Frontiers in Bioscience Landmark

Front. Biosci. (Landmark Ed) **2022**; 27(3): 080
https://doi.org/10.31083/j.fbl2703080

Original Research

Predicting Ischemic Stroke in Patients with Atrial Fibrillation Using Machine Learning

Seonwoo Jung[1,†], Min-Keun Song[2,†], Eunjoo Lee[3], Sejin Bae[3], Yeon-Yong Kim[3], Doheon Lee[4,5], Myoung Jin Lee[1,*], Sunyong Yoo[1,*]

[1]Department of ICT Convergence System Engineering, Chonnam National University, 61186 Gwangju, Republic of Korea
[2]Department of Physical & Rehabilitation Medicine, Chonnam National University Medical School & Hospital, 61469 Gwangju, Republic of Korea
[3]Big Data Steering Department, National Health Insurance Service, 26464 Wonju, Republic of Korea
[4]Bio-Synergy Research Center, 34141 Daejeon, Republic of Korea
[5]Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology (KAIST), 34141 Daejeon, Republic of Korea
*Correspondence: syyoo@jnu.ac.kr (Sunyong Yoo); mjlee@jnu.ac.kr (Myoung Jin Lee)
†These authors contributed equally.
Academic Editor: Alexandros G. Georgakilas
Submitted: 18 November 2021 Revised: 15 February 2022 Accepted: 21 February 2022 Published: 4 March 2022

Abstract

Background: Atrial fibrillation (AF) is a well-known risk factor for stroke. Predicting the risk is important to prevent the first and secondary attacks of cerebrovascular diseases by determining early treatment. This study aimed to predict the ischemic stroke in AF patients based on the massive and complex Korean National Health Insurance (KNHIS) data through a machine learning approach. **Methods**: We extracted 65-dimensional features, including demographics, health examination, and medical history information, of 754,949 patients with AF from KNHIS. Logistic regression was used to determine whether the extracted features had a statistically significant association with ischemic stroke occurrence. Then, we constructed the ischemic stroke prediction model using an attention-based deep neural network. The extracted features were used as input, and the occurrence of ischemic stroke after the diagnosis of AF was the output used to train the model. **Results**: We found 48 features significantly associated with ischemic stroke occurrence through regression analysis (p-value < 0.001). When the proposed deep learning model was applied to 150,989 AF patients, it was confirmed that the occurrence ischemic stroke was predicted to be higher AUROC (AUROC = 0.727 ± 0.003) compared to CHA_2DS_2-VASc score (AUROC = 0.651 ± 0.007) and other machine learning methods. **Conclusions**: As part of preventive medicine, this study could help AF patients prepare for ischemic stroke prevention based on predicted stoke associated features and risk scores.

Keywords: atrial fibrillation; stroke; national health insurance service; machine learning; deep neural network; attention

1. Introduction

In Korea, cerebrovascular diseases are the fourth leading cause of death [1]. It can lead to various functional impairments such as motor weakness, sensory deficit, dysphagia, dysarthria, aphasia, cognitive impairment, and emotional disturbances [2–4]. Therefore, it is important to prevent primary and secondary stroke by providing appropriate treatments, such as oral anticoagulation in patients with atrial fibrillation (AF), through early detection.

AF is a common risk factor of cardioembolic cerebral infarction [5]. It accounts for 7 to 31 percent of stroke patients aged 60 years or older [6–8]. Thromboembolism in the left atrium caused by AF would increase the risk of stroke by four to five times [7–9]. The recent population-based study presented AF as an independent predictor of 30-day and one-year mortality after a first ischemic stroke [10]. Approximately 17 percent of all deaths were attributable to the ischemic stroke with AF. The previous observation study showed that stroke with AF would affect the functional limitation and compromised quality of life [11]. Due to the high risk of recurrent embolism, the development of risk calculating methods for stroke with AF is in progress. In particular, because the pathophysiology of stroke in AF is different from that of non-AF, there is a need for a method that considers these characteristics [12–14].

Most previous studies for predicting ischemic stroke risk in patients with AF were based on statistical methods, such as $CHADS_2$ and CHA_2DS_2-VASc scores [15–17]. The $CHADS_2$ score would reflect the representation of incidence risk for stroke using five factors, including congestive heart failure, hypertensive diseases, more than 74 years of age, diabetes mellitus, and previous cerebrovascular attack [18]. However, $CHADS_2$ has a limitation in that it is difficult to accurately evaluate low-risk groups. To improve the predictive performance in the low-risk group, a CHA_2DS_2-VASc score was proposed considering the presence or absence of vascular diseases, ages 65–74 years, and female gender. CHA_2DS_2-VASc scores have guided many clinicians on using oral anticoagulants as an indicator of bleeding risk, which could suggest its low use for stroke with AF owing to high CHA_2DS_2-VASc scores [19]. It was devised to compensate for the defects of $CHADS_2$, but there are still other limitations. First, CHA_2DS_2-VASc scores are limited in considering various characteristics of ischemic stroke. For example, only vascular diseases were considered, and other mechanisms, such as large-artery atherosclerosis, and small-vessel occlusion, were not considered. Second, CHA_2DS_2-VASc scores have modest performance in stroke risk prediction [20,21].

Herein, we present a machine learning-based method to predict the occurrence of ischemic stroke in AF patients based on the Korean National Health Insurance Service (KNHIS) data. Recent studies have demonstrated that accumulated data of patients in electronic medical record (EMR) can be utilized to predict potential disease risk [22,23]. In Korea, more than 97% of the population is covered by the KNHIS program and the remaining three percent are covered by a medical aid program operated by the KNHIS [24]. The KNHIS contains information on Korean demographic, health examination, and medical use/transaction information. Therefore, it was hypothesized that the accumulated large-scale KNHIS information of the AF patients can be used to predict the further occurrence of ischemic stroke. To handle the massive and complex KNHIS information, we adapted a deep neural network that can express the degree of influence on the output by weight for each column through several hidden layers to identify patterns in the data. The evaluation results showed that many ischemic stroke patients were identified with high AUROC.

