Prediction Model of User Physical Activity using Data Characteristics-based Long Short-term Memory Recurrent Neural Networks

Joo-Chang Kim[1], Kyungyong Chung[2*]
[1]Data Mining Lab., Department of Computer Science, Kyonggi University
154-42, Gwanggyosan-ro, Yeongtong-gu, Suwon-si, Gyeonggi-do, 16227, South Korea
[e-mail: kjc2232@naver.com]
[2]Division of Computer Science and Engineering, Kyonggi University
154-42, Gwanggyosan-ro, Yeongtong-gu, Suwon-si, Gyeonggi-do, 16227, South Korea
[e-mail: dragonhci@gmail.com]
*Corresponding author: Kyungyong Chung

Received June 26, 2018; revised October 10, 2018; accepted October 28, 2018; published April 30, 2019

Abstract

Recently, mobile healthcare services have attracted significant attention because of the emerging development and supply of diverse wearable devices. Smartwatches and health bands are the most common type of mobile-based wearable devices and their market size is increasing considerably. However, simple value comparisons based on accumulated data have revealed certain problems, such as the standardized nature of health management and the lack of personalized health management service models. The convergence of information technology (IT) and biotechnology (BT) has shifted the medical paradigm from continuous health management and disease prevention to the development of a system that can be used to provide ground-based medical services regardless of the user's location. Moreover, the IT–BT convergence has necessitated the development of lifestyle improvement models and services that utilize big data analysis and machine learning to provide mobile healthcare-based personal health management and disease prevention information. Users' health data, which are specific as they change over time, are collected by different means according to the users' lifestyle and surrounding circumstances. In this paper, we propose a prediction model of user physical activity that uses data characteristics-based long short-term memory (DC-LSTM) recurrent neural networks (RNNs). To provide personalized services, the characteristics and surrounding circumstances of data collectable from mobile host devices were considered in the selection of variables for the model. The data characteristics considered were ease of collection, which represents whether or not variables are collectable, and frequency of occurrence, which represents whether or not changes made to input values constitute

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. NRF-2016R1D1A1A09917313)

significant variables in terms of activity. The variables selected for providing personalized services were activity, weather, temperature, mean daily temperature, humidity, UV, fine dust, asthma and lung disease probability index, skin disease probability index, cadence, travel distance, mean heart rate, and sleep hours. The selected variables were classified according to the data characteristics. To predict activity, an LSTM RNN was built that uses the classified variables as input data and learns the dynamic characteristics of time series data. LSTM RNNs resolve the vanishing gradient problem that occurs in existing RNNs. They are classified into three different types according to data characteristics and constructed through connections among the LSTMs. The constructed neural network learns training data and predicts user activity. To evaluate the proposed model, the root mean square error (RMSE) was used in the performance evaluation of the user physical activity prediction method for which an autoregressive integrated moving average (ARIMA) model, a convolutional neural network (CNN), and an RNN were used. The results show that the proposed DC-LSTM RNN method yields an excellent mean RMSE value of 0.616. The proposed method is used for predicting significant activity considering the surrounding circumstances and user status utilizing the existing standardized activity prediction services. It can also be used to predict user physical activity and provide personalized healthcare based on the data collectable from mobile host devices.

Keywords: Data Mining, Neural Networks, LSTM, Prediction, Mobile Healthcare

1. Introduction

The number of degenerative/metabolic diseases has been increasing owing to aging, urbanization, and lifestyle changes. This increase has resulted in growing attention to health management in society and individuals. Accordingly, the smart healthcare industry, which specializes in providing continuous healthcare in everyday life, has attracted public interest. The healthcare big data linkage platform currently provides users with valid information through the collection and analysis of series of data related to users' health by means of ambient sensors. Smart health services provide diagnosis/prevention-based medical services in the field of disease treatment-based medicine [1]. Recently, precision medicine focusing on personal diagnosis/disease prevention has attracted public interest. Meanwhile, the healthcare industry is developing personalized health management and lifestyle improvement models and services with the aim of providing universal precision medicine [2]. Biosensors, networks, and knowledge bases were investigated for collecting and analyzing data, such as users' status, surrounding environment, and surrounding circumstances [3]. A technology featuring integrated data collection is being developed in the field of biosensors to allow the transition from a single-modal system capable of using one sensor to collect one object data to a multi-modal system capable of collecting diverse data through ambient sensor networks [4]. In the network for the multi-modal system, the data simultaneously collected with an ambient sensor in the same environment are used to learn the associated relation among multi-modal systems. Based on this process, a deep learning network capable of extracting the representation shared among modes from the multi-modal input is constructed. Deep learning-based artificial intelligence (AI) has limitations, such as a lack of information in

single-modal systems, lack of effective convergence among multi-modal input information, and the knowledge bottleneck phenomenon. These problems can be resolved by devising a multi-modal machine learning method where shared representation and artificial intelligence AI factors extracted based on modal input information are converged [5]. In the field of networking, a common data model was investigated that uses multi-modal-based knowledge expression, acquisition, and reasoning in a mesh-form hybrid peer-to-peer (P2P) network system based on the host device [6]. This is a dynamic structure that is advantageous in terms of sharing and distributed processing of data based on diverse connectivities and is used to integrate data collected from wearable devices. The structure consists of users, servers, and gateways and is capable of efficient data management and collection through diverse routes [7].

Diverse technologies capable of collecting, supplying, processing, and analyzing large-scale data received from multi-modal sensors, mobile devices, RSS, and XML are being developed in the field of knowledge bases. In the current healthcare industry, data quantified through connections between electronic medical records (EMRs), personal health devices (PHDs), and personal health records (PHRs) are being integrated and preprocessed [8–10]. The big data in healthcare generated from mobile-based wearable devices are collected by different means according to the characteristics of the structured and non-structured data. A deep learning method capable of producing knowledge learned based on machine learning is being used to acquire logic-based knowledge and expand the ontology and logic knowledge bases. This method allows the extraction and expansion of significant knowledge through the acquisition of knowledge from a knowledge base and then the refinement of this knowledge. Accordingly, a technology must be developed that is capable of integrating the rapidly increasing healthcare big data and lifelogs and integrating and processing heterogeneous big data that are closely related to health, such as nutrition, environment, and meteorological data. In addition, given the rapid increase in the number of patients per medical worker due to overpopulation resulting from urbanization, it is necessary to universalize medical services through the supply of smart health. Currently, the universal wearable mobile-based devices are the smartwatch and health band, and the market size of these devices is expanding continuously. However, the integration of the standards or specifications for managing collected data remains to be achieved. In addition, as the mobile-based wearable devices lack an understanding of the users' tendency, purpose, use, and behavioral changes, they provide only standardized health management services. This issue can be resolved through machine learning that mimics the high-dimensional cognitive skills of humans, active machine learning based on knowledge of the real world, multi-modal-based knowledge expression, and acquisition and reasoning technology.

