Journal of Computer Technology and Software

ISSN:2998-2383

Vol. 4, No. 1, 2025

Disease Prediction Using Electronic Health Records with Long Short-Term Memory Networks

Elara Vandermeer[1], Kairav Mahadevan[2]

University of Alberta, Edmonton, Canada[1], University of Alberta, Edmonton, Canada[2]

ElaraVandermeer@ualberta.ca[1], Kairav.Mahadevanpp@gmail[2]

Abstract: With the widespread application of electronic health records (EHR), data-driven disease prediction has become an important research direction in the medical field. This study proposes a disease prediction model based on long short-term memory network (LSTM) to analyze time series data in electronic health records and predict patients' future disease risks. As a deep learning model with long-term dependency modeling capabilities, LSTM can effectively process complex time series features in electronic health data. We used the public MIMIC-III database, which contains a large number of patients' diagnosis, treatment and physiological data, and built a disease prediction system through data preprocessing, feature selection and model training. Experimental results show that LSTM shows superior performance in evaluation indicators such as mean square error (MSE), root mean square error (RMSE) and mean absolute error (MAE) compared with traditional machine learning models such as support vector machine (SVM), random forest (RF) and multi-layer perceptron (MLP). By further optimizing the LSTM model, the accuracy of disease prediction can be improved, providing clinicians with a scientific and reliable auxiliary decision-making tool.

Keywords: Long short-term memory networks, Electronic health records, Disease prediction, Deep learning

1. Introduction

With the continuous advancement of informatization in the medical field, electronic health records (EHR) have become an indispensable part of the modern medical system. EHR contains a large amount of patient health data, including diagnostic records, treatment history, laboratory test results, imaging data, drug prescriptions, etc., which provide rich resources for disease prediction, personalized treatment and health management. With the rapid increase in the amount of data, traditional statistical methods and machine learning techniques often face problems such as difficulty in extracting data features and insufficient modeling of long-term dependencies when processing these complex and time-series health data. Therefore, how to effectively use the time series data in electronic health records for disease prediction has become a key research topic[1].

As a special recurrent neural network (RNN), the long short-term memory network (LSTM) has been widely used in various fields due to its unique advantages in time series data processing, especially in natural language processing and time series prediction. It has achieved remarkable results. LSTM can effectively capture long-term dependencies in data through its gating mechanism (such as input gate, forget gate and output gate), solving the problem of gradient vanishing or exploding in traditional RNN in long sequence data. Therefore, when processing time series data in electronic health records, LSTM can better learn the dynamic changes of patients' health status and make early predictions of diseases based on historical health data[2].

The data in electronic health records usually include patients' historical diagnosis, treatment process, medication records, examination results, etc., which are usually highly time-series[3-4]. LSTM can effectively encode these time series data through its memory units, capture long-term dependencies, and help the model understand the patient's medical history and health change trends. Compared with traditional machine learning methods, LSTM can automatically learn potential features in data through end-to-end training without too much manual feature engineering. Especially when facing complex health data, LSTM can gradually extract useful information through its deep network structure, so that the model can have strong adaptability and accuracy when predicting diseases.

In addition, LSTM can process multi-dimensional and multi-type input data to adapt to the diverse health data of different patients. In electronic health records, the medical records of different patients vary greatly and may involve different diseases, treatment methods, and medication regimens[5]. The LSTM network can learn various complex patterns in patients' health data through training and identify potential disease risk factors. In the process of disease prediction, LSTM can predict the patient's future health trends by learning the time dependency of historical data, thus providing a scientific basis for disease prevention and personalized treatment. For example, the LSTM model can predict the probability of a certain disease in the future based on the patient's previous examination data, take intervention measures in advance, and reduce the patient's health risks[6].

Another advantage of the long short-term memory network is that it can effectively process medical data with a long span. In traditional medical research, patients' health data often require long-term tracking and observation, and LSTM is designed to address this problem. Through its meticulous memory mechanism, it can process data with a long span. In the application of electronic health records, the patient's health changes are a long-term accumulation process, and the occurrence of diseases is often not sudden, but the result of the gradual accumulation of multiple potential factors. LSTM can capture these potential cumulative effects through historical data and provide important prediction basis for long-term health management.

