A Study on Clinical Diagnosis of Hemorrhagic Stroke Based on Random Forest and LSTM Neural Networks

Yuqian Zeng *, Mingkun Li, Peixuan Han and Ningyu Bian

Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China

* Corresponding Author Email: zengyuqian330227@163.com

Abstract. This study is dedicated to the application of artificial intelligence technology to assist the clinical management of hemorrhagic stroke patients. By combining patients' clinical information and imaging features, we successfully predicted the probability of hematoma expansion, which provides an important reference for clinical decision-making. Further, we modeled the progression of perihematomal edema and investigated the effect of therapeutic interventions on edema progression. Through K-means clustering and least squares, we fitted the model effectively, which provided a basis for the development of personalized treatment plans. In addition, we used LSTM neural network to predict patients' prognostic MRS scores, which provided new ideas for optimizing patients' rehabilitation pathways. These findings will help to improve the treatment outcome of hemorrhagic stroke patients and reduce the burden on society and families, which is of great clinical significance.

Keywords: Random forest; K-means clustering; LSTM neural network.

1. Introduction

As an acute cerebrovascular disease, hemorrhagic stroke is characterized by high morbidity, disability, mortality, and recurrence, which imposes a heavy burden on individuals, families, and society [1]. Hypertension is an important causative factor, and the monitoring of hematoma extent and edema expansion is crucial in the treatment and prognosis of hemorrhagic stroke [2]. The application of medical imaging combined with artificial intelligence provides new ideas for clinical diagnosis, and machine learning technology assists in predicting patient prognosis, optimizing treatment plans, and improving patient survival and quality of life.

In this paper, hematoma expansion and edema progression in hemorrhagic stroke patients are discussed. First, the risk of hematoma expansion in patients was effectively predicted by principal component analysis and random forest prediction model, which provided an important basis for clinical decision-making. Second, K-means clustering algorithm and least squares method were applied to explore the relationship between the occurrence of perihematomal edema and therapeutic intervention, which provided a new perspective for personalized treatment. Finally, LSTM neural network was utilized to predict the MRS score, which provided a reliable tool for assessing patient prognosis. The aim of this paper is to integrate imaging features, patient information, and treatment options to accurately predict the prognosis of hemorrhagic stroke patients, and to provide support for improving patient survival and neurological recovery. By combining the cutting-edge technologies of artificial intelligence and medical imaging, we will explore more effective clinical diagnosis and treatment strategies to promote the recovery and quality of life of hemorrhagic stroke patients.

2. Determination of Hematoma Expansion

The time point for hematoma expansion determination was 48 h after the onset of the disease, but in clinical practice, it is not guaranteed to be able to monitor punctually after 48 h. Therefore, the monitoring data within 48 h were commonly used for judgment. The maximum value of the hematoma volume monitored within 48 h after the onset of disease in the first 100 patients, except for the first monitoring, was selected and compared with the hematoma volume monitored for the first time, and those patients who met the requirement of an absolute volume increase of ≥6 mL or a

relative volume increase of ≥33% were considered to have undergone hematoma dilatation, and this monitoring time point was the time point of hematoma dilatation.

Next, a hematoma expansion prediction model was constructed by combining the personal situation, disease history, and characteristics related to the onset and treatment of the first 100 patients, as well as the first imaging findings and hematoma shape characteristics and gray-scale characteristics of the first 100 patients, to obtain the probability of hematoma expansion within 48 h of the onset of the disease in all patients after treatment. Because of the high dimensionality of the data to be considered, the data were first subjected to principal component analysis to obtain the dimensionality-reduced data. Then the dimensionality-reduced data were inputted into the random forest regression prediction model for solving, and finally the probability of hematoma expansion within 48 h of onset after treatment was obtained for all patients.