2. Materials and Methods

2.1 Data Sources

This study used KNHIS data from January 1, 2005 to December 31, 2018. Since 1995, KNHIS, the single national health insurer, has provided health examinations for all Koreans. The KNHIS database contains complete health information about approximately 50 million Koreans [25]. In this study, case subjects were defined as patients with AF who were newly diagnosed with ischemic stroke, and control subjects were those with AF who had not been diagnosed with ischemic stroke. We used the International Classification of Disease, 10th revision (ICD-10) codes to identify patients with AF and those who had experienced ischemic stroke from the health claim records [26]. We obtained patients diagnosed with AF (ICD-10: I48) between 2005 and 2013. Subsequently, we checked if the selected patients were hospitalized for ischemic stroke (ICD-10: I63) within five years after the diagnosis of AF. Next, we collected demographic, health examination, and medical history information of subjects from the KNHIS database. Demographic information contains gender, age, occupational status, and income level. The medical history includes information on the occurrence of 43 diseases (e.g., hypertensive disease, hemolytic anemia, chronic gastritis, hyperlipidemia, and thyroid diseases). The medical history information in the KNHIS database is built using the medical bills that were claimed by medical service providers for the expenses. Health examination includes results of nine general laboratory tests (e.g., blood pressure, and urinary protein) and six questionnaires on lifestyle and behavior (e.g., smoking, exercise, and drinking). A detailed description of the extracted information was provided in **Supplementary Table 1**.

The study protocol was approved by the Institutional Review Board of the National Health Insurance Service in Korea (NHIS-2020-4-109). The authors confirm that all methods were performed in accordance with relevant guidelines and regulations. The need for informed consent from participants was waived by the ethics committee of the Chonnam National University because this study involved routinely collected medical data that were anonymized at all stages to protect an individual's privacy.

2.2 Regression-based Statistical Analysis

Logistic regression is a statistical technique that estimates the causal relationship between categorical dependent variables and several independent variables and is divided into two types according to the number of categories of dependent variables [27]. A binary logistic regression is used when the dependent variable has two categories of 0 or 1, and polynomial logistic regression is used when the dependent variable is composed of two or more categories. The binary logistic regression, used as a statistical technique in this study, was expressed by defining logistic functions in reverse using logits as shown below in Eqns. 1,2, to express linear relationships between independent and dependent variables [28].

$$\begin{aligned}&\text{Minimize}\;\; p(x)=\sigma(t)=\frac{1}{1+e^{-(\beta_0+\beta_1 x)}}\\&\text{Subjectto}\;\; (\because t=\beta_0+\beta_1 x_1+\cdots+\beta_k x_k)\end{aligned}\tag{1}$$

$$\begin{aligned}&\text{Minimize}\;\; g(p(x))=\sigma^{-1}(p(x))=\text{logit}(p(x))=\\&\text{Subjectto}\;\; \ln\left(\frac{P(x)}{1-P(x)}\right)=\beta_0+\beta_1 x_1+\cdots+\beta_k x_k\end{aligned}\tag{2}$$

Regression coefficient, standard error, Wald chi-square, and p-value were used for binary logistic regression analysis with maximum likelihood estimation. The regression coefficient implies that the dependent variable increases or decreases in proportion to the estimated value when the independent variable increases by one unit [29]. If the coefficient is a positive value, it has a positive correlation and vice versa. As the coefficient was close to zero, the effect of the independent variable decreased [29,30]. The standard error is the standard deviation of the sample means used to determine whether the regression coefficient occurs by accident, revealing the closeness of the sample mean values to the population mean [31]. The smaller the standard error, the closer it is to the population mean, and it spreads closer to the regression line, which implies that the probability of a regression coefficient being accidental is less likely to occur. This shows that the causal relationship between the independent variable and dependent variable is significant. The Wald chi-square is an index for evaluating the importance of each independent variable [27].

$$W = \left(\frac{\beta}{SE(\beta)}\right)^2 \tag{3}$$

where β is the coefficient, and SE is its standard error. The Wald chi-square refers to the ratio of the square of the regression coefficient to its standard error and is expressed as a chi-square distribution [27]. The higher the value, the lower the significance level, indicating that it is an important variable in explaining the dependent variable. The *p*-value is the probability that a value equal to or more than that in the sample is observed, assuming that the null hypothesis is correct [32]. Moreover, a *p*-value less than a certain significance level implies that the observed result is improbable under the null hypothesis and that there is a significant association between the dependent and the corresponding independent variables. However, a *p*-value greater than a certain significance level indicates that there is no significant association between the dependent and the corresponding independent variables. Therefore, the *p*-value for each feature tests the null hypothesis that the feature does not correlate with the occurrence of ischemic stroke. In this study, we set the significance of the *p*-value as 0.001. Multicollinearity was detected by the tolerance and variance inflation factor (VIF). The tolerance is defined as 1-R^2,where R^2 is the coefficient of determination for the regression of a variable on the other independent variables. The VIF is defined as the reciprocal of tolerance. If the VIF value exceeds 10, it is considered to indicate multicollinearity.

