Predicting the risk of ischemic stroke in patients with atrial fibrillation using heterogeneous drug–protein–disease network-based deep learning

Zhiheng Lyu,[1,2] Jiannan Yang,[3] Zhongzhi Xu,[1,a)] Weilan Wang,[4] Weibin Cheng,[2,5] Kwok-Leung Tsui,[6] and Qingpeng Zhang[3,7,8,a)]

AFFILIATIONS

[1]School of Public Health, Sun Yat-sen University, Guangzhou, China
[2]Department of Data Science, City University of Hong KongChina
[3]Musketeers Foundation Institute of Data Science, The University of Hong Kong, Hong Kong SAR, China
[4]Centre for Healthy Longevity, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
[5]Institute for Healthcare Artificial Intelligence, Guangdong Second Provincial General Hospital, Guangzhou, China
[6]Department of Manufacturing, Systems, and Industrial Engineering, University of Texas, Arlington, Texas 76019, USA;
[7]Department of Pharmacology and Pharmacy, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China
[8]HKU Shanghai Intelligent Computing Research Center, Shanghai, China

[a)]Authors to whom correspondence should be addressed: xuzhzh26@mail.sysu.edu.cn and qpzhang@hku.hk

ABSTRACT

Current risk assessment models for predicting ischemic stroke (IS) in patients with atrial fibrillation (AF) often fail to account for the effects of medications and the complex interactions between drugs, proteins, and diseases. We developed an interpretable deep learning model, the AF-Biological-IS-Path (ABioSPath), to predict one-year IS risk in AF patients by integrating drug–protein–disease pathways with real-world clinical data. Using a heterogeneous multilayer network, ABioSPath identifies mechanisms of drug actions and the propagation of comorbid diseases. By combining mechanistic pathways with patient-specific characteristics, the model provides individualized IS risk assessments and identifies potential molecular pathways involved. We utilized the electronic health record data from 7859 AF patients, collected between January 2008 and December 2009 across 43 hospitals in Hong Kong. ABioSPath outperformed baseline models in all evaluation metrics, achieving an AUROC of 0.7815 (95% CI: 0.7346–0.8283), a positive predictive value of 0.430, a negative predictive value of 0.870, a sensitivity of 0.500, a specificity of 0.885, an average precision of 0.409, and a Brier score of 0.195. Cohort-level analysis identified key proteins, such as CRP, REN, and PTGS2, within the most common pathways. Individual-level analysis further highlighted the importance of PIK3/Akt and cytokine and chemokine signaling pathways and identified IS risks associated with less-studied drugs like prochlorperazine maleate. ABioSPath offers a robust, data-driven approach for IS risk prediction, requiring only routinely collected clinical data without the need for costly biomarkers. Beyond IS, the model has potential applications in screening risks for other diseases, enhancing patient care, and providing insights for drug development.

I. INTRODUCTION

Atrial fibrillation (AF) is the most commonly occurring cardiac arrhythmia, impacting an estimated 1%–2% of the global population.[1] Among the most severe complications of AF are ischemic stroke (IS) and systemic embolism, which can occur regardless of the specific type or pattern of AF. IS is a major cause of death globally, often leading to severe outcomes such as vision loss, speech impairment, paralysis, and cognitive deficits. Approximately 15×10^6 people globally experience strokes each year, leading to 5×10^6 fatalities and 5×10^6 instances of permanent disability.[2] The burden of stroke extends beyond patients and their families, posing significant social and economic challenges.

Oral anticoagulation therapy has been the primary pharmacological approach for reducing stroke risk in AF patients, both diagnostically and prognostically, but it raises the risk of major bleeding.

Fortunately, evidence suggests that IS is not an inevitable consequence of aging if risks are accurately predicted and managed in time.[3] Retrospective and prospective studies have identified modifiable risk factors for IS, which can reduce stroke risk when properly managed.[3,4] In clinical practice, the CHADS$_2$ and CHA$_2$DS$_2$-VASc scoring systems guide anticoagulation decisions in AF patients.[5] However, these systems have shown only moderate predictive accuracy, often resulting in overtreatment or undertreatment.[6]

Two major gaps contribute to the limitations in predicting IS risk in AF patients: (a) The co-occurrence of diseases can influence stroke risk through a complex network of shared biological pathways and multifunctional proteins.[7] Key risk factors for stroke include hypertension, diabetes, coronary heart disease, AF, heart valve disease, and carotid artery disease,[8] yet the biological mechanisms underlying these associations remain poorly understood;[9] and (b) medications taken by patients can affect stroke development, either positively or negatively. For instance, while statins are primarily prescribed as lipid-lowering agents, they have been linked to better outcomes post-IS in AF patients, likely due to their pleiotropic effects on the cardiovascular system beyond lipid control.[10] Conversely, drugs targeting the IL-6 protein pathway may increase IS mortality risk,[11] and certain HIV treatments have been linked to a higher risk of IS.[12]

These gaps underscore the necessity for research that systematically integrates heterogeneous drug–protein–disease interactions to enhance IS risk prediction. To address this need, we constructed a multilayer network incorporating existing knowledge of disease–protein–drug interactions. Utilizing a comprehensive Hong Kong-based cohort focused on an Asian population, we developed a deep learning model called **AF-Bio**logical-**IS-Path** (ABioSPath) to predict IS risk in AF patients. Unlike traditional studies that analyze comorbidities and medications in isolation, our approach systematically integrates drug–protein and disease–protein interactions, along with protein–protein and disease–disease relationships, bridging previously siloed data.

This integration provides a foundation for both IS risk prediction and modeling of other diseases, enabling exploration of novel molecular pathways in disease progression. We propose a graph attention network (GAT)-based model that leverages this multilayer network and large-scale, real-world inpatient data, facilitating robust linkage between network information and patient data with multiple comorbidities and prescriptions. Our findings demonstrate that this framework significantly outperforms traditional risk models while offering interpretable insights at both cohort and individual levels. At the cohort level, proteins such as CRP, REN, and PTGS2 emerged as central factors associated with IS, while at the individual level, pathways including PI3K/Akt signaling and cytokine-/chemokine-induced signaling proved critical. These discoveries validate our approach and establish new directions for integrative methods in stroke risk prediction.

II. RESULTS

The characteristics of the research samples are summarized in Table I. The mean age of the study population was 79.82 years, with ages ranging from 65 to 107 years. Of the study population, 56.5% were male. On average, each patient had 4.11 prescriptions during the study period. Additionally, the average number of diagnoses per patient throughout the study period was 6.98.

TABLE I. Demographic and clinical characteristics of 7856 AF patients. Age, number of prescriptions, and diagnoses are presented as mean ± standard deviation, median [interquartile range], and range. Sex and comorbidities (identified by ICD-9 codes) are shown as counts (percentages).

Characteristic	Value
Demographics	
Age (years)	79.82 ± 7.25; 80.0 [75–85]; range 65–107
Male sex, n (%)	4,439 (56.5)
Clinical characteristics	
Number of prescriptions (CID codes)	4.12 ± 2.87; 3.0 [2.0–6.0]; range 1–23
Number of diagnoses (ICD-9 codes)	6.98 ± 3.67; 6.0 [4.0–9.0]; range 1–35
Comorbidities, n (%)	
Essential hypertension (ICD-9)	4352 (55.4)
Diabetes mellitus (ICD-9)	2175 (27.7)

A. Model performance

The performance of the ABioSPath model compared to baseline models is detailed in Table II. Overall, ABioSPath outperformed all baseline models across evaluation metrics. In the test set, ABioSPath achieved an AUROC of 0.782 (95% CI: 0.735–0.829), with sensitivity of 0.500, specificity of 0.885, PPV of 0.430, NPV of 0.870, average precision of 0.409, and a Brier score of 0.195. Among the baseline models, logistic regression performed best in testing (AUROC 0.725, 95% CI: 0.666–0.782), followed by LASSO (AUROC 0.697, 95% CI: 0.642–0.748), while CHADS$_2$ and CHA$_2$DS$_2$-VASc showed lower discriminative ability. The DeLong test demonstrated statistically significant differences (p < 0.01) between ABioSPath and all baseline models. To assess the impact of incomplete data, we conducted a simulation study where we randomly removed 50% of the information from patients with multiple records. The results, presented in Table SIII in the supplementary material, demonstrate a significant decrease in model performance, with the AUROC dropping to 0.684 on the test set.

