Diabetes Mellitus Treatment Ontology (DMTO) [68]	It provides interoperable facts for diabetes treatment.	10700	315	63
Human Disease Ontology (DOID) [67]	It represents the concept of rare diseases.	12694	15	0
Drug Target Ontology (DTO) [69]	It provides information for the classification of drug target data.	10075	0	0
FHIR And SSN-based Type 1 Diabetes Ontology (FASTO) [6]	It provides management details on insulin for diabetes patients.	9577	822	460

In this section, the system retrieves all the ontologies and attempts to integrate the ontology information with word embedding for classifiers. Generally, the bag-of-words model is used for features. Therefore, we consider each ontology as a bag-of-words. To extract information from the ontology, we employ well-known statistical methods called term frequency (TF), and term frequency and inverse document frequency (TF-IDF) [35,55,70]. Here, we consider each concept or feature of the ontology as a term, and the ontology as a document. Therefore, TF is the terms (ambiguous words) found in the ontology. TF is mathematically defined in the following equation:

$$TF(Term, Onto) = 1 + \log(Feature_{term}, Onto) \quad (4)$$

If the $TF(Term, Onto)$ value is greater than zero, this step will be repeated to extract the specific concept. TF-IDF selects the illustrative terms for information extraction. The IDF reduces the significance of the word that appears mostly in all ontologies (e.g., patient, disease, hospital, etc.). If the term occurs in more ontologies or in more concepts in a single ontology, it means it is a regular term and may not be the required term for information extraction. Therefore, the result of the log function for the input word will decrease to zero. This shows that the value of TF-IDF is small for this term. The statistical description of the IDF is shown in the following equation:

$$IDF\ (Term, Ontos) = \log \frac{N_f}{1 + |\{Onto\ \in\ Ontos : Term\ \in\ Onto\}|} \quad (5)$$

where N_f indicates the total number of ontologies in the database or the total number of concepts in the ontology, e.g., $N_f = |O|$ and $|\{Onto \in Ontos : Term\ \in Onto\ \}|$ is the number of ontologies or the number of concepts in the ontology where *Term* appears. We remove the common terms using the following TF-IDF equation:

$$\mathrm{TF} - \mathrm{IDF} = TF(Term, Onto)\ .IDF(Term, Ontos) \quad (6)$$

The TF-IDF result shows how essential the ontology feature is to the ontology in the ontologies corpus, or shows how important the ontology feature is to the concept in the ontology. Based on the values of TF-IDF terms (X=2), the top concept can be extracted. Let us suppose that the system gets the text related to diabetes, as shown in Fig. 5. In this figure, the ambiguous words for information extraction are in bold type (e.g., *hyperglycemia*). Once the semantic information about these words is identified, the system then extracts that specific concept from the ontology to provide additional information to word embedding and ML classifiers (e.g*., hyperglycemia is a disease complication*).

3.6 Features extraction from wearable sensors data

The huge size of the data is another main issue associated with sensor-based healthcare monitoring systems. Raw data from the patient's body extracted in large quantities is a burden for data processing. It is important to decrease the size of the data without losing useful information. The dataset contains a large number of attributes about diabetes and BP patients. However, all the attributes will not be needed for patient classification. Unnecessary attributes are time consuming and decrease the accuracy of classification. Different methods are utilized for feature selection, such as the dragonfly algorithm and recursive feature elimination [71,72]. We use the Information Gain (IG) method, which affects the classifier by reducing noise and irrelevant features.

Information gain (IG): IG chooses features based on the information's contribution related to the variables of the class without seeing attribute interactions [73]. Every attribute or feature in the dataset has importance, and based on that importance, the system can learn about specific problems. Waikato Environment for Knowledge Analysis (WEKA) is utilized to compute the IG [74]. It contains various methods for data processing. We apply the IG filter *"infoGainAtrributeVal"* as an evaluator on the diabetes and BP dataset to obtain the results. However, IG cannot be used for a numerical dataset. Therefore, it is important to convert numeric data into nominal data before using IG. When the value of the attribute is identified, the IG measure is linked with a decrease in entropy in the training dataset. This approach finds the value of an attribute by calculating the IG according to classification. The proposed IG method utilizes entropy to measure system uncertainty and finds the difference between prior entropy and post entropy [75]. It specifies the amount of extra information about A provided by B, as shown in Eq. 7:

$$IG(A|B) = H(A) - H(A|B) \tag{7}$$

where A and B are discrete variables. A is a feature, and its prior entropy can be measured using Eq. 8:

$$H(A) = -\sum_{i} P(A_i) \log_2 P(A_i) \tag{8}$$

where $P(A_i)$ represents the prior probability for the discrete value of A_i. The conditional entropy of A, after given the post entropy B, can be defined as shown in Eq. 9 and Eq. 10:

$$H(A|B) = -\sum_{j} P(B_j)\, H(A|B_j) \tag{9}$$

$$= -\sum_{j} P(B_j) \sum_{i} \left(P\left(A_i|B_j\right) log_2\, P\left(A_i|B_j\right)\right) \tag{10}$$

The IG can be computed by putting Eq. 8 and Eq. 10 into Eq. 7, as shown in Eq. 11:

$$IG(A|B) = -\sum_{i} P(A_i) \log_2 P(A_i) - \left(- \sum_{j} P(B_j) \sum_{i} \left(P\left(A_i|B_j\right) \log_2 P\left(A_i|B_j\right)\right)\right) \tag{11}$$

3.7 Principal component analysis

The dimensionality of the dataset is increased after preprocessing and extra-attribute generation. This creates problems, such as a decrease in classification accuracy, over-fitting, and time complexity. Therefore, we use principal component analysis (PCA), which is a statistical approach to dimensionality reduction [76]. It converts p-dimensional features into q-dimensional features (q < p) with the least loss. The q-dimensional features are entirely new orthogonal features, called principal comments. The new orthogonal feature has its own exclusive meaning. The information of features can be imitated in variance. Feature with high variance rate shows that the feature contains the main information. This can be identified by cumulative variance rate. The main aim of PCA is to find the projection vectors with the highest variance. Projection vector X on vector Y on the same plane can be illustrated using Eq. 12:

$$f = \frac{\langle X, Y \rangle}{|X|} . Y \tag{12}$$

Let X is the sample of n-dimensional drug side effect dataset, where $X = \{x_1, x_2, x_3, \ldots\ldots, x_m\}$. The sample of the dataset is arranged and therefore, it has a zero mean. In addition, we assume that after projection transformation the new coordinate system is $W = \{w_1, w_2, w_3, \ldots\ldots, w_n\}$, where w_i represents the standard vector of orthogonal basis, which is $\|w_i\|_2 = 1$. The projection of x_i on the hyper-plane in the new space is $W^T X$, where $W^T X = w_1 x_1 \ldots \ldots w_n x_m$. However, the main objective is to find the projection of all points of X on the hyper-plane with maximum variance. The projected points with maximum variance can be calculated using Eq.13.

$$u = \underset{\|w\|_2=1}{\operatorname{argmax}} \sum_{i=1}^{n} (W^T X)^2 \tag{13}$$

Where u is an eigenvector related to the maximum eigenvalue (λ) of $X^T X$. It is important to achieve eigenvalue decomposition on the covariance matrix $X^T X$, and arrange the eigenvalues in descending order. Usually, the number of principal components is selected based on the cumulative variance rate. In this paper, the PCA with cumulative variance rate of 90% is applied to reduce the dimensionality of datasets. The original drug side effect dataset contains 200 feature dimensions. By using PCA and IG, the dimensionality of the input data is reduced to 67 features (principal components). In addition, it shortens the processing time by reducing the dimension of the dataset before feeding it to Bi-LSTM models for classification.

3.8 Bi-LSTM–based diabetes and BP classification and drug side effect prediction

Machine learning approaches, such as the CNN, MLP, the SVM, logistic regression, decision trees, and KNN, are useful algorithms for classification, feature selection, data normalization, and statistical analysis [21,77–79]. These approaches conduct tasks using supervised and unsupervised methods. However, they are time consuming, and their performance decreases when the data size is large [58]. Therefore, in the proposed healthcare monitoring system, Bi-LSTM is used to classify the patient's diabetes, BP, mental health, and side effects from diabetes drugs. The recommendation system is activated in cases of diabetes, high BP, stress, or depression, and from severe side effects of diabetes drugs. Fig. 6 shows the architecture of Bi-LSTM for the classification of diabetes, BP, mental health, and drug side effects. The grid search optimization algorithm is applied to identify optimal values for the hyper-parameters of the LSTM model. The selected optimal values for the hyper-parameter dropout rate, epochs, batch size, and learning rate are 0.3, 30, 32, and 0.001, respectively.

