Electronic Research Archive

https://www.aimspress.com/journal/ERA

ERA, 32(5): 3202–3229.
DOI: 10.3934/era.2024147
Received: 31 October 2023
Revised: 08 March 2024
Accepted: 10 April 2024
Published: 07 May 2024

Research article

Deep autoencoders and multivariate analysis for enhanced hypertension detection during the COVID-19 era

Khongorzul Dashdondov[1], **Mi-Hye Kim**[2] **and Mi-Hwa Song**[3,*]

[1] Department of Computer Engineering, College of IT Convergence, Gachon University, Seongnam 13120, South Korea

[2] Department of Computer Engineering, Chungbuk National University, Cheongju 28644, South Korea

[3] Division of Computer Engineering, College of IT Engineering, Hansung University, Seoul 02876, South Korea

* **Correspondence:** Email: mhsong@hansung.kr.

Abstract: The incidence of hypertension has increased dramatically in both elderly and young populations. The incidence of hypertension also increased with the outbreak of the COVID-19 pandemic. To enhance hypertension detection accuracy, we proposed a multivariate outlier removal method based on the deep autoencoder (DAE) technique. The method was applied to the Korean National Health and Nutrition Examination Survey (KNHANES) database. Several studies have identified various risk factors for chronic hypertension. Chronic diseases are often multifactorial rather than isolated and have been associated with COVID-19. Therefore, it is necessary to study disease detection by considering complex factors. This study was divided into two main parts. The first module, data preprocessing, integrated external features for COVID-19 patients merged by region, age, and gender for the KHNANE-2020 and Kaggle datasets. We then performed multicollinearity (MC)-based feature selection for the KNHANES and integrated datasets. Notably, our MC analysis revealed that the "COVID-19 statement" feature, with a variance inflation factor (VIF) of 1.023 and a p-value < 0.01, is significant in predicting hypertension, underscoring the interrelation between COVID-19 and hypertension risk. The next module used a predictive analysis step to detect and predict hypertension based on an ordinal encoder (OE) transformation and multivariate outlier removal using a DAE from the KNHANES data. We compared each classification model's accuracy, F1 score, and area under the curve (AUC). The experimental results showed that the proposed XGBoost model achieved the best results, with an accuracy rate of 87.78% (86.49%–88.1%, 95% CI), an F1 score of 89.95%, and an

AUC of 92.28% for the COVID-19 cases, and an accuracy rate of 87.72% (85.86%–89.69%, 95% CI), an F1 score of 89.94%, and an AUC of 92.23% for the non-COVID-19 cases with the DAE_OE model. We improved the prediction performance of the classifiers used in all experiments by developing a high-quality training dataset implementing the DAE and OE in our method. Moreover, we experimentally demonstrated how the steps of the proposed method improved performance. Our approach has potential applications beyond hypertension detection, including other diseases such as stroke and cardiovascular disease.

Keywords: COVID-19; KNHANES; hypertension; deep autoencoder; outlier; machine-learning

1. Introduction

Hypertension is a chronic disease that can lead to serious conditions in the heart, brain, or kidney [1,2]. Hypertension, which occurs due to high blood pressure, is a condition in which blood vessels have constantly increased pressure, making it difficult for the human heart to pump blood. It is a major cause of early death worldwide, affecting up to one in four men and one in five women, and occurring in over a billion people worldwide [3]. In 2019, the coronavirus disease (COVID-19) threatened the world. Many studies have been conducted on COVID-19, and the use of machine learning methods to diagnose the causes of hypertension has increased in recent years. In this study, we aim to infer the association between COVID-19 and hypertension using machine learning (ML) methods, to identify hypertension based on the characteristics of COVID-19. Machine learning is a process that starts with observations or data, such as cases, real-world experiences, or instructions, to look for patterns in the data and make better decisions based on the examples that we supply. ML helps to make decisions automatically using models learned from data without subjectivity. These can also be used to diagnose various diseases. In a previous study [4], we proposed a multivariate outlier detection Mahalanobis-distance-based XGBoost model to predict hypertension complications. The accuracy, F1 score, and area under the curve (AUC) for the Korean National Health and Nutrition Examination Survey (KNHANES) dataset were 99.51%, 99.58%, and 99.65%, respectively. Liao et al. [5] compared several algorithms on electronic medical record (EMR) datasets to determine the main cause of hypertension and hyperlipidemia, effectively enhancing the interpretability of the model. Based on the KNHANES dataset [6], it is recommended that those who are susceptible to the pandemic pay special attention to COVID-19 protection and maintain suitable nutritional status to promote outstanding immune function. Kim et al. [7] studied how Korean adults perceived their health, happiness, and life satisfaction with different forms of relaxation activities during the COVID-19 pandemic. The authors of [8] studied changes in physical activity and energy consumption in Korean adult males before and after COVID-19 in relation to abdominal obesity. The increase in hypertension among children and adolescents in Korea during the COVID-19 outbreak was investigated using data from the KNHANES 2018–2020 [9,10]. Nguyen et al. [11] examined the association between berry rice consumption and cardiovascular disease, type 2 diabetes, arthritis, and depression among 18-year-old adolescents using NHANES data. In [12–14], the authors suggested that early diagnosis, monitoring, and treatment for high-risk COVID-19 could be identified using hierarchical clustering-based outlier detection with a deep-stacked autoencoder model. Some researchers have employed stacked autoencoder–based feature selection techniques for data mining and pattern detection in the education sector [15] to enhance data quality.

In this study, we propose a multivariate outlier removal–based hypertension prediction method using a deep autoencoder neural network. The proposed method improves the performance of machine learning–based classifiers by performing deep autoencoder (DAE) and ordinal encoder (OE) transformations on the training dataset. We evaluated the proposed method using two datasets from 2015 to 2019 and 2020. Given that COVID-19 began in 2019, Dataset II was created by combining 2020 health data with different COVID-19 features.

Contributions of this research include:

- Statistical and machine learning methods were used to determine whether the characteristics of COVID-19 influence hypertension.
- The performance of machine learning algorithms was improved by preparing training datasets using multivariate outlier removal based on DAE and OE transformations.
- The proposed method was evaluated by conducting experiments on an open dataset.
- The KNHANES dataset and COVID-19 features from Kaggle were integrated using feature matching.

The remainder of this study is structured as follows: Section 2 reviews the recent literature, encompassing a survey of related works in the field. In Section 3, we delineate the proposed methodology. Section 4 is dedicated to the experimental study, which includes an in-depth exploration of the dataset, the procedures employed for comparative analysis, the evaluation metrics utilized, and a comprehensive presentation of the comparative results. Finally, Section 5 concludes the research and offers a detailed examination of correlation, multicollinearity, and descriptive analyses pertinent to Datasets I and II.

2. Related works

Recent studies on hypertension have contributed to the field's evolving understanding and methodological advances. Fang et al. examined hypertension prevalence and management challenges, particularly in lower-income countries, and introduced a predictive model combining KNN and LightGBM based on individualized risk factors [16]. Conversely, [17] applied a novel approach utilizing machine learning techniques, including XGBoost and ensemble methods, juxtaposed with traditional logistic regression, using Japanese health checkup data to forecast the onset of hypertension. AlKaabi et al. conducted a cross-sectional analysis in Qatar, employing a spectrum of machine learning algorithms—decision tree, random forest, and logistic regression—rigorously evaluated against multiple performance metrics for hypertension prediction [18].

In [19], the focus shifted to leveraging artificial neural networks, aiming to estimate hypertension risk based on demographic and health-related variables, thus highlighting their potential in health management and patient categorization. Further, Zhang et al. delved into applying retinal fundus photography combined with neural network models for hypertension prediction in rural Chinese settings, showcasing the method's efficacy in detecting chronic disease [20]. In [21], deep learning algorithms analyzing ECG data were explored, providing insights into the early detection and subtype differentiation of pulmonary hypertension.

Regarding congenital heart disease, [22] ventured into detecting pulmonary hypertension using phonocardiogram recordings, employing a computer-aided diagnosis system that integrated diverse features and utilized an XGBoost classifier. Finally, [23] critically evaluated noninvasive testing methods, including the innovative use of the VITRO diagnostic tool, for detecting clinically significant portal hypertension in patients with chronic liver diseases, offering a nuanced perspective on diagnostic approaches. These studies collectively illuminate the diverse and complex research on hypertension, from predictive modeling and machine learning applications to noninvasive diagnostic methodologies, enriching the tapestry of contemporary medical research.

3. Methodology

In this section, the components of the proposed prediction method are described. Figure 1 shows the proposed framework based on a DAE-based outlier removal method. The proposed framework consists of two main modules: data preprocessing and predictive analysis.

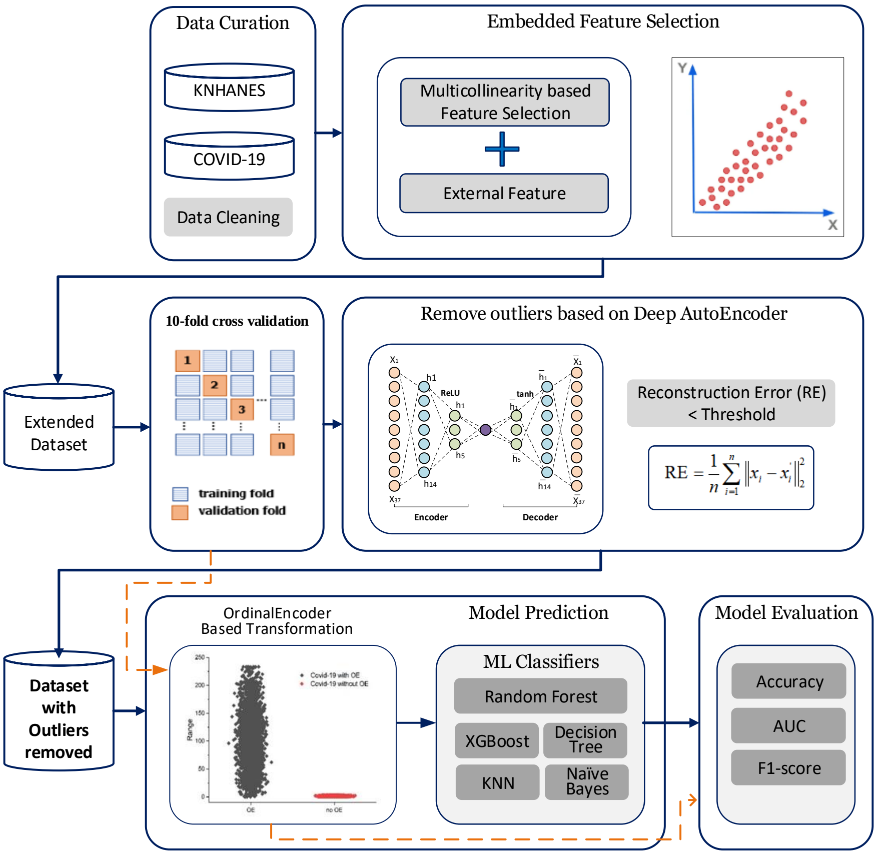


Figure 1. Deep autoencoder-based hypertension classification framework with multivariate outlier detection.

3.1. Data preprocessing

The data preprocessing module consists of three main parts: data cleaning, feature selection, and feature embedding. This study aims to predict hypertension using the KNHANES dataset. This dataset contains a health survey of a variety of diseases, health issues, and nutritional information on the Korean population with 1193 attributes.

