IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 16 December 2022, accepted 31 January 2023, date of publication 9 February 2023, date of current version 17 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3243722

APPLIED RESEARCH

Multi-Context Mining-Based Graph Neural Network for Predicting Emerging Health Risks

JI-WON BAEK[1] **AND KYUNGYONG CHUNG**[2]
[1]Department of Computer Science, Kyonggi University, Suwon, Gyeonggi 16227, South Korea
[2]Division of AI Computer Science and Engineering, Kyonggi University, Suwon, Gyeonggi 16227, South Korea

Corresponding author: Kyungyong Chung (dragonhci@gmail.com)

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2020R1A6A1A03040583). Additionally, this work was supported by Kyonggi University's Graduate Research Assistantship 2021.

ABSTRACT Patients with similar diseases are able to have similar treatments, care, symptoms, and causes. Based on these relations, it is possible to predict latent risks. Therefore, this study proposes Graph Neural Network-based Multi-Context mining for predicting emerging health risks. The proposed method first, collects and pre-processes chronic disease patients' disease information, behavioral pattern information, and mental health information. After that, it performs context mining. This is a multivariate regression analysis for predicting multiple dependent variables, it extracts a regression model and generates a feature map. Then, the initial graph is created by defining the number of clusters as nodes and constructing edges through correlation. By expanding the graph according to the results of context mining, it is possible to predict that a user has a similar chronic disorder and similar symptoms through users' connection relations. For performance evaluation, the validity of the regression analysis of context mining used in the proposed method, and the suitability of the clustering technique are evaluated.

INDEX TERMS Multi-context mining, graph neural network, emerging health risk, healthcare, knowledge, recommendation.

I. INTRODUCTION

Today, inappropriate living habits cause an increase in the number of patients with chronic diseases. Various changes in living environments influence people's mental health, and individuals have different living patterns. Also, with the increase in life expectancy, people are more interested in healthcare. Chronic disease is a long-term health condition that may improve and worsen repeatedly. Unless it is cared for or prevented, it can cause complications [1]. Therefore, it is necessary to urge patients to pay attention to such health risks.

The development of information technology and artificial intelligence draws a lot of attention. Along with that, there has been active research on data analysis for predicting results using reinforcement learning and machine learning based on different data comprising numbers, images, videos, etc., and the subsequent extraction of significant information. Data analysis methods have differences depending on their purpose [2]. For example, the regression analysis method is applied to analyze causal relations between dependent and independent variables. Accordingly, Baek et al. [3] proposed the multiple regression-based ContextDNN for predicting the risk of depression. Aimed at predicting the risk of mental health, the proposed method is to design a Context that represents a set of context information including surrounding conditions and time, and to apply it to a neural network. Furthermore, it establishes neural networks and connects the individually learned neural networks via the regression formula. Thus, it paves the way for predicting latent situations that influence mental health. However, this approach is limited by a dependent variable. In practice, because there are multiple dependent variables, it is necessary to consider them all. Clustering analysis is used to classify similar or related data into multiple groups. It has no pre-defined special purposes. Nevertheless, its advantage is that it relies on data and obtains meaningful information for all data. Jung et al. [4] proposed the social mining-based clustering process for big data integration. For a reliable model, the proposed method is used to apply different weight values through static model

The associate editor coordinating the review of this manuscript and approving it for publication was Dian Tjondronegoro.

information and information obtained from social networks, depending on user relations. Clustering for health conditions of survivors of an illness enables the prediction of health risks, thereby helping to improve health conditions based on the risk of medical accidents and expectancy. In such a case, a sufficient amount of data is required for modeling. Because of repeated scanning for pattern extraction, it takes longer to draw the analysis results. For this reason, it is necessary to devise a method for analyzing continuously growing data efficiently.

Social network analysis, recommendation systems, and knowledge graphs have been actively employed in practice. With the increase in graph applicability, the graph neural network (GNN) has been actively researched. A graph is the result of a set of nodes connected with directional or unidirectional edges. In addition, relationships or interrelationships can be structured and presented visually [5]. Because the node size, number of neighboring nodes, and features differ, a graph's structure is irregular. To solve this problem, GNNs are applied. GNN is a neural network for graphs that support node classification, connection prediction, and graph classification. Node classification is the process of classifying a node through node embedding under the condition that part of the graph is labeled. Linked prediction is the process of finding relations between nodes and predicting the degree of association and correlation between two nodes. It is widely applied in recommendation systems [6]. In addition, research is being conducted in the field of healthcare based on graphs. For example, Dong et al. [7] proposed to applying graph representations to recognize similar symptoms of influenza based on people's daily mobility, social interaction, and physical activity. However, the analysis of the dynamic interaction between disease symptoms and human behavior is insufficient. Therefore, when GNNs are formed based on social networks, it is possible to identify interactions between people with similar symptoms and behavior. Graph classification is the process of classifying graphs into various categories. A graph is defined as a connection between neighboring nodes. Accordingly, if a particular node removes a connection with its neighbor, the node is isolated and is meaningless in the graph. GNNs are categorized into recurrent graph neural networks, spatial convolutional networks, and spectral convolutional networks. In RNN, the hidden layer of the past time step and the input layer of the present time step are connected to predict current data. The node used in a recurrent graph neural network is an RNN unit, and the network is designed differently depending on the edge form. Accordingly, all nodes can obtain information about their neighboring nodes [8]. A spatial convolutional network is employed for image classification or region segmentation. Therefore, its structure is similar to that of CNN, using the features of the nodes connected in the graph. A spectral convolutional network is developed based on graph signal processing, including mathematical factors. By sharing and updating node information effectively, it expands a graph. For example, Tao et al. [9] proposed the item trend learning method for a sequence-based recommendation system using a gate graph neural network. The proposed method learns item trend information from interaction logs of implicit users and integrates recommendations with trend information of items. Consequently, it improves the accuracy of representation through a self-attention mechanism. By integrating a user's short-term preference with recommendation items, it is possible to improve the accuracy of recommendations and offer custom representations to a user. In other words, the method is used to integrate item trend information to improve the current recommendation item. Because it designs a model with the existing data, it performs poorly in predicting new data. Therefore, it is necessary to provide a solution to the cold start problem in the recommendation system.

This study proposes a multi-context mining-based graph neural network for predicting emerging health risks. The proposed method aims to determine the similarity relations between chronic disease patients according to their behavioral patterns and mental health to predict the risks of chronic disease patients who have similar features and increased awareness of health care and prevention. The contributions of the method proposed in this paper are as follows:

- Identify the causal relationship between chronic diseases. It is possible to grasp the causal relationship by generating a feature map based on the influencing factors of chronic diseases.
- Early graphs were constructed through clustering and correlation of similar diseases through context mining. Therefore, relationships such as similar behavior patterns, diseases, and symptoms may be formed in each clustered user and users between clusters.
- Graph relationship representations allow us to measure potential risks in other users.