LSTM is a type of RNN that consists of five main components: memory cell ($\bar{C}^t$), input gate (i^t), forget gate (f^t), current memory cell (C^t), and output gate (O^t). These components control the update and use of previous data. The LSTM output can be computed using the following equations [58]:

$$\bar{C}^t = tanh\,(w_{xc} X^t + w_{hc} h^{t-1} + b_c) \tag{14}$$

$$i^t = \sigma\,(w_{xi} X^t + w_{hi} h^{t-1} + w_{Ci} C^{t-1} + b_i) \tag{15}$$

$$f^t = \sigma\left(w_{xf} X^t + w_{hf} h^{t-1} + w_{Cf} C^{t-1} + b_f\right) \tag{16}$$

$$C^t = f^t \odot C^{t-1} + i^t \odot \bar{C}^t \tag{17}$$

$$O^t = \sigma\,(w_{xo} X^t + w_{ho} h^{t-1} + w_{Co} C^{t-1} + b_o) \tag{18}$$

where X^t, w_x and w_h, and b are input, weight matrices, and bias vectors of LSTM, respectively; $tanh\,(.)$ and $\sigma\,(.)$ are the hyperbolic tangent function and sigmoid function, respectively. The final output of LSTM can be calculated using the following equation:

$$h^t = O^t \odot tanh\,(C^t) \tag{19}$$

where $\odot$ is features-wise multiplication between the output gate and input cell state. The LSTM output is linked with a softmax function to identify the probabilistic outputs (0, 1, and 2), where 0, 1, and 2 show that the patient is normal, pre-diabetes, or diabetes, respectively. The softmax function can be calculated using the following equation:

$$softmax^{st} = \frac{\exp\,(x^{st})}{\sum_k^c \exp(x^{\acute{s}t})} \tag{20}$$

where c and x^{st} are feature category and the input of time step k, respectively. After preprocessing, feature extraction, dimensionality reduction, and word embedding, a sequence of inputs (features and word vectors) representing the patient's physiological information and drug reviews is then fed into the Bi-LSTM layer. We developed four Bi-LSTM–based classifier models: LSTM^a, LSTM^b, LSTM^c, and LSTM^d.

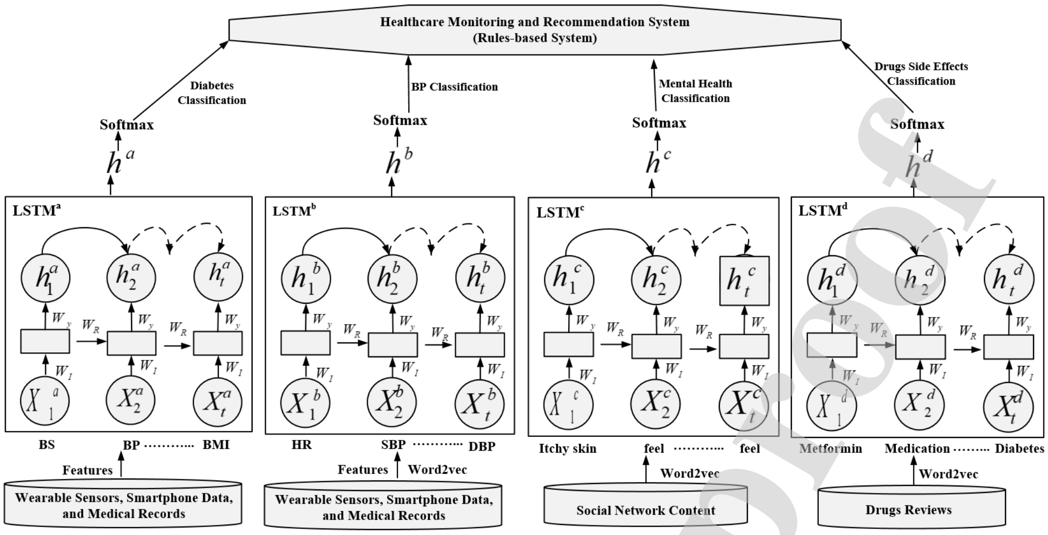

Fig. 6. LSTM-based healthcare data classification. The system consists of four LSTM models that predict the patient health condition.

In this study, wearable devices and smartphones are utilized to collect the patient's physiological and personal information. The collected data are preprocessed and converted into structured form for further processing. At each time step, *t*, feature X_t is fed into the LSTM^a units to predict diabetes class {normal, pre-diabetes, and diabetes}. The first LSTM unit of the LSTM^a model may predict the patient category based on a specific feature. However, all the features should be used to get important information for classification. Therefore, the sequence of diabetes features, along with the results of the previous unit, is assigned to the next LSTM unit. This procedure is repeated with each input feature. In this way, the units of LSTM^a save the valuable features and generate the output. The output is linked with a softmax activation function, which predicts the diabetes patient category. The medical rules to classify diabetes and blood pressure patients are presented in Table 3. Wearable sensors for blood pressure and an ECG are used to identify the patient's BP and heart rate (HR), respectively. The patient can be evaluated from systolic BP and diastolic BP values in millimeters of mercury, and HR values in beats per minute, as shown in Table 3. However, these parameters are sometimes affected by physical activity, sleep, temperature, stress, eating habits, etc. [5]. In addition, patients may have BP and HR problems due to other factors, such as family profile and disease history. For example, HR is always high during exercise. This means that an abnormal health condition in the patient is sometimes not dangerous. Therefore, it is essential to use the patient's medical records and daily activities to detect abnormal conditions. We select main features from medical records, including age, gender, and family history, from the diagnosis and medical history. All the diagnosis and medical records are searched to extract information related to diabetes and BP.

Table 3. Medical rules to classify diabetes and BP patients.

	Category	**Age (years)**	**Family (Y/N)**	**Gender (M/F)**	**Activity (in steps)**	**BMI**	**Blood sugar (MG/dL)**	**Heart rate (bpm)**	**Systolic BP (mmHg)**	**Diastolic BP (mmHg)**
Diabetes classification	Normal	20-50	N	M/F	5000+	18-25	≤ 100 (fasting)	60-80	-	-
	Pre-diabetes	41-65	Y	M/F	3001-5000	25-30 or more	100 to 125 (fasting)	140-160	-	-
	Diabetes	60+	Y	M/F	≤ 3000	30+	≥126 (fasting)	140-160 or ≥160	-	-
Blood pressure classification	Normal	20-50	N	M/F	≤ 3000	-	-	60-100	90-119	60-79
	Hypotension (Low BP)	41-65	Y	M/F	3001-5000	-	-	-	≤ 90	≤ 60
	Hypertension (High BP)	60+	Y	M/F	5000+	-	-	-	≥140	≥ 90

The extracted information is then used for feature construction. For example, if the BP data are found in the patient family profile, then a 1 appears as the Family feature. If data about BP are not found in diagnosis and medical records, then a null value is used in that feature. Later, all null values are replaced with a zero. After generating and preprocessing the features, we developed a model based on Bi-LSTM, named LSTM^b, which is capable of keeping temporal patterns present in historical sequenced data. For the classification of BP, all the parameters and generated features at each time step *t* are converted into category vectors {0, 1, or 2}.

In mental health monitoring, patient messages and posts are extracted from social networks and are filtered using supervised and unsupervised approaches. However, text related to depression and stress are usually short, unstructured, contain negative emotions, and low sentiment polarity. Therefore, the text mining and ML approaches are used to efficiently handle the social network content and identify the sentiment polarity of a text with the terms happy, normal, depression, or stress. Here, first the emotional texts are identified, and the sentiment analysis method is then applied to classify the text [55,58,80]. The Word2vec model with ontologies is used in this work to represent the text for deep learning classifiers. We trained a 200-dimensional Word2vec model, and then fed the sequence of words into the LSTM^c model. The output of LSTM^c is then assigned to the softmax function, which predicts the sentiment label (such as positive, neutral, negative, or strong negative) for each sentence of the published content. The rules to classify the patient's mental health based on the predicted polarity are shown in Table 4.

In drug side effect prediction, the current intake diabetes and BP drugs are used as input queries, and we retrieved reviews about them from different websites. After filtering and preprocessing, the reviews are automatically labeled as having no side effects, moderate side effects, and severe side effects. For this labeling, the sentiment polarity of each sentence is identified. Then, positive, neutral, and negative sentiment polarity are considered as having no side effects, moderate side effects, and severe side effects, respectively, as shown in Table 4. The 200-dimensional Word2vec model is trained to represent the drug reviews as a sequence of words in the LSTM^d model. The output of LSTM^d is then fed to the softmax function, which predicts drugs side effects accurately.

Table 4. Rules to classify a patient's mental health and drug results.

	Category	**Sentiment polarity about patient posts and comments (Positive/Neutral/Negative/Strong Negative)**	**Sentiment polarity about anti-diabetic drugs (Positive/Neutral/Negative)**
Patient mental health classification	Happy	Positive	-
	Normal	Neutral	-
	Depressed	Negative	-
	Stressed	Strong Negative	-
Drug side effect prediction	No side effects	-	Positive
	Moderate side effects	-	Neutral
	Severe side effects	-	Negative

4. Experiment and results

In this section, the evaluation procedure of the proposed healthcare monitoring system is presented, and the results are discussed. Experiments were conducted on four different types of dataset, which were extensively discussed in subsection 3.1.

4.1 Performance evaluation

To evaluate the efficiency of the proposed approach, various experiments were conducted on the four different datasets. The datasets were extracted from the patient's body and from social networks using wearable sensors and APIs, respectively. The extracted datasets were then sent to the Hadoop cloud environment, which is a platform for big data analysis and processing. Hadoop MapReduce analyzed the extracted datasets for further processing. Furthermore, the generated datasets were then used to build the models for the classification of diabetes, blood pressure, mental health, and drug side effects. The word embedding approach was used to represent the social networking data and drug reviews for the classification models. The number of features utilized from diabetes and BP datasets, and the dimensionality of word embedding for drug reviews and social networks datasets are discussed as follows.

