Original Paper

Enhanced Prediction of Atrial Fibrillation in Patients With Ischemic Stroke Through Electronic Medical Records and Text Mining: Algorithm Development and Validation

Yu-Wei Chen[1,2,3*], MD, PhD; Sheng-Feng Sung[4,5*], MD, PhD; Ya-Han Hu[6,7], PhD; Yu-Hsuan Yang[6], MBA

[1]Department of Neurology, Landseed International Hospital, Taoyuan City, Taiwan
[2]Department of Neurology, National Taiwan University Hospital, Taipei, Taiwan
[3]Center for General Education, National Central University, Taoyuan City, Taiwan
[4]Division of Neurology, Department of Internal Medicine, Ditmanson Medical Foundation Chia-Yi Christian Hospital, Chiayi, Taiwan
[5]Department of Nursing, Fooyin University, Kaohsiung, Taiwan
[6]Department of Information Management, National Central University, Taoyuan City, Taiwan
[7]Asian Institute for Impact Measurement and Management, National Central University, Taoyuan City, Taiwan
*these authors contributed equally

Corresponding Author:

Ya-Han Hu, PhD
Department of Information Management
National Central University
No. 300, Zhongda Rd, Zhongli District
Taoyuan City 320317
Taiwan
Phone: 886 3 4227151 ext 66560
Email: yhhu@mgt.ncu.edu.tw

Related Article:
This is a corrected version. See correction statement in: https://medinform.jmir.org/2026/1/e107705

Abstract

Background: Stroke remains one of the leading causes of mortality and long-term disability worldwide. Atrial fibrillation (AF) is a major and often underdiagnosed risk factor for ischemic stroke as it is frequently asymptomatic and may remain undetected until a catastrophic cerebrovascular event occurs. The lack of timely identification and preventive treatment for AF substantially increases stroke risk. Although previous studies have proposed various predictive models for AF detection, many rely primarily on structured clinical variables and are developed using data from a single institution, which limits their generalizability and real-world applicability across different health care settings.

Objective: The objective of this study was to develop a robust and generalizable AF risk prediction model for patients with stroke using electronic medical records. By integrating structured clinical variables with features derived from unstructured clinical text, this study aimed to construct a more comprehensive representation of patient health status. Furthermore, this study emphasized systematic internal and external validation, along with calibration assessment, to evaluate model stability and generalizability across multiple hospital datasets, thereby supporting its potential use in routine clinical practice.

Methods: This study analyzed datasets from 2 hospitals in Taiwan: Landseed International Hospital (LIH), with 3988 patients, and Chia-Yi Christian Hospital (CYCH), with 5821 patients. We applied 5 feature engineering techniques to extract features from unstructured electronic medical record data, addressed data imbalance using 6 distinct resampling methods, and used 9 classification algorithms to compare model performance across both internal and external validation sets. This study identified the top 20 most important features from the best-performing models for both the LIH and CYCH datasets.

Results: The optimal predictive model for LIH was based solely on structured variables, whereas the model for CYCH achieved superior results by integrating structured variables with text-derived variables obtained from unstructured clinical notes using term frequency–inverse document frequency. Notably, feature importance analysis consistently identified the ratio of E- to A-wave velocities, left atrial size, and age as the top 3 predictive factors across both datasets, underscoring their critical role in AF risk assessment among patients with stroke.

Conclusions: This study demonstrated the development of predictive models for AF in patients with ischemic stroke. Notably, the integration of structured variables with variables derived from unstructured clinical text improved predictive performance in selected model configurations. Rigorous internal and external validation processes confirmed the superior performance of ensemble learning–based machine learning models compared with alternative algorithms, underscoring the potential of this approach for AF risk prediction.

JMIR Med Inform 2026;14:e78117; doi: 10.2196/78117

Keywords: atrial fibrillation; stroke; electronic medical records; text mining; clinical decision support

Introduction

Stroke ranks as the second leading cause of death and the third leading cause of disability worldwide, with both related mortality rates and the global disease burden continuing to rise [1,2]. The most common types of recurrent stroke are acute ischemic stroke (AIS) and transient ischemic attack (TIA) [3]. Within 1 year following an ischemic stroke, the risk of recurrence exceeds 7% [4-6]. Moreover, patients who experience a recurrent stroke within 10 years face more than twice the risk of death or disability compared with those who experience a first stroke [7].

Atrial fibrillation (AF) is a common sustained or paroxysmal cardiac arrhythmia [8,9]. Treatment may often be delayed as AF can be asymptomatic or episodic, making it challenging to detect [10]. AF is closely associated with cerebrovascular disease: irregular heartbeats over time can impair efficient blood contraction and clearance, potentially leading to blood clot formation. These clots may then travel to the brain and cause a stroke if AF remains untreated, increasing the risk of acute complications [11,12]. Patients with AF have twice the risk of dying from cardiovascular disease and up to a 5-fold increased risk of stroke [13]. Consequently, AF impacts patients with stroke significantly in terms of risk, treatment, and daily care requirements [14-17]. Regarding medication, for patients with stroke and AF, oral anticoagulants (eg, factor IIa and Xa inhibitors and warfarin) have been shown to reduce mortality by one-quarter and the risk of recurrent stroke by two-thirds [12]. Comparatively, antiplatelet drugs can reduce the risk of recurrent stroke by three-fifths [18]. Newer direct oral anticoagulants such as apixaban, rivaroxaban, edoxaban, and dabigatran are considered relatively safer and more effective, with a reduced risk of major bleeding compared with warfarin [19,20]. Physicians should find more underdiagnosed patients with AF, mostly paroxysmal AF, among those with AIS and TIA to administer appropriate safe and effective treatment for stroke prevention [21].

Risk assessment scales such as CHA2DS2-VASc [22] and its 2024 updated CHA2DS2-VA version [23] are widely used as clinical tools for evaluating AF risk based on variables collected from patient data. However, these assessments often require additional time or supplementary screening items to thoroughly identify associated risks. Long-term screening studies reveal that the diagnosis rate of AF, particularly paroxysmal AF, tends to increase with extended electrocardiographic (electrocardiogram; ECG) monitoring, although this also raises costs. Furthermore, collecting ECG data is time-intensive and can be cumbersome. Some researchers advocate for the use of continuous ECG monitoring (eg, Holter monitors) for AF detection, noting that recordings exceeding 24 hours may improve detection rates. However, given limited medical resources [7], the presence or absence of AF in patients with stroke significantly influences both treatment outcomes and subsequent risk levels. Thus, it is crucial to develop an efficient screening method that enables physicians to more accurately identify high-risk patients [10].

Electronic medical records (EMRs) contain rich structured and unstructured information that can be leveraged to support risk stratification and early detection of AF among patients with ischemic stroke [24]. While prior studies have demonstrated the potential of machine learning models for AF prediction, many have relied primarily on structured variables and data from a single institution, which may limit model generalizability and real-world applicability across different clinical settings [25-27]. In addition, the incremental value of incorporating unstructured clinical text into predictive models has not been systematically evaluated using rigorous external validation.

The aim of this study was to develop and validate a robust and generalizable machine learning model for predicting AF during hospitalization among patients with ischemic stroke using EMRs. Specifically, we sought to (1) integrate structured clinical variables with features derived from unstructured clinical text; (2) systematically evaluate multiple feature extraction methods, resampling strategies, and classification algorithms; and (3) assess model performance and calibration through rigorous internal and external validation across 2 independent hospital datasets. By emphasizing generalizability and real-world applicability, this study aimed to provide an effective decision support tool to facilitate early identification of high-risk patients and support clinical screening strategies for AF after stroke.

Methods

Study Population

This retrospective study was conducted using EMR datasets from 2 hospitals in Taiwan. The study population included all inpatients who were diagnosed with or suspected of having ischemic stroke by a physician during hospitalization, as well as those who presented with stroke-related symptoms. The dataset from Landseed International Hospital (LIH) covered the period from 2018 to 2022 and included a total of 3988 patients, while the dataset from Chia-Yi Christian Hospital (CYCH) spanned 2007 to 2020 and comprised 5821 patients.

