ISSN 2706-6819 Vol.6, Issue 6: 45-52, DOI: 10.25236/IJFM.2024.060607

Multivariate Data Analysis-Based Study on Clinical Decision Support for Hemorrhagic Stroke

Wenteng Huo[1], Zunshu Li[2,*], Xueying Liu[1]

[1]*School of Economics and Management, Yanshan University, Qinhuangdao, 066004, China*
[2]*Silesian College of Intelligent Science and Engineering, Yanshan University, Qinhuangdao, 066004, China*
[*]*Corresponding author: leezszero@163.com*

Abstract: *Improvements in the management and treatment of hemorrhagic stroke, one of the leading causes of disability and death worldwide, are of great public health importance. In this study, we investigated the process of hematoma expansion and the development of peripheral edema by analyzing clinical case data to improve accurate prediction of patient prognosis. Hematoma expansion within 48 hours of onset was identified by careful analysis of the first imaging data, and risk analysis and prediction were performed using an integrated tree model, such as the Random Forest algorithm. The model optimization significantly improved the prediction accuracy and provided strong support for clinical decision-making. In exploring the evolution of perihematoma edema, polynomial function fitting K-means clustering and particle swarm algorithms were used to reveal the individual differences in edema progression, which provides a new scientific basis for clinical treatment. This study provides an in-depth analysis of the pivotal clinical features of hemorrhagic stroke and offers new insights into hematoma expansion and the development of peripheral edema through advanced data processing and modeling, which in turn provides a scientific basis for clinical treatment decisions, enhances the understanding of this complex brain disease, and contributes to improved patient outcomes.*

Keywords: *Hematoma Expansion; Edema Development; Random Forest Algorithm; Apriori Algorithm; Model Prediction*

1. Introduction

Stroke is a global health challenge affecting physiological, social, and economic health. Hemorrhagic stroke, characterized by its sudden onset and rapid progression, requires immediate medical intervention to limit brain damage and improve survival chances. Medical imaging technologies, such as CT scans and MRI, are crucial for identifying the location of the hemorrhage, estimating the volume of bleeding, and monitoring hematoma changes. These imaging techniques provide valuable insights for assessing the severity of the condition, guiding clinical treatment decisions, and predicting patient outcomes, thus playing a central role in the management of hemorrhagic stroke. Recent advances in the field of medical imaging for hemorrhagic stroke have highlighted the importance of advanced imaging techniques in improving diagnosis and treatment[1]. Non-contrast-enhanced head CT scans are widely recognized as the most commonly used initial imaging study for hemorrhagic stroke, capable of rapidly identifying intracerebral hemorrhage. With the development of magnetic resonance imaging (MRI) technology, the use of MRI in hemorrhagic stroke is increasing, providing more detailed information on brain damage and bleeding characteristics. As technology continues to advance and be applied, significant progress is expected in the future diagnosis, treatment, and prognosis assessment of hemorrhagic stroke [2].

This study employed the random forest algorithm to address the issues of early detection and prognostic assessment of hemorrhagic stroke. Random forest is a robust machine-learning algorithm and has been widely used in the field of medical image analysis [3]. Random forest is acclaimed for its excellent classification performance and ability to handle high-dimensional data, particularly in medical imaging analysis. In hemorrhagic stroke research, the random forest algorithm is used to analyze complex imaging data to identify characteristics of hemorrhagic stroke and predict the evolution of the condition[4]. For instance, Hakimi and Garg (2016) emphasized the importance of combining advanced imaging techniques with data analysis tools, where the random forest algorithm played a crucial role in analyzing imaging data[5]. Huang et al. (2017) explored the potential of TAI in

ISSN 2706-6819 Vol.6, Issue 6: 45-52, DOI: 10.25236/IJFM.2024.060607

detecting hemorrhagic stroke, noting that combining algorithms like random forest could further enhance detection accuracy[6]. Moreover, the random forest has been applied to evaluate hematoma transformation after stroke, as Zaheer et al. (2000) used MRI data combined with the random forest algorithm to predict hemorrhagic stroke[7]. In this study, besides employing the random forest algorithm, the K-means clustering analysis method was explicitly used further to investigate hemorrhagic stroke's imaging characteristics and pathological patterns. In the research of hemorrhagic stroke, K-means clustering analysis effectively groups patients into different subtypes based on imaging features, aiding in understanding the heterogeneity of hemorrhagic stroke and providing precise evidence for etiological research and therapeutic strategy formulation. Additionally, using the Apriori algorithm to explore the relationship between treatment preferences and survival is analogous to studying the relationship between treatment interventions and PHE[8]. The research indicates that the expansion rate of perihematomal edema predicts outcomes after intracerebral hemorrhage, suggesting the need for targeted therapeutic interventions[9].

