IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 24, 2021, accepted September 15, 2021, date of publication October 1, 2021, date of current version October 8, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3116974

An Efficient Prediction Method for Coronary Heart Disease Risk Based on Two Deep Neural Networks Trained on Well-Ordered Training Datasets

TSATSRAL AMARBAYASGALAN[1], VAN-HUY PHAM[2], NIPON THEERA-UMPON[3,4], (Senior Member, IEEE), YONGJUN PIAO[5,6], AND KEUN HO RYU[2,3], (Life Member, IEEE)

[1]Database and Bioinformatics Laboratory, School of Electrical and Computer Engineering, Chungbuk National University, Cheongju 28644, South Korea
[2]Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City 700000, Vietnam
[3]Biomedical Engineering Institute, Chiang Mai University, Chiang Mai 50200, Thailand
[4]Department of Electrical Engineering, Faculty of Engineering, Chiang Mai University, Chiang Mai 50200, Thailand
[5]School of Medicine, Nankai University, Tianjin 300071, China
[6]Tianjin Key Laboratory of Human Development and Reproductive Regulation, Tianjin Central Hospital of Gynecology Obstetrics, Tianjin 300199, China

Corresponding authors: Keun Ho Ryu (khryu@tdtu.edu.vn) and Yongjun Piao (ypiao@nankai.edu.cn)

This work was supported in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) by the Ministry of Science, ICT, and Future Planning under Grant 2019K2A9A2A06020672 and Grant 2020R1A2B5B02001717, in part by the National Natural Science Foundation of China under Grant 61802209, and in part by the Open Fund of Tianjin Central Hospital of Gynecology Obstetrics/Tianjin Key Laboratory of Human Development and Reproductive Regulation under Grant 2020XHY03.

ABSTRACT This study proposes an efficient prediction method for coronary heart disease risk based on two deep neural networks trained on well-ordered training datasets. Most real datasets include an irregular subset with higher variance than most data, and predictive models do not learn well from these datasets. While most existing prediction models learned from the whole or randomly sampled training datasets, our suggested method draws up training datasets by separating regular and highly biased subsets to build accurate prediction models. We use a two-step approach to prepare the training dataset: (1) divide the initial training dataset into two groups, commonly distributed and highly biased using Principal Component Analysis, (2) enrich the highly biased group by Variational Autoencoders. Then, two deep neural network classifiers learn from the isolated training groups separately. The well-organized training groups enable a chance to build more accurate prediction models. When predicting the risk of coronary heart disease from the given input, only one appropriate model is selected based on the reconstruction error on the Principal Component Analysis model. Dataset used in this study was collected from the Korean National Health and Nutritional Examination Survey. We have conducted two types of experiments on the dataset. The first one proved how Principal Component Analysis and Variational Autoencoder models of the proposed method improves the performance of a single deep neural network. The second experiment compared the proposed method with existing machine learning algorithms, including Naïve Bayes, Random Forest, K-Nearest Neighbor, Decision Tree, Support Vector Machine, and Adaptive Boosting. The experimental results show that the proposed method outperformed conventional machine learning algorithms by giving the accuracy of 0.892, specificity of 0.840, precision of 0.911, recall of 0.920, f-measure of 0.915, and AUC of 0.882.

INDEX TERMS Coronary heart disease, deep neural network, machine learning, principal component analysis, reconstruction error, variational autoencoder.

I. INTRODUCTION

Coronary heart disease (CHD) is a type of Cardiovascular Disease (CVD), and 85% of CVD deaths are due to CHD. According to the report by the World Health Organization, CHD is the top cause of death globally with regard to 2017; an estimated 15.2 million people died from CHD as of 2016 [1]. It is also highly ranked in South Korea, being ranked second of all deaths [2]. If suffering from CHD, a waxy substance called plaque will be built up inside the coronary arteries that

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

deliver oxygen and nutrients to the heart muscle. This plaque narrows arteries, and the flow of oxygen-rich blood to the heart muscle is limited [3]. Over time, heart arteries are more narrowed and block the blood flow. Then, a heart attack or sudden death can occur because of the blockage. It usually progresses over many years without any symptoms.

Therefore, most people are diagnosed in the middle or late stage after feeling some symptoms, such as chest pain, shortness of breath, or fatigue. If CHD reaches serious condition, the most advanced treatments are necessary, such as stent surgery for keeping coronary arteries open and reducing the occurring of a heart attack, and coronary artery bypass grafting for supporting blood flow to the heart muscle, and heart transplant [4]. In the early stage, a healthy diet, active exercise, and appropriate medicines and care can help prevent suffering from CHD.

Recently, many studies have been conducted to predict the risk of CHD using machine learning and deep learning approaches. The machine learning-based methods mainly proposed single or ensemble classification algorithms [6], and some of them used feature selection or feature extraction techniques to improve the performance [7], [13], [14]. Nowadays, deep learning techniques have been successfully used to diagnose CHD [15]–[19]. Most existing methods first split an experimental dataset into two parts for training and testing. Then, they build predictive models from the whole or randomly sampled training dataset using classification algorithms. As a result, the models are more fitted on the regularly distributed dataset and misclassify irregularly distributed (biased) data.

Therefore, we focused on this problem by using distinct predictive models for the regular and biased inputs. In our previous study [20], the proposed method consisted of four deep learning models, including two Stacked Autoencoder (SAE) models and two deep neural network (DNN) models. First, we divided a training dataset into two groups based on their reconstruction errors given from the first SAE model. Next, two DNN models were trained on these divided groups by combining a reconstruction error-based new feature with other risk factors to predict the risk of developing CHD. The main idea was extracting the reconstruction error-based new feature from the second SAE model for two DNN models. In this study, the presented method does not perform feature extraction for DNN models. Instead, it is focused on the data distribution to improve the performance. It successfully improved the performance of the previous study.

We propose a prediction method for CHD risk based on a combination of DNN, Variational Autoencoder (VAE), and Principal Component Analysis (PCA). We addressed the following problems related to improving the prediction performance: (1) Previous studies used the whole or randomly sampled training datasets for model training. However, some data can be significantly different from the same labeled dataset. It degrades the performance of predictive models if the training dataset includes this highly biased subset. Therefore, the proposed method divides the training dataset into two groups, regular and highly biased using reconstruction error (RE) of the PCA. (2) The grouped highly biased subset from the training dataset may not be sufficient for model building due to accounts for a small percentage of the total dataset. The proposed method decides this problem by enriching the highly biased subset via two deep VAE models.

In this study, we improve the prediction performance by preparing the training dataset efficiently by solving the previously mentioned problems. The main contributions of this study are as follows:

- We propose a novel method for predictive analysis and applied it to the Korea National Health and Nutrition Examination Survey (KNHANES) dataset to predict CHD risk. The proposed method consists of one PCA and four deep learning models, including two VAE and two DNN models. The combination of the used models together is more effective. In other words, the performance of a single DNN model was improved by using these models.
- The proposed method was evaluated through two kinds of experiments. First, each model was experimented with independently and proved how they improve the performance. Second, the proposed method contrasted with several machine learning algorithms.

The rest of the paper is organized as follows. Section II provides an overview of existing methods for CHD risk prediction. The proposed method is detailed in Section III. Section IV presents evaluation metrics, experimental dataset, and parameter tuning of the compared algorithms. Section V provides a performance evaluation of the compared algorithms on the KNHANES dataset. Finally, Section VI concludes the paper.

II. LITERATURE REVIEW OF CHD RISK PREDICTION METHODS

The early detection of CHD increases the chance of successful treatment. Many researchers have focused on finding and inventing efficient algorithms to perform the CHD prediction task. In this section, an overview of CHD prediction methods is provided. First, we talk about machine learning-based methods used for CHD. Then, an overview of deep learning-based CHD prediction methods is given. At last, CHD risk prediction methods experimented on the KNHANES dataset are discussed.

Machine learning-based approaches have been used commonly for predicting CHD. Soni *et al.* compared several algorithms, such as Decision Tree (DT), Naïve Bayes (NB), K-Nearest Neighbors (KNN), and Neural Network (NN) on the Cleveland Heart Disease dataset using a free data mining software named Tanagra. As a result, DT showed the highest accuracy of 89%, followed by NB [6]. The authors of [7] compared classification methods, namely NN, Support Vector Machine (SVM), Classification based on Multiple Association Rule (CMAR), DT, and NB to predict CHD on two kinds of datasets consisted of ultrasound images of Carotid Arteries (CAs) and Heart Rate Variability (HRV)

of the electrocardiogram signal. First, they extracted feature vectors from the CAs dataset, HRV dataset, and a combination of CA and HRV datasets. As a result, the extracted vector from the CA+HRV dataset showed higher accuracy than the separated feature vectors of CAs and HRV. As a result, SVM and CMAR classifiers outperformed other compared classifiers by the accuracy of 89.51% and 89.46%, respectively. Gonsalves *et al.* studied NB, SVM, and DT algorithms on the South African Heart Disease dataset with 462 instances. Based on 10-fold cross-validation, the NB algorithm gave a promising result for detecting CHD with a sensitivity of 63% and specificity of 76% [8]. Beunza *et al.* compared DT, RF, SVM, NN, and Linear Regression (LR) on the Framingham Heart Study dataset for predicting CHD risk. According to the Area under the ROC Curve (AUC), the SVM algorithm showed the highest performance with 75% [9]. Joloudari *et al.* implemented several data classification models, including Chi-squared Automatic Interaction Detection, SVM, C5.0, and Random Tree (RT) for CHD prediction using the Z-Alizadeh Sani dataset with 303 records from the UCI machine learning repository [10]. As a result, the RT model showed the best accuracy of 91.47% and an AUC of 96.70%. These studies generally proposed and compared conventional machine learning algorithms on publicly available heart disease datasets.

