Enhanced Disease Detection via Graph Clustering and Centroid-based Representations

David Xu*
davidxu2026@u.northwestern.edu
Northwestern University
Evanston, IL, US

Sanaz Matinmehr*
sbardoo@iu.edu
Indiana University-Purdue University
Indianapolis
Indianapolis, IN, USA

Xiao Luo
luo25@iupui.edu
Indiana University-Purdue University
Indianapolis
Indianapolis, IN, USA

ABSTRACT

According to the World Health Organization (WHO), 15 million people around the world become victims of a stroke every year. Of those 15 million people, 5 million succumb to the disease and another 5 million who survive become permanently disabled. Though it may not be common for those under the age of 40, anyone can become susceptible to a stroke, as factors such as high blood pressure can play a major role increasing one's risk. It is crucial to identify such symptoms to discover patterns within a patient that may lead to a stroke. In this paper, we utilize five different graph clustering techniques to analyze patient data in order to evaluate and find the most effective and accurate method of detecting a stroke. Patient data from both stroke and non-stroke patients are used to identify phenotypes, which then form clusters to uncover overlapping patterns in the symptoms. Finally, we conclude our discussion with our final thoughts to the study and other methods that could be evaluated in the future.

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

KEYWORDS

Graph Clustering, Node2Vec, Random Forest, Stroke, K-Means

ACM Reference Format:
David Xu, Sanaz Matinmehr, and Xiao Luo. 2023. Enhanced Disease Detection via Graph Clustering and Centroid-based Representations. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

A cerebrovascular accident, most commonly known as a stroke, is one of the most fatal brain diseases worldwide. It functions in a similar fashion as a heart attack, except for the primary location of its function-the brain. The most common type, an ischemic stroke is triggered due to the lack of blood flow to the brain, cutting off the precious oxygen and nutrients needed to function. The less common, however more catastrophic type of stroke is called a hemorrhagic stroke, which occurs when the blood vessel in the brain completely ruptures. The impact of the stroke may result from many different reasons, most frequently from the buildup of fatty deposits within the walls of the blood vessel, which then continue to narrow over time until it is fully blocked or bursts. According to the Center for Disease Control, about 1 in 6 deaths from cardiovascular disease were solely due to strokes in 2021. [https://www.cdc.gov/stroke/facts.htm]

The use of machine and deep learning techniques is no stranger within the medical world. As such, using artificial intelligence to assist medical professionals with the detection of diseases prior to its arrival [https://link.springer.com/article/10.1007/s11227-020-03481-x] is a widely known and exploited concept. Dependent on the information being fed into the model, it can prove to produce ground-breaking predictions that could potentially save the lives of millions who may or may not be aware to their susceptibility to a disease.

In this paper, we present our findings based on our attempt to discover the most accurate and efficient method of detecting stroke. We offer the following contributions:

- We begin by normalizing based on medications, lab work, and diagnosis', which are then compared directly to the demographics pool. After the data has been analyzed, we split our data into a training data set and a testing data set, to create graphs.
- In addition to our initial weight of 6 months, within these graphs, we have also added nodes and edges are being produced with varying weights from 1 month, 3 months, and 1 year.
- We then apply Node2Vec to each time frame and applied a silhouette score analysis to ultimately rehabilitate the data to form the patients' graph cluster and centroids.
- Based on the clusters that are formed, we find an assemblage of the phenotypes (nodes) vectors, which are then transformed into a binary representation of that data. The binary data representation is then fed into our chosen classifiers and processed to be exhibited.
- Finally, we display our findings and draw our conclusions from comparisons made from the given data. Additionally,

[*] Both authors contributed equally to this research.

Conference'17, July 2017, Washington, DC, USA
© 2023 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

we offer an ablation study and case study in which we demonstrate stroke patients' graph clusters to offer insight into the process.

2 RELATED WORKS

Below, we offer relevant works pertaining to our topic of study, including expanding upon the methods that will be utilized in this paper. We will then ultimately draw comparisons between each method to identify the most robust method of stroke disease detection.