This paper is structured as follows. Section 2 discusses mobile healthcare provided by means of wearable sensors. Section 3 presents the proposed prediction model of user physical activity that implements a data characteristics-based long short-term memory (DC-LSTM) recurrent neural network (RNN). Section 4 describes the performance evaluation and results. Section 5 concludes the paper.

2. Mobile Healthcare Provided by Means of Wearable Sensors

The sensors of health bands, smart caps, and ECG-measuring smartwear are used to collect context information, without the user's awareness, on a mobile health platform regardless of

time and the user's location. Wearable sensors allow the collection, recording, and storage of lifelog data, such as time, place, location, movement, biosignals, and calories.

The sensors of previously developed health bands that are worn on the wrist and frequently used by individuals while exercising [1] collect the user's biosignals and provide the user with his/her context-based health status. In addition, the smart cap, which is worn in everyday life [8], collects location-based context information and provides five types of health-weather indices and eight types of life-weather indices for every geographical region through a sensor attached to it. Weather indices can be used to provide services according to a certain location and to allow the user to actively prepare his or her physical body for climate changes. Moreover, phased context awareness-based cautions according to the current location of the user, detected through a GPS, can be provided. The level of the influence of meteorological elements on health or the level of the possibility of an adverse weather occurrence, displayed as an index that refers to the probability of such an occurrence under specific conditions, is quantified using a prediction model developed based on meteorological data. The public data portal [11], which contains actualized index information per context, provides an open API based on weather index information that is applicable to everyday life and healthcare.

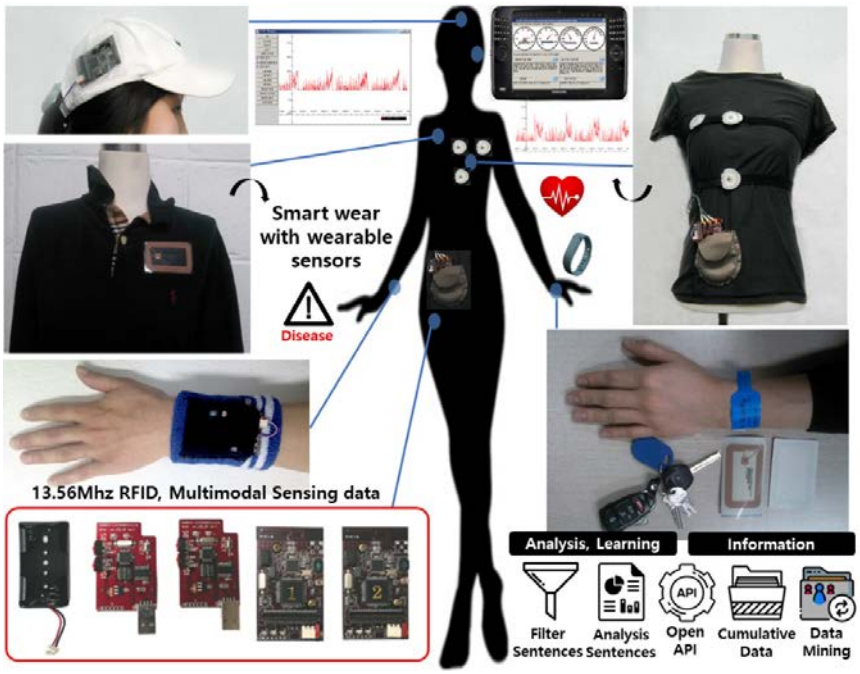

Fig. 1. Smartwear with a wearable sensor

Wearable sensors are capable of measuring body temperature, humidity, illumination intensity, temperature, and UV and, using Bluetooth communication technology, can transfer these data from mobile devices featuring GPS reception. These sensors use 2.4 GHz ZigBee, Atmega 128L, on TinyOS 2.X to share Bluetooth communication with mobile devices featuring GPS reception for the transfer of data packets. In the research conducted by Kim et al. [7] and Chung et al. [1], a wearable sensor was used for context recognition and healthcare service provision. An ontology-based context information model can be created using the OWL reasoning process based on the Jena API [12]. A knowledge base can be constructed based on service reasoning, health reasoning, and context reasoning rules, and the context per

user can be recognized to establish an evolutionary reasoning rule. The data packet consists of packet values starting with 7E and ending with 7E for serial communication with wireless sensor networks and can be expressed in the form of hexadecimal numbers based on the consecutive numerical values indicating body temperature, humidity, illumination intensity, and UV. The packets received from a serial port can be expressed as "7E 46 04 FF FF 03 0A 00 00 08 00 03 1B 00 38 00 05 00 70 00 72 D4 7E." In addition, a wearable sensor can be manufactured that can be attached to and detached from body regions, thus causing no hindrance to the user's movements, which resolves the structural problems of complicated smartwear. The really simple syndication (RSS) based weather data [13] from the Korea Meteorological Administration and the data received through IEEE 802.15.4 Standard Wireless Transfer can be analyzed according to the location in a region to provide a health-weather index. Given that meteorological and context information varies and users' health status changes fluidly according to their current location, a GPS receiving module is used to provide adequate services in real-time. RSS-based meteorological data are forms of data established based on XML and JSON, the two switched data standards, and provide updated information every three hours using the interface standard REST(Get) method. To collect meteorological data, queries are used to save data in the form of an index and a display, and a Document Object Model (DOM) parser is used to extract, exchange, and process data into an XML format [14]. This method expresses contents and structures as objects and provides a standard interface that can be managed by the user. **Fig. 1** shows smartwear with an attached wearable sensor, including a health band [15], smart cap [10], and ECG measuring smart wear [16] previously developed on a mobile health platform.

Chung et al. [1] developed a user-adaptive decision-making simulation for which smartwear to which a sensor was attached and a meteorological WebBot were implemented. In their research study, emotions that change according to meteorological elements were analyzed and applied to decision making. In a previous research study, ECG measuring smartwear was developed that uses cardiac biosignal data to monitor heart rate variability (HRV) in real-time [16]. Because it requires an attached wearable sensor and a circuit for transferring the ECG and heart rate (HR) signal data, it was manufactured in the form of a small attachable pocket. In addition, the design of the smartwear considered the electrode's position and volume, the wearer's movements, battery, clothing pressure, and clothing type. To provide smartwear capable of producing stable ECG measurements, an electrode was fixed to a highly elastic band in the form of embroidery, which allowed stable contact of the smartwear and biosensor with the body, and thus, the ECG and HR signal data could be analyzed accurately. This ECG measuring smartwear was worn by a subject and the ECG/HR data were collected in the 1–100 Hz frequency band. To examine the ECG waveform according to the user status, the power spectrum was analyzed using the low frequency/high frequency ratio. The time series analysis according to time was conducted by limiting the size of the window within the ECG data. A peak detection algorithm was used to calculate the R–R value, and Fourier transform was used to analyze low frequencies. **Fig. 2** shows a mobile health service in which wearable sensors are implemented. The figure shows the HRV-based stress index service [6], chatbot-based mobile healthcare [17], and dietary nutrition recommendation service developed on a mobile health platform for the management of obesity among teenagers [18].

