International Journal of Data Science and Analytics
https://doi.org/10.1007/s41060-024-00597-8

REGULAR PAPER

Identifying clinical feature clusters toward predicting stroke in patients with asymptomatic carotid stenosis

David Xu[1] · Sanaz Matinmehr[2] · Alan Sawchuk[3] · Xiao Luo[3,4]

Received: 2 April 2024 / Accepted: 24 June 2024

Abstract

Despite the widespread application of machine learning models and feature selection methods to identify important clinical features in electronic health records (EHR) for disease prediction, the use of graph neural networks (GNNs) to uncover significant clinical features associated with a disease remains largely unexplored. In this investigation, we developed a computational method utilizing EHR data from Indiana University Medical Hospital to predict stroke in patients with asymptomatic carotid stenosis. We first constructed a patient clinical feature graph for each patient based on the co-occurrence of features (medications, diagnoses, and results of laboratory tests) in the EHR data within a predefined timeframe (e.g., 6 months before the detection of the disease). Then, we applied an unsupervised GNN-based clustering approach and our algorithm to select notable clinical feature clusters crucial for stroke prediction. These clinical features served as the basis for constructing patient representation for prediction. Various supervised learning models were evaluated for their prediction capabilities. Unlike conventional feature selection methods, our GNN-based feature selection approach relies solely on positive cases. We compared our method against baseline models for stroke prediction and achieved robust performance metrics, including an AUC of 0.87 and an F1 score of 0.80, surpassing all baselines. Additionally, we conducted an ablation study on the amount of EHR data, measured in months, to determine the most effective approach for generating patient clinical feature graphs. By capturing inherent relationships between clinical features using the graph model, our approach offers a promising avenue for advancing disease prediction, particularly in scenarios with limited positive cases available. Our code can be found on Github (https://github.com/xudav001/Identifying-Phenotype-Clusters)

Keywords Disease prediction · Asymptomatic carotid stenosis · Graph clustering · Stroke · Electronic health records

1 Introduction

Stroke is a significant global cause of both mortality and disability, greatly affecting quality of life. Every year in the USA, approximately 610,000 individuals experience their first stroke, contributing significantly to long-term disability [1]. Between 11 and 15% of strokes are linked to asymptomatic carotid stenosis (ACS) as the initial neurological event. Treating asymptomatic carotid stenosis poses notable risks as well. Although current randomized clinical trials offer guidance on when to intervene in asymptomatic stenosis, they require a high number of potentially unnecessary procedures [2, 3]. The available literature on the natural progression of asymptomatic carotid stenosis demonstrates the value of adopting a management strategy for asymptomatic carotid stenosis. Such an approach would help reduce unnecessary interventions and effectively preventing strokes. Therefore, it is imperative to identify asymptomatic carotid

✉ Xiao Luo
xiao.luo@okstate.edu

David Xu
davidxu2026@u.northwestern.edu

Sanaz Matinmehr
sbardoo@iu.edu

Alan Sawchuk
asawchuk@iupui.edu

1 Department of Computer Science, Northwestern University, Mudd Hall, Evanston, IL 60208, USA

2 Department of Statistics, Indiana University Indianapolis, Myles Brand Hall, Indianapolis, IN 47408, USA

3 School of Medicine, Indiana University, 340 W 10th St, Indianapolis, IN 47408, USA

4 Department of Management Science and Information Systems, Oklahoma State University, 370 Business Building, Stillwater, OK 74078, USA

Published online: 09 August 2024

stenosis (ACS) patients who are at a heightened risk of stroke and administer appropriate invasive treatment accordingly. In recent decades, extensive research has been conducted to assess ACS patients, with the aim of identifying high-risk characteristics that predispose certain individuals to strokes, such as plaque-related attributes [4] and stroke-relevant genotypes [5] and clinical features [6]. However, conducting detailed ultrasound examinations to identify all plaque features is challenging in busy clinical settings and rural areas. With advancements in AI techniques and the availability of extensive medical data stored in electronic health records (EHRs), the development of a predictive stroke model using clinical features collected from EHRs has become feasible. This research is the first to predict the risk of developing stroke in patients with ACS based on clinical features stored in real-world EHRs, including diagnoses, medications, and laboratory tests.

Feature selection plays a crucial role in analyzing EHR data for disease prediction by identifying the most relevant variables that contribute to the predictive model's accuracy and efficiency. In the context of EHR data, studies have shown that using feature selection methods can improve the performance of machine learning models by focusing on the most informative clinical variables, such as laboratory results, medication history, and patient demographics [7–9]. The traditional feature selection methods that are applied to EHR data include information gain [10], tree-based methods [11], principal component analysis (PCA) [12], etc. The limitation of these traditional feature selection methods does not consider the complex relations between features. Graph neural networks (GNNs) have significantly advanced in recent years and are increasingly being applied to feature selection tasks due to their ability to capture complex relationships and dependencies in data. In the context of feature selection, GNNs can be used to model the relationships between features as a graph, where nodes represent features and edges represent correlations or interactions between them. Recent research has illustrated the potential efficacy of GNNs in feature selection for disease prediction using EHR data [13]. For instance, GNNs can effectively model the interactions between different clinical variables, such as symptoms, diagnoses, and treatments, to select the most relevant features for predicting patient outcomes [13–16]. This research is the first to utilize GNN in an unsupervised manner to identify the significant clinical variables that contribute to the stroke in patients with ACS.

In this research, we aim to develop a computational method to first identify a set of clinical variables that are significantly associated with development of stroke reflected by the medical history of individuals. This novel clinical variable selection approach is based on the unsupervised GNN model. Then, we utilize these selected clinical variables to predict the development of stroke in patients with ACS using supervised learning models. We compared our framework against the traditional clinical variable selection to demonstrate the advance of our approach. Our research has the following main contributions:

- We pioneered the use of a graph model to represent the relationships among clinical features in a patient's medical record.
- We introduced the use of weighted GNN to detect clusters of clinical variables in an unsupervised manner, thereby identifying significant clinical variables associated with patients who subsequently develop strokes.
- We evaluated various supervised learning models using the selected variables for stroke prediction and demonstrated that the performances surpassed the baseline models.

2 Related works

In this section, we explore related work by addressing the main aspects most relevant to this research: (1) conventional approaches used to select important clinical variables from EHR data for disease prediction, (2) the application of graph neural networks to medical data for predictive analysis and relation extraction, and (3) the cutting-edge methodologies specifically designed for predicting stroke or other diseases.

2.1 Clinical variable selection for disease prediction

Within electronic health records (EHR), clinical variables can stem from structured fields such as medication and diagnosis, or unstructured fields like textual symptoms and examination findings. Chronic conditions like diabetes, chronic obstructive pulmonary disease (COPD), and hypertension can significantly contribute to carotid stenosis, a condition linked to stroke risk [17]. Traditional computational methods for assessing the significance of clinical features in medical histories often rely on vector-based approaches [18–21] including REF, LASSO, most tree-based approaches, where clinical variables are treated as independent attributes forming vectors. However, such approaches fail to capture relationships between variables in an unsupervised manner and struggle to quantify the importance of heterogeneous relationships within the data. For instance, intricate temporal relationships between chronic diseases, laboratory tests, medical procedures, and medications cannot be fully interpreted when represented as entries in a single vector. While some research utilizes techniques like SHAP [22] or local interpretable model agnostic explanations (LIME) [23] for feature interpretation, these methods often focus on individual features rather than clusters of related features. Given that patients with extensive medical histories may develop

multiple chronic diseases with varying sequences, a data-driven quantification approach to identifying crucial clinical features or feature associations could aid clinicians in identifying groups of key risk factors within EHR that may predict disease onset. It is crucial to develop an improved approach for selecting clinical variables that can effectively analyze the intricate relationships among these variables.

2.2 Graph neural networks for clinical data analysis

Graph neural networks (GNNs) have been increasingly utilized for healthcare outcome prediction and analysis due to their ability to model complex relationships between medical entities. GNNs effectively capture the intricate dependencies within healthcare data, such as patient records, disease pathways, and molecular interactions, facilitating more accurate predictions and insights. For example, GNNs have been applied to predict patient outcomes by modeling the relationships between clinical variables in electronic health records (EHRs), such as symptoms, diagnoses, and treatments [15, 16, 24]. Studies have demonstrated that GNNs outperform traditional machine learning models in predicting disease progression and patient readmission by leveraging the graph structure of EHR data [14]. In addition to EHR data, GNNs have been used to analyze genomic data for predicting disease susceptibility and drug response. Furthermore, GNNs have been utilized in pharmacogenomics to predict drug–target interactions and optimize drug discovery processes by integrating various biological networks, including protein–protein interaction networks and molecular graphs [25]. These advancements highlight the potential of GNNs to revolutionize health care by providing deeper insights into disease mechanisms and enabling personalized medicine approaches [26]. Unlike previous studies, our research emphasizes the application of GNNs to identify significant clinical variables that contribute to disease prediction, utilizing the electronic health records (EHR) data of individuals.

