IoT-Based Health Big-Data Process Technologies: A Survey

Hyun Yoo[1], Roy C. Park[2], and Kyungyong Chung[3*]
[1] Contents Convergence Software Research Center, Kyonggi University
154-42, Gwanggyosan-ro, Yeongtong-gu, Suwon-si, Gyeonggi-do, 16227, South Korea
[e-mail: rhpa0916@gmail.com]
[2] Department of Computer Information Software Engineering, Sangji University
83, Sangjidae-gil, Wonju-si, Gangwon-do, 26339, South Korea
[e-mail: roypark1984@gmail.com]
[3] Division of AI Computer Science and Engineering, Kyonggi University
154-42, Gwanggyosan-ro, Yeongtong-gu, Suwon-si, Gyeonggi-do, 16227, South Korea
[e-mail: dragonhci@gmail.com]

Received September 28, 2020; revised January 20, 2020; accepted February 8, 2020; published March 31, 2021

Abstract

Recently, the healthcare field has undergone rapid changes owing to the accumulation of health big data and the development of machine learning. Data mining research in the field of healthcare has different characteristics from those of other data analyses, such as the structural complexity of the medical data, requirement for medical expertise, and security of personal medical information. Various methods have been implemented to address these issues, including the machine learning model and cloud platform. However, the machine learning model presents the problem of opaque result interpretation, and the cloud platform requires more in-depth research on security and efficiency. To address these issues, this paper presents a recent technology for Internet-of-Things–based (IoT-based) health big data processing. We present a cloud-based IoT health platform and health big data processing technology that reduces the medical data management costs and enhances safety. We also present a data mining technology for health-risk prediction, which is the core of healthcare. Finally, we propose a study using explainable artificial intelligence that enhances the reliability and transparency of the decision-making system, which is called the black box model owing to its lack of transparency.

Keywords: Data Mining, XAI, Cloud, IoT, Healthcare, WBAN, Big Data, Deep Learning

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-2020R1A6A1A03040583).

http://doi.org/10.3837/tiis.2021.03.009 **ISSN : 1976-7277**

1. Introduction

Recently, various services and platforms have been developed worldwide based on the development of the Internet of Things (IoT), artificial intelligence (AI), and mobile and cloud technologies, which are fourth-industrial-revolution technologies [1]. In particular, IoT-based healthcare services can be used to detect physical and chemical changes in patients via the use of various sensors. These have also resulted in the development of medical services for patient management in real time by delivering information to hospitals and healthcare-related medical institutions. In addition, technology for the continuous monitoring of a patient's condition and wireless communication technology have been developed. On the basis of this, there is growing interest in wireless body area network (WBAN) technology, which is a near field communication technology used around or of the human body [2]. The medical WBAN technology used for healthcare services is different from that of existing healthcare and can provide a wide range of services in the form of a platform. It is a technology that can be used to measure bio signals both inside and outside the human body through sensors and various medical devices and provide patient-health information via various medical networks. The introduction of this platform has resulted in an increase in the production speed and volume of health big data at the big-data scale [3]. The majority of health big data comprises personal information, and health big data features a wide range of information protection and security in the storage and transmission processes. As a result, healthcare platforms are required to maintain high levels of information protection and security in addition to being efficient in processing big data. However, it is difficult to consider and evaluate both these aspects simultaneously.

Observations of every process of an individual's life are contained in the big data used in healthcare. Observational data of the individual's environment, eating habits, and biological activities from birth are the basis of healthcare. In addition, medical information generated by medical institutions comprises the core data of health big data. Medical information can comprise various forms of atypical data, such as medical records, images of computed tomography (CT) and magnetic resonance imaging (MRI), results of ultrasound, and endoscopy images. Health data includes physical-activity and surrounding-environment information [3]. Owing to technological developments, the amount of health data has increased rapidly to the big-data scale. Recently, various studies have been conducted on disease prediction and new drug developments using deep learning have occurred in the healthcare field [4, 5]. As discussed above, data mining research in the healthcare field differs from a statistical analysis. Medical knowledge of the studied disease is required in the former. The implemented system also requires steps to be performed to confirm its clinical utility. The majority of medical information is described using atypical text; it is characterized by many medical abbreviations and symbols. Therefore, research on the effective implementation of atypical text analysis techniques, such as natural language processing, is crucial, and a health big-data-processing technique is required to process various types of big data [6]. Data mining, machine learning, and reinforcement learning, which are represented as the element technologies of decision-making systems, process data using the following techniques: association rules, correlation, regression analysis, clustering, classification, and prediction. Research tasks, such as the ensemble technique, which operate based on this and explainable AI technologies, are possibly important factors for future health big-data processing [7].

The objective of this study is to analyze recent research trends to help health big-data researchers access and use big data correctly. Section 2 describes cloud-based IoT health platform, and Section 3 presents health big-data-processing technology. In Section 4, we

present a health-risk-prediction technology comprising data mining. The conclusions of this study are presented in Section 5.

2. Cloud-based IoT Health Platform

In the past, the focus was on disease treatment and saving a patient's life. The current concept of medical services is to make innovative changes in the medical industry using technologies such as the cloud, IoT, big data, and AI. Among these, the cloud-based IoT health platform reduces the costs associated with human and material resources that are unnecessarily consumed in the medical treatment process through medical devices using IoT. Collected information is stored in the cloud to provide easy access to patient information, thus making it possible to provide more efficient and effective services [8]. In addition, the doctor-patient relationship that is focused on a doctor's diagnosis has evolved into a cloud-based IoT medical system in which a patient knows his/her own condition, and a consensus is drawn through sustainable discussions with the doctor. This allows doctors and patients to communicate, and appropriate medical services are determined. **Table 1** presents recent studies conducted on cloud-based IoT health systems.

Table 1. Recent research on cloud-based IoT health systems

Author (year)	Research content
A. Omar et al. [9] (2019)	- Developed a patient-centered medical data management system using blockchain technology for storage to achieve privacy protection - Encryption function used to encrypt health data and ensure anonymity
Y. Karaca et al. [10] (2019)	- Converged the mobile cloud environment with cloud computing for the purpose of medical information processing
M. Rahman et al. [11] (2019)	- Developed a lossless deoxyribo-nucleic-acid-sequence (DNA-sequence) hiding method to ensure the authenticity of DNA sequences in mobile cloud-based medical systems
R. Ganiga et al. [12] (2019)	- Presented a secure cloud architecture by building a private cloud - Managed patient data in the medical environment by building a personal cloud using open source tools
M. Pham et al. [13] (2018)	- Developed a cloud-based smart home environment - Collected physiological, motion, and voice signals via non-invasive wearable sensors and provided situational awareness services
S. Miah et al. [14] (2018)	- Evaluated patient data and medical histories - Developed diagnostic skills through health professionals and community clinics in a cloud-based solution
P. Verma et al. [15] (2018)	- Developed a cloud-centered IoT-based health-diagnosis system - Defined a smart, interactive health system for IoT environments
T. Bhardwaj et al. [16] (2018)	- Developed technology to provide services to WBAN users based on sensory data volume and application type - Developed a computing system for maintenance at the Edge of Things - Developed a framework to regulate computing resources in the cloud