PCA has been widely used in the dimension reduction of high-dimensional data. Recently, several studies have used PCA as a feature extractor for improving classification performance [13], [14]. The authors of [13] improved the performance of SVM, NB, DT algorithms by reducing data dimension from 10 to 6 using PCA on Cleveland heard disease dataset. In [14], the combination of Chi-square and PCA showed promising results to detect CHD. First, they obtained important features using the Chi-square test and reduced their dimension using PCA. Another application of PCA is to use it for detecting anomalies. Hoffmann [21] modeled the distribution of the training dataset by kernel-PCA for detecting an anomaly. The proposed approach computed the RE in feature space and used it as a novelty measure. In [22], the authors detected anomalies by computing errors when reconstructing the original image on PCA projections for hyperspectral imagery.

VAE is one kind of neural network that is not only used as a generative model but also used as a classifier. In [23] paper, VAE was proposed for generating synthetic electronic health records (EHR). They confirmed the performance of the LSTM model trained on the synthetic data is similar to those trained on real EHRs containing over 250,000 records. The authors of [24] paper used generated data from VAE for missing data imputation to identify the abnormal carotid arteries. They also removed a few labels from the test dataset and generated the incomplete labels by the VAE. As a result, VAE based classifier outperformed other supervised classifiers, including SVM, LR, and RF algorithms.

Tama *et al.* proposed a two-tier ensemble model for CHD prediction and evaluated it on the Z-Alizadeh Sani, Statlog, Cleveland, and Hungarian datasets [11]. The first-tier was constructed by the RF, Gradient Boosting (GBM), and Extreme Gradient Boosting Machine (XGBoost) classifiers. These classifiers predicted CHD in a parallel manner, and its output fed the second-tier. The final prediction was made by Generalized Linear Model (GLM). The comparison results showed the proposed method outperformed DT, RT, Classification and Regression Trees (CART), RF, GBM, and XGBoost algorithms on all datasets except the Cleveland dataset. Wang *et al.* designed a two-level stacking based model and evaluated it on the Z-Alizadeh Sani dataset [12]. The predicted outputs of the first level (base-level) classifiers, including RF, Extra Trees (ET), AdaBoost, SVM, Multi-layer Perceptron (MLP), XGBoost, Gaussian Process Classification (GPC), NB, and LR were given as an input of the second level (meta-level) classifier based on LR. The proposed method outperformed the compared machine learning algorithms by the accuracy, sensitivity, and specificity of 95.43%, 95.84%, and 94.44%. By using an ensemble approach, these studies outperformed the single machine learning algorithms on the Z-Alizadeh Sani dataset.

In recent years, deep learning techniques have been successfully used to diagnose and predict disease. Deep learning is derived from the conventional neural network but it is designed for using numerous hidden layers without requiring any human-designed rules [25]. Atkov *et al.* developed an NN-based model with two hidden layers (four neurons in each hidden layer) for predicting CHD using genetic and non-genetic CHD risk factors [15]. The authors built ten predictive models from different risk factors; the accuracy reached 93% on 487 patients' data in Central Clinical Hospital No. 2 of Russian railways. Samuel *et al.* proposed a combination of an Artificial Neural Network (ANN) and Fuzzy Analytic Hierarchy Process (Fuzzy-AHP) techniques for heart failure risk prediction [16]. Fuzzy-AHP technique was used to compute the global weights for the attributes based on the fuzzy triangular membership function. Then, global weights that represent the contributions of attributes were applied to train the ANN. The performance of the proposed method was evaluated on the Cleveland Heart Disease dataset with 297 patients. As a result, the proposed method showed an accuracy of 91.10%, which is 4.4% higher than conventional ANN. Darmawahyuni *et al.* used DNN for CHD prediction on the Cleveland Heart Disease dataset [17]. The authors chose the number of hidden layers from one to five, and each layer had a hundred neurons. The best-performed model had three hidden layers, and its accuracy, sensitivity, and specificity reached 96%, 99%, 92%, respectively. Ayon *et al.* compared LR, SVM, DNN, DT, NB, RF, and KNN for predicting CHD [18]. The Statlog and Cleveland heart disease datasets retrieved from the UCI machine learning repository were used in an experimental study. As a result, the DNN with four hidden layers (the number of neurons in hidden layers were 14, 16, 16, and 14, respectively) showed the highest accuracy of 98.15%, sensitivity of 98.67%, and precision of 98.01%. Khaneja *et al.* focused on the class imbalance problem using

NNs with two hidden layers; each layer had 256 nodes [19]. The proposed method consisted of two identical NNs that work together and share their weights. First, an input pair was prepared by the combination of two records based on the generation of random numbers. Next, the prepared pair was given to the model, and the NNs received one sample each from the pair. Then the distance was calculated from outputs of these NNs and used in the calculation of loss. In the experiment on the Framingham Heart Study dataset that contains 4,240 samples with 16 columns, the accuracy of the proposed method was 99.66%. M.A. Khan proposed an Internet of Things (IoT) framework for CHD prediction based on a deep Convolutional Neural Network (MDCNN) classifier that optimized by an Adaptive Elephant Herd Optimization (AEHO) algorithm [26]. First, the smart watch and heart monitor devices were attached to a patient for monitoring the blood pressure and Electrocardiogram (ECG). Then MDCNN was utilized for classifying the received sensor data into normal and abnormal. It outperformed the compared algorithms such as DNN and LR classifiers. And its accuracy reached 93.3%, 98.2%, and 96.3% on Cleveland, Framingham, and Sensor datasets, respectively.

Recently, several studies have been conducted using the KNHANES dataset related to the Korean population. Kim *et al.* developed a CHD prediction model based on Fuzzy Logic and DT for Koreans [27]. The model used the Framingham risk factors (gender, age, Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP), Total Cholesterol (TCL), High-Density Cholesterol (HDL), obesity, smoking) and diabetes as input. The proposed model contrasted with NN, SVM, LR, and DT classifiers and gave the highest accuracy and sensitivity scores with 69.51% and 93.10%, respectively. Lim *et al.* proposed the optimized DBN model to predict CHD risk on the KNHANES-VI dataset with 748 instances using the Framingham risk factors [28]. The optimum number of nodes and layers in the DBN was derived through the genetic algorithm. They compared the result of the Optimized-DBN with NB, LR, RF, and FRS algorithms. The proposed approach showed the highest performance with an accuracy of 89.24%, specificity of 74.40%, sensitivity of 85.49%, and AUC of 76.20%. Amarbayasgalan *et al.* proposed a deep learning-based CHD risk prediction model (DAE-NNs) consisted of a Deep Autoencoder (DAE) and two DNN models [29]. The DAE-NNs used the Framingham risk factors as an input of the model, and it was evaluated on the fifth and sixth KNHANES datasets, including 25,990 patients. First, the training dataset was divided into two groups by a RE-based threshold from the DAE model. Then, DNN classifiers were trained on each group. As a result, the performance measurements, including accuracy, f-measure, and AUC reached 83.53%, 84.36%, and 84.02%, respectively. NN with a feature correlation analysis (NN-FCA) approach has been presented [30]. They have performed a statistic-based feature selection for the sixth KNHANES dataset with 4,146 records. The selected features such as age, Body Mass Index (BMI), TCL, HDL, SBP, DBP, triglyceride, smoking status, and diabetes were given as an input of the NN model with three hidden layers. Compared to the results of the Framingham Risk Score (FRS) and LR model, their proposed model has shown high performance with an accuracy of 82.51% and AUC of 74.9%. According to the KNHANES-VI dataset with 4,244 records, Kim *et al.* proposed a CHD risk prediction method based on the Statistics and Deep Belief Network (DBN) [31]. First, important features such as age, SBP, DBP, HDL, diabetes, and smoking were selected by the statistical analysis. Then, DBN with two hidden layers was worked as a predictor using the selected features. As a result, the Statistical-DBN outperformed NB, LR, SVM, RF, and DBN, and its accuracy and AUC reached 83.9% and 79.0%, respectively. The authors of [20] proposed a CHD risk prediction model using Autoencoder and DNN models. The first Autoencoder model was trained on a dataset labeled as risky for feature extraction. The second Autoencoder model was trained on the whole dataset to select an appropriate prediction model from two DNN classifiers. They selected fourteen risk factors, such as age, knee joint pain status, lifetime smoking status, waist circumference, neutral fat, BMI, weight change in one-year status, SBP, TCL, obesity status, frequency of eating out, HDL, marital status, and diabetes from the KNHANES dataset using an Extremely Randomized Tree classifier. As a result, the proposed method outperformed machine learning algorithms; its accuracy, precision, recall, f-measure, and AUC score reached 86.33%, 91.37%, 82.90%, 86.91%, and 86.65%, respectively.