2.3 Machine learning models for disease prediction

While previous studies have explored stroke prediction using machine learning and neural networks, our research stands as the first endeavor to forecast stroke in patients with asymptomatic carotid stenosis (ACS), a clinically more challenging task than predicting stroke in the general population. For instance, Dev et al. utilized a perceptron neural network to predict stroke using only four patient attributes: age, heart disease, average glucose level, and hypertension [27]. Similarly, Nwosu et al. [28] employed all patient attributes as input features for stroke prediction, utilizing machine learning algorithms such as decision trees, random forests, and neural networks. Additionally, several deep learning algorithms have been developed for stroke prediction. Hung et al. [29] explored the efficacy of a deep neural network in stroke prediction using claims data, comparing its performance with traditional machine learning algorithms. While the results indicated a slight improvement with the deep neural network compared to traditional methods, the difference was not significant. Moreover, instead of relying solely on clinical features in medical records, other studies [30, 31] have employed electrocardiograms or CT scan (computed tomography) image datasets for stroke prediction. Our study is the first to develop a novel computational framework and applied on the complete EHR data for stroke prediction.

In summary, our study is the first to use a graph model to represent the relationships between clinical variables and employ unsupervised graph clustering with GNNs to identify key clinical variables for stroke prediction in patients with ACS using EHR data. The developed computational framework can be adapted to other disease prediction using EHR data.

3 Methodology

Figure 1 shows the computational framework of our approach. Given a dataset including diagnoses, medication, and laboratory tests of each patient in the EHR. We first began by applying a data preprocessing step that normalizes clinical variables using standard coding and normalization methods commonly applied to diagnoses, medications, and laboratory tests. Then, we developed an innovative unsupervised method to select the important clinical variables based on the positive cases using a GNN-based approach. In this approach, a clinical variable graph is constructed for each patient who experienced a stroke in the training data. Weighted graph clustering is then employed to identify clusters of clinical features, which help characterize these patients. Ultimately, patient representation is crafted using the cluster centroids and the associated clinical features, upon which diverse supervised learning methodologies are applied for stroke prediction.

3.1 Clinical variable preprocessing

The structured data within the EHR, such as diagnosis codes (International Classification of Diseases (ICD) 10 or 9)), laboratory test results, and medications, were incorporated as clinical variables in our study. Concerning diagnoses, we initially convert all ICD version 9 codes to their corresponding ICD version 10 codes. Rather than utilizing the complete ICD code, we opted for the second-level categorization down to the leaf nodes of the ICD version 10 code hierarchy, which aggregates diagnoses into broader categories [32]. Specifically, we employed the ICD version 10 codes without the decimal point and subsequent numbers. For instance, the ICD

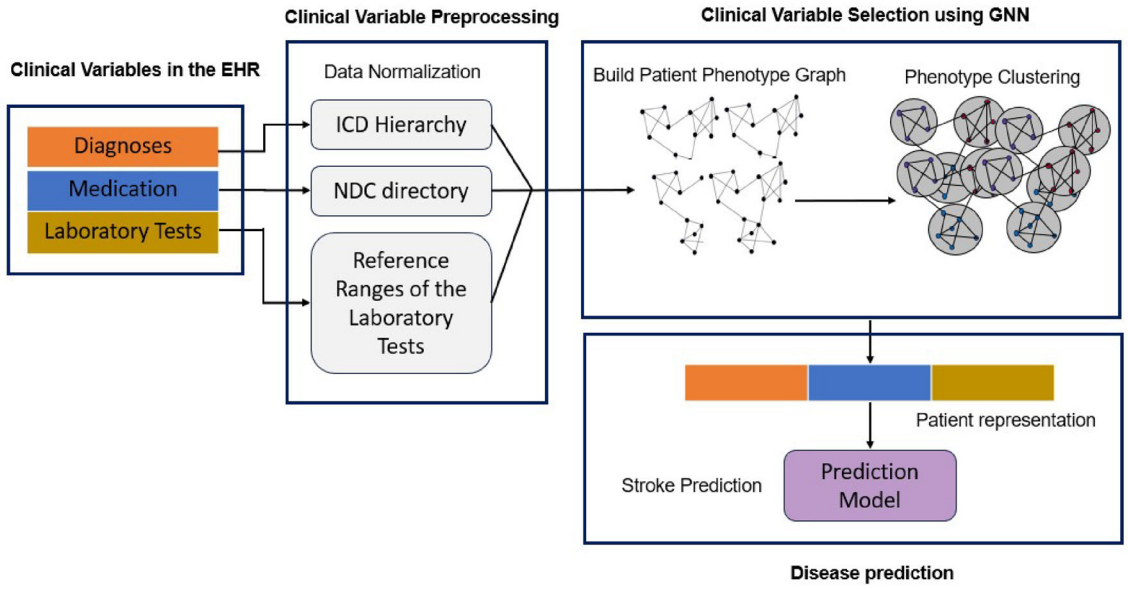

Fig. 1 Illustration of the system framework

version 10 code E78.1 denoting pure hypertriglyceridemia was converted to E78, representing disorders of lipoprotein metabolism. Regarding laboratory tests, we standardized the values to descriptors such as "low," "normal," "high," or "abnormal" based on the reference ranges provided by the laboratory test references. For medications, instead of utilizing the original drug names alongside dosages, we categorized them according to the drug groups specified in the National Drug Code (NDC) Directory, excluding categories such as "medical devices and supplies," "diagnostic products," "Nasal Agents-Systemic Topical," and "dermatologicals." For example, the medication Heparin is mapped to the medication category of anticoagulants.

3.2 GNN-FS: clinical variable selection using graph neural network

To identify significant clinical variables associated with patients who developed stroke, we first constructed a patient clinical variable graph to capture the relationships among the variables. We then applied a GNN approach to identify clusters of clinical variables that are relevant to patients who experienced a stroke.

3.2.1 Build a patient clinical variable graph

To form a patient clinical variable graph, each node within the graph corresponds to a clinical variable in the patient's medical history, such as a diagnosis, medication, or laboratory test. For laboratory tests, each node represents both the specific test and its result, categorized as "low," "high," or "abnormal." Edges connecting the nodes are established when they occur simultaneously within a predefined timeframe, e.g., 1 month or 6 months. The timeframe is calculated based on the discharge date associated with a diagnosis, the order date of a medication, and the performance date of a laboratory test. In a patient's medical history, laboratory tests may be performed and medications ordered more than once within a timeframe. Thus, the frequencies of co-occurrences of clinical features determine the weights on the edges ($w_{L_1L_2}, w_{M_1L_2}, \ldots$). For instance, if a patient undergoes laboratory test L_1 and receives medication M_2 on day 1, and the same procedures are repeated on day 2 a few months after day 1, the edge weight between L_1 and M_2 is 2 if the defined timeframe for co-occurrence calculation is less than 3 months. Figure 2 illustrates a patient graph comprising nodes: D_i nodes represent diagnoses, M_i nodes represent medications, and L_i nodes represent laboratory tests. This graphical depiction of a patient elucidates the correlations and patterns of co-occurrence among various clinical variables, providing a structured approach to analyzing the connections between diagnoses, prescribed medications, and laboratory results for each individual patient, thereby enhancing our understanding of patient health and medical relationships.

3.2.2 Generate clinical variable clusters using GNN and select important variables

To generate clinical variable clusters from the clinical variable graph, we first created node embeddings and then applied the k-means clustering algorithm, an unsupervised approach, to identify the clinical variable clusters. In this research, each node in the graph represents a clinical variable.