B. Identifying IS risk in patients using specific medications

Antiplatelet drugs and ACE inhibitors are recognized for their ability to reduce the risk of ischemic stroke.[13] In our dataset, we focus on aspirin and ACE inhibitors as the effective medical therapies available. However, despite the use of these treatments, the risk of ischemic stroke persists. To investigate this further, we identified a subset of 6072 patients who were prescribed both antiplatelet medications (including aspirin, clopidogrel, and dipyridamole) and ACE inhibitors (such as lisinopril, ramipril, enalapril, captopril, fosinopril sodium, and perindopril tert-butylamine).

Table III presents the performance of the ABioSPath model and baseline models on this subset. In assessing residual IS risk within this group, ABioSPath demonstrated superior performance compared to other methods, achieving an AUROC of 0.8067, a PPV of 0.504, an NPV of 0.893, a sensitivity of 0.528, a specificity of 0.883, an average precision of 0.470, and a Brier score of 0.186. In contrast, the CHADS$_2$

TABLE II. Comparative performance analysis of IS risk prediction models. Model performance metrics evaluated across training, validation, and test sets include area under receiver operating characteristic curve (AUROC) with 95% confidence intervals, positive predictive value (PPV), negative predictive value (NPV), sensitivity, specificity, average precision, Brier score, and DeLong test p-values (compared to ABioSPath). Models compared: ABioSPath, logistic regression, LASSO, $CHADS_2$, and CHA_2DS_2-VASc. Bold values represent each model's results from the test set.

Model	Cohort	AUROC	PPV	NPV	Sensitivity	Specificity	Average precision	Delong test p-value	Brier score
Logistic regression	Train	0.745 (0.727, 0.762)	0.449	0.896	0.499	0.876	0.391	<0.01	0.213
	Valid	0.735 (0.704, 0.766)	0.473	0.899	0.527	0.879	0.398	<0.05	0.213
	Test	**0.725 (0.666, 0.782)**	**0.366**	**0.905**	**0.495**	**0.855**	**0.339**	**<0.05**	**0.221**
LASSO	Train	0.704 (0.684, 0.723)	0.463	0.881	0.500	0.869	0.325	<0.01	0.292
	Valid	0.726 (0.692, 0.760)	0.473	0.880	0.527	0.870	0.344	<0.01	0.280
	Test	**0.697 (0.642, 0.748)**	**0.343**	**0.881**	**0.426**	**0.873**	**0.254**	**<0.01**	**0.304**
$CHADS_2$	Train	0.667 (0.647, 0.685)	0.383	0.840	0.366	0.840	0.262	<0.01	0.231
	Valid	0.668 (0.633, 0.703)	0.400	0.844	0.375	0.824	0.266	<0.01	0.230
	Test	**0.675 (0.625, 0.725)**	**0.302**	**0.862**	**0.322**	**0.837**	**0.231**	**<0.01**	**0.224**
CHA_2DS_2-VASc	Train	0.661 (0.642, 0.680)	0.332	0.841	0.314	0.862	0.255	<0.01	0.232
	Valid	0.658 (0.623, 0.693)	0.350	0.840	0.327	0.852	0.255	<0.01	0.232
	Test	**0.673 (0.624, 0.722)**	**0.311**	**0.863**	**0.287**	**0.866**	**0.221**	**<0.01**	**0.220**
ABioSPath	Train	0.813 (0.798, 0.827)	0.503	0.898	0.500	0.900	0.451		0.183
	Valid	0.772 (0.737, 0.807)	0.479	0.900	0.530	0.882	0.447		0.196
	Test	**0.781 (0.735, 0.829)**	**0.421**	**0.910**	**0.500**	**0.885**	**0.410**		**0.195**

TABLE III. Model performance metrics. Performance metrics were calculated on 6072 patients using antiplatelet and ACE inhibitor medications. Computed model values for area under receiver operating characteristic curve (AUROC) (with 95% confidence intervals), positive predictive value (PPV), negative predictive value (NPV), sensitivity, specificity, average precision, DeLong test p-values, and Brier score are listed in the table. Bold values indicate best performance.

Model	AUROC	PPV	NPV	Sensitivity	Specificity	Average precision	Delong test p-value	Brier score
Logistic regression	0.748 (0.731,0.765)	0.463	0.890	0.528	0.863	0.400	<0.01	0.213
LASSO	0.692 (0.676,0.708)	0.454	0.877	0.528	0.858	0.325	<0.01	0.293
$CHADS_2$	0.667 (0.650,0.685)	0.388	0.830	0.383	0.824	0.279	<0.01	0.233
CHA_2DS_2-VASc	0.661 (0.643,0.678)	0.352	0.828	0.334	0.846	0.271	<0.01	0.234
ABioSPath	**0.807 (0.794,0.822)**	**0.504**	**0.893**	**0.528**	**0.883**	**0.470**		**0.186**

and CHA_2DS_2-VASc models produced the poorest results, followed by LASSO and then logistic regression.

C. Assessing model robustness for patients not receiving specific therapies

To assess ABioSPath's broader clinical utility, we evaluated its performance in 1787 patients not receiving antiplatelet drugs or ACE inhibitors. This analysis is crucial as real-world settings often involve incomplete guideline adherence or limited treatment access. The evaluation aimed to verify the model's generalizability across different therapeutic regimens and its ability to predict IS risk under varied treatment conditions.

Table IV shows ABioSPath's performance in patients without specific therapies, achieving an AUROC of 0.751. Among comparators, logistic regression performed highest (AUROC 0.662),

followed by CHA_2DS_2-VASc (0.640), $CHADS_2$ (0.639), and LASSO (0.624). The model demonstrated balanced metrics: sensitivity (0.606), specificity (0.758), PPV (0.237), and NPV (0.938). It also showed superior calibration with the highest average precision (0.287) and favorable Brier score (0.190). DeLong tests confirmed significant performance differences between model pairs (p < 0.01), validating ABioSPath's advantage in patients without using antiplatelet drugs and ACE inhibitors.

D. Identified pathway results for drugs

ABioSPath incorporates a deep learning approach to understanding the mechanisms of drug actions (MODA) and disease comorbidity propagations (DCP) in disease risk prediction.[14] Among the identified pathways for all patients, the ten most important pathways for each patient were selected based on their ranked weights. The aggregated

TABLE IV. Model performance metrics in patients without antiplatelet or ACE inhibitor therapy. Performance metrics of prediction models in patients not receiving antiplatelet drugs or ACE inhibitors. Performance evaluation includes for area under receiver operating characteristic curve (AUROC) (with 95% confidence intervals), positive predictive value (PPV), negative predictive value (NPV) sensitivity, specificity, average precision, DeLong test p-values (vs ABioSPath), and Brier score. Bold values indicate best performance.

Model	AUROC	PPV	NPV	Sensitivity	Specificity	Average precision	Delong test p-value	Brier score
Logistic regression	0.662 (0.616,0.709)	0.178	0.928	0.601	0.658	0.240	<0.01	0.220
LASSO	0.624 (0.590,0.658)	0	0.889	0.338	0.910	0.182	<0.01	0.278
$CHADS_2$	0.639 (0.598,0.680)	0.211	0.905	0.560	0.661	0.170	<0.01	0.220
CHA_2DS_2-VASc	0.640 (0.600,0.680)	0.213	0.901	0.469	0.723	0.162	<0.01	0.216
ABioSPath	**0.751 (0.715,0.787)**	**0.237**	**0.938**	**0.606**	**0.758**	**0.287**		**0.190**

pathways from all patients revealed the ten most frequently occurring ones, as shown in Fig. 1(a). Out of the ten identified pathways, four originated from CID-2244 (aspirin), three from CID-5362119 (lisinopril), two from CID-3440 (furosemide), and one from CID-4510 (nitroglycerin). The key proteins identified in these pathways included Entrez-1636 (ACE), Entrez-5743 (PTGS2), Entrez-59272 (ACE2), Entrez-476 (RNA5SP476), Entrez-5972 (REN), Entrez-4843 (NOS2), Entrez-1559 (CYP2C9), Entrez-1401 (CRP), and Entrez-4311 (MME). Among the pathways extracted from each patient, the ten most prevalent pathways across the entire cohort had counts of 1500, 1282, 1243, 1196, 1173, 990, 982, 942, 902, and 894. After extracting the top twenty important pathways from each patient, the mean occurrence frequency of all extracted pathways was 19.229 (SD = 66.26). One-tailed, one-sample t-tests confirmed that these top ten pathway counts were significantly higher than the overall mean occurrence frequency (p < 0.01), indicating their consistent prevalence across the patient cohort.