2.3 Deep Neural Network for Predicting The Occurrence of Ischemic Stroke in AF Patients

In this study, we used a deep neural network to predict the occurrence of ischemic stroke in AF patients based on KNHIS data (Fig. 1). The deep neural network is composed of multiple hidden layers between an input layer and an output layer. The multiple hidden layers enable the modeling of complex nonlinear relationships through the learning function of a high-level layer formed by combining the features of the lower layer, and learning complex functions mapping the input to the output from data [33]. Among the 75 features extracted from KNHIS, we used 4 demographic information, 31 medical histories, and 13 health examination features, which were considered statistically significant through regression analysis. The dataset was divided into 6:2:2 as training, validation, and test set, respectively. Then, the self-attention mechanism was applied to the deep learning model. The self-attention mechanism improves the prediction performance by estimating the importance of the feature [34]. Input features were fed to the fully-connected and the softmax layers to calculate the self-attention scores.

$$a = \text{softmax}(\text{g}(X)) \tag{4}$$

where X is the selected input features, and g(•) is the fully-connected layer without activation. g(•) can be represented as below.

$$g(x) = Wx + b \tag{5}$$

where $W = [w_1, w_2, ..., w_n]$ is the weight matrix, and b is the bias of each unit. In this study, the output of linear operator g(•) is the same size as the input; therefore, $W \in \mathbb{R}^{48\times48}$ and $b \in \mathbb{R}^{48}$. g(•) is then fed into the softmax function which return a vector of numbers with equal to one.

$$\text{softmax}(z) = \frac{e^{z_i}}{\sum_i e^{z_i}} \tag{6}$$

Then, the component-wise multiplication between input features and self-attention score vector was performed.

$$o = a \bigodot X \tag{7}$$

where $\odot$ is the component-wise multiplication operator. Then, we concatenated the output vector o and input features x as $[o_i \; ; \; x_i]$. The concatenated vector was used to train the three-layer fully-connected neural network for predicting the occurrence of ischemic stroke of AF patients. The hyperparameters, additional techniques, and library information used to build a fully-connected network are described in **Supplementary Section 1** and **Supplementary Table 2**.

3. Results

3.1 General Characteristic of the Study Population

From June 2005 to March 2013, a total of 754,949 patients were diagnosed with AF, of which 62,226 (8.24%) were diagnosed with ischemic stroke five years after diagnosis of AF. Table 1 shows the frequency and proportion of stroke and non-stroke groups in each variable. Exceptionally, insurance fee is continuous data, we reported mean $\pm$ SD (range). The CHA_2DS_2-VASc score ranged from 0 to 9, indicating that the risk increases as the score increases. The mean CHA_2DS_2-VASc score of the non-stroke group was 2.15 points, and the stroke group was 3.01 points. As expected, it was confirmed that the stroke group had higher CHA_2DS_2-VASc scores than the non-stroke group. Next, we checked the individual risk factors of CHA_2DS_2-VASc scores, including five medical history factors, age, and sex. It was confirmed that the stroke group had a high proportion compared with the non-stroke group for the medical history of five diseases considered in the CHA_2DS_2-VASc score. The mean ($\pm$ SD) age of the patients was 64.6 $\pm$ 13.3 years

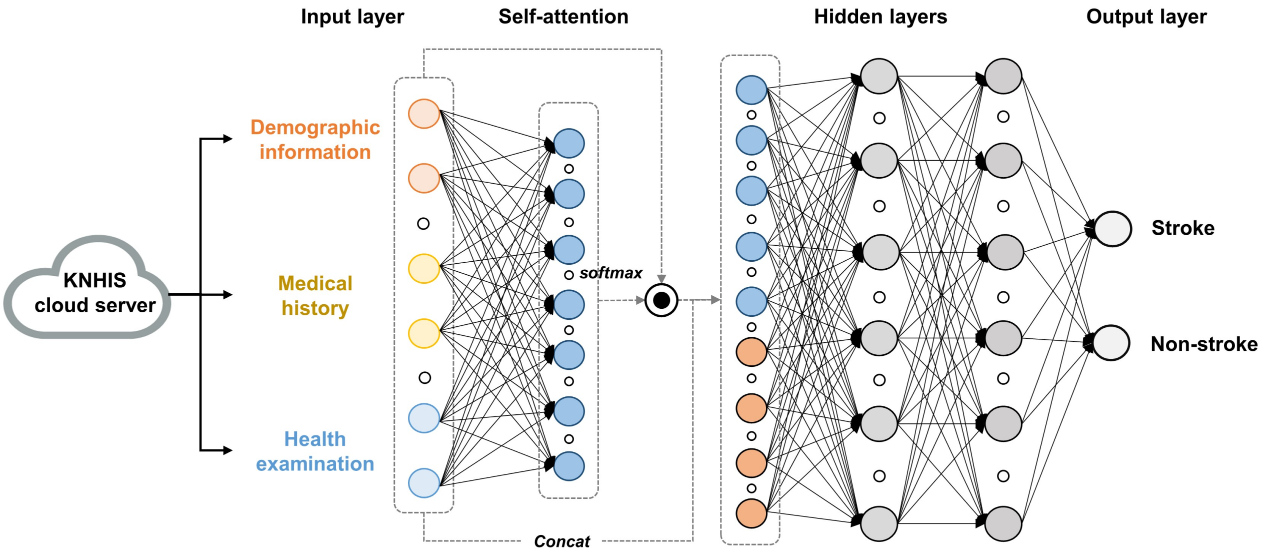

Fig. 1. A systematic overview of the deep neural network-based model that predicts the occurrence of ischemic stroke in AF patients. Demographic, medical history, and health examination information was used as input, and occurrence of ischemic stroke in AF patients was used as output. We calculated attention scores for the input features and concatenated the attention scores with input features. The occurrence of ischemic stroke was predicted by a three-layer fully-connected neural network with non-linear activation function.

in the non-stroke group and 71.5 $\pm$ 9.5 years in the stroke group. We also observed that the stroke group had a higher proportion of patients aged from 65 to 74 years and above 75 years than the non-stroke group. Regarding sex, the non-stroke group had a higher proportion of males (59.21%), whereas the stroke group had a higher proportion of females (50.34%). These results indicate that the CHA_2DS_2-VASc score reflects the characteristics of stroke because it gave a high score for the five medical history features, elderly, and women in the stroke group. However, the difference between the two groups was not significant, and the proportion of patients who scored five or higher in the stroke group did not reach 20%. To estimate the strength of the association between independent variables and occurrence of ischemic stroke, relative risk were provided. When the relative risk value is 1, it means that the independent variable does not affect the result. When the Relative risk value is higher than 1, it means that the risk of the occurrence of ischemic stroke is increased by the independent variable. Conversely, if the relative risk value is less than 1, the risk is decreased by the independent variable.