2.1. Principal Component Analysis

The hematoma expansion was first determined by the following formula:

$$r_i = \begin{cases} 1, b_{ij} - b_{i1} \geq 6000 or \left(b_{ij} - b_{i1}\right)/b_{i1} \geq 33\% \\ 0, b_{ij} - b_{i1} < 6000 or \left(b_{ij} - b_{i1}\right)/b_{i1} < 33\% \end{cases} \tag{1}$$

Due to the high dimensionality of the data, principal component analysis was used to downscale the data. Principal component analysis is essentially a dimensionality reduction process of the data, where the original variables with correlation are transformed and recombined into a set of mutually uncorrelated variables through orthogonal transformations [3]. Mathematically, it is a linear combination of the original variables, which realizes the dimensionality reduction of the data while generating new variables. The specific calculation process is as follows:

(1) Original data standardization

To improve the effect of principal component analysis, first, the standardization of the original data, so that it is converted into a dimensionless and similar data size indicators. Firstly, m samples $P_i = (p_{i1}, p_{i2}, p_{i3}, \dots, p_{in})^T$ of the given n-dimensional random vector$P = (p_1, p_2, p_3, \dots, p_n)^T, i = 1, 2, 3, \dots, m, m > n$ are constructed to construct the sample matrix, and the data in the matrix are standardized:

$$q_{ij} = \frac{p_{ij} - \overline{p_j}}{s_j}, i = 1,2,3, \dots, m, j = 1,2,3, \dots, n \tag{2}$$

(2) Solution of the correlation coefficient matrix R of the standardized matrix

The correlation coefficient matrix R of the normalized data:

$$R = \left[r_{ij}\right]_{n \times n} \tag{3}$$

(3) Calculate the eigenvalues and eigenvectors of the correlation coefficient matrix$R$.

Calculate the nonzero eigenroots $(\lambda_1, \lambda_2, \lambda_3, \dots, \lambda_n)$ and eigenvectors of the matrix R using the characteristic equation:

$$a_i = (a_{i1}, a_{i2}, \dots, a_{in}), (i = 1,2, \dots, n) \tag{4}$$

(4) Factor analysis fitness test

The KMO test is usually used to compare the correlation between variables and the expression is:

$$KMO = \frac{\sum\sum_{i \neq j} {r_{ij}}^2}{\sum\sum_{i \neq j} {r_{ij}}^2 + \sum\sum_{i \neq j} {s_{ij}}^2} \tag{5}$$

(5) Selection of principal components

After principal component analysis, we get n principal components, but there is a decreasing effect of principal component information utilization, in the actual analysis process will be based on the size of the cumulative contribution rate of the principal components to select the first k principal components. The contribution rate size c_i is:

$$c_i = \frac{\lambda_i}{\sum_{i=1}^{n} \lambda_i}, (i = 1,2, \dots, n) \quad (6)$$

The contributions of the principal components are ranked from largest to smallest, and the eigenvalues and their contributions are shown in Table 1 and Table 2.

Table 1. Patient hematoma and edema volume and location data eigenvalues and their contribution rates.

Eigenvalue	Contribution(%)	Cumulative contribution(%)
5.5950	25.4318	25.4318
3.2915	14.9614	40.3932
2.6497	12.0443	52.4374
1.9033	8.6511	61.0886
1.6032	7.2874	68.3760
1.4621	6.6459	75.0219
1.0026	4.5574	79.5793
0.9095	4.1342	83.7135

Table 2. Eigenvalues and their contribution to the hematoma shape and gray scale distribution data of patients.