Most proposed methods in previous studies were based on the whole training dataset. The proposed method in this study is unlike them. It focuses on data distribution to prepare training datasets efficiently using PCA and VAE models. First, the training dataset is partitioned into two groups by their divergence using the RE-based threshold estimated from the PCA model. PCA is used to reduce the dimensionality of a dataset by projecting high dimensional space into lower-dimensional space. RE occurs when transforming back the lower-dimensional representation of data to its original dimension. In other words, data with high RE (highly biased) and low RE (regular) are grouped separately. For VAE models, they employ to enrich the highly biased group by generating similar samples with normal (labeled as 0) and risky (labeled as 1) data in the highly biased group. Finally, the first DNN classifier learns from the regular group that includes data with low RE, and the second DNN classifier learns from the enriched highly biased group. At the prediction time, only one appropriate classifier employs to predict CHD risk from the given input. For selecting an appropriate DNN classifier, the proposed method checks the given input is whether closer to the highly biased group based on its reconstruction error on the PCA model. First, input data is given to the PCA model to get reconstruction error. If returned reconstruction error exceeds the threshold calculated by equation (2), the DNN model that was trained on the highly biased training group will be used; otherwise, the DNN model based on the regular training group will predict the class label. By preparing

well-ordered two training groups, the proposed method successfully improved the performance of a single DNN classifier that is based on the whole training dataset.

III. THE PROPOSED METHOD FOR CHD RISK PREDICTION

The proposed method consists of three modules, as shown in FIGURE 1. The first module (Preparation of two groups) splits the whole training dataset into highly biased and regular groups, the second module (Enrichment of the highly biased training group) generates samples similar with both normal and risky datasets in the highly biased group, and the third module (CHD risk predictor) builds two DNN classifiers to predict output from the given unseen data as normal or risky.

A. PREPARATION OF TWO TRAINING GROUPS

In this module, two groups of training datasets are prepared from the initial training dataset. The whole training dataset is divided into two subsections by their divergence using the PCA model. PCA is a dimensionality reduction technique that transforms input variables into a lower-dimensional space that contains most information of the input variables. It is possible to reconstruct back original space of the input from the lower-dimensional data. RE is a difference between input data and its inverse transformation (reconstruction) on the PCA model. The proposed method uses RE for distinguishing the highly biased subset from the training dataset. First, the PCA model is trained on the whole training dataset. Thus, it is more suitable for commonly distributed data than for highly biased data. And it projects common data into lower dimensional space with less information loss and reconstructs back with a smaller error. It is possible to separate the highly biased subset from the dataset based on the RE of the PCA model. RE is calculated through the mean of the squared difference between the input features and its reconstruction; it can be defined as (1):

$$\mathrm{RE} = \frac{1}{n}\sum_{i=1}^{n} \left\| x_i - x'_i \right\|_2^2 \tag{1}$$

where n is the number of input features; x_i is the i-th feature; x'_i is the reconstruction of the i-th feature.

First, we calculate the RE of the training dataset on the PCA model. Then, a threshold to split training dataset is estimated by the mean and standard deviation of these REs; it can be described as (2):

$$\mathrm{TRE} = \frac{1}{k}\sum_{i=1}^{k} RE_i + \sqrt{\frac{1}{k}\sum_{i=1}^{k}\left(RE_i - \frac{1}{k}\sum_{i=1}^{k} RE_i\right)^2} \tag{2}$$

where k is the number of instances in the training dataset; REi is the reconstruction error of the i-th training instance. As a result of this module, two different groups of training datasets are prepared, as well as the RE-based threshold is estimated for further analysis. Later, the threshold is also used to select an appropriate CHD risk prediction model from the DNN models trained on the prepared two groups.

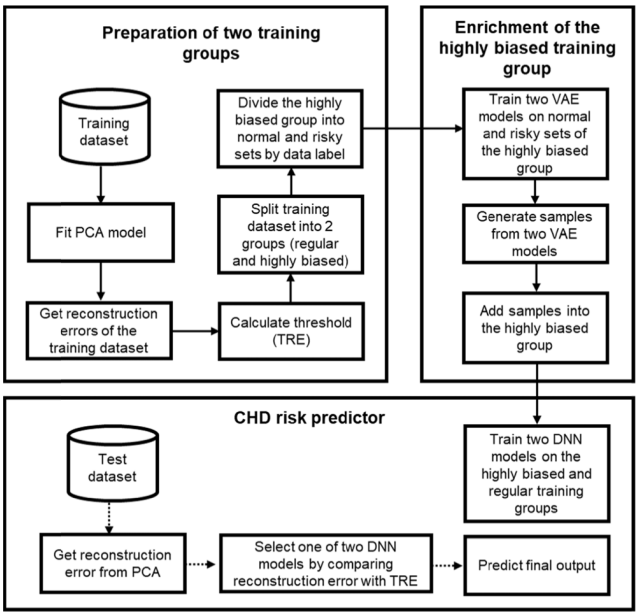


FIGURE 1. General architecture of the proposed method. Solid line indicates the training process and dotted line shows the prediction steps.

B. ENRICHMENT OF THE HIGHLY BIASED TRAINING GROUP

Instead of using prepared training groups directly, two VAE models enrich the highly biased training group. The first VAE model is for generating samples labeled as risky, remained one is for generating samples labeled as normal. FIGURE 2 presents the process of enrichment of the highly biased training group using two VAE models. In this figure, datasets in green are the results of the previous module (preparation of two training groups). All data with the RE greater than or equal to the TRE are assigned to the highly biased group. And it is again bisected into the normal and risky sections according to the class label. Each section is used to train VAE models named VAE-normal and VAE-risky, as shown in FIGURE 2.

VAE was first introduced by [32], and its architecture consists of encoder and decoder parts. The encoder compresses the data to the encoded space, also named latent space, whereas the decoder decompresses them. In VAE, the encoder part is trained to return mean and variance that describe the normal distribution, and it encodes an input as a distribution instead of encoding it as a fixed vector. The loss function of the VAE is defined by two terms, such as the reconstruction loss calculated by the difference between original data and its reconstructed output and the Kullback-Leibler (KL) Divergence score that quantifies how much latent distribution differs from the standard normal distribution. VAE minimizes the loss during training to learn the latent distribution to be as close as possible to the standard normal distribution. The loss can be calculated as (3):

$$loss = \frac{1}{n}\sum_{i=1}^{n}(x_i - x'_i)^2 + \mathrm{KL}[(\mathrm{N}(\mu, \sigma)||\mathrm{N}(0, 1)] \tag{3}$$

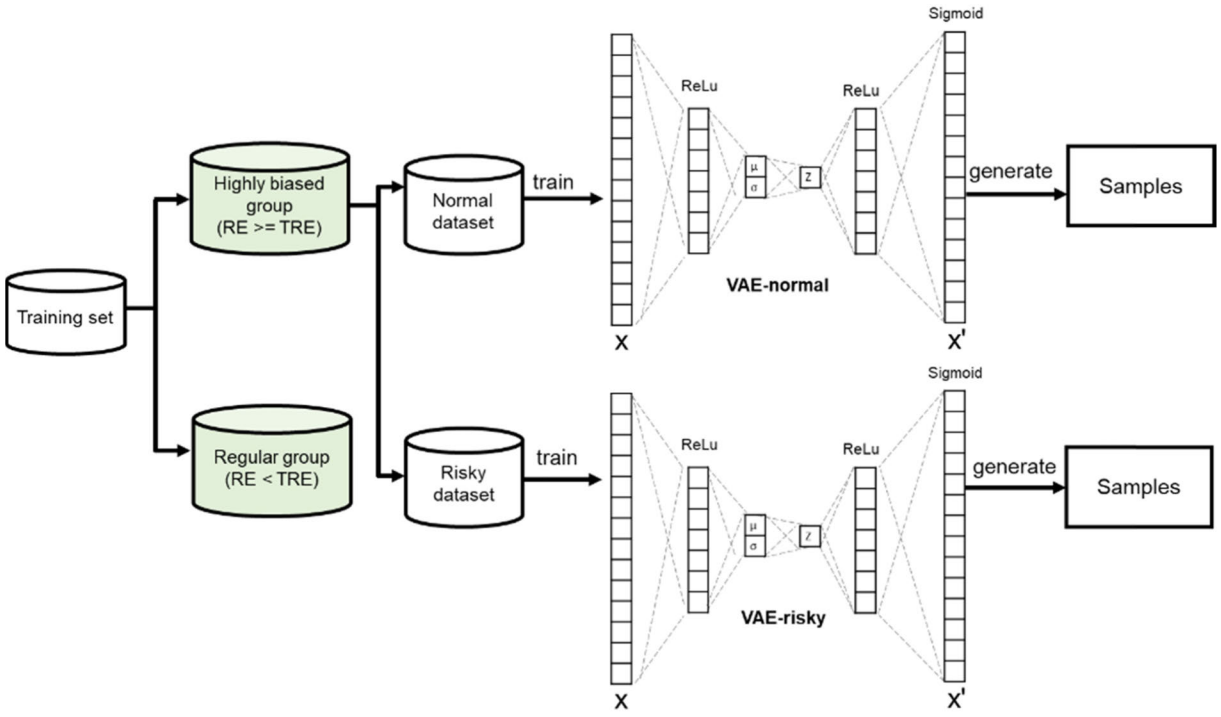


FIGURE 2. Process of enriching the highly biased training group by VAE models.

where n is the number of instances; x_i is the i-th instance; x_i' is the reconstruction of x_i; μ and σ are mean and variance of the latent distribution.

FIGURE 3 represents the architecture of the used two VAE models. Each hidden layer uses the ReLu activation function as given in (4), and the output layer uses the sigmoid activation function shown in (5).