3.2 Identifying Relationships between Features and Ischemic Stroke Occurrence

We used coefficient values and *p*-values of logistic regression to identify the features related to ischemic stroke occurrence. The VIF values of independent variables are reported in **Supplementary Table 3**. The result indicated that VIF values of all variables did not exceed 3. The results showed that age, sex, and occupational status were important factors in demographic information. We identified important features from medical history, including thyroid diseases, other cardiac arrhythmias, chronic lower respiratory diseases, hemolytic anemia, cancer, hemorrhoids, diabetes mellitus, hypertensive diseases, chronic kidney diseases, heart failure, hyperlipidemia, peripheral vascular disease, gout, noninflammatory gynecological problems, pulmonary embolism, and chronic gastritis. Significant features found by the *p*-values through tests were analyzed based on coefficient, Wald chi-square, and odds ratio with 95% confidence interval (CI) (**Supplementary Section 2**, **Supplementary Table 3** and **Supplementary Fig. 1**).

3.3 Predicting the Occurrence of Ischemic Stroke in Patients with AF

Our method predicts the occurrence of ischemic stroke risk in AF patients based on KNHIS data. We evaluated the area under the curve scores of the receiver operating characteristic (AUROC) and corresponding SD for the average of 10-fold cross validation to assess the predictive performance. We tested the performance for five different types of input feature sets: (i) using all features without feature selection (FS); (ii) using all features with FS; (iii) using demographic features with FS only; (iv) using medical his-

Table 1. Baseline characteristics of the study cohort.

Characteristic	Non-ischemic stroke patients (n = 692,723)	Ischemic stroke patients (n = 62,226)	Relative risk
CHA_2DS_2-VASc scores			
0	105,511 (15.23%)	3416 (5.49%)	
1	169,613 (24.48%)	8912 (14.32%)	
2	156,863 (22.64%)	13,174 (21.17%)	
3	124,433 (17.96%)	14,055 (22.59%)	
4	77,692 (11.22%)	11,071 (17.79%)	
5	36,049 (5.2%)	6210 (9.98%)	
6	15,303 (2.21%)	3427 (5.51%)	
7	5879 (0.85%)	1565 (2.52%)	
8	1275 (0.18%)	356 (0.57%)	
9	105 (0.02%)	40 (0.06%)	
Age			
Age ≥75	106,976 (15.45%)	16,682 (26.81%)	
Age 65–74	184,629 (26.65%)	19,907 (31.99%)	
Age ≤64	401,118 (57.90%)	25,637 (42.00%)	
Sex			
Male	410,197 (59.21%)	30,900 (49.66%)	
Female	282,526 (40.79%)	31,326 (50.34%)	
Insurance type			
Employee insured	129,787 (18.74%)	5404 (8.68%)	
Self-employed insured	562,936 (81.26%)	52,978 (85.14%)	
Income level			
Insurance fee	111,091.83 ± 111,710.27 (3610–1,445,400)	110,919.66 (±108,925.21) (3450–1,391,500)	
Medical history			
Heart failure	92,580 (13.36%)	11,301 (18.16%)	1.3908
Hypertensive diseases	335,126 (48.38%)	36,252 (58.26%)	1.4415
Diabetes mellitus	122,587 (17.70%)	13,802 (22.18%)	1.2927
Vascular disease	47,126 (6.8%)	5561 (8.94%)	1.3081
Operation history	15,076 (2.18%)	850 (1.37%)	0.6426
Hyperlipidemia	55,133 (7.96%)	4589 (7.37%)	0.9268
Dorsopathies	348,529 (50.31%)	35,364 (56.83%)	1.2725
Vision loss	197,680 (28.54%)	22,570 (36.27%)	1.3817
Osteoarthritis	222,716 (32.15%)	24,863 (39.96%)	1.3637
Thyroid disease	57,176 (8.25%)	4366 (7.02%)	0.8502
Cardiac arrhythmias	95,386 (13.77%)	7164 (11.51%)	0.8277
Obesity	399 (0.06%)	22 (0.04%)	0.6339
Gout	24,065 (3.47%)	2367 (3.8%)	1.0899
Hyperplasia of prostate	97,552 (14.08%)	8662 (13.92%)	0.9877
Liver disease	69,604 (10.05%)	5321 (8.55%)	0.8487
Asthma and chronic obstructive pulmonary disease (COPD)	222,117 (32.06%)	21,769 (34.98%)	1.1275
Gynecological problems	42,009 (6.06%)	3170 (5.09%)	0.8433
Osteoporosis	56,796 (8.2%)	6750 (10.85%)	1.3239
Chronic kidney failure	11,506 (1.66%)	1210 (1.94%)	1.1575
Pulmonary thromboembolism	2345 (0.34%)	220 (0.35%)	1.0407
Hearing loss	34,061 (4.92%)	3754 (6.03%)	1.2175
Disorders of gallbladder, biliary tract and pancreas	12,520 (1.81%)	1322 (2.12%)	1.1622

Table 1. Continued.