Figure 1(b) illustrates the drug pathways identified for an individual patient as an example. This patient had a medical history of ICD9-250 (diabetes mellitus), ICD9-272 (disorders of lipoid metabolism), ICD9-401 (essential hypertension), ICD9-427 (cardiac dysrhythmias), ICD9-428 (heart failure), ICD9-438 (late effects of cerebrovascular disease), and ICD9-496 (chronic airway obstruction, not elsewhere classified). The medications taken included CID-4917 (prochlorperazine maleate), CID-3746 (ipratropium bromide), CID-3108 (dipyridamole), CID-2244 (aspirin), CID-2153 (theophylline), CID-2083 (salbutamol (sulphate), CID-5362119 (lisinopril), CID-38853 (methyldopa), CID-5754 (hydrocortisone), and CID-39186 (diltiazem). The patient was later diagnosed with IS within a one-year follow-up.

The twenty most significant pathways from drugs to stroke were identified based on ranked weights, as detailed in Fig. 1(b). Among these pathways, eight originated from CID-2153 (theophylline), five from CID-4917 (prochlorperazine maleate), three from CID-2083 (salbutamol), two from CID-38853 (methyldopa), one from CID-2244 (aspirin), and one from CID-39186 (diltiazem). The pathways ranged in length from three to four, with fifty-two proteins appearing in these pathways. The five most frequently occurring proteins were Entrez-5290 (PIK3CA), Entrez-7124 (TNF), Entrez-1129 (CHRM2), Entrez-4804 (NGFR), and Entrez-153 (ADRB1), with frequencies ranging from one to five.

III. DISCUSSIONS

In this study, we initially constructed a multilayer network that integrates existing knowledge about disease–protein–drug interactions. We then developed a deep learning model, ABioSPath, utilizing this multilayer network to enhance the prediction of IS risk in individuals with AF. Experiments conducted on a cohort of 7859 AF inpatients demonstrated that ABioSPath improved the AUROC by 8% compared to a standard logistic regression model that independently considered prescriptions and historical diagnoses.

In subset analyses, our model outperformed existing clinical standards, which often overlook drug usage, by accurately identifying high-risk patients even under treatments considered optimal by current standards. This capability allows ABioSPath to effectively pinpoint patients with residual risk. Additionally, ABioSPath provided insights into the potential mechanistic pathways that contribute to its predictive performance. The model's high NPV indicates its accuracy in identifying patients with low stroke risk, which can help clinicians avoid unnecessary medications and associated risks, such as internal bleeding.

ABioSPath demonstrated robust performance in patients not receiving specific therapies, achieving an AUROC of 0.751 (95% CI: 0.715–0.787) in this challenging subgroup, as shown in Table IV. While performance metrics were modestly lower compared to patients on specific therapies, ABioSPath consistently outperformed both traditional risk scores and conventional machine learning approaches. The performance difference between medicated and non-medicated patients underscores the importance of incorporating medication data into the model. Notably, even with this performance gap, ABioSPath maintained superior predictive accuracy compared to existing clinical risk scores, demonstrating its reliability as a risk assessment tool regardless of treatment status.

The integration of heterogeneous biological networks enables ABioSPath to model disease risk propagation by utilizing rich drug–protein–disease interactions from a novel perspective. Employing GAT and Bi-LSTM networks, the model achieves end-to-end prediction while integrating complex network data—a capability beyond the reach of traditional machine learning algorithms. Incorporating both node and path attention mechanisms allows the model to identify significant biological pathways, distinguishing critical pathways more effectively than conventional methods. This approach provides detailed insights into underlying pathways, enabling clinicians to make
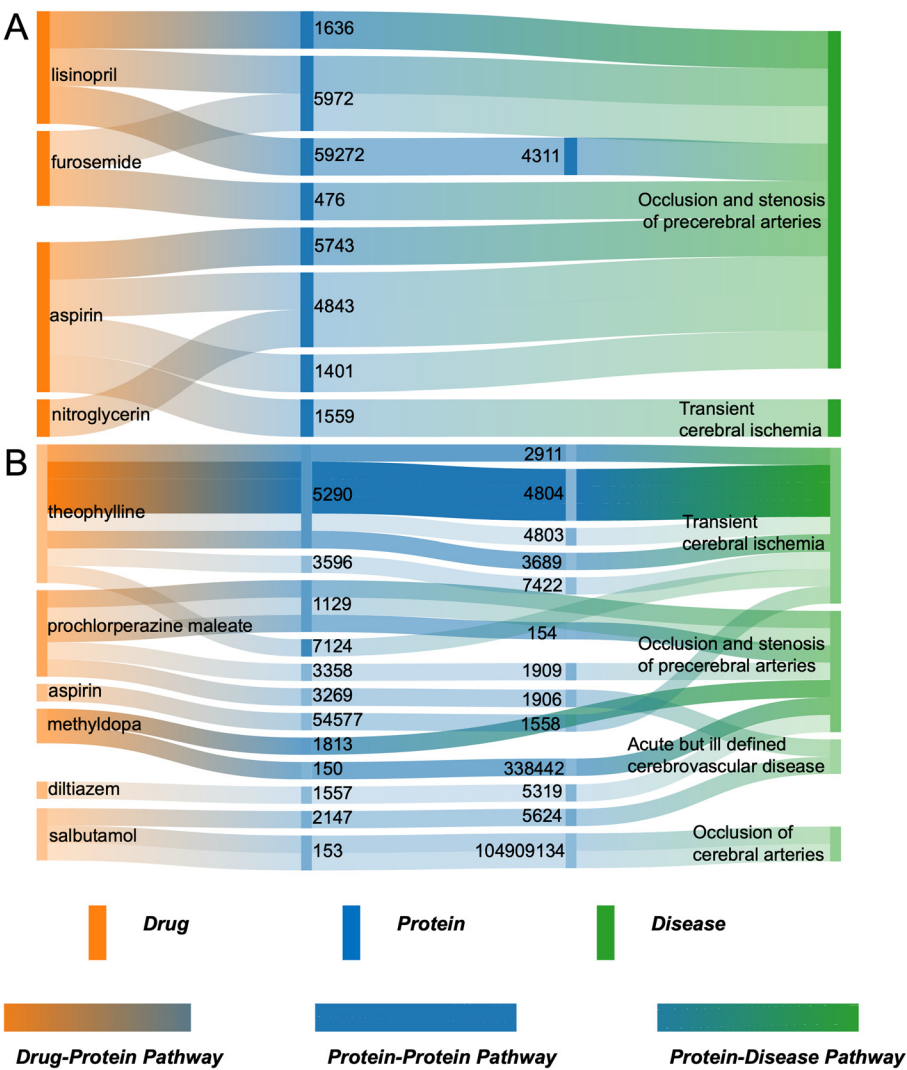
FIG. 1. Protein-mediated pathways linking medications to ischemic stroke. (a) Sankey diagram illustrating the ten most prevalent protein pathways connecting medications to ischemic stroke across the study cohort. The width and color intensity of connections represent pathway occurrence and association strength, respectively. (b) Patient-specific pathway analysis demonstrating the mechanistic relationships between prescribed medications and stroke outcome. Node classification: medications (orange), proteins (blue), and diseases (green). Node size indicates pathway occurrence, while color intensity reflects association strength. Link thickness and darkness correspond to connection strength between network elements.

more informed medication adjustments based on deeper biological understanding. Understanding the rationale behind predictions is crucial in healthcare, as it aids clinicians in decision-making. Our model tracks the shortest paths within a complex, multilayered network, from specific drugs to IS. We used the cuGraph shortest path search function—a GPU-accelerated graph analytics package—to expedite the process, given the network's size.[15] ABioSPath identifies pathways that rationalize risk predictions, highlighting the significance of protein–protein interactions in predicting drug effects and target protein functions at the molecular level. In drug discovery, such interactions are increasingly critical, offering therapeutic potential for targeting specific disease mechanisms. Compared to baseline models, ABioSPath excels in identifying critical pathways from drugs to diseases. These identified pathways, validated through existing evidence, bolster the model's credibility at both the cohort and individual levels.