Disease Detection Models

As the field of machine learning continues to be evolve and expand, as does its desire to be exploited. As mentioned, the use of machine learning methods and ai is not new with respects to its application in disease detection. Some [] may even have the potential for success, such as this study that was able to forecast patient symptoms to find possible underlying illnesses. This method also uses demographic information with conclusions drawn from patterns seen in their medical data. However, this [repeat cite] provides the added bonus of extending past stroke prediction to other life-threatening illnesses. There is also the consideration of unique life-style factors that each patient experiences through their lives. [https://link.springer.com/article/10.1007/s11227-020-03481-x] There is the added complexity of understand that each person has their own unique experience, whether in regards to impacts on illnesses or external factors from their environment. Each of those patients process it differently, especially when it pertains to a current timeline which is always frequently evolving. This leaves the room for critique on such methods and how they are able to keep up with, and how well they can correctly identify a possible disease.

Additional Clustering-based Methods There is also the integration of more up-to-date methods, such as [] XGBoost, a classifier that is younger in comparison to random forest. Random forest was heavily employed within this study, and given for any future improvements or expansions upon this topic, other classifiers like XGBoost can offer possible improvements when continuing this experiment. In fact, there has been recent data showing the use of methods such as XGBoost [https://onlinelibrary.wiley.com/doi/epdf/10.4218/etrij.2022-0271] for stroke prediction as a fairly decent prediction method, along with Naive Bayes, Decision Tree, Lightgbm, and Cat Boost.

3 PROPOSED METHOD

In this section, we begin our discussion over our proposed methodology. We expand upon our framework and how our data collection procedure and its contribution into the construction of the overall graphs.

3.1 Notations

Let our graph $G = (V, E)$ be defined by a set of nodes $V = (v_1, \ldots, v_n)$, $|V| = n$, and edges $E \subseteq V \times V$, $|E| = m$:

Node2Vec [https://arxiv.org/pdf/1607.00653.pdf] Node2Vec has the potential to express the diversity of how patterns that are formed within the data connect in a meaningful fashion to one another [https://arxiv.org/pdf/1607.00653.pdf]. This method was inintally introduced in 2016 by A. Grover and J. Leskovec and possess short, yet in consequential scope.

The first step is implemented through the use of generating Random Walks to explore the entirety of the graph. For each node v in V, Node2Vec generates multiple random walks of a specified length. The walks are generated by transitioning between nodes based on predefined transition probabilities, which are designed to balance between exploring local neighborhoods and exploring distant parts of the graph. The transition probabilities are controlled by two parameters: p and q. The parameter p encourages the random walk to revisit nodes it has previously visited, while q encourages the random walk to explore uncharted territory.

The second and final step involves learning node embeddings through the use of Skip-Gram Models. The main idea is idea is to predict context (neighboring) nodes given a target node. Each node is represented as a one-hot encoded vector in the input layer. The skip-gram model is trained using stochastic gradient descent to maximize the likelihood of predicting context nodes for each target node. The optimization process adjusts the node embeddings in a way that similar nodes have similar embeddings. The optimization of node embeddings using skip-gram models effectively captures the structural characteristics of the graph, such as node similarities and relationships.

Silhouette Analysis

The silhouette score is a metric used to evaluate the quality of clusters in unsupervised machine learning, such as clustering algorithms like K-means. It measures how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette score for a data point is calculated using the formula:

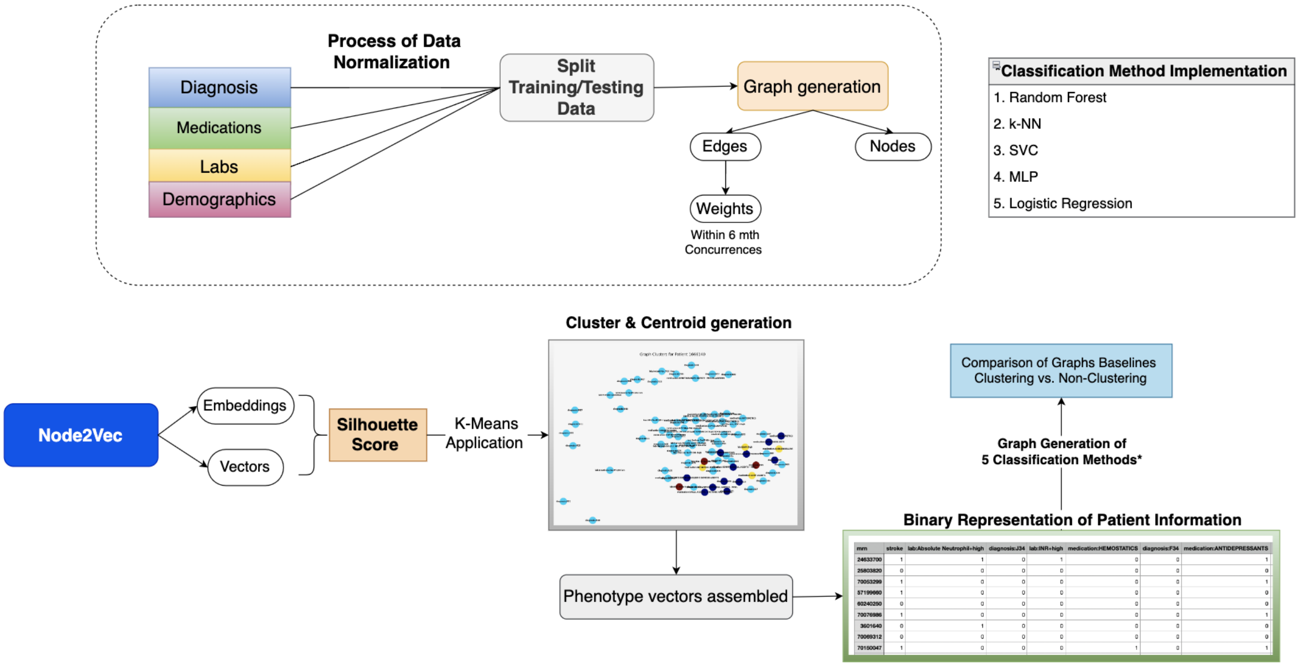
$$s = \frac{b - a}{\max(a, b)}$$

Where a is the average distance between the data point and all other data points in the same cluster and b is the smallest average distance between the data point and any other cluster. The silhouette score ranges from -1 to +1, with a higher score indicating better cluster quality.

ROC Curve Receiver Operating Characteristic (ROC) Curves, has been widely used in medical data, especially within the area of radiology. It has proven to show promising results when used in the comparison of different data processing methods, and especially for prediction purposes [cite]. Specifically, in this study, we will be using ROC curves to assess and compare the overall performance of our chosen classifiers to see how accurately is can determine the risk of a disease arising. To fully understand the application of ROC curves to one's data, we must take into consideration some of its elements and how it affects the overall results [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8831439/]. The necessary features to achieving the best ROC curve is through the analysis of the sensitivity, specificity, false positives, and false negative aspects of the data. Below, we define our True Positive Rate and our False Positive Rate:

True Positive Rate (TPR):

$$TPR = \frac{\text{True Positives}}{\text{True Positives + False Negatives}}$$


Process of Data Normalization
Diagnosis
Medications
Labs
Demographics
Split Training/Testing Data
Graph generation
Edges
Nodes
Weights
Within 6 mth Concurrences
Classification Method Implementation
1. Random Forest
2. k-NN
3. SVC
4. MLP
5. Logistic Regression
Node2Vec
Embeddings
Vectors
Silhouette Score
K-Means Application
Cluster & Centroid generation
Phenotype vectors assembled
Comparison of Graphs Baselines Clustering vs. Non-Clustering
Graph Generation of 5 Classification Methods*
Binary Representation of Patient Information

mrn	stroke	lab:Absolute Neutrophil+high	diagnosis:J34	lab:INR+high	medication:HEMOSTATICS	diagnosis:F34	medication:ANTIDEPRESSANTS
24633700	1	1	0	1	0	0	1
25803820	0	0	0	0	0	0	0
70053299	1	0	0	0	0	0	1
57199660	1	0	0	0	0	0	0
60240250	0	0	0	0	0	0	0
70076986	1	0	0	0	0	0	1
3601640	0	1	0	0	0	0	0
70069312	0	0	0	0	0	0	0
70150047	1	0	0	0	1	0	1

Figure 1: Pipeline explanation

False Positive Rate (FPR):

$$FPR = \frac{\text{False Positives}}{\text{False Positives + True Negatives}}$$

3.2 Pre-processing data

Our model takes patient diagnoses, medications, and lab results as its inputs. This data is then normalized to consider only patients that are defined as stroke or non-stroke patients. We split the patients into training and testing sets and generate graph representations for all patients.

Nodes are defined as the diagnosis, medication, or lab results. Edges are formed when two nodes co-occur within six months and the edge weights are determined by the number of co-occurrences within six months. For each stroke patient in the training dataset, we generate node embeddings using Node2Vec and identify clusters and centroids using K-means.