In the research study conducted by Yoo et al. [1], HRV was used to analyze the activity of sympathetic and parasympathetic nerves on a mobile health platform to determine users' stress levels. According to the input biosignal data, the HRV-based frequency domain can be divided into a positive status and negative status based on the negative feedback provided from the sympathetic and parasympathetic nerves of the user. Using this process, users can monitor

their cardiac impulse and ECG analysis in real-time and set their health management and exercise based on this analysis. In addition, this process is valuable when used to monitor the status of patients with heart diseases, as well as for detecting incidences of respiratory disturbance-related diseases. The chatbot-based mobile health service developed by Park et al. [17] is an intelligent chatting interface that provides prompt treatment for emergencies that may occur in everyday life and responds to the changing status of patients with chronic diseases. The diagnosis/treatment program, which can be installed in diverse types of mobile devices, was expanded through the analysis of the interaction among data using natural language processing. The dietary nutrition recommendation system for obesity management developed by Jung et al. [18] collects context information from mobile device health data and uses knowledge base-based cooperative filtering to predict missing values in the {User, Diet} matrix. Using the constructed {User, Diet}-merged matrix, a customized diet is recommended for obesity management according to context. This service allows recipes and diets to be provided through a mobile device regardless of time and the user's location.

Fig. 2. Mobile health service implementing wearable sensors

3. Prediction Model of User Physical Activity using Data Characteristics-based Long Short-term Memory Recurrent Neural Networks

The current activity of a user is predicted using his/her record of past activity according to a date or day of the week. For example, if the average activity of a user on previous Sundays burned 1,000 calories, this method predicts that the activity of the user on the subsequent Sunday will also burn 1,000 calories. However, this prediction method does not consider the user context, and the error of recommending outdoor activity in a context where such activity is impossible occurs if the recorded activity is insufficient, that is, in this case burns less than 1,000 calories. In order to avoid such an error, data that are highly relevant to the user's activity and can easily be collected from healthcare data are selected and used for predicting his/her future activity. The range of the numeric data in the domain is clear; however, in the case of time series data, an LSTM model is used in the recommendation method in order to analyze the data in time intervals and determine whether such data are normal or abnormal. Given that users' health data and surrounding context data change in real-time in everyday life, the renewal cycle, collection method, overlapping significance, acquisition convenience, and

utility must be considered. The level of difficulty of the data sequence and the collection of mobile health data vary according to the collection methods. In addition, the scope and management of collection vary according to the user's field of interest or owned devices. Accordingly, to analyze health data efficiently in a mobile environment the characteristics of the data must be considered. Health data are specific in that they are mutually influenced by each other, either directly or indirectly. For instance, users' BMI and body fat percentage change because they are influenced by the users' height and weight, and users' blood sugar changes because it is influenced by their intake of sugar. Given the mutual, direct relationship between health data, it is difficult to find a hidden significance through data analysis. The maximum and minimum temperatures are used to calculate the daily temperature difference using an arithmetic operation. However, it is possible that an overlapping significance may be shared by the variables. Between data that do not indicate a clear association, such as daily temperature difference, activity, temperature, and sleep, it is possible to find a significant relationship by analyzing their hidden relationship [19–22]. In order to provide efficient health services, these relationships must be constructed based on data that can be conveniently collected by users. In this paper, we propose a user physical activity prediction model that uses the DC-LSTM RNN. The proposed model is a prediction method that implements a neural network, the construction of which is based on the characteristics of health-related data collectable from mobile host devices. Health data and weather data continuously change over time because of their time series characteristics. Accordingly, an old state affects a new state. Therefore, in this study an RNN was used to predict user activity. The structure of an RNN is such that the old state is used in the learning process as additional input. In the case of an RNN, it is likely that the long-range dependence issue that the slope disappears as learning continues will occur, and therefore, only short-term data can be considered. Therefore, an LSTM RNN model, which allows long-term data to be considered, was selected. **Fig. 3** shows the configuration of the proposed physical activity prediction model.

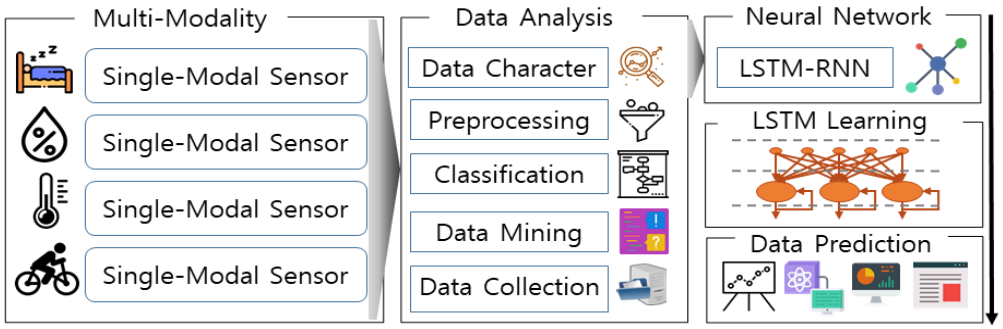

Fig. 3. Configuration of the proposed physical activity prediction model

3.1 Selection and Collection of Multi-modal Health Data

The term multi-modal health data refers to all health-related data for the same target that are collectable from mobile host devices. These data can be divided according to the collection methods and data sources, such as PHRs, Meteorological Information Systems (MISs), and lifelogs. In addition, because they are obtained through diverse collection means, these data are multi-modal. Data representing the same target are collected using an ambient sensor network. The collected data have mutually different characteristics related to the same target and show incompleteness, noise, and inconsistency problems [20,23]. A PHR consists of personal health-related variables, such as gender, age, height, weight, family medical history,

allergy, smoking status, drinking status, and health-related medical examination information. Meteorological information consists of weather, temperature, humidity, UV, daily temperature difference, fine dust, and weather index. The term lifelog refers to all recorded data on the environments surrounding an individual's everyday life. The data are collected by means of diverse wearable devices, personal health devices, and location-based services. Wearable devices, such as health bands [1], smart caps [10], and ECG measuring smartwear [16], use an ambient sensor to collect health-related lifelogs. These devices can be used to collect and analyze multi-modal health data. They use short-distance communication means, such as Bluetooth, WiFi direct, and Beacon to achieve real-time interaction with mobile host devices. Ambient sensors can be divided into photoplethysmography (PPG) sensors, pulse wave velocity (PWV) sensors, gyro sensors, and GPS. PPG sensors measure the HR by detecting the blood flow with an LED light. ECG refers to electrocardiography, which measures the electrical activity of the heart. The users' personal HR, sleep, activity, altitude, atmospheric pressure, and cadence can be measured using these sensors. A personal health device is a small device capable of measuring the user's bio status and health status and is used to measure the user's blood pressure, blood sugar, heart rate, and sleep. Location-based service refers to all the services based on the users' location, such as travel routes, weather, traffic conditions, surrounding facilities, and emergency facilities. **Table 1** shows a list of collectable multi-modal health data.