Figure 1(a) illustrates significant pathways identified by ABioSPath, highlighting drugs such as aspirin, lisinopril, furosemide, and nitroglycerin that have a greater impact on IS risk compared to other medications. Lisinopril, an ACE inhibitor, promotes vascular relaxation and dilation, lowering blood pressure.[16] Research has shown a connection between ACE inhibitors and vascular diseases, including IS.[17] REN is a key component of the renin-angiotensin system, which plays a vital role in regulating blood pressure and maintaining fluid balance.[18] Aspirin is widely used to prevent recurrent IS.[19] Numerous studies have suggested that PTGS2 may serve as a critical risk factor in IS, with genetic variation in PTGS2 potentially contributing to the progression of cardiovascular events.[20] Aspirin use may alter the association between the PTGS2 G-765C polymorphism and coronary heart disease (CHD) risk.[21] The PTGS2 G765C mutation may also be associated with aspirin resistance.[22] Additionally, aspirin can influence the activity of nitric oxide synthase (NOS), including NOS2.[22] NOS2 has been identified as a stroke-related gene.[23] C-reactive protein (CRP) is also an important risk factor for IS.[24] In AF patients, CRP levels correlate with stroke risk and other cardiovascular diseases, serving as a key biomarker for IS risk assessment.[24,25] Aspirin's therapeutic efficacy has been observed in individuals with high CRP levels.[24] Evidence further

supports a connection between furosemide, nitroglycerin, and IS via identified pathways.[26]

In a specific case study, Fig. 1(b) reveals that among the top 20 significant pathways, theophylline and prochlorperazine maleate had the most substantial impact on the patient's IS risk, as determined by ABioSPath. The relationship between theophylline and IS is complex and multifaceted. Our pathway analysis suggests that theophylline is associated with IS through proteins such as phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha (PIK3CA) and nerve growth factor receptor (NGFR). While theophylline is commonly prescribed for asthma and chronic obstructive pulmonary disease (COPD),[27] its association with IS remains under-explored. However, some evidence indicates that theophylline may have a neuroprotective effect in IS.[28] The PI3K/Akt pathway, inhibited by theophylline, is linked to the development of diseases such as stroke, with PIK3CA, an isoform of PI3K, potentially playing a role.[29,30] NGFR is also involved in the development of IS, with studies suggesting it could be a risk factor.[31] Tumor necrosis factor (TNF) has been identified as a potential stroke marker due to its role in stroke pathophysiology and involvement in cytokine- and chemokine-induced signaling pathways in neuroinflammation.[30,32]

Other drugs, such as aspirin, identified by ABioSPath, are well-known for their role in preventing blood clots, a leading cause of IS. Aspirin significantly reduces clot density and strength.[33] Similarly, methyldopa may decrease IS risk through its antihypertensive effects.[34] Another drug associated with IS risk is prochlorperazine maleate. Although current research has not definitively established a relationship between prochlorperazine maleate and IS, clinical case reports suggest a potential connection.[35] Our model's findings align with this observation, indicating that prochlorperazine maleate might increase IS risk in this patient. This conclusion is based on a rise in the risk score from 0.559 to 0.566 following drug administration. Thus, our model can identify potential risks associated with less-studied IS-related drugs, such as prochlorperazine maleate. While the pathways identified by ABioSPath were algorithmically derived rather than clinically observed, they align with existing research, suggesting that some under-explored pathways may present new targets for IS prevention.

Cohort-level analyses reveal the occurrence and strength of pathways linking drugs to diseases, with statistical testing confirming the significance of ten key mechanistic pathways. In contrast, individual-level analyses offer insights into unique drug interactions and relations, highlighting that different drugs may exert varying influences on specific diseases. This information highlights the distinction between cohort-level and individual-level analyses. At the cohort level, certain drugs and proteins may exert a more significant impact, guiding therapy plans that are broadly applicable to the entire group. In contrast, individual-level analyses leverage patient-specific information, uncovering unique drug–disease interactions that enable tailored therapeutic strategies. This difference is further illustrated in our example. Despite the patient taking common medications such as aspirin, similar to the cohort, they also take uncommon drugs like prochlorperazine maleate. These uncommon medications introduce unique physiological characteristics and drug interactions, resulting in different pathway outcomes. This discrepancy emphasizes the value of personalized analysis in clinical decision-making. This distinction underscores the strength of our model in integrating complex relationships and delivering personalized predictions, representing a significant advancement in predictive modeling.

We evaluated our model using the ProBast checklist, ensuring that the design and validation process minimized bias.[36] First, in terms of the dataset, the model utilizes comprehensive, territory-wide data from all public hospitals in Hong Kong, reflecting the overall population status in the region. This approach mitigates representation bias compared to datasets sourced from a limited number of hospitals, thereby enhancing the model's generalizability. Second, our methodology is rigorous, employing strict criteria in developing and validating the model, including random sampling and internal validation through a separate dataset. This approach minimizes the risk of overfitting and improves the model's robustness. Third, the model underwent a thorough evaluation, demonstrating superior performance in calibration, discrimination, and classification. The DeLong test results further support our model's effectiveness. Finally, our model ensures the timeliness and relevance of predictors, as all are derived from historical EHR data and open-source datasets, thus avoiding look-ahead bias. The 1-year prediction horizon was strategically chosen to balance clinical utility and temporal constraints, ensuring practical applicability in clinical settings.

While ABioSPath demonstrates superior predictive performance by leveraging routinely collected EHR data and drug–protein interaction networks, we acknowledge important practical considerations for clinical implementation. Traditional scoring systems like $CHADS_2$ and CHA_2DS_2-VASc offer the key advantage of rapid manual calculation at the point of care without requiring computational resources or complex interfaces. This allows physicians to quickly stratify risk during time-sensitive clinical encounters. Please refer to Sec. IV in the supplementary material for detailed comparisons between ABioSPath and other existing models for predicting IS risk.

The ABioSPath model aims to maintain clinical utility while providing deeper mechanistic insights. It can be integrated into existing EHR systems to automatically calculate IS risk using standard clinical data inputs like disease history, medications, demographics, and lab values. An intuitive interface will allow clinicians to review personalized risk assessments, key contributing factors, and relevant molecular pathways—capabilities unavailable in current scoring systems. However, successful implementation will require proper IT infrastructure and workflow integration to ensure the model enhances clinical decision-making.

This study has several important limitations that should be acknowledged. First, our analysis was based exclusively on patients from Hong Kong, which may limit the generalizability of our findings to other regions. It is essential to validate our model with data from patients across diverse geographic areas to identify potential variations and ensure broader applicability. Second, the study did not account for the temporal sequence of disease onset and prescriptions. While integrating the chronological order of diseases and the corresponding treatments could introduce additional complexity and noise, it may also reveal valuable insights, patterns, and relationships that would otherwise remain hidden. Future research should aim to incorporate temporal information to gain a more comprehensive understanding of disease progression and treatment effects. Third, the weighting of the knowledge graph used in our study presents a complex challenge that requires further exploration. Currently, fixed weights are assigned to the connections between drugs and proteins, without considering the

relative strength of these associations. These weights provide valuable information about the strength of relationships between entities. To enhance the accuracy and richness of our model, future work will focus on assigning weights based on the significance and relevance of different chemicals. Fourth, despite incorporating as many predictors as possible, there are still potential confounders that may have been overlooked or are difficult to quantify. Fifth, while the adoption of electronic health record (EHR) systems has increased data availability, access remains limited in underdeveloped regions, potentially affecting our model's performance. Sixth, while shortest paths often capture primary mechanisms, our approach may overlook complex feedback loops and secondary pathways that could contribute to drug effects and disease progression. Further research and analysis are needed to address these gaps. Refining the model to include more comprehensive information could improve the knowledge graph's quality and the model's overall performance.

IV. CONCLUSION

This study presents the development of ABioSPath, a deep learning model designed to predict the risk of ischemic stroke (IS) in patients with atrial fibrillation (AF). The model demonstrated superior predictive performance compared to traditional machine learning models and clinical risk scores. Its use of routinely collected electronic health records (EHRs), which eliminates the need for expensive biomarker testing, makes it a practical option for widespread clinical application. Additionally, the model's ability to pinpoint specific pathways associated with IS risk offers new opportunities for targeted prevention strategies and personalized care.