Characteristic	Non-ischemic stroke patients	Ischemic stroke patients	Relative risk
	(*n* = 692,723)	(n = 62,226)	
Hemorrhoids	35,785 (5.17%)	2537 (4.08%)	0.7948
Diverticular disease of intestine	2721 (0.39%)	253 (0.41%)	1.0322
Rheumatoid arthritis	25,106 (3.62%)	2518 (4.05%)	1.1104
Heart valve disease	18,801 (2.71%)	1840 (2.96%)	1.0840
Neuropathy	60,090 (8.67%)	6342 (10.19%)	1.1762
Dizziness	99,750 (14.4%)	10,999 (17.68%)	1.2489
Incontinence	7442 (1.07%)	893 (1.44%)	1.3042
Ureter stones	14,185 (2.05%)	1128 (1.81%)	0.8917
Anemia	19,363 (2.8%)	1741 (2.8%)	1.0009
Psoriasis	7016 (1.01%)	585 (0.94%)	0.9331
Headache	59,717 (8.62%)	5577 (8.96%)	1.0398
Parkinson's disease	6077 (0.88%)	803 (1.29%)	1.4215
Cancer	50,916 (7.35%)	4471 (7.19%)	0.9778
Allergy	256,651 (37.05%)	22,689 (36.46%)	0.9771
Chronic gastritis/GERD	292,591 (42.24%)	25,678 (41.27%)	0.9640
Sexual dysfunction	2163 (0.32%)	143 (0.23%)	0.7518
Insomnia	35,727 (5.16%)	3598 (5.78%)	1.1168
Hypotension	2855 (0.41%)	300 (0.48%)	1.1544
Depression	32,539 (4.7%)	3354 (5.39%)	1.1413
Somatoform disorder	14,162 (2.04%)	1362 (2.19%)	1.0659
Dementia	18,037 (2.6%)	2942 (4.73%)	1.7362
Anxiety disorder	37,483 (5.41%)	3442 (5.53%)	1.0216
(INP)Afloqualone	355,648 (51.34%)	37,304 (59.95%)	1.3789
Operation history	338,130 (48.81%)	31,516 (50.65%)	1.0697
Emergency Care (EDI)	47,849 (6.91%)	4046 (6.5%)	0.9421
Health examination			
High blood pressure ($\geq$140 mmHg)	275,011 (39.7%)	26,321 (42.3%)	1.1035
Urinary protein ($\geq$300 mg/dL)	2772 (0.41%)	304 (0.49%)	1.2000
Smoking	76,892 (11.1%)	7404 (11.9%)	1.0745
Exercise	310,339 (44.8%)	26,010 (41.8%)	0.8938
Drinking	390,365 (56.4%)	36,526 (58.7%)	1.0922

tory features with FS only; and (v) using health examination features with FS only (Fig. 2a). From the results, we found that using all features with FS (AUROC = 0.722 $\pm$ 0.004) exhibited better performance than using all features without FS (AUROC = 0.714 $\pm$ 0.003) and using a single feature set only (AUROC = 0.613~0.679). These results indicate that the proposed model considers the complex associations of large-scale feature sets. Next, we compared our method with other machine learning methods, including logistic regression, XGBoost, and random forest (Fig. 2b). Both XGBoost and random forest optimized hyperparameters with Bayesian optimizers. In both methods, the maximum depth of the tree was tuned in the range of 5 to 10, and the number of classifiers was tuned in the range of 10 to 500. To prevent overfitting, XGBoost tuned minimum child weight (1–10) and learning rate (0.01–0.1), and random forest tuned minimum sample split (2–10) and minimum sample leaf (0.01–0.5). Results indicate that the proposed deep learning model has better AUROC than logistic regression (AUROC = 0.691 $\pm$ 0.005), XGBoost (AUROC = 0.708 $\pm$ 0.004), and random forest (AUROC = 0.694 $\pm$ 0.009). Furthermore, we compared the prediction performance of our method with the CHA_2DS_2-VASc scores (Fig. 2c). Since CHA_2DS_2-VASc scores are predictors of cardioembolic sources, we screened the patients corresponding to the cerebral infraction due to embolism of cerebral arteries (ICD-10: I63.4) in ischemic stroke patients. Previous study indicated that patients with ICD-10 code of I63.4 includes about 73% subtype diagnosis of cerebral embolism [35]. Through this process, we finally selected 954 ischemic stroke patients and conducted the experiment. The results indicated that an AUROC value of the proposed method was 0.727 $\pm$ 0.003 and an AUROC value of CHA_2DS_2-VASc scores was 0.651 $\pm$ 0.007.

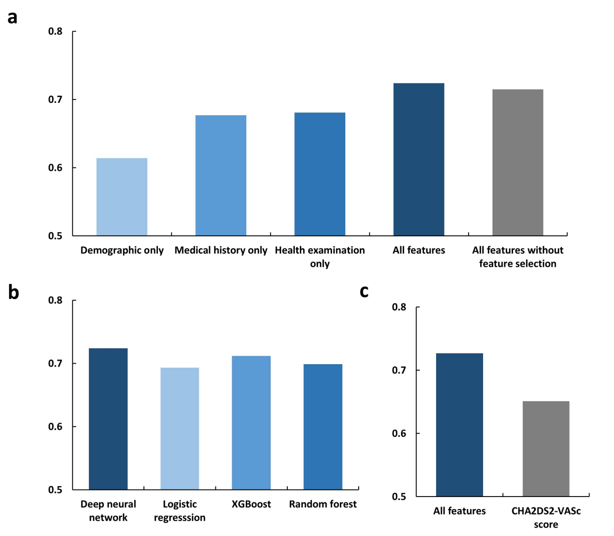

Fig. 2. AUROC value of models generated from different datasets and methods. (a) We compared the AUROC performance of the proposed method for four different input datasets, including all features, demographic feature only, medical history features only, and health examination features only. (b) We compared the performance of our method with other machine learning methods, including logistic regression, XGBoost, and random forest. (c) We compared the performance of our method with the CHA2DS2-VASc scores.

In the prediction of the occurrence of ischemic stroke in AF patients, we firstly calculated precisions for different positive/negative ratios to evaluate the precision performance in the various skewness of datasets (Table 2) [36]. To do this, a negative set was generated by random sampling of the AF patients without ischemic stroke at various rates. The negative sets were generated ten times at each rate, and the performance for each case was evaluated by averaging the results. Results indicated that the precision performance of the proposed deep learning model was decreased when skewness was increased. In realistic scenario, the precision score of the proposed method was 0.132 $\pm$ 0.011, logistic regression was 0.095 $\pm$ 0.013, XGBoost was 0.121 $\pm$ 0.012, and random forest was 0.109 $\pm$ 0.013. Next, we checked the recall performance. Out of 12,445 ischemic stroke patients of the test dataset, the proposed model covered 9508 patients (r = 0.764 $\pm$ 0.005). Furthermore, the proposed method performed better compared to other machine learning methods (r = 0.678~0.717).