In total, 9809 patients were initially identified across the 2 hospitals.

Because the objective of this study was to develop predictive models for AF among patients with ischemic stroke without a prior history of AF, several exclusion criteria were applied. Consistent with prior studies, patients with documented AF prior to the index stroke were classified as having known AF before stroke and were excluded from model development. Known AF before stroke was defined as an AF diagnosis recorded in the medical history at the time of admission, indicating that AF had been identified through outpatient visits or cardiac monitoring before the current hospitalization. In addition, patients with AF detected on admission were excluded. AF detected on admission was identified based on the initial ECG, admission records, or documentation in the present illness or past medical history.

After applying the inclusion and exclusion criteria, the LIH dataset consisted of 1969 inpatients with unstructured clinical records, 2032 inpatients with structured clinical data, and 1226 inpatients with both structured and unstructured data. Similarly, the CYCH dataset included 3319 inpatients with unstructured data, 1441 inpatients with structured data, and 1072 inpatients with both data types.

Patients with both structured and unstructured data were defined as the intersecting cohort, representing overlapping patients across the 2 data sources. This intersecting cohort constituted the final analytic population used for model development and evaluation to ensure consistent feature availability across all included patients. Patients with data available in only 1 data source were not included in the final modeling analyses.

In the LIH intersecting cohort, 7% (86/1226) of the patients were identified as having AF during hospitalization following ischemic stroke. In contrast, in the CYCH intersecting cohort, 32.6% (350/1072) of the patients met the criteria for AF. Therefore, both cohorts exhibited class imbalance with notably different proportions of AF cases, which motivated the use of resampling strategies in subsequent model development.

Ethical Considerations

The study protocol was approved by the institutional review boards of LIH (IRB-22-049) and Ditmanson Medical Foundation Chia-Yi Christian Hospital (IRB2022071). As this was a retrospective study using de-identified medical records, the requirement for informed consent was waived by both institutional review boards. To ensure patient confidentiality, all personal identifiers were removed and replaced with unique study identification numbers prior to analysis. No participants received any financial compensation, as no direct contact or intervention was involved.

Outcome Variable

The primary outcome variable in this study was the occurrence of AF during hospitalization following ischemic stroke among patients in the intersecting cohort. AF status was determined using a predefined set of criteria that integrated multiple data sources from the EMRs, as summarized in Table 1. Specifically, AF was ascertained based on evidence from structured diagnostic information, unstructured clinical text, and medication records. Structured data included *International Classification of Diseases* diagnosis codes indicative of AF, whereas unstructured sources comprised textual mentions of AF in clinical narratives and examination reports, such as Holter monitoring reports; cardiac ultrasound reports; ECG reports, excluding the first ECG on admission; and general medical records, including discharge summaries. Keyword-based text searches were conducted using commonly used terms referring to AF. In addition, prescriptions of antiarrhythmic agents and oral anticoagulants were incorporated as supportive evidence for AF identification [28]. A patient was classified as having AF if any one of the criteria listed in Table 1 was satisfied.

Table 1. Criteria used to ascertain atrial fibrillation (AF) status in the intersecting cohort.

Criterion	Conditions
AF-1	AF documented in the stroke registry during hospitalization
AF-2	AF-related keywords identified in Holter monitoring report text
AF-3	AF-related keywords identified in cardiac ultrasound report text
AF-4	AF-related keywords identified in ECG[a] report text, excluding the first ECG on admission
AF-5	AF-related keywords identified in clinical narratives, including hospital discharge summaries
AF-6	Presence of an AF-related *ICD*[b] diagnosis code (*ICD-9-CM*[c] 427.31 or *ICD-10-CM*[d] I48.91)
AF-7	Prescription of antiarrhythmic medications, including amiodarone, propafenone, or dronedarone
AF-8	Prescription of oral anticoagulants, including warfarin, apixaban, edoxaban, rivaroxaban, or dabigatran

[a]ECG: electrocardiogram.
[b]*ICD*: *International Classification of Diseases*.
[c]*ICD-9-CM*: *ICD, Ninth Revision, Clinical Modification*.
[d]*ICD-10-CM*: *ICD, 10th Revision, Clinical Modification*.

Predictor Variables and Text Preprocessing

Predictor variables comprised data obtained within 72 hours of admission. After excluding patients with a history of AF or diagnosed with AF upon admission, all variables were standardized across both hospitals.

The structured variables, detailed in Multimedia Appendix 1, comprised a comprehensive set of clinical predictors

that are routinely available at hospital admission and during early hospitalization. These structured predictors included demographic characteristics (sex, age, height, and weight); vascular imaging findings derived from carotid duplex sonography; echocardiographic parameters reflecting cardiac structure and function (eg, aortic diameter, left atrial size, ventricular dimensions, ejection fraction, and diastolic function indexes); vital signs (body temperature, heart rate, respiratory rate, and blood pressure); and laboratory measurements covering hepatic function, renal function, coagulation profiles, inflammatory markers, lipid profiles, and hematological indexes. In addition, neurological severity was captured using individual National Institutes of Health Stroke Scale (NIHSS) subitems, as well as the overall NIHSS score, enabling fine-grained representation of stroke-related neurological deficits. Family medical history variables related to hypertension, diabetes mellitus, stroke, and ischemic heart disease were also included as nominal predictors. Together, these structured variables provided a multidimensional characterization of patients' demographic profiles, cardiovascular status, metabolic conditions, neurological severity, and familial risk factors.

Unstructured data encompassed magnetic resonance imaging results, chief concerns, and present illness. Text preprocessing of variables derived from unstructured clinical text involved Chinese-English translation, spell-checking, abbreviation expansion, nonword symbol removal, AF-suggestive word deletion, lowercase conversion, word form reduction, negation marking as "_NEG" using additional negation words [29,30] and the Natural Language Toolkit's (Team NLTK) *mark_negation* function, and stop word removal.

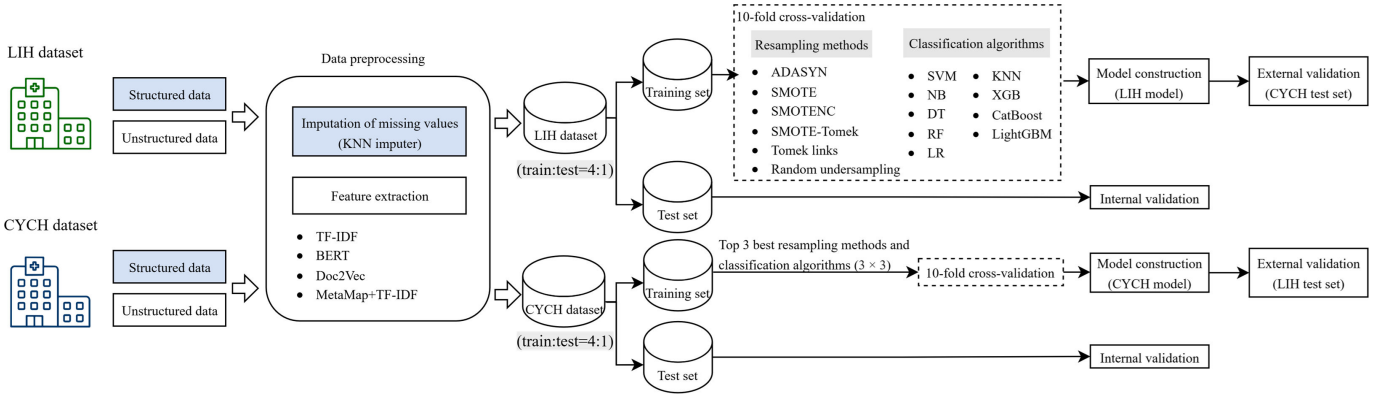
For text feature extraction, we used bidirectional encoder representations from transformers (BERT) [31,32], Doc2Vec [33,34], term frequency–inverse document frequency (TF-IDF) [35,36], and MetaMap [37,38]. These 4 methods were selected to represent complementary paradigms in clinical natural language processing, ranging from deep contextualized language models to traditional statistical and knowledge-based approaches. BERT has demonstrated strong performance in modeling contextual semantics and capturing complex linguistic patterns in clinical narratives and has been widely adopted in biomedical and clinical natural language processing tasks. Doc2Vec was included as a representative neural embedding method that captures document-level semantics with lower computational complexity, making it suitable for large-scale clinical text analysis. TF-IDF remains a widely used and robust baseline in clinical text mining due to its interpretability and effectiveness in high-dimensional sparse representations. MetaMap was incorporated to leverage domain-specific medical knowledge by mapping clinical text to standardized concepts in the Unified Medical Language System, which has been shown to enhance semantic normalization and downstream clinical prediction tasks.