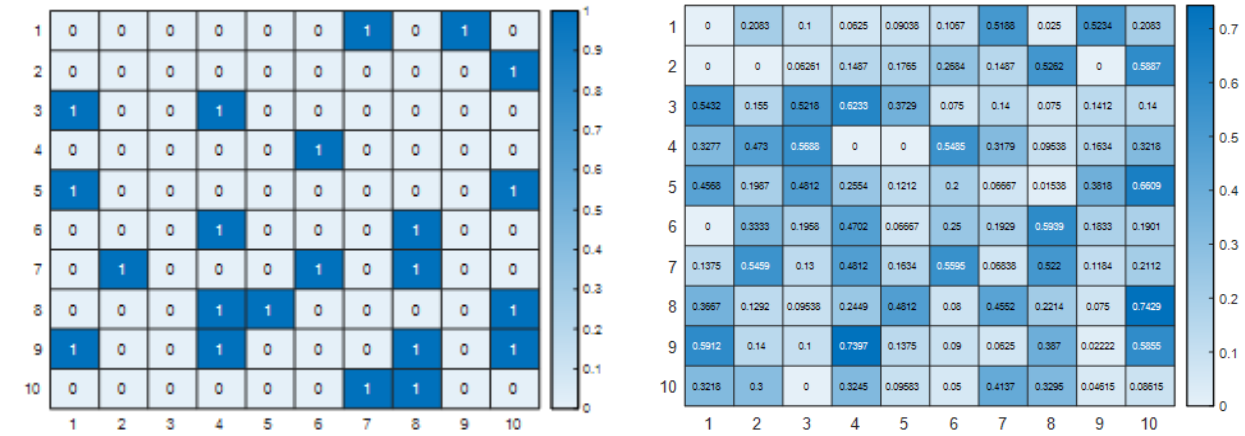
Eigenvalue	Contribution(%)	Cumulative contribution(%)
10.4755	33.7918	33.7918
8.2503	26.6139	60.4058
4.5422	14.6524	75.0581
2.5059	8.0836	83.1417
1.5741	5.0777	88.2193

Through Table 1 and Table 2, we can see that the hematoma and edema volume and location data of the patient's first detection were downgraded from 22 dimensions to 8 dimensions, and the hematoma shape and gray scale distribution data of the patient's first detection were downgraded from 31 dimensions to 5 dimensions.

2.2. Random Forest Prediction Model

Random Forest (RF) is a typical, highly flexible parallel integrated learning algorithm with self-contained feature filtering to assess the importance of each feature variable in the model solution process [4]. Random Forest is mainly divided into an autonomous sampling phase and a strategy combination phase. In the first stage, it is assumed that there are m samples in the original training set, and the method of put-back sampling is applied to carry out m times of extraction to obtain a new subset of samples, and the rest of the unsampled samples form the out-of-bag data, and a total of n groups of this kind of put-back sampling are carried out, which in turn form a subset of n new samples, and based on which n classification trees are constructed, and this method of extraction guarantees the independence between the classification trees. In the second stage, each decision tree makes decisions of equal importance, and the mean value is used as the result of the model regression prediction.

In this paper, the 33-dimensional correlation data of the first 100 patients after dimensionality reduction by principal component analysis is input into the random forest regression prediction model as a training set, and then the remaining 60 patients are used as a test set to verify the feasibility of the model. The results are shown in Fig. 1.



(a) Thermogram of actual hematoma expansion. (b) Thermogram of actual hematoma dilatation.

Fig 1. Comparison of predicted and real hematoma expansion.

As can be seen in Fig. 1, with a probability of 0.5 as the basis for whether hematoma expansion occurs, 90% of the first 100 patients were able to obtain a correct prediction, which is reliable at a confidence level of 90%, and adequately demonstrates the validity of the prediction model.

3. Edema and Hematoma Correlation

Through observation of the samples, we found that some of the samples were highly variable in terms of age, medical history, time interval between onset of disease and the first imaging examination, blood pressure, and volume of edema at the time of the initial examination, etc., and that the model obtained by fitting these data together would inevitably lose the characteristics of many of the data points. In this paper, K-means clustering is used to classify the samples into four categories based on the above-mentioned categories of data with large differences, and ordinary least squares is used to fit each category of samples separately, and the fitting results are compared with the real data in order to improve the fitting performance of the data.

3.1. K-means algorithm ideas

K-means algorithm has good stability, good clustering effect, and there is no relationship with the order of data input, so it can get better results when processing in disorder [5]. When clustering, first randomly select k samples in the data set, as the initial center point of clustering, and then calculate the Euclidean distance between the rest of the sample points and the k sample points, compare the sample points with the k distance values of the k center points, and the closest distance from which center point is categorized into which class, and after that, re-calculate the center point of the cluster, and have been repeating the previous steps until the position of the center point of the cluster is converged.

3.2. K-means algorithm steps

Selection of points: randomly select k non-repeating samples from the samples as the original cluster centers.