The ReLU activation function is usually used in hidden layers. It can be described as (4):

$$\mathrm{ReLU}(x) = \begin{cases} 0 & if\ x \leq 0 \\ x & if\ x > 0 \end{cases} = max\{0, x\} \tag{4}$$

The sigmoid activation function converts input x into a value between 0 and 1, and it is especially used to predict the probability of output. It can be described as (5):

$$\mathrm{sigmoid}(x) = \frac{1}{1 + e^{-x}} \tag{5}$$

First, the input is encoded as a distribution over the latent space. Second, an input of the decoder (z) is randomly sampled from the latent distribution. Then, the sampled point z is decoded to the output. In this study, the latent distribution is chosen to be normal, and the encoder is trained to return the mean and variance that describes the normal distribution. To generate samples using the VAE model, ε is sampled randomly from the standard normal distribution and add it to the mean value (μ) by multiplying it by the standard deviation (σ) of its latent distribution for obtaining z,

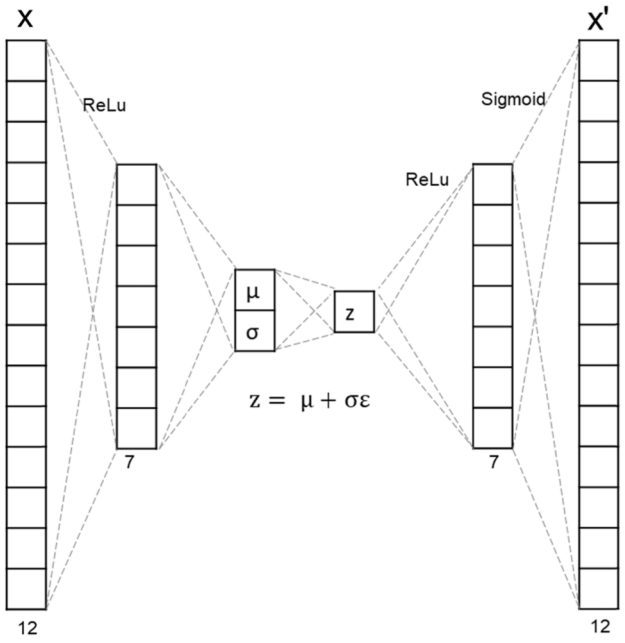


FIGURE 3. Architecture of the VAE models used to generate samples.

as described in (6). Finally, the sampled point z is decoded to get new data. The decoded output of z is a generated sample.

$$z = \mu + \sigma\varepsilon \tag{6}$$

where ε is the random value from the standard normal distribution; μ and σ are the latent distribution's mean and standard deviation.

C. CHD RISK PREDICTOR

In this study, we use DNN model to predict CHD risk. NN model was first proposed by Warren McCullough and Walter Pitts in 1943 [33]. It has been applied successfully to speech recognition [34], emotion recognition [35], disease predictions [36], and so on.

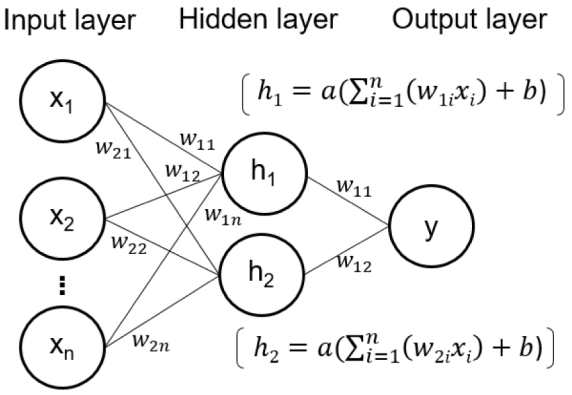


FIGURE 4. Example of NN architecture with one hidden layer.

FIGURE 4 shows the NN example that has an input layer with n neurons, a hidden layer with two neurons, and an output layer with one neuron. The input layer is composed of neurons that describe input features, whereas neurons in hidden and output layers receive results of activation function that converts the weighted summation of the neurons of the previous layer. The output of the NN represented in FIGURE 4 can be written in (7):

$$y = a(w_{11}a(\sum_{i=1}^{n}(w_{1i}x_i) + b) + w_{12}a(\sum_{i=1}^{n}(w_{2i}x_i) + b) + b) \tag{7}$$

where a is an activation function, w is the weight matrix, x is the input vector, and b is the bias.

In the CHD risk predictor module, two DNN models are trained on the prepared training groups by splitting the whole training dataset. In practice, a dataset can include a subset that is higher variance than most data. This highly biased subset degrades the performance of predictive models. Therefore, we isolate a highly biased subset from the common subset using the RE of the PCA model. It also gives a possibility to train two distinct predictive models targeted for regular input and biased input separately. By using two distinct classifiers for regular and biased distributions, it can classify the regular as well as biased input data well. Moreover, we augmented the biased section with new samples generated from VAE models. As a result, the performance of the predictive model trained on the biased section is more improved.

The architecture of the proposed DNN models is the same as each other, as shown in FIGURE 5. Each model has six hidden layers with 71, 51, 31, 11, 5, and 3 neurons, respectively, and all of the hidden layers use the ReLU activation function. The input layer consists of 12 neurons for CHD risk factors to predict the target variable. The output layer uses the sigmoid activation function for the binary classification problem; it returns the probability associated with class 1 as a value from 0 to 1. Finally, a high probability class is selected for the output result.

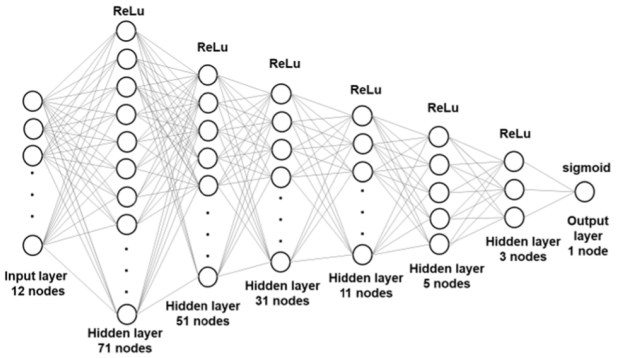


FIGURE 5. Proposed DNN architecture.

FIGURE 6 shows the prediction steps. In the process of CHD risk prediction, first, input data is given to the PCA model, and its RE on PCA is calculated. If the RE exceeds the threshold (TRE) estimated by (2), then DNN-biased trained on the highly biased training group employs; otherwise, the DNN-regular trained on regular training group with low RE predict the CHD risk.

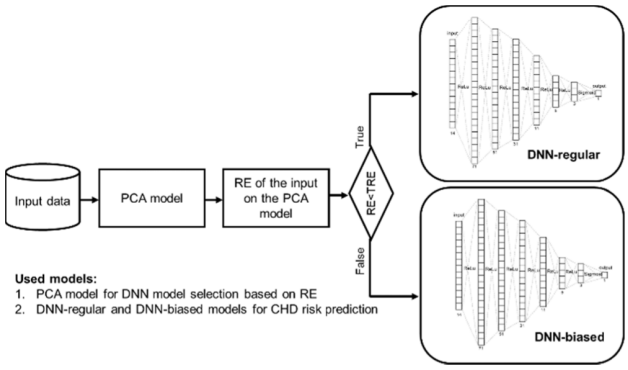


FIGURE 6. Steps to predict the CHD risk of the proposed method.

IV. EXPERIMENTAL STUDY

To evaluate the proposed method, we have conducted two types of experiments. In the first type of experiment, we proved how each model improves the prediction performance. In other words, the purpose of this experiment is to show the contribution of the performance improvement of used models. Therefore, first, we trained a predictive model based on DNN from the whole training dataset without any other models, and it was used as a baseline model. In the proposed method, we prepared training groups from the initial

training dataset using PCA and VAE models to improve the baseline model. We showed how the prediction performance was improved using the PCA model first and then the VAE model step by step. The following models were compared in this experiment:

- The single DNN model trained on the whole training dataset and its architecture was the same as the two DNNs used in the proposed method. It was used as a baseline model.
- Two DNN models that were trained on the training groups divided by the PCA model (the first step of preparing well-ordered training groups in the proposed method).
- Two DNN models that were trained on the training groups divided by the PCA model. However, the highly biased training group was enriched by two VAE models (the second step of preparing well-ordered training groups in the proposed method).

The comparison between the baseline model and the two-DNN shows that the two-DNN improves the performance of the single DNN model significantly. After that, we showed the performance improvement of the two-DNN by enriching the highly biased training group using generated samples from VAE models.

In the second kind of experiment, we compared the proposed method with machine learning-based algorithms, including NB, RF, KNN, DT, SVM, and AdaBoost.

A. EVALUATION METRICS

This section describes performance measurements for prediction models on the test dataset. The confusion matrix is a table to visualize the performance of classification models when data labels are available. It represents the total number of correct (True Positives (TP) and True Negatives (TN)) and incorrect predictions (False Positives (FP) and False Negatives (FN)).

Accuracy is the proportion of correct predictions among all data. It defined by (8):

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

True Positive Rate (TPR) known as ''Sensitivity'' or ''Recall'' is defined as the fraction of positive instances predicted correctly by the model, which is defined by (9):

$$\text{Recall} = \frac{TP}{TP + FN} \tag{9}$$

Precision is a fraction of TP predictions among all positive predictions. It evaluates the effectiveness of TP predictions. Precision can be defined as (10):

$$\text{Precision} = \frac{TP}{TP + FP} \tag{10}$$

However, it is difficult to compare models with low precision with high recall or high precision with low recall. Thus, F-measure is used to measure precision and recall together, where a high value indicates a good result. It can be defined as (11):

$$\text{F} - \text{measure} = \frac{2 * precision * recall}{precision + recall} \tag{11}$$

ROC curve is a graphical representation of the balance between the TPR (y-axis) and FPR (x-axis) of a classifier. It compares the performance of several classifiers together and evaluates which model is better on average. It indicates how much a classification model is capable of distinguishing between classes [37]. If the model is perfect, the area under the ROC curve (AUC) is close to 1. A model with a larger AUC is better.