In this study, we comprehensively analyzed the data provided by the website (https://cpipc.acge.org.cn/cw/hp/4). We harness the robust capabilities of the Random Forest and Apriori algorithms to enhance the predictive accuracy and clinical decision-making processes for hemorrhagic stroke. By integrating sophisticated machine learning models with extensive clinical imaging data, we provide novel insights into the dynamics of hematoma expansion and the progression of perihematoma edema. This research not only refines the predictive models for patient outcomes but also deepens the understanding of stroke pathology, offering a significant contribution to the fields of medical imaging and neurology. These advancements promise to support more informed and effective therapeutic strategies, potentially improving prognosis and management of this critical condition.

2. Model Preparation and Preprocessing

2.1 Data pre-processing

Data preprocessing was initially performed before building the model in this research. To enhance clarity, the team members organized and analyzed the data based on extensive patient information provided by the website. When examining various data sets, it was noted that patient sub074's first admission imaging examination serial number was shown as 20180719000630 in dataset 1, which differed from the records in other datasets.

Due to the limited sample size, arbitrarily deleting data could negatively impact the research results. To ensure the scientific validity and accuracy of the model, we opted to replace this serial number in the data processing with 20180719000020, as recorded in other datasets. For patient sub131, the first admission imaging examination serial number was recorded as 20160413000006 in datasets 1 and 4 but appeared as 20171220002173 in dataset 2 and Appendix 1. Upon further analysis of dataset 2, it was found that the follow-up record of patient sub131 could be crucial for later data analysis. Therefore, we decided to retain the serial number from dataset 2 and make the necessary corrections in datasets 1 and 4. The data processing strategy for patient sub132 was similar to that for sub131. During this process, we noticed that the first admission imaging examination record for patient sub131 with serial number 20171220002173 was missing in dataset 3. Comparing this with the data in dataset 1, we found that the clinical information matched that of patient sub3. Based on this observation, we hypothesized that these records might belong to the same patient but were recorded in different years. For patient sub132, we adopted the same processing approach as for sub131, assuming they were the same individual as patient sub4. Through this process, we cleaned the data, ensuring the accuracy and reliability of subsequent analyses.

2.2 Determination of hematoma expansion in patients

After data preprocessing, the research team obtained a dataset containing 160 independent patient samples, each with a serial number for the first admission imaging examination. These serial numbers enabled the team to determine the specific timing of each imaging examination. By comparing the interval between the onset of symptoms and the time of the first imaging examination and the intervals between subsequent imaging examinations, the team could assess whether these imaging examinations were completed within 48 hours after the onset of symptoms. This assessment is crucial for an in-depth analysis of the timeliness of imaging examinations related to the onset of the disease. In further studies, the team will use these data to explore the imaging examination conditions at the onset of the disease

ISSN 2706-6819 Vol.6, Issue 6: 45-52, DOI: 10.25236/IJFM.2024.060607

and evaluate their effectiveness in clinical applications.

3. Machine learning model construction for predicting hematoma expansion

The research team selected methods based on machine learning models. To enhance the model's fitting and predictive capabilities, the team chose the ensemble tree model, which is highly interpretable in machine learning, for regression analysis. Specifically, the research team employed the random forest and AdaBoost algorithms for comparative analysis.