The optimal number of clusters for each patient is determined using the silhouette score. We identify overlapping centroids between stroke patients and use the nodes within these clusters to generate a phenotype vector to generate the foundation of the vector representation for each patient.

We then use both training and test data to generate vector representation for each patient. If a patient has the signature phenotype it is classified as 1, otherwise it is classified as 0. For the baseline model, we do not perform graph clustering and generate the vector representation foundation using all phenotypes present in the training data. The same method as above is used to form the binary representation for all patients. Finally, we run random forest and four other classification methods on the training data and then on the testing data to evaluate the performance of the model.

4 EXPERIMENTAL SETTINGS

Table 1: Patient Cohort Demographics

Demographic	Number
Gender	
Male	116
Female	91
Race	
White	175
Black	30
Asian	1
Unknown	1
Ethnicity	
Hispanic/Latino	2
Non-Hispanic/Latino	205
Age	
Minimum Age	46
Maximum Age	90
Average Age	70.31

4.1 Data set

For our experiment, we used a proprietary dataset obtained from the Indiana University Medical Hospital. It contains 208 patient's demographic information and diagnosis, medication, and lab results data for several hundred patients. Each diagnosis, medication, and lab results was associated with a date. Out of the 208 patients, 92 are considered stroke patients. Any patient with a stroke diagnosis code, a Transient ischemic attack (TIA), stroke, or Amaurosis, or doctor notes was considered a stroke patient. To generate our patient graph representations, we removed all patients from the diagnosis, medication, and lab result data sits that were not part of the 208 patients in the demographic dataset.

4.2 Baselines and Parameter settings

We compare the performance of 5 machine learning classification models for a vector representation generated with and without graph clustering methods:

1. **Random Forest:** 100 forest trees, a minimum sample split of 2, and no maximum depth
2. **k-Nearest Neighbors:** k = 21, uniform weight, and leaf size of 30
3. **Support Vector Classification (SVP):** regularization parameter or 1.0, an RBF kernel, and a degree of 3
4. **Multilayer Perceptron (MLP):** hidden layer size of (100, 50) and 1000 max iterations
5. **Logistic Regression:** l2 penalty, primal formulation, and a stopping tolerance of 0.0001

5 EXPERIMENTAL RESULTS

5.1 Performance and Comparison

View table 1 for comparison results

Our approach utilizing graph-based clustering to construct patient vectors demonstrated marked enhancements in classification accuracy compared to a baseline approach. The graph clustering technique allowed us to create vector representations that captured intricate medical attribute relationships, resulting in improved accuracy.

Comparing our approach with the baseline, where patient vectors were constructed without graph clustering, significant accuracy improvements were observed across most classification methods. Methods such as Random Forest, Multi-Layer Perceptron (MLP), Linear Regression, and k-Nearest Neighbors (k-NN) exhibited notably increased accuracy with our graph-based approach. This suggests that shared phenotypic clusters in patient vectors better captured predictive patterns.

Interestingly, the Support Vector Machine (SVM) method maintained consistent accuracy levels with both approaches. This finding underscores SVM's resilience to noise and outliers.

Our study highlights the utility of phenotypic clustering in constructing patient vectors. Enhanced accuracy indicates the importance of incorporating shared phenotypic clusters for more accurate stroke prediction, offering potential for improved personalized treatments and outcomes.

5.2 Case Study

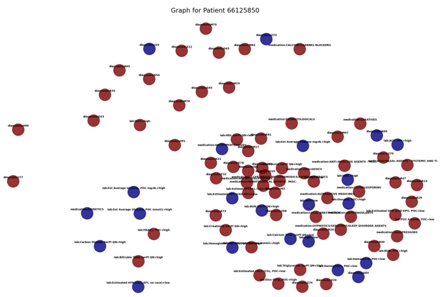


Figure 2: 79 year old White male stroke patient #6612850

For our first case study we use the patient with a MRN of 66125850 from our dataset. This patient is a 79 year-old White male who has developed stroke. This patient has two phenotype clusters. The centroids of these clusters are Antihypertensive medication and an ICD-10 diagnosis code of I10 which represents essential hypertension. Hypertension, commonly known as high blood pressure is known to be a main risk factor for stroke. The relationship between hypertension and stroke is multifaceted, involving various physiological mechanisms that contribute to the increased risk of cerebrovascular events.