Table 1. Multi-modal health data

Division	Name	Source	Type	Sequence	Management
Personal Health Record	Sex	User	Category	Once	User
	Age	User	Integer	Once a year	User
	Height	User	Integer	Measuring users	User
	Weight	User	Integer	Measuring users	User
	Allergy	User	Category	Measuring users	User/doctor
	Smoking	User	Category	Measuring users	User
	Screenings	Center	Doc.	1–2/year	National
	… …				
Meteorological Information	Weather	Agency	Category	Real-time	National
	Temperature	Agency	Integer	Real-time	National
	Humidity	Agency	Integer	Real-time	National
	UV-rays	Agency	Integer	Real-time	National
	Daily temp.	Agency	Integer	1/day	National
	Fine dust	Agency	Integer	Real-time	National
	Weather index	Agency	Integer	Different by index	National
	… …				
Lifelog	Heart rate	Band	Integer	User settings	User
	Sleep time	Device	Complex	Daily	User
	Activity	Band	Integer	Daily	User
	Altitude	Watch	Integer	Real-time	User
	Atmospheric	Watch	Integer	Real-time	User
	Steps	Device	Integer	Real-time	User
	GPS	Device	Location	Real-time	User
	Body fat	Device	Integer	Measuring users	Center

	Blood sugar	PHD	Integer	Measuring users	User
	Blood pressure	PHD	Integer	Measuring users	User
	… …				

In general, more types and a greater volume of data allow a more accurate analysis. Given that different users own different devices, universally collectable data must be structured and the scope of the data must be set. In this study, related data were set as variables for predicting user physical activity. Activity refers to the data collected from smart devices on calories burned during physical activities such as walking and running. In general, activity is calculated as the activities accumulated from morning to the instant the user falls asleep. To be consistent with general users' everyday life patterns, in this study the data on calories burned during the day were collected at 12:00, 18:00, and 24:00. As the definition of availability, which is applicable to actual circumstances, ease of collection and frequency of occurrence were considered in the selection of the variables. Ease of collection confirms whether or not a variable is collectable by general users. For frequency of occurrence, variables indicating a noteworthy decrease in the input value, such as gender, age, and height, were excluded from neural network learning. In this study, the variables were selected by comparing the ease of collection and frequency of occurrence of an activity and the output variable with those of other variables. As a result, the collected variables were as follows: activity, weather, temperature, mean daily temperature, humidity, UV, fine dust, asthma and lung disease probability index, skin disease probability index, cadence, travel distance, mean HR, and sleep hours.

3.2 Preprocessing and Classification According to Data Characteristics

Mobile health data are preprocessed using data integration, data cleaning, data conversion, and data reduction. Data integration is a process in which users are designated as identifiers, unnecessary data are removed from diverse data sources, and data are integrated based on users. In this process, data are integrated into a group of consistent units using a map-reduce method to repeat the mapping and reduction of heterogeneous data. Given that heterogeneous big data are diversely utilized, medical institutions and hospitals are attempting to use diverse methods and programs for collecting and utilizing medical data [24,25]. Lifelogs are continuously accumulated over time, and a common data model is used to manage users' lifelogs. In general, data access methods consume substantial time and costs owing to the required repeated collection and analysis processes and context-based knowledge acquisition and refinement. However, a common data model can be used to process data in real-time without any loss and to construct a data network quickly. Data cleaning is a method used to integrate one sequence into one transaction. Given that overlapping health data can be simultaneously collected via a smartwatch, health band, and smartphone, one eigenvalue is stored and the overlapping data are deleted. Data conversion is a process in which the value of a property is separated from non-structured data containing a number of properties diversely generated and collected according to wearable devices. Data reduction is a method used to display diversely expressed health data in an integer format for the purpose of calculation and analysis. For instance, a code can be assigned to data, such as allergy, weather, and family medical history, so that they can be displayed in an integer format.

To construct a neural network, the variables were classified according to their characteristics identified during preprocessing. Characteristics refer to data properties, such as data type and collection method. Codes were assigned to data by means of data classification. In the initial phase, the variables were classified according to their data type. Based on their

data type, categorical variables were classified into weather (c01), asthma and lung disease probability index (c02), and skin disease probability index (c03). Variables other than the categorical variables, such as document type variables, coordinate type variables, and compound type variables, were excluded, because they are not suitable for neural network learning. Integer-type variables were classified according to their collection method. In the second phase, the variables were classified according to their collection method. The variables collected through the Korea Meteorological Administration were classified into temperature (w01), mean daily temperature (w02), humidity (w03), UV (w04), fine dust (w05), and ultra-fine dust (w06). Variables collected through users' devices were classified into cadence (d01), travel distance (d02), mean HR (d03), sleep hours (d04), and activity (d05). **Table 2** shows the data prior to processing, where TID denotes the transaction ID and seq. denotes the sequence. For TID, 18060711256 signifies that User u1256 is the 1st sequence on 7th June 2018, 18060722334 signifies that User u1256 is the 2nd sequence on 7th June 2018, 18060712334 signifies that User u2334 is the 1st sequence on 7th June 2018, and so on. Data collected from 26 users by means of smart bands and smart phone applications for 110 days from March 2, 2018 to June 19, 2018 were used.

Table 2. Data prior to processing

TID	**18060711256**	**21806071334**	**18060713553**	**18060711256**	**18060712011**	**18060718031**	**.....**
user	u1256	u2334	u3553	u1256	u2011	u8031	.....
seq.	1	1	1	2	1	1	.....
c01	1	1	2	1	1	1	.....
c02	1	1	1	1	1	1	.....
c03	3	3	2	3	2	2	.....
w01	16.5	15.4	17.3	19.1	16.1	15.7	.....
w02	20.3	18.5	19.5	20.3	19.7	18.5	.....
w03	50	55	55	75	60	40	.....
w04	8	9	7	9	8	8	.....
w05	45	43	51	45	41	43	.....
w06	30	30	28	34	27	32	.....
d01	1133	1246	1426	2789	1347	1235	.....
d02	519.7	658.7	626.3	1237.9	585.3	550.6	.....
d03	74	77	75	75	76	78	.....
d04	310	350	360	310	290	325	.....
d05	65	84	71	136	75	73	.....