3.1 Comparative analysis of models

The random forest algorithm is part of the ensemble tree model family. It reduces the risk of overfitting by constructing multiple decision trees and integrating the results of these trees for predictions. Due to the adoption of multiple decision trees, random forest exhibits high robustness and generalization capability, which is especially effective in reducing the risk of overfitting large datasets. AdaBoost is an ensemble learning model that linearly combines multiple base learners to minimize the loss function. This model corrects its errors by increasing the weights of incorrectly classified data points, iteratively trains a series of weak classifiers, and dynamically adjusts their weights based on accuracy, thus forming a solid classifier[10].

When plotting the ROC curve, the curve is generated through different classification thresholds, with the valid positive rate (TPR) as the y-axis and the false positive rate (FPR) as the x-axis. In comparing the ROC curves of different models, the closer the curve is to the top left corner, the better the model's classification performance. Maintaining consistency in the training and testing sets when comparing models is crucial. The AUC (Area Under Curve) ranges between [0,1], with larger AUC values indicating better model classification performance.

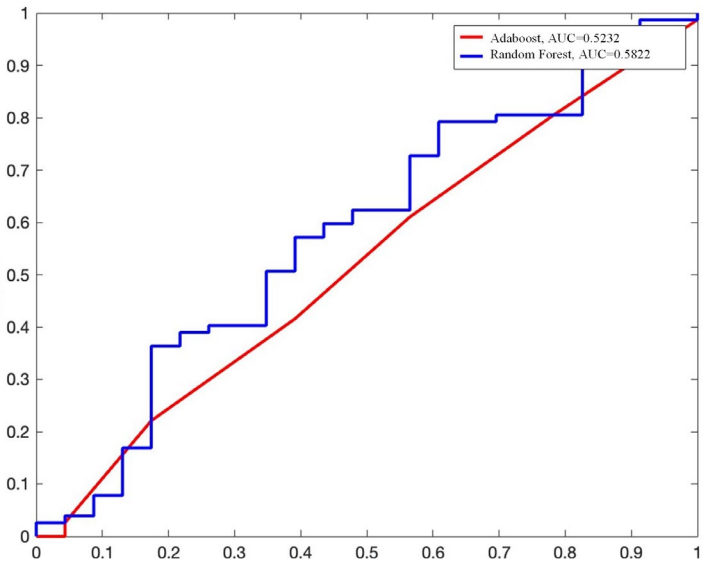

Figure 1: Algorithm comparing AUC values

As shown in figure1, it is observed that the AdaBoost algorithm performs weaker in terms of AUC value compared to the random forest algorithm (Fig.1). This indicates that there might be a higher risk of error when making predictions with the AdaBoost model. Therefore, in the context of the current problem, selecting the random forest model is more appropriate as it demonstrates smaller error margins and higher prediction accuracy. The random forest model is more stable and reliable in processing complex datasets because it can handle large amounts of data and consider multiple input variables. This model significantly enhances the accuracy of predictions, making the research findings more universally applicable and valuable.

3.2 Determining the model - Random Forest Model

Through a comprehensive comparison of the above models, this study employed the random forest algorithm to predict the risk of hematoma expansion in patients. The feature data was derived from

ISSN 2706-6819 Vol.6, Issue 6: 45-52, DOI: 10.25236/IJFM.2024.060607

patients' personal history, medical history, onset-related data, and initial imaging examination results, serving as input variables for the random forest model. During the model training process, each decision tree randomly selected subsets of features, enriching the model's diversity and enhancing its generalization ability on new data. In the model training phase, cross-validation methods were used for parameter optimization, such as the number and depth of decision trees and feature selection strategies. The optimized model was trained on the training set and validated on an independent test set to assess its predictive performance.

4. Analysis of the impact of therapeutic interventions on the progression of perihematomal edema

4.1 Effect of time factors on perihematomal edema progression

In neuroscience, understanding the dynamic changes in perihematomal edema is essential for optimizing the treatment strategies for hemorrhagic stroke. This research focuses on how time factors influence the volume changes of perihematomal edema, aiming to provide accurate timing evidence for clinical intervention. A comprehensive model was developed by conducting a detailed analysis of the data for the first 100 patients recorded in Figure 2 to describe the pattern of edema volume changes over time. The relationship curve between edema volume and time progression was established to delve into their dynamic correlation. Since each patient underwent multiple follow-up examinations during the study, multiple examination time points were recorded. The change in edema volume between these consecutive examination time points was the focus of the analysis, revealing the relationship between time progression and edema volume, with examination time points serving as independent variables and edema volume as the dependent variable, thus creating a series of data points for each patient.