V. METHODS

A. Data

The data used in this study comprises two main parts. The first part consists of electronic health records (EHRs) from inpatients, collected from the Hong Kong Hospital Authority (HA), which oversees all 43 public hospitals in Hong Kong. For each hospital admission, the EHRs recorded detailed patient information, including a unique patient identifier, gender, age, up to 15 diagnoses coded by ICD-9-CM, admission and discharge dates, and prescribed medications. For each medication, additional details were recorded: prescription date, British National Formulary code, drug name, frequency, dosage (value and unit), quantity, type, and duration of treatment. The dataset spans three years, from January 1, 2008, to December 31, 2010, and includes over 5.2×10^6 health records corresponding to 1 764 094 inpatients.

The second part of the dataset is a three-layer drug–protein–disease network, derived from iDPath and the comorbidity network developed in our previous research, as illustrated in Fig. 2.[14,37]

1. Disease comorbidity network

The disease comorbidity network (DCN) was constructed using seven years of electronic health record (EHR) data (2000–2007) from the Hong Kong Hospital Authority. For each hospital admission, up to 15 disease diagnoses were recorded, focusing exclusively on primary and secondary admissions. The resulting comorbidity network is represented as an undirected graph, where each node corresponds to a disease and the weighted edges indicate the co-occurrence frequency of disease pairs. To minimize noise, co-occurrences with frequencies lower than 10 were excluded from the network. All diseases in the comorbidity network are encoded using the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) codes.

2. Disease–protein interaction dataset

The disease–protein interaction dataset was developed by linking diseases to protein-coding genes. It includes associations between diseases and both genes and genetic variants, sourced from DisGeNET's expert-reviewed repositories.[38] Diseases are identified by ICD-9-CM codes, and proteins are represented by corresponding genes using Entrez Gene IDs.

3. Protein–protein interactions

The protein–protein interaction network (PPIN) consists of two key data sources. The first is the STRING database,[39] selected for human-specific interactions with high confidence (score > 0.7). The second source is from Cheng et al.,[40] which aggregates experimentally supported interactions from various databases. Proteins in this network are identified by their corresponding Entrez Gene IDs.

4. Drug–protein interactions

The drug–protein interactions dataset integrates multiple sources, including the PCI network in STITCH,[41] data from Cheng et al.,[40] the therapeutic target database (TTD),[42] and DrugBank.[43] STITCH provides comprehensive information on chemical–protein interactions, while TTD focuses on therapeutic targets and associated diseases. DrugBank offers a detailed database on drugs and their targets. In this dataset, drugs are identified by PubChem CIDs, and proteins by their Entrez Gene IDs.

The three-layer network is composed of the disease comorbidity network (DCN), the protein–protein interaction network (PPIN), and the drug network (DN), as illustrated in Fig. 2. The DCN provides insights into how diseases may progress from one to another, illustrating patterns of comorbidity. The PPIN reveals the interactions between proteins, mapping the complex web of protein relationships. In the DN, there are no direct links between drugs; instead, connections between drugs and proteins are mediated through drug–protein interactions. Additionally, interlayer links connect the DCN and PPIN, highlighting disease–protein relationships through the disease–protein interaction dataset. The network comprises 13 758 nodes in the PPIN, 1077 nodes in the DCN, and 27 230 nodes in the DN.

B. Sample selection

The sample selection criteria are detailed in Fig. 3. Patients with AF were identified using the ICD-9-CM code 427.31. For IS, the ICD-9-CM codes 433 (occlusion and stenosis of precerebral arteries), 434 (occlusion of cerebral arteries), 435 (transient ischemic attack), and 436 (acute, but ill-defined, cerebrovascular disease) were used.[44] During the study period, 7859 patients were diagnosed with AF, of whom 1309 (25%) developed IS within 12 months. These patients were included in the analysis as positive cases. All prescriptions prior to the first occurrence of IS and diagnoses preceding the last diagnosis of AF were included in the analysis.

We acknowledge that the observed stroke rate in our study was relatively high compared to other studies. Several factors may account
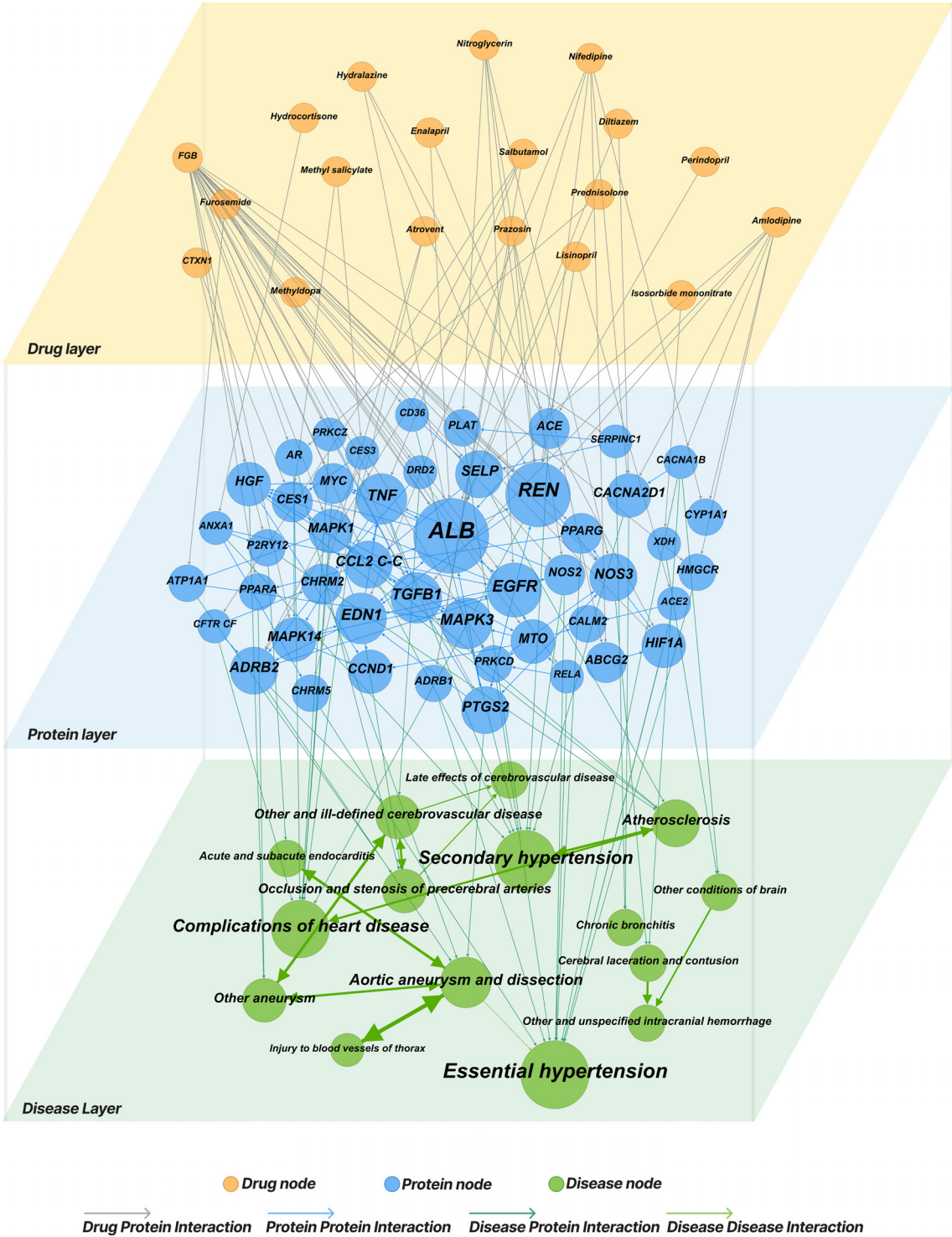
FIG. 2. Three-layer network visualization of drug–protein–disease interactions. Network diagram showing the relationships between selected nodes across three layers: drugs (top layer, 19/27 230 nodes), proteins (middle layer, 137/13 758 nodes), and diseases (bottom layer, 53/1077 nodes). Node colors indicate layer type (orange: drug, blue: protein, and green: disease) with connecting lines representing interactions between layers.
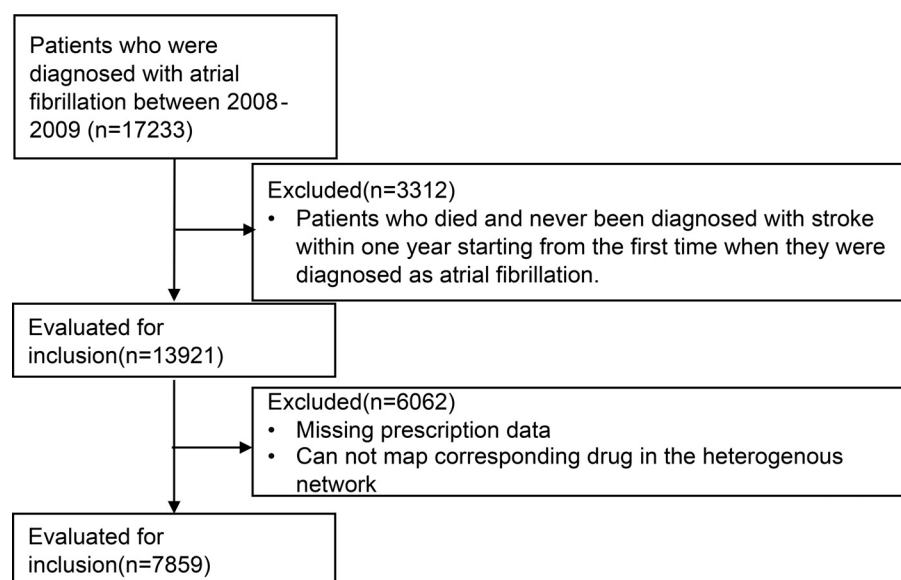
FIG. 3. Study population selection flow chart. The flow chart shows the selection process of AF patients from initial diagnosis (n = 17 233) through exclusion criteria to final study cohort (n = 7859). Exclusion criteria include early mortality without stroke diagnosis (n = 3312) and incomplete prescription data (n = 6062).