Diabetes and BP datasets: For diabetes classification, the Pima Indians Diabetes dataset was acquired from the UCI machine learning repository [81]. The Pima Indians dataset contains the records of 768 patients, out of which 268 tested positive for diabetes, while 500 tested normal. There are eight input attributes in the dataset. However, only six attributes are used for training the diabetes classification model. The age, family, gender, activities, BMI, blood pressure, and blood sugar features are used for diabetes classification. The dataset from the PhysioNet multi-parameter intelligent monitoring in intensive care II (MIMIC-II) database was used to train the BP classification model [82]. This dataset comprises a large number of features, including BP and HR. However, only nine attributes are utilized for training the BP classification model. The age, gender, BMI, heart rate, systolic BP, diastolic BP, activities, blood sugar, and family history features are used for BP classification. Furthermore, we also combined our new datasets regarding diabetes and BP with the abovementioned datasets. Thus, the total number of instances for diabetes and BP classification are 868 and 550, respectively.

Drug reviews and social networking datasets: The dataset for drug side effect prediction was acquired from the UCI repository [32]. This dataset consists of six attributes. However, only two attributes (the name of the drug and patient reviews) are considered in the proposed work. We extracted patient messages and posts from social networks for patient mental health monitoring. The social networking and drug reviews dataset are extensively discussed in subsection 3.1. We trained a 200-dimensional Word2vec model to represent both the social networking and drugs reviews datasets for classification.

Machine learning algorithms such as logistic regression and SVM have been utilized to classify the structured and unstructured data [83] [84] [85]**.** These algorithms are shallow and trained on sparse and high-dimensional features, which may not achieve the required results. Deep learning algorithms including MLP and CNN have been utilized for the purpose of textual data classification. However, MLP performed well for small amount of input data. MLP input is constant and link only with the current instant. This algorithm may not handle the textual data to generate the accurate results. It consumes a lots of time for data processing and the classification results are worse than others [78]. CNN has been compared with RNN in terms of sentiment classification [77]. The authors described that CNN and RNN are useful for informative features extraction and sequences of unit modeling, respectively. CNN utilizes the fixed size of a window that moves over a textual data to extract features from a sequence of terms. However, CNN misses the semantic meaning of words and unable to extract the valuable words from lengthy sentences. In this evaluation, we selected the following algorithms based on the comparative study and discussions along with their limitations reported in our recent work [55,58].

- Convolution neural network (CNN): CNN has been applied for various classification tasks using sensors and textual data [86–88]. We compared it with our proposed model in order to understand the efficiency of the proposed classifier. We used the CNN of the WEKA [74] library along with a sigmoid activation function.
- Multilayer perceptron (MLP): MLP is a neural network model that learns the data pattern using several layers with connected perceptrons [89]. We trained the MLP of WEKA library with 3 hidden layers and a squared error function. The sigmoid function was used as an activation function.
- Support vector machine (SVM): SVM with linear kernel function is used to classify the datasets, and compare it with the proposed model [83]. SVM is suitable for light training data, which comprises non-zero values. However, it is time-consuming during sparse data training. SVM with a training parameter kernel-type radial basis function is considered in the proposed work.
- Fuzzy classifier: This classifier is a rules–based classifier that categorizes the data using fuzzy rules in the form of *if-then*. We utilized fuzzy classifier with fuzzy unordered rule induction algorithm [90].
- Logistic regression: This algorithm is applied to define the association between a binary response variables [84]. It predicts the categorical outcomes based on certain predictors. It utilizes the relationship function that transforms the probability range [0, 1] into (-∞*, +*∞). We applied logistic regression classifier with the training parameter ridge estimator.
- Random forest: This algorithm constructs a forest of random trees for classification [53]. It combines multiple decision trees to achieve the accurate final results. We used it with number of iterations 100 and seed 1.
- K-nearest neighbors (KNN): The KNN is non-parametric approach that is utilized for the purpose of both regression and classification. KNN first stores all cases, and then classifies the new cases based on the similarity measures. We utilized KNN at k = 3 with the function of Euclidean distance.

The above-mentioned models were trained by randomly distributing the datasets into training and testing sets at 70% and 30%, respectively. The proposed models were used both with an ontology and without an ontology, and performance was evaluated and compared, each model with the others. In addition, PCA and IG were utilized with the proposed models and the results are compared. The WEKA API was applied to train the word embedding model and to evaluate the effectiveness of the classifiers. This system was developed employing WEKA and the Protégé OWL tool with Java.

4.2 Results

To assess the usefulness of the aforementioned models, different evaluation metrics were used, including precision, recall, accuracy, function measures, root mean square error (RMSE), and mean absolute error (MAE). In this section, the results of the above-mentioned experiments are presented. The obtained results from diabetes classification are shown in Table 5. The different classifiers (CNN, MLP, SVM, fuzzy logic, logistic regression, random forest, and KNN) were compared with the proposed LSTM^a model using the Pima Indians dataset. The results show that the proposed LSTM^a obtained the highest accuracy (75%) and the lowest MAE (26%) in comparison with the other classifiers. This high accuracy indicates that LSTM^a stores more important information in the memory cells for diabetes classification. In addition, the lowest MAE shows the better performance of the LSTM^a model. The fuzzy classifier obtained the lowest RMSE (46) compared to the other classifiers. Based on this experiment, we observe that the other classifiers are time consuming, and their performance decreased for even a small number of features. However, LSTM outperformed the other classifiers in the prediction of diabetes.

Table 6 shows the results obtained by the proposed model and the other classifiers in terms of BP classification. These classifiers were trained using the PhysioNet MIMIC-II dataset. We observe that LSTM^b achieved the highest accuracy (88%) compared to the CNN (70%), MLP (80%), the SVM (73%), the fuzzy classifier (83%), logistic regression (72%), random forest (70%), and KNN (58%). However, the MAE and RMSE of the proposed model were 12 and 34, respectively, which are lower than the other classifiers. As shown in Table 6, the accuracy of the fuzzy classifier is surprisingly high. This is because the training data are handled by generated fuzzy rules. In addition, the results with KNN are surprisingly poor. The value of K affected the results from KNN. If the value of K is too big, then KNN misclassifies the test sample. This is because of the large distance between data points and the test sample of the neighborhood.

Table 5. Comparison of the LSTM^{a} model with other classifiers in terms of diabetes classification based on six features.

Proposed Models and Other Classifiers	**Precision (P) (%)**	**Recall (R) (%)**	**Function Measure (FM) (%)**	**Accuracy (Ac) (%)**	**RMSE**	**MAE**
CNN	62	66	63	66	58	34
MLP	67	67	67	68	51	34
SVM	67	70	68	70	54	30
Fuzzy classifier	73	72	65	72	52	26
Logistic regression	70	68	69	69	55	32
Random forest	67	70	66	70	46	42
KNN	57	65	59	65	52	50
LSTM^{a}	74	75	75	**75**	50	**26**

Table 6. Comparison of the LSTM^{b} model with other classifiers in terms of BP classification based on nine features.

Proposed Models and Other Classifiers	**Precision (P) (%)**	**Recall (R) (%)**	**Function Measure (FM) (%)**	**Accuracy (Ac) (%)**	**RMSE**	**MAE**
CNN	71	70	69	70	54	30
MLP	80	80	80	80	37	22
SVM	83	75	73	74	50	25
Fuzzy classifier	84	83	83	83	40	16
Logistic regression	72	72	71	72	51	29
Random forest	78	70	67	70	45	42
KNN	60	58	56	58	64	42
LSTM^{b}	89	87	87	**88**	34	**12**

We extracted patient emotional comments and posts, and drug reviews from social networking and drug-related webpages, as discussed in subsection 3.1. The emotional posts were labeled automatically using our proposed method, as explained in subsection 3.4. The patient's mental health was classified using social network data, and the side effects of diabetes drugs were predicted using drug reviews. To accomplish these tasks, Word2vec models were trained to represent these data in the form of word vectors with high dimensionality.

Table 7 shows the performance of the classifiers in terms of mental health classification. According to Table 7, LSTM^{c} using softmax shows the highest accuracy (89%) in comparison with the other classifiers. In addition, the MAE and RMSE of LSTM^{c} are 15 and 35, respectively, which is lower than the other classifiers. However, the fuzzy classifier and random forest obtained lower accuracy in comparison with the other classifiers. Random forest faces over-fitting problems due to its complex representation and noise in the dataset. Based on this experiment, we observe that Word2vec and LSTM with softmax are better than the other classifiers in terms of emotional text classification. LSTM^{c} presented the best results because its memory function arranges the text in two directions that affect the text classification results. It is important to note that the LSTM^{c} outputs of positive, neutral, negative, and strong negative indicate the patient's mental health as happy, normal, depressed, and stressed, respectively, as listed in Table 4.

Table 7. Comparison of the LSTM^{c} model with other classifiers in terms of mental health classification based on 200-dimensional Word2vec model.