3.1.1. Dataset preparation

The KNHANES dataset was used to build a model to predict hypertension. The KNHANES data were collected from the Korean Disease Control and Prevention (KDCA) [24]. It comprises health examinations for various diseases, health interviews, and nutritional surveys of the Korean population. We analyzed two datasets: the first dataset from 2015 to 2019 and the second from 2020 for COVID-19 [25]. We generated a target value for > 19-year-old patients with hypertension. The target hypertension group included participants with a history of diabetes, pre-diabetes, heart disease, heart attack, or stroke. There were two types of KNHANES datasets: Datasets I and II. The KNHANES datasets were released for public use within one year from the end of each survey year. Figure 2 illustrates the procedure used to create the target dataset. As the number of datasets increased, the results of the machine learning algorithm improved, as shown in the experimental results.

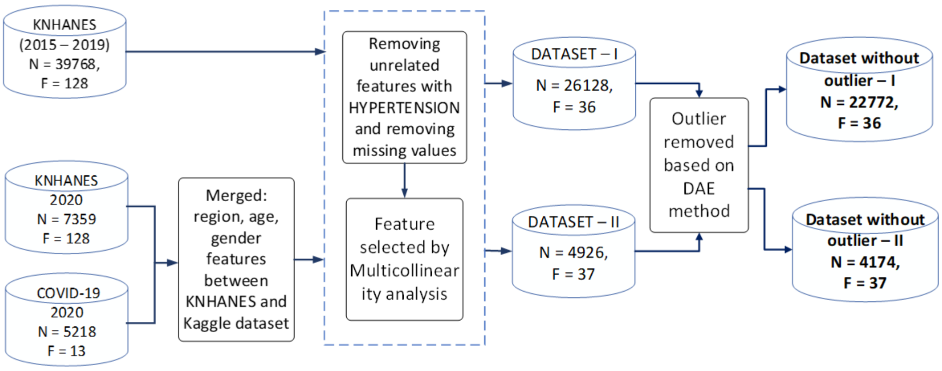

KNHANES
(2015 – 2019)
N = 39768,
F = 128
KNHANES
2020
N = 7359
F = 128
COVID-19
2020
N = 5218
F = 13
Merged: region, age, gender features between KNHANES and Kaggle dataset
Removing unrelated features with HYPERTENSION and removing missing values
Feature selected by Multicollinearity analysis
DATASET – I
N = 26128,
F = 36
DATASET – II
N = 4926,
F = 37
Outlier removed based on DAE method
Dataset without outlier – I
N = 22772,
F = 36
Dataset without outlier – II
N = 4174,
F = 37

Figure 2. Experimental datasets based on the KNHANES and Kaggle.

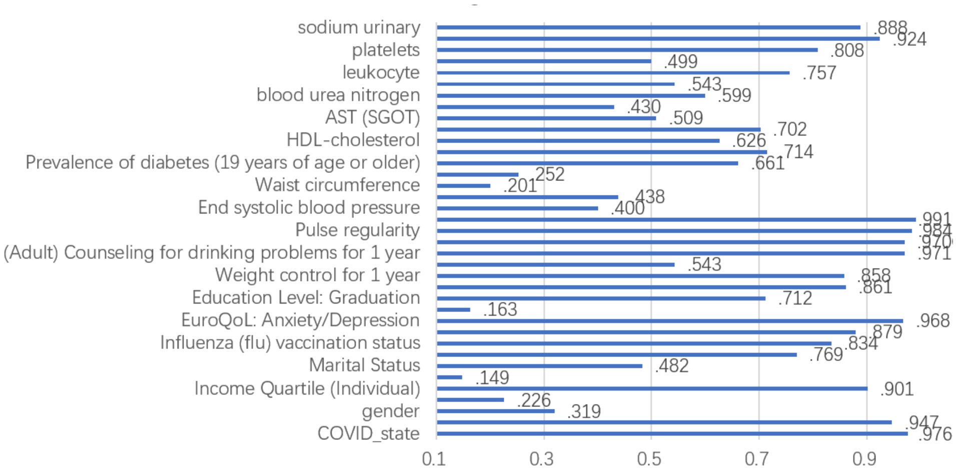


Figure 3. Tolerance value of the multicollinearity analysis of the selected features on experimental datasets.

The 2015−2019 dataset consisted of 39,758 records with 1192 columns. From this data, after removing features not associated with hypertension, we began data analysis in Dataset I, initially with 39,758 records and 128 attributions. Subsequently, we selected features using multicollinearity analysis based on the attribution significance dataset after removing several missing values and features unrelated to hypertension. There were 26,128 records and 36 related features with hypertension used in our experimental target Dataset I. We removed outliers based on DAE and then included 22,772 records and 36 features for Dataset I for the experiment.

Next, we used an integrated dataset called Dataset II, which was created by combining the 2020

COVID-19 datasets with the KNHANES 2020 survey data using a feature alignment approach. The categorization of COVID-19 patient statuses is essential for analyzing the progression and impact of the disease within a population. The data encompasses 11,468 confirmed cases between January 20 and May 31, 2020, compiled by the Korean Centers for Disease Control and Prevention. Most patients were released from isolation, followed by a smaller percentage in isolation and a small fraction resulting in fatalities [2]. Data from 5218 non-overlapping individuals were used. In other words, dead people were excluded from the data at the confirmed date.

From this data, after removing features not associated with hypertension, we began data analysis in Dataset II, initially with 7359 records and 128 attributions for 2020 KNHANES and 5218 records and 13 attributions for 2020 COVID-19. We removed missing values and several features unrelated to hypertension using MC analysis, including 4926 records and 37 features for Dataset II. Then, outliers were removed based on the DAE method, leaving 4174 records and 37 features of Dataset II to continue testing.

In COVID-19, statements are as follows:

1) Isolated: This designation is assigned to individuals diagnosed with COVID-19 who are undergoing isolation. The isolation setting could be a hospital, a care facility, or a home, determined by the patient's symptom severity and the prevailing healthcare guidelines in the region.

2) Released: Patients categorized as "released" have recovered from COVID-19 and are no longer infectious. Transitioning to this category typically requires the cessation of symptoms and a negative virus test result, indicating successful recovery from the infection.

3) Deceased: This category accounts for those who have lost their lives due to the virus, and it represents the mortality rate associated with COVID-19.

3.1.2. External feature

The daily frequency of COVID-19 cases in South Korea between January 20, 2020, and June 30, 2020, was reported by Kaggle [26,27]. The values of the COVID-19 features ranged from 1–3, indicating isolated, released, or deceased patients. By merging the COVID-19 features with the 2020 KNHANES dataset, we matched the location, age, and sex of the Kaggle and KNHANES datasets. This dataset is called Dataset II. The tolerance value of Dataset II is shown in Figure 3. Columns 1–36 of Dataset I and II are identical and distinct from the column COVID-19. The target feature descriptions, mean, and standard deviations (Std. Dev) for each dataset (I and II) are listed in Table 1.

3.1.3. Feature selection based on multicollinearity analysis

The feature selection module was performed using multicollinearity analysis (MC). Therefore, we selected important features to develop a simple and accurate model. We verified the collinearity between selected features of health examination, nutrition, basic information, and hypertension using MC in the regression analysis. MC is a statistical term used when the values of two or more input attributes are extremely correlated [28]. If highly correlated variables are present, the attributes should be removed. The tolerance results and variance inflation factor (VIF) were examined using the MC analysis. If the VIF value is greater than 10 and the tolerance is less than 0.10, then an MC problem occurs. At the end of this analysis, 36 features were used as inputs for subsequent analysis of the two sets of data. Figure 3 shows the tolerance values for the selected features. In our study, we conducted an MC analysis on two distinct datasets to draw comparative insights. The results, including p-values, tolerance, and VIF scores are detailed in Tables 2 and 3. Notably, the highest VIF scores were observed for three variables: “health checkup status in adults” (4.435), “influenza vaccination status” (4.118), and “number of walking days per week” (4.016). This last finding was particularly unexpected, as frequent walking is often linked with hypertension management, as illustrated in Table 2. Other significant predictors, such as “waist circumference” and “age” showed VIF scores of 3.233 and 3.167, respectively. The VIF range of 1–5 for these predictors suggests a lack of correlation, affirming their suitability for inclusion in the hypertension prediction model.

Table 1. Feature descriptions of Dataset I and II.

Features	Descriptions	Dataset I		Dataset II	
		Mean	Std. Dev.	Mean	Std. Dev.
region	Region	7.39	4.940	7.61	4.989
sex	Gender	1.55	.497	1.54	0.498
age	Age	51.41	16.526	51.21	16.847
incm	Income quartile (individual)	2.52	1.115	2.53	1.112
marri_1	Marital status	1.17	.374	1.21	0.405
Graduate	Education level	1.75	1.773	1.57	1.295
DI2_ag	When to diagnose dyslipidemia	735.94	321.547	705.23	344.66
BH9_11	Influenza (flu) vaccination status	1.70	1.113	1.53	0.510
BH1	Health check-up status	1.45	1.127	1.29	0.466
LQ_5EQL	EuroQoL: anxiety/depression	1.25	1.109	1.10	0.316
EC_wht_0	Employment (full-time or not)	5.27	3.234	5.17	3.237
BO2_1	Weight control for 1 year	2.38	1.395	2.26	1.292
BD2_32	Frequency of female binge drinking	5.85	2.978	5.95	2.920
BD7_5	Counseling for drinking problems for 1 year	2.15	.985	2.00	.059
BE3_31	Number of walking days per week	6.77	13.685	4.90	3.490
HE_rPLS	Pulse regularity	1.01	.109	1.01	.098
HE_nARM	Blood pressure measuring arm	1.01	.101	1.00	.065
HE_sbp	End systolic blood pressure	119.22	16.721	119.08	16.209
HE_dbp	Final diastolic blood pressure	75.55	10.011	75.94	9.781
HE_wc	Waist circumference	82.84	10.174	84.672	10.630
HE_obe_BMI	Prevalence of obesity	2.71	.934	3.04	1.065
HE_DM	Prevalence of diabetes	1.54	.711	1.73	.718
HE_chol	Total cholesterol	192.28	37.427	189.90	38.721
HE_HDL_st2	HDL cholesterol	51.35	12.776	51.56	12.523
HE_TG	Triglycerides	134.51	107.119	130.57	107.287
HE_ast	GOT	23.51	13.878	25.06	16.744
HE_alt	GPT	22.50	18.440	23.87	19.873
HE_BUN	Blood urea nitrogen	14.92	4.759	15.05	4.833
HE_crea	Blood creatinine	1.00	.209	.808	.241
HE_WBC	Leukocyte	6.25	1.785	6.178	1.699
HE_RBC	Red blood cells	4.60	.587	4.545	.489
HE_Bplt	Platelets	258.36	64.117	254.07	59.729
HE_Uph	Uric acid	5.97	.881	5.850	.7414
HE_UNa	Sodium urinary	113.49	48.130	111.28	47.776
COVID_state	COVID-19 statement	-	-	1.44	0.528

Table 2. Multivariate logistic regression analysis results of risk for hypertension of KNHANES 2015−2019, Dataset I.