In this study, the team included the time interval from the onset to the first imaging examination as part of the time analysis to ensure the completeness of the time calculations. The timeline for each patient was from the onset to the last imaging examination, ensuring the comprehensiveness of the independent variables. Drawing a scatter plot was necessary to assess the correlation between edema volume and time. Through the analysis presented in Figure 2, the pattern and trend of changes in edema volume over time can be intuitively identified, providing a basis for further statistical analysis and model development.

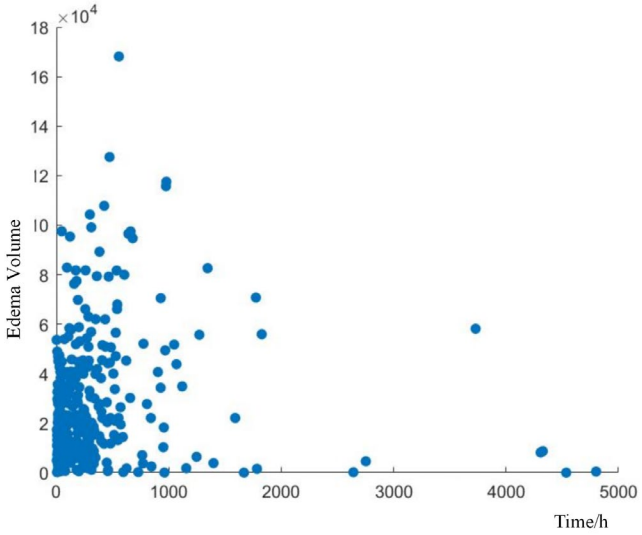

Figure 2: Patient edema volume trends

4.2 K-means clustering-based approach to study the impact of therapeutic interventions and progression of perihematomal edema

K-means clustering is a widely used algorithm in data science and statistics. Its primary goal is to divide a dataset into several categories or groups so that the similarity among data points within the same group is higher than that among data points in different groups.

In clinical research, particularly in studies on the progression of perihematomal edema, K-means

ISSN 2706-6819 Vol.6, Issue 6: 45-52, DOI: 10.25236/IJFM.2024.060607

can reveal patterns in how various therapeutic interventions affect patients' radiological outcomes, thereby assisting physicians in devising more precise treatment plans[11]. In recent years, with the advancement of machine learning technologies, K-means clustering has been applied in various aspects of the healthcare sector, including disease classification, patient stratification, and evaluation of treatment outcomes. In figure 3, the specific results of the algorithm operation are as follows:

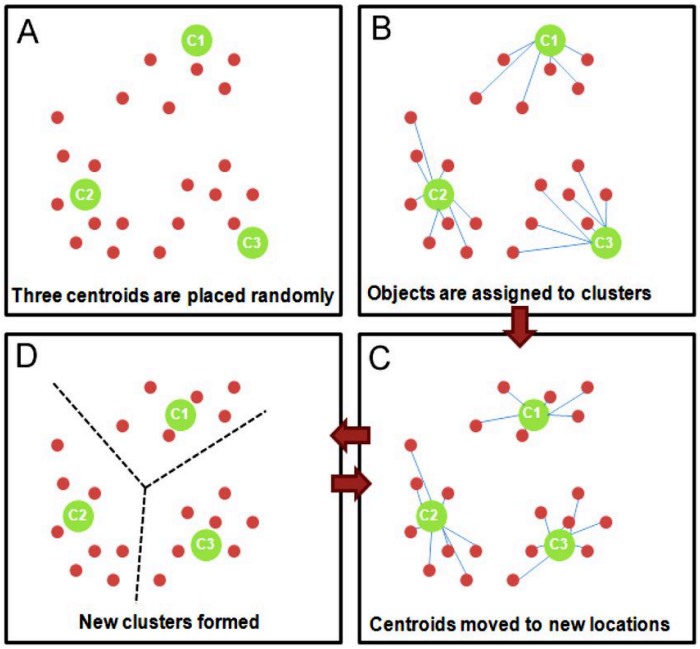

Figure 3: Cluster analysis

K-means clustering, an unsupervised classification method, divides a set of samples into meaningful groups based on specific data mining objectives. Cluster validity is often measured in terms of intra-cluster and inter-cluster metrics. The ideal clustering result should exhibit the smallest possible intra-cluster cohesion and the highest possible inter-cluster separation. Cohesion measures the closeness among samples within a cluster, while separation measures the degree of difference between clusters. Here is the formula for calculating cohesion based on prototypes:

Cohesion: The cohesion of a cluster reflects the similarity or closeness among samples within the cluster. A standard method to calculate cohesion is the average intra-cluster distance. The formula is as follows:

$$Cohesion = \frac{1}{n} \sum_{1 \le i < j \le n} distance\left(x_i, x_j\right) \tag{1}$$

Separation: The separation of clusters measures the difference or degree of separation between samples of different clusters. A standard method for calculating separation is the average inter-cluster distance. The formula is as follows:

$$Separation = \frac{1}{mm'} \sum_{1 \le i < j \le n} distance\left(x_i, x_j\right) \tag{2}$$

Cluster cohesion and separation are not independent; their sum is constant, equal to the total sum of squares, which is the sum of the squared distances of each sample from the overall mean. From this conclusion, it can be inferred that minimizing cohesion is equivalent to maximizing separation. Most currently proposed validity functions are based on combining cohesion and separation and their weighted improvements. Therefore, the silhouette coefficient is an effective solution for addressing this issue.

ISSN 2706-6819 Vol.6, Issue 6: 45-52, DOI: 10.25236/IJFM.2024.060607

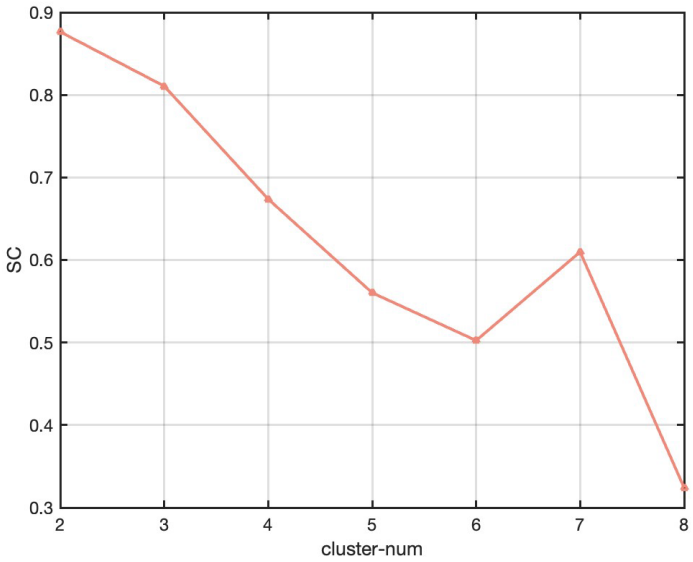

Figure 4: Contour Coefficient

Based on the analysis in Figure 4, this study further investigates the impact of seven treatment methods on the progression pattern of edema volume: "ventricular drainage," "hemostatic treatment," "intracranial pressure reduction treatment," "antihypertensive treatment," "sedation and analgesic treatment," "antiemetic and gastric protection," and "neurotrophic nutrition." Initially, the progression pattern of edema volume was defined in this study. However, considering that some patients may only undergo 1-2 follow-up visits, the follow-up data for these patients were based on the last follow-up during the visualization process. The visualization of the follow-up change rates for different patients is shown in figure 5:

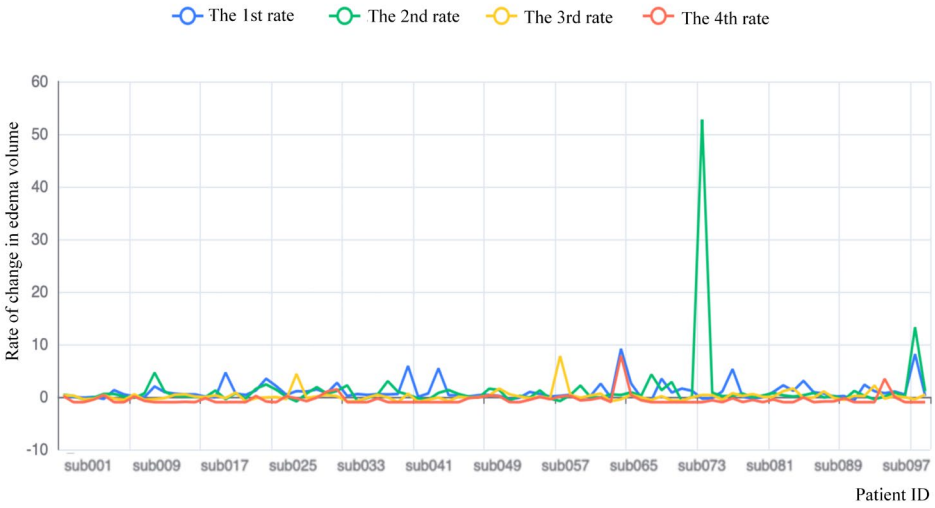

Figure 5: Comparison of hematoma changes at four follow-up visits

From Figure 5, the study observed that the edema volume of most patients significantly increased during the first follow-up. However, with ongoing treatment, the edema volume was effectively controlled. By the second follow-up, the edema volume had stabilized. During the third and fourth follow-ups, the condition of the patients was effectively managed, with the edema volume showing a downward trend compared to earlier. Additionally, as the treatment progressed, the change rate in the follow-ups gradually approached zero, suggesting the effectiveness of the treatment approaches. In the subsequent analysis, the results of the last follow-up for each sample will be taken as definitive.

4.3 Exploring the relationship between therapeutic interventions and perihematomal edema based on the Apriori algorithm

In this chapter, the study further analyzes the relationships among hematoma volume, edema

ISSN 2706-6819 Vol.6, Issue 6: 45-52, DOI: 10.25236/IJFM.2024.060607

volume, and treatment methods. According to the previous analysis, it is known that the condition of patients tends to stabilize with treatment, during which the volumes of hematoma and edema change accordingly. Thus, this study uses the difference in hematoma and edema volumes between the last and first follow-up visits as an indicator. If the difference is negative, it is marked as 0, indicating effective treatment; if positive, it is marked as 1, indicating effective treatment. This marking method clearly shows whether the treatment is effective and allows for comparing the correlation between hematoma volume and edema volume. After encoding all data with 0-1 coding, the Apriori algorithm is introduced to construct the association mining model.

Association rule mining is one of the most essential functions in data mining. The core of data mining research is the extraction of association rules, and mining association rules in the transaction database between various itemsets is an essential field of dataset research[12]. The most significant advantage of mined association rules is their ability to detect unknown relationships and produce results that provide a basis for decision-making and prediction. The discovery process of association rules can be divided into two stages: detecting frequent itemsets and generating association rules[13].

In generating association rules, the support and confidence thresholds must be applied to all association rules as constraints, ensuring that only rules meeting the minimum threshold requirements are generated. Among the association rule data mining algorithms, the Apriori algorithm is relatively simple and easy to execute and is used to mine frequent item sets in the database.

The basic idea of the Apriori algorithm is first to identify all frequent item sets that appear at least as frequently as a predefined minimum support level. Then, strong association rules are generated from these frequent item sets, which must satisfy both minimum support and minimum confidence thresholds. The desired rules are produced using the frequent itemsets found in the first step, generating rules that only contain set items, where each rule's right-hand side contains only one item, following the definition of median rules. Once these rules are generated, only those with a confidence level higher than the user's minimum threshold are retained. A recursive approach is used to create all frequent item sets.

After inputting the influencing factors of this study into the Apriori algorithm, the results lead to the conclusion that "hemostatic treatment," "intracranial pressure reduction treatment," "antihypertensive treatment," "sedation and analgesic treatment," "antiemetic and gastric protection," and "neurotrophic nutrition" have a strong association with edema volume. Furthermore, "hemostatic treatment," "antihypertensive treatment," "sedation and analgesic treatment," "antiemetic and gastric protection," and "neurotrophic nutrition" are strongly associated with hematoma volume. In contrast, the association between hematoma volume and edema volume is not strong.