3.3 Recurrent Neural Network-based Time Series Data

Mobile health data are specific given that that they are consecutively collected in time series as time progresses. In everyday life health data, sequentially earlier data (*s-1*) frequently influence the subsequent data. Valuable information can be provided to users by predicting the change in mobile health data according to time series. In this study, the RNN described in [19] was used to learn mobile health data and predict users' context information changes. The RNN uses actual data learning up to a certain sequence to predict the $s + 1, s + 2, s + 3, \ldots, s + n$ sequence. **Fig. 4** shows the RNN-based prediction of time series data [26–29].

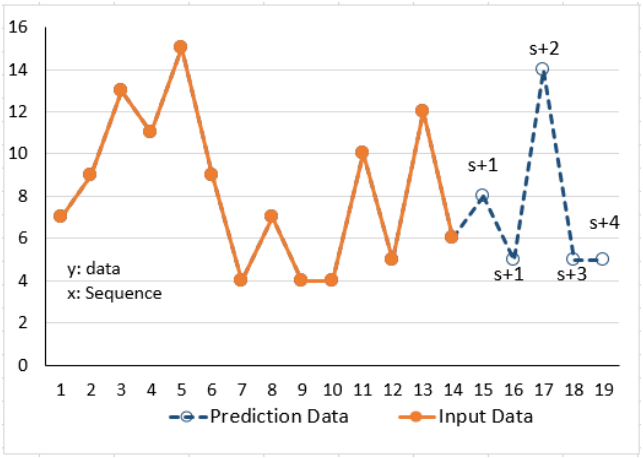

Fig. 4. Recurrent neural network-based prediction of time series data [19]

In an RNN, data runs from the input layer to the hidden layer, and the obtained results are entered back into the input layer. In an RNN, the previous status influences the subsequent status according to the data sequence. An RNN operates in the feedforward neural network (FNN)-based structure, and the data flow operates in the order of input layer–hidden layer–output layer. RNN models can be divided into fully connected RNN (FRNN), recurrent multilayer perceptron (RMLP), and simple recurrent network (SRN) models according to their feedback method. For an RNN, the pattern of the arrangement appearing according to sequence is used for calculating the arrangement of the next sequence [26–29]. **Fig. 5** shows an RNN module, where x_s represents input and y_s and h_s represent new outputs according to node function. f_w represents the node function. h_{s-1} represents the recurring old output. New input can be calculated by substituting the new state (x_s) and the old state (h_{s-1}) for node function.

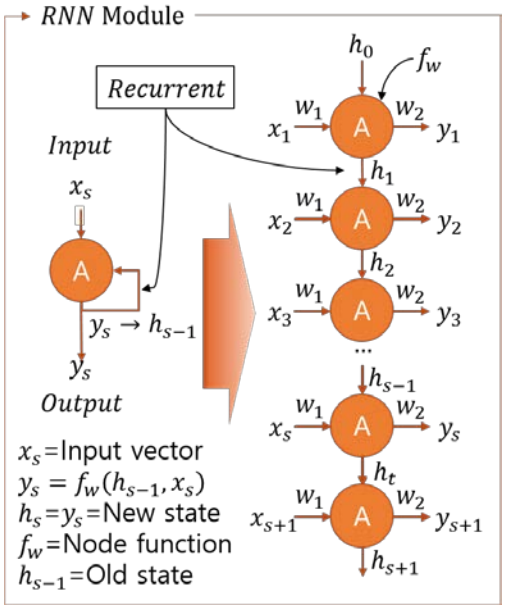

Fig. 5. RNN module [26–29]

3.4 Long Short-term Memory Recurrent Neural Network Modeling for Prediction

In the prediction of mobile health data, the old state influences the new state because of their time series characteristics. To consider both states, mobile health data are used in the RNN structure. General RNNs are likely to cause the long-term dependency problem followed by the vanishing gradient problem. Mobile health data generate a large number of sequences as time progresses and are constructed based on LSTM [30–33], where the long-term dependency problem is ameliorated. An LSTM model uses a gate mechanism to resolve the vanishing gradient problem. It is constructed based on a number of gate-connected cells, which can be used to read/write information. **Fig. 6** shows an LSTM neural network module.

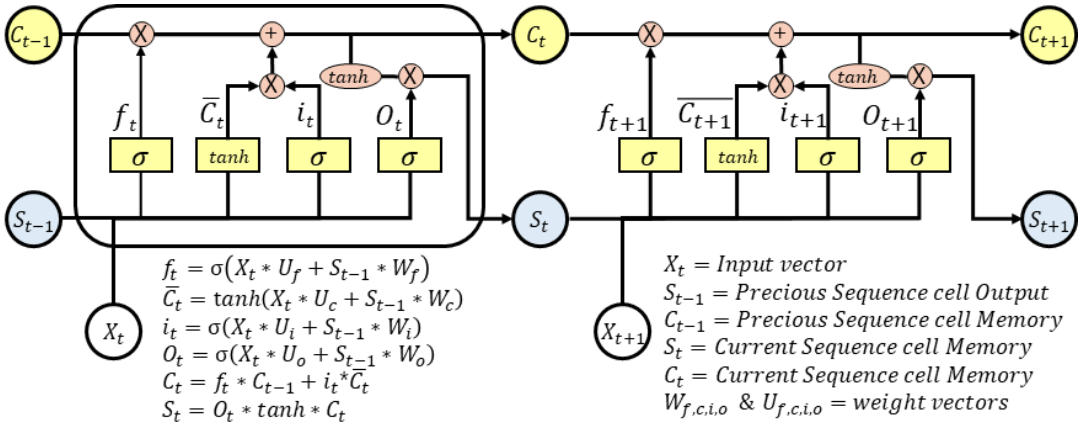

Fig. 6. Long short-term memory neural network module [30-33]

In **Fig. 6**, *i* represents the input gate, *f* represents the forget gate, *o* represents the output gate, *g* represents the candidate hidden state, c_t represents the unit's internal memory, and s_t represents the hidden state. *i*, *f*, and *o*, serving as gates, are converted into a value between 0 and 1 using the sigmoid function to conduct an element-wise calculation for each element of the input vector. The input gate adjusts the information transfer rate to confirm the extent to which the current input value is transferred. The forget gate adjusts whether the input value is long-term or short-term memory. The output gate adjusts the information transfer rate to confirm the extent to which the state information is output. The candidate hidden state (*g*) displays the candidate output value calculated using the activation function *tanh* and the pre-existing state. Instead of providing the candidate output value itself as the output, LSTM provides only part of it as the output through gate calculation. c_t is the sum of c_{t-1} stored in the previous memory, the value calculated for each element of the hidden gate value, and the value calculated for each element of the candidate state and input gate. It represents the combination of the previous memory and current input. s_t calculates the final output value by calculating the c_t value and each element of the output gate.