Features	p-value	Tolerance	VIF
Health check-up status	0.284	0.225	4.435
Influenza (flu) vaccination status	**< 0.0001**	0.243	4.118
Number of walking days per week	0.815	0.249	4.016
Waist circumference	**< 0.0001**	0.309	3.233
Age	**< 0.0001**	0.316	3.167
Prevalence of obesity	**< 0.0001**	0.374	2.674
ALT(SGPT)	0.694	0.380	2.630
Gender	0.067	0.395	2.534
End systolic blood pressure	**< 0.0001**	0.425	2.355
AST (SGOT)	0.116	0.426	2.345
Final diastolic blood pressure	**< .0001**	0.483	2.068
Frequency of (adult) female binge drinking	0.118	0.518	1.929
Marital status	**< 0.0001**	0.577	1.734
Education level: graduation	**< 0.0001**	0.608	1.645
Red blood cells	**< 0.0001**	0.631	1.584
HDL cholesterol	**< 0.0001**	0.642	1.559
Triglycerides	**< 0.0001**	0.665	1.503
Blood urea nitrogen	**< 0.0001**	0.701	1.426
Prevalence of diabetes (including glycated hemoglobin)	**< 0.0001**	0.727	1.376
Total cholesterol	**< 0.0001**	0.735	1.361
Survey year	**< 0.0001**	0.762	1.312
Leukocyte	**< 0.001**	0.792	1.262
Employment (full-time or not)	.323	0.828	1.208
Platelets	**< 0.0001**	0.829	1.206
When to diagnose dyslipidemia	**< 0.0001**	0.831	1.204
Weight control for 1 year	0.221	0.833	1.201
Counseling for drinking problems for 1 year	0.714	0.858	1.165
Blood creatinine	0.613	0.864	1.157
Uric acid	0.034	0.926	1.080
Sodium urinary	**< 0.0001**	0.929	1.076
Income quartile (individual)	0.626	0.955	1.047
Region	0.561	0.966	1.035
Pulse regularity	**< 0.012**	0.981	1.020
Blood pressure measuring arm	0.664	0.993	1.007

Table 3 lists the results of the multivariate logistic regression analysis for the risk of hypertension according to COVID-19 with an adjustment for significant variables, VIF, and tolerance. The highest VIF was waist circumference (4.978), which is surprising because it is one of the most common symptoms of hypertension. Predictors such as the prevalence of obesity, age, and sex had the next highest VIF with scores of 3.968, 3.775, and 3.074, respectively. Additionally, with the COVID-19 status VIF at 1.023, the p-value was < 0.011, indicating that the predictors were not correlated and could be considered when building the hypertension prediction model.

Table 3. Multivariate logistic regression analysis results of risk for hypertension of integrated Dataset II of KNHANES 2020 and COVID-19.

Features	p-value	Tolerance	VIF
Waist circumference	**< 0.010**	0.201	4.978
Prevalence of obesity (19 years of age or older)	**< 0.008**	0.252	3.968
Age	**< 0.0001**	0.265	3.775
Gender	**< 0.018**	0.325	3.074
End systolic blood pressure	**< 0.0001**	0.402	2.488
ALT(SGPT)	0.297	0.430	2.324
Final diastolic blood pressure	**< 0.0001**	0.439	2.276
Marital status	**< 0.0001**	0.490	2.042
Red blood cells	**< 0.0001**	0.499	2.004
AST (SGOT)	0.751	0.510	1.961
Frequency of (adult) female binge drinking	0.902	0.544	1.838
Blood creatinine	**< 0.050**	0.545	1.834
Blood urea nitrogen	**< 0.012**	0.601	1.664
HDL cholesterol	0.268	0.626	1.598
Prevalence of diabetes (including glycated hemoglobin)	**< 0.0001**	0.661	1.512
Triglycerides	0.161	0.703	1.423
Total cholesterol	**< 0.0001**	0.719	1.391
Leukocyte	0.126	0.758	1.320
When to diagnose dyslipidemia	**< 0.0001**	0.771	1.296
Platelets	**< 0.042**	0.809	1.236
Influenza (flu) vaccination status	0.362	0.839	1.191
Employment (full-time or not)	0.851	0.862	1.161
Weight control for 1 year	0.168	0.863	1.159
Health check-up status	0.063	0.882	1.134
Sodium urinary	0.097	0.890	1.124
Education level: graduation	**< 0.0001**	0.910	1.099
Uric acid	0.557	0.925	1.082
Income quartile (individual)	0.576	0.940	1.064
Region	0.544	0.952	1.051
EuroQoL: anxiety/depression	0.324	0.968	1.033
Counseling for drinking problems for 1 year	0.927	0.973	1.028
Number of walking days per week	0.864	0.974	1.027
COVID-19 statement	**< 0.011**	0.978	1.023
Pulse regularity	0.411	0.984	1.016
Blood pressure measuring arm	0.751	0.991	1.009

3.2. Predictive analysis

3.2.1. Outlier detection based on DAE

The DAE was used to clean the data. The AE is an unsupervised artificial neural network that learns how to efficiently compress and encode data and then reconstructs the data from the reduced encoded representation to a representation as close to the original input as possible [29]. The AE structure consists of encoder and decoder components. The encoder compresses the input data by reducing the data dimensions, whereas the decoder constructs the compressed data as the output. The reconstruction error of the autoencoder is the difference between the input and reconstructed outputs [30,31].

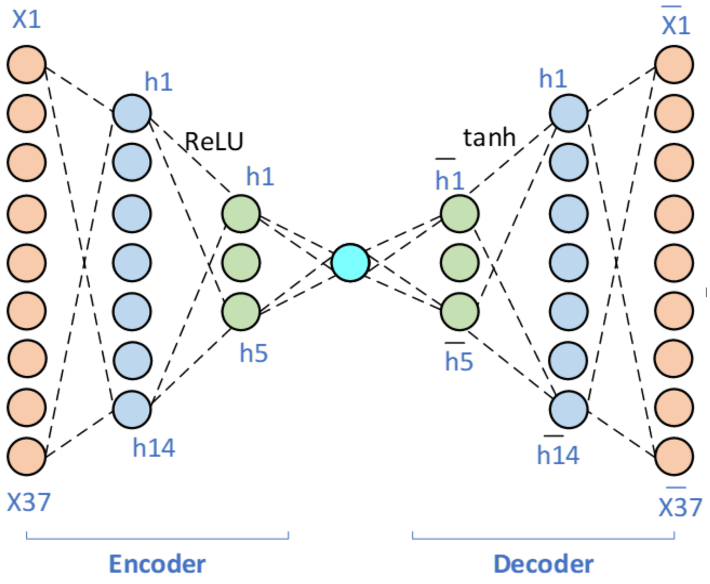


Figure 4. Architecture of DAE neural network used in the proposed method.

Figure 4 shows the structure of the proposed autoencoder model. The proposed autoencoder has six hidden layers with 37, 20, 10, 5, 1, 5, 10, 20, and 37 nodes. Moreover, the hidden layers in the encoder part use the "ReLU" activation function, and hidden layers in the decoder part use the "tanh" activation function. First, we trained the DAE model using the entire dataset. We then calculated the reconstruction errors (RE) using the mean of the squared difference between the input and output as described in the following equation [32]:

$$RE = \frac{1}{n}\sum_{i=1}^{n}\|x_i - x_i'\|_2^2 \tag{1}$$

where n is the number of records, x is the original input, and x is the reconstructed input. First, the RE of the training dataset was calculated using the DAE model. The mean and standard deviation of these REs were then used to estimate the threshold for splitting the training dataset, which can be described as follows [32].

$$Threshold = \frac{1}{n}\sum_{i=1}^{n} RE_i + \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[RE_i - \frac{1}{n}\sum_{i=1}^{n} RE_i\right]}, \tag{2}$$

where k is the number of instances in the training dataset, and RE_i is the reconstruction error of the i-th training instance. Consequently, two different training datasets were prepared, and the RE-based threshold was estimated for further analysis. Subsequently, a threshold was used to select an appropriate hypertension-prediction model from the DAE models trained on the two datasets.

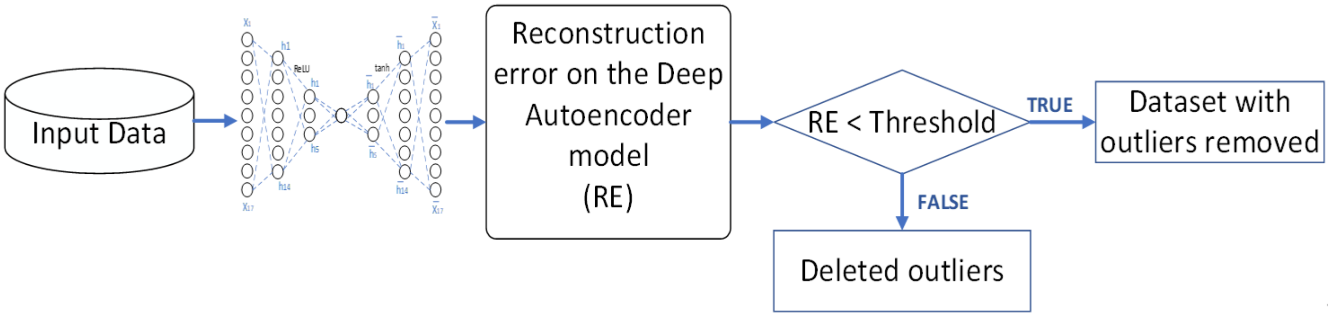

Input Data
ReLU
tanh
Reconstruction error on the Deep Autoencoder model (RE)
RE < Threshold
TRUE
Dataset with outliers removed
FALSE
Deleted outliers

Figure 5. Process of outlier detection using the DAE.

Figure 5 shows the outlier detection steps. In this process, the first input data was provided to the DAE model, and its RE on DAE was calculated. If the RE exceeded the threshold estimated using Eq (2), the outlier threshold value was estimated by summing the average reconstruction error and the standard deviation. If the reconstruction error of the data exceeded the threshold value, it was excluded from the dataset. For the DAE model, the learning rate was configured to minimize the mean squared error to 0.001, and the Adamax optimizer was employed [30]. The batch size was set to 32, and the number of epochs was specified to 1000. The performance of the DAE model was compared under different threshold values for the HP (hypertension) feature in Table 4, which helps to understand how the DAE model performs with varying threshold settings.

Table 4. Comparative results of threshold values for DAE model.

Statistics	Dataset with varying threshold							
	Dataset I				**Dataset II**			
	Original	High	Medium	**Low**	**Original**	High	Medium	**Low**
	HP	0.953	0.816	**0.544**	**HP**	0.414	0.349	**0.237**
	feature				**feature**			
N	26128	25662	24352	**22772**	4925	4364	4214	**4174**
Mean	1.89	1.89	1.88	1.87	1.91	1.93	1.89	1.85
Std. Dev.	0.862	0.862	0.859	0.853	0.852	0.862	0.847	0.839
Std.E.Mean	0.005	0.005	0.005	0.006	0.12	0.013	0.013	0.013
t-statistic	355.13	351.587	342.347	330.00	157.07	148.24	144.92	142.12
p-value	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
95% CI	1.88−1.91	1.88−1.90	1.87−1.89	1.85−1.88	1.88−1.93	1.91−1.96	1.86−1.92	1.82−1.87

Table 4 summarizes statistical analyses conducted on the two separate Datasets I and II, across three thresholds: High, Medium, and Low regarding an "HP' feature. This provides information regarding the mean values, variability, and statistical significance of the model's performance. The statistical measures provided include the sample size (N), mean, standard deviation (Std. Dev.), standard error of the mean (Std.E.Mean), t-statistic, p-value, and 95% confidence interval (95% CI). For Dataset I, the original HP feature value threshold has 0.953 for high, 0.816 for medium, and 0.544 for low. Additionally, for dataset II, the original HP feature values were 0.414 for high, 0.349 for medium, and 0.237 for low. The sample size decreased as the threshold decreased, indicating possible criteria-based selection within the dataset. Despite varying thresholds, the mean values were consistent, suggesting a relatively stable central tendency. The standard deviation was also consistent, indicating that the dispersion of the data around the mean was similar across the thresholds. The small standard error of the mean provides high precision for the mean estimates. The t-statistics were high, and the p-values were extremely low (0.0001). This indicates that the means at different thresholds are statistically significantly different from the hypothesized population mean, assuming the null hypothesis to be zero or another baseline value. The 95% CI are narrow and overlap slightly, which provides a precise estimate of where the population's true mean is expected to have a 95% confidence level.