This of paper is composed as follows: in chapter 2, we describe deep learning-based relationship prediction in recommendation systems and GNN-based relation prediction applications. In chapter 3, the proposed multi-context mining-based graph neural network is described for predicting emerging health risks. In chapter 4, the recommendation results and performance evaluation are described, and in chapter 5, the conclusions drawn from this study are described.

II. RELATED WORK

A. RECOMMENDATION METHOD USING RELATIONSHIP PREDICTION BASED ON GRAPH NEURAL NETWORK

With the development of ICT, it is possible to collect multiple forms of data in various ways anytime and anywhere. As a result, the amount of data generated is huge and diverse, so each data has a variety of features [10]. In various fields including medical service and traffic control, summary, statistics, decision making, knowledge search, and pattern analysis are applied to extract and employ significant information. However, data missing, bias, contingency, and other problems arise according to real-time data collection methods, devices, and collection targets. As a result, a data shortage problem occurs during analyses [11]. Data augmentation is

applied to solve this problem. It is a technique of augmenting small data through diverse algorithms. With the technique, it is possible to solve the data set shortage problem and to consider diverse situations that change differently in real-time. In fact, data for considering these situations have high dimensions and are complex, so that it is hard to find relations between data [12]. To solve the problem a graph is employed. A general graph analysis method requires empirical or preliminary knowledge through breadth-first search, depth-first search, shortest path algorithm, and clustering. Therefore, if there are multiple graphs, it is hard to extract significant information. To continue to present massive data in a graph, it is easy to lose the structure of the graph. For these reasons, Graph Neural Network (GNN), in which deep learning is applied to graphs, is employed [13]. A graph consists of nodes and edges. By connecting data that has various types of relations, it is possible to conduct an analysis. GNN is an effective framework for learning graph representation [14]. In a GNN, a node calculates a new feature vector while collecting information from its neighboring nodes repeatedly. After K repeated operations, the node expresses the structural information in its k-hop neighbor as the captured and converted functional vector. Accordingly, through pooling, it is possible to obtain the entire graph representation. A new design of GNN is mostly based on an empirical, heuristic, and experimental results. In addition, the features and theoretical results of GNN are merely found, and GNN uses a lot of data. Graph embedding [15] of GNN is a process of embedding a graph in a vector space in order for easier data analysis. With that, it is possible to solve diverse network problems in a vector space and to use the technique in a recommender system [16]. For example, Liu et al. [17] proposed the real-time social recommendation method based on graph embedding and temporal context. The proposed method is a new dynamic graph-based embedding model for recommending users and items of interest. For real-time recommendation, it establishes a heterogeneous user item (HUI) network. Dynamic graph embedding (DGE) shares the HUI network and builds it in a low-dimension space. Accordingly, it captures visual meaning effect, social relationship, and sequential pattern of user behavior in an integrated way. Through simple search or similarity calculation, it employs encoded expressions and generates recommendation items. For node representation learning, it, however, takes into account neighbor information only, so it is necessary to devise a method of considering the similarities of various users. Given that a graph has multiple features, it is necessary to employ a method for learning features. Figure 1 shows the Recommendation method using relationship prediction based on graph neural network process.

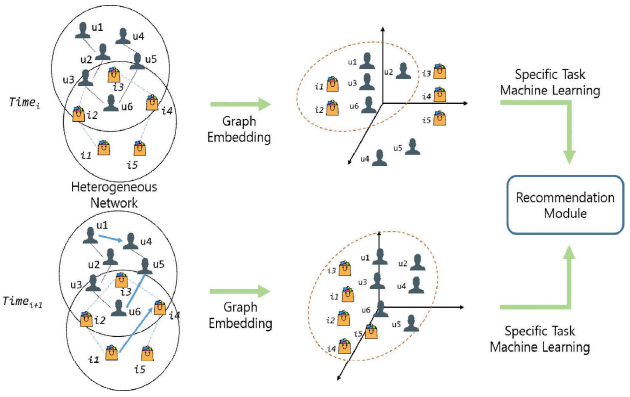


FIGURE 1. Recommendation method using relationship prediction based on graph neural network process.

III. MULTI-CONTEXT MINING-BASED GRAPH NEURAL NETWORK FOR PREDICTING EMERGING HEALTH RISKS

The proposed multi-context mining-based graph neural network for predicting emerging health risks aims to identify relationships between data on chronic diseases, mental health, and behavioral patterns and suggest knowledge for preventing latent emerging health risks. The knowledge recommendation model for emerging health risks consists of three steps: The first step is data preprocessing. In this step, the data from the National Health and Nutrition Examination Survey are collected; information on patients who have high blood pressure, diabetes, and dyslipidemia is obtained. In addition, these patients' information on mental health and behavioral patterns is collected. Unnecessary variables are removed. The second step is multi-context mining. In this step, multivariate and multiple regression analysis is conducted to detect the variables influencing high blood pressure, diabetes, dyslipidemia, mental health, and behavioral patterns from preprocessed data. Accordingly, each one of the prediction models is generated. Patients with similar chronic diseases have similar mental health symptoms and behavioral patterns. For this reason, chronic disease patients are clustered according to their mental health and behavioral patterns, and user relations are analyzed. The last step is a graph representation of relations between chronic disease patients and the expansion of the graph based on the results of context-mining. This step is aimed at finding not only invisible user relations but other user risks that appear in specific users. Figure 2 shows the process of the multi-context mining-based GNN for predicting emerging health risks.

A. DATA COLLECTION AND PREPROCESSING

The used in this study are the 7th raw data offered by national health and nutrition examination survey. These three-year data (2016 to 2018) are nationally representative and reliable data for people's health levels, health behavioral patterns, and food & nutrition [18]. The raw data of the National Health and Nutrition Examination Survey are collected in the health questionnaire surveys, health examination surveys, and other surveys, so that there are missing data. Therefore, improving the accuracy and reliability of the analysis is required to do preprocessing. From the raw data, data about chronic diseases (diabetes, high blood pressure, and dyslipidemia), mental health, and physical activities are collected. From them,

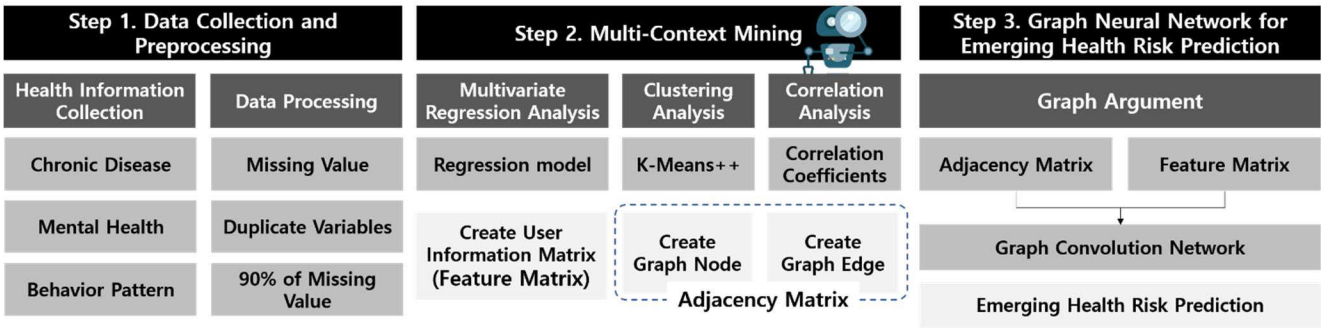


FIGURE 2. The process of the multi-context mining-based graph neural network for predicting emerging health risks.