The model output can be interpreted as an approximate probability of ischemic stroke occurrence and has a value between 0 and 1. In general, the decision threshold that predicts ischemic stroke occurrence based on the model output value is often 0.5. However, the default threshold may not represent an optimal interpretation of the predicted probabilities [37]. In this study, the class distribution of the dataset is skewed, and predicted probabilities are not calibrated. This is a classification problem with imbalanced classes [38]. To solve this, we identified the optimal threshold value of the model output to judge the occurrence of ischemic stroke. We calculated the F1-scores, which is the harmonic mean of precision and recall, by changing the threshold of the model output. The best performance (F1-score = 0.223) was when the threshold value was 0.519.

4. Discussion

AF is the most common sustained arrhythmia. $CHADS_2$ and CHA_2DS_2-VASc scores are the most popular methods for predicting the risk of ischemic stroke in patients with AF. However, these scores may not be enough to predict the incidence of stroke as they only use five to seven features based on limited information. Previous studies demonstrated that the pathogenesis of stroke in AF patients is complex and involves various factors, such as hypertension, diabetes, dementia, and obesity [39,40]. Moreover, based on KNHIS data, this study found that 32 features were statistically significantly associated with stroke. Therefore, more accurate predictions will be possible if the information from ischemic stroke patients with AF can be completely utilized.

The EMR data accumulated in the hospital applies to this approach because it contains various medical information about patients. However, sharing or releasing EMR data is very difficult owing to privacy and confidentiality issues [41,42]. Therefore, there is a limit to analyzing past medical information of patients who have used several hospitals. In recent years, the Observational Health Data Sciences and Informatics (OHDSI) project has been attempting to standardize and expand the EMR information of hospitals; however, it is currently being conducted only for a few hospitals, and technical and institutional improvements are needed [43]. In Korea, the records of diagnosis, prescriptions, and health examination generated by all medical institutions are collected by KNHIS. This can overcome the limitations of EMR data being accessible only to certain hospitals. Another strength is that it is a model specific to a particular race and region. Most previous risk prediction models were developed in different cohort studies [44]. They are not suitable for the Korean population as their clinical trial cohorts include information on people from different races and regions. This study is significant as the prediction model was developed by considering the characteristics of the Korean population with high AUROC comparing with the CHA_2DS_2-VASc score.

There are additional considerations that may improve our study. First, the type I error (false positive) in the study subjects increases because only ICD-10 codes are considered without prescription information when extracting the AF and ischemic stroke patients. Type I error refers to a situation in which the result incorrectly indicates the presence of a disease, and type II error (false negative) is an opposite situation in which the result does not indicate the

Table 2. Precision and recall performance of various machine learning models in predicting the risk of ischemic stroke occurrence in AF patients.

Measure	Skewness	Deep neural network	Logistic regression	XGBoost	Random forest
Precision	1:01	0.927 ± 0.002	0.882 ± 0.003	0.908 ± 0.002	0.896 ± 0.002
	1:05	0.420 ± 0.007	0.317 ± 0.006	0.375 ± 0.007	0.351 ± 0.009
	All	0.132 ± 0.011	0.095 ± 0.013	0.121 ± 0.012	0.111 ± 0.013
Recall	All	0.764 ± 0.005	0.678 ± 0.004	0.717 ± 0.004	0.713 ± 0.007

presence of a disease [45]. If we use both ICD-10 codes and prescription information, the type I error decreases but the type II error increases. In this study, only ICD-10 codes were used because it was considered important to reduce the type II error. However, it may be more important to reduce Type I errors depending on the researcher's research design. Therefore, we plan to conduct additional analysis considering various combinations of subject selection. The best way to solve this is to use medical examination results, but currently, KNHIS does not collect medical examination results. Second, the prediction results can vary significantly depending on the definition of the case target. In this study, we selected the case subjects by checking whether ischemic stroke occurred during the 5-year period after AF. The period was an option to extract as many ischemic stroke patients as possible. If the purpose of the study was to predict acute stroke, the period should have been shorter. To predict the various occurrence characteristics of ischemic stroke, we can consider learning the model for various datasets and applying ensemble methods. Third, our method had better performance than conventional methods but requires improvement for practical application. Many existing studies typically report predictive performance at approximately 0.850 AUROC value [46]. However, due to different study design, case subject definition, and used datasets, there is a limit to simply comparing them with performance values. Therefore, we have shown how well the proposed method performs compared to conventional methods. However, since this is also only fragmentary performance on the particular dataset, it is necessary to consider the number of different cases in which ischemic stroke occurs in AF patients. In addition, improvement of precision is necessary. The precision performance of the proposed method indicates that only 13.2% (n = 9508) of those who were predicted to have ischemic stroke (n = 72,032) have actually experienced ischemic stroke. In order to be used in practical application, it is necessary to increase true positive predictions and decrease false positive predictions of the model.

In the future, we plan to construct diverse subject datasets, which will allow us to ensemble various machine learning models considering the characteristics of the subjects to improve the predictive performance. This study will allow more accurate prediction through future experiments, and will be useful for patients with AF to prepare for ischemic stroke prevention as part of preventive medicine.

5. Conclusions

This study proposes a new model to predict the occurrence of ischemic stroke in patients with AF. To prevent ischemic stroke, a system for early detection of occurrence should be established. This study predicted the occurrence of ischemic stroke in AF patients based on a machine learning approach by utilizing the massive and complex KNHIS data. The validation results showed that the proposed machine learning model has high AUROC compared to CHA_2DS_2-VASc scores. However, in order for the proposed method to be used in practice, the challenge of improving predictive performance remains. Nevertheless, this study suggested a method of using NHIS data in the development of healthcare applications. In addition, it is expected that further studies will be able to be applied not only to predict ischemic stroke but also to predict various diseases.