Based on our analysis, we added the mean and standard deviation, which we named "low", to calculate a threshold value for comparison. We calculated the 75th percentile as "high" and the 50th percentile as "medium" using this threshold value and compared the results in Table 4. Our findings revealed that a lower threshold value of 0.544 and 0.237 led to better outcomes for the selected two datasets.

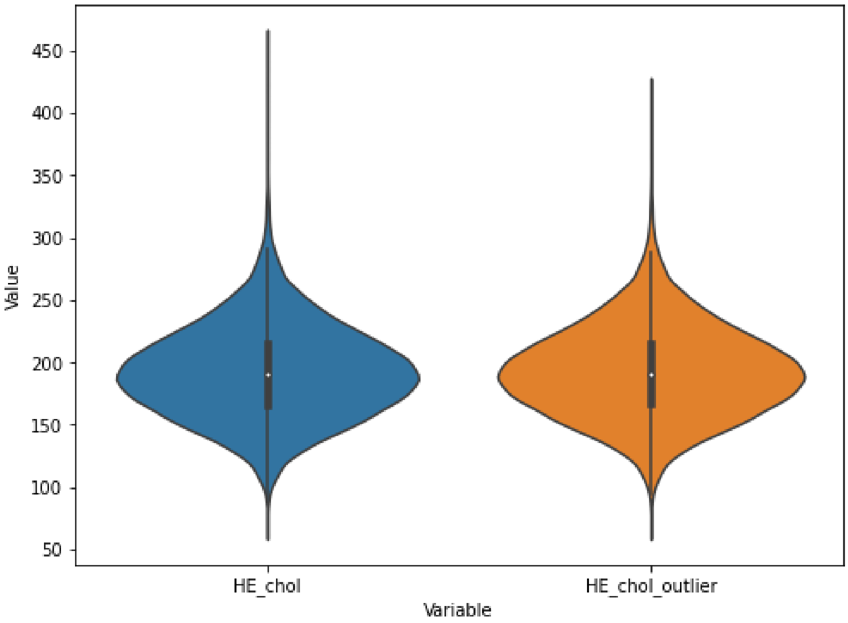


Figure 6. Distribution of COVID-19 features in the original dataset and dataset without outliers.

Figure 6 shows the data with and without outliers from the dataset using several values based on the DAE method. The left violin plot illustrates the distribution of the COVID-19-state feature in the original dataset. The width of the plot at various points indicates the density of data points; more comprehensive sections mean more data points at that value. Based on the DAE method, the right violin plot shows the distribution of the same COVID-19-state feature after removing outliers, including outliers. The absence of the broader sections at the extremes suggests that removing outliers leads to a more concentrated distribution around the median, with fewer extreme values.

3.2.2. Ordinal encoder transformation

The following step in this module involves transforming the data from which outliers have been removed. This is achieved by applying the ordinal encoding (OE) transformation technique. In this process, categorical variables are encoded as integer arrays. The input for this transformation is consistent with either an integer array or a string array, each element of which corresponds to a value determined based on the data's categorical (discrete) attributes. Subsequently, this section is dedicated to converting these features into ordinal integers. This conversion creates a single integer column for each element, where the integers range from 0 to n-1, with *n* representing the total number of categories [32]. Among the 37 features analyzed, 16 categorized features were transferred to OE transformation. These features include Region, Gender, Age, Marital status, Education level, Influenza (flu) vaccination status, EuroQoL: anxiety/depression, Employment (full-time or not), Weight control for one year, Frequency of female binge drinking, Counseling for drinking problems for one year, Number of walking days per week, Prevalence of obesity, Prevalence of diabetes, COVID-19 statement, and Prevalence of hypertension. Figure 7 graphically illustrates the distribution of "marri_1" (marital status), showcasing the data both with and without the application of OE, categorized by the number of classes.

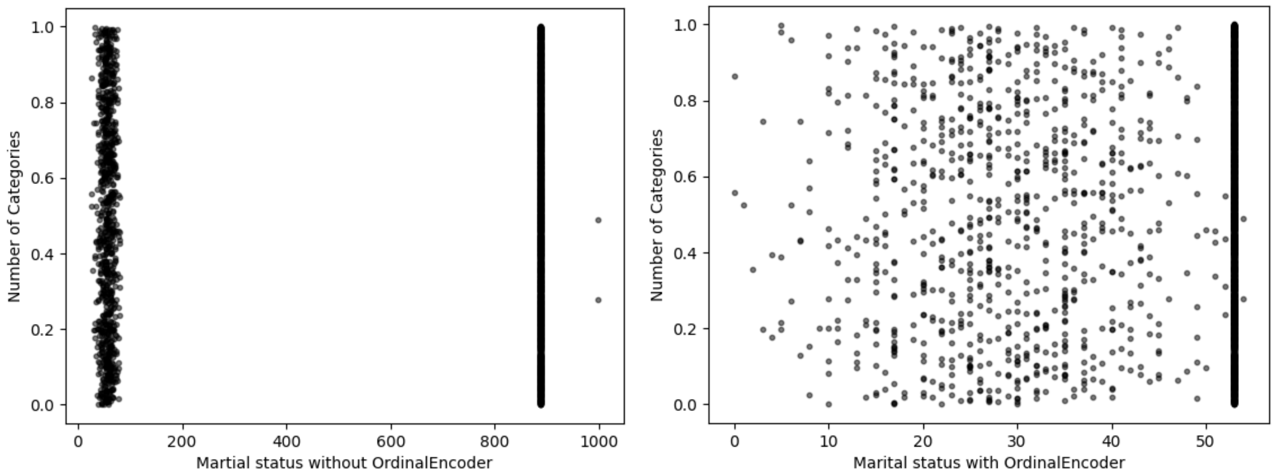


Figure 7. OE-based transformation on the *marital status* feature in the without-outlier dataset.

3.2.3. Classifiers

To improve the performance of the predictive analysis, we focused on a training dataset. That is, instead of directly training the classifiers, outliers were removed from the training dataset using the DAE-OE method. Subsequently, the RF, KNN, XGBoost, DT, and NB algorithms were applied to the prepared training datasets [33,34].

NB: Naïve Bayes is a probability-based classification algorithm. It calculates the probability of each class label and selects the class label with the highest probability. It calculates the probability by considering each feature separately; this is called conditional independence.

KNN: The k-nearest neighbor algorithm was used for classification. First, the user defines the value of parameter k, which is the number of nearest samples used for prediction. Then, all distances between the test data and training dataset are calculated and sorted in descending order. Finally, the top k instances from the ordered dataset are used to predict the class labels. A majority-voted class label is assigned to the output label.

DT: The decision tree classifier is an interpretable label and a commonly used algorithm. It builds a model to predict the target variable using decision rules trained from the data.

RF: The random forest is an ensemble algorithm. It consists of several decision-tree classifiers trained on different subsamples of the entire dataset. For prediction, the majority-voted class label of these decision trees was chosen as the output.

XGBoost: XGBoost uses a method called classification and regression (CART) in which all leaves are related to the final score of a model, unlike a decision-making tree, which only considers the result values of leaf nodes. Although a common decision-making tree is interested in how well the classification has been performed, CART enables a comparison of the superiority of models that retain identical classification results.

3.3. Evaluation metrics

To evaluate the performance of our predictive models, we adopt a comprehensive set of metrics: accuracy, AUC, F1 score, and mean squared error (MSE) [30]. Each metric offers a unique lens to assess the effectiveness of our models, from accuracy in predictions to the balance between precision and recall, providing a holistic view of our model's performance as follows:

$$Precision\ = \frac{TP}{(TP+FP)}\ and\ Recall\ = \frac{TP}{(TP+FN)} \tag{3}$$

Precision and recall are important metrics for evaluating classification models. Precision measures the accuracy of positive predictions, while recall measures the model's ability to identify all positive instances. These metrics are useful for imbalanced datasets or when the cost of false positives and negatives varies.

The F1 score is the harmonic mean of precision and recall as follows:

$$F1 = \frac{2 \cdot Precision \cdot Recall}{(Precision + Recall)} \tag{4}$$

We studied the multiclass case, and the average of the F1 score of each class label with weighting depends on the average parameter, as shown in Eq (4).

Accuracy is a measure of the degree of closeness of the calculated value to its actual value. Accuracy is the sum of the true positive fraction and true negative fraction among all the test data, as shown in Eq (5).

$$Accuracy\ = \frac{TP+TN}{TP+FP+FN+TN} \tag{5}$$

The AUC is a crucial metric for multiple classification models as shown in Eq (6). It is calculated by finding the area under the ROC (receiver operating characteristic) curve. A higher AUC value indicates better model performance, facilitating distinguishing between positive and negative instances. It is beneficial for imbalanced datasets or when the cost of false positives and negatives varies.

$$AUC = \sum_{i=1}^{n} \frac{(FPR_i+FPR_{i+1})\cdot(TPR_{i+1}-TPR_i)}{2} \tag{6}$$

In addition, one of our evaluated metrics is the MSE for the predicted leaks relative to actual values:

$$MSE = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{i=0}^{n-1} [X(i,j) - Y(i,j)]^2 \tag{7}$$

with *m* and *n* being the number of observations, where *m* is the number of data points and *n* is predicting diabetes. *X* and *Y* are the actual and predicted values for the *i*, *j* - th data points, respectively.

4. Experimental study

4.1. Experimental dataset exploration

In our study, we categorized the target characteristics of the experimental dataset into three distinct labels: normal, pre-hypertension, and hypertension. Detailed descriptions of these target features are provided in Table 1, with the descriptive statistics outlined in Tables 4 and 5. In this article, hypertension is defined for patients exhibiting a systolic blood pressure (SBP) greater than 140 mmHg, a diastolic blood pressure (DBP) greater than 90 mmHg, or those under a physician-prescribed antihypertensive drug (AHD) regimen. The criteria for the hypertension classification labels were established as follows: A label of *Normal* was assigned for DBP in the range of 0–80 mmHg and SBP in the range of 0–120 mmHg. A label of *Pre-hypertension* was designated for DBP ranging from 80 to 90 mmHg and SBP from 120 to 140 mmHg. In cases exceeding these ranges, the classification label is *Hypertension*.

Box plots for the relationship between hypertension, COVID-19, age, sex, and survey year are shown in Figure 8(a),(b) for Datasets I and II, respectively. The difference between the datasets is that Dataset II included COVID-19 features. Therefore, we present a feature description based on Dataset II in Table 1. Cross-tabulation analysis was performed on both datasets for comparison. The normal, pre-hypertensive, and hypertensive values for each dataset are listed in Tables 5 and 6.