Medical concepts (concept unique identifiers) were identified via MetaMap and subsequently transformed into feature vectors using TF-IDF applying a term frequency threshold with maximum and minimum values set to 0.9 and 0.1, respectively. In addition, Chinese text translation and spell-checking were conducted using *googletrans* (version 3.1.0a0) and *pyspellchecker* (version 0.8.1) to improve text consistency prior to feature extraction.

Prediction Model Development and Evaluation

Figure 1 delineates a comprehensive machine learning model validation workflow encompassing both internal and external validation phases. The internal validation phase uses the LIH dataset and a 10-fold cross-validation strategy. During this process, various combinations of resampling methods and modeling techniques are evaluated, with the top 3 best-performing combinations (totaling 9 configurations) selected for further analysis. This step is crucial for identifying the most appropriate models and parameter settings for the given data.

Figure 1. The process of machine learning model construction. ADASYN: adaptive synthetic sampling; BERT: bidirectional encoder representations from transformers; CatBoost: categorical boosting; CYCH: Chia-Yi Christian Hospital; DT: decision tree; KNN: k-nearest neighbor; LightGBM: light gradient-boosting machine; LIH: Landseed International Hospital; LR: logistic regression; NB: naive Bayes; RF: random forest; SMOTE: synthetic minority oversampling technique; SMOTENC: SMOTE for nominal and continuous features; SVM: support vector machine; TF-IDF: term frequency–inverse document frequency; XGB: extreme gradient boosting.



The external validation phase leverages an independent dataset from CYCH. In this stage, the optimal model and resampling method combinations identified during internal validation are used to train models, which are subsequently evaluated on the external dataset. This phase is essential for assessing the models' generalizability, ensuring their robustness and efficacy when applied to novel, unseen data. This rigorous 2-stage validation approach enhances the reliability and external validity of the developed predictive models, providing a robust framework for clinical application in AF risk assessment among patients with stroke. For both the internal and external validation phases, we used k-nearest neighbor imputation to address missing values within structured data. Additionally, we applied multiple feature extraction techniques to the unstructured text, including TF-IDF, BERT, Doc2Vec, and MetaMap combined with TF-IDF, followed by subsequent evaluations to compare model performance. To comprehensively evaluate the predictive power of different data types, we constructed and compared three distinct model configurations: (1) models using only structured variables, (2) models using only variables derived from unstructured clinical text (ie, text-derived variables), and (3) models combining both structured and text-derived variables.

To address the class imbalance present in the datasets from both hospitals, we implemented various resampling techniques during the model training phase, including adaptive synthetic sampling, synthetic minority oversampling technique (SMOTE), SMOTE for nominal and continuous features, SMOTE-Tomek, Tomek links, and random undersampling. We stratified and split the data into training and test sets with a 4:1 ratio using the parameter *stratify=df[['AF']]* to maintain consistent class ratios across sets.

We evaluated multiple classifiers, including support vector machine, Gaussian naïve Bayes, k-nearest neighbor, decision tree, random forest, logistic regression, Extreme Gradient Boosting (XGBoost), categorical boosting (CatBoost), and light gradient-boosting machine (LightGBM). The training process incorporated 10-fold cross-validation with stratification (*stratify=df[['AF']]*) to obtain robust performance estimates. We assessed model performance using multiple evaluation metrics, including the area under the receiver operating characteristic curve (AUC), sensitivity, specificity, and expected calibration error. Following the identification of optimal model combinations, we conducted both internal and external validation. The final results reflected the outcomes of the test set.

To enhance model interpretability, we used Shapley values (using *shap* version 0.45.0) to explain the best-performing models in both the LIH and CYCH datasets. This approach provides insights into the relative importance of different features in the prediction task.

Baseline Models

The baseline model was compared with machine learning models using traditional risk-scoring models (AS5F [age, stroke severity NIHSS>5 to find atrial fibrillation] [39-41] and CHASE-LESS [coronary, heart failure, age, stroke severity–lipidemia, sugar, and prior stroke] [15,42,43]). CHASE-LESS was developed using data derived from a Taiwanese population, which is consistent with the population of this study, and its variables include age, NIHSS score, coronary heart disease, congestive heart failure, hyperlipidemia, diabetes mellitus, and prior stroke or TIA. The threshold value of CHASE-LESS was set to 6, and the scoring items were based on the variables in the EMR at the time of admission. The AS5F score is designed to determine the need for long-term ECG monitoring in patients with AIS or TIA.

Results

We conducted modeling and prediction using both structured and text-derived variables from the two hospitals. In the internal validation for LIH (Table 2), the combination of structured variables with logistic regression yielded the best performance, achieving an AUC of 0.896. For text-derived variables, the model using Doc2Vec, random undersampling, and LightGBM achieved an AUC of 0.674. When integrating both structured and text-derived variables, the model incorporating Doc2Vec, SMOTE, and CatBoost performed well, with an AUC of 0.885. Some models showed enhanced performance when integrating both types of variables. For instance, in the LightGBM model implementation, the integration of structured and text-derived variables, coupled with the adaptive synthetic sampling technique, demonstrated superior performance, with an AUC of 0.876. This feature-integrated approach substantially outperformed both models using only structured variables (AUC=0.850) and those using only text-derived variables (AUC=0.468). However, not all combined approaches demonstrated such improvement.

Table 2. Comparative analysis of Landseed International Hospital (LIH) models for atrial fibrillation prediction for the area under the receiver operating characteristic curve (AUC) metric: internal validation using the LIH dataset.

Model	Structured variables, AUC	Text-derived variables, AUC				Structured+text-derived variables, AUC			
		BERT[a]	Doc2Vec	MetaMap	TF-IDF[b]	BERT	Doc2Vec	MetaMap	TF-IDF
AS5F[c]	0.597	—[d]	—	—	—	—	—	—	—