B. DATASET

KNHANES is a nationwide program to evaluate Koreans' health and nutritional status. It has been continuously conducted by the Korea Centers for Disease Control and Prevention (KCDC) since 1998 [38]. KNHANES dataset consists of 3 parts: medical examination, health survey, and nutrition survey described in TABLE 1.

TABLE 1. KNHANES dataset survey contents.

Types	Contents
Medical examination	Basic examination, blood pressure measurement, body measurement, blood test, urine test, lung function test, chest x-ray, oral health examination, Ear-Nose-Throat examination, eye examination, grip test
Health survey	Household survey, Subjective Health status, medical use, Health checkup and vaccination, Activity restrictions and quality of life, Damage (accidents and poisoning), Hospitalization, Outpatient use, Patient experience, Education and economic activity, Obesity and weight control, Drinking, Safety awareness, Mental health, smoking, Physical activity, Women's health
Nutrition survey	Dietary Survey, Food intake frequency survey, food intake, Food stability survey

We have analyzed samples spanning the years 2010-2015. In this experiment, we used a total of 25,340 records without a history of previous myocardial infarction or angina from the KNHANES dataset. If a patient has been diagnosed with myocardial infarction or angina and the first diagnosed age is younger than the current age, we removed the record. The final output of the proposed method is to predict whether there is a risk of CHD from the given input. The final experimental dataset consisted of 10,991 men and 14,349 women. From them, 15,796 records were high risk, and 9,544 records were normal. Risk factors including age, knee joint pain status, waist circumference, neutral fat, BMI, weight change in one-year status, SBP, TC, obesity status, frequency of eating out, HDL, and marital status were used to predict CHD risk [20]. The general descriptions of the risk factors used in the experimental study are shown in TABLE 2.

TABLE 2. Description of the CHD risk factors.

Risk factors	Normal (9,544 records)	High risk (15,796 records)
Age (year)	38.95 (17.28)	53.85 (16.17)
Body mass index (BMI) (kg/m2)	22.27 (3.24)	24.30 (3.38)
Total cholesterol (TC) (mg/dL)	86.32 (41.65)	192.78 (39.38)
High-density lipoprotein cholesterol (HDL) (mg/dL)	54.22 (10.17)	47.68 (12.13)
Systolic blood pressure (SBP) (mmHg)	111.35 (14.43)	122.80 (17.07)
Waist circumference (WC) (cm)	75.04 (9.09)	83.80 (9.38)
Neutral fat (NF) (mg/dL)	86.32 (41.65)	154.84 (117.59)
Obesity status		
1. Underweight	999	424
2. Normal	6827	9280
3. Obesity	1718	6092
Knee joint pain status		
1. Yes	603	2134
2. No	2476	8305
8. Non − applicable (below the 50 years of age)	6462	0
9. No response	3	5
Weight change in one year status		
1. No change	5171	10459
2. Weight loss	956	2299
3. Weight gain	1868	2811
9. No response	1549	16
Frequency of eating out year status		
1. More than twice a day	664	1206
2. Once a day	1392	2363
3. 5 to 6 times a week	2002	2016
4. 3 to 4 times a week	944	1383
5. 1 to 2 times a week	2250	3292
6. 1 to 3 times a month	1735	3620
7. Less than once a month	555	1914
9. No response	2	2
Marital status		
1. Married, living together	5786	11781
2. Married, living separately	36	99
3. Bereavement	444	1584
4. Divorced	185	537
8. Response refused	243	0
9. No response	0	0
88. Non − applicable	2850	1501

C. PARAMETER TUNING FOR COMPARED MACHINE LEARNING ALGORITHMS

To compare the proposed method with other machine learning algorithms, we used the sklearn library [39]. The following Python implementations were used for the compared machine learning algorithms:

- K-Nearest Neighbors: sklearn.neighbors.KNeighborsClassifier
- Naïve Bayes: sklearn.naive_bayes.GaussianNB
- Support Vector Machine: sklearn.svm.SVC
- Decision Tree: sklearn.tree.DecisionTreeClassifier
- Random Forest: sklearn.ensemble.RandomForestClassifier
- Adaboost: sklearn.ensemble.AdaBoostClassifier

We chose optimal values for input parameters of the compared algorithms by changing values until decreasing the model performance. The configurations of the parameters for each algorithm are shown in TABLE 3.

TABLE 3. Parameter configuration of the compared algorithms.

Algorithm	Parameter configuration	Optimal values
NB	Default configuration	
KNN	n_neighbors: The number of neighbors. n_neighbors parameter was configured between 2 and 20.	n_neighbors = 24
DT	criterion: "gini" for the Gini impurity and "entropy" for the information gain measurements; they were used to identify the best decision tree splitting candidate.	criterion = "entropy"
RF	n_estimators: The number of trees in the forest. It was configured between 10 and 200 and increased by 10. criterion: "gini" and "entropy" were used for splitting criteria.	n_estimators = 80 criterion = "entropy"
SVM	kernel: It specifies the kernel type to be used in the algorithm. It must be one of "linear", "poly", "rbf", or "sigmoid".	kernel = "linear"
AdaBoost	n_estimators: The maximum number of estimators at which boosting is terminated. It was configured between 10 and 200 and increased by 10.	n_estimators = 90

The proposed DNN model was trained with Adam optimizer [40], a learning rate of 0.001, batch size of 32, and epochs of 1000. Early stopping [41] with the validation accuracy as a stopping criterion and patience of 500 epochs is applied. The proposed method uses 90% of data for training, 10% of the training set for validating, and remained 10% of data for testing. For the VAE models, they trained with Adam optimizer, a learning rate of 0.001, batch size of 8, and epochs of 1000.

V. EXPERIMENTAL RESULTS

We conducted two types of experiments to evaluate the proposed method. The first kind of experiment proved how the components of the proposed method work together more efficiently and improve the prediction performance. The comparison between the proposed method and machine learning algorithms such as NB, RF, KNN, DT, SVM, and AdaBoost was shown in the second kind of experiment.

A. RESULTS OF THE FIRST EXPERIMENT

In this section, the performance improvement of the single DNN model that is the baseline model using two DNN models learned from the prepared training groups is detailed. We compared the baseline model to Two-DNN and VAE-Two-DNN models.

The single DNN was trained on the whole training dataset. The Two-DNN was trained on the highly biased and regular training groups separated by the PCA model. For VAE-Two-DNN, we enriched the highly biased training group with samples generated from VAE models. There were 3,000 samples generated for each class label (normal and risky), respectively.

1) PREPARE HIGHLY BIASED AND REGULAR GROUPS BY SPLITTING THE TRAINING DATASET USING PCA

According to the proposed method, the training dataset is divided into two groups based on the PCA model. First, the PCA model was trained on the whole training dataset, and it fitted more for the common dataset. Therefore, the data that is different from most data gives higher RE than common data on the PCA model. Based on this characteristic, we distinguished the highly biased section of the training dataset. By separating highly biased data and commonly distributed data, the dataset with high divergence can be modeled independently to improve the prediction performance. In this experiment, the number of principal components of the PCA model was 6, which can explain 95% of the input variance.

Each group consisted of a dataset labeled as both risky and normal. According to the partitioned groups in 10-fold cross-validation, the highly biased group accounted for approximately 9.07% of the whole training dataset (from 8.94% to 9.42% in each fold), and about 82.6% was labeled as risky. As shown in FIGURE 7, the mean values (and standard deviation in parentheses) of risk factors such as age, waist circumference, neutral fat, body mass index, systolic blood pressure, total cholesterol, and high-density lipoprotein cholesterol were 47.66 (18.12), 80.12 (9.81), 123.90 (79.77), 23.40 (3.28), 117.52 (15.71), 186.29 (33.19), 50.06 (10.74) in G1 and 54.03 (16.64), 84.26 (12.87), 180.41 (218.29), 24.85 (4.79), 128.18 (24.91), 206.10 (55.52), 55.10 (20.15) in G2, respectively. However, the deviation of risk factors in the G1 was lower than G2, and about 60.3% were labeled as risky. Moreover, we can see the average neutral fat, total cholesterol, and systolic blood pressure increased significantly in G2.

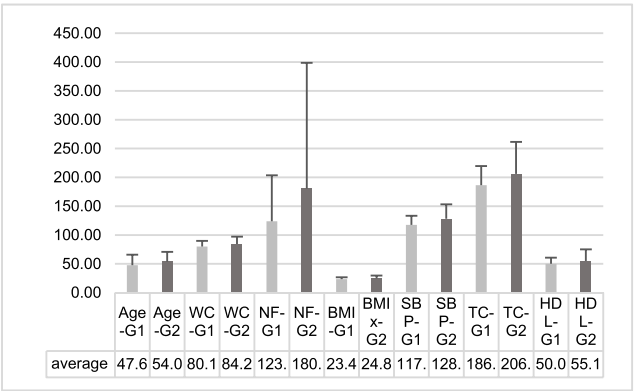


FIGURE 7. Mean and standard deviation of risk factors for two training groups (G1 is group of common subset and G2 is group of highly biased subset).

FIGURE 8 shows the comparison between the Single-DNN and Two-DNN. The accuracy, precision, recall, specificity, and f-measure of the Single-DNN were increased from 0.836, 0.867, 0.870, 0.772, 0.868 to 0.873, 0.900, 0.899, 0.826, and 0.899, respectively by using Two-DNN.

TABLE 4. Results of the ROC curve analysis of the single-DNN and two-DNN on the KNHANES dataset.