Categorization: calculate the Euclidean distance between the remaining samples and the k cluster centers, and compare the size of each distance value, and categorize the samples to the centers with the closest distance to them, i.e., assign the samples to the most similar clusters.

Computation: recalculate the mean value of the sample points in each cluster and use the mean value as the new k cluster centers.

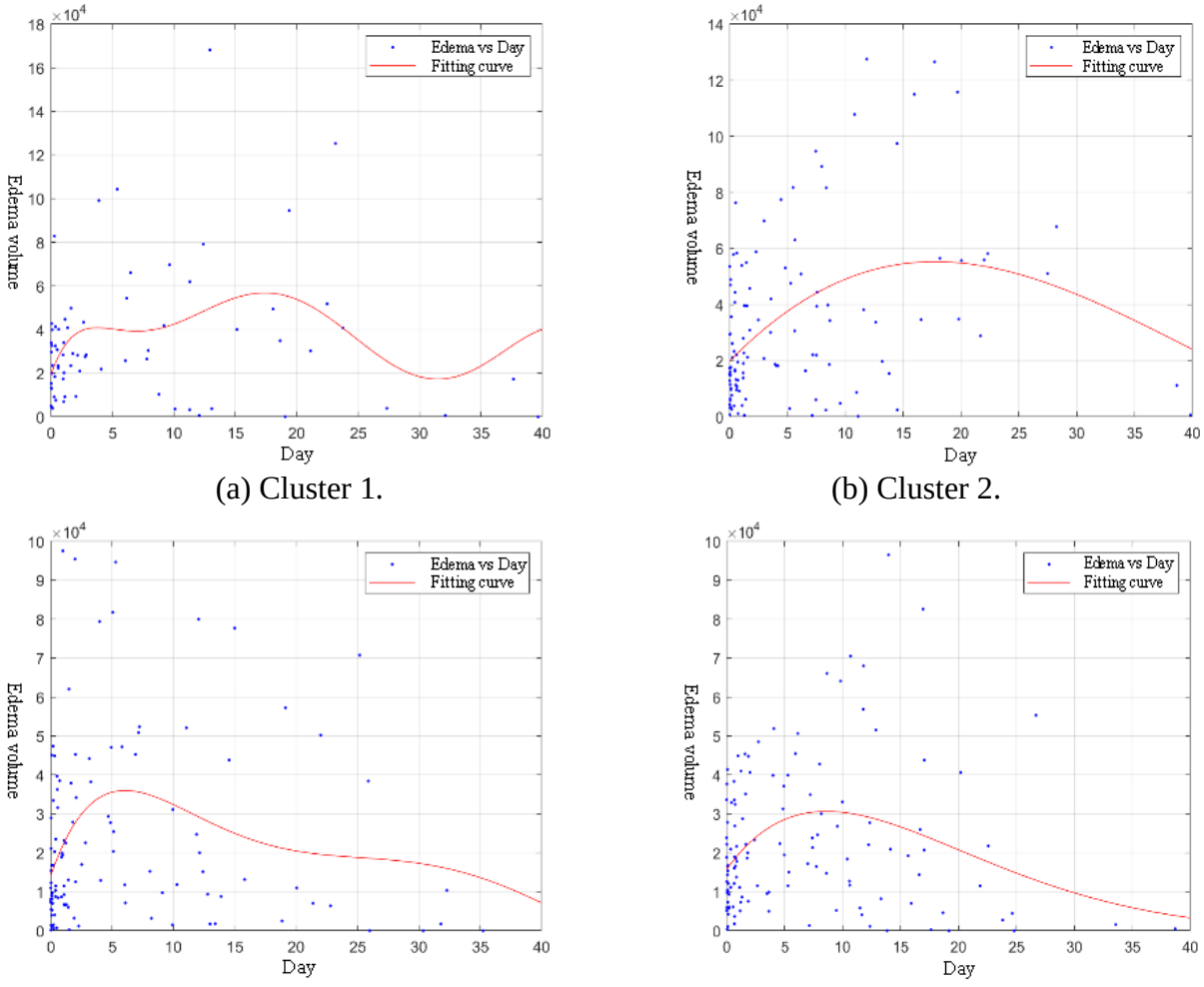
End: keep repeating the previous steps until the end when the positions of the cluster centroids converge (the new cluster centers are less than or equal to the original cluster centers).