Proposed Classifier Model and Other Classifiers	**Precision (P) (%)**	**Recall (R) (%)**	**Function Measure (FM) (%)**	**Accuracy (Ac) (%)**	**RMSE**	**MAE**
CNN	66	66	66	66	56	34
MLP	72	71	71	72	47	29
SVM	68	68	68	68	56	32
Fuzzy classifier	60	58	56	59	67	45
Logistic regression	70	75	72	71	44	34
Random forest	60	60	59	60	58	40
KNN	64	63	64	64	60	36
LSTM^{c}	84	83	85	**84**	35	**15**

To evaluate the classification results of the proposed model from using drug reviews, the LSTM^{d} model was compared with other classifiers, as shown in Table 8. LSTM^{d} and logistic regression obtained high accuracy at 90% and 83%, respectively. Other classifiers achieved lower accuracy in comparison with LSTM^{d} and logistic regression. RMSE and MAE of LSTM^{d} are 32 and 13, respectively, which is lower than the other classifiers. The accuracy of the CNN is very low, and its RMSE is very high, compared to the other classifiers. This indicates that LSTM^{d} can handle longer sequences of word vectors as the dimensions of word vectors increase. In contrast, the CNN results in over-fitting when the sequence of words and the dimension of word embedding increase. It is important to note that the LSTM^d model outputs of positive, neutral, and negative show the drug side effects as having no side effects, moderate side effects, and severe side effects, respectively.

Table 8. Comparison of the LSTM^d model with other classifiers in terms of drug side effect classification based on 200-dimensional Word2vec model.

Proposed Classifier Models and Other Classifiers	**Precision (P) (%)**	**Recall (R) (%)**	**Function Measure (FM) (%)**	**Accuracy (Ac) (%)**	**RMSE**	**MAE**
CNN	76	68	65	68	62	39
MLP	82	81	81	81	39	20
SVM	82	82	82	82	36	18
Fuzzy classifier	79	78	78	78	41	28
Logistic regression	84	82	83	83	36	17
Random forest	76	66	63	66	58	33
KNN	77	75	75	76	49	24
LSTM^d	88	90	89	**90**	32	**14**

4.3 Accuracy and MAE of the proposed models with ontology, PCA, and IG

We utilized ontology-based semantic knowledge with Bi-LSTM in order to classify the social networking data and drug reviews. The proposed ontology presented semantic knowledge of the features related to depression and diabetes drugs. Fig. 7 presents the accuracy and MAE of the proposed LSTM^c and LSTM^d models using an ontology and without using an ontology. As can be seen, the proposed ontology-LSTM shows significant improvement over simple LSTM. In terms of emotional text classification, the accuracy of LSTM^c was 84%, which increased to 87% when using the ontology. However, MAE decreased by just 1. In addition, the accuracy and MAE of LSTM^d was 90% and 14, respectively, in terms of drug side effect classification. However, this accuracy increased to 93%, and MAE decreased by 1 when using the ontology with LSTM^d. The obtained results indicate that the proposed system can detect the most efficient features in the text and provide additional information about them. In addition, the proposed ontologies explore the domain information of diabetes drugs and depression. Therefore, they identify depression and drug features in the dataset, and help word embedding and LSTM to understand the exact meanings of the features in the task of text classification. The generated results indicate that the proposed system outperforms in comparison with simple LSTM and the other classifiers.

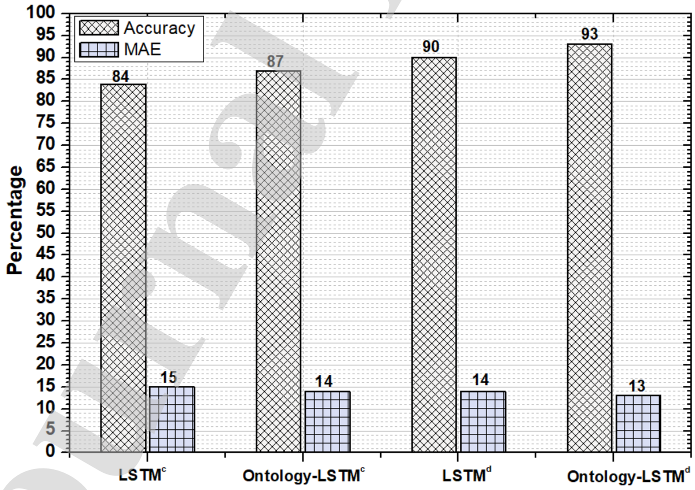

Fig. 7. Accuracy and MAE comparisons of the LSTM^c and LSTM^d models with ontology-LSTM^c and ontology-LSTM^d.

We applied PCA and IG to further improve the accuracy of the proposed models. PCA combined the original features of each dataset to generate new independent features. However, only a few features are selected from the newly generated features. The selected features are less in number than the original features and that is how the dimensionality is reduced.

PCA first normalized features using mean and standard deviation function, which is the square root of the variance. PCA then used covariance function in order to know the correlation between those features. Finally, the new features are selected based on their eigenvalues to reduce the dimension. For example, PCA reduced the dimensionality of the diabetes dataset by transforming six features into five features, as shown in table 9. Vector Y_1, Y_2, Y_3, Y_4, and Y_5 are called principal components.

These new vectors are selected because of their high eigenvalues compared to other vectors. In this research work, the PCA with cumulative variance rate of 90% is applied to reduce the dimensionality of datasets.

Table 9. PCA-based selected features from diabetes dataset.

Vectors	New features	Eigenvalues
Y_1	0.205 Glucose = 172+0.205, Diabetes Pedigree Function = 0.702+0.205, Insulin = 579-0.194, Pregnancies = 17-0.194, Diabetes Pedigree Function = 0.817...	2.799
Y_2	-0.304 Diabetes Pedigree Function = 0.817-0.304, Pregnancies = 17-0.299, BMI = 40.9-0.27, Insulin = 114-0.224, Glucose = 163...	2.107
Y_3	0.311 Insulin = 240+0.269, BMI = 45.4+0.269, Diabetes Pedigree Function = 0.721+0.242, Blood Pressure = 110+0.217, Glucose = 171...	1.500
Y_4	0.216 Glucose = 172+0.216, Diabetes Pedigree Function = 0.702+0.216, Insulin = 579+0.212, Insulin = 240+0.201, Diabetes Pedigree Function = 0.721...	1.092
Y_5	0.185 Skin Thickness = 7+0.176, Pregnancies = 12+0.162, Diabetes Pedigree Function = 0.926+0.162, Insulin = 258-0.158, BMI = 43.1...	0.677

After applying PCA, IG evaluated the value of each feature by computing the IG with respect to the class. IG did not reduce the PCA-based selected features in case of diabetes classification, and considered that all the five features provide maximal information. Similar procedure of PCA and IG is applied for other three datasets. However, IG selected the best features in case of other datasets that give useful information, and removed the unrelated features. The dataset of diabetes, BP, mental health, and drug side effect contain 6, 9, 200, and 200 features, respectively. The PCA and IG reduced the input data features to 5, 7, 99, and 67, respectively. The obtained results using the classification models along with PCA and IG are shown in Fig. 8. PCA and IG along with ontology-LSTM reduced the dimensionality of a large dataset that still contains most of the useful information. This shows that PCA leads to an approximately 2x reduction in training time of LSTM models. Based on the obtained results in Fig. 8, we note that the accuracy of LSTM^a, LSTM^b, ontology-LSTM^c, and ontology-LSTM^d with PCA and IG increased by 4%, 1%, 3%, and 1%, respectively. This indicates that using PCA and IG with ontology-LSTM is more efficient than the other classifiers in terms of numeric and textual data classification. In addition, it also shows that Word2vec and LSTM without an ontology miss the semantic meanings of features related to the drug and disease domains. The main reason is that many hidden features of diabetes and depression represent different notions.

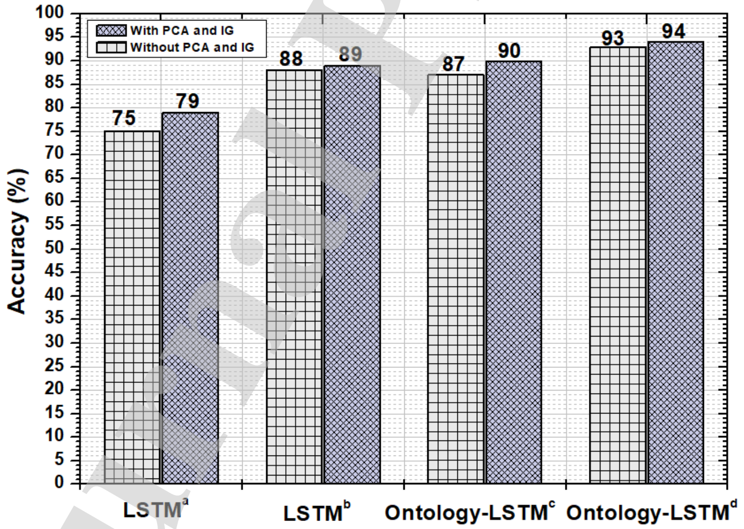

Fig. 8. Proposed classification model accuracy with and without PCA and IG.