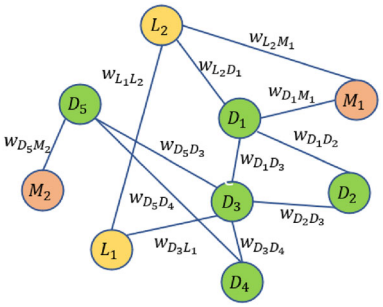

Fig. 2 Weighted clinical feature graph of a patient

Node2Vec [33] is one of the state-of-the-art unsupervised graph neural networks with an identity feature matrix. Its training objective is to generate node embeddings that capture the relationships among nodes in a given graph. In this research, we utilized a weighted Node2Vec model to produce clinical variable embeddings derived from the patient clinical variable graph.

Consider a graph $G = (V, E)$ denotes a set of nodes $V = (v_1, \ldots, v_n)$, $|V| = n$, and edges $E \subseteq V \times V$, $|E| = m$. Node2Vec produces multiple random walks of a specified length for every node v within the set V. The random walks are created by moving from one node to another according to predetermined transition probabilities ($\pi_{v_i,v_j} = \alpha_{pq} \cdot w_{v_i,v_j}$). The calculation of α_{pq} is shown as Eq. (1), which are designed to balance between exploring local neighborhoods and exploring distant parts of the graph.

$$\alpha_{pq} = \begin{cases} \frac{1}{p}, & \text{if } d_{v_i,v_j} = 0 \\ 1, & \text{if } d_{v_i,v_j} = 1 \\ \frac{1}{q}, & \text{if } d_{v_i,v_j} = 2 \end{cases} \tag{1}$$

where w_{v_i,v_j} is the weight of the edge between nodes v_i and v_j, and d_{ij} is the shortest path distance between nodes v_i and v_j. Here, d_{ij} represents the shortest path distance between nodes i and j. The transition probability is controlled by two parameters: p and q:

- p defines the probability of the random walk to revisit nodes it has previously visited
- q defines the probability of the random walk to explore new nodes

In our experiment, we opted for unbiased walks, thus setting both p and q to 1. This choice guarantees that the random walks offer an equitable exploration of the graph's topology. Employing a weighted Node2Vec algorithm, we accounted for edge weights during random walk generation. This means that the algorithm is inclined to traverse edges with greater weights, indicating stronger relationships between nodes. In our setup, we conducted random walks with a length of 30 and execute 200 random walks for each node.

After generating the random walk paths, Node2Vec uses the skip-gram model to generate node embeddings. The primary aim of the learning task is to forecast context nodes, those neighboring a specific target node, where the target node is initially transformed into a one-hot encoded vector within the input layer. The training of the skip-gram model employs stochastic gradient descent to maximize the likelihood of correctly predicting context nodes for each given target node. Through this optimization process, the embeddings of nodes are iteratively adjusted, resulting in embeddings that reflect similarity among nodes, thereby capturing key structural aspects of the graph, including node resemblances and connections. The objective of the embedding generation process is to maximize the logarithmic probability of observing a network neighborhood V for a node u based on its feature representation, which is shown as Eq. (2).

$$\mathcal{L} = \sum_{v \in V} \sum_{u \in \mathcal{N}_v} \log P(u|v) \tag{2}$$

here V is the set of all nodes, u represents context nodes, v is the target node, and $\mathcal{N}_v$ is the set of context nodes for target node v. $P(u|v)$ is the probability of context node u given target node v, and it is defined using the softmax function, shown in Eq. (3).

$$P(u|v) = \frac{\exp(e_u^T \cdot e_v)}{\sum_{w \in V} \exp(e_w^T \cdot e_v)} \tag{3}$$

where e_u and e_v are the nodes embeddings u and v, respectively. For our experiment, the generated embeddings had a dimension of 64.

After the Node2Vec model is trained, we employ the k-means clustering algorithm on the node embeddings generated by the Node2Vec model. By applying the k-means clustering algorithm to these embeddings, we aim to identify clinical variable clusters derived from the patient graphs. This approach helps discover groups of clinical features that share similar traits and potentially indicate similar disease predispositions. The k-means algorithm aims to partition the embeddings into "k" distinct clusters, where each cluster is represented by a centroid point. These centroids serve as representatives of the patients within their respective clusters, enabling us to categorize patients based on their clinical feature similarities. The k-means algorithm minimizes the proximity of a clinical feature (x) to the centroid of the corresponding cluster, shown as Eq. (4), where μ_i is the centroid of cluster C_i and k is the number of clusters. This process orga-

nized similar attributes into clusters, enabling us to uncover patterns and relationships within the dataset. The resulting clusters provided a glimpse into the commonalities among certain groups of patient attributes related to stroke. Essentially, k-means clustering served a crucial function in our investigation by pinpointing centroids and organizing clusters among the dataset of attributes associated with stroke patients. This methodology facilitated the discovery of latent patterns and correlations, thus enhancing the field of stroke risk prediction and associated studies. To ascertain the appropriate number of clusters, we compute silhouette scores [34] across different values of k and subsequently employ the elbow method [35] to pinpoint the optimal number of clusters.

$$\text{Minimize } J = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \tag{4}$$

Algorithm 1: Clinical feature selection using graph clustering

Data: Clinical Feature Clusters ($C_{0i}, \ldots, C_{ni}$) of each patient ($i \in P$) who developed stroke in the training data
Result: A set of important clinical features (S)

$S = \emptyset$;
$centroid\_occurrences = \emptyset$;
foreach $i \in P$ **do**
  foreach $cluster \in C_i$ **do**
    centroid = calculate_centroid($cluster$) ;
    if $centroid \in centroid\_occurrences$ **then**
      $centroid\_occurrences$[centroid] = $centroid\_occurrences$[centroid] + 1 ;
    end
    else
      $centroid\_occurrences$[centroid] = 1 ;
    end
  end
end
foreach $(centroid, occurrences) \in centroid\_occurrences$ **do**
  if $occurrences \geq 2$ **then**
    $important\_cluster = \emptyset$;
    foreach $i \in P$ **do**
      foreach $cluster \in C_i$ **do**
        if $centroid = calculate\_centroid(cluster)$ **then**
          $important\_cluster = cluster$;
          break ;
        end
      end
    end
    foreach $clinical\_feature \in important\_cluster$ **do**
      $S = S \cup \{clinical\_feature\}$;
    end
  end
end

We formed clinical variable clusters based on patients who have experienced a stroke, aiming to characterize these patients in a way that facilitates stroke prediction. Our hypothesis is that clinical variables with centroids shared by multiple patients' clinical feature graphs indicate groups of patients with similar underlying characteristics. Therefore, we utilized all clinical variables within these clusters where shared centroids exist among multiple stroke patients, identifying them as important clinical variables. This process is applied exclusively to the training data to ensure that the resulting patient clinical features are robust and applicable to new, unseen data. Algorithm 1 details the procedure for identifying significant clinical variables from stroke patients in the training dataset.

3.3 Patient representation and stroke prediction

After identifying a set of important clinical variables to characterize patients who developed stroke, we constructed patient representations using these selected variables. For each patient (i), a one-hot vector representation is generated based on the presence ($v_{ij} = 1$) or absence ($v_{ij} = 0$) of clinical variable trait j: $V_i = [v_{i1}, v_{i2}, \ldots, v_{iN}]$ where N is the total number of clinical variables. Our patient representation approach differs from traditional methods, which consider clinical variables from both positive and negative patients. In the traditional approach, negative patients help identify negative clinical features. Conversely, our method relies solely on positive patients to determine important clinical variables for disease prediction. This strategy is particularly beneficial for identifying significant clinical variables from a small number of known positive cases, such as in the context of rare diseases.

To comprehensively evaluate the identified clinical variables for stroke prediction, we implemented five supervised learning algorithms on top of patient representations. These algorithms include random forest (RF), multilayer perceptron (MLP), support vector machines (SVM), linear regression (LR), and k-nearest neighbors (k-NN). In the subsequent experimental section, we provide a detailed description of the hyperparameters associated with these prediction models.

4 Experimental results

4.1 Dataset

This study was approved by Indiana University Institutional Review Board (IRB) and was classified as an exempt study (IRB number #14845; approved on March 25, 2022). The target population for this study is patients who have had CTA examinations to confirm the carotid stenosis status and who also have concurrent carotid ultrasound and clinical data stored in the EHR. Patients are categorized as having ACS versus symptomatic carotid stenosis (SCS) based

on the medical chart review. Patients who have not had a stroke, transient ischemic attack, or amaurosis fugax monocular which is blindness-temporary or permanent blindness in one eye are categorized as having ACS [36]. The EHR data is extracted from an Indiana University Medical Hospital. We have included 207 patients in this research. Out of the 207 patients, 92 are considered stroke patients. Table 1 shows the basic statistics of our dataset. Our dataset contains slightly more non-stroke patients than stroke patients. The mean age of both groups is similar. There are slightly more male patients in the dataset. The race distribution is imbalanced, with a significantly higher number of white patients and fewer patients from other racial groups. We stratified our data and divided it into training (70%), validation (10%), and testing (20%).