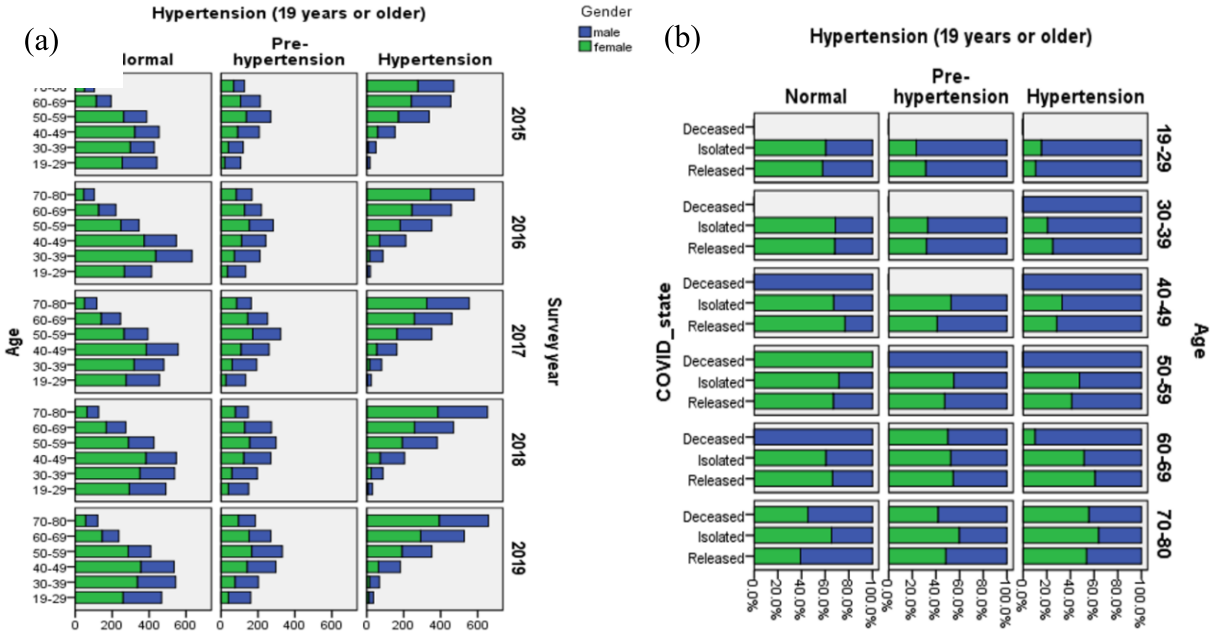


Figure 8. Correlation of categorical features of the experimental datasets. (a) Dataset I; (b) Dataset II.

Table 5. Descriptive statistics for KNHANES (2015−2019).

Features	Hypertension (19 years or older)			
	Normal 10021 (44.0%)	Pre-hypertension 5786 (25.4%)	Hypertension 6965 (30.6%)	Total 22772 (100%)
Gender				
Male	3671 (16.1)	3272 (14.4)	3752 (16.5)	10695 (47.0)
Female	6350 (27.9)	2514 (11.0)	3213 (14.1)	12077 (53.0)
Survey year				
2015	1626 (7.1)	802 (3.5)	960 (4.2)	3388 (14.9)
2016	2027 (8.9)	1151 (5.1)	1483 (6.5)	4661 (20.5)
2017	2107 (9.3)	1258 (5.5)	1436 (6.3)	4801 (21.1)
2018	2233 (9.8)	1262 (5.5)	1605 (7.0)	5100 (22.4)
2019	2028 (8.9)	1313 (5.8)	1481 (6.5)	4822 (21.2)
Age				
19−29 years old	2151 (9.4)	642 (2.8)	77 (0.3)	2870 (12.6)
30−39 years old	2481 (10.9)	876 (3.8)	326 (1.4)	3683 (16.2)
40−49 years old	2433 (10.7)	1197 (5.3)	797 (3.5)	4427 (19.4)
50−59 years old	1684 (7.4)	1393 (6.1)	1528 (6.7)	4605 (20.2)
60−69 years old	919 (4.0)	1084 (4.8)	2017 (8.9)	4020 (17.7)
70−80 years old	353 (1.6)	594 (2.6)	2220 (9.7)	3167 (13.9)
Marital status				
Married	7357 (32.3)	4794 (21.1)	6630 (29.1)	18781 (82.5)
Single	2664 (11.7)	992 (4.4)	335 (1.5)	3991 (17.5)
Influenza (flu) vaccination status				
Yes	3207 (14.1)	2152 (9.5)	4048 (17.8)	9407 (41.3)
No	6693 (29.4)	3539 (15.5)	2825 (12.4)	13057 (57.3)
No response	121 (0.5)	95 (0.4)	92 (0.4)	308 (1.4)
Prevalence of obesity (19 years or older)				
Underweight	601 (2.6)	155 (0.7)	88 (0.4)	844 (3.7)
Normal	5966 (26.2)	2499 (11.0)	2345 (10.3)	10810 (47.5)
Pre-obesity stage	2110 (9.3)	1728 (7.6)	2411 (10.6)	6249 (27.4)
1st stage obesity	1235 (5.4)	1232 (5.4)	1796 (7.9)	4263 (18.7)
2nd stage obesity	101 (0.4)	164 (0.7)	293 (1.3)	558 (2.5)
3rd stage obesity	8 (0.0)	8 (0.0)	32 (0.1)	48 (0.2)
Prevalence of diabetes (19 years of age or older)				
Normal	7863 (34.5)	3435 (15.1)	2580 (11.3)	13878 (60.9)
Pre-diabetes	1922 (8.4)	1905 (8.4)	2797 (12.3)	6624 (29.1)
Diabetes	236 (1.0)	446 (2.0)	1588 (7.0)	2270 (10.0)
Region				
Seoul	2080 (9.1)	1116 (4.9)	1204 (5.3)	4400 (19.3)
Busan	697 (3.1)	335 (1.5)	417 (1.8)	1449 (6.4)
Daegu	516 (2.3)	271 (1.2)	335 (1.5)	1122 (4.9)
Incheon	556 (2.4)	333 (1.5)	419 (1.8)	1308 (5.7)
Gwangju	389 (1.7)	163 (0.7)	186 (0.8)	738 (3.2)

Continued on next page

Daejeon	365 (1.6)	198 (0.9)	237 (1.0)	800 (3.5)
Ulsan	215 (0.9)	114 (0.5)	134 (0.6)	463 (2.0)
Sejong	640 (2.8)	333 (1.5)	384 (1.7)	1357 (6.0)
Gyeonggi	2196 (9.6)	1308 (5.7)	1499 (6.6)	5003 (22.0)
Gangwon	241 (1.1)	209 (0.9)	304 (1.3)	754 (3.3)
Chungbuk	256 (1.1)	179 (0.8)	259 (1.1)	694 (3.0)
Chungnam	289 (1.3)	185 (0.8)	276 (1.2)	750 (3.3)
Jeonbuk	284 (1.2)	177 (0.8)	221 (1.0)	682 (3.0)
Jeonnam	303 (1.3)	203 (0.9)	262 (1.2)	768 (3.4)
Gyeongbuk	419 (1.8)	286 (1.3)	348 (1.5)	1053 (4.6)
Gyeongnam	449 (2.0)	292 (1.3)	368 (1.5)	1109 (4.9)
Jeju	126 (0.6)	84 (0.4)	112 (0.5)	322 (1.4)

Our study analyzed data from the KNHANES conducted between 2015 and 2019, encompassing 22,772 participants. The average age of these participants was 51.41 years old, comprising 10,695 (47%) men and 12,077 (53%) women. Additionally, we included data from the KNHANES 2020 survey, which involved 4174 participants with a mean age of 51.21 years old, including 2007 men (48.1%) and 2167 women (51.9%).

Table 6. Descriptive statistics for KNHANES 2020 year.

Features	Hypertension (19 years or older)			
	Normal	Pre-hypertension	Hypertension	Total
	1840 (44.1%)	1136 (27.2%)	1198 (28.7%)	4174 (100%)
Gender				
Male	883 (21.2)	560 (13.4)	564 (13.5)	2007 (48.1)
Female	957 (22.9)	576 (13.8)	634 (15.2)	2167 (51.9)
Age				
19−29 years old	458 (11.0)	171 (4.1)	23 (0.6)	652 (15.6)
30−39 years old	404 (9.7)	166 (4.0)	54 (1.3)	624 (14.9)
40−49 years old	422 (10.1)	218 (5.2)	138 (3.3)	778 (18.6)
50−59 years old	319 (7.6)	235 (5.6)	256 (6.1)	810 (19.4)
60−69 years old	173 (4.1)	221 (5.3)	375 (9.0)	769 (18.4)
70−80 years old	64 (1.5)	125 (3.0)	352 (8.4)	541 (13.0)
Marital status				
Married	1506 (36.1)	967 (23.2)	965 (23.1)	3438 (82.4)
Single	334 (8.0)	1136 (27.2)	1198 (28.7)	736 (17.6)
Influenza (flu) vaccination status				
Yes	829 (19.9)	553 (13.2)	557 (13.3)	1939 (46.5)
No	1011 (24.2)	583 (14.0)	641 (15.4)	2235 (53.5)
Prevalence of obesity (19 years or older)				
Underweight	138 (3.3)	18 (0.4)	11 (0.3)	167 (4.0)
Normal	924 (22.1)	349 (8.4)	246 (5.9)	1519 (36.4)
Pre-obesity stage	391 (9.4)	282 (6.8)	292 (7.0)	965 (23.1)

Continued on next page

1st stage obesity	351 (8.4)	406 (9.7)	544 (13.0)	1301 (31.2)
2nd stage obesity	35 (0.8)	74 (1.8)	96 (2.3)	205 (4.9)
3rd stage obesity	1 (0.0)	7 (0.2)	9 (0.2)	17 (0.4)
Prevalence of diabetes (19 years of age or older)				
Normal	1210 (29.0)	475 (11.4)	253 (6.1)	1938 (46.4)
Pre-diabetes	578 (13.8)	555 (13.3)	612 (14.7)	1745 (41.8)
Diabetes	52 (1.2)	106 (2.5)	333 (8.0)	491 (11.8)
Region				
Seoul	418 (10.0)	240 (5.7)	208 (5.0)	866 (20.7)
Busan	101 (2.4)	55 (1.3)	58 (1.4)	214 (5.1)
Daegu	72 (1.7)	23 (0.6)	41 (1.0)	136 (3.3)
Incheon	76 (1.8)	50 (1.2)	58 (1.4)	184 (4.4)
Gwangju	67 (1.6)	37 (0.9)	43 (1.0)	147 (3.5)
Daejeon	47 (1.1)	24 (0.6)	49 (1.2)	120 (2.9)
Ulsan	40 (1.0)	28 (0.7)	24 (0.6)	92 (2.2)
Sejong	40 (1.0)	27 (0.6)	22 (0.5)	89 (2.1)
Gyeonggi	518 (12.4)	290 (6.9)	314 (7.5)	1122 (26.9)
Gangwon	65 (1.6)	62 (1.5)	57 (1.4)	184 (4.4)
Chungbuk	57 (1.4)	47 (1.1)	44 (1.1)	148 (3.5)
Chungnam	56 (1.3)	43 (1.0)	29 (0.7)	128 (3.1)
Jeonbuk	43 (1.0)	48 (1.1)	52 (1.2)	143 (3.4)
Jeonnam	47 (1.1)	34 (0.8)	37 (0.9)	118 (2.8)
Gyeongbuk	75 (1.8)	52 (1.2)	51 (1.2)	178 (4.3)
Gyeongnam	90 (2.2)	59 (1.4)	85 (2.0)	234 (5.6)
Jeju	28 (0.7)	17 (0.4)	26 (0.6)	71 (1.7)

4.2. Chi-square test analysis results

The chi-square test is a statistical test that is used to determine if there is any relation between two categorical variables. For example, in our study, we performed a chi-square test to investigate the potential connection between COVID-19 and hypertension, or to determine whether COVID-19 affects hypertension [35]. The test statistic for the chi-square test of independence is denoted by χ^2 and is computed as

$$X^2 = \sum_{i=1}^{R}\sum_{j=1}^{C}\frac{\left(o_{ij}-e_{ij}\right)^2}{e_{ij}} \tag{8}$$

where o_{ij} is the observed cell count in the i-th row and j-th column of the table, and e_{ij} is the expected cell count in the i-th row and j-th column of the table, computed as

$$e_{ij} = \frac{row\ i\ total \cdot column\ j\ total}{grand\ total} \tag{9}$$

The quantity (o_{ij}-e_{ij}) is sometimes referred to as the residual of cell (i, j), denoted as r_{ij}. The calculated X^2 value was compared with the critical value from the X^2 distribution table with degrees of freedom $df = (R-1)(C-1)$ and the chosen confidence level. If the calculated X^2 value was greater than the critical X^2 value, the null hypothesis was rejected. Based on these results, we can state the value of the test statistic as $X^2 = 30.742$, degrees of freedom $df = 4$, and p-value of the test statistic $p < 0.001$.