5. Conclusion

In this work, a novel healthcare monitoring framework for chronic patients was presented, which integrates advanced technologies, including data mining, cloud servers, big data, ontologies, and deep learning. The proposed framework enhances the performance of heterogeneous data handling and processing, and improves the accuracy of healthcare data classification. The proposed method correctly examines diabetes and blood pressure (BP) patients using various sources for their data, such as smartphones, wearable sensors, medical records, and social networks (called big data). Various reasonable issues were discussed, including big data storage in a cloud server, MapReduce-based data processing, data preprocessing using data mining techniques, feature extraction, and dimensionality reduction, and the role of domain-specific ontologies in deep learning–based classification. The proposed big data analytics engine automatically detects valuable information, extracts useful features from healthcare data, reduces the dimensionality of data, classifies patient health conditions, and predicts drug side effects. This system helps warn diabetes and BP patients before risks to their health reach a high level. In addition, this system supports physicians to offer actual treatments to their patients by smartly monitoring their patients' health conditions. This method can store and analyze healthcare big data, extract valuable features, and provide the semantic meanings of features in order to improve the performance of health condition classification. In this context, this framework can be useful in healthcare sectors for monitoring chronic patients using their structured and unstructured data.

Despite its potentials, this work has a couple of limitations. First, the SentiWordNet lexicons are not enough to precisely understand the textual data for the prediction of patient mental health and drug side effects. Therefore, it would be worth to use different sentiment lexicons for the improvement of the proposed system. Second, the proposed framework does not consider the multimodal data such as videos, emoticons, images, etc., which might be integrated with the said model to further improve the system performance. Also, the presented model can be extended taking a fuzzy ontology with fuzzy LSTM into account to enhance the classification performances in healthcare context.

Acknowledgment

This research was supported in part by the MSIT (Ministry of Science, ICT), Korea, under the ITRC (IITP-2018-2014-1-00729) supervised by the IITP, and in part by Sejong university through its faculty research program. Imran's work is supported by the Deanship of Scientific Research at King Saud University through research group project number RG-1435-051. Shaker's work supported by the Spanish Ministry of Science, Innovation and Universities (grants RTI2018-099646-B-I00, TIN2017-84796-C2-1-R, TIN2017-90773-REDT, and RED2018-102641-T) and the Galician Ministry of Education, University and Professional Training (grants ED431F 2018/02, ED431C 2018/29, ED431G/08, and ED431G2019/04), with all grants co-funded by the European Regional Development Fund (ERDF/FEDER program).

References

[1] C.W. Song, H. Jung, K. Chung, Development of a medical big-data mining process using topic modeling, Cluster Comput. (2017) 1–10. https://doi.org/10.1007/s10586-017-0942-0.

[2] J. Peral, A. Ferrandez, D. Gil, R. Munoz-Terol, H. Mora, An ontology-oriented architecture for dealing with heterogeneous data applied to telemedicine systems, IEEE Access. 6 (2018) 41118–41138. https://doi.org/10.1109/ACCESS.2018.2857499.

[3] T. Nguyen Gia, I. Ben Dhaou, M. Ali, A.M. Rahmani, T. Westerlund, P. Liljeberg, H. Tenhunen, Energy efficient fog-assisted IoT system for monitoring diabetic patients with cardiovascular disease, Futur. Gener. Comput. Syst. 93 (2019) 198–211. https://doi.org/10.1016/j.future.2018.10.029.

[4] M. Saravanan, R. Shubha, A.M. Marks, V. Iyer, SMEAD: A Secured Mobile Enabled Assisting Device for Diabetics Monitoring, 2017 IEEE Int. Conf. Adv. Networks Telecommun. Syst. (2017) 1–6.

[5] G. Alfian, M. Syafrudin, M.F. Ijaz, M.A. Syaekhoni, N.L. Fitriyani, J. Rhee, A personalized healthcare monitoring system for diabetic patients by utilizing BLE-based sensors and real-time data processing, Sensors (Switzerland). 18 (2018). https://doi.org/10.3390/s18072183.

[6] S. El-Sappagh, F. Ali, S. El-Masri, K. Kim, A. Ali, K.S. Kwak, Mobile Health Technologies for Diabetes Mellitus: Current State and Future Challenges, IEEE Access. 7 (2019) 21917–21947. https://doi.org/10.1109/ACCESS.2018.2881001.

[7] F. Ali, S.M.R. Islam, D. Kwak, P. Khan, N. Ullah, S. jo Yoo, K.S. Kwak, Type-2 fuzzy ontology–aided recommendation systems for IoT–based healthcare, Comput. Commun. 119 (2018) 138–155. https://doi.org/10.1016/j.comcom.2017.10.005.

[8] S.A. Siddiqui, Y. Zhang, J. Lloret, H. Song, Z. Obradovic, Pain-Free Blood Glucose Monitoring Using Wearable Sensors: Recent Advancements and Future Prospects, IEEE Rev. Biomed. Eng. 11 (2018) 21–35. https://doi.org/10.1109/RBME.2018.2822301.

[9] B.Y. Su, M. Enayati, K.C. Ho, M. Skubic, L. Despins, J. Keller, M. Popescu, G. Guidoboni, M. Rantz, Monitoring the Relative Blood Pressure Using a Hydraulic Bed Sensor System, IEEE Trans. Biomed. Eng. 66 (2019) 740–748. https://doi.org/10.1109/TBME.2018.2855639.

[10] T. Arakawa, Recent research and developing trends of wearable sensors for detecting blood pressure, Sensors (Switzerland). 18 (2018). https://doi.org/10.3390/s18092772.

[11] F. Ali, D. Kwak, P. Khan, S.H.A. Ei-Sappagh, S.M.R. Islam, D. Park, K.S. Kwak, Merged Ontology and SVM-Based Information Extraction and Recommendation System for Social Robots, IEEE Access. 5 (2017) 12364–12379. https://doi.org/10.1109/ACCESS.2017.2718038.

[12] M. Chen, J. Yang, J. Zhou, Y. Hao, J. Zhang, C.H. Youn, 5G-Smart Diabetes: Toward Personalized Diabetes

Diagnosis with Healthcare Big Data Clouds, IEEE Commun. Mag. 56 (2018) 16–23. https://doi.org/10.1109/MCOM.2018.1700788.

[13] A. Adeli, M. Neshat, A Fuzzy Expert System for Heart Disease Diagnosis, Proc. Int. MultiConference Engineeers Comput. Sci. I (2010) 1–6.

[14] A.R.M. Forkan, I. Khalil, A. Ibaida, Z. Tari, BDCaM: Big Data for Context-Aware Monitoring—A Personalized Knowledge Discovery Framework for Assisted Healthcare, IEEE Trans. Cloud Comput. 5 (2015) 628–641. https://doi.org/10.1109/tcc.2015.2440269.

[15] H.F. Nweke, Y.W. Teh, U.R. Alo, G. Mujtaba, Analysis of Multi-Sensor Fusion for Mobile and Wearable Sensor Based Human Activity Recognition, (2018) 22–26. https://doi.org/10.1145/3224207.3224212.

[16] G. Manogaran, R. Varatharajan, M.K. Priyan, Hybrid Recommendation System for Heart Disease Diagnosis based on Multiple Kernel Learning with Adaptive Neuro-Fuzzy Inference System, Multimed. Tools Appl. 77 (2018) 4379–4399. https://doi.org/10.1007/s11042-017-5515-y.

[17] M. Chen, W. Li, Y. Hao, Y. Qian, I. Humar, Edge cognitive computing based smart healthcare system, Futur. Gener. Comput. Syst. 86 (2018) 403–411. https://doi.org/10.1016/j.future.2018.03.054.

[18] N. Yuvaraj, K.R. SriPreethaa, Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster, Cluster Comput. (2017) 1–9. https://doi.org/10.1007/s10586-017-1532-x.

[19] S. Asthana, A. Megahed, R. Strong, A Recommendation System for Proactive Health Monitoring Using IoT and Wearable Technologies, Proc. - 2017 IEEE 6th Int. Conf. AI Mob. Serv. AIMS 2017. (2017) 14–21. https://doi.org/10.1109/AIMS.2017.11.

[20] Z.C. Lipton, D.C. Kale, C. Elkan, R. Wetzel, Learning to Diagnose with LSTM Recurrent Neural Networks, (2015) 1–18. https://doi.org/10.14722/ndss.2015.23268.

[21] K. Vimalkumar, N. Radhika, A Big Data Framework for Intrusion Detection, (2017) 198–204.

[22] S. Qingnan, M. V. Jankovic, B. Joao, B.L. Moore, P. Diem, C. Stettler, S. Mougiakakou, Predicting blood glucose with an LSTM and Bi-LSTM based deep neural network, 14th IEEE Symp. Neural Networks Appl. (2019).

[23] L. Tsao, L. Li, L. Ma, Human work and status evaluation based on wearable sensors in human factors and ergonomics: A review, IEEE Trans. Human-Machine Syst. 49 (2019) 72–84. https://doi.org/10.1109/THMS.2018.2878824.

[24] M.S. Hossain, G. Muhammad, Emotion-aware connected healthcare big data towards 5G, IEEE Internet Things J. 5 (2018) 2399–2406. https://doi.org/10.1109/JIOT.2017.2772959.

[25] F. Ahamed, F. Farid, Applying Internet of Things and Machine-Learning for Personalized Healthcare: Issues and Challenges, 2018 Int. Conf. Mach. Learn. Data Eng. (2019) 19–21. https://doi.org/10.1109/icmlde.2018.00014.