69 variables and 24,269 persons' data are extracted. Items that have the answer of no idea, no answer, or no availability are defined as missing data and are preprocessed. Firstly, in terms of the missing value process, all the data of the population with missing values are removed. Secondly, if over 90% of data for a certain variable includes missing values, the variable is removed due to its unnecessary influence on the analysis. Thirdly, variables with duplicate meanings are removed. In the end, 50 variables and 189 persons' data are used. Table 1 shows preprocessed health data. This table is consisting of variables, a description of variables, and the content of variables.

TABLE 1. Preprocessed health data.

Variable	Variable Description	Contents
DI1_dg	Whether or not to diagnose high blood pressure	0: None 1: Exists
DI1_pr	Current prevalence of high blood pressure	0: None 1: Exists
DI1_pt	High blood pressure treatment	0: None 1: Exists
DE1_dg	Whether or not to diagnose diabetes	0: None 1: Exists
DE1_pr	Current prevalence of diabetes	0: None 1: Exists
DE1_pt	Diabetes treatment	0: None 1: Exists
...	...	...

B. FEATURE EXTRACTION USING MULTI-CONTEXT MINING

People face challenging situations in their daily activities because of health issues. These situations include influential latent factors that change in real-time. Along with changes in situations and context, preferences are changed in line with users' states and situations. Therefore, it is necessary to consider users' context changes over time and make recommendations through context mining. To analyze causal relations between variables, a probability for users' health conditions is generated in regression analysis. General linear regression analysis estimates relations between one dependent variable and one independent variable in a straight line. Multiple linear regression analysis estimates relations between at least two independent variables and one dependent variable [19], [20]. However, in reality, an independent variable is influenced by different variables, or there are multiple dependent variables. Therefore, multivariate multiple linear regression analysis is applied to generate a probability model of users' health conditions. This regression model identifies relations between variables when there are at least two dependent variables [21].

Firstly, Diagnosis of each chronic disease is set as a dependent variable, and then relations with independent variables analyze. A regression formula is generated with variables that meet the significance level of 0.05. Regression results present with the uses of the dependent variable, independent variable, estimated value, standard error, t-value, and Signif. (significance level). Table 2 shows the regression results for high blood pressure.

TABLE 2. The regression results for diabetes.

Dependent Variable	High Blood Pressure				
Independent Variable	DI1_pr	DI1_pt	HE_sbp1	HE_dbp1	...
Estimate	-0.6517	2.025	0.001736	-0.002351	...
Std. Error	0.1595	0.3649	0.002016	0.002708	...
t-value	-4.086	5.55	-0.861	-0.868	...
Signif.	0.00708	0.00124	0.3904	0.3866	...

Equation (1) shows the regression formula for high blood pressure. HBP (High Blood Pressure) represents high blood pressure, and 1.591 is a value of y-intercept.

$$\mathrm{HBP} = 1.591 + (-0.6517 \times \mathrm{DI1\_pr}) + (2.025 \times \mathrm{DI1\_pt}) \quad (1)$$

In Equation (1), variables that meet the significance level and influences high blood pressure, whether to have high blood pressure at present and whether to treat high blood pressure is used. Table 3 shows the regression results for diabetes.

TABLE 3. The regression results for diabetes.

Dependent Variable	Diabetes				
Independent Variable	HE_TG	HE_HTG	HE_glu	HE_HbA1c	···
Estimate	4.64E-17	-4.35E-15	3.77E-17	-1.66E-15	···
Std. Error	2.52E-17	4.41E-15	4.40E-17	1.56E-15	···
t-value	1.84E+00	-9.88E-01	8.57E-01	-1.07E+00	···
Signif.	0.0676	0.3249	0.0393	0.2876	···

Equation (2) shows the regression formula for diabetes. DM represents Diabetes Mellitus, and 3 is a value of y-intercept.

$$\begin{aligned} \mathrm{DM} = {} & 3 + ((3.77E-17) \times \mathrm{HE\_glu}) \\ & + ((4.12E-17) \times Total\_slp\_wk) \\ & + ((-1.66E-15) \times HE\_HbAlc) \end{aligned} \tag{2}$$

In Equation (2), variables that meet the significance level and influence diabetes, fasting blood glucose, average daily sleep hours in weeks, and glycated hemoglobin is used. Table 4 shows the regression results for dyslipidemia.

TABLE 4. The regression results for dyslipidemia.

Dependent Variable	Dyslipidemia				
Independent Variable	DI1_pt	DI1_2	DI2_pt	DI2_2	···
Estimate	1.009	0.2886	-0.8275	-0.4076	···
Std. Error	0.355	0.08181	0.3924	0.1005	···
t-value	2.842	3.527	-2.108	-4.057	···
Signif.	0.005	0.000555	0.036629	0.0000791	···

Equation (3) shows the regression formula for dyslipidemia. Dys represents Dyslipidemia, and 0.3858 is a value of y-intercept.

$$\begin{aligned} \mathrm{Dys} = {} & 0.3858 + (1.009 \times \mathrm{DI1\_2}) + (-0.8275 \times \mathrm{DI2\_pt}) \\ & + (-0.4076 \times \mathrm{DI2\_2}) \\ & + (0.002468 \times \mathrm{HE\_chol}) \end{aligned} \tag{3}$$

In Equation (3), variables that meet the significance level and influence dyslipidemia, high blood pressure treatment, blood pressure control drug intake, dyslipidemia treatment, dyslipidemia drug intake, and total cholesterol is used.

Secondly, a model for mental health is extracted. To do that, multiple regression analysis (for one dependent variable and multiple independent variables) is applied. As a dependent variable, the prevalence of perceived stress is used. With independent variables that meet the significance level of 0.05, a regression model is established. Table 5 shows the regression results for mental health.

TABLE 5. The regression results for mental health.

Dependent Variable	Stress				
Independent Variable	HE_obe	DI2_pr	HE_HTG	DE1_33	···
Estimate	0.050051	0.27029	-0.15524	0.188145	···
Std. Error	0.021401	0.1056	0.073586	0.065028	···
t-value	2.339	2.56	-2.11	2.893	···
Signif.	0.02066	0.01146	0.03653	0.00438	···

Equation (4) presents the regression model for mental health. The value of y-intercept is 2.901.

$$\begin{aligned} \mathrm{mental\_health} = {} & 2.901 + (0.050051 \times HE\_obe) \\ & + (0.27029 \times DI2\_pr) \\ & + (-0.15524 \times HE\_HTG) \\ & + (0.188145 \times DE1\_3) \\ & + (-0.41919 \times BP1) \end{aligned} \tag{4}$$

In Equation (4), mental health is predicted with the following variables: obesity, whether to have dyslipidemia at present, whether to have hypertriglyceridemia, blood glucose care treatment for diabetes (non-pharmacological therapy), and bedtime in weeks.