4.2 Hyperparameter setting

To construct the patient clinical feature graph, we set the co-occurrence timeframe to be 6 months at the beginning. That means the edge between two clinical features is created when they co-occur within 6 months based on the dates in records. All the machine learning models are fine-tuned to gain the optimal performance. Random forest has 100 decision trees and no maximum depth. k-Nearest neighbors are used with 21 neighbors and a leaf size of 30. The default parameters are used for support vector machine which include a regularization parameter of 1.0, a radius basis function kernel, and a degree of 3. The implemented multilayer perceptron has two hidden layers with a size of 100 and 50, respectively, and 1000 max iterations. The logistic regression is implemented with a l2 regularization and 100 max iterations.

4.3 Evaluation metrics

The performance metrics adopted in this study include F1, precision, recall, and area under the receiver operating characteristic (ROC) curve (AUC). After training and testing, we calculated these metrics to evaluate the performance for each technique, as each patient is classified as having had a stroke or not. These metrics were derived from the confusion matrices of both the training and test evaluations for each technique employed in the study.

The F1 score, a conventional metric for assessing model performance on a given dataset, is derived from precision and recall, with precision indicating the proportion of relevant true positives and recall indicating the proportion of true positives out of all true positives in the data. Additionally, the area under the receiver operating characteristic (ROC) curve (AUC) assesses a model's ability to discriminate between classes by computing the area under the curve of true positive rates plotted against false positive rates across different classification thresholds. The formulae for precision, recall, and F1 are given in the following Eqs. (5)–(7) [37].

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{5}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{6}$$

$$F1 = \frac{2\mathrm{PR}}{P + R} \tag{7}$$

4.4 Overall performance comparison

Table 2 shows the performance of our approach (GNN-FS) against the baseline (BS) using a 6-month timeframe. The results of other timeframe are included in the Sect. 4.5. The baseline approach includes all clinical variables without applying the proposed GNN-FS feature selection approach.

The findings reveal notable enhancements in accuracy across the majority of classification methods. Particularly, methods like multilayer perceptron (MLP), random forest (RF), k-nearest neighbors (k-NN), and logistic regression (LR) demonstrate significant improvements in accuracy when incorporating our graph-based feature selection. This suggests that this graph-based feature selection approach enhances the capture of predictive patterns. Among the different classification techniques evaluated, random forest consistently outperforms others in accuracy, regardless of the clinical feature selection strategy employed. Meanwhile, support vector machine (SVM) gains the best precision with or without applying the proposed GNN-FS feature selection approach.

4.5 Performance evaluation on timeframe and comparison with traditional feature selection

Our patient clinical feature graph is constructed based on their co-occurrences within a specified timeframe. While our initial experiments indicate superior performance compared to the baseline, conducting an ablation study is crucial to investigate how varying duration of timeframes influence our proposed graph-based clinical feature clustering for stroke prediction. Altering the timeframe impacts the number of edges and weights between clinical features, consequently affecting the clustering results and the identification of significant clinical features for stroke prediction. Initially, our approach utilized a 6-month window between nodes to establish edges and edge weights between clinical features. In this ablation study, we explore 1-month, 3-month, 1-year, and 2-year timeframes to assess the significance of temporal granularity and its implications for stroke prediction. For comparison, we employ traditional feature selection (traditional FS), specifically Chi-square feature selection [38],

Table 1 Statistics of the data set

	Stoke	Non-stroke	Total
# of patients age	92 (44.4%)	115 (55.5%)	207
Mean (std)	70.1	70.5	70.3
Gender			
Male (%)	48 (23.2%)	68 (32.8%)	116 (56.0%)
Female (%)	44 (21.3%)	47 (22.7%)	91 (44.0%)
Race			
White (%)	76 (36.7%)	99 (47.8%)	175 (84.5%)
African American (%)	15 (7.2%)	15 (7.2%)	30 (14.5%)
Other (%)	1 (0.5%)	1 (0.5%)	2 (1.0%)

Table 2 Performance comparison against the baseline (BS) models without GNN-FS using AUC, recall (R), precision (P), and F1

Model type	AUC		R		P		F1	
	BS	GNN-FS	BS	GNN-FS	BS	GNN-FS	BS	GNN-FS
RF	**0.82**	**0.85**	**0.72**	**0.89**	0.72	0.67	**0.72**	**0.76**
k-NN	0.63	0.71	0.33	0.39	0.75	0.58	0.46	0.47
SVM	0.81	0.81	0.56	0.61	**0.77**	**0.79**	0.65	0.69
MLP	0.71	0.74	0.61	0.67	0.65	0.63	0.63	0.65
LR	0.79	0.80	**0.72**	0.72	0.72	0.68	**0.72**	0.70

Bold numbers indicate the best performed model in terms of evaluation metrics

to identify a set of important clinical features based on training data and compare it with our clinical feature clustering approach. Chi-square feature selection has been widely utilized in clinical informatics research [38, 39] and has demonstrated favorable performance in various studies.

Given that the random forest classifier demonstrates superior performance compared to other classifiers, we opt to utilize it for our ablation study. The results are detailed in Table 3. Notably, our approach outperforms the baseline across all examined timeframes, except the 2-year timeframe. Analysis of the AUC and F1 scores reveals that extending the timeframe to 3 months or beyond enhances performance till 1-year timeframe. In our approach, maintaining the recall while increasing the timeframe from 3 to 12 months suggests consistent predictive power, while precision improves with the 12-month timeframe. The performance of our approach declined when a 2-year timeframe was applied, whereas the traditional FS maintained similar performance to the 1-year timeframe in terms of AUC and F1 scores. These results underscores the importance of temporal granularity in capturing stroke-related patterns effectively. The enhanced accuracy observed with a 12-month timeframe suggests critical relationships between clinical features for stroke prediction, emphasizing the necessity of incorporating extended temporal contexts in patient graph creation. However, when longer timeframe is employed the correlations between the clinical features are not as significant for stroke prediction. This phenomenon can also be explained by literature, indicating that high accuracy in stroke prediction may not be reliable for long-term forecasting before the event occurs [40]. However, it is crucial to acknowledge that the optimal timeframe may vary across diseases and datasets. Future research endeavors should explore diverse temporal resolutions for predicting other medical conditions to broaden the applicability of our findings.

4.6 Performance on subsets based on gender and race

Given the study cohort's higher proportion of male patients and predominantly white racial composition, we delve into model performances across sub-cohorts delineated by gender and race. Figures 3 and 4 show the AUC curves illustrating random forest performance using test data. Figure 3 indicates that although male patients slightly outnumber female patients in the study cohort, the model exhibits superior performance with male patients compared to females. We posit that this discrepancy may stem from male patients having more prominent clinical feature clusters than their female counterparts. Conversely, Fig. 4 reveals that our model performs more effectively with white patients. We attribute this outcome to the African American patient subset's smaller size, which potentially renders their clinical feature clusters less representative. It is important to note that the inherent imbalance in the dataset regarding gender and race could influence the differences in performance as well..