2.1 Cloud-based WBAN Healthcare

Many WBAN studies have been conducted because cloud technology is useful for big-data management, processing, and analysis. In recent years, many studies have been conducted on the benefits of the cloud for medical applications. The hierarchical structure of the cloud network consists of three service layers, which provide various services, between the client and server layers [10, 13]. **Fig. 1** shows the service layer of a cloud network. First, the infrastructure as a service (IaaS) provides the network technology, such as the load balancer and virtual private network (VPN). It is divided into a physical and a virtual network layer [16]. IaaS is mechanically different from traditional physical network devices because it is serviced virtually to each user. The second is a platform as a service (PaaS), which is a platform layer that provides a virtual technology for development platforms. Third, software as a service (SaaS) is the software layer, which provides the user's medical information via virtual software services such as web applications. Peer-to-peer networking physically connects these services to the server [17]. The overlay cloud computing service is then set up between the client and software layers such that the system can provide an automatic network configuration. Observation areas are set up across layers to collect observed values from each layer. The control area sends a virtual network configuration command to the virtual network layer. In contrast to conventional peer-to-peer networks, the network of the infrastructure layer is divided into a physical and virtual network layer in the cloud network. In addition, overlay cloud computing services that provide automatic network configurations operate between the software and client layers.

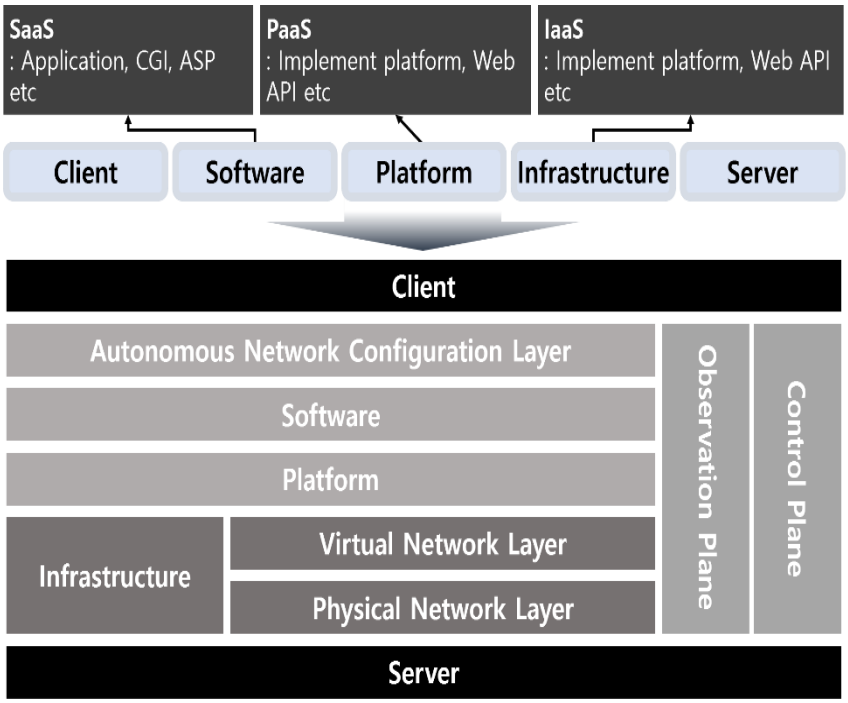

Fig. 1. Service layer of a cloud network

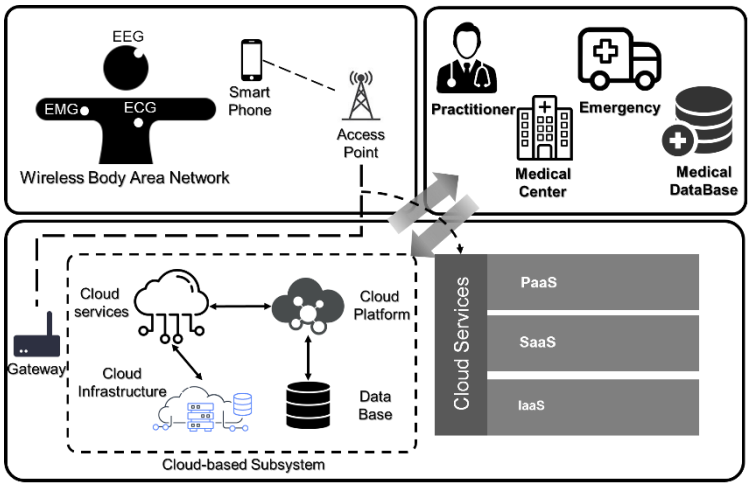

Fig. 2. Health prediction system using scalable cloud and big-data technology

P. Sahoo et al. [18] conducted a study on a health-prediction technology and an analyzing healthcare big-data technology for future health conditions. **Fig. 2** shows a health-prediction system comprising the use of scalable cloud and big-data technology. The patient's health status is monitored via their WBAN, and the data is stored in the scalable cloud. A signature-based access control mechanism prevents unauthorized users from accessing data. The patient sets up profiles, configures those who can access data, and determines whether to monitor continuously, only on request, or periodically. Furthermore, machine learning was applied to signals collected by a WBAN sensor to classify congestive heart failure among system users. However, as the amount of data increased, the data traffic grew rapidly. Therefore, it is necessary to consider a method of reducing the waiting time as the data inflow rate and volume increase.

2.2 Medical IoT Healthcare Network Platform using a WBAN

The IoT-based healthcare network is a key element of WBANs used in healthcare applications. **Fig. 3** depicts an IoT-based healthcare network. The IoT-based healthcare network consists of a topology, architecture, and platform.

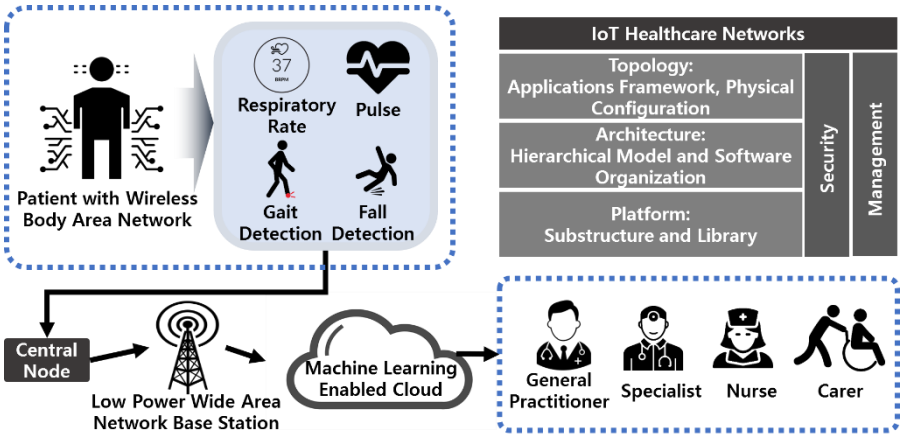

Fig. 3. IoT-based healthcare network

The topology represents the flow of the network wherein the WBAN sensors and data collected through the sensors are transmitted in an IoT-based healthcare system. As the healthcare environment comprises dynamic environments owing to the mobility of users, the network interface conditions are changed. In addition, the server and WBAN sensors that constitute the topology determine the value of the optimal condition through the session setup process that periodically sends and receives control signals. In other words, the operating environment of the WBAN sensor, sensor type and communication method, traffic types and patterns, and reliability and delay requirements for communication are used to determine the optimal values for the parameters required at each protocol layer and to maintain an efficient healthcare environment. In addition, various features related to the device's movement, channel status, communication status, information, and amount of data transmitted are reflected in the pattern management of the communication and traffic between the WBAN devices in a continuous healthcare environment.