for this discrepancy. First, our model accounted for recurrent strokes, which could increase the observed incidence rate. Second, the age demographic of our study cohort was skewed toward older individuals, who inherently have a higher risk of stroke. Additionally, we excluded patients for whom subsequent medical records were unavailable; many of these exclusions were likely due to patients remaining healthy and not experiencing a stroke in the year following their initial data entry. Moreover, 6057 patients were excluded due to incomplete prescription records, and an additional 5 patients were excluded because their medications could not be mapped in the network, suggesting that the actual stroke rate may be lower than reported. Furthermore, our analysis included patients who did not survive beyond one year. Considering all these factors provides a comprehensive explanation for the elevated stroke rate observed in our study.

C. The ABioSPath model

Figure 4 illustrates the overall design of ABioSPath, which consists of two main modules: a wide module and a deep module, inspired by the popular wide and deep model developed by Google.[45] The deep module processes pathways with heads and targets related to a patient's prescription information and historical diagnoses to predict the probability of developing IS within 12 months, using a deep neural network architecture. The underlying assumption of the model is that the risk of IS tends to propagate along the shortest paths within the drug–disease–protein network, which often represent the most direct and significant mechanisms of action for a drug target, a concept supported by several studies.[14,40,46,47] While longer paths exist in biological systems, focusing on the shortest paths helps identify biologically plausible and interpretable connections that simplify complex causal relationships. This approach provides two key advantages in our knowledge graph framework: it enhances model interpretability by reducing pathway complexity, and it maintains biological relevance by capturing primary mechanistic relationships. Though other clinically relevant pathways exist, the shortest path methodology offers an effective balance between biological significance and computational

tractability, enabling the identification of critical pathways while preserving the clarity of causal relationships in the network. The model specifically identifies the shortest paths between source nodes (e.g., historical drugs and diagnoses) and the target node (IS) within the network, based on a patient's prescriptions and diagnosis history. The deep module comprises two identical submodules: a drug module and a disease module.

In the drug module, a graph attention network (GAT) is applied to the heterogeneous graph to capture both local and global topological connectivity, representing each node as a vector or embedding.[48] For each node in the graph, GAT generates unique embeddings. These embeddings replace the nodes in each patient's specific pathways, following the sequence of nodes to form raw path vectors. These raw path vectors are then utilized by a bidirectional long short-term memory (Bi-LSTM) neural network to capture sequential information.[49] The Bi-LSTM processed path embeddings are further refined through a node attention layer, which generates new path embeddings for each pathway. A subsequent path attention layer processes these node-refined path embeddings to identify the most significant pathways, producing aggregated final drug path embeddings.

The disease module employs the same framework independently, following the same principles to generate disease path embeddings. Finally, both the drug and disease path embeddings are concatenated and passed through a three-layer multilayer perceptron to predict the "deep probability" of developing IS within the next 12 months. Additional discussion and experimental results regarding the selection of two distinct components for biological and comorbidity pathways are available in the supplementary material.

The wide module employs a one-hot encoding technique to integrate relevant patient information, including previous diagnoses, prescriptions, age, and gender. In this encoding scheme, diseases are represented by a vector of length 555, drugs by a vector of length 109, age by a vector of length 3 (categorizing age groups as below 65, 65–75, and above 75), and gender by a binary vector of length 2. These encodings are combined into a single vector of length 669. This
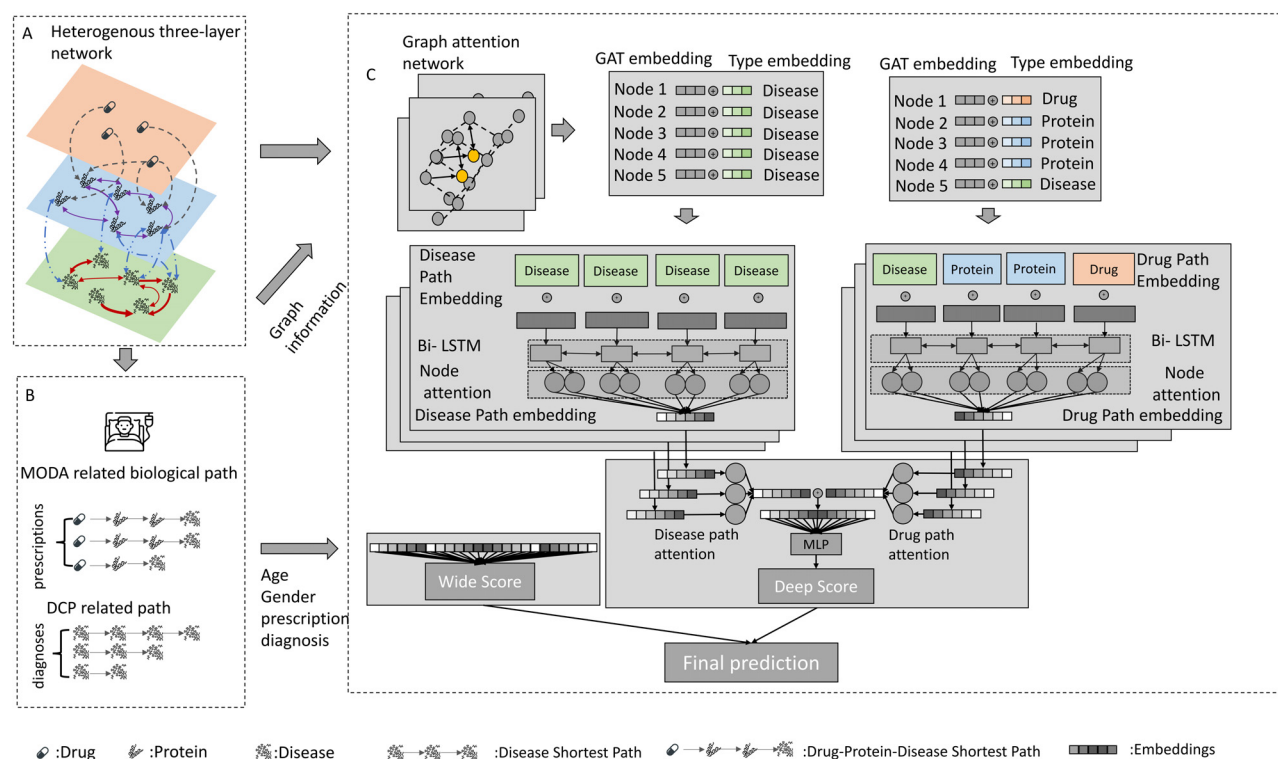
FIG. 4. Overview of ABioSPath model architecture. (a) Heterogeneous biological network comprising drug–protein network (orange), protein–protein interaction network (blue, bidirectional purple arrows), and disease comorbidity network (green, bidirectional red arrows). Inter-network connections show drug–protein interactions (dashed arrows) and protein–disease associations (dotted arrows). (b) Pathway identification module extracting MODA-related biological pathways (shortest paths between drugs and IS) and DCP pathways (past diagnoses to IS), integrated with demographic features. (c) Neural network architecture featuring three-layer GAT for node embedding generation, dual Bi-LSTM networks processing MODA and DCP pathway embeddings, followed by attention mechanisms for node and pathway importance weighting. The deep score from pathway analysis combines with a wide score (derived from clinical features through $CHADS_2$-based encoding) to produce the final risk prediction.

comprehensive vector is then input into a fully connected layer to compute the "wide module probability" of developing IS within the next 12 months.