Mobile health data have a direct/indirect association established in the collection or deduction process depending on the variables, which needs to be considered. Associations can be divided into direct relations and indirect relations. A direct relation is a relation with a BMI or daily temperature difference, which are variables calculated from other variables. An indirect relation is exemplified by the relation between travel distance and weather, which are variables that change as the user's location changes. An LSTM network was constructed based on the characteristics of mobile health data. **Fig. 7** shows the DC-LSTM RNN model. The proposed model is a complex neural network, where the classified input variables c{c01~c03},

w{w01~w06}, and d{d01~d05} are used to construct the many-to-many LSTM for each classified variable group, the output values are totaled by fully connecting the feedforward fusion layer, and the totaled data are again set as the input to LSTM to obtain the final user physical activity prediction.

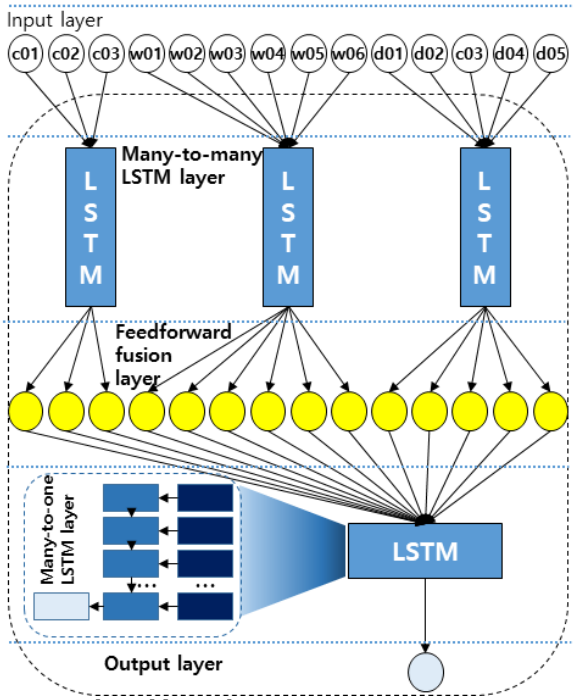

Fig. 7. Data characteristics-based long short-term memory recurrent neural network model

4. Performance Evaluation and Results

The physical activity prediction model uses keras on RStudio for learning. The software environment comprises Windows 10 Pro, R 3.5.0 [34], RStudio 1.1.453 [35], Keras 2.1.6 [36], and TensorFlow 1.8.0 [37]. The hardware comprises Intel i5-4690 CPU 3.50 GHz (4 CPUs), 16384 MB RAM, and GeForce GTX 970. RStudio is an integrated development environment for R and provides diverse functions through its packages [38]. Keras is a library that can be easily used based on TensorFlow and is capable of using a sequential model to actualize multi-layers [36]. In this study, the R packages used were timetk, cowplot, recipes, rsample, yardstick, and keras. The timetk package is a time series processing tool. The cowplot package is a data visualization tool. The recipes package is a preprocessing tool for design matrices. The Rsample package is a general resampling infrastructure tool. The yardstick package is a tool for attaining accuracy. The keras package is an R-based neural network API and is a modeling tool that supports both circuit-based and recurrent networks [34,35,38].

The neural network learning using R was operated in the following order: package call, data input, separation between learning data and test data, data preprocessing, LSTM modeling, model learning, prediction, and model evaluation. In R, the package call was made in a library (keras) format. The data were divided into 80% training data and 20% testing data, and the key values assigned to the two data groups were “train” and “test,” respectively. The sequential data for 22 days following the old state were used as the testing data in order to consider their

time series characteristics. The sequential data for the remaining 88 days was used as training data. The LSTM algorithm requires that input data be centrally aligned and scaled and uses the recipes package for data preprocessing. It uses the recipes package's step_sqrt for data conversion and singular value reduction. The proposed LSTM model was built in a many-to-one structure [30–33] that is capable of predicting one activity from various inputs. The input value had 14 columns, which were divided into a group of 3, a group of 6, and a group of 5 according to the data characteristics. Accordingly, three LSTMs were constructed having sequence lengths of 3, 6, and 5, respectively, and their output was modeled as an LSTM with a sequence length of 3. In the LSTM model, *tanh* was used for the activation function. The constructed LSTM RNN model used the data having the key value "train" for learning. When the learning process was complete, the data having the key value "test" were used to predict and evaluate user physical activity.

User physical activities are data with time series characteristics and the most representative time series prediction methods are statistical regression methods and neural network methods. The statistical regression methods used were the autoregressive (AR) model and autoregressive integrated moving average (ARIMA) model, which are modified versions of the moving average (MA) model [39]. These statistical models use past data for tendency prediction. The neural network methods include multi-layer perceptron (MLP), convolutional neural networks (CNNs), and RNNs. In this study, an evaluation was conducted to measure the performance of the activity prediction method using the ARIMA model, CNN, RNN, and the proposed DC-LSTM RNN. The activity of a user was predicted using each method and the difference between the actual value and the predicted value was evaluated as an error. The testing data were collected from 26 users for 22 days. The activity from March 2 to March 13 was predicted according to the input variables c{c01~c03}, w{w01~w06}, and d{d01~d05} of the user collected on March 2, 2018 and compared with the actual value. A total of 572 errors, that is, 22 errors for each user, were evaluated using each method and the root mean square error (RMSE) was calculated as the average value according to the prediction time (+1 day, +2 days, +3 days, …, +11 days). **Fig. 8** shows the user physical activity prediction according to the time series prediction method.

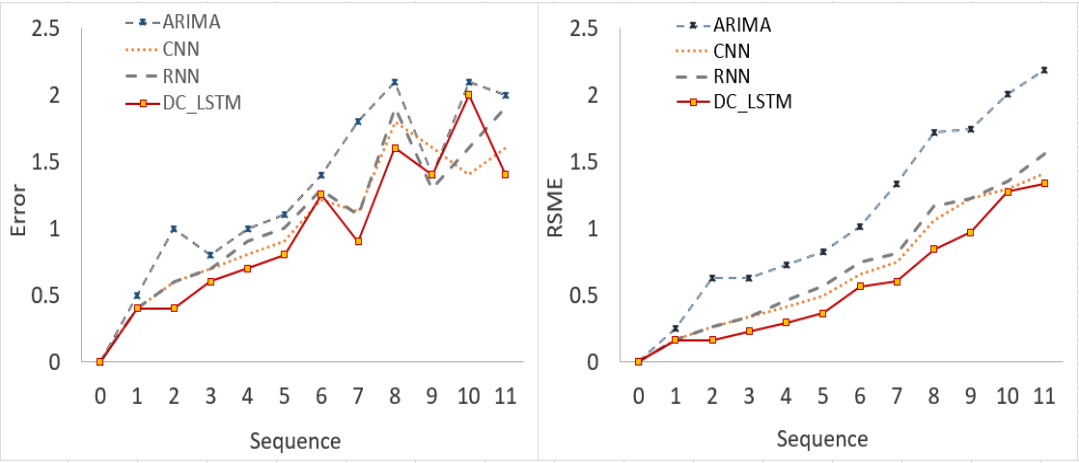

Fig. 8. User physical activity prediction based on the time series prediction method