Model	Structured variables, AUC	Text-derived variables, AUC				Structured+text-derived variables, AUC			
		BERT[a]	Doc2Vec	MetaMap	TF-IDF[b]	BERT	Doc2Vec	MetaMap	TF-IDF
CHASE-LESS[e]	0.657	—	—	—	—	—	—	—	—
SVM[f]									
Imbalanced	0.835	0.403	0.496	0.499	0.64	0.654	0.837	0.809	0.862
Undersampling	0.778	0.579	0.61	0.48	0.595	0.482	0.76	0.779	0.841
ADASYN[g]	0.812	0.548	0.514	0.466	0.602	0.504	0.846	0.81	0.868
SMOTE[h]	0.803	0.54	0.511	0.51	0.672	0.504	0.824	0.783	0.796
Gaussian NB[i]									
Imbalanced	0.52	0.439	0.604	0.572	0.466	0.404	0.535	0.566	0.468
Undersampling	0.763	0.495	0.671	0.483	0.574	0.616	0.764	0.551	0.599
ADASYN	0.455	0.402	0.533	0.534	0.643	0.452	0.456	0.545	0.52
SMOTE	0.458	0.507	0.532	0.55	0.525	0.477	0.42	0.547	0.609
KNN[j]									
Imbalanced	0.618	0.463	0.47	0.519	0.666	0.486	0.642	0.58	0.567
Undersampling	0.646	0.398	0.451	0.491	0.527	0.526	0.576	0.628	0.648
ADASYN	0.627	0.529	0.45	0.483	0.506	0.421	0.603	0.691	0.625
SMOTE	0.594	0.446	0.427	0.511	0.548	0.405	0.577	0.582	0.608
DT[k]									
Imbalanced	0.542	0.529	0.512	0.535	0.503	0.613	0.576	0.531	0.553
Undersampling	0.621	0.433	0.509	0.49	0.539	0.673	0.634	0.501	0.531
ADASYN	0.666	0.439	0.584	0.521	0.448	0.723	0.631	0.567	0.675
SMOTE	0.545	0.45	0.599	0.574	0.439	0.713	0.673	0.604	0.548
RF[l]									
Imbalanced	0.815	0.548	0.535	0.533	0.586	0.627	0.758	0.757	0.756
Undersampling	0.846	0.376	0.658	0.553	0.523	0.609	0.838	0.738	0.829
ADASYN	0.82	0.523	0.532	0.487	0.476	0.6	0.832	0.79	0.821
SMOTE	0.86	0.502	0.539	0.522	0.565	0.621	0.845	0.815	0.851
LR[m]									
Imbalanced	0.896	0.523	0.659	0.504	0.641	0.798	0.862	0.819	0.805
Undersampling	0.84	0.462	0.664	0.492	0.558	0.769	0.783	0.805	0.827
ADASYN	0.865	0.463	0.635	0.537	0.621	0.752	0.847	0.789	0.804
SMOTE	0.866	0.471	0.624	0.543	0.644	0.757	0.833	0.803	0.787
XGB[n]									
Imbalanced	0.845	0.402	0.541	0.571	0.653	0.813	0.812	0.822	0.822
Undersampling	0.829	0.408	0.617	0.537	0.547	0.797	0.787	0.824	0.826
ADASYN	0.849	0.41	0.5	0.46	0.504	0.865	0.868	0.86	0.874
SMOTE	0.856	0.449	0.591	0.474	0.581	0.882	0.868	0.857	0.819
CatBoost[o]									
Imbalanced	0.869	0.41	0.609	0.517	0.652	0.774	0.85	0.778	0.849
Undersampling	0.868	0.407	0.657	0.536	0.573	0.772	0.854	0.806	0.835
ADASYN	0.864	0.524	0.522	0.481	0.558	0.809	0.875	0.857	0.864
SMOTE	0.855	0.46	0.567	0.529	0.564	0.833	0.885	0.84	0.866
LightGBM[p]									
Imbalanced	0.829	0.524	0.46	0.619	0.538	0.801	0.75	0.837	0.798
Undersampling	0.783	0.361	0.674	0.435	0.513	0.777	0.79	0.783	0.81
ADASYN	0.85	0.471	0.642	0.604	0.468	0.852	0.843	0.773	0.876
SMOTE	0.858	0.488	0.665	0.616	0.525	0.811	0.873	0.845	0.837

[a]BERT: bidirectional encoder representations from transformers.
[b]TF-IDF: term frequency–inverse document frequency.
[c]AS5F: age, stroke severity National Institutes of Health Stroke Scale >5 to find atrial fibrillation.
[d]Not applicable.
[e]CHASE-LESS: coronary, heart failure, age, stroke severity–lipidemia, sugar, and prior stroke.
[f]SVM: support vector machine.
[g]ADASYN: adaptive synthetic sampling.
[h]SMOTE: synthetic minority oversampling technique.
[i]NB: naïve Bayes.
[j]KNN: k-nearest neighbor.
[k]DT: decision tree.
[l]RF: random forest.
[m]LR: logistic regression.
[n]XGB: extreme gradient boosting.
[o]CatBoost: categorical boosting.
[p]LightGBM: light gradient-boosting machine.

In the external validation of the LIH model using the CYCH dataset (Table 3), the best performance was achieved using a combination of BERT and SMOTE, with an AUC of 0.795, which exceeded the AUC for models using only structured variables (AUC=0.781) and those using only text-derived variables. In the internal validation for CYCH (Table 4), the best-performing model was a combination of BERT with undersampling and XGBoost, achieving an AUC of 0.861. In the external validation, the optimal combination involved an imbalanced technique with CatBoost, resulting in an AUC of 0.832. These findings suggest that combining structured and text-derived variables can improve predictive performance in some models, highlighting the importance of selecting appropriate models and resampling techniques to enhance model outcomes.

Figure 2 presents the external validation results from the 2 hospitals, indicating that the combination of CatBoost and SMOTE ranked third in both cases as determined by the AUC metric. Overall, ensemble learning demonstrated superior performance throughout the validation process. While the incorporation of text-derived variables did not consistently yield the highest AUC values in both internal and external validation, it is noteworthy that their inclusion generally resulted in improved outcomes based on the expected calibration error during external validation.

Table 3. Performance of the top 3 Landseed International Hospital models incorporating structured and unstructured feature combinations: external validation using the Chia-Yi Christian Hospital dataset.

Model	AUC[a]	Sensitivity	Specificity	ECE[b]
XGB[c]				
Structured	0.781	0.651	0.776	0.187
Structured+MRI[d]	0.771	0.549	0.843	0.174
Structured+CC[e]	0.776	0.623	0.796	0.171
Structured+PI[f]	0.773	0.597	0.798	0.178
Structured+MRI+CC+PI	0.775	0.571	0.821	0.168
CatBoost[g]				
Structured	0.789	0.463	0.892	0.09
Structured+MRI	0.781	0.311	0.942	0.147
Structured+CC	0.788	0.483	0.895	0.093
Structured+PI	0.79	0.46	0.91	0.104
Structured+MRI+CC+PI	0.788	0.443	0.911	0.107
LightGBM[h]				
Structured	0.781	0.611	0.81	0.161
Structured+MRI	0.794	0.509	0.889	0.111
Structured+CC	0.772	0.5	0.864	0.124
Structured+PI	0.784	0.509	0.873	0.124
Structured+MRI+CC+PI	0.795	0.574	0.855	0.133

[a]AUC: area under the receiver operating characteristic curve.
[b]ECE: expected calibration error.
[c]XGB: extreme gradient boosting.
[d]MRI: magnetic resonance imaging.
[e]CC: chief concern.
[f]PI: present illness.