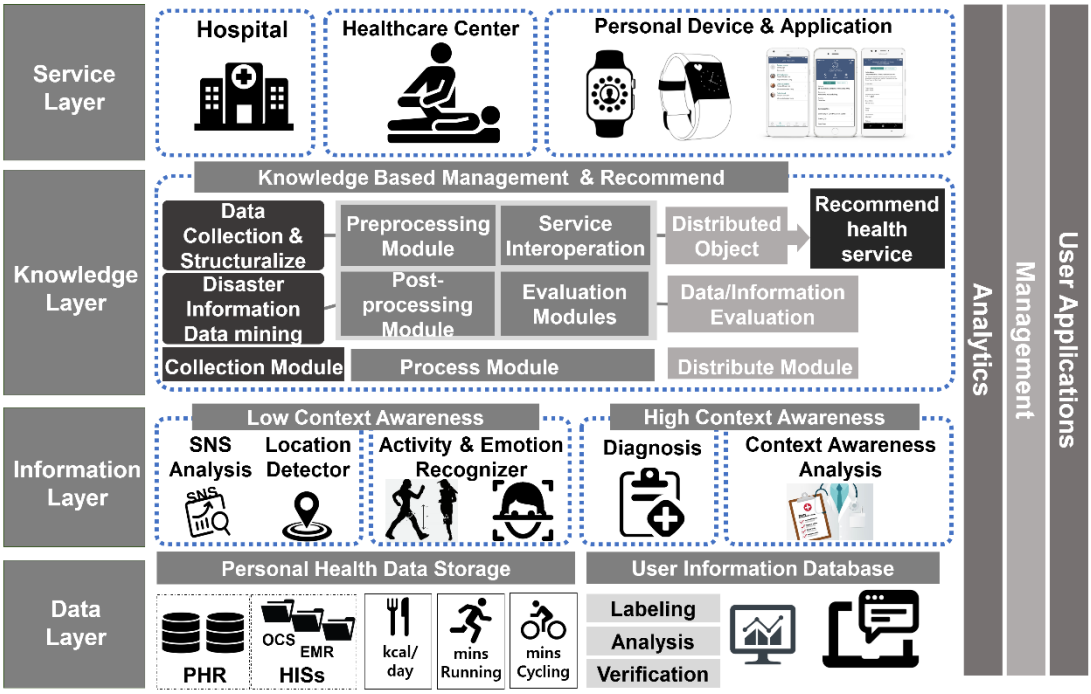

Fig. 4. WBAN-based IoT healthcare service platform and framework model

Fig. 4 shows the proposed WBAN-based IoT healthcare service platform and framework model. The framework model developed by Chung et al. [17, 19] consists of four layers and forms a healthcare platform via the interactions between layers. The data layer consists of components for the storage and processing of data collected via the WBAN sensor and undergoes real-time data filtering to increase the reliability and consistency of the data analysis. In this process, users are made to go through user authentication and encryption procedures to enhance privacy. The information layer analyzes the user's behavior and performs situation inference modeling, which can be used to predict a user's situation and behavior pattern through life-pattern recognition and inference based on the collected data. The knowledge layer analyzes the user's health information based on the medical information database established for healthcare services and creates and manages knowledge according to

the situation. The service layer provides customized services by converting knowledge information collected and newly processed through each layer into the user healthcare service.

2.3 Cloud-based Health Big-Data Management

Cloud-based health big-data management comprises the use of data mining, machine learning, and other forms of detailed analyses on the vast amounts of collected data to find meaningful associations between a patient's symptoms and conditions and to further determine effective treatments for various conditions. Moreover, a doctor can remotely provide treatment methods and medical-care-related feedback to the patient. In this process, data processing can be guaranteed through the delay processing of the cloud system to solve errors and losses caused by the delay. This provides services, such as scalability and link connectivity, to the total system capacity when the demand for the system increases by sharing resources such as bandwidth and storage space provided by all clients of the cloud network, thus increasing the accessibility and reliability of the network [20].

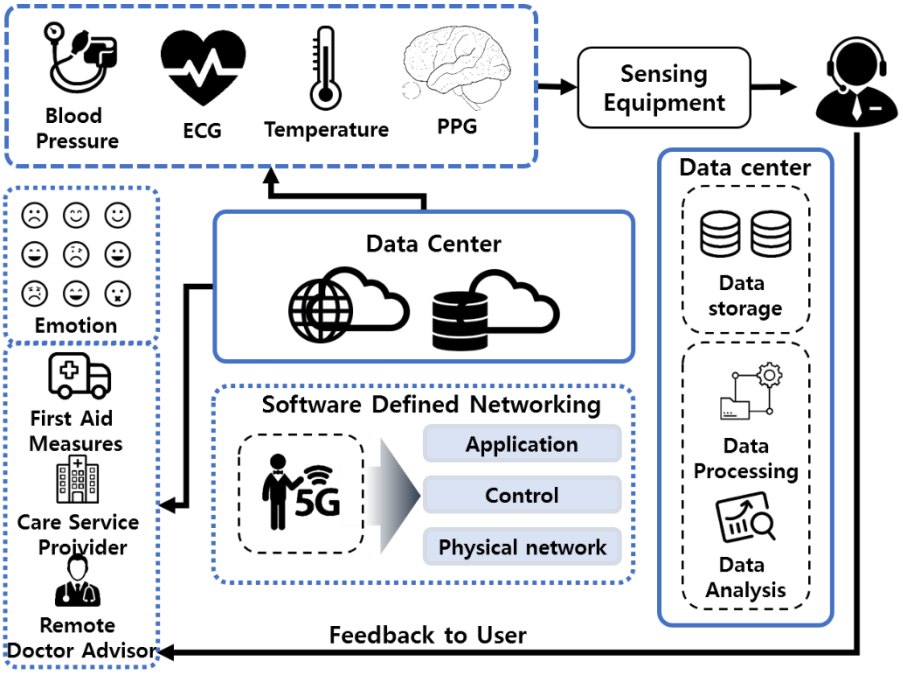

Fig. 5. Cloud-WBAN system for monitoring the emotional status of a patient

H. Kalantarian et al. [21] considered cloud-based patient emotional state monitoring and developed a cloud-WBAN system for monitoring the emotional state of patients. It measures a users' physical and mental changes through the WBAN sensor and conducts a big data analysis on the relationships between them. **Fig. 5** shows the cloud-based patient-emotional-state monitoring system. The patient data collected via the WBAN sensor are stored in the cloud module, and data mining is used to extract key data based on the patient's condition, which is then used to infer knowledge. To maximize the storage space, a pre-treatment process is performed to remove and integrate redundant or unwanted data from the database. In the system, cloud storage is not just a function for maintaining health records, but a knowledge base that can be used to construct new knowledge via deduction and inference through

machine learning, reinforcement learning, and data mining. The recommendation information created through the knowledge base is reprocessed based on the user's living environment and health-status information. The created information is then provided to the user after an authentication process for duplication prevention and rule consistency verification to increase the reliability.