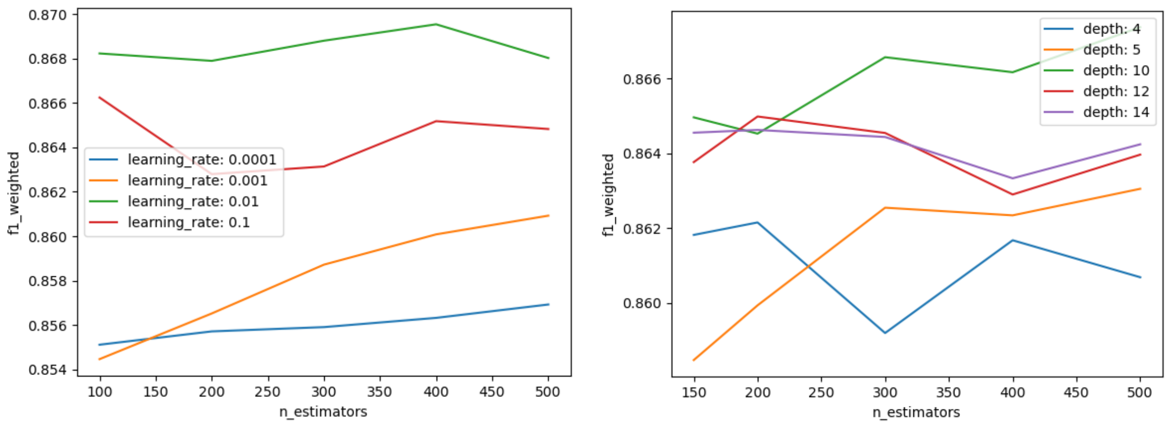
Table 7. Crosstabulations of COVID-19 and hypertension.

		Prevalence of hypertension (19 years or older)			Total
		Normal (%)	Pre-hypertension (%)	Hypertension (%)	
COVID-19 state	Released	1034 (24.8)	631 (15.1)	708 (17.0)	2373 (56.9)
	Isolated	799 (19.1)	504 (12.1)	470 (11.3)	1773 (42.5)
	Deceased	7 (0.2)	1 (0.0)	20 (0.5)	28 (0.7)
Total		1840 (44.1)	1136 (27.2)	1198 (28.7)	4174 (100)

Because the p-value is lower than our chosen significance level of α = 0.05, we reject the null hypothesis and conclude that there is a significant association between hypertension and the COVID-19 statement. The cross-tabulations of COVID-19 and hypertension are shown in Table 7.

4.3. *Hyper-parameter results*

To obtain better results, we tuned some XGBoost hyperparameters on a target dataset using the grid search infrastructure in scikit-learn [34]. Figure 9(a) presents a plot of each learning rate as a series, showing the F1-weighted performance as the number of trees variation. It also shows that the best result observed was a learning rate of 0.01 with 400 trees. Clearly, the expected general trend holds, where the performance improves as the number of trees increases. Figure 9(b) shows the relationship between the number of trees in the model and the depth of each tree. We created a grid of nine different n-estimator values (100–500) and six different maximum depth values (2, 4, 6, 8, 10, and 12), and each combination was evaluated using a 10-fold cross-validation. A total of 9 × 6 × 10 or 540 models were trained and evaluated. The best result was achieved with 500 estimators and a maximum depth of 10 in an F1-weighted score.



(a) (b)

Figure 9. Comparison charts of the datasets. (a) Learning rate and n_estimator, (b) depth and n_estimator.

4.4. Classifier results

The data preprocessing and predictive analysis modules were implemented in Python using the sklearn library [36]. The data preprocessing module was implemented using SPSS 23.0. First, we measured the performances of the baseline models for comparison with the proposed method. We trained the baseline models directly on the raw dataset using ML algorithms shown in Figure 1. We then trained the baseline OE-based models on a dataset with outliers removed. The model was verified using a 10-fold cross-validation, and the results were compared with those of XGB, KNN, DT, RF, and NB. Tables 8 and 9 show the baseline models' and proposed methods' compared performances, where the highest values of evaluation scores are marked in bold. As a result, DAE-based data can improve the performance of models trained on experimental datasets. Moreover, the combination of DAE-based outlier removal and OE-based data transformation in the proposed methods outperformed all compared baselines.

Table 8. Evaluation comparison of the proposed methods on experimental Dataset I.

Methods	Algorithms	Accuracy (%)	ROC (%)	F1 score (%)
Baseline models	XGB	86.13	90.78	87.60
	KNN	63.60	74.67	68.20
	DT	74.02	87.20	83.00
	RF	85.02	90.37	88.75
	NB	57.26	78.59	66.93
Outlier removing method	DAE_XGB	86.49	91.26	88.18
	DAE _KNN	64.64	75.31	68.94
	DAE _DT	76.59	88.41	84.67
	DAE _RF	86.54	91.26	89.86
	DAE _NB	59.64	79.74	69.08
Proposed method	**DAE_OE_XGB**	**87.25**	**91.64**	**88.66**
K = 10-fold	DAE _OE _KNN	64.74	75.38	69.12
	DAE_OE _DT	76.56	88.31	84.46
	DAE_OE _RF	**86.63**	**91.28**	**89.97**
	DAE _OE _NB	50.62	79.97	66.99

Tables 8 and 9 list the comparative performances of the baseline model and the proposed method. Consequently, the OE-based data transformation can improve the performance of the models trained on raw datasets, as summarized in Table 8. Furthermore, the combination of DAE-based outlier removal and OE-based data transformation in the proposed method outperformed all the compared baselines. The accuracy, F1 score, and ROC measurements of the performance results are presented in Table 8, with the highest scores in bold. The XGB model exhibited the best accuracy of 86.13%, which improved to 87.252% when the OE-based transformation was applied to the baseline model. The XGBoost algorithm yielded the best results among all compared models, with an accuracy rate of 87.252%, F1 score of 88.663, MSE of 0.075, and ROC of 91.64%. The DAE-OE-NB model exhibited lower results than the other proposed predictive models in terms of the evaluation metrics.

Table 9. Evaluation comparison of the proposed methods on experimental Dataset II.

Methods	Features	Accuracy (%)			AUC (%)			F1 score (%)		
		COVID-19	No COVID-19	Difference from baseline	COVID-19	No COVID-19	Difference from baseline	COVID-19	No COVID-19	Difference from baseline
Baseline models	XGB	84.49	84.42	0.07	90.15	90.12	0.03	87.05	86.97	0.08
	KNN	59.00	59.00	0	71.64	71.62	0.02	64.71	64.71	0
	DT	73.08	72.44	0.64	86.21	86.05	0.16	81.67	81.45	0.22
	RF	82.21	81.23	0.98	88.88	88.53	0.35	88.28	88.61	−0.33
	NB	58.90	59.99	1.09	78.77	79.08	−0.31	69.18	69.76	−0.58
Outlier removing method	DAE_XGB	85.93	85.93	0	90.64	90.64	0	87.68	87.68	0
	DAE_KNN	47.51	48.55	−1.04	64.17	64.82	−0.65	54.44	55.44	−1
	DAE_DT	74.19	74.35	−0.16	87.21	87.16	0.05	82.98	82.66	0.32
	DAE_RF	82.54	81.97	0.57	89.03	88.92	0.11	88.16	88.96	−0.8
	DAE_NB	45.92	45.85	0.07	77.50	77.49	0.01	66.45	66.41	0.04
Proposed method K = 10-fold	DAE_OE_XGB	**87.78**	**87.72**	**0.06**	**92.28**	**92.23**	**0.05**	**89.95**	**89.94**	**0.01**
	DAE_OE_KNN	61.80	61.81	−0.01	73.40	73.40	0	66.82	66.82	0
	DAE_OE_DT	78.58	78.53	0.05	89.21	89.32	−0.11	85.57	85.41	0.16
	DAE_OE_RF	**86.34**	**85.58**	**0.76**	**91.50**	**91.05**	**0.45**	**90.88**	**90.58**	**0.3**
	DAE_OE_NB	63.24	63.15	0.09	80.68	80.65	0.03	72.12	72.09	0.03

In future research, validating these preliminary findings through experiments using targeted clinical data supplemented by open data sources is imperative. Table 9 offers a detailed comparative analysis of the dataset outcomes, distinguishing between scenarios with and without integrating COVID-19 features. When evaluating the accuracy, the proposed method applied to the dataset that includes COVID-19 features outperformed XGB, KNN, DT, RF, and NB by margins of 0.06%, -0.01%, 0.05%, 0.76%, and 0.09%, respectively. In terms of AUC, the method exhibited enhancements of 0.05%, 0%, −0.11%, 0.45%, and 0.03% when compared with these algorithms. Furthermore, the F1-score analysis reveals that the proposed method, when applied to the dataset with COVID-19 features, achieved superior results over XGB, KNN, DT, RF, and NB, showing improvements of 0.01%, 0%, 0.16%, 0.3%, and 0.03%, respectively.

Table 10. Statistical significance of the overall mean accuracy, p-values, and CI values for hypertension risk prediction using ML algorithms on the target Datasets I and II.

	Dataset I			**Dataset II**		
Algorithms	**Accuracy (%)**	**p-value**	**95% CI**	**Accuracy (%)**	**p-value**	**95% CI**
XGB	87.25	2.52×10^{-35}	86.49~88.01	87.78	1.02×10^{-23}	85.86~89.69
KNN	64.74	2.15×10^{-35}	63.49~65.99	61.80	7.98×10^{-24}	59.63~63.98
DT	76.56	1.93×10^{-35}	75.28~78.08	78.58	5.86×10^{-24}	75.70~80.88
RF	86.62	3.97×10^{-35}	85.63~87.62	86.34	1.98×10^{-34}	84.12~88.56
NB	50.62	2.77×10^{-35}	47.89~53.38	63.24	9.88×10^{-24}	60.48~66.00