Abbreviations

AF, atrial fibrillation; KNHIS, Korean national health insurance; AUROC, area under receiver operating characteristic; EMR, electronic medical record; ICD, international classification of diseases; ATC, anatomical therapeutic chemical; CI, confidence interval; FS, feature selection; OHDSI, Observational Health Data Sciences and Informatics.

Author Contributions

SY proposed the objective and motivation of this work and designed overall method. SJ, MKS, EL, YYK and SB performed data-preprocessing. SJ, MJL, and SY performed preliminary study. DL, MKS and SY helped to write the main manuscript text and provided comments that improve introduction and method parts. SJ, EL, and SY performed evaluation process. EL and MKS provided some ideas in discussion. MJL, and SY supervised this work.

Ethics Approval and Consent to Participate

The study protocol was approved by the Institutional Review Board of the National Health Insurance Service in Korea (NHIS-2020-4-109). The authors confirm that all methods were performed in accordance with relevant guidelines and regulations. The need for informed consent from participants was waived by the ethics committee of the Chonnam National University because this study involved routinely collected medical data that were anonymized at all stages to protect an individual's privacy.

Acknowledgment

Not applicable.

Funding

This research was funded by the Bio-Synergy Research Project (NRF-2012M3A9C4048758) of the Ministry of Science, ICT, and Future Planning, through the National Research Foundation, and supported by the National Research Foundation of Korea grant funded by the Korea government (MSIT) (NRF-2020R1C1C1006007).

Conflict of Interest

The authors declare no conflict of interest.

Supplementary Material

Supplementary material associated with this article can be found, in the online version, at https://www.imrpress.com/journal/FBL/27/3/10.31083/j.fbl2703080.

References

[1] KOREA S. Causes of death statistics in 2020. Statistics Korea: Daejeon. 2020. Available at: http://kostat.go.kr/portal/eng/pressReleases/8/10/index.board?bmode=read&bSeq=&aSeq=414516&pageNo=1&rowNum=10&navCount=10&currPg=&searchInfo=&sTarget=title&sTxt= (Accessed: 28 September 2021).

[2] Kim K, Kim H, Chun I. Correlations between the sequelae of stroke and physical activity in Korean adult stroke patients. Journal of Physical Therapy Science. 2016; 28: 1916–1921.

[3] Mukherjee D, Levin RL, Heller W. The Cognitive, Emotional, and Social Sequelae of Stroke: Psychological and Ethical Concerns in Post-Stroke Adaptation. Topics in Stroke Rehabilitation. 2006; 13: 26–35.

[4] Schneider AT, Pancioli AM, Khoury JC, Rademacher E, Tuchfarber A, Miller R, *et al*. Trends in Community Knowledge of the Warning Signs and Risk Factors for Stroke. Journal of the American Medical Association. 2003; 289: 343.

[5] Boehme AK, Esenwa C, Elkind MSV. Stroke Risk Factors, Genetics, and Prevention. Circulation Research. 2017; 120: 472–495.

[6] Go AS, Hylek EM, Phillips KA, Chang Y, Henault LE, Selby JV, *et al*. Prevalence of diagnosed atrial fibrillation in adults: national implications for rhythm management and stroke prevention: the AnTicoagulation and Risk Factors in Atrial Fibrillation (ATRIA) Study. Journal of the American Medical Association. 2001; 285: 2370–2375.

[7] Waldo AL, Becker RC, Tapson VF, Colgan KJ. Hospitalized patients with atrial fibrillation and a high risk of stroke are not being provided with adequate anticoagulation. Journal of the American College of Cardiology. 2005; 46: 1729–1736.

[8] The Boston Area Anticoagulation Trial for Atrial Fibrillation Investigators*. The effect of low-dose warfarin on the risk of stroke in patients with nonrheumatic atrial fibrillation. New England Journal of Medicine. 1990; 323: 1505–1511.

[9] Friberg L, Bergfeldt L. Atrial fibrillation prevalence revisited. Journal of Internal Medicine. 2013; 274: 461–468.

[10] Marini C, De Santis F, Sacco S, Russo T, Olivieri L, Totaro R, *et al*. Contribution of atrial fibrillation to incidence and outcome of ischemic stroke: results from a population-based study. Stroke. 2005; 36: 1115–1119.

[11] Dalen JE, Alpert JS. Silent Atrial Fibrillation and Cryptogenic Strokes. The American Journal of Medicine. 2017; 130: 264–267.

[12] Kamel H, Okin PM, Elkind MSV, Iadecola C. Atrial Fibrillation and Mechanisms of Stroke: Time for a New Model. Stroke. 2016; 47: 895–900.

[13] D'Souza A, Butcher KS, Buck BH. The Multiple Causes of Stroke in Atrial Fibrillation: Thinking Broadly. The Canadian Journal of Cardiology. 2018; 34: 1503–1511.

[14] Violi F, Soliman EZ, Pignatelli P, Pastori D. Atrial Fibrillation and Myocardial Infarction: a Systematic Review and Appraisal of Pathophysiologic Mechanisms. Journal of the American Heart Association. 2016; 5: e003347.

[15] Malone DC, Charland SL, Agatep BC, Herrera V, Hawk GS, Schrader BJ, *et al*. PRM26 the Use of Claims-Based CHA2DS2-VASc and ATRIA Scores to Predict Stroke/Systemic Embolism and Bleeding Rates among Anticoagulated Patients with Atrial Fibrillation (AFIB) in a Pharmacy-Benefit Management (PBM) Environment. Value in Health. 2012; 15: A464.