[26] R. Lopes Rosa, G. M. Schwartz, W. Vicente Ruggiero, D. Zegarra Rodriguez, A Knowledge-Based Recommendation System that includes Sentiment Analysis and Deep Learning, IEEE Trans. Ind. Informatics. 3203 (2018) 1–12. https://doi.org/10.1109/TII.2018.2867174.

[27] Y. Chen, B. Zhou, W. Zhang, W. Gong, G. Sun, Sentiment analysis based on deep learning and its application in screening for perinatal depression, Proc. - 2018 IEEE 3rd Int. Conf. Data Sci. Cyberspace, DSC 2018. (2018) 451–456. https://doi.org/10.1109/DSC.2018.00073.

[28] M. Villari, L. Carnevale, A. Celesti, A. Galletta, G. Fiumara, Applying Artificial Intelligence in Healthcare Social Networks to Identity Critical Issues in Patients' Posts, (2018) 680–687. https://doi.org/10.5220/0006750606800687.

[29] D. Bell, E. Laparra, A. Kousik, T. Ishihara, M. Surdeanu, S. Kobourov, Detecting Diabetes Risk from Social Media Activity, (2018) 1–11.

[30] J.R. Reichert, K.L. Kristensen, R.R. Mukkamala, R. Vatrapu, A supervised machine learning study of online discussion forums about type-2 diabetes, 2017 IEEE 19th Int. Conf. e-Health Networking, Appl. Serv. Heal. 2017. 2017-Decem (2017) 1–7. https://doi.org/10.1109/HealthCom.2017.8210815.

[31] E. Tutubalina, Z. Miftahutdinov, S. Nikolenko, V. Malykh, Medical concept normalization in social media posts with recurrent neural networks, J. Biomed. Inform. 84 (2018) 93–102. https://doi.org/10.1016/j.jbi.2018.06.006.

[32] F. Gräßer, S. Kallumadi, H. Malberg, S. Zaunseder, Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning, (2018) 121–125. https://doi.org/10.1145/3194658.3194677.

[33] J. Liu, X. Jiang, Q. Chen, M. Song, J. Li, Adverse drug reaction related post detecting using sentiment feature, Iran. J. Public Health. 47 (2018) 861–867. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85048895897&partnerID=40&md5=e603602caba00dab4ffdbfa1c85b8e32.

[34] M.D.P. Salas-Zárate, J. Medina-Moreira, K. Lagos-Ortiz, H. Luna-Aveiga, M.Á. Rodríguez-García, R. Valencia-García, Sentiment Analysis on Tweets about Diabetes: An Aspect-Level Approach, Comput. Math. Methods Med. 2017 (2017). https://doi.org/10.1155/2017/5140631.

[35] M. Chandrashekar, R. Nagulapati, Y. Lee, Ontology Mapping Framework with Feature Extraction and Semantic Embeddings, Proc. - 2018 IEEE Int. Conf. Healthc. Informatics Work. ICHI-W 2018. (2018) 34–42. https://doi.org/10.1109/ICHI-W.2018.00012.

[36] R. Maldonado, T.R. Goodwin, M.A. Skinner, S.M. Harabagiu, Deep Learning Meets Biomedical Ontologies: Knowledge Embeddings for Epilepsy., AMIA ... Annu. Symp. Proceedings. AMIA Symp. 2017 (2017) 1233–1242. http://www.ncbi.nlm.nih.gov/pubmed/29854192%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5977726.

[37] B. Pang, L. Lee, S.C.-1118704 D.O.-10. 3115/1118693. 111870. Vaithyanathan, Thumbs up?: sentiment classification using machine learning techniques, Proc. ACL-02 Conf. Empir. Methods Nat. Lang. Process. - Vol. 10. (2002) 79–86.

[38] A. Lamurias, L.A. Clarke, F.M. Couto, BO-LSTM: Classifying relations via long short-term memory networks along biomedical ontologies, BioRxiv. (2018) 336719. https://doi.org/10.1101/336719.

[39] V. Jagadeeswari, V. Subramaniyaswamy, R. Logesh, V. Vijayakumar, A study on medical Internet of Things and Big Data in personalized healthcare system, Heal. Inf. Sci. Syst. 6 (2018) 1–20. https://doi.org/10.1007/s13755-018-0049-x.

[40] A. Onal, O.B. Sezer, M. Ozbayoglu, E. Dogdu, A. Can Onal, Weather data analysis and sensor fault detection using an extended IoT framework with semantics, big data, and machine learning SpEnD Project View project A Multi-Agent Simulation and Backtesting Workbench Software for Deep Learning and Evolutionary Algor, (2017). https://doi.org/10.1109/BigData.2017.8258150.

[41] J.C. Kim, K. Chung, Mining health-risk factors using PHR similarity in a hybrid P2P network, Peer-to-Peer Netw. Appl. 11 (2018) 1278–1287. https://doi.org/10.1007/s12083-018-0631-7.

[42] S. Chakrabarti, A. Swetapadma, P.K. Pattnaik, Smart Techniques for a Smarter Planet, Springer International Publishing, 2019. https://doi.org/10.1007/978-3-030-03131-2.

[43] M.I. Razzak, M. Imran, G. Xu, Big data analytics for preventive medicine, Springer London, 2020. https://doi.org/10.1007/s00521-019-04095-y.

[44] F. Amalina, I.A. Targio Hashem, Z.H. Azizul, A.T. Fong, A. Firdaus, M. Imran, N.B. Anuar, Blending Big Data Analytics: Review on Challenges and a Recent Study, IEEE Access. 8 (2020) 3629–3645. https://doi.org/10.1109/ACCESS.2019.2923270.

[45] P.K. Sahoo, S.K. Mohapatra, S.L. Wu, SLA based healthcare big data analysis and computing in cloud network, J. Parallel Distrib. Comput. 119 (2018) 121–135. https://doi.org/10.1016/j.jpdc.2018.04.006.

[46] K. Muhammad, A.K. Sangaiah, A. Abdelaziz, M. Elhoseny, A.S. Salama, A.M. Riad, A hybrid model of Internet of Things and cloud computing to manage big data in health services applications, Futur. Gener. Comput. Syst. 86 (2018) 1383–1394. https://doi.org/10.1016/j.future.2018.03.005.

[47] N. El aboudi, L. Benhlima, Big Data Management for Healthcare Systems: Architecture, Requirements, and Implementation, Adv. Bioinformatics. 2018 (2018) 1–10. https://doi.org/10.1155/2018/4059018.

[48] A. Jindal, A. Dua, N. Kumar, A.K. Das, A. V. Vasilakos, J.J.P.C. Rodrigues, Providing Healthcare-as-a-Service Using Fuzzy Rule Based Big Data Analytics in Cloud Computing, IEEE J. Biomed. Heal. Informatics. 22 (2018) 1605–1618. https://doi.org/10.1109/JBHI.2018.2799198.

[49] H. Habibzadeh, A. Boggio-Dandry, Z. Qin, T. Soyata, B. Kantarci, H.T. Mouftah, Soft Sensing in Smart Cities: Handling 3Vs Using Recommender Systems, Machine Intelligence, and Data Analytics, IEEE Commun. Mag. 56 (2018) 78–86. https://doi.org/10.1109/MCOM.2018.1700304.

[50] P. Verma, S.K. Sood, Cloud-centric IoT based disease diagnosis healthcare framework, J. Parallel Distrib. Comput. 116 (2018) 27–38. https://doi.org/10.1016/j.jpdc.2017.11.018.

[51] T. Lenc, P.E. Keller, M. Varlet, S. Nozaradan, Frequency tagging is sensitive to the temporal structure of signals, 9264 (2017) 17–24. https://doi.org/10.5281/zenodo.

[52] A. Dwivedi, R.P. Pant, S. Pandey, K. Kumar, Internet of Things' (IoT's) Impact on Decision Oriented Applications of Big Data Sentiment Analysis, Proc. - 2018 3rd Int. Conf. Internet Things Smart Innov. Usages, IoT-SIU 2018. (2018) 1–10. https://doi.org/10.1109/IoT-SIU.2018.8519922.

[53] V. Vijayakumar, K. Shankar, S.K. Lakshmanaprabu, N. Chilamkurti, M. Ilayaraja, A.W. Nasir, Random forest for

big data classification in the internet of things using optimal features, Int. J. Mach. Learn. Cybern. 0 (2019) 0. https://doi.org/10.1007/s13042-018-00916-z.

[54] G. Manogaran, R. Varatharajan, D. Lopez, P.M. Kumar, R. Sundarasekar, C. Thota, A new architecture of Internet of Things and big data ecosystem for secured smart healthcare monitoring and alerting system, Futur. Gener. Comput. Syst. 82 (2018) 375–387. https://doi.org/10.1016/j.future.2017.10.045.

[55] F. Ali, D. Kwak, P. Khan, S. El-Sappagh, A. Ali, S. Ullah, K.H. Kim, K.-S. Kwak, Transportation sentiment analysis using word embedding and ontology-based topic modeling, Knowledge-Based Syst. (2019). https://doi.org/10.1016/j.knosys.2019.02.033.

[56] A. Teixeira, Data extraction and preparation to perform a The example of a Facebook fashion brand page, (n.d.).