The final layer of the model combines the scores from both the wide and deep modules using a weighted sum to generate a comprehensive prediction of an individual's risk of developing IS within the next 12 months. This approach leverages not only patient-specific demographic and clinical information but also integrates general knowledge about drug–protein–disease interactions embedded within the multilayer network. We hypothesize that this enriched knowledge incorporation enhances the accuracy of predicting an individual's risk of developing IS.

Importantly, the model is designed to provide explanations for how each predicted result is generated. Transparency is crucial in any model, especially in the healthcare industry, as it enhances credibility and trustworthiness among healthcare professionals and patients.[47] We utilize node and path attention layers to combine node embeddings along pathways, generating a comprehensive embedding for each path.

The node attention mechanism uses a trainable linear layer network and a softmax function to identify the most significant nodes within each path during the aggregation process. Similarly, path attention follows a similar calculation schema as node attention to identify the most influential pathways, based on the integrated node vectors processed through node attention. Following this, the model assigns attention weights to indicate the relative importance of nodes and paths. This approach allows the attention layers to (a) differentiate the significance of nodes within each path and (b) evaluate the contribution of each path to the final prediction. Through these mechanisms, the model enhances interpretability, enabling healthcare professionals to better understand the factors contributing to the risk prediction, thereby facilitating more informed clinical decision-making.

D. Model evaluation

The research samples were randomly divided into three sets: 70% (5503 patients) for model derivation, 20% (1571 patients) for model validation, and 10% (785 patients) for testing. Two baseline models were adopted for comparison: a logistic regression model and a least absolute shrinkage and selection operator (LASSO) model, both utilizing one-hot embeddings along with clinical risk scores.[50] The LASSO and logistic regression models utilize the same input features as the wide component of our architecture. The feature space consists of a 669-dimensional vector created through one-hot encoding of patient characteristics. Specifically, this vector comprises binary encodings of 555 distinct diagnoses, 109 different medications, age stratified into

three categories (<65, 65–75, and >75 years), and patient sex. Each model processes this high-dimensional input to produce a single output probability score for stroke risk prediction. Please refer to Table SI in the supplementary material for further details. The performance of all models was evaluated using several metrics: sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), average precision, and the area under the receiver operating characteristic curve (AUROC). The logistic regression model was trained using fivefold cross-validation on the training dataset, employing the elastic net method for regularization. The optimal alpha value was determined to be 10^{-6}, with an L1 ratio of 0. The model's performance was subsequently tested on both the validation and test datasets. Similarly, the LASSO model was trained with fivefold cross-validation, utilizing L1 norm regularization. A parameter search identified the optimal alpha value as 0.01. The LASSO model's performance was also tested on the validation and test datasets. Both models used the same vector length, consistent with that of the wide module. To evaluate pathway significance, we first identified the top twenty most important pathways for each patient based on their predictive contribution to stroke risk. We then quantified the occurrence frequency of each identified pathway across the entire patient cohort, where the occurrence count represents how many times a specific pathway was selected as one of the top twenty important pathways among all patients. This systematic approach enabled us to identify consistently important biological pathways in stroke risk prediction by measuring their recurrence across the study population.

SUPPLEMENTARY MATERIAL

See the supplementary material for details on baseline models, the rationale for ABioSPath's two-component design, ABioSPath's performance with incomplete data, and comparative analyses with existing IS risk prediction studies.

ACKNOWLEDGMENTS

AUTHOR DECLARATIONS
Conflict of Interest

The authors have no conflicts to disclose.

Ethics Approval

Ethics approval for experiments reported in the submitted manuscript on animal or human subjects was granted. This study was approved by the City University of Hong Kong Human Subjects Ethics Sub-Committee (2–3-201804_03).

Author Contributions

Zhiheng Lyu and Jiannan Yang contributed equally to this work.

Zhiheng Lyu: Data curation (equal); Formal analysis (lead); Investigation (equal); Methodology (equal); Validation (equal); Visualization (equal); Writing – original draft (lead). **Jiannan Yang:** Conceptualization (equal); Data curation (equal); Formal analysis (equal); Investigation (equal); Methodology (equal); Validation (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal). **Zhongzhi Xu:** Conceptualization (equal); Formal analysis (equal); Funding acquisition (equal); Investigation (equal); Methodology (equal); Project administration (equal); Validation (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal). **Weilan Wang:** Investigation (equal); Validation (equal); Writing – review & editing (equal). **Weibin Cheng:** Investigation (equal); Validation (equal); Writing – review & editing (equal). **Kwok-Leung Tsui:** Investigation (supporting); Supervision (equal); Validation (supporting). **Qingpeng Zhang:** Conceptualization (lead); Data curation (equal); Formal analysis (equal); Funding acquisition (equal); Investigation (equal); Methodology (equal); Project administration (supporting); Supervision (supporting); Validation (equal); Writing – original draft (equal); Writing – review & editing (equal).

DATA AVAILABILITY

The data that support the findings of this study are openly available in GitHub at https://github.com/lyuzhathk/abiospath_lyu, Ref. 51.

REFERENCES

[1]G. Lippi, F. Sanchis-Gomar, and G. Cervellin, Int. J. Stroke **16**(2), 217 (2021).

[2]World Health Organization Regional Office for the Eastern Mediterranean, "Stroke, cerebrovascular accident" (2020), https://www.emro.who.int/health-topics/stroke-cerebrovascular-accident/index.html (accessed March 15, 2025).

[3]E. Hunter and J. D. Kelleher, Front. Neurol. **13**, 803749 (2022).

[4]S. Aspberg, Y. Chang, A. Atterman, M. Bottai, A. S. Go, and D. E. Singer, Eur. Heart J. **37**(42), 3203 (2016).

[5]A. Amin, Clin. Interventions Aging. **8**, 75–84 (2013).

[6]M. S. Dzeshka, D. A. Lane, and G. Y. Lip, Clin. Cardiol. **37**(10), 634 (2014); V. H. W. van der Endt, J. Milders, B. B. L. P. de Vries, S. A. Trines, R. H. H. Groenwold, O. M. Dekkers, M. Trevisan, J. J. Carrero, M. van Diepen, F. W. Dekker, and Y. de Jong, EP Europace **24**(11), 1739–1753 (2022).

[7]J. Menche, A. Sharma, M. Kitsak, S. D. Ghiassian, M. Vidal, J. Loscalzo, and A. L. Barabási, Science **347**(6224), 1257601 (2015).

[8]A. K. Boehme, C. Esenwa, and M. S. Elkind, Circ. Res. **120**(3), 472 (2017); A. Arboix, World J. Clin. Cases **3**(5), 418 (2015).

[9]L. R. Øie, T. Kurth, S. Gulati, and D. W. Dodick, J. Neurol., Neurosurg. Psychiatry **91**(6), 593 (2020).

[10]W. T. Fang, H. J. Li, H. Zhang, and S. Jiang, Br. J. Clin. Pharmacol. **74**(5), 744–756 (2012); A. Oesterle, U. Laufs, and J. K. Liao, Circ. Res. **120**(1), 229 (2017).

[11]H. Zhu, S. Hu, Y. Li, Y. Sun, X. Xiong, X. Hu, J. Chen, and S. Qiu, Front. Immunol. **13**, 828447 (2022).