Thirdly, a model for behavioral patterns is extracted. Whether to have physical activities is set as a dependent variable, and multiple regression analysis is conducted. Table 6 shows the regression results for behavioral patterns.

TABLE 6. The regression results for behavioral patterns.

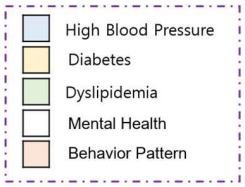
Dependent Variable	Behavioral patterns				
Independent Variable	HE_nARM	HE_sbp1	HE_obe	BD1_11	···
Estimate	-0.3247	-0.0002899	-0.02749	-0.03169	···
Std. Error	0.07203	0.001693	0.01081	0.009814	···
t-value	-4.508	-0.171	-2.543	-3.229	···
Signif.	0.0000131	0.86424	0.01199	0.00153	···

Equation (5) shows the regression model for mental health. The value of the y-intercept is 2.26.

$$\begin{aligned} \mathrm{Behavior\_pattern} = {} & 2.26 + (-0.03169 \times \mathrm{BD1\_11}) \\ & + (-0.3247 \times \mathrm{HE\_nARM}) \\ & + (-0.02749 \times \mathrm{HE\_obe}) \end{aligned} \tag{5}$$

In Equation (5) for behavioral patterns, blood pressure measurement (arm), whether to have obesity and yearly drinking frequency are used as variables. Through equations (1) to (5), a matrix is expressed to extract easily relations between users' chronic disease data, mental health data, and behavioral patterns data. With the matrix, it is possible to express data relations easily and to execute operations

conveniently [22], [23]. The values in the matrix are the values of a regression formula. Figure 3 shows the user data matrix.

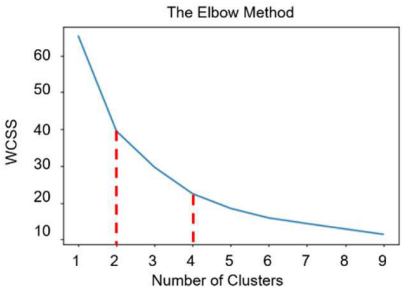


User1	2.9	3	0.0	1.9	1.6
User2	1.5	3	-0.0	2.0	1.6
User3	2.9	3	0.7	1.8	1.7
User4	1.5	3	-0.1	2.0	1.7
user5	2.9	3	0.1	1.6	1.6
...	...	...	...	...	...

FIGURE 3. The user data matrix.

Also, among users have similar relations. It is necessary to increase the accuracy of knowledge recommendations by using similarity relations between users. To find user similarities, clustering is performed. As for clustering, a K-means algorithm with low time complexity is employed. However, a cluster's core point is randomly selected so that a test result becomes different. In short, it is impossible to draw a consistent result through clustering [24]. To solve the problem, the K-means++ algorithm is used. Although it is similar to the K-means algorithm, its step of initializing a core point is different. The procedure of the K-means++ algorithm is as follows:

First, a core point is randomly specified. Next, the distance from a core point closest to each one of the remaining data is calculated. The next core point is specified according to the probability in proportion to the distance from the closest core point. In this way, it is possible to prevent a core point from approaching closely the core point already specified. Therefore, the general K-means++ algorithm is more strategic than the K-means algorithm when initializing the central point, and optimized clustering is possible [25]. To determine the most appropriate number of groups for user clustering, the Elbow method is applied. According to the result of Within Cluster Sum of Squares (WCSS), it presents a section in which the sum of distances between clusters is sharply reduced. Such a point is used as the number of clusters. Nevertheless, in the Elbow method, there is still an unclear part in determining the number of clusters. For this reason, Silhouette is applied to evaluate the validity of clusters and determine the number of clusters to use. Silhouette is a measure of how similar an object is to its own cluster compared to other clusters. It ranges from –1 to +1, where the higher a value is, the more appropriate clustering occurs; the lower the value is, the less appropriate clustering occurs [26]. Figure 4 shows the results of Elbow, where the vertical axis represents WCSS, and the horizontal axis represents the number of clusters. As shown in the results of Elbow evaluation in Fig. 4, the distance between clusters reduces when the number of clusters is 2 or 4. However, in order to determine the number of most suitable clusters, it is determined through Silhouette results.



Number of K	WCSS
1	65.20474
2	39.42379
3	29.627986
4	22.441515
5	18.452466
6	15.870806
7	14.351538
8	12.886551
9	11.442350

FIGURE 4. The results of the Elbow method.

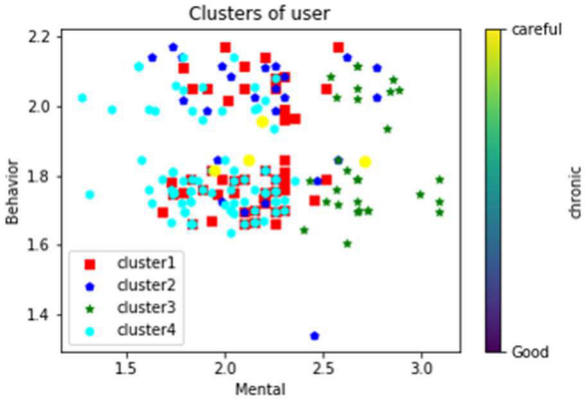


FIGURE 5. The results of user clustering based on the K-means++ algorithm.

Table 7 shows the results of Silhouette according to the number of clusters. In short, the result of Silhouette is different depending on K, the number of clusters.

TABLE 7. Silhouette results by number of clustering.

Number of K	Silhouette Score
2	0.34825
3	0.34392
4	0.54783
5	0.32643
6	0.33548
7	0.29543
8	0.29737
9	0.31573

As shown in Table 7, a result of Silhouette scored the highest when the number of clusters is 4. Therefore, four clusters are generated. Figure 6 shows the results of user clustering based on the K-means++ algorithm. Persons who suffer from similar chronic diseases have similar mental health and behavioral patterns. For this reason, chronic disease patients are clustered on the basis of mental health and behavioral patterns.

In the Fig. 6, the horizontal axis represents mental health, and the vertical axis means behavioral patterns.

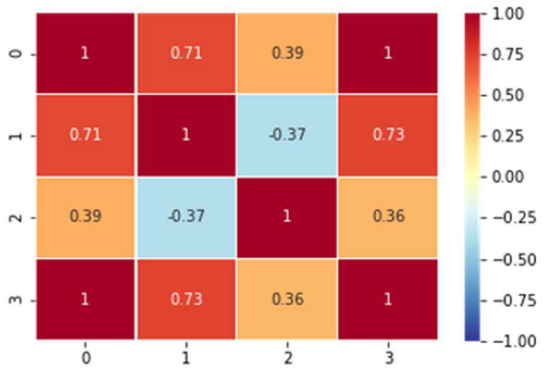


FIGURE 6. Coefficients of correlations between clusters for an adjacency matrix.