Table 3 shows the RMSE of the user physical activity prediction that was yielded using the ARIMA model, CNN, RNN, and DC-LSTM. A sequence refers to a sequential flow, such as time flow. *s* represents the current time point and *s* + *n* represents the current time point plus

the next *n* time point. For example, Sequence 0 of DC-LSTM indicates the activity of a user at the current time and Sequence 1 indicates the activity of a user one day after that time, which is predicted at the current time. The user physical activity prediction results achieved using the prediction methods were evaluated using the RMSE. The evaluation results showed that the activity prediction methods incurred greater errors because more sequences were implemented. Overall, the proposed DC-LSTM showed an outstanding RMSE as compared to other methods despite the fact that more sequences were implemented.

Table 3. Root mean square error of user physical activity prediction using ARIMA, CNN, RNN, and DC-LSTM

Sequence	**RMSE**			
	ARIMA	CNN	RNN	DC-LSTM
***s* + 1**	0.250	0.160	0.160	0.160
***s* + 2**	0.625	0.260	0.260	0.160
***s* + 3**	0.630	0.337	0.337	0.227
***s* + 4**	0.722	0.413	0.455	0.292
***s* + 5**	0.820	0.492	0.564	0.362
***s* + 6**	1.010	0.656	0.747	0.563
***s* + 7**	1.329	0.742	0.813	0.599
***s* + 8**	1.714	1.055	1.163	0.844
***s* + 9**	1.741	1.225	1.221	0.968
***s* + 10**	2.008	1.299	1.355	1.271
***s* + 11**	2.189	1.414	1.560	1.334
Avg.	1.185	0.732	0.785	0.616

5. Conclusion

In this paper, a prediction model of user physical activity was proposed that uses the DC-LSTM RNN. The proposed method is an LSTM RNN constructed by selecting, collecting, preprocessing, and classifying data according to their characteristics. Various methods exist for predicting or calculating the activity level of a user; however, such methods most frequently use step count and traveling distance data. These methods predict the activity for the current day or future days using the data for activity performed in the previous week or month. If low activity due to unfavorable weather or the surrounding context is measured, a warning is given to the user. These methods suffer the problem that outdoor activity may be recommended through a comparison with the activity in the past based on the date, even if the current surrounding context is not favorable for such activity. In this study, such errors were minimized by means of the learning of various variables to obtain an activity prediction that reflects the surrounding context and personal context. In the case of mobile health data, the old state affects the new state because of the data's time series characteristics. In order to consider this, an LSTM model that allows the use of long-term memory was configured. Based on data characteristics, data that can be conveniently collected and utilized by universal users were selected. The selected data were collected from wearable devices, the Korea Meteorological Administration, and meteorological information applications. The collected data were constructed into a transaction by preprocessing, including data integration, data cleaning, data conversion, and data reduction. Data were classified according to their types and collection methods revealed during preprocessing and were set as the variables of a neural network. Data

classification was used to construct an LSTM RNN and the learning process was performed. The RMSE of the activity prediction method was evaluated using the ARIMA model, CNN, RNN, as well as the DC-LSTM RNN. The evaluation results showed that the mean RMSE value of the proposed DC-LSTM RNN model, 0.616, was the best value.

In the future, we plan to conduct in-depth research on differentiated personalized smart health services by expanding the scope of data collection and increasing the number of target prediction variables. In addition, the development of an application is planned for providing user-based smart health services. This application would overcome the problems of existing standardized health services and provide information that is more valuable to users.

References

[1] K. Chung, Y. Na, J. H. Lee**, "**Interactive Design Recommendation using Sensor based Smart Wear and Weather WebBot," *Wireless Personal Communications*, Vol. 73, No. 2, pp. 243–256, 2013. Article (CrossRef Link).

[2] G. Bartlett, M. Dawes, Q. Nguyen, M. S. Phillips, M. S., "Precision Medicine in Primary Health Care," in *Proc. of Progress and Challenges in Precision Medicine*, pp. 101-113, 2017. Article (CrossRef Link).

[3] H. Yoo, K. Chung, "Heart Rate Variability based Stress Index Service Model using Bio-Sensor," *Cluster Computing*, vol.21, no.1, pp.1139-1149, 2017. Article (CrossRef Link).

[4] A. Nasrollahi, W. Deng, Z. Ma, P. Rizzo, "Multimodal Structural Health Monitoring based on Active and Passive Sensing," *Structural Health Monitoring*, Vol. 17, No. 2, pp. 395–409. 2018. Article (CrossRef Link).

[5] C. Cadena, A. Dick, I. Reid, "Multi-modal Auto-encoders as Joint Estimators for Robotics Scene Understanding," in *Proc. of 2016 Robotics: Science and Systems XII Conference 2016*, pp. 1–9, 2016. Article (CrossRef Link).

[6] J. Kim, H. Jang, J. T. Kim, H. J. Pan, R. C. Park, "Big-Data Based Real-Time Interactive Growth Management System in Wireless Communications," *Wireless Personal Communications*, vol.105, no.2, pp.655-671, 2018. Article (CrossRef Link).

[7] J. C. Kim, K. Chung, "Mining Health-Risk Factors using PHR Similarity in a Hybrid P2P Network," *Peer-to-Peer Networking and Applications*, Vol. 11, No. 6, pp. 1278–1287, 2018. Article (CrossRef Link).

[8] Observational Health Data Sciences and Informatics. Article (CrossRef Link).

[9] J. H. Kim, J. Kim, D. Lee, K. Chung, "Ontology Driven Interactive Healthcare with Wearable Sensors," *Multimedia Tools and Applications*, Vol. 71, No. 2, pp. 827–841, 2014. Article (CrossRef Link).

[10] I. J. Jun, K. Jung, "Life Weather Index Monitoring System using Wearable based Smart Cap," *Journal of the Korea Contents Association*, Vol. 9, No. 12, pp. 477–484, 2009. Article (CrossRef Link).

[11] Open data Portal. Article (CrossRef Link).

[12] Jena. Article (CrossRef Link).

[13] Korea Meteorological Administration. Article (CrossRef Link).

[14] S. M. Jo, K. Chung, "Design of Access Control System for Telemedicine Secure XML Documents," *Multimedia Tools and Applications*, Vol. 74, No. 7, pp. 2257–2271, 2015. Article (CrossRef Link).