[12]J. P. Marto, D. Strambo, F. Livio, and P. Michel, Stroke **52**(10), e646–e659 (2021).

[13]S. Kumar, S. Savitz, G. Schlaug, L. Caplan, and M. Selim, Neurology **66**(8), 1153 (2006).

[14]J. Yang, Z. Li, W. K. K. Wu, S. Yu, Z. Xu, Q. Chu, and Q. Zhang, Briefings Bioinf. **23**(6), bbac469 (2022).

[15]T. Hricik, D. Bader, and O. Green, "Using RAPIDS AI to accelerate graph data science workflows," in IEEE High Performance Extreme Computing Conference (HPEC), 2020.

[16]A. C. Cameron, N. N. Lang, and R. M. Touyz, Drugs **76**, 1529 (2016).

[17]M. Selim, S. Savitz, I. Linfante, L. Caplan, and G. Schlaug, BMC Neurol. **5**(1), 10 (2005).

[18]N. Muñoz-Durango, C. A. Fuentes, A. E. Castillo, L. M. González-Gómez, A. Vecchiola, C. E. Fardella, and A. M. Kalergis, Int. J. Mol. Sci. **17**(7), 797 (2016).

[19]P. M. Rothwell, A. Algra, Z. Chen, H. C. Diener, B. Norrving, and Z. Mehta, Lancet **388**(10042), 365 (2016).

[20]C. R. Lee, K. E. North, M. S. Bray, D. J. Couper, G. Heiss, and D. C. Zeldin, Clin. Pharmacol. Ther. **83**(1), 52 (2008).

[21]J. Jiang and R. Dingledine, Trends Pharmacol. Sci. **34**(7), 413 (2013).

[22]Z. H. Xu, J. R. Jiao, R. Yang, B. Y. Luo, X. F. Wang, and F. Wu, J. Int. Med. Res. **40**(1), 282 (2012).

[23]D. Z. Liu, G. C. Jickling, B. P. Ander, H. Hull, X. Zhan, C. Cox, N. Shroff, C. Dykstra-Aiello, B. Stamova, and F. R. Sharp, J. Cereb. Blood Flow Metab. **36**(8), 1374 (2016).

[24]G. J. Blake and P. M. Ridker, Circ. Res. **89**(9), 763 (2001).

[25]N. S. Rost, P. A. Wolf, C. S. Kase, M. Kelly-Hayes, H. Silbershatz, J. M. Massaro, R. B. D'Agostino, C. Franzblau, and P. W. F. Wilson, Stroke **32**(11) 2575–2579 (2001); P. M. Ridker, Circulation **108**(12), e81–e85 (2003).

[26]R. Greco, D. Amantea, F. Blandini, G. Nappi, G. Bagetta, M. Tiziana Corasaniti, and C. Tassorelli, Int. Rev. Neurobiol. **82**, 423 (2007); M. E. Maniskas, J. M. Roberts, R. Trueman, A. E. Learoyd, A. Gorman, J. F. Fraser, and G. J. Bix, J. Neurointerventional Surg. **10**(11), 29 (2018).

[27]A. J. Pesce, M. Rashkin, and U. Kotagal, Clin. Chem. **44**(5), 1124–1128 (1998).

[28]B. Modrau, G. Andersen, I. K. Mikkelsen, A. Nielsen, M. B. Hansen, M. B. Johansen *et al.*, Stroke **51**(7), 1983 (2020).

[29]L. C. Foukas, N. Daniele, C. Ktori, K. E. Anderson, J. Jensen, and P. R. Shepherd, J. Biol. Chem. **277**(40), 37124 (2002); V. Asati, S. K. Bharti, D. K. Mahapatra, V. Asati, and A. K. Budhwani, Curr. Pharm. Des. **22**(39), 6039–6054 (2016); D. A. Fruman, H. Chiu, B. D. Hopkins, S. Bagrodia, L. C. Cantley, and R. T. Abraham, Cell **170**(4), 605 (2017).

[30]C. Qin, S. Yang, Y.-H. Chu, H. Zhang, X.-W. Pang, L. Chen, L.-Q. Zhou, M. Chen, D.-S. Tian, and W. Wang, Signal Transduct. Target. Ther. **7**(1), 215 (2022).

[31]A. Stepanyan, R. Zakharyan, A. Simonyan, G. Tsakanova, and A. Arakelyan, BMC Med. Genet. **19**(1), 33 (2018).

[32]Y. Xue, X. Zeng, W.-J. Tu, and J. Zhao, Dis. Markers **2022**, 2395269.

[33]S. N. Stanford, A. Sabra, L. D'Silva, M. Lawrence, R. H. K. Morris, R. Storton, M. R. Brown, V. Evans, K. Hawkins, P. R. Williams, S. J. Davidson, M. Wani, J. F. Potter, and P. A. Evans, BMC Neurol. **15**(1), 35 (2015).

[34]C. M. M. Lawes, D. A. Bennett, V. L. Feigin, and A. Rodgers, Stroke **35**(3), 776 (2004).

[35]Z. Coralic, A. S. Kim, and D. R. Vinson, West J. Emerg. Med. **16**, 572 (2015).

[36]R. F. Wolff, K. G. M. Moons, R. D. Riley, P. F. Whiting, M. Westwood, G. S. Collins, J. B. Reitsma, J. Kleijnen, S. Mallett, and PROBAST Group, Ann. Intern. Med. **170**(1), 51 (2019).

[37]Z. Xu, J. Zhang, Q. Zhang, Q. Xuan, and P. S. F. Yip, IEEE Trans. Cybern. **52**(9), 9809 (2022).

[38]J. Piñero, À. Bravo, N. Queralt-Rosinach, A. Gutiérrez-Sacristán, J. Deu-Pons, E. Centeno, J. García-García, F. Sanz, and L. I. Furlong, Nucl. Acids Res. **45**(D1), D833–D839 (2016).

[39]D. Szklarczyk, A. L. Gable, K. C. Nastou, D. Lyon, R. Kirsch, S. Pyysalo, N. T. Doncheva, M. Legeay *et al.*, Nucl. Acids Res. **49**(D1), D605 (2021).

[40]F. Cheng, I. A. Kovács, and A.-L. Barabási, Nat. Commun. **10**(1), 1197 (2019).

[41]D. Szklarczyk, A. Santos, C. V. Mering, L. J. Jensen, P. Bork, and M. Kuhn, Nucl. Acids Res. **44**(D1), D380 (2016).

[42]Y. Zhou, Y. Zhang, X. Lian, F. Li, C. Wang, F. Zhu, Y. Qiu, and Y. Chen, Nucl. Acids Res. **50**(D1), D1398 (2022).

[43]D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, and Z. Sayeeda, Nucl. Acids Res. **46**(D1), D1074 (2018).

[44]E. Bosco, L. Hsueh, K. W. McConeghy, S. Gravenstein, and E. Saade, BMC Med. Res. Methodol. **21**, 241 (2021); K. C. Abbott, J. C. Reynolds, A. J. Taylor, and L. Y. Agodoa, Am. J. Transplant. **3**(4), 471 (2003).

[45]H.-T. Cheng, L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. Aradhye, G. Anderson, G. Corrado, W. Chai, and M. Ispir, paper presented at the Proceedings of the 1st workshop on deep learning for recommender systems, 2016.

[46]E. Guney, J. Menche, M. Vidal, and A.-L. Barábasi, Nat. Commun. **7**(1), 10331 (2016).

[47]S. Sadegh, J. Skelton, E. Anastasi, J. Bernett, D. B. Blumenthal, G. Galindez, M. Salgado-Albarrán, O. Lazareva, K. Flanagan, S. Cockell, C. Nogales, A. I. Casas, H. H. H. W. Schmidt, J. Baumbach, A. Wipat, and T. Kacprowski, Nat. Commun. **12**(1), 6848 (2021).

[48]P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, arXiv:1710.10903 (2017).

[49]Z. Huang, W. Xu, and K. Yu, arXiv:1508.01991 (2015).

[50]N. Gauraha, Resonance **23**(4), 439 (2018); J. Y. Chen, A. D. Zhang, H. Y. Lu, J. Guo, F. F. Wang, and Z. C. Li, J. Geriatr. Cardiol. **10**(3), 258–266 (2013).

[51]See https://github.com/lyuzhathk/abiospath_lyu for "GitHub."