3. Health Big-Data Processing Technology

3.1 Health Big-Data Processing using Explainable Artificial Intelligence (XAI)

A big-data-processing algorithm processes health big data and supports user decision making through prediction and recommendation models. However, owing to the complexity of the algorithm, the inside of the model is a black box. Therefore, it is difficult to clearly explain the rationale and process of the derived result. In health big-data processing, the reliability and accuracy of the obtained results are important. Therefore, a clear explanation is required for the validity of all processes and results generated from the decision-making system. Users should provide a clear description of the results of health big data, and researchers and experts should provide a step-by-step description of the characteristics and advantages/disadvantages of the algorithm. Therefore, explainable AI (XAI) technology has attracted attention as a new research field [22]. In the USA, the Defense Advanced Research Projects Agency (DARPA) is the leader in XAI research and has predicted the development of AI [23]. **Fig. 6** presents the research conducted on XAI for health big data.

	Characteristic	Implementati On method	Data type	Case	Problem
1th Generation AI	Logical, Association analysis through rules	Knowledge-based expert system	Sets of rules	Deep Blue (AI Chess)	Exception outside the rules
2th Generation AI	Probabilistic, Understanding patterns through big data analysis	Sensor / Data-based deep learning system	Training dataset	AlphaGo (AI Go) Autonomous driving	Opacity for the result
3th Generation AI	Understanding of outcomes through describing relationships	Intelligent system using visual verbal expression and cognitive agent	Description of model, Interface	eXplainable LIME	-

Fig. 6. XAI research on health big data [23]

The study being conducted as part of DARPA's XAI program will continue until 2021 and will comprise 11 sub-projects after 2017. Among the typical companies, H2O.ai is representatively studying explainable AI [24], and Microsoft will provide it through Azure. In particular, Kyndi is conducting research on XAI in the healthcare field. DAPRA has divided XAI into an explainable model that shows the interior of the aforementioned black box and an explanation interface for users. To develop an explainable model in the XAI study, the development strategy of the explainable model is possible, as shown in **Table 2**.

Table 2. Strategy for the development of the explainable model

Algorithm	Characteristics
In-depth explanation learning [25]	- Develop a deep-learning technology that improves the method by attaching the explanation label to the nodes of the hidden layer of the neural network and transforms or supplements the existing neural network into a hybrid form - Perform semantic interpretation to reach the final conclusion by backtracking the nodes on which the network focuses its attention
Decision-tree [26]	- Use machine learning to learn decision-tree logic to explain the neural-network operation in connection with the decision-tree process - Check the consistency of the results via a combination of the learning method with high interpretation, such as decision trees - Use an explanatory model in the form of a tracer
Model Inference [27]	- Infer and explain the results of the black box model through experiments and observations as a separate statistical model

There are two methodologies that can be used for explaining the operation of big-data-processing algorithms: sensitivity analysis (SA) [28] and layer-wise relevance propagation (LRP) [29]. The SA evaluates the change in a result depending on the type of input data. The contribution of the final result is quantified and explained according to each part or item of data. LRP explains the final result by describing the decomposition of layers in a hierarchical model, such as a deep neural network. It works as a method for identifying the amount of change in the result as the input changes in each layer. In this methodology, the contribution of each item or layer is visualized as graphs and images, which are further provided to users. LRP comprises the use of a backpropagation algorithm that is implemented during the learning phase of neural networks for the purpose of visualization. The general neural network algorithm backpropagates the contribution to each node of the previous layer based on the learned weights. For visualization, the contribution of the hierarchical model is constructed in the form of a heat map in the backpropagation step. The heat map of each layer can be visualized and expressed comprehensively. The user can intuitively observe which parts of the neural network have had a significant effect on the results. LRP is more useful in image analyses and provides a clear basis for disease judgment in medical image analyses, thus facilitating its effective use by medical personnel for verifying information. It is also useful for explaining the operation results of PilotNet, which is NVIDIA's deep-learning-based autonomous driving control system [30].

3.2 Internal Analysis of the Health Big-Data Algorithm

Representative big-data-algorithm analysis models include local interpretable model-agnostic explanations (LIME). LIME provides a technique for interpreting the results of a big-data-processing model [31]. There are various ways to understand the results of image classification in big data. Ribeiro [32] presented a method for identifying the major factors using images. In this method, visualization was used to determine which parts of the image were important. LIME comprises the use of a method of dividing an image into several smaller parts and checking the score change. This method is called the super pixel method [33]. LIME can be applied to various algorithms, such as neural networks, random forests, support vector machines (SVMs), and heterogeneous forms (e.g., numerical data, images, and text). Therefore, the results of various black box models can be interpreted in a reliable way. LIME identifies variables that are important for predicting results by approximating the model as an

interpretable linear model. **Fig. 7** illustrates the LIME process.

The LIME model has a process for predicting the expression of a specific disease, which is presented through an explainer. In the process of implementing the algorithm, the explainer analyzes the impact of the input data and output results. The explainer analyzes the influence of the input data list and the prediction by weight. The magnitude of the weight and the positive and negative effects are relatively analyzed to highlight the important symptoms that affect the results. This helps medical practitioners to definitively diagnose a patient's condition. In addition, algorithms such as Shapley additive explanations (SHAP) have been studied for general use in machine learning [34]. SHAP measures the importance of attributes. To this end, LIME is complemented by the integration of a number of algorithms, such as game theory and local explanations. R packages, such as the XGBoost Explainer, show the inside of an algorithm comprising XGBoost as a white box. The XGBoost Explainer outputs the effect analysis at the terminal of the decision tree in a table form. This allows the ensemble model to be organized in the form of a transparent and easy-to-understand graph and to analyze the internal tree structure.

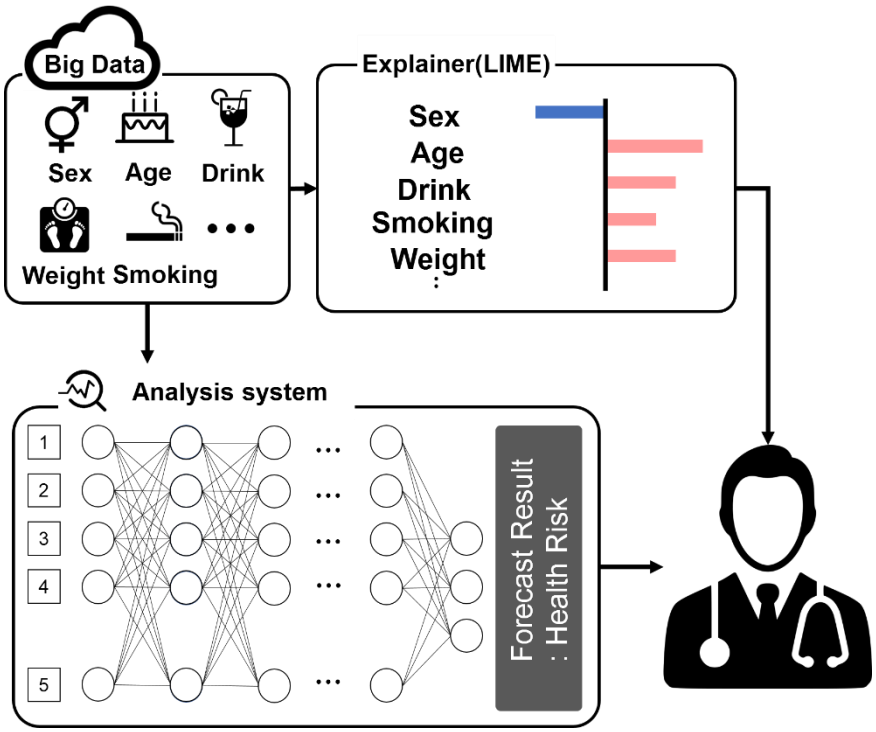

Fig. 7. LIME process

As shown above, XAI provides information to explain the interpretation of the algorithm analysis. The user can understand the system results. Moreover, the researcher can check the result of the model's predictive evaluation more intuitively beyond a simple accuracy evaluation. It also provides assistance in understanding the internal workings of the model.