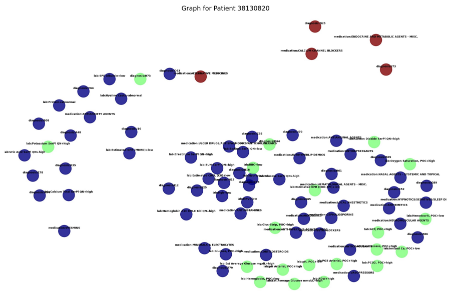


Figure 3: 76 year old White female stroke patient #38130820

We will now look at the patient with MRN of 38130820. This patient is a 76 year-old White female who has developed stroke. This patient has three phenotype clusters. The centroids of these clusters are a lab test of high high uric acid, an ICD-10 diagnosis code of D64, which represents anemia, and endocrine and metabolic agent medication. The connection between anemia and increased stroke risk lies in the disruption of normal blood flow dynamics, the potential for thrombus formation, and the resulting tissue oxygen deprivation.

5.3 Ablation study

We conducted an ablation study to explore how different time frames influence our proposed graph-based clustering approach's accuracy. The study involved varying the temporal window for creating edges between medical attributes in patient graphs. We explored 1-month, 3-month, and 1-year time frames to understand their impact on predicting developed stroke.

Table 2: Model Performance Comparison

Model Type	AUC (Baseline)	AUC (Clustering)
Random Forest	0.82	0.85
k-NN	0.72	0.77
SVC	0.81	0.81
MLP	0.71	0.74
Logistic Regression	0.79	0.80

For each time frame, we adjusted the edge creation process while keeping other methodology aspects constant. The baseline approach used a 6-month window, while the 1-month, 3-month, and 1-year settings considered relevant medical attributes within those time frames.

The Random Forest classification consistently demonstrated the highest accuracy. Interestingly, we observed that the 3-month time frame increased accuracy, while the 1-month frame decreased it compared to the 6-month baseline. The most significant improvement occurred with a 1-year time frame. This outcome underscores the importance of temporal granularity in capturing stroke-related patterns. The enhanced accuracy with a 1-year frame suggests stable phenotypic relationships over longer periods, emphasizing the value of extended temporal context in constructing patient graphs. This success highlights the significance of considering long-term medical trends for stroke prediction. However, it's crucial to acknowledge that the optimal time frame may vary across diseases and datasets. Future studies should explore temporal resolutions for other medical conditions to validate the generalizability of our findings.

5.4 Traditional Feature Selection

To refine the accuracy of our baseline approach, we introduced traditional feature selection techniques, specifically leveraging the chi-squared test. Our aim was to enhance the stroke prediction model by identifying the most pertinent medical attributes.

The feature selection process was specifically applied to the baseline model, ensuring the integrity of our novel graph-based clustering methodology. Just before employing the Random Forest classifier, we independently conducted feature selection on both training (X_train) and testing (X_test) datasets to safeguard model generalization.

The chi-squared test, assessing the independence between categorical variables, guided our selection process. We gauged each phenotype's association with the target variable (stroke or non-stroke). Elevated chi-squared values denoted stronger associations, implying a phenotype's potential as a stroke predictor.

Selecting phenotypes with the highest chi-squared scores—indicative of robust stroke associations—we curated a subset. This selection, performed separately for training and testing data, averted data leakage and maintained fairness in evaluation.

Employing the chi-squared test-based feature selection and consequent vector reduction, we witnessed significant enhancement in the baseline model's performance. This streamlined model, comprising relevant phenotypes, surpassed the original baseline encompassing all attributes. This approach showcased the potential of targeted phenotypic focus to elevate model discernment.

6 LIMITATIONS AND FUTURE WORK

Our study, while contributing valuable insights, is not without limitations. Firstly, the study cohort is predominantly composed of individuals from a specific racial background, potentially limiting the ability to generalize our findings. Moreover, the absence of certain data types, such as genetic information and socioeconomic factors, may hinder a comprehensive understanding of disease risk factors. Additionally, the temporal resolution explored in our ablation study offers a limited snapshot of potential disease progression patterns.

Looking ahead, future research could address these limitations and propel the field forward. Exploring advanced graph-based techniques like Graph Neural Networks (GNNs) could unlock complex relationships within patient graphs, potentially enhancing the accuracy of our disease prediction model. Furthermore, incorporating patient demographics, including race, gender, age, and ethnicity, could provide a more nuanced view of disease risk and ensure fairness in predictive outcomes. Additional considerations to the patient's living environment, along with potential diaspora, there is the possibility of providing more insight of how their socioeconomic background can come into play. Enriching the dataset by integrating patient family medical history and individual medical history would provide a more holistic perspective on disease predisposition, enabling a more comprehensive predictive framework.