K-means clustering was performed on the information related to the first 100 patients to obtain four cluster centers, as shown in Table 3.

Table 3. K-means clustering centers.

	1	2	3	4
Age	61.379	58.169	79.032	64.143
Sex	0.72414	0.72308	0.54839	0.71429
Pre-onset mRS score	0	0.12308	0.096774	0.31429
History of hypertension	1	0.92308	0.80645	0.77143
History of stroke	0.10345	0.21538	0.25806	0.31429
History of diabetes mellitus	0.24138	0.10769	0.19355	0.11429
history of atrial fibrillation	0.034483	0.030769	0.064516	0.057143
history of coronary artery disease	0.10345	0.030769	0.22581	0.057143
history of smoking	0.31034	0.24615	0.16129	0.057143
history of alcohol consumption	0.17241	0.18462	0.096774	0.028571
Time between onset and first imaging	2.831	3.1015	3.7742	5.1977
High blood pressure	208.97	171.78	164.87	137.37
Low blood pressure	113.9	101.68	77.419	81.657

After the clustering was completed, ordinary least squares were fitted to each of the four classes of data, and the results were obtained as shown in Fig. 2 and Fig. 3



(a) Cluster 1.

(b) Cluster 2.

(c) Cluster 3.

(d) Cluster 4.

Fig 2. Fitting results for each cluster group.

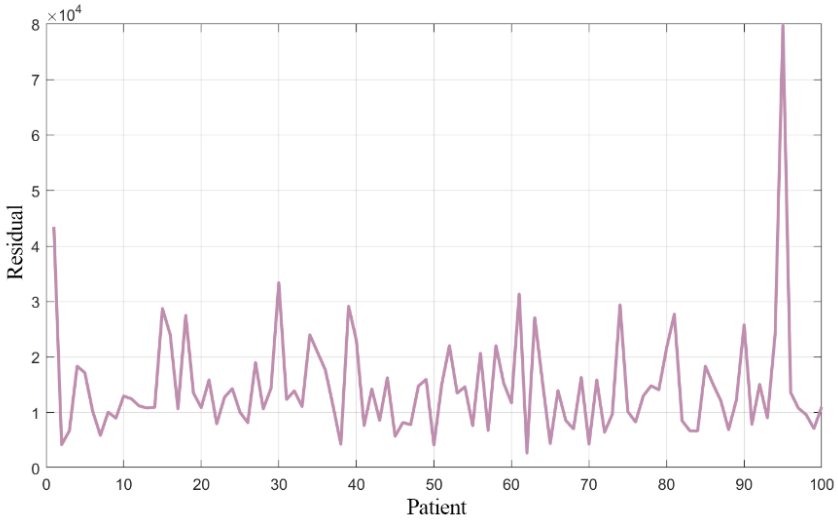


Fig 3. Plot of residuals by patient.

As can be seen from Fig. 2, the general trend of the clustering groups is the same, and the edema volume all shows the trend of increasing and then decreasing, cluster group 1 has a rising trend in the late stage, which is mainly due to the fact that some of the patients' edema volume is still higher in the late stage; cluster group 2 is relatively later in edema volume inflection point compared with clusters 3 and 4. As can be seen from Fig. 3, the residuals of each patient are mainly distributed in the range of 10,000-30,000, and the fitting performance is good.

4. Prediction of mRS Scores

In this section, LSTM neural network is used to predict mRS scores. RNN is a kind of chain-structured neural network commonly used in sequence data processing, and the phenomena of gradient vanishing and gradient explosion will occur when dealing with long sequence training, while LSTM, as a special type of recurrent neural network, is able to solve the time-dependence problem of RNN which needs to carry out long term learning, and to avoid the gradient vanishing and gradient explosion problems [6]. LSTM adds or deletes information about cell states by introducing memory structure under each time step combined with the output of the previous node through the gate structure. Gradient explosion problems. LSTM adds or removes information about the cell state by introducing a memory structure that combines the output of the previous node at each time step with a gate structure.