The clustering results in Fig. 6 reveal that although clusters 2 and 3 have similar mental health, they have different behavioral patterns and chronic disease risk; that although clusters 1 and 3 have similar behavioral patterns and chronic disease risk, they have different mental health.

C. GRAPH NEURAL NETWORK BASED ON CONTEXT-MINING

Persons with similar disorders have latent associations and similarities. A graph is employed to find them. A graph is made up of a set of node (vertex) V and edge E. It helps to represent relations or interactions between objects, to express a complex issue in a simple way, and to make expressions from multiple perspectives to solve problems [27], [28]. Also, the general graph analysis method that needs empirical and preliminary knowledge has difficulty extracting information from multiple graphs. Accordingly, deep learning or machine learning-based GNN is applied to a graph. It takes into account similarities with a particular user's distant neighbors as well as near neighbors and maintains a graph's structure. Graph Convolutional Network (GCN) is a sort of GNN. It extracts abstract features for input data with no use of neighboring nodes' information. Accordingly, by multiplying an adjacency matrix (A), it combines neighboring nodes' information. An adjacency matrix specifies a graph shape at the beginning. In the matrix, the presence of a connection between nodes represents '1', and no presence represents '0'. Accordingly, an adjacency matrix has no edge from a vertex to itself, so the diagonal elements of the matrix are all zeros. Since a weighted kernel, which uses no its own information, is generated, an identity matrix (I) is added to solve the problem [29]. As for the weight used in a graph, the user data attributes extracted from regression analysis are used. Through relations between users, such as similar chronic diseases, disorders, and behaviors, the grounds for recommendation are found. Figure 6 shows coefficients of correlations between clusters for an adjacency matrix.

An adjacency matrix represents two-dimensional arrays of graph connections. The generation of an adjacency matrix is defined through connections of the clusters that are found to be related in cluster correlation analysis. Based on a core point of user similarity cluster, a primary node is used. Based on a weight representing a correlation between clusters, edges are connected. As shown in Fig. 6, since the diagonal elements represent clusters, coefficients of correlations are 1. In an adjacency matrix, a node's own information is excluded. For this reason, connections are made when coefficients of correlations are a positive number except for 1. An initial graph is designed. Figure 7 shows the initial graph and adjacency matrix.

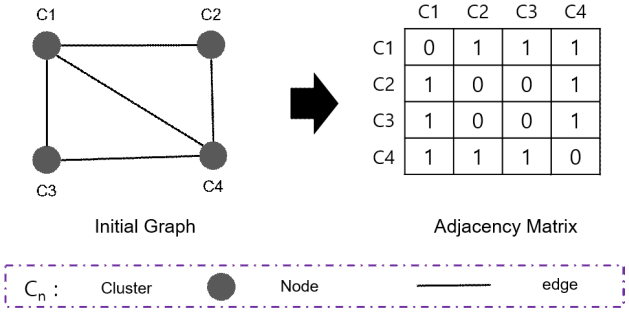


FIGURE 7. Coefficients of correlations between clusters for an adjacency matrix.

In the case of a feature matrix for augmenting a graph with the use of node information sharing, a user's data matrix in Fig. 4 is used. Algorithm 1 shows a graph augmentation algorithm based on context mining. Its input value is an initial graph. Its output value is an augmented graph.

The first step in algorithm 1 is to design an initial graph. Nodes are established according to the number of clusters. Edges are connected according to a positive coefficient of correlation between clusters and an adjacency matrix is generated. The second step is to augment a graph through multi-context mining. A feature matrix for graph augmentation is a user feature matrix based on the regression model that represents users' chronic disease information, mental health, and behavioral patterns. In graph convolution operations, features are extracted, and accordingly, a graph is augmented. However, depending on the degree of connection between nodes, reliability is different. Therefore, weights are newly applied to edges. In this way, it is possible to obtain not only the primary neighboring information but k^{th} neighboring information. Equation (6) shows an edge weight based on context. As a result, with the reliable weight, node and edge information is updated.

$$\text{Weight} = update(\text{A} \times U^{(k)} \times F^{(k)}) \tag{6}$$

In Equation (6), 'A' means an adjacency matrix; U(k) represents k^{th} user information; F(k) means kth user's feature matrix. Figure 8 shows the structure of a graph-based on multi-context mining. The input graph in Fig. 8 is user data. Context mining is performed through preprocessing of raw data. Accordingly, the adjacency matrix and feature matrix

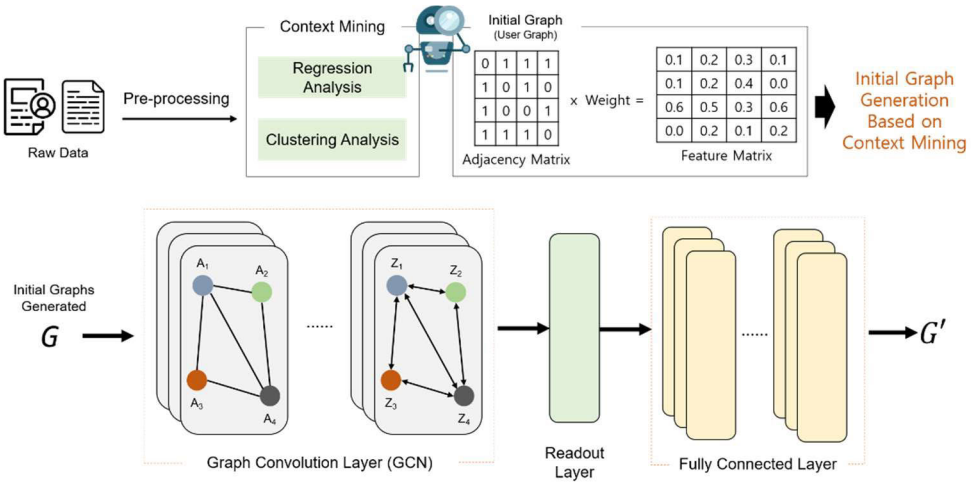


FIGURE 8. The structure of a graph-based on multi-context mining.