7 CONCLUSION

In conclusion, our study introduces a novel approach to disease prediction using graph clustering techniques, demonstrating its potential in identifying shared phenotype clusters among developed stroke patients. The integration of patient diagnosis, medication history, and lab results within patient graphs sheds light on complex relationships within medical data. Through graph embeddings and clustering, our model showcases promising results for disease prediction, outperforming traditional methods. The insights gained from the ablation study emphasize the importance of temporal granularity.

As we pave the way for future research, the integration of advanced techniques like GNNs, the consideration of patient demographics, and the inclusion of medical history emerges as paramount. These factors hold the promise of crafting a more comprehensive and accurate disease prediction framework, ensuring equitable and personalized healthcare interventions.

Our commitment to enhancing medical data analysis remains unwavering, with the goal of providing healthcare professionals with powerful tools for improved disease prognosis and tailored care.

ACKNOWLEDGMENTS

This research was supported by the National Science Foundation REU program and the National Institute of Health (grants 1R01AR077273-01 and 1R15GM139094-01A1). We would like to extend our gratitude

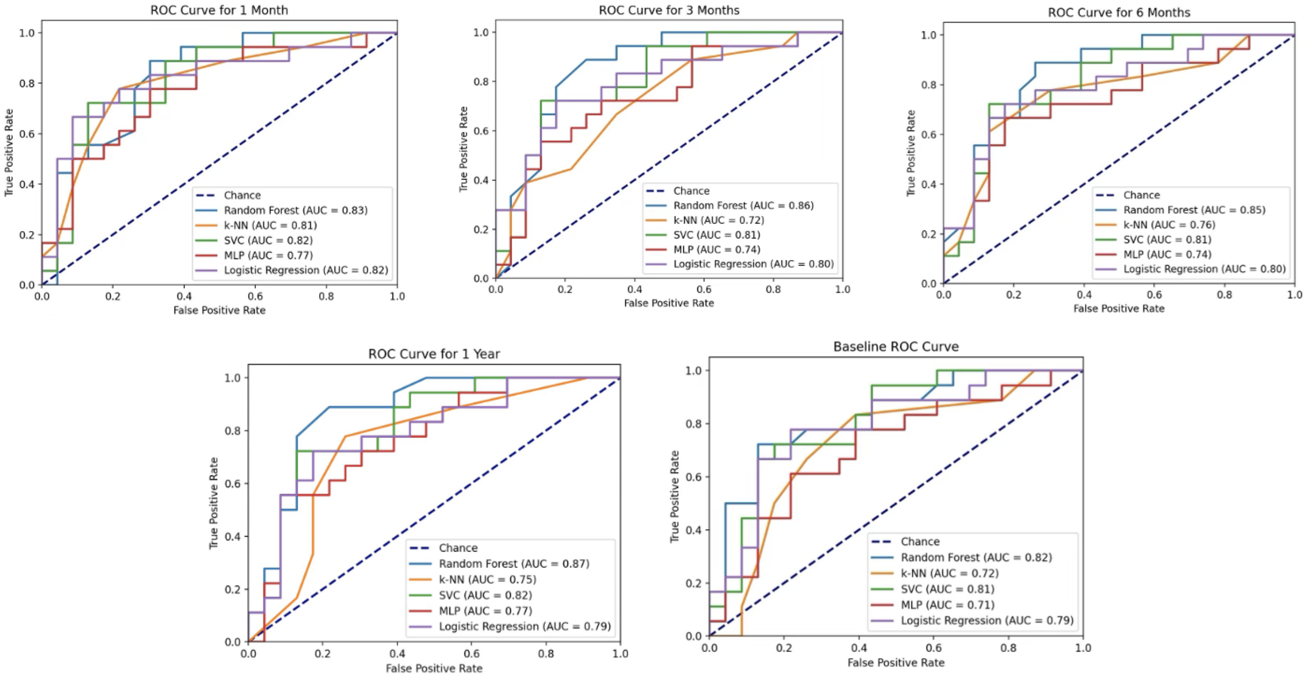


Figure 4

to Dr. Xiao Luo for her guidance. The authors would also like to thank Professor Feng Li and Sheila Walter for their support.

REFERENCES

Received August 2023