[57] F. Ali, D. Kwak, P. Khan, S.M.R. Islam, K.H. Kim, K.S. Kwak, Fuzzy ontology-based sentiment analysis of transportation and city feature reviews for safe traveling, Transp. Res. Part C Emerg. Technol. 77 (2017) 33–48. https://doi.org/10.1016/j.trc.2017.01.014.

[58] F. Ali, S. El-Sappagh, D. Kwak, Fuzzy ontology and LSTM-based text mining: A transportation network monitoring system for assisting travel, Sensors (Switzerland). 19 (2019). https://doi.org/10.3390/s19020234.

[59] H. Htet, S.S. Khaing, Y.Y. Myint, Big Data Analysis and Deep Learning Applications, Springer Singapore, 2019. https://doi.org/10.1007/978-981-13-0869-7.

[60] S. Din, A. Paul, Smart health monitoring and management system: Toward autonomous wearable sensing for Internet of Things using big data analytics, Futur. Gener. Comput. Syst. 91 (2019) 611–619. https://doi.org/10.1016/j.future.2017.12.059.

[61] F. Ali, E.K. Kim, Y.G. Kim, Type-2 fuzzy ontology-based opinion mining and information extraction: A proposal to automate the hotel reservation system, Appl. Intell. 42 (2015) 481–500. https://doi.org/10.1007/s10489-014-0609-y.

[62] S. Baccianella, A. Esuli, F. Sebastiani, SentiWordNet 3 . 0 : An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining SentiWordNet, Analysis. 0 (2010) 1–12. https://doi.org/10.1.1.61.7217.

[63] D.C. Cavalcanti, R.B.C. Prudêncio, S.S. Pradhan, J.Y. Shah, R.S. Pietrobon, Good to be bad? Distinguishing between positive and negative citations in scientific impact, Proc. - Int. Conf. Tools with Artif. Intell. ICTAI. (2011) 156–162. https://doi.org/10.1109/ICTAI.2011.32.

[64] F. Ali, S. Ei-Sappagh, P. Khan, K.S. Kwak, Feature-based Transportation Sentiment Analysis Using Fuzzy Ontology and SentiWordNet, 9th Int. Conf. Inf. Commun. Technol. Converg. ICT Converg. Powered by Smart Intell. ICTC 2018. (2018) 1350–1355. https://doi.org/10.1109/ICTC.2018.8539607.

[65] F. Ali, S. Ei-sappagh, L. Feng, K.S. Kwak, ONEMLI! - Word2vec and LSTM-based Offensive Content Detection, (2019) 1480–1481.

[66] F. Vitali, R. Lombardo, D. Rivero, F. Mattivi, P. Franceschi, A. Bordoni, A. Trimigno, F. Capozzi, G. Felici, F. Taglino, F. Miglietta, N. De Cock, C. Lachat, B. De Baets, G. De Tré, M. Pinart, K. Nimptsch, T. Pischon, J. Bouwman, D. Cavalieri, ONS: An ontology for a standardized description of interventions and observational studies in nutrition, Genes Nutr. 13 (2018) 1–9. https://doi.org/10.1186/s12263-018-0601-y.

[67] W.A. Kibbe, C. Arze, V. Felix, E. Mitraka, E. Bolton, G. Fu, C.J. Mungall, J.X. Binder, J. Malone, D. Vasant, H. Parkinson, L.M. Schriml, Disease Ontology 2015 update: An expanded and updated database of Human diseases for linking biomedical knowledge through disease data, Nucleic Acids Res. 43 (2015) D1071–D1078. https://doi.org/10.1093/nar/gku1011.

[68] S. El-Sappagh, D. Kwak, F. Ali, K.S. Kwak, DMTO: A realistic ontology for standard diabetes mellitus treatment, J. Biomed. Semantics. 9 (2018) 1–30. https://doi.org/10.1186/s13326-018-0176-y.

[69] Y. Lin, S. Mehta, H. Küçük-McGinty, J.P. Turner, D. Vidovic, M. Forlin, A. Koleti, D.T. Nguyen, L.J. Jensen, R. Guha, S.L. Mathias, O. Ursu, V. Stathias, J. Duan, N. Nabizadeh, C. Chung, C. Mader, U. Visser, J.J. Yang, C.G. Bologa, T.I. Oprea, S.C. Schürer, Drug target ontology to classify and integrate drug discovery data, J. Biomed. Semantics. 8 (2017) 1–16. https://doi.org/10.1186/s13326-017-0161-x.

[70] M.Á. Rodríguez-García, R. Valencia-García, F. García-Sánchez, J.J. Samper-Zapater, Ontology-based annotation and retrieval of services in the cloud, Knowledge-Based Syst. 56 (2014) 15–25. https://doi.org/10.1016/j.knosys.2013.10.006.

[71] C. Verma, M. Hart, S. Bhatkar, A. Parker-Wood, S. Dey, Improving Scalability of Personalized Recommendation Systems for Enterprise Knowledge Workers, IEEE Access. 4 (2016) 204–215. https://doi.org/10.1109/ACCESS.2015.2513000.

[72] A. Derungs, S. Soller, A. Weishaupl, J. Bleuel, G. Berschin, O. Amft, Regression-based, mistake-driven movement

skill estimation in Nordic Walking using wearable inertial sensors, 2018 IEEE Int. Conf. Pervasive Comput. Commun. PerCom 2018. (2018). https://doi.org/10.1109/PERCOM.2018.8444576.

[73] H. Uğuz, A hybrid system based on information gain and principal component analysis for the classification of transcranial Doppler signals, Comput. Methods Programs Biomed. 107 (2011) 598–609. https://doi.org/10.1016/j.cmpb.2011.03.013.

[74] M. Hall, E. Frank, G. Holmes, P. Bernhard, P. Reutemann, I. Witten, The WEKA Data Mining Software: An Update, 11 (2009) 10–18.

[75] G. Rehman, D. Khan, A. Hussain, W. Ahmad, M. Hamayun, A. Ahmad, S. Khan, A. Iqbal, U.U. Khan, L. Huang, Intelligent hepatitis diagnosis using adaptive neuro-fuzzy inference system and information gain method, Soft Comput. (2018). https://doi.org/10.1007/s00500-018-3643-6.

[76] S. Balli, E.A. Sağbaş, M. Peker, Human activity recognition from smart watch sensor data using a hybrid of principal component analysis and random forest algorithm, Meas. Control (United Kingdom). 52 (2018) 37–45. https://doi.org/10.1177/0020294018813692.

[77] W. Yin, K. Kann, M. Yu, H. Schütze, Comparative Study of CNN and RNN for Natural Language Processing, (2017). https://doi.org/10.14569/IJACSA.2017.080657.

[78] B.T. Pham, D. Tien Bui, I. Prakash, M.B. Dholakia, Hybrid integration of Multilayer Perceptron Neural Networks and machine learning ensembles for landslide susceptibility assessment at Himalayan area (India) using GIS, Catena. 149 (2017) 52–63. https://doi.org/10.1016/j.catena.2016.09.007.

[79] S. Kim, D. Jin, H. Lee, Predicting drug-target interactions using drug-drug interactions, PLoS One. 8 (2013) 1–12. https://doi.org/10.1371/journal.pone.0080129.

[80] F. Ali, E.-S. Shaker, A. Ali, K.S. Kwak, D. Kwak, Sentiment analysis of transportation using word embedding and LDA approaches, N.D. (2018) 1111–1112.

[81] J.W. Smith, J.E. Everhart, W.C. Dicksont, W.C. Knowler, R.S. Johannes, Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus, Proc. Annu. Symp. Comput. Appl. Med. Care. Am. Med. Informatics Assoc. (1988) 261–265.

[82] C.M. Cabello, W.B. Bair, S.D. Lamore, S. Ley, S. Alexandra, S. Azimian, G.T. Wondrak, NIH Public Access, 46 (2010) 220–231. https://doi.org/10.1016/j.freeradbiomed.2008.10.025.The.

[83] F. Ali, K.S. Kwak, Y.G. Kim, Opinion mining based on fuzzy domain ontology and Support Vector Machine: A proposal to automate online review classification, Appl. Soft Comput. J. 47 (2016) 235–250. https://doi.org/10.1016/j.asoc.2016.06.003.

[84] G. Stewart, A. Kamata, R. Miles, E. Grandoit, F. Mandelbaum, C. Quinn, L. Rabin, Predicting mental health help seeking orientations among diverse Undergraduates: An ordinal logistic regression analysis☆, J. Affect. Disord. 257 (2019) 271–280. https://doi.org/10.1016/j.jad.2019.07.058.

[85] T. Young, D. Hazarika, S. Poria, E. Cambria, Recent trends in deep learning based natural language processing [Review Article], IEEE Comput. Intell. Mag. 13 (2018) 55–75. https://doi.org/10.1109/MCI.2018.2840738.

[86] S. Poria, E. Cambria, A. Gelbukh, Aspect extraction for opinion mining with a deep convolutional neural network, Knowledge-Based Syst. 108 (2016) 42–49. https://doi.org/10.1016/j.knosys.2016.06.009.

[87] M. Porumb, S. Stranges, A. Pescapè, L. Pecchia, Precision Medicine and Artificial Intelligence: A Pilot Study on Deep Learning for Hypoglycemic Events Detection based on ECG, Sci. Rep. 10 (2020) 1–17. https://doi.org/10.1038/s41598-019-56927-5.