4. Health-Risk-Prediction Technology Comprising Data Mining

4.1 Mining Multi-layer Association Rules in Health Transactions

Multi-layer association rule mining is a method of discovering multi-dimensional relationships between variables and attributes that occur frequently in a health transaction and that have undergone the pre-treatment step. As a method of discovering multidimensional association rules from frequent item sets, it determines association rules through the static discretization of quantitative attributes. This is a method of extracting frequent items from a transaction and finding a rule as per which the association between different independent items satisfies a minimum level of support. To improve the efficiency and scalability, frequent item set mining with transaction reduction, splitting, and sampling or without candidates has been developed in recent years. Data mining can be used in health transactions to discover hidden relationships such as the cause of a disease, complications, treatments, and the relationship with a disease. Multi-layer association rule mining also finds associations between items for health-risk prevention [35]. Hypertension data is collected from the Korea Centers for Disease Control and Prevention (KCDC). KCDC provides health information on disease definitions, causes and risk factors, symptoms, complications, and treatment [36]. The collected hypertension data comprises transactions through preprocessing to determine the association rule that satisfies the minimum map of 0.7 and the reliability of 0.8. **Fig. 8** shows the support and lift of multi-layer association rule mining performed for hypertension data [35, 37].

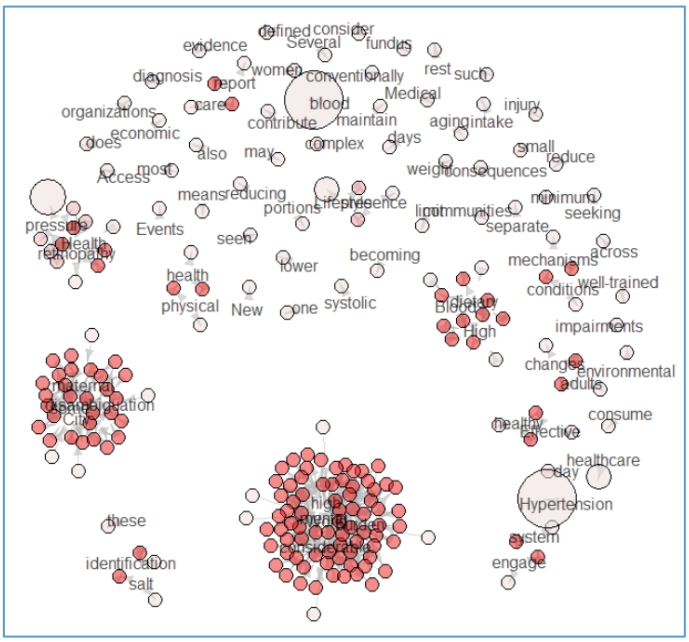

Fig. 8. Support and lift of multi-layer association rule mining in hypertension data [35, 37]

The size and color of a node represent the level of support and the size of the lift, respectively. The darker the color of the node, the higher the lift. The larger the size of the node, the higher the level of support.

K. Xia et al. [38] developed methods for the treatment and diagnosis of chronic diseases through association mining analyses. In the optimization stage of the diagnosis and treatment,

the association rule of frequent pattern growth and the Apriori algorithm were used to find correlations in the clinical data, and the association rules of the clinical treatment were generated. This optimized the clinical pathway, thus improving the associated cost and medical quality.

4.2 Disease-Risk Prediction and Classification using a Regression Analysis

A regression analysis can be used to mathematically estimate linear correlations in health data and model them using independent and dependent variables. Independent variables, also called explanatory variables, are causative variables that are necessary for obtaining predictions. Dependent variables, also called target and response variables, are the results of predictions. A regression analysis used for disease-risk prediction determines the extent to which independent variables affect the dependent variables through causal relationships. A regression analysis uses linear, multiple, and nonlinear regression. A linear regression models the linear correlation of dependent and independent variables and is classified as either simple linear or multiple linear regression depending on the number of independent variables [39]. Regression analyses can be used on a patient's medical data to predict the risk of disease. Colon-cancer-patient information uses colon data from R's survival package [37]. The attributes of the data consist of age, sex, cancer status, censorship status, etc. in the form of categorical or continuous numbers. For example, a gender category of 1 or 2 indicates male or female, respectively. Colon data are extracted from independent variables to predict them using censorship status as target variables, and then the influence and predictability of the independent variables are identified. **Fig. 9** presents the results of a regression analysis of colon-cancer-patient data. Here, the dotted line represents the Cook's distance, and the residuals and leverage that have undergone normalization describe the influence on the data. The horizontal axis represents the influence of the variable, and the vertical axis represents the Pearson residual, which indicates how well the model predicts the observed values [39].

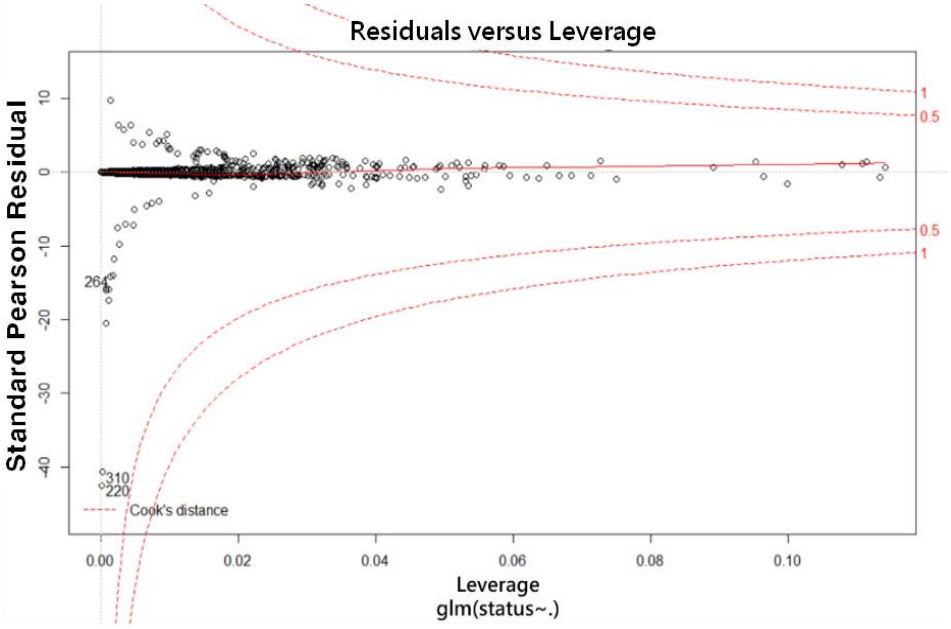

Fig. 9. Results of a regression analysis on colon-cancer-patient data [39]

Fig. 9 shows the results of the application of a linear regression model that generalized the treatment progress of colon-cancer patients as cure, recurrence, and death. The dependent variable was designated as the status, and the independent variable comprised the remaining influence factors. The predicted results for the treatment effect of colon-cancer patients had a small Pearson residual value, and therefore, the predictive model of the regression analysis model was considered to be appropriate. G. Manogaran et al. [40] used the stochastic gradient descent (SGD) method and a scalable logistic regression analysis to analyze the health risk. An SGD algorithm was used to develop scalable diagnostic and logistic regression models. They also developed a scalable data structure and disease prediction model for cloud computing to determine the health risk.