Memory gates:

$$\Gamma_f^{<t>} = \sigma\left(W_f \cdot [h^{a-1>}, x^{a>}] + b_f\right) \tag{7}$$

Input gates:

$$\Gamma_i^{<t>} = \sigma(W_i \cdot [h^{a-1>}, x^{a>}] + bi) \tag{8}$$

Output gates:

$$\Gamma_o^{<t>} = \sigma(W_o \cdot [h^{<t-1>}, x^{<t>}] + b_o) \tag{9}$$

Memory cell structure:

$$c^{<t>} = \Gamma_f^{<t>} \cdot c^{<t-1>} + \Gamma_i^{<t>} \cdot x^{<t>} \tag{10}$$

$$h^{<t>} = \Gamma_o^{<t>} \cdot {}_{tanh}(c^{<t>}) \tag{11}$$

$$y^t = \sigma(Wh^{c>}) \tag{12}$$

Where: W_f, W_r, W_o and b_f, b_i, b_o are the network structural parameters through the LSTM layer, i.e., the parameters that are adjusted through the feedback of the loss function during training, is the forgetting gating $\Gamma_f^{<t>}$, which controls the specific forgetting part of $c^{<t-1>}$ of the previous state, $\Gamma_i^{<t>}$ is the input gating, which controls the memorization of the input $x^{<t>}$, $\Gamma_o^{<t>}$ is the output gating, which controls the output of the current state. The $c^{<t>}$ result is transformed into a value between -1 and 1 by multiplying the splicing vector by the above structural parameter matrix and then by the sigmoid excitation function to a value between 0 and 1, and by the nonlinear function $tanh$ to a value between -1 and 1.

The personal history, disease history, morbidity-related and first imaging result data of the first 100 patients (sub001 to sub100) were used as inputs to the LSTM neural network, and the known mRS scores of the first 100 patients were used as outputs to build the prediction model. The first 100 patients (sub001 to sub100) were the training set, and the last 60 patients (sub101 to sub160) were the test set. The input data are 104 indicators, and the output is 0-6 classification.

For the determination of whether the network training converges or not is mainly judged by the convergence curve of the selected loss function, in the case of a large magnitude of input data, a longer training is required to achieve convergence, which is not in line with the needs of practical engineering applications. Therefore, after the load data is imported, the data is normalized, and the original data is transformed into data falling between [0, 1] by min-max normalization, and the data magnitude is normalized, which can achieve faster convergence of the loss function, and the formula used is:

$$x^* = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{13}$$

Where: x is the value after normalization; x_{max} is the maximum value in the data, $x_{min}$is the minimum value in the data.

Based on the LSTM neural network model to predict the training set, the LSTM neural network parameters are set, and the settings are shown in Table 4. In order to reduce the running time and improve the computational accuracy, the model uses three hidden layers through several attempts. The rest of the parameters in the model are the best values obtained through several attempts.

Table 4. LSTM neural network parameter settings.

Parameter name	Parameter values
Number of neurons in input layer	104
Number of neurons in the first hidden layer	450
Number of neurons in the second hidden layer	150
Number of neurons in the third hidden layer	20
Number of neurons in output layer	1
Training method	adam
Maximum number of rounds	300
Initial learning rate	0.001
Cycle interval to reduce learning rate	100
Learning rate reduction factor	0.2
Hardware resources	Single GPU

The model achieved the best prediction effect through the debugging of hyperparameters, and the progress of LSTM neural network training as well as the training effect are shown in Fig. 4, during the training process, the small batch RMSE and function loss gradually decreased, indicating that the model parameters were optimized with the deepening of the training. A training set prediction accuracy of 100% was obtained, which in turn predicted the 90-day mRS scores of patients (sub001 to sub160).