[88] G. Hussain, M.K. Maheshwari, M.L. Memon, M.S. Jabbar, K. Javed, A CNN Based Automated Activity and Food Recognition Using Wearable Sensor for Preventive Healthcare, Electronics. 8 (2019) 1425. https://doi.org/10.3390/electronics8121425.

[89] T. Moon, S. Hong, H.Y. Choi, D.H. Jung, S.H. Chang, J.E. Son, Interpolation of greenhouse environment data using multilayer perceptron, Comput. Electron. Agric. 166 (2019) 105023. https://doi.org/10.1016/j.compag.2019.105023.

[90] L.J.B. Caluza, Fuzzy Unordered Rule Induction Algorithm Application Basic Programming Language Competence: A Rule-Based Model, Indian J. Sci. Technol. 12 (2019) 1–10. https://doi.org/10.17485/ijst/2019/v12i12/142575.

Author photo

Farman Ali

Shaker El-Sappagh

S.M. Riazul Islam

Amjad Ali

Muhammad Attique

Muhammad Imran

Kyung-Sup Kwak

- Smartphones, wearable sensors, and social networks provide a new approach to collect patient data.
- Continuous patient monitoring generates a large amount of unstructured healthcare data.
- Existing approaches cannot deal with huge amounts of healthcare data extracted from various sources.
- Traditional ML techniques are unable to handle extracted healthcare data for abnormality prediction.
- A big data analytics engine is proposed to precisely analyze different sources of healthcare data.

Author Biographies

Farman Ali received the B.S. degree in computer science from the University of Peshawar, Pakistan, in 2011, the M.S. degree in computer science engineering from Gyeongsang National University, South Korea, in 2015, and the Ph.D. degree in information and communication engineering from Inha University, South Korea, in 2018, where he was a Post-Doctoral Fellow with the UWB Wireless Communications Research Center from September 2018 to August 2019. He is currently an Assistant Professor with the Department of Software, Sejong University. His current research interests include data mining, deep learning, sentiment analysis, recommendation systems, healthcare monitoring systems, ontology, information extraction, information retrieval, and fuzzy logic.

Shaker El-Sappagh was born in El-Behara, Egypt, in 1977. He received the bachelor degree in computer science from Information Systems Department, Faculty of Computers and Information, Cairo University, Cairo, Egypt, in 1997, and the master degree from the same university in 2007. He received the Ph.D. degrees in computer science from Information Systems Department, Faculty of Computers and Information, Mansura University, Mansura, Egypt in 2015. In 2003, he joined the Department of Information Systems, Faculty of Computers and Information, Benha University, Banha, Egypt as a teaching assistant. In 2009, he joined the Collage of Science, King Saud University as a lecturer. Since June 2016, he has been with the Department of Information Systems, Faculty of computers and Information, Benha University as a lecturer. He has publications in clinical decision support systems and semantic intelligence. His current research interests include medical informatics, ontology engineering, distributed and hybrid clinical decision support systems, semantic data modeling, distributed database systems, big data, semantic query languages, medical data encoding, medical terminology, semantic interoperability, description logic, fuzzy logic, fuzzy mathematics, fuzzy database, semantic database, cloud computing, data integration, semantic web, and fuzzy expert systems. Dr. El-Sappagh is a reviewer in many journals, and he is very interested in the diseases diagnosis and treatment researches. He has built some publicly available ontologies including diabetes diagnosis ontology which is publicly available in BioPortal site at *https://bioportal.bioontology.org/ontologies/DDO*, SNOMED CT OWL 2 ontology at *https://bioportal.bioontology.org/ontologies/SCTTO*, and DMTO diabetes treatment OWL 2 ontology available at *https://bioportal.bioontology.org/ontologies/DMTO*.

S.M. Riazul Islam (M'10) received the B.S. and M.S. degrees in Applied Physics and Electronics from University of Dhaka, Bangladesh in 2003, and 2005, respectively and the Ph.D. degree in Information and Communication Engineering from Inha University, South Korea in 2012. He has been working at Sejong University, south Korea as an Assistant Professor at the Department of Computer Science and Engineering since March 2017. From 2014 to 2017, he worked at Inha University, South Korea as a Research Professor at the UWB Wireless Communications Research Center. Dr. Islam was with the University of Dhaka, Bangladesh as an Assistant Professor and Lecturer at the Dept. of Electrical and Electronic Engineering (formerly Dept. of Applied Physics, Electronics & Communication Engineering) for the period September 2005 to March 2014. In 2014, In 2014, he worked at the Samsung R&D Institute Bangladesh (SRBD) as a Chief Engineer at the Dept. of Solution Lab for six months. His research interests include wireless communications, signal processing for communications, and enabling technologies for 5G and beyond.

Amjad Ali received the B.S. and M.S. degrees in computer science from the COMSATS Institute of Information Technology, Pakistan, in 2006 and 2008, respectively, and the Ph.D. degree from the Electronics and Radio Engineering Department, Kyung Hee University, South Korea, in June 2015. Since July 2015, he has been an Assistant Professor with the Department of Computer Science, COMSATS University Islamabad, Lahore Campus, Pakistan. From 2018 to 2019, he was a Post-Doctoral Research Scientist with the UWB Wireless Communications Research Center (formerly Key National IT Research Center), Department of Information and Communication Engineering, Inha University, South Korea, and also with the Mobile Network and Communications Lab, School of Electrical Engineering, Korea University, Anam-dong, Seoul, South Korea. His main research interests include multimedia transmission, cognitive radio networks, manchine learning, Internet of Things, smart grid, and vehicular networks. He was a recipient of the Excellent Research Award from the UWB Wireless Communications Research Center and from the Mobile Network and Communications Lab during his Post-doctorate studies.

Muhammad Attique received the bachelor's degree in information and communication systems engineering from the National University of Science and Technology, Pakistan, in 2008, and the Ph.D. degree from Ajou University, South Korea, in 2017. He is currently an Assistant Professor with the Department of Software, Sejong University, South Korea. His research interests include location-based services, spatial queries in the road networks, and big data analysis.

Muhammad Imran is an Associate Professor in the College of Applied Computer Science at King Saud University, Saudi Arabia. He received a Ph. D in Information Technology from the University Teknologi PETRONAS, Malaysia in 2011. His research interest includes Mobile and Wireless Networks, Internet of Things, Big Data Analytics, Cloud computing, and Information Security. His research is financially supported by several national and international grants. He has completed a number of international collaborative research projects with reputable universities. He has published more than 250 research articles in peer-reviewed, well-recognized international conferences and journals. Many of his research articles are among the highly cited and most downloaded. He served as an Editor in Chief for European Alliance for Innovation (EAI) Transactions on Pervasive Health and Technology. He is serving as an associate editor for top ranked international journals such as IEEE Communications Magazine, IEEE Network, Future Generation Computer Systems, and IEEE Access. He served/serving as a guest editor for about two dozen special issues in journals such as IEEE Communications Magazine, IEEE Wireless Communications Magazine, Future Generation Computer Systems, IEEE Access, and Computer Networks. He has been involved in about one hundred peer-reviewed international conferences and workshops in various capacities such as a chair, co-chair and technical program committee member. He has been consecutively awarded with **Outstanding Associate Editor of IEEE Access** in 2018 and 2019 besides many others.

Kyung-Sup Kwak (M'81) received the Ph.D. degree from the University of California at San Diego in 1988. From 1988 to 1989, he was with Hughes Network Systems, San Diego, CA, USA. From 1989 to 1990, he was with the IBM Network Analysis Center, Research Triangle Park, NC, USA. Since then, he has been with the School of Information and Communication Engineering, Inha University, South Korea, as a Professor, where he had been the Dean of the Graduate School of Information Technology and Telecommunications from 2001 to 2002. He has been the Director of the UWB Wireless Communications Research Center (formerly Key National IT Research Center), South Korea, since 2003. In

2006, he served as the President of Korean Institute of Communication Sciences, and in 2009, the President of Korea Institute of Intelligent Transport Systems. In 2008, he had been selected for Inha Fellow Professor and now for Inha Hanlim Fellow Professor. Dr. Kwak published more than 200 peer-reviewed journal papers and served as TPC/Track chairs/organizing chairs for several IEEE related conferences. His research interests include wireless communications, UWB systems, sensor networks, WBAN, and nano communications. He was a recipient of the number of awards, including the Engineering College Achievement Award from Inha University, the LG Paper Award, the Motorola Paper Award, the Haedong Prize of research, and various government awards from the Ministry of ICT, the President, and the Prime Minister of Korea, for his excellent research performances

Author Statement

Farman Ali: Conceptualization, Methodology, Software, Visualization, Writing-Original Draft. **Shaker El-Sappagh:** Resources, Validation, Investigation**.: S.M. Riazul Islam:** Writing- Review and Editing**.: Amjad Ali:** Formal analysis, Software**.: Muhammad Attique:** Data Curation, Resources**.: Muhammad Imran:** Writing-Review and Editing.**: Kyung-Sup Kwak:** Funding acquisition, Project administration, Supervision.

Conflict of interest

The authors of this manuscript declare no conflicts of interest.