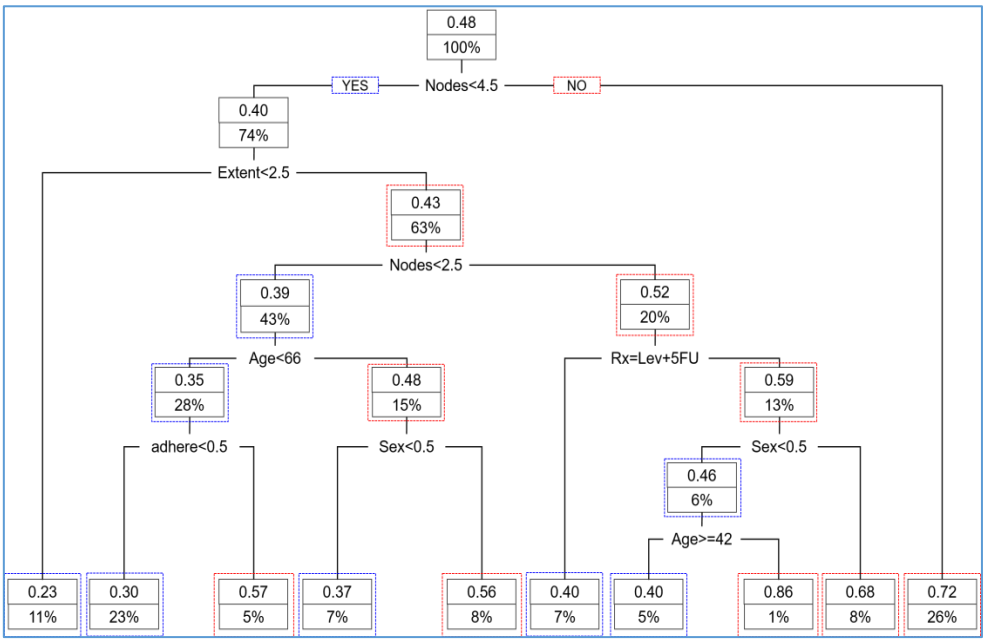

Fig. 10. Decision tree of the classification results of colon-cancer-patient data [35, 37, 41]

Classification is a method of constructing health data via data purification, relevance analysis, and data transformation, and belongs to the class label according to a range of predefined attributes [42]. The classification of health data comprises the use of decision trees, random forests, the k-nearest neighbors algorithm, SVMs, and neural networks. The patient data are classified into each class, such as the presence or absence of diabetes and hypertension, and the characteristics are extracted. As a result, diet, exercise, treatment, or other suitable recommendations are provided for patient management. **Fig. 10** shows the classification results presented in a decision tree for colon-cancer-patient data from R's survival package [37]. K. Dauda et al. [41] developed a decision-making model for survival data that includes competing risks. The decision tree is constructed using the classification and regression tree algorithm to process the validated data for the regression and classification trees. R. Vijayarajeswari et al. [43] developed a classification method for the early detection of breast cancer using an SVM classifier and the Hough transform. The Hough transform extracts features of a particular shape from an image and classifies them using the SVM. This method can be effectively used to classify images of X-rays that have been abnormally obtained.

4.3 Ensemble Technique for Predictions

The ensemble model method is used to derive the most appropriate prediction results using the prediction results of various models. It is a method of creating various prediction models based on the given health data. The main evaluation methods include bagging predictors using a simple majority vote method, random forest method, and weighting boosting method.

L. Breiman [44] introduced the bagging of predictors, which is known as bagging and is a bootstrap aggregating algorithm and type of ensemble method. After the creation of bootstrap data and a corresponding predictive model, the ensemble method is applied to the result. A simple majority vote method includes a random forest. The random Forest method comprises creating several decision trees with randomness and decorrelation and then determining the result by a majority vote. This structure is also useful for data that includes random forest noise. The randomization of the tree is constructed through the bagging process, trains the tree through the training dataset, and combines it by a majority vote method. This addresses the disadvantage of decision trees being likely to poorly overfit new data. Boosting is an ensemble method that comprises the use of weights and was studied by Y. Freund and R. Schapire [45]. This method weights the error data with poor predictions of the boosting model. By modifying models that present negative results, the susceptibility to overfitting is reduced. In addition, even if the performances of individual models are poor, the final model provides improved results. Adaptive boosting (AdaBoost) is a basic boosting method. AdaBoost can be used with algorithms such as decision tree learning, and it learns with a focus on more difficult data.

The gradient boosting machine (GBM) is a machine learning technique that combines gradient descent with boosting [46]. The GBM is a concept that connects many simple models of shallow trees. The GBM is constructed in a manner that compensates for errors in the previous tree, such that the core of the GBM comprises error correction of the previous tree. Gradient descent and the learning rate are used for error correction. Complex models can be constructed according to the learning rate. Using a relatively shallow tree, GBM uses less memory, performs better, and is able to perform regression and classification analyses. In particular, it performs well for X–Y grid-type data and provides an excellent prediction performance as compared with other machine learning algorithms [47].

Recently, various derived algorithms and packages have been developed to take advantage of the superior performance of the GBM, e.g., the Python-based packages such as XGBoost [48], Light GBM [49], and CatBoost [50]. These improve the performance of GBMs and are applied to big data processing, which requires a significant number of computations [51]. Various methods have been attempted to implement the hardware efficiently. **Table 3** shows the types of boosting algorithms.

Table 3. Types of boosting algorithms

Year	Algorithm	Characteristics
2019	XGBoost	- Distributed, parallel processing combined with performance verification through Kaggle
2019	Light GBM	- Improved performance and minimized resource consumption compared to XGBoost - Improved performance through approximations of the split

In general, the health big-data algorithm presents the problem of deep dimensionality occurring in the learning form [52]. There is a stereotypical data form with an effective performance according to the type of machine learning. Therefore, the form or setting value should be adjusted for the algorithm performance. The concept of the boosting algorithm is more general and provides an effective performance with the use of fewer parameters. In addition, by selecting effective feature data, it reduces the number of dimensions of the health big-data learning network and improves the execution time.

5. Conclusion

The key to researching the health big-data system is the acquisition of various data and accuracy of data analyses. Recently developed health big-data analysis algorithms show positive effects in terms of their accuracy and speed. These provide personalized healthcare services while reducing medical expenses and time required. In addition, it is possible to provide medical professionals with analyses, research results in a short time, simulations, and predictions of the toxicity and side effects of drugs. A healthcare cloud system protects personal privacy and improves data management costs efficiently. In addition, more advanced explainable big-data-processing technologies provide users with explainable predictive results. Recently, using the mining multi-layer association rules and regression analysis, an attempt has been made to develop a method for predicting the risk of disease and the hidden relationships such as the cause of the disease, complications, treatment, and the relationship with the disease. In addition, ensemble models use the prediction results of various models to derive more effective prediction results. In particular, XAI technology can be used to visualize the decision process of AI models and explain the elements of deep-learning models involved in decision making. In the future, XAI is expected to be developed in the direction of creating an automated report or interactively by combining it with the technology of expressing human sentences. This would allow experts to understand the contents of an analysis and provide a reasonable basis for decision-making.