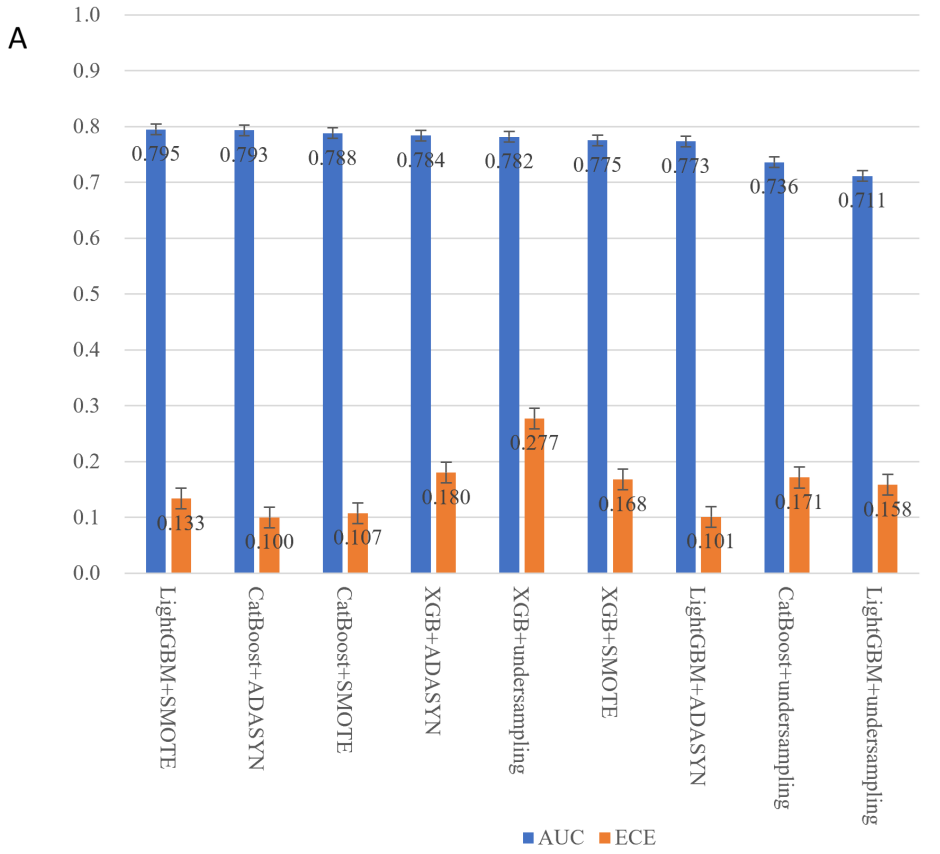
[g]CatBoost: categorical boosting.
[h]LightGBM: light gradient-boosting machine.

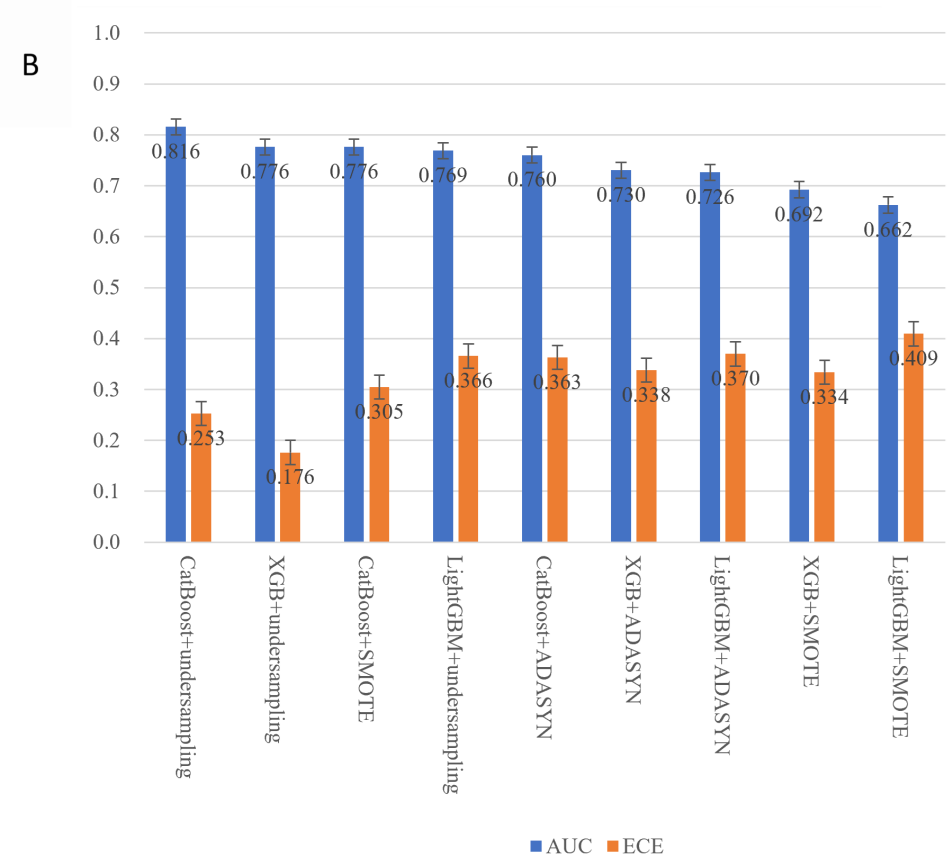
Table 4. Performance of the top 3 Chia-Yi Christian Hospital (CYCH) models incorporating structured and unstructured (magnetic resonance imaging+chief concerns+present illness) feature combinations.

Model	Internal validation (CYCH dataset), AUC[a]	External validation (LIH[b] dataset)	
		AUC	ECE[c]
XGB[d]			
Imbalanced	0.858	0.816	0.08
Undersampling	0.861	0.776	0.176
ADASYN[e]	0.852	0.73	0.338
SMOTE[f]	0.845	0.692	0.334
CatBoost[g]			
Imbalanced	0.851	0.832	0.129
Undersampling	0.859	0.816	0.253
ADASYN	0.843	0.76	0.363
SMOTE	0.843	0.776	0.305
LightGBM[h]			
Imbalanced	0.84	0.764	0.225
Undersampling	0.816	0.769	0.366
ADASYN	0.789	0.726	0.37
SMOTE	0.778	0.662	0.409

[a]AUC: area under the receiver operating characteristic curve.
[b]LIH: Landseed International Hospital.
[c]ECE: expected calibration error.
[d]XGB: extreme gradient boosting.
[e]ADASYN: adaptive synthetic sampling.
[f]SMOTE: synthetic minority oversampling technique.
[g]CatBoost: categorical boosting.
[h]LightGBM: light gradient-boosting machine.

Figure 2. Ranking of the top 3 model and resampling method combinations for external validation: validation of the Landseed International Hospital (LIH) model using the Chia-Yi Christian Hospital (CYCH) dataset (A) and validation of the CYCH model using the LIH dataset (B). ADASYN: adaptive synthetic sampling; AUC: area under the receiver operating characteristic curve; CatBoost: categorical boosting; ECE: expected calibration error; LightGBM: light gradient-boosting machine; SMOTE: synthetic minority oversampling technique; XGB: extreme gradient boosting.





Discussion

Summary of Main Findings

This study developed and validated machine learning–based prediction models for identifying AF during hospitalization among patients with ischemic stroke by leveraging EMRs from 2 independent hospitals. The main findings can be summarized in 3 points. First, models trained on structured clinical variables achieved strong predictive performance, and the integration of unstructured clinical text further improved discrimination and calibration in selected model configurations, particularly in the external validation setting. Second, ensemble learning methods, including CatBoost, XGBoost, and LightGBM, consistently outperformed traditional risk scores and conventional classifiers across internal and external validation. Third, feature importance analysis using SHAP revealed that echocardiographic parameters, including the ratio of E- to A-wave velocities, left atrial size, and age, were among the most influential predictors across both datasets. Together, these findings demonstrate the feasibility and generalizability of incorporating structured and unstructured EMR data to support early risk stratification for AF after ischemic stroke.

Implications of Integrating Structured and Text-Derived Variables

Although the integration of structured variables with variables derived from unstructured clinical text improved predictive performance in several model configurations, the results in Table 2 indicate that such integration did not uniformly outperform models based solely on structured variables. This finding warrants careful interpretation rather than being viewed as a limitation of the proposed framework. In this study, structured clinical variables already encompassed well-established and highly discriminative predictors of AF, including echocardiographic parameters such as the ratio of E- to A-wave velocities, left atrial size, and age. When such strong predictors are present, models trained exclusively on structured variables may achieve near-optimal performance, leaving limited room for additional gains from text-derived variables.

Moreover, variables derived from unstructured clinical text are inherently high-dimensional and potentially noisy, particularly when generated using representation techniques such as TF-IDF or Doc2Vec. For certain classifiers that are sensitive to feature dimensionality or redundancy, the inclusion of text-derived variables may dilute the contribution of strong structured predictors, resulting in marginal or inconsistent performance gains. This phenomenon is consistent with prior observations in studies integrating clinical text with structured EMR data.