Algorithm 1 Graph Augmentation Algorithm Based on Context Mining

Input: IG // initial Graph
Output: AG //Argmented Graph
int Adj_matrix = [[0,0,1,0], [0,0,1,1], [1,1,0,1], [0,1,1,0]]
int V, E, w
Graph (V, E, w) ← 0
Step 1: Prime Graph Type //node is number of cluster
  for each v ∈ V and each e ∈ E
    for v **to** Number of cluster **do**
      if v <= 4
        node++
      else
        stop creating vertex
      endif
    end
    for e **to** using Adjacency Matrix **do**
      if e = 1
    add edge and connection vertex
      else if e = 0
     not add edge and not connection vertex
    end
end
Step 2: Graph Argument using graph convolution
  User Matrix = using result of regression model
  for F **to** User Matrix and Adjacency Matrix **do**
    F[i][j] = Convolution (User Matrix$_i$ × Adjacency Matrix$_j$)
    G = F[i][j] //Argument Graph
  end
  return

of an initial graph is generated. In graph networks, weights are updated with graph convolution operation and the weight operation of Equation (6). A graph is augmented in the course of sharing a particular node's information with a different node. In this way, it is possible to classify a node, find relations between nodes, and predict a degree of association. Neighboring nodes have similar attributes. Therefore, based on the node and edge information in layer 1, it is possible to obtain a different node's information gradually. Therefore, GNN based emerging health risk prediction is a method of predicting latent risks on the basis of neighboring information. Figure 9 shows the prediction process using Multi-Context-GNN.

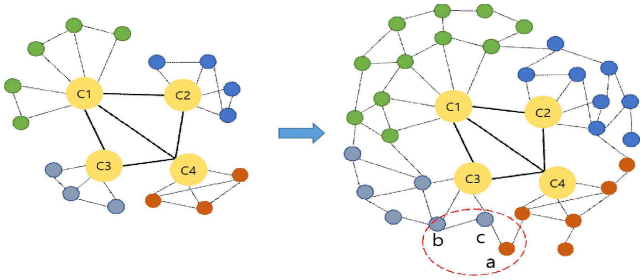


FIGURE 9. The prediction process using multi-context-GNN.

For example, in Fig. 9, users *a* and *b* are not directly connected with each other, but it is found that they are associated with each other on the basis of information of user *c*. In other words, with graph augmentation, user relationship continues to be generated. Accordingly, even if users are not directly associated with each other, it is possible to predict a latent risk with the use of a neighbor's information.

IV. PERFORMANCE EVALUATION

Raw data from the National Health and Nutrition Survey [18] were used in the experiments. The purpose of the national health survey is to calculate statistics with national representation and reliability regarding the health level, health behavior, and food and nutrition intake of the people. Accordingly, it provides basic data for health policies, such as goal setting and evaluation of the comprehensive national health promotion plans, and health promotion program developments. For the sampling frame of the National Health and Nutrition Examination Survey, the most recent data from the Population and Housing Census available at the time of sample design was used as the basic extraction frame. It supplements the

basic extraction frame by adding and improving the population inclusion rate. In this study, chronic disease, behavioral patterns, and mental health-related data were selected from the raw data present in the National Health and Nutrition Examination Survey, and preprocessing was performed on the data related to 69 variables and 24,269 people. Through preprocessing, missing values are processed, and unnecessary variables and variables with duplicate meanings are removed. As a result, preprocessed data are divided into training data (70%), test data (20%), and validation data (10%). In terms of performance, the validity of the regression analysis used for a user feature matrix is assessed, a clustering method is evaluated, and the proposed model is compared with a conventional model.

Firstly, the validity of regression analysis is assessed. In this study, to establish a feature matrix of a graph, a regression model is generated as the result of multivariate analysis. To find the validity of the multivariate analysis-based regression model, univariate analysis is compared with multivariate analysis. As for performance indexes, the coefficient of determination denoted R2, Adjusted R-Squared, and MSE are used. R2 is used to evaluate prediction performance based on distribution [30]. The larger the coefficient of determination is, the more the actual value is similar to the predicted value, and the better the data explanation. Equation (7) states the expression for R^2.

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST} \tag{7}$$

In Equation (7), the Total Sum of Squares (SST) is the sum of the differences between the mean of observed values and an observed value. Explained Sum of Squares (SSE) is the sum of the differences between the mean of observed values and a predicted value. The Residual Sum of Squares (SSR) is the sum of the residuals between a predicted value and an observed value. The closer the coefficient of determination (R^2) is to 1, the better performance occurs. This influences the number of independent variables. In short, it increases with the number of independent variables. For this reason, it is difficult to evaluate performance accurately. To solve the problem, an Adjusted R-Squared is used. Too lower than Adjusted R-Squared means that unnecessary independent variables are included [31]. Equation (8) states the expression for the adjusted R-squared value.

$$Adj\_R = 1 - \frac{(n-1)(1-R^2)}{n-p-1} \tag{8}$$

In Equation (8), n represents the number of data, and p means the number of independent variables. The closer the Adjusted R-Squared is to 1, the more prediction is accurate. A negative value of Adjusted R-Squared means that a regression model is useless. For the evaluation of a regression analysis model, MSE is employed. It represents the mean of the square of the difference between actual and predicted values [32]. Equation (9) shows the expression of MSE.

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{9}$$

As data used in the regression analysis for designing a feature map of the proposed model, high blood pressure, diabetes, and dyslipidemia data among raw data of the National Health and Nutrition Examination Survey are used. It is necessary to prove that performance is better when a feature map is generated via multivariate regression analysis considering all three chronic diseases than that obtained when a feature map is generated via univariate regression analysis considering each disease individually. Table 8 shows the results of the regression analysis. It shows the results from the comparison between univariate regression analysis on each one of diabetes, high blood pressure, and dyslipidemia) and multivariate regression analysis on all three chronic diseases.

TABLE 8. The results of regression analysis evaluation.

Method	Division	R^2	R^2 Adj.	MSE
Univariate	High Blood Pressure	0.795	0.677	0.06
	Diabetes	1.000	0.552	0.05
	Dyslipidemia	0.848	0.776	0.07
Multivariate (Ours)	Chronic diseases	0.899	0.839	0.01

In Table 8, the coefficient of determination for diabetes in the univariate analysis is the highest but has the largest difference from R^2 adj. It means that explanatory variables for diabetes include unnecessary variables. Therefore, in the proposed method, multivariate analysis is the most effective.

Secondly, the clustering method for creating an initial graph is evaluated. In the proposed method, an initial graph is generated through K-means++ algorithm-based clustering. When a cluster's core point is randomly selected, it is difficult to obtain a consistent result. Therefore, the K-means++ algorithm prevents the problem [33], [34]. In the case of the general K-means algorithm, a cluster's core point is changed such that the initial graph loses consistency. To prove that the K-means++ algorithm used in the proposed model is more appropriate than the K-means algorithm, it is necessary to compare K-means and K-means++ algorithms in terms of Precision, Recall, and F-measure. Figure 10 shows the F-measure values according to the number of clusters of K-means and K-means++.

Table 9 shows the precision, recall, and f-measure of K-means and K-means++ when the number of clusters with excellent F-measure in Figure 12 is 4.

TABLE 9. Comparison results of precision, recall, and f-measure of clusters.