Research on these techniques may contribute to the development of AI systems in various fields, including law, finance, economics, and medical treatments. In addition, these are expected to negate the concerns regarding automation systems and provide highly reliable information to a user.

References

[1] T. Muhammed, R. Mehmood, A. Albeshri, and I. Katib, “UbeHealth: A Personalized Ubiquitous Cloud and Edge-Enabled Networked Healthcare System for Smart Cities,” *IEEE Access*, vol. 6, pp. 32258-32285, June 2018. Article (CrossRef Link)

[2] M. Alam, H. Malik, M. Khan, T. Pardy, A. Kuusik, and Y. Moullec, “A Survey on the Roles of Communication Technologies in IoT-based Personalized Healthcare Applications,” *IEEE Access*, vol. 6, pp. 36611-36631, July 2018. Article (CrossRef Link)

[3] A. Alaiad and L. Zhou, “Patients' Adoption of WSN-Based Smart Home Healthcare Systems: An Integrated Model of Facilitators and Barriers,” *IEEE Transactions on Professional Communication*, vol. 60, no. 1, pp. 4-23, Mar. 2017. Article (CrossRef Link)

[4] M. Chen, W. Li, Y. Hao, Y. Qian, and I. Humar, “Edge cognitive computing based smart healthcare system,” *Future Generation Computer Systems*, vol. 86, pp. 403-411, Sep. 2018. Article (CrossRef Link)

[5] G. Manogaran, C. Thota, D. Lopez, V. Vijayakumar, K. Abbas, and R. Sundarsekar, "Big Data Knowledge System in Healthcare," *Internet of Things and Big Data Technologies for Next Generation Healthcare*, vol. 23, pp. 133-157, Jan. 2017. Article (CrossRef Link)
[6] R. Irfan, Z. Rehman, A. Abro, C. Chira, and W. Anwar, "Ontology Learning in Text Mining for Handling Big Data in Healthcare Systems," *Journal of Medical Imaging and Health Informatics*, vol. 9, no. 4, pp. 649-661, May 2019. Article (CrossRef Link)
[7] E. Beulah, S. Rajini, N. Selvaraj, and R. Narayanan, "Application of Data Mining in Healthcare: A Survey," *Asian Journal of Microbiology, Biotechnology and Environmental Sciences*, vol. 18, no. 4, pp. 999-1001, Dec. 2016. Article (CrossRef Link)
[8] M. Suguna, M. G. Ramalakshmi, J. Cynthia, and D. Prakash, "A Survey on Cloud and Internet of Things based Healthcare Diagnosis," in *Proc. of Computing Communication and Automation*, pp. 1-4, Dec. 2018. Article (CrossRef Link)
[9] A. Omar, Z. Bhuiyan, A. Basu, S. Kiyomoto, and M. Rahman, "Privacy-friendly Platform for Healthcare Data in Cloud based on Blockchain Environment," *Future Generation Computer Systems*, vol. 95, pp. 511-521, June 2019. Article (CrossRef Link)
[10] Y. Karaca, M. Moonis, Y. D. Zhang, and C. Gezgez, "Mobile Cloud Computing based Stroke Healthcare System," *International Journal of Information Management*, vol. 45, pp. 250-261, Apr. 2019. Article (CrossRef Link)
[11] M. Rahman, I. Khalil, and X. Yi, "A Lossless DNA Data Hiding Approach for Data Authenticity in Mobile Cloud based Healthcare Systems," *International Journal of Information Management*, vol. 45, pp. 276-288, Apr. 2019. Article (CrossRef Link)
[12] R. Ganiga, R. M. Pai, and R. Sinhaa, "Private Cloud Solution for Securing and Managing Patient Data in Rural Healthcare System," *Procedia Computer Science*, vol. 135, pp. 688-699, 2018. Article (CrossRef Link)
[13] M. Pham, Y. Mengistu, H. Do, and W. Sheng, "Delivering Home Healthcare through a Cloud-based Smart Home Environment (CoSHE)," *Future Generation Computer Systems*, vol. 81, pp. 129-140, Apr. 2018. Article (CrossRef Link)
[14] S. Miah, J. Hasan, and J. Gammack, "On-Cloud Healthcare Clinic: An E-health Consultancy Approach for Remote Communities in a Developing Country," *Telematics and Informatics*, vol. 34, no. 1, pp. 311-322. Feb. 2017. Article (CrossRef Link)
[15] P. Verma and S. Sood, "Cloud-Centric IoT based Disease Diagnosis Healthcare Framework," *Journal of Parallel and Distributed Computing*, vol. 116, pp. 27-38, June 2018. Article (CrossRef Link)
[16] T. Bhardwaj and S. Sharma, "Cloud-WBAN: An Experimental Framework for Cloud-enabled Wireless Body Area Network with Efficient Virtual Resource Utilization," *Sustainable Computing: Informatics and Systems*, vol. 20, pp. 14-33, Sep. 2018. Article (CrossRef Link)
[17] K. Chung and R. Park, "P2P Cloud Network Services for IoT based Disaster Situations Information," *Peer-to-Peer Networking and Applications*, vol. 9, no. 3, pp. 566-577, May 2016. Article (CrossRef Link)
[18] P. Sahoo, S. Mohapatra, and S. Wu, "Analyzing Healthcare Big Data with Prediction for Future Health Condition," *IEEE Access*, vol. 4, pp. 9786-9799, Nov. 2016. Article (CrossRef Link)
[19] K. Chung and R. Park, "Cloud based U-healthcare Network with QoS Guarantee for Mobile Health Service," *Cluster Computing*, vol. 22, no. 1, pp. 2001-2015, Jan. 2019. Article (CrossRef Link)
[20] M. Zayoud, Y. Kotb, and S. Ionescu, "β Algorithm: A New Probabilistic Process Learning Approach for Big Data in Healthcare," *IEEE Access*, vol. 7, pp. 78842-78869, June 2019. Article (CrossRef Link)
[21] H. Kalantarian, K Jedoui, P Washington, Q Tariq, K Dunlap, J Schwartz, and D. P.Wallab, "Labeling Images with Facial Emotion and the Potential for Pediatric Healthcare," *Artificial Intelligence in Medicine*, vol. 98, pp. 77-86, July 2019. Article (CrossRef Link)
[22] A. Adadi and M. Berrada, "Peeking inside the Black-box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138-52160, Sep. 2018. Article (CrossRef Link)