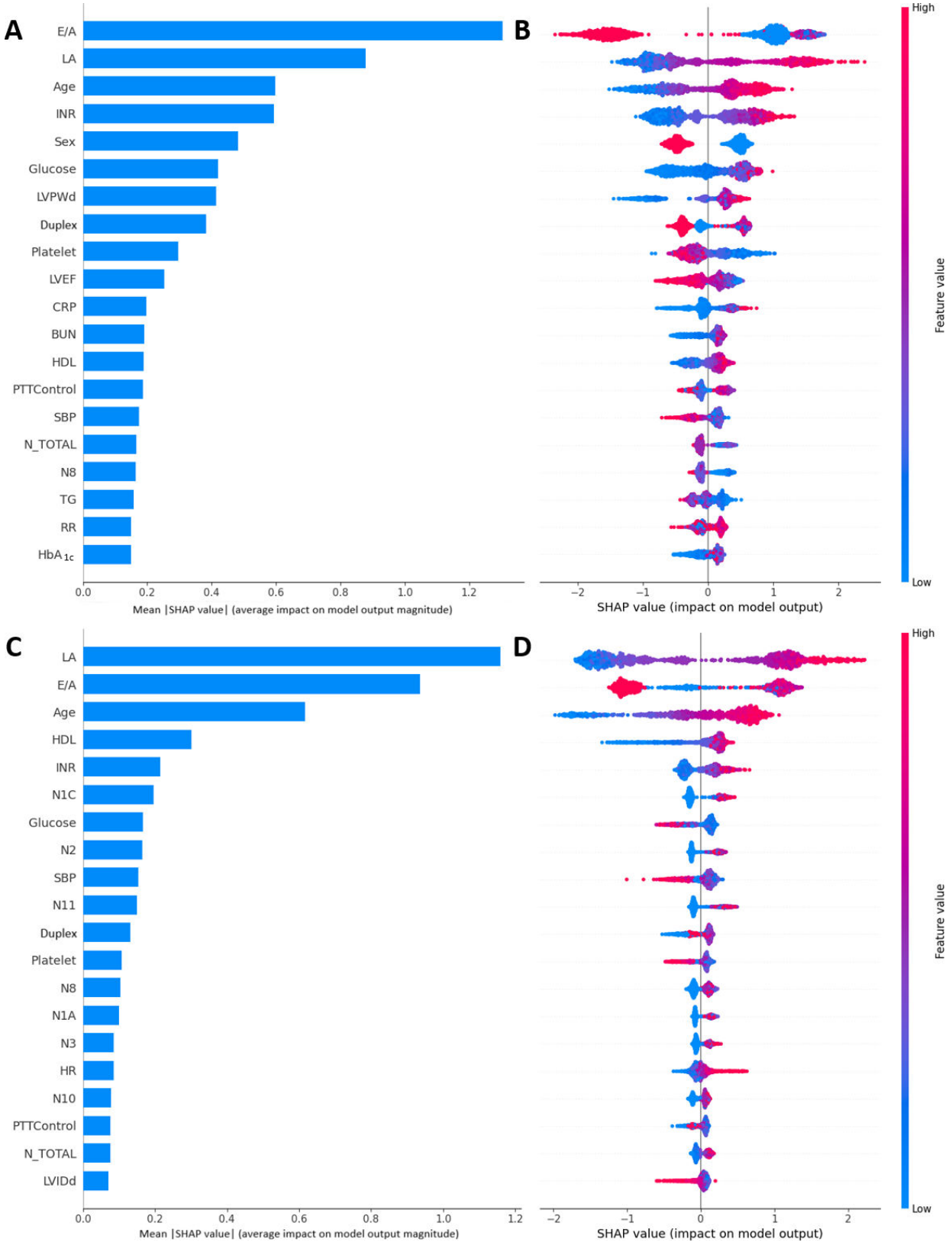
Importantly, our results suggest that the benefit of feature-level integration is highly model dependent. Ensemble learning–based algorithms, including LightGBM, XGBoost, and CatBoost, demonstrated a greater capacity to exploit complementary information from heterogeneous feature sources, particularly when combined with appropriate resampling strategies. In contrast, simpler or less flexible classifiers did not consistently benefit from the inclusion of text-derived variables. These findings indicate that integrating unstructured clinical text should be viewed as a complementary strategy whose effectiveness depends on the interaction among feature representation, classifier architecture, and imbalance-handling methods.

In addition, Table 2 indicates that resampling did not consistently improve performance for all classifiers as some models achieved comparable results when trained on the original imbalanced data. This finding suggests that the necessity of resampling is model dependent. Ensemble learning–based classifiers may inherently mitigate class imbalance effects, whereas other models benefit more substantially from resampling techniques. Therefore, resampling should be considered a robustness-oriented strategy rather than a mandatory preprocessing step.

Significant Features

Figure 3 illustrates the 20 most significant features identified for the LIH and CYCH datasets using shap (version 0.45.0). In Figure 3, features are ranked according to their mean absolute Shapley values, which quantify the global importance of each feature in relation to the model's output. Figure 3A presents the feature ordering for LIH, whereas Figure 3C depicts the ordering for CYCH. Figure 3 shows beeswarm plots that visualize the Shapley values for each patient concerning these features. The beeswarm plots convey the contribution of each feature to the model's output based on the magnitude and direction of the Shapley values. Figure 3B represents the contributions of LIH patients, whereas Figure 3D illustrates those of CYCH patients. Notably, the graphs reveal that the top 3 features consistently identified across both datasets were ratio of E- to A-wave velocities, left atrial size, and age.

Figure 3. The top 20 most important features identified by the model based on structured variables. Panels A and C display bar charts representing the mean absolute Shapley values, which indicate the average contribution of each feature to the model's output. Panels B and D present beeswarm plots of individual Shapley values for each feature across patients, where the position of each dot on the x-axis reflects that feature's contribution to the prediction for a given patient. The color of each dot corresponds to the relative value of the associated feature. BUN: blood urea nitrogen; CRP: C-reactive protein; E/A: ratio of E- to A-wave velocities; HbA_{1c}: hemoglobin A_{1c}; HDL: high-density lipoprotein; HR: heart rate; INR: international normalized ratio; LA: left atrial size; LVEF: left ventricular ejection fraction; LVIDd: left ventricular internal dimension in diastole; LVPWd: left ventricular posterior wall in diastole; N1A: level of consciousness responsiveness; N1C: level of consciousness commands; N2: best gaze; N3: visual fields; N8: sensory; N10: dysarthria; N11: extinction and inattention; PTTControl: partial thromboplastin time control; RR: respiratory rate; SBP: systolic blood pressure; SHAP: Shapley additive explanations; TG: triglyceride.



Clinical Applications and Significance

Predicting AF following a stroke is critical for effective secondary stroke prevention as it directly influences medical decision-making and patient outcomes. Given the constraints of health care resources, it is essential to optimize their use by incorporating efficient screening protocols. In this context, machine learning offers a highly effective approach. This study identified the optimal combination of variables derived from unstructured clinical text through rigorous external validation supplemented by internal validation using the LIH dataset and further validated using EMR data from CYCH. Although models built on internal data are often preferred for clinical applications, our experiments revealed that models developed using the CYCH data occasionally outperformed those based on LIH data. This discrepancy may be attributed to missing information and the relatively small sample size of the LIH dataset. The models developed in this study aim to support clinicians by identifying high-risk patients who would benefit from additional screening measures, such as ECG, ultimately reducing the risk of missed diagnoses and ensuring timely and appropriate treatment.

Limitations

This study has several limitations that should be acknowledged. First, the data source poses a potential issue as the data were derived from EMRs within the hospitals. It is possible that patients were diagnosed with AF outside these institutions, introducing bias, particularly given the older age demographic of the hospital population. Second, regarding the data themselves, while generalizability was tested across 2 different hospitals, the variability in clerical styles between the hospitals may affect the consistency of the variables derived from unstructured clinical text. For structured variables, a high proportion of missing values in factors known to be associated with stroke or AF (eg, smoking history), uncertainties about the accuracy of the records, and the small and variable sample size could all impact the overall efficacy of the model. Finally, it should be noted that the predictors contributing to the model in this study do not imply causality.