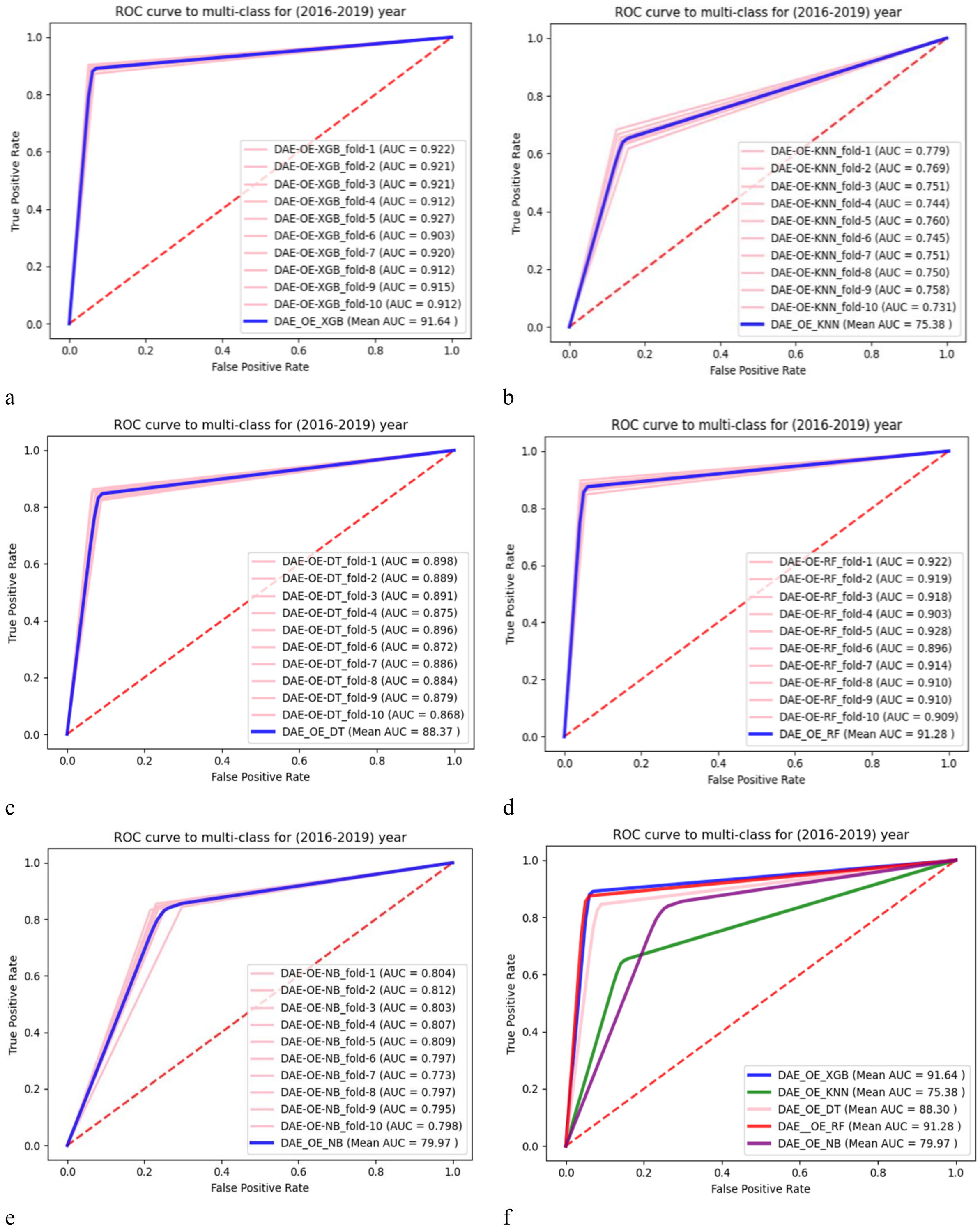


Figure 10. ROC curves of compared algorithms on outlier-removed dataset I using DAE with OE. (a) XGBoost; (b) KNN; (c) DT; (d) RF; (e) NB; (f) Average ROC curves of compared algorithms for DAE with OE (DAE_OE)-based algorithm.

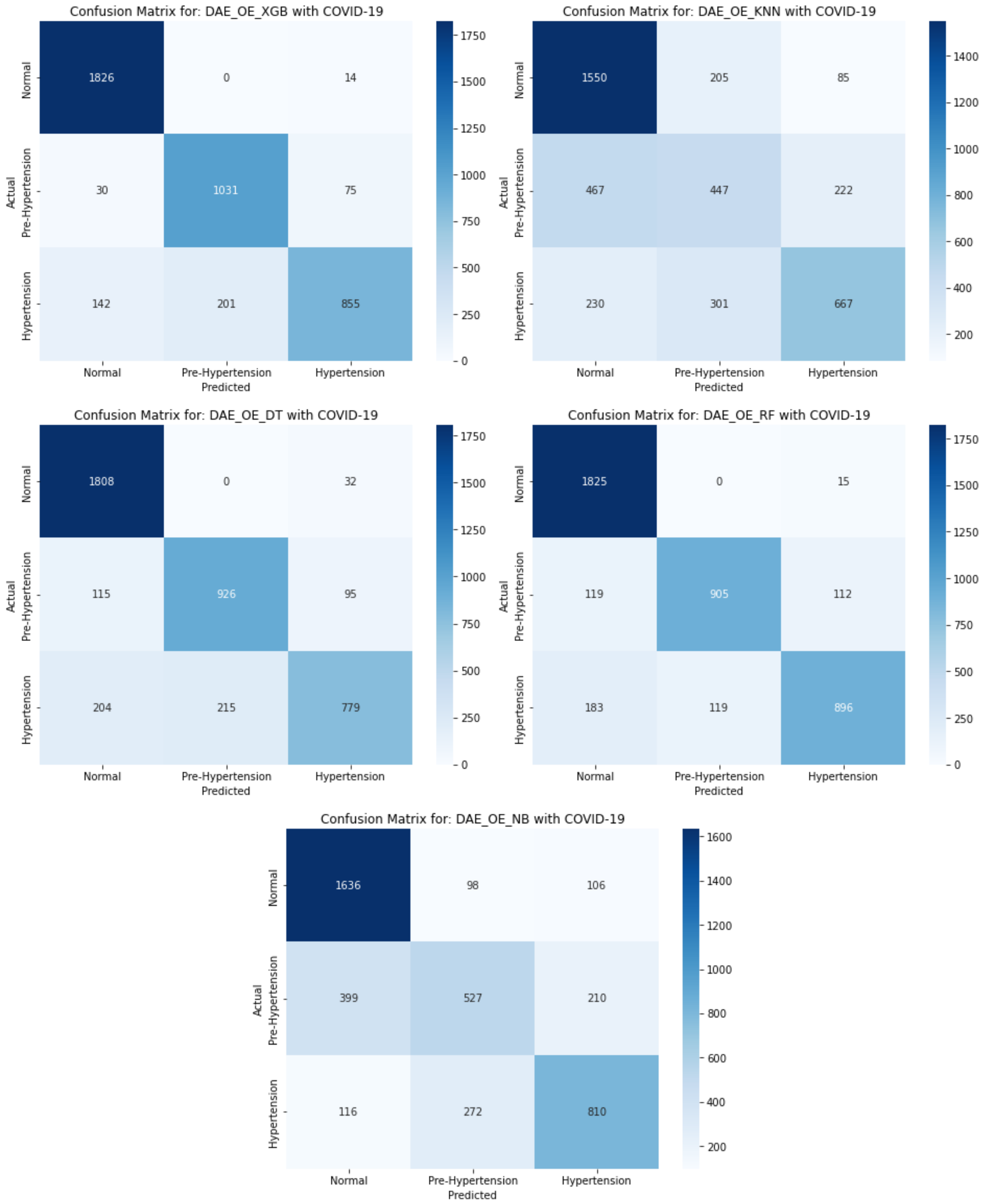


Figure 11. Comparative confusion matrices for hypertension risk prediction models.

Table 10 provides a detailed evaluation of various machine learning algorithms in the context of hypertension risk prediction, focusing on COVID-19 features. A key observation is the statistical significance of accuracy across all evaluated methods ($p < 0.00001$), indicating the reliability and validity of our model's performance by established statistical benchmarks [35]. Notably, implementing the DAE technique significantly enhances the performance of individual algorithms. Among these, the DAE_OE_XGB model (XGBoost augmented by DAE and OE) is the most productive. The XGB algorithm demonstrates a remarkable accuracy of 87.25% (95% CI: 99.43 to 99.71) with Dataset I, and 87.78% (95% CI: 97.09 to 97.59) with Dataset II, as part of our DAE_OE-based framework, as illustrated in Table 9. This investigation further highlights the critical contribution of the COVID-19 feature in enhancing the predictive accuracy of several machine learning algorithms, particularly XGB. Our analysis reveals that XGB, closely followed by RF and DT, significantly benefits from integrating the COVID-19 feature. This contrasts with algorithms like KNN and NB, which exhibit minimal or marginally decreased accuracy improvements when incorporating these features. Such variations underline the distinct influence that the COVID-19 feature exerts on different algorithms, with XGB, DT, and RF showing notably favorable responses.

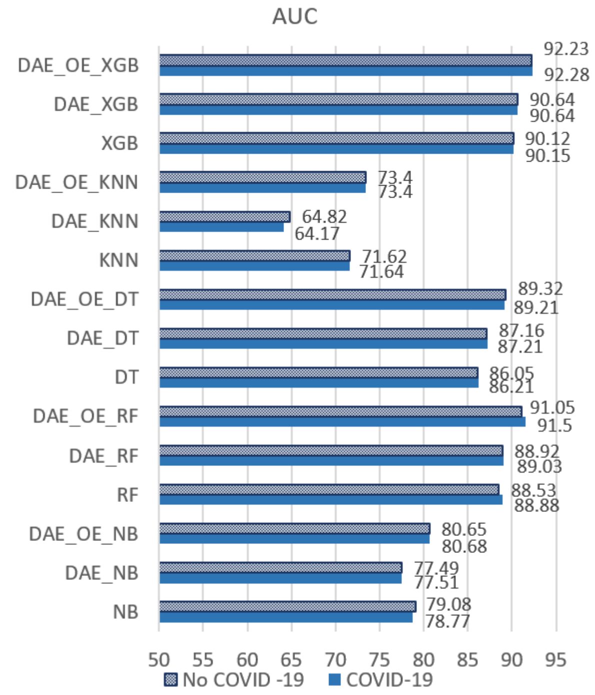


Figure 12. Comparison of hypertension prediction models on dataset II with and without COVID-19 feature.

Figure 10 presents multiclass receiver operating characteristic (ROC) curves for each comparative method applied to the integrated Dataset I, validated using 10-fold cross-validation. By partitioning the training set into several subsets in the 10-fold cross-validation process, we could calculate the mean AUC and observe the variance in the ROC curves. This approach comprehensively evaluated the model performance across different data segments. Figure 10(a)–(e) shows the validation using 10-fold cross-validation as XGB, KNN, DT, RF, and NB for DAE_OE-based algorithms, and Figure 10(f) shows the average ROC curves of compared algorithms for DAE_OE-based algorithm. The XGBoost model with DAE_OE showed superior performance with a mean AUC score of 91.64, followed closely by the RF model with a mean AUC of 91.28. The KNN model exhibited significantly poorer performance, with a mean AUC of 75.38. The ROC curves revealed that the DAE_OE_XGB model was the most effective classifier among the tested models. Previously, in Table 8, we identified XGB as the superior model for predicting outcomes in Dataset I.

Figure 11 shows the confusion matrices representing the performance of five different machine learning models on a classification task to distinguish between Normal, Pre-Hypertension, and

Hypertension states, possibly in a health study related to COVID-19. Each matrix corresponds to a different model: XGBoost (DAE_OE_XGB), K-Nearest Neighbors (DAE_OE_KNN), Decision Tree (DAE_OE_DT), Random Forest (DAE_OE_RF), and Naive Bayes (DAE_OE_NB). The main diagonal of each matrix shows the number of correct predictions for each class, while the off-diagonals show misclassifications. For example, the DAE_OE_XGB model performs strongly with many true positives and fewer misclassifications. In contrast, the DAE_OE_NB model seems to struggle with a higher rate of misclassifying *Normal* instances as *Pre-Hypertension* or *Hypertension*. The variance in these matrices reflects the differing abilities of each model to accurately predict the classes in the dataset.

Finally, Figure 12 shows how the AUC score of the proposed DAE improved by DAE with the OE-based feature. By combining the DAE-OE-based hypertension prediction with the COVID-19 feature, the AUC of the ML-based XGB, KNN, DT, RF, and NB approaches improved by 2.3%, 9.23%, 3%, 2.62%, and 3.18%, respectively.