[23] Defense Advanced Research Projects Agency, "Explainable Artificial Intelligence (XAI)," *DARPA presentation*, Nov. 2017. Article (CrossRef Link)
[24] P. Hall, M. Kurka, and A. Bartz, "Using H2O Driverless AI," *H2O.ai*, 2018. Article (CrossRef Link)
[25] J. Choo and S. Liu, "Visual Analytics for Explainable Deep Learning," *IEEE Computer Graphics and Applications*, vol. 38, no. 4, pp. 84-92, 2018. Article (CrossRef Link)
[26] S. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 21, no. 3, pp. 660-674, 1991. Article (CrossRef Link)
[27] H. Kuwajima, M. Tanaka, and M. Okutomi, "Improving transparency of deep neural inference process," *Progress in Artificial Intelligence*, vol. 8, pp. 273-285, Apr. 2019. Article (CrossRef Link)
[28] E. Borgonovo and E. Plischke, "Sensitivity analysis: A review of recent advances," *European Journal of Operational Research*, vol. 248, no. 3, pp. 869-887, Feb. 2016. Article (CrossRef Link)
[29] S. Bach, A. Binder, G. Montavon, F. Klauschen, K. Müller, and W. Samek, "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation," *PloS ONE*, July 2015. Article (CrossRef Link)
[30] M. Bojarski, P. Yeres, A. Choromanska, K. Choromanski, B. Firner, L. Jackel, and U. Muller, "Explaining How a Deep Neural Network Trained with End-to-End Learning Steers a Car," *arXiv:1704.07911*, Apr. 2017. Article (CrossRef Link)
[31] K. Sokol, A. Hepburn, R. Santos-Rodriguez, and P. Flach, "bLIMEy: Surrogate Prediction Explanations Beyond LIME," *arXiv:1910.13016*, Oct. 2019. Article (CrossRef Link)
[32] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You?": Explaining the Predictions of Any Classifier," in *Proc. of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135-1144, Aug. 2016. Article (CrossRef Link)
[33] R. Achanta, A. Shaji, K. Smith, A Lucchi, P. Fua, and S. Süsstrunk, "SLIC Superpixels Compared to State-of-the-Art Superpixel Methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2274-2282, May 2012. Article (CrossRef Link)
[34] S. Lundberg and S. Lee, "A Unified Approach to Interpreting Model Predictions," in *Proc. of the 31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017. Article (CrossRef Link)
[35] N. Kumar, A. Gangopadhyay, and G. Karabatis, "Supporting Mobile Decision Making with Association Rules and Multi-layered Caching," *Decision Support Systems*, vol. 43, no. 1, pp. 16-30, Feb. 2007. Article (CrossRef Link)
[36] Korea Centers for Disease Control and Prevention (KCKC). [Online]. Available: http://health.cdc.go.kr/
[37] CRAN Packages, Survival: Survival Analysis. [Online]. Available: https://cran.rstudio.com/web/packages/survival/
[38] K. Xia, X. Zhong, L. Zhang, and J. Wang, "Optimization of Diagnosis and Treatment of Chronic Diseases based on Association Analysis Under the Background of Regional Integration," *Journal of medical systems*, vol. 43, no. 46, pp. 1-8, Mar. 2019. Article (CrossRef Link)
[39] M. Sato-Ilic, "Knowledge-based Comparable Predicted Values in Regression Analysis," *Procedia Computer Science*, vol. 114, pp. 216-223, 2017. Article (CrossRef Link)
[40] G. Manogaran and D. Lopez, "Health Data Analytics using Scalable Logistic Regression with Stochastic Gradient Descent," *International Journal of Advanced Intelligence Paradigms*, vol. 10, no 1-2, pp. 118-132, Jan. 2018. Article (CrossRef Link)
[41] K. Dauda, B. Pradhan, B. Shankar, and S. Mitra, "Decision Tree for Modeling Survival Data with Competing Risks," *Biocybernetics and Biomedical Engineering*, vol. 39, no. 3, pp. 697-708, 2019. Article (CrossRef Link)
[42] M. Hosni, I. Abnane, A. Idri, J. Gea, and J. Alemán, "Reviewing ensemble classification methods in breast cancer," *Computer Methods and Programs in Biomedicine*, vol. 177, pp. 89-112, Aug. 2019. Article (CrossRef Link)

[43] R. Vijayarajeswari, P. Parthasarathy, S. Vivekanandan, and A. Basha, "Classification of Mammogram for Early Detection of Breast Cancer using SVM Classifier and Hough Transform," *Measurement*, vol. 146, pp. 800-805, Nov. 2019. Article (CrossRef Link)
[44] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123-140, Aug. 1996. Article (CrossRef Link)
[45] Y. Freund and R. Schapire, "A Decision-theoretic Generalization of On-line Learning and An Application to Boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119-139, Aug. 1997. Article (CrossRef Link)
[46] J. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189-1232, Oct. 2001. Article (CrossRef Link)
[47] S. Oh, K. Chung, and J. Han, "Towards Ubiquitous Health with Convergence," *Technology and Health Care*, vol. 24, no. 3, pp. 411-413, 2016. Article (CrossRef Link)
[48] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. of KDD '16: the 2nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785-794, Aug. 2016. Article (CrossRef Link)
[49] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T. Liu, "LightGBM: a highly efficient gradient boosting decision tree," in *Proc. of the 31st International Conference on Neural Information Processing Systems*, pp. 3149-3157, Dec. 2017. Article (CrossRef Link)
[50] L. Prokhorenkova, G. Gusev, A. Vorobev, A. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," in *Proc. of NIPS'18: the 32nd International Conference on Neural Information Processing Systems*, pp. 6639-6649, Dec. 2018.
Article (CrossRef Link)
[51] K. Chung and J. Kim, "Activity based Nutrition Management Model for Healthcare using Similar Group Analysis," *Technology and Health Care*, vol. 27, no. 5, pp. 473-485, Sep. 2019.
Article (CrossRef Link)
[52] C. Zhang and J. Zhang, "A local boosting algorithm for solving classification problems," *Computational Statistics & Data Analysis*, vol. 52, no. 4, pp. 1928-1941, Jan. 2008.
Article (CrossRef Link)

Hyun Yoo has received B.S., M.S., and Ph.D. degrees in 1999, 2011, and 2019, respectively, all from the Department of Computer & Information Engineering, Sangji University, South Korea. He has worked for Fujitsu and XCE (SK). He has been a researcher at Data Mining Lab., Kyonggi University, South Korea. He is currently a research professor in the Contents Convergence Software Research Center, Kyonggi University, South Korea. His research interests include Deep Learning, Artificial Intelligent, Big Data Mining, Ambient Intelligence, Medical Decision System, Healthcare, Recommendation, and HCI.

Roy C. Park has received the B.S. degrees from Dept. of Industry Engineering, and M.S., Ph. D. degrees from Dept. of Computer Information Engineering, Sangji University, South Korea, in 2010 and 2015. From 2015 to 2018, he was a professor in the Division of Computing Engineering, Dongseo University, Korea. Since 2019, he is currently a professor in the Department of Information Communication Software Engineering, Sangji University, Wonju, South Korea. His research interests include WLAN System, Heterogeneous Network, Ubiquitous Network Service, Human-Inspired Artificial Intelligent and Computing, Health Informatics, Knowledge System, Peer-to-Peer, and Cloud Network.

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 in 2000, 2002, and 2005, respectively. He has worked for the Software Technology Leading Department, Korea IT Industry Promotion Agency (KIPA). From 2006 to 2016, he was a professor at the School of Computer Information Engineering, Sangji University, South Korea. Since 2017, he is currently a professor in the Division of AI Computer Science and Engineering, Kyonggi University, South Korea. He was named a 2017 Highly Cited Researcher by Clarivate Analytics. His research interests include data mining, artificial intelligence, healthcare, knowledge systems, HCI, and recommendation systems.