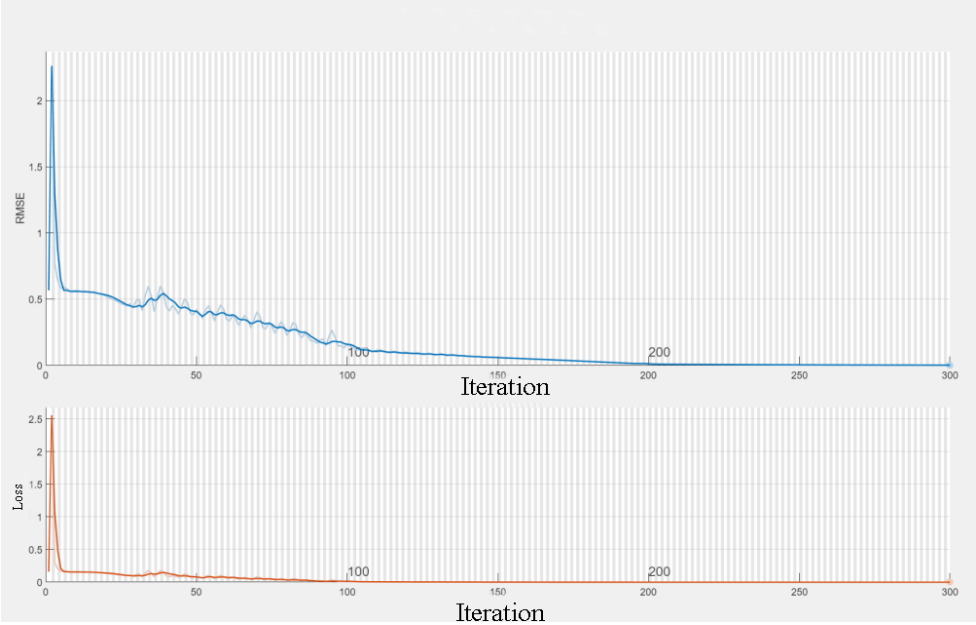


Fig 4. Speed of LSTM neural network training.

5. Summary

In this study, we successfully predicted and modeled hematoma expansion, edema progression, and prognosis of hemorrhagic stroke patients effectively based on machine learning and deep learning techniques, combined with clinical data and medical imaging features. Through principal component analysis and random forest prediction model, we realized the probabilistic prediction of hematoma expansion events, which provides an important reference for clinical decision-making. In the modeling of perihematoma edema, the application of K-means clustering algorithm and least squares method demonstrated the good fitting effect of the model. In addition, by predicting the MRS score with LSTM neural network, we constructed a comprehensive prediction model, which provides a new idea for patient prognosis assessment. These results provide strong support for the individualized treatment and management of hemorrhagic stroke patients and are expected to achieve significant application and promotion in clinical practice. By integrating medical imaging and artificial intelligence technologies, we can more accurately assess the condition and prognosis of patients, contributing to the improvement of patients' quality of life and the reduction of treatment risk.

References

[1] El Hajj M, Abdo R, Assaf S, et al. Stroke Management in Developing Countries[M] Handbook of Medical and Health Sciences in Developing Countries: Education, Practice, and Research. Cham: Springer International Publishing, 2023: 1-31.

[2] Gil-Garcia C A, Flores-Alvarez E, Cebrian-Garcia R, et al. Essential topics about the imaging diagnosis and treatment of hemorrhagic stroke: a comprehensive review of the 2022 AHA guidelines [J]. Current Problems in Cardiology, 2022, 47(11): 101328.

[3] Hasan B M S, Abdulazeez A M. A review of principal component analysis algorithm for dimensionality reduction [J]. Journal of Soft Computing and Data Mining, 2021, 2(1): 20-30.

[4] Elavarasan D, Vincent D R, Sharma V, et al. Forecasting yield by integrating agrarian factors and machine learning models: A survey [J]. Computers and electronics in agriculture, 2018, 155: 257-282.

[5] Mavroeidis D, Marchiori E. Feature selection for k-means clustering stability: theoretical analysis and an algorithm [J]. Data Mining and Knowledge Discovery, 2014, 28: 918-960.

[6] Liu Y, Zhang Q, Song L, et al. Attention-based recurrent neural networks for accurate short-term and long-term dissolved oxygen prediction [J]. Computers and electronics in agriculture, 2019, 165: 104964.