Table 3 Ablation study on duration of the timeframes using RF

Timeframe	# of clinical features	Method	AUC	Recall	Precision	F1
	1257	Baseline	0.82	0.83	0.60	0.70
1-month	583	Traditional FS	0.82	0.78	0.61	0.74
		GNN-FS	0.83	0.83	0.68	0.75
3-month	645	Traditional FS	0.81	0.78	0.67	0.72
		GNN-FS	0.86	0.89	0.70	0.78
6-month	687	Traditional FS	0.84	0.83	0.68	0.75
		GNN-FS	0.85	0.89	0.67	0.76
12-month	595	Traditional FS	0.84	0.78	0.61	0.68
		GNN-FS	**0.87**	**0.89**	**0.73**	**0.80**
24-month	540	Traditional FS	0.84	0.61	0.79	0.69
		GNN-FS	0.70	0.16	0.75	0.27

Bold numbers indicate the best performed timeframe in terms of AUC and F1

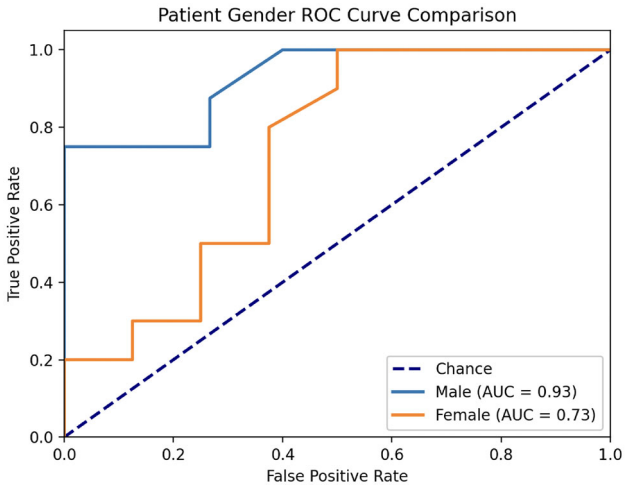

Fig. 3 Performance on different gender

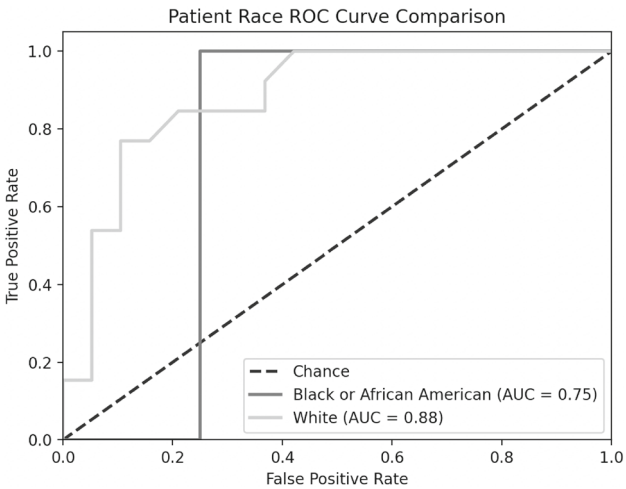

Fig. 4 Performance on different race

4.7 Analysis of the significant clinical features

After we generate clinical feature clusters based on all the training data, we identify the top frequent diagnosis, medication, and laboratory tests that are centroids of the clusters as significant clinical features associated with stroke or non-stroke patients. Tables 4, 5, and 6 show the top four clinical features of each category and their significance toward stroke population. The p values are calculated using Chi-square test.

Based on our literature review, these top clinical features are highly relevant to stroke. The disorders of lipoprotein metabolism and other lipidemias include LDL-hypercholesterolemia, hypertriglyceridemia, mixed hyperlipoproteinemia, lipoprotein deficiency, low HDL cholesterol, etc. Research [41] shows that cholesterol levels are associated with stroke, and diets high in fat-inducing hyperlipidemia are linked to alterations in brain calcium and lipid metabolism with susceptibility to stroke [42]. Hypertension is one of the most prevalent risk factors for stroke [43]. The hypertension-related stroke has been a research topic for many years in various populations [44–46]. These top diagnoses show a significant association with stroke, as indicated by the p values shown in Table 4.

From the list of medication groups, the literature shows that anticoagulant, antihyperlipedemics, and antihypertensives are the groups of medication often used to prevent stroke [47, 48]. Patients who take these medication often have associated diseases that increase the risks of developing stroke. Recent research [49] shows that hematological disorders can cause stroke which indicates that if a patient is taking medication in the category of hematological agents, he/she has a higher risk of developing a stroke. These top medications demonstrate a significant association with stroke, as indicated by the p values in Table 5.

The identified top laboratory tests are also relevant to stroke which is reflected by the previous literature. For example, there is an increased risk of stroke seen in hyperglycemic

Table 4 Top diagnoses related to stroke

Diagnosis	Stroke	Non-stroke	p value
E78: disorders of lipoprotein metabolism and other lipidemias	71 (71.2%)	68 (59.1%)	0.00046
I10: essential (primary) hypertension	69 (75.0%)	52 (45.2%)	$1.15 * 10^{-6}$
Z79: long-term (current) drug therapy	66 (71.7%)	53 (46.1%)	0.0011
Z86: personal history of certain other diseases	60 (65.2%)	55 (47.8%)	0.022

Table 5 Top medication related to stroke

Medication groups	Stroke	Non-stroke	p value
Anticoagulant	71	83	$1.27 * 10^{-6}$
Hematological agents	68 (73.9%)	63 (54.8%)	$1.19 * 10^{-7}$
Antihyperlipedemics	70 (76.1%)	61 (53.0%)	$9.18 * 10^{-6}$
Antihypertensives	58 (63.0%)	50 (43.5%)	$6.36 * 10^{-5}$

Table 6 Top laboratory test results related to stroke

Laboratory test results	Stroke	Non-stroke	p value
Glucose (high)	51 (55.4%)	73 (63.5%)	0.27
Low-density lipoprotein (LDL) cholesterol (high)	30 (32.6%)	27 (23.5%)	0.21
Blood urea nitrogen (BUN) (high)	46 (50.0%)	55 (47.9%)	0.91
Calcium total (low)	42 (45.7%)	46 (40.0%)	0.53

patients [50, 51]. Research shows that elevated levels of BUN are associated with greater risks of total stroke and ischemic stroke [52]. Excessive blood glucose can result in increased fatty deposits or clots in blood vessels and trigger stroke [50, 51]. Research shows that dietary calcium intake was found to be associated with a reduced incidence of stroke among men [53], and low calcium intake may contribute to increased risk of ischemic stroke in middle-aged American women [54]. High LDL (low-density lipoprotein) cholesterol level can potentially be associated with an increased risk of stroke [55]. However, the significance of these laboratory tests is not reflected by the calculated p values. Our hypothesis is that the characteristics of the patient cohort in this study may have led to the finding being not significant.

4.8 Case study

Since our approach is rooted in a graph-based model, it possesses the ability to generate visual representations that aid in exploring the connections among clinical features through the patient-level clinical feature graph. To demonstrate the practical applications, we present two case studies. These examples illustrate how the clinical feature graph and its clusters can help healthcare professionals interpret the implications of predictive outcomes.

The first case study, shown in Fig. 5, is a 76-year-old female patient who developed a stroke. This patient has four clinical feature clusters. The centroids of four clusters are medication group antihypertensives, diagnosis I10 essential hypertension, diagnosis F32 depressive episode, and medication group cephalosporins. Hypertension increases the risk of stroke through various physiological mechanisms, such as accelerating buildup of fats and cholesterol in the artery walls, as well as increasing intraluminal pressure and peripheral vascular resistance, all of which can cause cerebrovascular events [56]. Depression is associated with a higher risk of stroke as well as higher stroke mortality due to its neuroendocrine and inflammatory effects [57]. Cephalosporin does not appear to be associated with an increased risk of stroke. Indeed, the first three centroids (antihypertensives, I10 essential hypertension, diagnosis, and F32 depressive episode) are also seen in other patients who developed stroke. Hence, the clinical features within these three clusters are used to build the patient representations for prediction.

The second case study, shown in Fig. 6, is a 60-year-old male patient who developed a stroke. This patient has two clinical feature clusters. The centroids of these clusters are diagnosis E11 Type 2 diabetes mellitus, and medication analgesic. Type 2 diabetes can significantly increase the risk of ischemic stroke [58]. The literature shows that nonsteroidal