Models	AUC	p-value	95% CI
Single-DNN	0.821	2.51E-72	0.803-0.839
Two-DNN	0.862	2.13E-52	0.846-0.878

TABLE 4 shows the results of ROC curve analysis for Single-DNN and Two-DNN models. We tested whether the observed AUC differs significantly from the AUC of 0.5 by Hanley and McNeil test [25]. For both of compared models, AUC (p-value$<$0.000001) was statistically significant. TABLE 4 shows how the AUC of the single DNN model improved from 0.821 to 0.862 using two DNN models. It increased by 4.1% by separating a biased section from the training dataset.

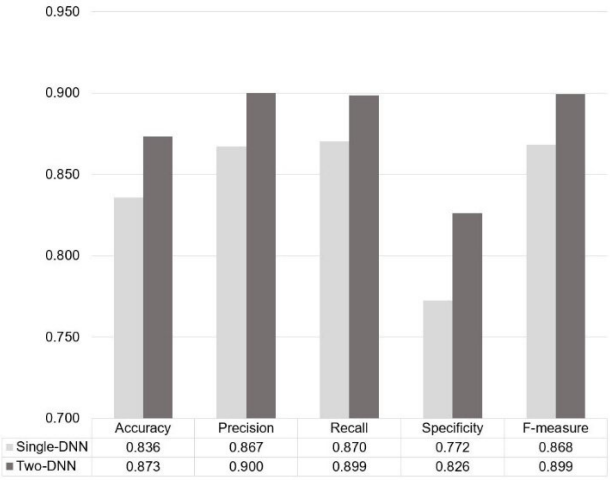


FIGURE 8. Comparison of Single-DNN and Two-DNN on the KNHANES dataset.

2) ENRICH HIGHLY BIASED TRAINING GROUP USING VARIATIONAL AUTOENCODERS

The distinguished highly biased training group is one of the prepared groups based on the PCA model. It consists of data with high RE and may not be sufficient for building the model due to accounts for a small percentage of the total dataset. In the experiment, 9.07% of the whole training dataset belonged to the highly biased group. And 82.6% were risky, 17.4% were normal in this group. The proposed method decides this problem by enriching both risky and normal instances in the highly biased training group via two deep VAE models.

This section describes the performance improvement of the Two-DNN model introduced in the previous section by the enriched training group. We generated 3,000 samples for each class label (normal and risky), respectively, using two different VAE models. Even if 82.6% of the dataset was risky in the highly biased group, the number of risky instances is not big enough for training. Therefore, we used two VAE models for data generation of both risky and normal data.

FIGURE 9 shows the comparison of Two-DNN and VAE-Two-DNN methods. In Two-DNN, there were two DNN models independently trained on the divided training groups directly. For VAE-Two-DNN, the highly biased training group was augmented by newly generated samples. Then two DNN models were independently trained on these groups. As a result, all performance of VAE-Two-DNN outperformed Two-DNN. It increased accuracy, precision, recall, specificity, and f-measure of Two-DNN by 1.9%, 1.07%, 2.12%, 1.77%, and 1.59%, respectively in VAE-Two-DNN.

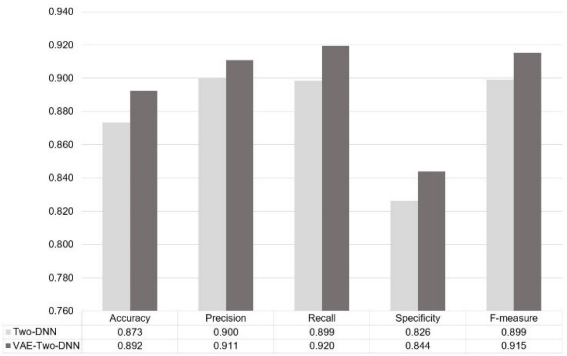


FIGURE 9. Comparison of Two-DNN and VAE-Two-DNN on the KNHANES dataset.

TABLE 5 shows the results of ROC curve analysis for Two-DNN and VAE based Two-DNN methods. For all of the compared methods, AUC (p-value<0.000001) was statistically significant. Although the Two-DNN improved the performance of single DNN, the VAE-Two-DNN outperformed the Two-DNN. In the case of Two-DNN, AUC was 0.862 (95% CI, 0.846-0.878), and it has been improved to 0.894 (95% CI, 0.881-0.906) by using VAE based enriched training group (VAE-Two-DNN), shown in TABLE 5.

TABLE 5. Results of ROC curve analysis of two-DNN and VAE-two-DNN on the KNHANES dataset.

Classifier	AUC	p-value	95%CI
Two-DNN	0.876	9.51E-94	0.862-0.889
VAE-Two-DNN	0.894	3.5751E-103	0.881-0.906

As a result, all of the steps in the proposed method improved the prediction performance of the baseline model.

In the first step, the irregular dataset (highly biased) was distinguished from the training dataset using RE from the PCA model. Then, two DNN models were trained on the separated groups one by one. When predicting the CHD risk using these two models, the PCA model received the input first and returned the RE error. If the returned RE was higher than the threshold, a DNN model based on the irregular training group predicted the CHD risk. In the opposite case, a DNN model learned from the regular training group was employed to predict CHD risk. In the second step, the irregular training group was enriched by samples generated from the VAE models because it consisted of insufficient instances to build a predictive model. After that, two DNN models were trained on the regular and enriched highly biased training groups separately. FIGURE 10 shows the improvement of the baseline model step by step.

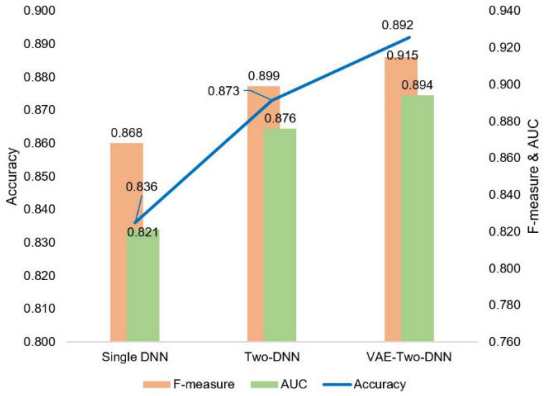


FIGURE 10. Performance improvement of the baseline model by the proposed method step by step.

B. RESULTS OF THE SECOND EXPERIMENT

We compared six machine learning algorithms such as KNN, NB, DT, RF, AdaBoost, and SVM with the proposed method using the 10-fold cross-validation. In order to compare them, the whole training dataset was used without splitting to train these six machine learning algorithms. TABLE 6 shows the average confusion matrix obtained from the 10-fold cross-validation. The parameter configurations of these algorithms are shown in TABLE 3.

TABLE 7 shows the results of CHD risk prediction models based on risk factors shown in TABLE 2, and the highest values of the performance are marked in bold. According to the compared machine learning algorithms, the RF algorithm showed the highest performance than KNN, NB, DT, SVM, and AdaBoost algorithms. Its accuracy, precision, recall, specificity, f-measure were 0.827, 0.859, 0.863, 0.760, and 0.861, sequentially. However, the results show that the proposed VAE-Two-DNN method achieves the best performance. The accuracy of the RF algorithm increased by 6.56% in the proposed method. Also, it improved the specificity by 8.37%. It is a fraction of the true normal predictions over the total amount of dataset labeled by normal. The recall measures what proportion of the dataset labeled as risky was predicted correctly, and the precision evaluates how many percent of total risky predictions is correct. The proposed VAE-Two-DNN incremented the recall, precision as well as f-measure by 5.68%, 5.2%, and 5.44%, respectively. Therefore, VAE-Two-DNN successfully improved the prediction of both normal and risky cases.

FIGURE 11 shows the AUC of compared seven algorithms together. The proposed VAE-Two-DNN method outperformed the AUC of all compared algorithms by giving the AUC of 0.882.

TABLE 8 shows the results of ROC curve analysis for all compared algorithms. The AUC of the proposed method was 0.881 (95% CI, 0.867-0.896), and it improved the highest AUC of compared algorithms (AUC of RF) by 7.2%.

TABLE 6. Average confusion matrix obtained from the 10-fold cross-validation.

KNN	True class: Positive	True class: Negative
Predicted class: Positive	1312.8	277.3
Predicted class: Negative	266.8	677.1

NB	True class: Positive	True class: Negative
Predicted class: Positive	1289.1	319.3
Predicted class: Negative	290.5	635.1

SVM	True class: Positive	True class: Negative
Predicted class: Positive	1347.8	325.2
Predicted class: Negative	231.8	629.2

RF	True class: Positive	True class: Negative
Predicted class: Positive	1363.8	223.5
Predicted class: Negative	215.8	730.9

AdaBoost	True class: Positive	True class: Negative
Predicted class: Positive	1349.2	220.2
Predicted class: Negative	230.4	734.2

DT	True class: Positive	True class: Negative
Predicted class: Positive	1279	304.4
Predicted class: Negative	300.6	650

VAE-Two-DNN	True class: Positive	True class: Negative
Predicted class: Positive	1449.8	143.2
Predicted class: Negative	129.8	811.2

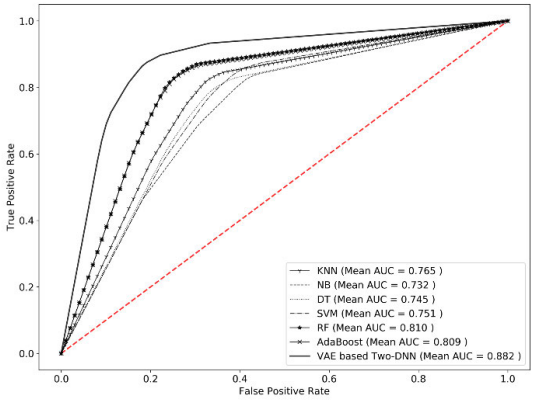


FIGURE 11. Average AUC of the compared seven algorithms together.