Our proposed DAE_OE_XGB method, compared with other advanced ML-based techniques for hypertension detection, demonstrates superior performance. As detailed in Table 11, this method outperforms various state-of-the-art ML-based algorithms in accuracy and AUC metrics. For example, DAE_OE_XGB achieved a remarkable accuracy of 87.78% and an AUC of 92.28% on Dataset I, exceeding other models like KNN, LightGBM [16], XGBoost, Ensemble [17], RF [18], NN [19], hypertension [20], pulmonary hypertension (PH) [21], precapillary PH and Group 3 PH [21], and VITRO [23]. This indicates the exceptional efficacy of DAE_OE_XGB in predicting diabetes risk, setting a new benchmark in the field.

Table 11. Comparison of the classification applications of ML models using other methods for hypertension risk prediction.

Algorithms	Accuracy (%)	AUC (%)
KNN [16]	83.5	94.6
LightGBM [16]	86.54	92.9
XGBoost [17]	-	87.7
Ensemble [17]	-	88.1
RF [18]	82.1	86.9
NN [19]	73.2	77.0
Hypertension [20]	68.8	76.6
Pulmonary hypertension (PH) [21]	-	89.0
Precapillary PH [21]	-	88.0
Group 3 PH [21]	-	80.0
VITRO [23]	-	90.9
XGB with COVID-19	**87.78**	**92.28**
XGB without COVID-19	**87.72**	**92.23**

5. Conclusions

This study proposed a comprehensive method for predicting hypertension consisting of three key modules. The first module focused on data preprocessing, in which we integrated external features for COVID-19 patients and performed multicollinearity-based feature selection. The second module involved the application of OE transformation and a DAE for multivariate outlier removal, which were used to detect and predict hypertension using the KNHANES data. Our method demonstrated a significant relationship between COVID-19 and hypertension, as confirmed by the proposed ML and chi-squared multivariate statistics methods. This finding provides a new perspective on the interplay between infectious diseases and chronic conditions such as hypertension. Despite these promising results, our study has some limitations, including the fact that our method was tested only on two datasets and compared using five classifiers. However, our approach showed the potential to improve the predictive performance of the classifiers used in all experiments, as demonstrated by the increased accuracy of the models trained on the original and outlier-removed datasets. Our study offers a novel approach for hypertension prediction, demonstrating the potential of integrating advanced data preprocessing techniques and machine learning methods. Future research should further explore this approach, potentially extending it to detect other diseases. We experimentally demonstrate how the steps of the proposed method improve performance.

Use of AI tools declaration

The authors declare that they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This study was supported by Innovative Human Resource Development for Local Intellectualization program through the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (IITP-2024-2020-0-01462, 80%) for Mi-Hye Kim. Also, this research was financially supported by Hansung University (20%) for Mi-Hwa Song.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. Korea Centers for Disease Control & Prevention. http://knhanes.cdc.go.kr. Accessed: February 4, 2014.
2. C. Wang, P. W. Horby, F. G. Hayden, G. F. Gao, A novel coronavirus outbreak of global health concern, *Lancet*, **395** (2020),470–473. https://doi.org/10.1016/S0140-6736(20)30185-9
3. World Health Organization, https://www.who.int/health-topics/hypertension/#tab=tab_1
4. D. Khongorzul, M. H. Kim, Mahalanobis distance based multivariate outlier detection to improve performance of hypertension prediction, *Neural Process. Lett.*, (2021), 1–13.
5. B. Liao, X. Jia, T. Zhang, R. Sun, DHDIP: An interpretable model for hypertension and hyperlipidemia prediction based on EMR data, *Comput. Methods Programs Biomed.*, **226** (2022), 107088. https://doi.org/10.1016/j.cmpb.2022.107088
6. I. Baik, Region-specific COVID-19 risk scores and nutritional status of a high-risk population based on individual vulnerability assessment in the national survey data, *Clin. Nutr.*, **41** (2022), 3100–3105. https://doi.org/10.1016/j.clnu.2021.02.019

7. J. Kim, K. K. Byon, Leisure activities, happiness, life satisfaction, and health perception of older Korean adults. *Int. J. Ment Health Promot.*, **23** (2021), 155–166. https://doi.org/10.32604/IJMHP.2021.015232
8. J. Y. Kwon, S. W Song, Changes in the prevalence of metabolic syndrome in Korean adults after the COVID-19 outbreak, *Epidemiol. Health*, **5** (2022), e2022101. https://doi.org/10.4178/epih.e2022101
9. K. Song, S. Y. Jung, J. Yang, H. S. Lee, H. S. Kim, H. W. Chae, Change in prevalence of hypertension among Korean children and adolescents during the COVID-19 outbreak: A population-based study, *Children*, **10** (2023), 159. https://doi.org/10.3390/children10010159
10. H. Jeong, H. W. Yim, S. Y. Lee, Impact of the COVID-19 pandemic on gender differences in depression based on national representative data, *J. Korean Med. Sci.*, **38** (2023), 6. https://doi.org/10.3346/jkms.2023.38.e36
11. H. D. Nguyen, H. Oh, M. S. Kim, The association between curry-rice consumption and hypertension, type 2 diabetes, and depression: The findings from KNHANES 2012–2016, *Diabetes Metab. Syndr.*, **16** (2022), 102378. https://doi.org/10.1016/j.dsx.2021.102378
12. A. Sumathi, S. Meganathan, B. V. Ravisankar, An intelligent gestational diabetes diagnosis model using deep stacked autoencoder, *Comput. Mater. Contin.,* **69** (2021), 3109–3126. https://doi.org/10.32604/cmc.2021.017612
13. Y. D. Zhang, M. A. Khan, Z. Zhu, S. H. Wang, Pseudo zernike moment and deep stacked sparse autoencoder for COVID-19 diagnosis, *Comput. Mater. Contin.*, **69** (2021), 3145–3162. https://doi.org/10.32604/cmc.2021.018040
14. H. Dhahri, B. Rabhi, S. Chelbi, O. Almutiry, A. Mahmood, A. M. Alimi, Automatic detection of COVID-19 using a stacked senoising convolutional autoencoder, *Comput. Mater. Contin.*, **69** (2021), 3259–3274. https://doi.org/10.32604/cmc.2021.018449
15. M. A. Hamza, S. B. Hassine, I. Abunadi, F. N. Al-Wesabi, H. Alsolai, A. M. Hilal, et al., Feature selection with optimal stacked sparse autoencoder for data mining, *Comput. Mater. Contin.*, **72** (2022), 2581–2596. https://doi.org/10.32604/cmc.2022.024764
16. M. Fang, Y. Chen, R. Xue, H. Wang, N. Chakraborty, T. Su, et al., A hybrid machine learning approach for hypertension risk prediction, *Neural Comput. Appl.*, **35** (2023), 14487–14497. https://doi.org/10.1007/s00521-021-06060-0
17. H. Kanegae, K. Suzuki, K. Fukatani, T. Ito, N. Harada, K. Kario, Highly precise risk prediction model for new-onset hypertension using artificial intelligence techniques. *J. Clin. Hyper.*, **22** (2020), 445–450. https://doi.org/10.1111/jch.13759
18. L. A. AlKaabi, L. S. Ahmed, M. F. Al Attiyah, M. E. Abdel-Rahman, Predicting hypertension using machine learning: Findings from Qatar biobank study, *Plos One*, **15** (2020), e0240370. https://doi.org/10.1371/journal.pone.0240370
19. F. López-Martínez, E. R. Núñez-Valdez, R. G. Crespo, V. García-Díaz, An artificial neural network approach for predicting hypertension using NHANES data, *Sci. Rep.*, **10** (2020), 10620. https://doi.org/10.1038/s41598-020-67640-z
20. L. Zhang, M. Yuan, Z. An, X. Zhao, H. Wu, H. Li, et al., Prediction of hypertension, hyperglycemia and dyslipidemia from retinal fundus photographs via deep learning: A cross-sectional study of chronic diseases in central China, *PloS One,* **15** (2020), e0233166. https://doi.org/10.1371/journal.pone.0233166

21. M. A. Aras, S. Abreau, H. Mills, L. Radhakrishnan, L. Klein, N. Mantri, et al., Electrocardiogram detection of pulmonary hypertension using deep learning, *J. Cardiac Failure*. **29** (2023), 1017–1028. https://doi.org/10.1016/j.cardfail.2022.12.016
22. B. Ge, H. Yang, P. Ma, T. Guo, J. Pan, W. Wang, Detection of pulmonary hypertension associated with congenital heart disease based on time-frequency domain and deep learning features, *Biomed. Signal Proc. Control*, **81** (2023), 104316. https://doi.org/10.1016/j.bspc.2022.104316
23. M. Jachs, L. Hartl, B. Simbrunner, D. Bauer, R. Paternostro, B. Scheiner, et al., The sequential application of Baveno VII criteria and VITRO score improves diagnosis of clinically significant portal hypertension, *Clin. Gastroent. Hepatol.*, **21** (2023), 1854–1863. https://doi.org/10.1016/j.cgh.2022.09.032
24. G. B. Lee, Y. Kim, S. Park, H. C. Kim, K. Oh, Obesity, hypertension, diabetes mellitus, and hypercholesterolemia in Korean adults before and during the COVID-19 pandemic: A special report of the 2020 Korea National Health and Nutrition Examination Survey, *Epidemiol.Health*, **44** (2022), e2022041. https://doi.org/10.4178/epih.e2022041
25. J. H. Nam, J. I. Park, B. J. Kim, H. T. Kim, J. H. Lee, C. H. Lee, et al., Clinical impact of blood pressure variability in patients with COVID-19 and hypertension, *Blood Press. Monit.,* **26** (2021), 348–356. https://doi.org/10.1097/MBP.0000000000000544
26. J. Kim, S. Jang, W. Lee, J. K. Lee, D. H. Jang, DS4C patient policy province dataset: A comprehensive COVID-19 dataset for causal and epidemiological analysis, in *Proceedings of the 4th Conference on Neural Information Processing Systems (NeurIPS 2020)*, (2020).
27. [NeurIPS 2020] data science for COVID-19 (DS4C), in *DS4C: Data Science for COVID-19 in South Korea*, (2020). https://www.kaggle.com/kimjihoo/coronavirusdataset
28. N. A. Senaviratna, T. M. Cooray, Diagnosing multicollinearity of logistic regression model, *Asian J. Probab. Stat*., **5** (2019), 1–9. https://doi.org/10.9734/ajpas/2019/v5i230132
29. T. Amarbayasgalan, K. H. Park, J. Y. Lee, K. H. Ryu, Reconstruction error based deep neural networks for coronary heart disease risk prediction, *Plos One*, **14** (2019), e0225991. https://doi.org/10.1371/journal.pone.0225991
30. K. Dashdondov, M. H. Kim, K. Jo, NDAMA: A novel deep autoencoder and multivariate analysis approach for IOT-based methane gas leakage detection, *IEEE Access*, **11** (2023), 140740–140751, http://doi.org/10.1109/ACCESS.2023.3340240
31. C. Y. Liou, W. C. Cheng, J. W. Liou, D. R. Liou, Autoencoder for words, *Neurocomputing*, **2** (2014), 84–96. https://doi.org/10.1016/j.neucom.2013.09.055
32. D. Khongorzul, S. M. Lee, M. H. Kim, OrdinalEncoder based DNN for natural gas leak prediction, *J. Korea Converg. Soc.*, **10** (2019), 7–13. https://doi.org/10.15207/JKCS.2019.10.10.007
33. O. Maimon, L. Rokach, *Data Mining And Knowledge Discovery Handbook*, Spring, 2005.
34. J. Brownlee, *Machine Learning Algorithms From Scratch With Python*, 2016.
35. J. Han, J. Pei, H. Tong, Data mining: Concepts and techniques, in *2013 International Conference On Machine Intelligence And Research Advancement*, (2022).
36. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al., Scikit-learn: Machine learning in Python, *J. Mach. Learn Res.*, **12** (2011), 2825–2830.