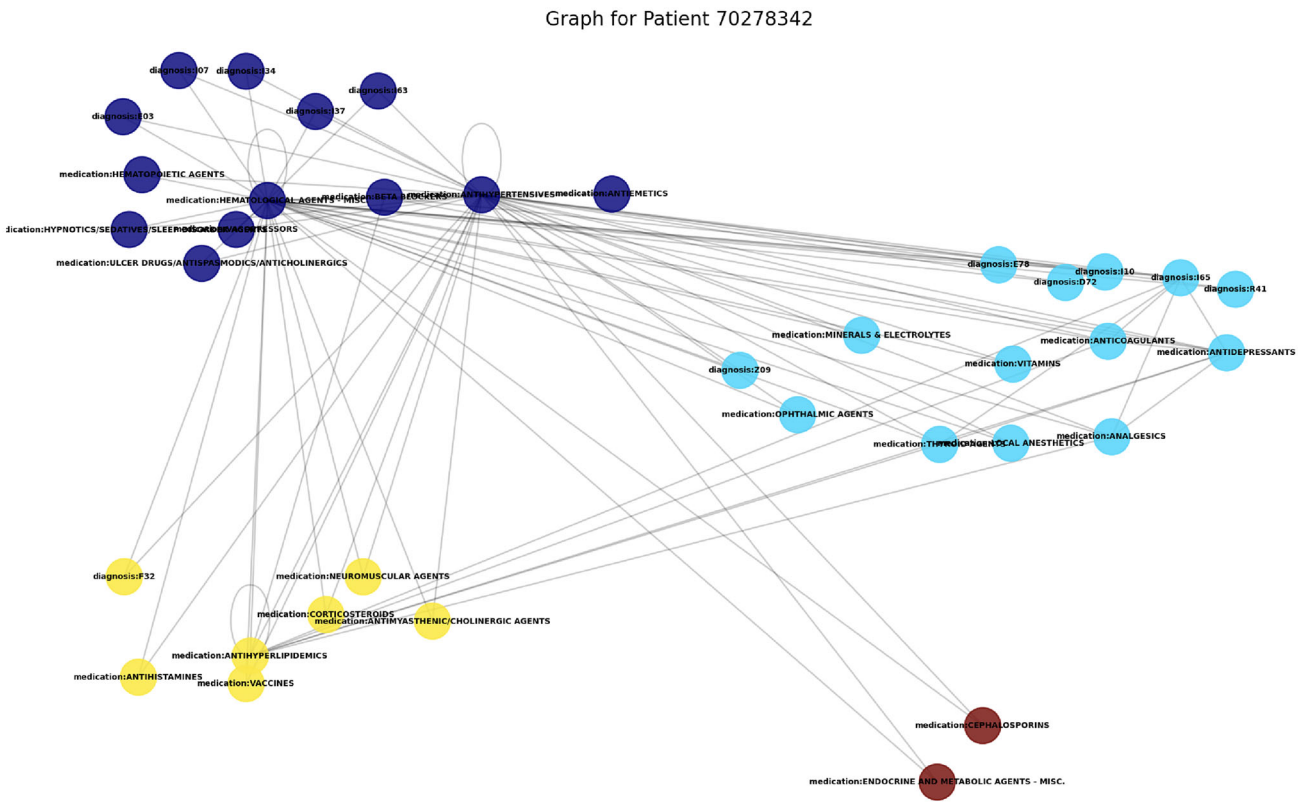

Fig. 5 76 Year-old female stroke patient

anti-inflammatory drugs, a class of analgesics, can increase the risks of stroke [59]. The clinical features within these two clusters are also included in the patient representations for prediction.

5 Conclusions, limitations, and future work

This study introduces a novel unsupervised GNN-based approach to identify critical clinical variables using positive cases, specifically the EHR data of ACS patients who developed a stroke. We constructed patient graphs that integrate patient diagnoses, medication history, and laboratory results, highlighting the intricate interconnections within a patient's EHR data. We then applied unsupervised graph clustering to select important clinical variables reflected in the data of stroke patients. By applying traditional supervised learning algorithms to these selected clinical variables, our model demonstrates promising outcomes in stroke prediction, surpassing conventional feature selection methods. This novel framework, which includes clinical variable selection and stroke prediction, can be easily adapted for other disease predictions using EHR data. The interconnections among clinical variables strongly correlate with the length and complexity of the medical history, impacting clinical variable selection. Our ablation study underscores the importance of considering temporal granularity in clinical variable selection for stroke prediction, specifically the number of months of a patient's medical history.

While our research demonstrates the effectiveness of using graph clustering for clinical variable selection in disease prediction, it is important to acknowledge its limitations. Firstly, our dataset is imbalanced, primarily consisting of white patients and more male patients. This imbalance poses challenges in optimizing performance for patients of other races or female patients. Although data balancing strategies can be applied to improve performance for minority groups, the size of our dataset is small which limits the generalizability of our findings to a larger population. Additionally, the absence of specific data types, such as socioeconomic factors, family history, and genetic information, may impede a comprehensive understanding of disease risk factors.

Subsequent research endeavors will involve the exploration of cutting-edge graph clustering methodologies, including dynamic graph clustering [60, 61], thereby further increasing the potential for uncovering intricate and dynamic associations within patient clinical feature graphs and potentially augmenting the precision of our disease prediction model. Additionally, expanding the graph construction to include family medical history and individual social history

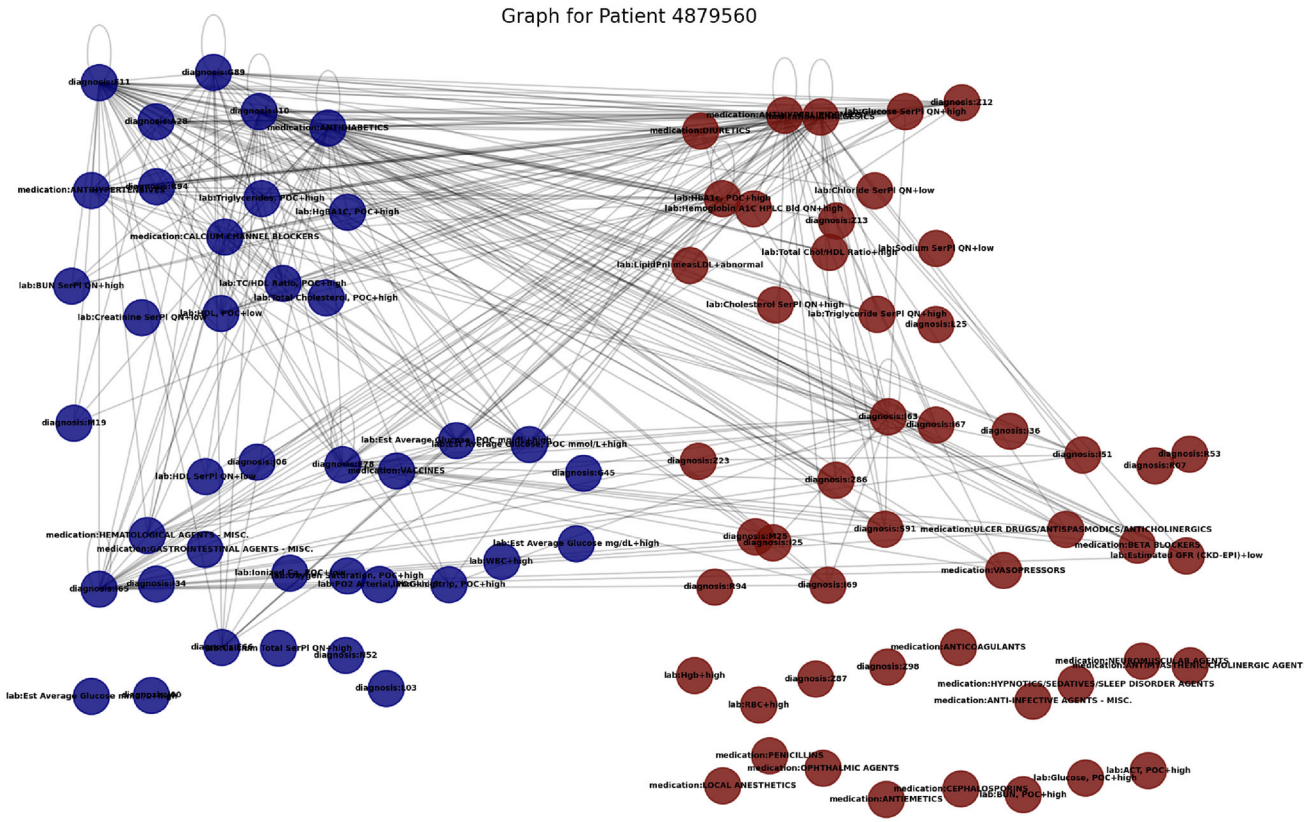

Fig. 6 60 Year-old male stroke patient

would contribute to a more comprehensive understanding of disease predisposition, ultimately enabling a more holistic and refined predictive framework. Lastly, advanced deep learning models [62–66], including transformer-based models, can be utilized to replace the traditional models used in this research to further improve the overall performance on prediction.

Acknowledgements This research is supported partially by Dr. Feng Li and Sheila Walter.

Author Contributions David Xu and Sanaz Matinmehr performed and collected the experimental results under guidance provided by Xiao Luo and Alan Sawchuk. All authors contributed to the manuscript in writing, editing, and reviewing. All authors have therefore read the manuscript.

Funding This research was supported by the National Science Foundation REU program (Award #1852105) and partially supported by the National Institute of Health (grant 1R15GM139094).