The comparative evaluation of the proposed method and existing CHD risk prediction methods on our experimental dataset are limited because the existing methods are not publicly available. Therefore, we did not run existing methods on our experimental dataset, and a comparison was made by taking the results from the papers. VI shows the comparison of the existing methods in previous studies and the proposed method on the KNHANES dataset, and the highest values of evaluation scores are marked in bold.

TABLE 7. Results of compared algorithms on the KNHANES dataset.

Classifier	Accuracy	Precision	Recall	Specificity	F-measure
KNN	0.785	0.825	0.829	0.702	0.827
NB	0.759	0.802	0.812	0.652	0.807
DT	0.761	0.807	0.810	0.674	0.808
SVM	0.780	0.805	0.852	0.650	0.828
RF	0.827	0.859	0.863	0.760	0.861
AdaBoost	0.822	0.859	0.854	0.763	0.857
VAE-Two-DNN	**0.892**	**0.911**	**0.920**	**0.844**	**0.915**

TABLE 8. Results of the ROC curve analysis for compared algorithms on the KNHANES dataset.

Classifier	AUC	p-value	95%CI
KNN	0.765	4.29E-52	0.745-0.786
NB	0.732	2.36E-33	0.711-0.753
DT	0.745	2.51E-38	0.721-0.762
SVM	0.751	2.02E-40	0.730-0.771
RF	0.810	1.90E-72	0.793-0.830
AdaBoost	0.809	5.35E-68	0.790-0.827
VAE-Two-DNN	0.882	1.34E-56	0.867-0.896

TABLE 9. Comparative evaluation of different algorithms in previous researches and the proposed method on the KNHANES dataset.

Algorithm	Accuracy	Precision	Recall	F-measure	AUC
Fuzzy logic and decision tree [27]	0.695	0.699	-	-	0.594
Neural network and feature correlation analysis [30]	0.825	-	-	-	0.749
Statistical deep belief network [31]	0.839	-	0.876	-	0.79
Optimized deep belief network [28]	0.774	-	0.855		0.829
DAE-DNNs [29]	0.825	0.895	0.797	0.843	0.840
AE-DNNs [20]	0.863	**0.913**	0.829	0.869	0.866
VAE-Two-DNN (proposed)	**0.892**	0.911	**0.920**	**0.915**	**0.894**

VI. CONCLUSION

In this study, we proposed the CHD risk prediction method based on two DNN models and applied it to the KNHANES dataset. The proposed method addressed preparing an efficient training dataset by distinguishing and enriching the highly biased subset that degrades the model performance using the PCA and VAE models. First, we grouped the highly biased subset from the whole training dataset using the PCA model. This is because the highly biased subset of the training dataset degrades the performance of predictive models. It is possible to improve the performance of a single predictive model trained on the whole training dataset by two different predictive models trained on the highly biased and remained common subsets. Therefore, to address this issue, we suggested using RE from the PCA model. As a result, the performance of CHD risk predictor based on the single DNN (accuracy: 0.836, precision: 0.867, recall: 0.870, specificity: 0.772, f-measure: 0.868, AUC: 0.821) improved by using two

DNN models learned from the partitioned training using the PCA model (accuracy: 0.873, precision: 0.90, recall: 0.899, specificity: 0.826, f-measure: 0.899, AUC: 0.862).

For improving the prediction performance by enriching the insufficient number of instances in the highly biased training group, the proposed method was designed to use two deep VAE models. The performance of CHD risk predictor based on two DNN models learned from the partitioned training groups was improved by using VAE based two DNN models learned from the enriched highly biased training group and regular training group (accuracy: 0.892, precision: 0.911, recall: 0.920, specificity: 0.844, f-measure: 0.915, AUC: 0.882).

To evaluate the proposed method, the proposed method was compared with various machine learning algorithms. The evaluation results showed that the proposed method improved the accuracy, specificity, f-measure, and AUC of NB, SVM, DT, KNN, AdaBoost, and RF by (13.3, 19.2, 18.3, 7.5), (11.2, 19.4, 16.5, 5.4), (11.2, 12.0, 16.5, 5.4), (10.7, 14.2, 15.0, 5.5), (7.0, 8.0, 10.7, 2.5), and (6.6, 8.4, 10.4, 2.1), respectively.

In shortly, this study proposed the comprehensive prediction method using PCA, VAE and DNN models. The two DNN trained on the partitioned training groups according to the PCA significantly improves the performance. Moreover, the proposed method raises the performance again using the VAE-based enriched training group. We show the performance improvement of the proposed method by using PCA and VAE models in the first experiment, and comparison between the proposed method and other machine learning algorithms in the second experiment.

The limitation of the proposed method is that it does not allow missing values. Therefore, we will focus on handling missing values by generating new values using the VAE model in our future study. Also, the reconstruction error-based threshold was estimated by the mean and standard deviation of the reconstruction errors of the training dataset on the PCA model. Finding the optimal threshold is challenging in this module.

REFERENCES

[1] *Cardiovascular Diseases (CVDs)*. Accessed: Jul. 9, 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] *Statistics Korea. Causes of Death Statistics in 2019*. Accessed: Jul. 9, 2021. [Online]. Available: http://kostat.go.kr/portal/eng/pressReleases/1/index.board?bmode=read&bSeq=&aSeq=385629&pageNo=1&rowNum=10&navCount=10&currPg=&searchInfo=srch&sTarget=title&sTxt=death

[3] *Coronary Heart Disease*. Accessed: Jul. 9, 2021. [Online]. Available: https://www.nhlbi.nih.gov/health-topics/coronary-heart-disease

[4] H. Hausmann, H. Topp, H. Siniawski, S. Holz, and R. Hetzer, "Decision-making in end-stage coronary artery disease: Revascularization or heart transplantation?" *Ann. Thoracic Surg.*, vol. 64, no. 5, pp. 1296–1302, Nov. 1997.

[5] K. S. Ryu, H. W. Park, S. H. Park, H. S. Shon, K. H. Ryu, D. G. Lee, M. E. Bashir, J. H. Lee, S. M. Kim, S. Y. Lee, and J. W. Bae, "Comparison of clinical outcomes between culprit vessel only and multivessel percutaneous coronary intervention for ST-segment elevation myocardial infarction patients with multivessel coronary diseases," *JGC*, vol. 12, no. 3, p. 208, May 2015.

[6] J. Soni, U. Ansari, D. Sharma, and S. Soni, "Predictive data mining for medical diagnosis: An overview of heart disease prediction," *Int. J. Comput. Appl.*, vol. 17, no. 8, pp. 43–48, 2011.

[7] H. Kim, M. Ishag, M. Piao, T. Kwon, and K. Ryu, "A data mining approach for cardiovascular disease diagnosis using heart rate variability and images of carotid arteries," *Symmetry*, vol. 8, no. 6, p. 47, Jun. 2016, doi: 10.3390/sym8060047.

[8] A. H. Gonsalves, F. Thabtah, R. M. A. Mohammad, and G. Singh, "Prediction of coronary heart disease using machine learning: An experimental analysis," in *Proc. 3rd Int. Conf. Deep Learn. Technol. (ICDLT)*, 2019, pp. 51–56.

[9] J. J. Beunza, E. Puertas, E. García-Ovejero, G. Villalba, E. Condes, G. Koleva, C. Hurtado, and M. F. Landecho, "Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease)," *J. Biomed. Informat.*, vol. 97, Sep. 2019, Art. no. 103257. Accessed: Jul. 9, 2021, doi: 10.1016/j.jbi.2019.103257.

[10] J. H. Joloudari, E. Hassannataj Joloudari, H. Saadatfar, M. Ghasemigol, S. M. Razavi, A. Mosavi, N. Nabipour, S. Shamshirband, and L. Nadai, "Coronary artery disease diagnosis; ranking the significant features using a random trees model," *Int. J. Environ. Res. Public Health*, vol. 17, no. 3, p. 731, Jan. 2020, doi: 10.3390/ijerph17030731.

[11] B. A. Tama, S. Im, and S. Lee, "Improving an intelligent detection system for coronary heart disease using a two-tier classifier ensemble," *Biomed Res. Int.*, vol. 2020, Apr. 2020, Art. no. 9816142. Accessed: Jul. 9, 2021, doi: 10.1155/2020/9816142.

[12] J. Wang, C. Liu, L. Li, W. Li, L. Yao, H. Li, and H. Zhang, "A stacking-based model for non-invasive detection of coronary heart disease," *IEEE Access*, vol. 8, pp. 37124–37133, 2020, doi: 10.1109/ACCESS.2020.2975377.

[13] A. Dey, J. Singh, and N. Singh, "Analysis of supervised machine learning algorithms for heart disease prediction with reduced number of attributes using principal component analysis," *Int. J. Comput. Appl.*, vol. 140, no. 2, pp. 27–31, Apr. 2016.

[14] A. K. Gárate-Escamila, A. Hajjam El Hassani, and E. Andrès, "Classification models for heart disease prediction using feature selection and PCA," *Informat. Med. Unlocked*, vol. 19, Jan. 2020, Art. no. 100330.