Conclusions

This study demonstrates the feasibility of developing machine learning–based predictive models to identify AF in patients with ischemic stroke using EMR data. The integration of structured variables with variables derived from unstructured clinical text, including medical history and magnetic resonance imaging findings extracted from free-text sources, showed potential to improve predictive performance in selected model configurations. Internal and external validation results indicate that ensemble learning–based models performed favorably compared with other algorithmic approaches, supporting their potential utility for AF risk prediction in clinical settings.

Funding

This research was supported in part by the National Central University–Landseed Hospital Research and Development Office (NCU-LSH-111-B-008) and the Ministry of Science and Technology of the People's Republic of China (grant MOST 111-2410-H-008-026-MY2).

Data Availability

The datasets generated or analyzed during this study are not publicly available due to patient privacy and institutional data protection regulations but are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: YWC, YHH, SFS
Data curation: YWC, SFS
Formal analysis: YWC, YHH, SFS, YHY
Investigation: YWC, SFS
Methodology: YWC, YHH, SFS
Supervision: YHH
Validation: YWC, YHH, SFS, YHY
Writing—original draft: YWC, YHH, SFS, YHY
Writing—review and editing: YWC, YHH, SFS, YHY
All authors had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Conflicts of Interest

Reviewers Szu-Yin Lin and Pei-Ju Lee were found to have undisclosed conflicts of interest with members of the author list under the terms of JMIR Publications policy. The editor in chief determined that the content of the reviews did not unduly influence the decision to accept the article. The decision to accept the submission was based primarily on the editor in chief's contemporary appraisal of the manuscript.

Multimedia Appendix 1

Detailed definitions and summary statistics of structured clinical variables extracted from electronic medical records at Landseed International Hospital and Chia-Yi Christian Hospital, including variable descriptions, data types, distributions, and proportions of missing values.
[DOCX File (Microsoft Word File), 32 KB-Multimedia Appendix 1]

References

1. Feigin VL, Brainin M, Norrving B, et al. World Stroke Organization (WSO): global stroke fact sheet 2022. Int J Stroke. Jan 2022;17(1):18-29. [doi: 10.1177/17474930211065917] [Medline: 34986727]
2. Feigin VL, Forouzanfar MH, Krishnamurthi R, et al. Global and regional burden of stroke during 1990–2010: findings from the Global Burden of Disease Study 2010. Lancet. Jan 18, 2014;383(9913):245-254. [doi: 10.1016/s0140-6736(13)61953-4] [Medline: 24449944]
3. Pennlert J, Eriksson M, Carlberg B, Wiklund PG. Long-term risk and predictors of recurrent stroke beyond the acute phase. Stroke. Jun 2014;45(6):1839-1841. [doi: 10.1161/STROKEAHA.114.005060] [Medline: 24788972]
4. Bergström L, Irewall AL, Söderström L, Ögren J, Laurell K, Mooe T. One-year incidence, time trends, and predictors of recurrent ischemic stroke in Sweden from 1998 to 2010: an observational study. Stroke. Aug 2017;48(8):2046-2051. [doi: 10.1161/STROKEAHA.117.016815] [Medline: 28706114]
5. Feigin VL, Norrving B, Mensah GA. Global burden of stroke. Circ Res. Feb 3, 2017;120(3):439-448. [doi: 10.1161/CIRCRESAHA.116.308413] [Medline: 28154096]
6. Lee M, Wu YL, Ovbiagele B. Trends in incident and recurrent rates of first-ever ischemic stroke in Taiwan between 2000 and 2011. J Stroke. Jan 2016;18(1):60-65. [doi: 10.5853/jos.2015.01326] [Medline: 26687123]
7. Lip GYH, Hunter TD, Quiroz ME, Ziegler PD, Turakhia MP. Atrial fibrillation diagnosis timing, ambulatory ECG monitoring utilization, and risk of recurrent stroke. Circ Cardiovasc Qual Outcomes. Jan 2017;10(1):e002864. [doi: 10.1161/CIRCOUTCOMES.116.002864] [Medline: 28096204]
8. Lee SR, Choi EK, Han KD, Cha MJ, Oh S. Trends in the incidence and prevalence of atrial fibrillation and estimated thromboembolic risk using the CHA_2DS_2-VASc score in the entire Korean population. Int J Cardiol. Jun 1, 2017;236:226-231. [doi: 10.1016/j.ijcard.2017.02.039] [Medline: 28233629]
9. Ball J, Carrington MJ, McMurray JJ, Stewart S. Atrial fibrillation: profile and burden of an evolving epidemic in the 21st century. Int J Cardiol. Sep 1, 2013;167(5):1807-1824. [doi: 10.1016/j.ijcard.2012.12.093] [Medline: 23380698]
10. Jones NR, Taylor CJ, Hobbs FD, Bowman L, Casadei B. Screening for atrial fibrillation: a call for evidence. Eur Heart J. Mar 7, 2020;41(10):1075-1085. [doi: 10.1093/eurheartj/ehz834] [Medline: 31811716]
11. Huang CK, Wang JC, Chung CH, Chen SJ, Liao WI, Chien WC. The risk and timing of acute ischemic stroke after electrical cardioversion for atrial fibrillation in Taiwan: a nationwide population-based cohort study. Int J Cardiol. Mar 15, 2022;351:55-60. [doi: 10.1016/j.ijcard.2021.12.035] [Medline: 34954280]
12. Mtwesi V, Amit G. Stroke prevention in atrial fibrillation: the role of oral anticoagulation. Med Clin North Am. Sep 2019;103(5):847-862. [doi: 10.1016/j.mcna.2019.05.006] [Medline: 31378330]
13. Lee E, Choi EK, Han KD, et al. Mortality and causes of death in patients with atrial fibrillation: a nationwide population-based study. PLoS ONE. Dec 26, 2018;13(12):e0209687. [doi: 10.1371/journal.pone.0209687] [Medline: 30586468]
14. Chao TF, Liu CJ, Tuan TC, et al. Lifetime risks, projected numbers, and adverse outcomes in Asian patients with atrial fibrillation: a report from the Taiwan nationwide AF cohort study. Chest. Feb 2018;153(2):453-466. [doi: 10.1016/j.chest.2017.10.001] [Medline: 29017957]
15. Hsieh CY, Kao HM, Sung KL, Sposato LA, Sung SF, Lin SJ. Validation of risk scores for predicting atrial fibrillation detected after stroke based on an electronic medical record algorithm: a registry-claims-electronic medical record linked data study. Front Cardiovasc Med. Apr 29, 2022;9:888240. [doi: 10.3389/fcvm.2022.888240] [Medline: 35571191]
16. Hsieh FI, Lien LM, Chen ST, et al. Get with the guidelines-stroke performance indicators: surveillance of stroke care in the Taiwan stroke registry: get with the guidelines-stroke in Taiwan. Circulation. Sep 14, 2010;122(11):1116-1123. [doi: 10.1161/CIRCULATIONAHA.110.936526] [Medline: 20805428]
17. Hsieh FI, Chiou HY. Stroke: morbidity, risk factors, and care in Taiwan. J Stroke. May 2014;16(2):59-64. [doi: 10.5853/jos.2014.16.2.59] [Medline: 24949310]
18. Garg N, Kumar A, Flaker GC. Antiplatelet therapy for stroke prevention in atrial fibrillation. Mo Med. 2010;107(1):44-47. [Medline: 20222295]
19. Själander S, Sjögren V, Renlund H, Norrving B, Själander A. Dabigatran, rivaroxaban and apixaban vs. high TTR warfarin in atrial fibrillation. Thromb Res. Jul 2018;167:113-118. [doi: 10.1016/j.thromres.2018.05.022] [Medline: 29803981]
20. Yang L, Brooks MM, Glynn NW, Zhang Y, Saba S, Hernandez I. Real-world direct comparison of the effectiveness and safety of apixaban, dabigatran, rivaroxaban, and warfarin in Medicare beneficiaries with atrial fibrillation. Am J Cardiol. Jul 1, 2020;126:29-36. [doi: 10.1016/j.amjcard.2020.03.034] [Medline: 32359718]