Data availability Because of the HIPAA requirement, the electronic health records cannot be published. Code will be published after the paper is accepted.

Declarations

Conflict of interest We have no conflict of interest to declare at the time of the submission of this publication.

Ethical approval This study was approved by the institutional review board (IRB) of Indiana University.

Consent to participate Not applicable

Consent for publication All authors consent to publication.

References

1. Tsao, C.W., Aday, A.W., Almarzooq, Z.I., Alonso, A., Beaton, A.Z., Bittencourt, M.S., Boehme, A.K., Buxton, A.E., Carson, A.P., Commodore-Mensah, Y., et al.: Heart disease and stroke statistics-2022 update: a report from the American heart association. Circulation **145**(8), 153–639 (2022)
2. Abbott, A.L.: Medical (nonsurgical) intervention alone is now best for prevention of stroke associated with asymptomatic severe carotid stenosis: results of a systematic review and analysis. Stroke **40**(10), 573–583 (2009)
3. Marquardt, L., Geraghty, O.C., Mehta, Z., Rothwell, P.M.: Low risk of ipsilateral stroke in patients with asymptomatic carotid stenosis on best medical treatment: a prospective, population-based study. Stroke **41**(1), 11–17 (2010)
4. Biasi, G.M., Froio, A., Diethrich, E.B., Deleo, G., Galimberti, S., Mingazzini, P., Nicolaides, A.N., Griffin, M., Raithel, D., Reid, D.B., et al.: Carotid plaque echolucency increases the risk of stroke in carotid stenting: the imaging in carotid angioplasty and risk of stroke (icaros) study. Circulation **110**(6), 756–762 (2004)
5. Khan, T.A., Shah, T., Prieto, D., Zhang, W., Price, J., Fowkes, G.R., Cooper, J., Talmud, P.J., Humphries, S.E., Sundstrom, J., et al.:

Apolipoprotein e genotype, cardiovascular biomarkers and risk of stroke: systematic review and meta-analysis of 14 015 stroke cases and pooled analysis of primary biomarker data from up to 60 883 individuals. Int. J. Epidemiol. **42**(2), 475–492 (2013)
6. Ding, L., Mane, R., Wu, Z., Jiang, Y., Meng, X., Jing, J., Ou, W., Wang, X., Liu, Y., Lin, J., et al.: Data-driven clustering approach to identify novel phenotypes using multiple biomarkers in acute ischaemic stroke: a retrospective, multicentre cohort study. EClinicalMedicine (2022). https://doi.org/10.1016/j.eclinm.2022.101639
7. Gronsbell, J., Minnier, J., Yu, S., Liao, K., Cai, T.: Automated feature selection of predictors in electronic medical records data. Biometrics **75**(1), 268–277 (2019)
8. Gajare, S., Sonawani, S.: Improved logistic regression approach in feature selection for EHR. In: Intelligent Systems Design and Applications: 18th International Conference on Intelligent Systems Design and Applications (ISDA 2018) Held in Vellore, India, December 6-8, 2018, vol. 1, pp. 325–334. Springer (2020)
9. Bhadra, S., Kumar, C.J.: Enhancing the efficacy of depression detection system using optimal feature selection from EHR. Comput. Methods Biomech. Biomed. Eng. **27**(2), 222–236 (2024)
10. Scheurwegs, E., Cule, B., Luyckx, K., Luyten, L., Daelemans, W.: Selecting relevant features from the electronic health record for clinical code prediction. J. Biomed. Inform. **74**, 92–103 (2017)
11. Chen, J., Aseltine, R.H., Wang, F., Chen, K.: Tree-guided rare feature selection and logic aggregation with electronic health records data. J. Am. Stat. Assoc. (2024). https://doi.org/10.1080/01621459.2024.2326621
12. Tsang, G., Zhou, S.-M., Xie, X.: Modeling large sparse data for feature selection: hospital admission predictions of the dementia patients using primary care electronic health records. IEEE J. Transl. Eng. Health Med. **9**, 1–13 (2020)
13. Lu, H., Uddin, S.: Disease prediction using graph machine learning based on electronic health data: a review of approaches and trends. Healthcare **11**, 1031 (2023)
14. Golmaei, S.N., Luo, X.: DeepNote-GNN: predicting hospital readmission using clinical notes and patient network. In: Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, pp. 1–9 (2021)
15. Tahabi, F.M., Storey, S., Luo, X.: Symptomgraph: identifying symptom clusters from narrative clinical notes using graph clustering. In: Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing, pp. 518–527 (2023)
16. Xiao, C., Pham, N., Imel, E., Luo, X.: Patient-gat: Sarcopenia prediction using multi-modal data fusion and weighted graph attention networks. In: Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing, pp. 614–617 (2023)
17. Işık, M., Velioğlu, Y.: Contribution of current comorbid conditions to carotid artery stenosis in patients undergoing coronary artery bypass and stroke distribution in carotid artery stenosis groups. Heart Surg. Forum **24**, 724–730 (2021)
18. Hirsch, J.S., Tanenbaum, J.S., Lipsky Gorman, S., Liu, C., Schmitz, E., Hashorva, D., Ervits, A., Vawdrey, D., Sturm, M., Elhadad, N.: Harvest, a longitudinal patient record summarizer. J. Am. Med. Inform. Assoc. **22**(2), 263–274 (2015)
19. Stirling, A., Tubb, T., Reiff, E.S., Grotegut, C.A., Gagnon, J., Li, W., Bradley, G., Poon, E.G., Goldstein, B.A.: Identified themes of interactive visualizations overlayed onto EHR data: an example of improving birth center operating room efficiency. J. Am. Med. Inform. Assoc. **27**(5), 783–787 (2020)
20. Anderson, A.E., Kerr, W.T., Thames, A., Li, T., Xiao, J., Cohen, M.S.: Electronic health record phenotyping improves detection and screening of type 2 diabetes in the general united states population: a cross-sectional, unselected, retrospective study. J. Biomed. Inform. **60**, 162–168 (2016)
21. Li, Q., Yang, X., Xu, J., Guo, Y., He, X., Hu, H., Lyu, T., Marra, D., Miller, A., Smith, G., et al.: Early prediction of Alzheimer's disease and related dementias using real-world electronic health records. Alzheimer's Dement. **19**(8), 3506–3518 (2023)
22. Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
23. Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining thepredictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144 (2016)
24. Wang, R.-H., Luo, T., Zhang, H.-L., Du, P.-F.: Pla-gnn: Computational inference of protein subcellular location alterations under drug treatments with deep graph neural networks. Computers in Biology and Medicine **157**, 106775 (2023)
25. R'eau, M., Renaud, N., Xue, L.C., Bonvin, A.M.: Deeprank-gnn: a graph neural network framework to learn patterns in protein–protein interfaces. Bioinformatics **39**(1), 759 (2023)
26. Vilela, J., Asif, M., Marques, A.R., Santos, J.X., Rasga, C., Vicente, A., Martiniano, H.: Biomedical knowledge graph embeddings for personalized medicine: Predicting disease-gene associations. Expert Systems **40**(5), 13181 (2023)
27. Dev, S., Wang, H., Nwosu, C.S., Jain, N., Veeravalli, B., John, D.: A predictive analytics approach for stroke prediction using machine learning and neural networks. Healthc. Anal. **2**, 100032 (2022)
28. Nwosu, C.S., Dev, S., Bhardwaj, P., Veeravalli, B., John, D.: Predicting stroke from electronic health records. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 5704–5707. IEEE (2019)
29. Hung, C.-Y., Chen, W.-C., Lai, P.-T., Lin, C.-H., Lee, C.-C.: Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database. In: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 3110–3113. IEEE (2017)
30. Xie, Y., Yang, H., Yuan, X., He, Q., Zhang, R., Zhu, Q., Chu, Z., Yang, C., Qin, P., Yan, C.: Stroke prediction from electrocardiograms by deep neural network. Multimed. Tools Appl. **80**, 17291–17297 (2021)
31. Reddy, M.K., Kovuri, K., Avanija, J., Sakthivel, M., Kaleru, S.: Brain stroke prediction using deep learning: a CNN approach. In: 2022 4th International Conference on Inventive Research in Computing Applications (ICIRCA), pp. 775–780. IEEE (2022)
32. Clapp, Mark A., James, Kaitlyn E., Friedman, Alexander M.: Identification of delivery encounters using international classification of diseases, tenth revision, diagnosis and procedure codes. Obstet. Gynecol. **136**(4), 765–767 (2020)
33. Grover, A., Leskovec, J.: node2vec: Scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 855–864 (2016)
34. Shahapure, K.R., Nicholas, C.: Cluster quality analysis using silhouette score. In: 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), pp. 747–748 (2020). IEEE
35. Liu, F., Deng, Y.: Determine the number of unknown targets in open world based on elbow method. IEEE Trans. Fuzzy Syst. **29**(5), 986–995 (2020)
36. Howard, D.P., Gaziano, L., Rothwell, P.M.: Risk of stroke in relation to degree of asymptomatic carotid stenosis: a population-based cohort study, systematic review, and meta-analysis. Lancet Neurol. **20**(3), 193–202 (2021)
37. Mohri, M., Rostamizadeh, A., Talwalkar, A.: Foundations of Machine Learning. The MIT Press, Cambridge (2012)