[15] O. Y. Atkov, S. G. Gorokhova, A. G. Sboev, E. V. Generozov, E. V. Muraseyeva, S. Y. Moroshkina, and N. N. Cherniy, "Coronary heart disease diagnosis by artificial neural networks including genetic polymorphisms and clinical parameters," *J. Cardiol.*, vol. 59, no. 2, pp. 190–194, Mar. 2012, doi: 10.1016/j.jjcc.2011.11.005.

[16] O. W. Samuel, G. M. Asogbon, A. K. Sangaiah, P. Fang, and G. Li, "An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction," *Expert Syst. Appl.*, vol. 68, pp. 163–172, Feb. 2017, doi: 10.1016/j.eswa.2016.10.020.

[17] A. Darmawahyuni, S. Nurmaini, and F. Firdaus, "Coronary heart disease interpretation based on deep neural network," *Comput. Eng. Appl. J.*, vol. 8, no. 1, pp. 1–12, Feb. 2019.

[18] S. I. Ayon, M. M. Islam, and M. R. Hossain, "Coronary artery heart disease prediction: A comparative study of computational intelligence techniques," *IETE J. Res.*, pp. 1–20, 2020, doi: 10.1080/03772063.2020.1713916.

[19] A. Khaneja, S. Srivastava, A. Rai, A. S Cheema, and P. K Srivastava, "Analysing risk of coronary heart disease through discriminative neural networks," 2020, *arXiv:2008.02731*. [Online]. Available: http://arxiv.org/abs/2008.02731

[20] T. Amarbayasgalan, K. H. Park, J. Y. Lee, and K. H. Ryu, "Reconstruction error based deep neural networks for coronary heart disease risk prediction," *PLoS ONE*, vol. 14, no. 12, Dec. 2019, Art. no. e0225991, doi: 10.1371/journal.pone.0225991.

[21] H. Hoffmann, "Kernel PCA for novelty detection," *Pattern Recognit.*, vol. 40, no. 3, pp. 863–874, 2007.

[22] J. A. Jablonski, T. J. Bihl, and K. W. Bauer, "Principal component reconstruction error for hyperspectral anomaly detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 8, pp. 1725–1729, Aug. 2015.

[23] S. Biswal, S. Ghosh, J. Duke, B. Malin, W. Stewart, and J. Sun, "EVA: Generating longitudinal electronic health records using conditional variational autoencoders," 2020, *arXiv:2012.10020*. [Online]. Available: http://arxiv.org/abs/2012.10020

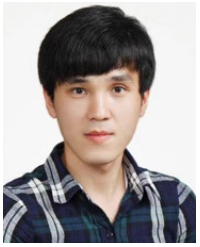
[24] X. Huang, G. Cui, D. Wu, and Y. Li, "A semi-supervised approach for early identifying the abnormal carotid arteries using a modified variational autoencoder," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2020, pp. 595–600.

[25] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions," *J. Big Data*, vol. 8, no. 1, pp. 1–74, Dec. 2021.
[26] M. A. Khan, "An IoT framework for heart disease prediction based on MDCNN classifier," *IEEE Access*, vol. 8, pp. 34717–34727, 2020, doi: 10.1109/ACCESS.2020.2974687.
[27] J. Kim, J. Lee, and Y. Lee, "Data-Mining-based coronary heart disease risk prediction model using fuzzy logic and decision tree," *Healthcare Informat. Res.*, vol. 21, no. 3, pp. 167–174, Jul. 2015.
[28] K. Lim, B. M. Lee, U. Kang, and Y. Lee, "An optimized DBN-based coronary heart disease risk prediction," *Int. J. Comput. Commun. Control*, vol. 13, no. 4, pp. 492–502, Jul. 2018, doi: 10.15837/ijccc.2018.4.3269.
[29] T. Amarbayasgalan, J. Y. Lee, K. R. Kim, and K. H. Ryu, "Deep autoencoder based neural networks for coronary heart disease risk prediction," in *Proc. VLDB DMAH*, Los Angeles, CA, USA, Aug. 2019, pp. 237–248.
[30] J. K. Kim and S. Kang, "Neural network-based coronary heart disease risk prediction using feature correlation analysis," *J. Healthcare Eng.*, vol. 2017, pp. 1–13, Sep. 2017, doi: 10.1155/2017/2780501.
[31] J. Kim, U. Kang, and Y. Lee, "Statistics and deep belief network-based cardiovascular risk prediction," *Healthcare Inform. Res.*, vol. 23, no. 3, pp. 169–175, 2017, doi: 10.4258/hir.2017.23.3.169.
[32] D. P Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*. [Online]. Available: http://arxiv.org/abs/1312.6114
[33] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bull. Math. Biophys.*, vol. 5, no. 4, pp. 115–133, 1943.
[34] A. L. Maas, P. Qi, Z. Xie, A. Y. Hannun, C. T. Lengerich, D. Jurafsky, and A. Y. Ng, "Building DNN acoustic models for large vocabulary speech recognition," *Comput. Speech Lang.*, vol. 41, pp. 195–213, Jan. 2017.
[35] E. Batbaatar, M. Li, and K. H. Ryu, "Semantic-emotion neural network for emotion recognition from text," *IEEE Access*, vol. 7, pp. 111866–111878, 2019.
[36] A. Das, U. R. Acharya, S. S. Panda, and S. Sabut, "Deep learning based liver cancer detection using watershed transform and Gaussian mixture model techniques," *Cognit. Syst. Res.*, vol. 54, pp. 165–175, May 2019.
[37] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.
[38] S. Kweon, Y. Kim, M.-J. Jang, Y. Kim, K. Kim, S. Choi, C. Chun, Y.-H. Khang, and K. Oh, "Data resource profile: The Korea national health and nutrition examination survey (KNHANES)," *Int. J. Epidemiol.*, vol. 43, no. 1, pp. 69–77, Feb. 2014.
[39] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.
[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[41] L. Prechelt, "Early stopping-but when," in *Neural Networks: Tricks of the Trade*, vol. 1254. Berlin, Germany: Springer, 1998, pp. 55–69.

TSATSRAL AMARBAYASGALAN received the B.Sc. and M.Sc. degrees from the National University of Mongolia, Mongolia, in 2010 and 2015, respectively, and the Ph.D. degree in computer science from Chungbuk National University, South Korea, in 2021. Her research interests include machine learning, artificial intelligence, deep learning, big data, and healthcare analytics.

VAN-HUY PHAM received the M.S. degree in computer science from the University of Science, Ho Chi Minh City, Vietnam, in 2007, and the Ph.D. degree in computer science from the University of Ulsan, South Korea, in 2015. Since 2015, he has been a Lecturer and a Researcher with the Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City. His main research interests include artificial intelligence, image processing, computer vision, and deep learning applications.

NIPON THEERA-UMPON (Senior Member, IEEE) received the B.Eng. degree (Hons.) from Chiang Mai University, the M.S. degree from the University of Southern California, and the Ph.D. degree from the University of Missouri, Columbia, all in electrical engineering. He has been with the Department of Electrical Engineering, Chiang Mai University, since 1993, where he is currently serving as the Director for the Biomedical Engineering Institute. He was the Associate Dean of engineering and the Chairman of graduate study in electrical engineering and graduate study in biomedical engineering. He has published more than 200 full research articles in international refereed publications. His research interests include pattern recognition, machine learning, artificial intelligence, digital image processing, neural networks, fuzzy sets and systems, big data analysis, data mining, medical signal, and image processing. He is a member of the IEEE-IES Technical Committee on Human Factors, Thai Robotics Society, the Biomedical Engineering Society of Thailand, and the Council of Engineers in Thailand. He has been bestowed several royal decorations and won several awards. He has served as an editor, a reviewer, the general chair, the technical chair, and a committee member for several journals and conferences. He has served as the Vice President for Thai Engineering in Medicine and Biology Society, and Korea Convergence Society.



YONGJUN PIAO received the Ph.D. degree in computer science from Chungbuk National University, South Korea, in 2017. He is currently an Assistant Professor with the School of Medicine, Nankai University, Tianjin, China, and a Principle Investigator with Tianjin Central Hospital of Gynecology Obstetrics, Tianjin. His research contributions focused on developing data mining and computational approaches for various biological datasets, including next generation sequencing data (RNA-seq, bisulfite-seq, and NOMe-seq), microarray, and other high-dimensional datasets.

KEUN HO RYU (Life Member, IEEE) received the Ph.D. degree in computer science and engineering from Yonsei University, South Korea, in 1988. He has served at the Reserve Officers' Training Corps (ROTC) of Korean Army. He was with The University of Arizona, Tucson, AZ, USA, as a Postdoctoral Researcher and a Research Scientist, and the Electronics and Telecommunications Research Institute, South Korea, as a Senior Researcher. He is currently a Professor with the Faculty of Information Technology, Ton Duc Thang University, Vietnam; the Emeritus and Endowed Chair Researcher with Chungbuk National University, South Korea; and an Adjunct Professor with Chiang Mai University, Thailand. He is also an Honorary Doctorate of the National University of Mongolia. He has been not only the Director of the Database and Bioinformatics Laboratory, South Korea, since 1986, but also the Co-Director of the Research Group, Data Science Laboratory, Ton Duc Thang University, since March 2019. He is also the former Vice-President of the Personalized Tumor Engineering Research Center. He has published more than 1000 refereed technical papers in various journals and international conferences, in addition to authoring a number of books. His research interests include databases, spatiotemporal databases, big data analysis, data mining, deep learning, biomedical informatics, and bioinformatics. He has been a member of IEEE since 1982 and ACM since 1983. He has served on numerous program committees, including roles as the Demonstration Co-Chair for VLDB, the Panel and Tutorial Co-Chair for APWeb, and the FITAT General Co-Chair.