[15] K. Jung, Y. H. Lee, J. K. Ryu, "Health Information Monitoring System using Context Sensors based Band," *Journal of the Korea Contents Association*, Vol. 11, No. 8, pp. 14–22, 2011. Article (CrossRef Link).

[16] K. Jung, "Correlation between Visual Sensibility and Vital Signal using Wearable based Electrocardiogram Sensing Clothes," *Journal of the Korea Contents Association*, Vol. 9, No. 12, pp. 496–503, 2009. Article (CrossRef Link).

[17] K. Chung, R. C. Park, "Chatbot-based Healthcare Service with a Knowledge Base for Cloud Computing," *Cluster Computing*, pp.1-13, 2018. Article (CrossRef Link).
[18] H. Jung, K. Chung, "Knowledge-based Dietary Nutrition Recommendation for Obese Management," *Information Technology and Management*, Vol. 17, No. 1, pp. 29–42, 2016. Article (CrossRef Link).
[19] H. Jung, K. Chung, "Sequential Pattern Profiling based Bio-Detection for Smart Health Service," *Cluster Computing*, Vol. 18, No. 1, pp. 209–219, 2015. Article (CrossRef Link).
[20] H. Jung, H. Yoo, K. Chung, "Associative Context Mining for Ontology-Driven Hidden Knowledge Discovery," *Cluster Computing*, Vol. 19, No. 4, pp. 2261–2271, 2016. Article (CrossRef Link).
[21] H. Yoo, K. Chung, "Mining-based Lifecare Recommendation using Peer-to-Peer Dataset and Adaptive Decision Feedback," *Peer-to-Peer Networking and Applications*, Vol. 11, No. 6, pp. 1309–1320, 2018. Article (CrossRef Link).
[22] F. Masseglia, M. Teisseire, P. Poncelet, "Sequential Pattern Mining," *Encyclopedia of Data Warehousing and Mining, IGI Global*, pp. 1028–1032, 2005. Article (CrossRef Link).
[23] H. G. Jun, G. S. Hyun, K. B. Lim, W. H. Lee, H. J. Kim, "Big Data Preprocessing for Predicting Box Office Success," *KIISE Transactions on Computing Practices*, Vol. 20, No. 12, pp. 615–622, 2014. Article (CrossRef Link).
[24] A. Kiourtis, A. Mavrogiorgou, D. Kyriazis, "Aggregating Heterogeneous Health Data through an Ontological Common Health Language," in *Proc. of International Conference on Developments in eSystems Engineering*, pp. 175–181, 2017. Article (CrossRef Link).
[25] J. Dean, S. Ghemawat, "MapReduce: simplified data processing on large clusters," *Communications of the ACM*, Vol. 51, No. 1, pp. 107–113, 2008. Article (CrossRef Link).
[26] A. Manashty, J. L. Thomson, "A New Temporal Abstraction for Health Diagnosis Prediction using Deep Recurrent Networks," In *Proc. of the 21st International Database Engineering & Applications Symposium*, pp. 14-19, 2017. Article (CrossRef Link).
[27] E. J. Lee, C. H. Min, T. S. Kim, "Development of the KOSPI (Korea Composite Stock Price Index) Forecast Model using Neural Network and Statistical Methods," *The Institute of Electronics Engineers of Korea, Computer and Information*, Vol. 45, No. 5, pp. 95–101, 2008.
[28] T. J. Hsieh, H. F. Hsiao, W. C. Yeh, "Forecasting Stock Markets using Wavelet Transforms and Recurrent Neural Networks: An Integrated System based on Artificial Bee Colony Algorithm," *Applied Soft Computing*, Vol. 11. No. 2, pp. 2510–2525, 2011. Article (CrossRef Link).
[29] S. Jelena, M. Nijole, M. Algirdas, "High-low Strategy of Portfolio Composition using Evolino RNN Ensembles," *Engineering Economics*, Vol. 28, No. 2, pp. 162–169, 2017. Article (CrossRef Link).
[30] A. Khosravi, R. N. N. Koury, L. Machado, J. J. G. Pabon, "Prediction of wind speed and wind direction using artificial neural network, support vector regression and adaptive neuro-fuzzy inference system," *Sustainable Energy Technologies and Assessments*, Vol. 25, pp. 146-160. 2018. Article (CrossRef Link).
[31] T. Fischer, C. Krauss, "Deep Learning with Long Short-term Memory Networks for Financial Market Predictions," *European Journal of Operational Research*, Vol. 270, No. 2, pp. 654-669, 2018. Article (CrossRef Link).
[32] F. A. Gers, N. N. Schraudolph, J. Schmidhuber, "Learning Precise Timing with LSTM Recurrent Networks," *Journal of Machine Learning Research*, 3, pp. 115–143, 2002.
[33] R. Cai, B. Zhu, L. Ji, T. Hao, J. Yan, W. Liu, "An CNN-LSTM Attention Approach to Understanding User Query Intent from Online Health Communities," in *Proc. of the IEEE International Conference on Data Mining Workshops*, pp. 430-437, 2017. Article (CrossRef Link).
[34] R: The R Project for Statistical Computing. Article (CrossRef Link).
[35] R Studio. Article (CrossRef Link).
[36] Keras: The Python Deep Learning library. Article (CrossRef Link).
[37] Tensorflow. Article (CrossRef Link).
[38] CRAN - R Project. Article (CrossRef Link).

[39] G. P. Zhang, "Time Series Forecasting using a Hybrid ARIMA and Neural Network Model," *Neurocomputing*, Vol. 50, pp. 159-175, 2003. Article (CrossRef Link).

Joo-Chang Kim has received B.S. and M.S. degrees from the School of Computer Information Engineering, Sangji University, South Korea in 2014 and 2016, respectively. He has worked for Network Management Department, Network O&S Co., Ltd. He is currently in the doctorate course of Department of Computer Science, Kyonggi University, South Korea. He has been a researcher at Data Mining Lab., Kyonggi University. His research interests include Data Mining, Machine Learning, Artificial Intelligent, Knowledge System, Medical Bigdata Mining, Healthcare, and Recommendation.

Kyungyong Chung has received B.S., M.S., and Ph.D. degrees in 2000, 2002, and 2005, respectively, all from the Department of Computer Information Engineering, Inha University, South Korea. He has worked for Software Technology Leading Department, Korea IT Industry Promotion Agency (KIPA). From 2006 to 2016, he was a professor in the School of Computer Information Engineering, Sangji University, South Korea. Since 2017, he is currently a professor in the Division of Computer Science and Engineering, Kyonggi University, South Korea. His research interests include Data Mining, Artificial Intelligent, Healthcare, Knowledge System, HCI, and Recommendation.