[16] Potpara TS, Olesen JB. Comparing the ATRIA, CHADS 2 , and CHA 2 DS 2 -VASc Scores for Stroke Prediction in Atrial Fibrillation. Journal of the American College of Cardiology. 2016; 67: 2316–2317.

[17] van den Ham HA, Klungel OH, Singer DE, Leufkens HGM, van Staa TP. Comparative Performance of ATRIA, CHADS2, and CHA2DS2-VASc Risk Scores Predicting Stroke in Patients with Atrial Fibrillation: Results from a National Primary Care Database. Journal of the American College of Cardiology. 2015; 66: 1851–1859.

[18] Group ISTC. The International Stroke Trial (IST): a randomised trial of aspirin, subcutaneous heparin, both, or neither among 19 435 patients with acute ischaemic stroke. The Lancet. 1997; 349: 1569–1581.

[19] Roldán V, Marín F, Manzano-Fernández S, Gallego P, Vílchez JA, Valdés M, *et al*. The has-BLED score has better prediction accuracy for major bleeding than CHADS2 or CHA2DS2-VASc scores in anticoagulated patients with atrial fibrillation. Journal of the American College of Cardiology. 2013; 62: 2199–2204.

[20] Melgaard L, Gorst-Rasmussen A, Lane DA, Rasmussen LH, Larsen TB, Lip GYH. Assessment of the CHA2DS2-VASc Score in Predicting Ischemic Stroke, Thromboembolism, and Death in Patients with Heart Failure with and without Atrial Fibrillation. Journal of the American Medical Association. 2015; 314: 1030–1038.

[21] Joundi RA, Cipriano LE, Sposato LA, Saposnik G. Ischemic Stroke Risk in Patients with Atrial Fibrillation and CHA2DS2-VASc Score of 1: Systematic Review and Meta-Analysis. Stroke. 2016; 47: 1364–1367.

[22] Perotte A, Ranganath R, Hirsch JS, Blei D, Elhadad N. Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis. Journal of the American Medical Informatics Association. 2015; 22: 872–880.

[23] Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: a Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. IEEE Journal of Biomedical and Health Informatics. 2017; 22: 1589–1604.

[24] Shin DW, Cho B, Guallar E. Korean National Health Insurance Database. JAMA Internal Medicine. 2016; 176: 138.
[25] Kwon S. Payment system reform for health care providers in Korea. Health Policy and Planning. 2003; 18: 84–92.
[26] Wilchesky M, Tamblyn RM, Huang A. Validation of diagnostic codes within medical services claims. Journal of Clinical Epidemiology. 2004; 57: 131–141.
[27] Park HA. An introduction to logistic regression: from basic concepts to interpretation with particular attention to nursing domain. Journal of Korean Academy of Nursing. 2013; 43: 154–164.
[28] Peng CJ, Lee KL, Ingersoll GM. An Introduction to Logistic Regression Analysis and Reporting. The Journal of Educational Research. 2002; 96: 3–14.
[29] Guthery FS, Bingham RL. A Primer on Interpreting Regression Models. Journal of Wildlife Management. 2007; 71: 684–692.
[30] Peng C-YJ, So T-SH, Stage FK, John EPS. The Use and Interpretation of Logistic Regression in Higher Education Journals: 1988–1999. Research in Higher Education. 2002; 43: 259–293.
[31] McHugh ML. Standard error: meaning and interpretation. Biochemia Medica. 2008; 18: 7–13.
[32] Wasserstein RL, Lazar NA. The ASA Statement on p-Values: Context, Process, and Purpose. The American Statistician. 2016; 70: 129–133.
[33] Bengio Y. Learning deep architectures for AI. Now Publishers Inc: Boston, USA. 2009.
[34] Kim H, Nam H. HERG-Att: Self-attention-based deep neural network for predicting hERG blockers. Computational Biology and Chemistry. 2020; 87: 107286.
[35] Goldstein LB. Accuracy of ICD-9-CM coding for the identification of patients with acute ischemic stroke: effect of modifier codes. Stroke. 1998; 29: 1602–1604.
[36] Jeni LA, Cohn JF, De La Torre F. Facing Imbalanced Data Recommendations for the Use of Performance Metrics. International Conference on Affective Computing and Intelligent Interaction and workshops. 2013; 2013: 245–251.
[37] Brownlee J. Imbalanced classification with Python: better metrics, balance skewed classes, cost-sensitive learning. Machine Learning Mastery. 2020.
[38] Fernández A, García S, Galar M, Prati RC, Krawczyk B, Herrera F. Learning from imbalanced data sets. Springer. 2018.
[39] Kim Y, Roh S. The Mechanism of and Preventive Therapy for Stroke in Patients with Atrial Fibrillation. Journal of Stroke. 2016; 18: 129–137.
[40] Sanoski CA. Prevalence, pathogenesis, and impact of atrial fibrillation. American Journal of Health-System Pharmacy. 2010; 67: S11–S16.
[41] Terry NP, Francis LP. Ensuring the privacy and confidentiality of electronic health records. University of Illinois Law Review. 2007; 681.
[42] Price WN, Cohen IG. Privacy in the age of medical big data. Nature Medicine. 2019; 25: 37–43.
[43] Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, *et al*. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Studies in Health Technology and Informatics. 2015; 216: 574–578.
[44] Senoo K, Lane D, Lip GY. Stroke and bleeding risk in atrial fibrillation. Korean Circulation Journal. 2014; 44: 281–290.
[45] Dekking FM KC, Lopuhaä HP, Meester LE. A Modern Introduction to Probability and Statistics: Understanding why and how. Springer Science & Business Media. 2005.
[46] Lip GYH, Tran G, Genaidy A, Marroquin P, Estes C, Landsheftl J. Improving dynamic stroke risk prediction in non-anticoagulated patients with and without atrial fibrillation: Comparing common clinical risk scores and machine learning algorithms. European Heart Journal - Quality of Care and Clinical Outcomes. 2021. (in press)