	Precision	Recall	F-measure
K-means	0.712	0.483	0.575
K-means++	0.753	0.771	0.761

As presented in Figure 10 and Table 9, when the K-means++ algorithm is used, the performance is excellent. And in Table 9, the K-means algorithm randomly selects a

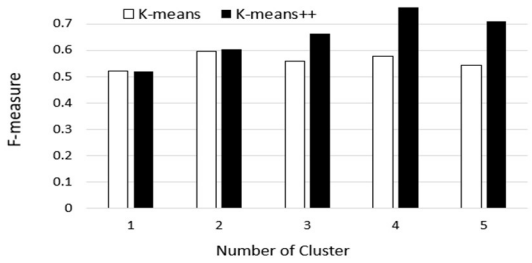


FIGURE 10. The F-measure values according to the number of clusters of K-means and K-means++.

cluster's core point, and consequently, recall is low. Given that, K-means++ is appropriate for generating an initial graph of the proposed model.

Lastly, to evaluate the excellence of the proposed method, it is compared with conventional GCN-based recommender models. For performance evaluation, the MSE and recall were employed. Wu et al. [35] proposed a graph convolutional network-based model for social recommendation, which was generated based on the expansion of social networks and user-item preferences. However, it struggles with predicting a latent factor that is not included in the user features. Yang et al. [36] proposed a graph network for solving social inconsistency problems. In their proposed method, a sampling probability is associated with the score of the neighbors' consistency, and thus, consistent neighbors are sampled. This method was limited owing to considering only the information of consistent neighbors. R. Wang et al. [37] applied a deep graph network for the analysis and prediction of patient health comorbidities from sparse health records. Their approach represents patient data including health examination categories, hospitalization, and injury accidents in a graph structure, and models patient health trends, disease prognosis, and potential correlations by recovering missing connections through connection prediction problems [38]. Table 10 shows the performance comparison results between the conventional models and the proposed method.

TABLE 10. The results of comparison with conventional models.

	MSE	Recall
L. Wu et al. [35]	0.132	0.587
L. Yang et al. [36]	0.252	0.622
R. Wang et al. [37]	0.355	0.598
Proposed Method	0.01	0.771

As shown in Table 10, the proposed method has excellent performance. The technique proposed by Wu et al. [35] considers only the neighbors' information, hence, its performance is low on the augmented graph. The method proposed by Yang et al. [36] considers the information of the most related user. Therefore, if neighbors have no consistent information or a new user, it is hard to make a graph-based analysis, and performance is evaluated as low. In addition, the method proposed by Wang et al. [37] does not include the semantic or causal component of the disease and uses sparse data. Therefore, the obtained accuracy of health risk and prediction is low. On the other hand, in the proposed method, a graph is augmented using neighbor information so that it is possible to obtain information according to a 2-hop relationship as well as a 1-hop relationship. Therefore, compared to conventional methods, the proposed method attains excellent performance.

V. CONCLUSION

A graph has the advantages of being convenient to expand and being able to intuitively present the relationship between nodes. This study proposed a multi-context mining-based graph neural network for predicting emerging health risks. The proposed method predicts and recommends potential emerging risks through a graph neural network based on information regarding similar symptoms, causes, and management methods for patients with chronic diseases. It consisted of three steps. The first step was to collect and preprocess health information, mental health information, and behavioral patterns information of chronic disease patients who suffer from high blood pressure, diabetes, and dyslipidemia. The second step, context mining was performed using preprocessed data to generate a feature map for the graph extension. In multivariate regression analysis, a regression model that has high blood pressure, diabetes, and dyslipidemia as dependent variables were extracted. In addition, in linear regression analysis, a regression model for mental health and behavioral patterns was generated, and a feature map was created. Through clustering, the nodes of the initial graph were created. According to the correlation coefficients, the edges of the initial graph were designed. The last step was to augment the graph with the feature map generated through context mining of the initial graph and to update weights for predicting latent risks not only in relations between distant neighbors but in relations between close neighbors. In this way, it was possible to find similar symptoms and causes of all users and to predict emerging risks. Performance evaluation was conducted in three ways. First, the validity of the regression analysis for patients with chronic disease was evaluated. As a result, performance was better in multivariate analysis than in univariate analysis, in which each chronic disease was set as a dependent variable. Second, the clustering method for determining the number of nodes in an initial graph was evaluated. As a result, the K-means++ algorithm achieved better performance because it overcame the problem of the K-means algorithm, where a cluster's core point varied. Finally, to evaluate the excellence of the proposed model, it was compared with conventional models in terms of MSE and recall. As a result, the proposed method solved the problems of conventional methods so that its MSE and Recall were about 0.1-0.2 higher than those of conventional ones. Therefore, it is possible to effectively predict the potential risk through the proposed model; accordingly, information that can prevent health risks can be provided.

REFERENCES

[1] H. Yoo and K. Chung, "Deep learning-based evolutionary recommendation model for heterogeneous big data integration," *KSII Trans. Internet Inf. Syst.*, vol. 14, no. 9, pp. 3730–3744, Sep. 2020.

[2] H. Yoo, R. C. Park, and K. Chung, "IoT-based health big-data process technologies: A survey," *KSII Trans. Internet Inf. Syst.*, vol. 15, no. 3, pp. 974–992, 2021.

[3] J.-W. Baek and K. Chung, "Context deep neural network model for predicting depression risk using multiple regression," *IEEE Access*, vol. 8, pp. 18171–18181, 2020.

[4] H. Jung and K. Chung, "Social mining-based clustering process for big-data integration," *J. Ambient Intell. Humanized Comput.*, vol. 12, no. 1, pp. 589–600, Jan. 2021.

[5] Z. Guo and H. Wang, "A deep graph neural network-based mechanism for social recommendations," *IEEE Trans. Ind. Informat.*, vol. 17, no. 4, pp. 2776–2783, Apr. 2021.

[6] M. Zhang, P. Li, Y. Xia, K. Wang, and L. Jin, "Labeling trick: A theory of using graph neural networks for multi-node representation learning," 2020, *arXiv:2010.16103*.

[7] G. Dong, L. Cai, D. Datta, S. Kumar, L. E. Barnes, and M. Boukhechba, "Influenza-like symptom recognition using mobile sensing and graph neural networks," in *Proc. Conf. Health, Inference, Learn.*, Apr. 2021, pp. 291–300.

[8] K. Guo, Y. Hu, Z. Qian, H. Liu, K. Zhang, Y. Sun, J. Gao, and B. Yin, "Optimized graph convolution recurrent neural network for traffic prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 2, pp. 1138–1149, Feb. 2021.

[9] Y. Tao, C. Wang, L. Yao, W. Li, and Y. Yu, "Item trend learning for sequential recommendation system using gated graph neural network," *Neural Comput. Appl.*, pp. 1–16, Feb. 2021.

[10] H. Jung and K. Chung, "Knowledge-based dietary nutrition recommendation for obese management," *Inf. Technol. Manage.*, vol. 17, no. 1, pp. 29–42, Mar. 2016.

[11] S. Kobayashi, "Contextual augmentation: Data augmentation by words with paradigmatic relations," 2018, *arXiv:1805.06201*.