21. Granger CB, Alexander JH, McMurray JJV, et al. Apixaban versus warfarin in patients with atrial fibrillation. N Engl J Med. Sep 15, 2011;365(11):981-992. [doi: 10.1056/NEJMoa1107039] [Medline: 21870978]
22. Lip GY, Nieuwlaat R, Pisters R, Lane DA, Crijns HJ. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the Euro Heart Survey on atrial fibrillation. Chest. Feb 2010;137(2):263-272. [doi: 10.1378/chest.09-1584] [Medline: 19762550]
23. Van Gelder IC, Rienstra M, Bunting KV, et al. 2024 ESC Guidelines for the management of atrial fibrillation developed in collaboration with the European Association for Cardio-Thoracic Surgery (EACTS). Eur Heart J. Sep 29, 2024;45(36):3314-3414. [doi: 10.1093/eurheartj/ehae176] [Medline: 39210723]
24. Karnik S, Tan SL, Berg B, et al. Predicting atrial fibrillation and flutter using electronic health records. Annu Int Conf IEEE Eng Med Biol Soc. 2012;2012:5562-5565. [doi: 10.1109/EMBC.2012.6347254] [Medline: 23367189]
25. Chamberlain AM, Roger VL, Noseworthy PA, et al. Identification of incident atrial fibrillation from electronic medical records. J Am Heart Assoc. Apr 5, 2022;11(7):e023237. [doi: 10.1161/JAHA.121.023237] [Medline: 35348008]
26. Hulme OL, Khurshid S, Weng LC, et al. Development and validation of a prediction model for atrial fibrillation using electronic health records. JACC Clin Electrophysiol. Nov 2019;5(11):1331-1341. [doi: 10.1016/j.jacep.2019.07.016] [Medline: 31753441]
27. Verhaeghe J, De Corte T, Sauer CM, et al. Generalizable calibrated machine learning models for real-time atrial fibrillation risk prediction in ICU patients. Int J Med Inform. Jul 2023;175:105086. [doi: 10.1016/j.ijmedinf.2023.105086] [Medline: 37148868]
28. Khurshid S, Keaney J, Ellinor PT, Lubitz SA. A simple and portable algorithm for identifying atrial fibrillation in the electronic medical record. Am J Cardiol. Jan 15, 2016;117(2):221-225. [doi: 10.1016/j.amjcard.2015.10.031] [Medline: 26684516]
29. Hu M, Liu B. Mining and summarizing customer reviews. Presented at: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Aug 22-25, 2004; ACM. Seattle, WA. [doi: 10.1145/1014052.1014073]
30. Liu B, Hu M, Cheng J. Opinion observer: analyzing and comparing opinions on the web. Presented at: Proceedings of the 14th International Conference on World Wide Web; May 10-14, 2005; Chiba, Japan. [doi: 10.1145/1060745.1060797]
31. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. Presented at: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; Jun 2-7, 2019; Minneapolis, MN. [doi: 10.18653/v1/N19-1423]
32. Gardazi NM, Daud A, Malik MK, Bukhari A, Alsahfi T, Alshemaimri B. BERT applications in natural language processing: a review. Artif Intell Rev. 2025;58(6):166. [doi: 10.1007/s10462-025-11162-5]
33. Le Q, Mikolov T. Distributed representations of sentences and documents. Proc Mach Learn Res. 2014;32(2):1188-1196. URL: https://proceedings.mlr.press/v32/le14.html [Accessed 2026-02-13]
34. Getzen E, Ruan Y, Ungar L, Long Q. Mining for health: a comparison of word embedding methods for analysis of EHRs data. In: Zhao Y, Chen DG, editors. Statistics in Precision Health: Theory, Methods and Applications. Springer International Publishing; 2024:313-338. [doi: 10.1007/978-3-031-50690-1_13]
35. Sparck Jones K. A statistical interpretation of term specificity and its application in retrieval. J Doc. Jan 1972;28(1):11-21. [doi: 10.1108/eb026526]
36. Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Inf Process Manag. 1988;24(5):513-523. [doi: 10.1016/0306-4573(88)90021-0]
37. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc. 2010;17(3):229-236. [doi: 10.1136/jamia.2009.002733] [Medline: 20442139]
38. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. Jan 1, 2004;32(Database issue):D267-D270. [doi: 10.1093/nar/gkh061] [Medline: 14681409]
39. Uphaus T, Weber-Krüger M, Grond M, et al. Development and validation of a score to detect paroxysmal atrial fibrillation after stroke. Neurology. Jan 8, 2019;92(2):e115-e124. [doi: 10.1212/WNL.0000000000006727] [Medline: 30530796]
40. Zheng X, Wang F, Zhang J, et al. Using machine learning to predict atrial fibrillation diagnosed after ischemic stroke. Int J Cardiol. Jan 15, 2022;347:21-27. [doi: 10.1016/j.ijcard.2021.11.005] [Medline: 34774886]
41. Sung SF, Sung KL, Pan RC, Lee PJ, Hu YH. Automated risk assessment of newly detected atrial fibrillation poststroke from electronic health record data using machine learning and natural language processing. Front Cardiovasc Med. Jul 29, 2022;9:941237. [doi: 10.3389/fcvm.2022.941237] [Medline: 35966534]

42. Hsieh CY, Lee CH, Sung SF. Development of a novel score to predict newly diagnosed atrial fibrillation after ischemic stroke: the CHASE-LESS score. Atherosclerosis. Feb 2020;295:1-7. [doi: 10.1016/j.atherosclerosis.2020.01.003] [Medline: 31972497]
43. Ratajczak-Tretel B, Lambert AT, Al-Ani R, et al. Prediction of underlying atrial fibrillation in patients with a cryptogenic stroke: results from the NOR-FIB Study. J Neurol. Aug 2023;270(8):4049-4059. [doi: 10.1007/s00415-023-11680-8] [Medline: 37162578]

Abbreviations

AF: atrial fibrillation
AIS: acute ischemic stroke
AS5F: age, stroke severity National Institutes of Health Stroke Scale>5 to find atrial fibrillation
AUC: area under the receiver operating characteristic curve
BERT: bidirectional encoder representations from transformers
CatBoost: categorical boosting
CHASE-LESS: coronary, heart failure, age, stroke severity–lipidemia, sugar, and prior stroke
CYCH: Chia-Yi Christian Hospital
ECG: electrocardiogram
EMR: electronic medical record
LightGBM: light gradient-boosting machine
LIH: Landseed International Hospital
NIHSS: National Institutes of Health Stroke Scale
SHAP: Shapley additive explanations
SMOTE: synthetic minority oversampling technique
TF-IDF: term frequency–inverse document frequency
TIA: transient ischemic attack
XGBoost: extreme gradient boosting

Edited by Arriel Benis; peer-reviewed by Pei-Ju Lee, Szu-Yin Lin; submitted 02.Jun.2025; final revised version received 10.Jan.2026; accepted 19.Jan.2026; published 10.Mar.2026

Please cite as:
Chen YW, Sung SF, Hu YH, Yang YH
Enhanced Prediction of Atrial Fibrillation in Patients With Ischemic Stroke Through Electronic Medical Records and Text Mining: Algorithm Development and Validation
JMIR Med Inform 2026;14:e78117
URL: https://medinform.jmir.org/2026/1/e78117
doi: 10.2196/78117

© Yu-Wei Chen, Sheng-Feng Sung, Ya-Han Hu, Yu-Hsuan Yang. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 10.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.