5. Conclusion

In this study, through meticulous data preprocessing and analysis, we established a machine-learning model to predict the probability of hematoma expansion. The comprehensive data verification and correction during data preprocessing ensured the accuracy and reliability of the model analysis. The comparative analysis between the random forest algorithm and the AdaBoost algorithm revealed that the random forest exhibited superior performance in this study, especially regarding the AUC value. Regarding therapeutic interventions, the association rule mining of the Apriori algorithm revealed the relationships between different treatment methods and the volumes of hematoma and edema. The results indicated a strong correlation between specific treatment methods and the volumes of hematoma and edema, providing valuable insights for clinical treatment and aiding physicians in making more precise and effective therapeutic strategy decisions. This research successfully built an efficient machine-learning model to predict the probability of hematoma expansion. It deepened the understanding of the relationship between therapeutic interventions and the progression of perihematomal edema through data mining techniques. These findings provide a scientific basis for improving the clinical treatment of patients with cerebral hemorrhage and are expected to play a significant role in future clinical practices.

References

[1] SMITH E E, ROSAND J, GREENBERG S M. Imaging of hemorrhagic stroke[J]. Medical Physics, 2006, 14(2): 127–140.
[2] BANDI V, BHATTACHARYYA D, MIDHUNCHAKKRAVARTHY D. Prediction of brain stroke

ISSN 2706-6819 Vol.6, Issue 6: 45-52, DOI: 10.25236/IJFM.2024.060607

severity using machine learning[J]. Revue d'Intelligence Artificielle, 2020, 34(6): 753–761.
[3] MUSCHELLI J, SWEENEY E M, ULLMAN N L, et al. PItcHPERFeCT: primary intracranial hemorrhage probability estimation using random forests on CT [J]. NeuroImage: Clinical, 2017, 14: 379–390.
[4] LIN C-H, HSU K-C, JOHNSON K R, et al. Evaluation of machine learning methods to stroke outcome prediction using a nationwide disease registry[J]. Computer Methods and Programs in Biomedicine, 2020, 190: 105381.
[5] HAKIMI R, GARG A. Imaging of hemorrhagic stroke[J]. CONTINUUM: Lifelong Learning in Neurology, 2016, 22(5): 1424–1450.
[6] HUANG L, LI T, JIANG H. Technical note: thermoacoustic imaging of hemorrhagic stroke: a feasibility study with a human skull [J]. Medical Physics, 2017, 44(4): 1494–1499.
[7] ZAHEER A, OZSUNAR Y, SCHAEFER P W, et al. Magnetic resonance imaging of cerebral hemorrhagic stroke[J]. CONTINUUM: Lifelong Learning in Neurology, 2016, 11(5): 288–299.
[8] BOUTS M J, TIEBOSCH I A, RUDRAPATNA U S, et al. Prediction of hemorrhagic transformation after experimental ischemic stroke using mri-based algorithms[J]. Journal of Cerebral Blood Flow & Metabolism, 2017, 37(8): 3065–3076.
[9] CHOI J-M, SEO S-Y, KIM P-J, et al. Prediction of hemorrhagic transformation after ischemic stroke using machine learning[J]. Journal of Personalized Medicine, 2021, 11(9): 863.
[10] Rong Fang, Gao Z ,Liu. An Improved AdaBoost Algorithm for Hyperparameter Optimization[J]. Journal of Physics: Conference Series, 2020.DOI:10.1088/1742-6596/1631/1/012048.
[11] TAO Y, WANG Y. Improved k means algorithm based on the selection of initial clustering centers [J]. Foreign Electronic Measurement Technology, 2022, 41(9): 54–59.
[12] ZHENG Z, GONG Q, ZHANG J. Improved association rule mining algorithm: mifp-apriori algorithm [J]. Science Technology and Engineering, 2019, 19(16): 216–220.
[13] LIU W, XU Y. Association rule mining of metro failures based on improved apriori algorithm[J]. Journal of Ordnance Equipment Engineering, 2021, 42(12): 210–215.