38. Gárate-Escamila, A.K., El Hassani, A.H., Andrès, E.: Classification models for heart disease prediction using feature selection and PCA. Inform. Med. Unlocked **19**, 100330 (2020)
39. Spencer, R., Thabtah, F., Abdelhamid, N., Thompson, M.: Exploring feature selection and classification methods for predicting heart disease. Digit. Health **6**, 2055207620914777 (2020)
40. Kasabov, N., Feigin, V., Hou, Z.-G., Chen, Y., Liang, L., Krishnamurthi, R., Othman, M., Parmar, P.: Evolving spiking neural networks for personalised modelling, classification and prediction of spatio-temporal patterns with a case study on stroke. Neurocomputing **134**, 269–279 (2014)
41. Ali, M.T., Martin, S.S.: Disorders of lipid metabolism. In: Aronow, W.S., Fleg, J.L., Fleg, J.L., Rich, M.W., Rich, M.W. (eds.) Tresch and Aronow's Cardiovascular Disease in the Elderly, pp. 111–127. CRC Press, Florida (2019)
42. Martins, I.J., Creegan, R.: Links between insulin resistance, lipoprotein metabolism and amyloidosis in Alzheimer's disease. Health (2014). https://doi.org/10.4236/health.2014.612190
43. Wajngarten, M., Silva, G.S.: Hypertension and stroke: update on treatment. Eur. Cardiol. Rev. **14**(2), 111 (2019)
44. Gorgui, J., Gorshkov, M., Khan, N., Daskalopoulou, S.S.: Hypertension as a risk factor for ischemic stroke in women. Can. J. Cardiol. **30**(7), 774–782 (2014)
45. Dubow, J., Fink, M.E.: Impact of hypertension on stroke. Curr. Atheroscler. Rep. **13**, 298–305 (2011)
46. Singh, R., Suh, I., Singh, V., Chaithiraphan, S., Laothavorn, P., Sy, R., Babilonia, N., Rahman, A., Sheikh, S., Tomlinson, B., et al.: Hypertension and stroke in Asia: prevalence, control and strategies in developing countries for prevention. J. Hum. Hypertens. **14**(10), 749–763 (2000)
47. Diener, H.-C., Aisenberg, J., Ansell, J., Atar, D., Breithardt, G., Eikelboom, J., Ezekowitz, M.D., Granger, C.B., Halperin, J.L., Hohnloser, S.H., et al.: Choosing a particular oral anticoagulant and dose for stroke prevention in individual patients with non-valvular atrial fibrillation: part 2. Eur. Heart J. **38**(12), 860–868 (2017)
48. Freedman, B., Potpara, T.S., Lip, G.Y.: Stroke prevention in atrial fibrillation. Lancet **388**(10046), 806–817 (2016)
49. Markus, H.S.: Hematological Disorders and Stroke. SAGE Publications, London (2023)
50. Wannamethee, S.G., Perry, I.J., Shaper, A.G.: Nonfasting serum glucose and insulin concentrations and the risk of stroke. Stroke **30**(9), 1780–1786 (1999)
51. Sui, X., Lavie, C.J., Hooker, S.P., Lee, D.-C., Colabianchi, N., Lee, C.-D., Blair, S.N.: A prospective study of fasting plasma glucose and risk of stroke in asymptomatic men. Mayo Clin. Proc. **86**, 1042–1049 (2011)
52. Peng, R., Liu, K., Li, W., Yuan, Y., Niu, R., Zhou, L., Xiao, Y., Gao, H., Yang, H., Zhang, C., et al.: Blood urea nitrogen, blood urea nitrogen to creatinine ratio and incident stroke: the Dongfeng–Tongji cohort. Atherosclerosis **333**, 1–8 (2021)
53. Adebamowo, S.N., Spiegelman, D., Flint, A.J., Willett, W.C., Rexrode, K.M.: Intakes of magnesium, potassium, and calcium and the risk of stroke among men. Int. J. Stroke **10**(7), 1093–1100 (2015)
54. Iso, H., Stampfer, M.J., Manson, J.E., Rexrode, K., Hennekens, C.H., Colditz, G.A., Speizer, F.E., Willett, W.C.: Prospective study of calcium, potassium, and magnesium intake and risk of stroke in women. Stroke **30**(9), 1772–1779 (1999)
55. DeBaun, M.R., Sarnaik, S.A., Rodeghier, M.J., Minniti, C.P., Howard, T.H., Iyer, R.V., Inusa, B., Telfer, P.T., Kirby-Allen, M., Quinn, C.T., et al.: Associated risk factors for silent cerebral infarcts in sickle cell anemia: low baseline hemoglobin, sex, and relative high systolic blood pressure. Blood J. Am. Soc. Hematol. **119**(16), 3684–3690 (2012)
56. Johansson, B.B.: Hypertension mechanisms causing stroke. Clin. Exp. Pharmacol. Physiol. **26**(7), 563–565 (1999)
57. Pan, A., Sun, Q., Okereke, O.I., Rexrode, K.M., Hu, F.B.: Depression and risk of stroke morbidity and mortality: a meta-analysis and systematic review. Jama **306**(11), 1241–1249 (2011)
58. Janghorbani, M., Hu, F.B., Willett, W.C., Li, T.Y., Manson, J.E., Logroscino, G., Rexrode, K.M.: Prospective study of type 1 and type 2 diabetes and risk of stroke subtypes: the Nurses' health study. Diabetes Care **30**(7), 1730–1735 (2007)
59. Vives, R., Gomez-Lumbreras, A., Fradera, M., Giner-Soriano, M., Garcia-Sangenis, A., Marsal, J., Morros, R.: Risk of ischemic stroke associated to analgesic drugs use: a real world data case-control study. Osteoarthr. Cartil. **26**, 225 (2018)
60. Tsitsulin, A., Palowitch, J., Perozzi, B., Müller, E.: Graph clustering with graph neural networks. J. Mach. Learn. Res. **24**(127), 1–21 (2023)
61. Tahabi, F.M., Luo, X.: Dynamicg2b: dynamic node classification with layered graph neural networks and BiLSTM. In: The International FLAIRS Conference Proceedings, vol. 36 (2023)
62. Hu, Z., Wang, Z., Jin, Y., Hou, W.: VGG-TSwinformer: transformer-based deep learning model for early Alzheimer's disease prediction. Comput. Methods Programs Biomed. **229**, 107291 (2023)
63. Dileep, P., Rao, K.N., Bodapati, P., Gokuruboyina, S., Peddi, R., Grover, A., Sheetal, A.: An automatic heart disease prediction using cluster-based bi-directional LSTM (C-BiLSTM) algorithm. Neural Comput. Appl. **35**(10), 7253–7266 (2023)
64. Alshehri, A., Owais, M., Gyani, J., Aljarbou, M.H., Alsulamy, S.: Residual neural networks for origin-destination trip matrix estimation from traffic sensor information. Sustainability **15**(13), 9881 (2023)
65. Owais, M.: Deep learning for integrated origin-destination estimation and traffic sensor location problems. IEEE Trans. Intell. Transp. Syst. (2024). https://doi.org/10.1109/TITS.2023.3344533
66. Owais, M., Alshehri, A., Gyani, J., Aljarbou, M.H., Alsulamy, S.: Prioritizing rear-end crash explanatory factors for injury severity level using deep learning and global sensitivity analysis. Expert Syst. Appl. **245**, 123114 (2024)

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH (“Springer Nature”).
Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users (“Users”), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use (“Terms”). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.
These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.
We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.
While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval , sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.
These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.
To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.
Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.
If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com