[12] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip, "A comprehensive survey on graph neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 4–24, Mar. 2020.

[13] E. Chien, J. Peng, P. Li, and O. Milenkovic, "Adaptive universal generalized PageRank graph neural network," 2020, *arXiv:2006.07988*.

[14] Y. Jing, J. Wang, W. Wang, L. Wang, and T. Tan, "Relational graph neural network for situation recognition," *Pattern Recognit.*, vol. 108, Dec. 2020, Art. no. 107544.

[15] B. Abu-Salih, M. Al-Tawil, I. Aljarah, H. Faris, P. Wongthongtham, K. Y. Chan, and A. Beheshti, "Relational learning analysis of social politics using knowledge graph embedding," *Data Min. Knowl. Discov.*, vol. 35, no. 4, pp. 1497–1536, 2021.

[16] X. Sha, Z. Sun, and J. Zhang, "Hierarchical attentive knowledge graph embedding for personalized recommendation," 2019, *arXiv:1910.08288*.

[17] P. Liu, L. Zhang, and J. A. Gulla, "Real-time social recommendation based on graph embedding and temporal context," *Int. J. Hum.-Comput. Stud.*, vol. 121, pp. 58–72, Jan. 2019.

[18] *The Seventh Korea National Health and Nutrition Examination Survey*. Accessed: Jun. 27, 2020. [Online]. Available: https://knhanes.kdca.go.kr/

[19] C. Wang, B. Zhao, L. Luo, and X. Song, "Regression analysis of current status data with latent variables," *Lifetime Data Anal.*, vol. 27, no. 3, pp. 413–436, Apr. 2021.

[20] C. Maheswari, E. B. Priyanka, S. Thangavel, S. V. R. Vignesh, and C. Poongodi, "Multiple regression analysis for the prediction of extraction efficiency in mining industry with industrial IoT," *Prod. Eng.*, vol. 14, no. 4, pp. 457–471, Jun. 2020.

[21] S. Ghosal, B. Sinha, M. Majumder, and A. Misra, "Estimation of effects of nationwide lockdown for containing coronavirus infection on worsening of glycosylated haemoglobin and increase in diabetes-related complications: A simulation model using multivariate regression analysis," *Diabetes Metabolic Syndrome, Clin. Res. Rev.*, vol. 14, no. 4, pp. 319–323, Jul. 2020.

[22] Z. Li, Z. Hu, F. Nie, R. Wang, and X. Li, "Matrix completion with column outliers and sparse noise," *Inf. Sci.*, vol. 573, pp. 125–140, Sep. 2021.

[23] A. Alvarez-Melcon, X. Wu, J. Zang, X. Liu, and J. S. Gomez-Diaz, "Coupling matrix representation of nonreciprocal filters based on time-modulated resonators," *IEEE Trans. Microw. Theory Techn.*, vol. 67, no. 12, pp. 4751–4763, Dec. 2019.

[24] K. P. Sinaga and M.-S. Yang, "Unsupervised K-means clustering algorithm," *IEEE Access*, vol. 8, pp. 80716–80727, 2020.

[25] J. Hämäläinen, T. Kärkkäinen, and T. Rossi, "Improving scalable K-means++," *Algorithms*, vol. 14, no. 1, p. 6, Dec. 2020.

[26] G. Ogbuabor and U. F. N, "Clustering algorithm for a healthcare dataset using silhouette score value," *Int. J. Comput. Sci. Inf. Technol.*, vol. 10, no. 2, pp. 27–37, Apr. 2018.

[27] K. Zhan, C. Niu, C. Chen, F. Nie, C. Zhang, and Y. Yang, "Graph structure fusion for multiview clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 10, pp. 1984–1993, Oct. 2019.

[28] Z. Zhang, J. Bu, M. Ester, J. Zhang, C. Yao, Z. Yu, and C. Wang, "Hierarchical graph pooling with structure learning," 2019, *arXiv:1911.05954*.

[29] L. Yao, C. Mao, and Y. Luo, "Graph convolutional networks for text classification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, Jul. 2019, pp. 7370–7377.

[30] J. D. Rights and S. K. Sterba, "New recommendations on the use of R-Squared differences in multilevel model comparisons," *Multivariate Behav. Res.*, vol. 55, no. 4, pp. 568–599, Jul. 2020.

[31] T. Hayes, "R-squared change in structural equation models with latent variables and missing data," *Behav. Res. Methods*, vol. 53, no. 5, pp. 2127–2157, Mar. 2021.

[32] A. de Myttenaere, B. Golden, B. Le Grand, and F. Rossi, "Mean absolute percentage error for regression models," *Neurocomputing*, vol. 192, pp. 38–48, Jun. 2016.

[33] K. Makarychev, A. Reddy, and L. Shan, "Improved Guarantees for k-means++ and k-means++ Parallel," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 16142–16152.

[34] J. Hämäläinen, T. Kärkkäinen, and T. Rossi, "Improving scalable K-means++," *Algorithms*, vol. 14, no. 1, pp. 6–25, Jun. 2020.

[35] L. Wu, P. Sun, R. Hong, Y. Fu, X. Wang, and M. Wang, "SocialGCN: An efficient graph convolutional network based model for social recommendation," 2018, *arXiv:1811.02815*.

[36] L. Yang, Z. Liu, Y. Dou, J. Ma, and P. S. Yu, "ConsisRec: Enhancing GNN for social recommendation via consistent neighbor aggregation," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2021, pp. 2141–2145.

[37] R. Wang, M. C. Chang, and M. Radigan, "Modeling latent comorbidity for health risk prediction using graph convolutional network," in *Proc. 33rd Int. Flairs Conf.*, 2020, pp. 341–346.

[38] D.-H. Shin, R. C. Park, and K. Chung, "Decision boundary-based anomaly detection model using improved AnoGAN from ECG data," *IEEE Access*, vol. 8, pp. 108664–108674, 2020.

JI-WON BAEK received the B.S. degree from the School of Computer Information Engineering, Sangji University, South Korea, in 2017, and the master's degree from the School of Department of Computer Science, Kyonggi University, South Korea, in 2020, where she is currently pursuing the doctorate degree with the Department of Computer Science. She was at the Data Management Department, Infiniq Company Ltd. She is a Researcher with the Data Mining Laboratory, Kyonggi University. Her research interests include data mining, data management, knowledge systems, automotive testing, deep learning, medical data mining, healthcare, and recommendation.

KYUNGYONG CHUNG received the B.S., M.S., and Ph.D. degrees from the Department of Computer Information Engineering, Inha University, South Korea, in 2000, 2002, and 2005, respectively. He was at the Software Technology Leading Department, South 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 has been a Professor with the Division of AI Computer Science and Engineering, Kyonggi University, South Korea. He was named a Highly Cited Researcher by Clarivate Analytics in 2017. His research interests include data mining, artificial intelligence, healthcare, knowledge systems, HCI, and recommendation systems.