Combined with other advantages of deep learning, LSTM can also be trained with large-scale data sets to obtain higher accuracy and generalization ability. With the increase in the amount of medical data, traditional machine learning models may face the risk of overfitting, while deep learning models, especially LSTM, can learn more abstract and effective features in large amounts of data through multi-layer network learning. Compared with rule-based models or traditional statistical methods, LSTM models can handle complex nonlinear relationships, helping doctors better understand and predict patients' health status[7]. By analyzing a large amount of patient data, LSTM can not only improve the accuracy of disease prediction for a single patient, but also extract more universal health trends from group data, providing data support for public health policies and large-scale disease prevention.

In general, long short-term memory networks provide a powerful tool for disease prediction in electronic health records. LSTM can make full use of the characteristics of time series data, capture long-term dependencies, and automatically extract features from massive data through deep learning methods, greatly improving the accuracy and adaptability of disease prediction. In the future of medical and health management, LSTM has broad application prospects. It can not only be used for early diagnosis of diseases, but also help realize personalized medicine and precision treatment, and contribute to the development of global medical and health.

2. Related Work

Deep learning has significantly improved disease prediction using electronic health records (EHR) by capturing long-term dependencies in time-series data. Traditional machine learning methods, such as support vector machines (SVM), random forests (RF), and multi-layer perceptrons (MLP), struggle with the complexity and sequential nature of medical records. Long short-term memory (LSTM) networks, with their gating mechanisms, have been widely adopted to model sequential dependencies effectively. Recent studies have explored enhanced deep learning architectures for EHR-based disease prediction. Gao et al. [8] introduced a multi-channel hypergraph-enhanced model for sequential visit prediction, demonstrating the advantages of structured deep learning in modeling patient trajectories. Similarly, Mei et al. [9] proposed collaborative hypergraph networks to assess disease risk, leveraging graph-based neural architectures to capture complex dependencies. The integration of retrieval-augmented generation (RAG)-based recommendation systems for analyzing medical test data further enhances the interpretability of predictive models [10].

Medical text processing plays a critical role in extracting meaningful features from EHR. Named entity recognition (NER) and deep learning-driven medical text classification have been extensively studied. Fei et al. [11] explored privacy-preserving mechanisms in NLP for medical records, highlighting security concerns in healthcare data processing. Cang et al. [12] investigated deep learning approaches for medical text analysis, improving information retrieval from unstructured clinical notes. Additionally, an ALBERT-driven ensemble learning method was proposed for medical text classification, showing enhanced performance in feature extraction [13]. Hu et al. [14] introduced specialized NLP models for medical named entity recognition, achieving high precision in identifying disease-related terms, while Zheng et al. [15] conducted a comparative study of advanced pre-trained NER models, refining structured feature extraction techniques for healthcare applications. Furthermore, Liang et al. [16] developed deep learning methods for sensitive information detection in medical documents, addressing privacy and security issues in EHR analysis.

Advancements in neural network architectures have further improved sequential modeling and disease prediction accuracy. Yan et al. [17] explored the synergistic role of deep learning and neural architecture search (NAS) in optimizing AI models, paving the way for more efficient LSTM-based architectures. Qi et al. [18] focused on optimizing multi-task learning to enhance large-scale deep learning performance, an approach that can improve the generalization ability of LSTM networks in healthcare. Transformer-based models have also seen optimizations for sequential data processing; Gao et al. [19] introduced a multi-level attention mechanism within an optimized Transformer for text classification, suggesting potential improvements for processing sequential EHR data.

In addition to textual data analysis, deep learning has made substantial progress in feature extraction from medical images, which could complement structured EHR-based disease prediction. Zheng et al. [20] explored fully convolutional neural networks (FCNNs) for high-precision medical image analysis, contributing to advanced feature extraction techniques. Sui et al. [21] proposed a U-Net-based channel squeeze structure for lung nodule detection and segmentation, which demonstrates the efficacy of deep networks in medical diagnostics. He et al. [22] evaluated the performance of VGG19 in complex image classification tasks, providing insights into convolutional architectures applicable to healthcare data. Additionally, Hu et al. [23] investigated few-shot learning with adaptive weight masking in conditional generative adversarial networks (GANs), presenting potential solutions for handling limited labeled medical data.

The literature demonstrates that deep learning, particularly LSTM-based models, has significantly improved disease prediction using EHR. Prior research has contributed to sequential modeling, NLP-based medical text analysis, neural architecture optimization, and feature extraction methodologies. Our study builds on these advancements by further optimizing LSTM networks to enhance disease risk assessment, aiming to improve clinical decision support systems through deep learning innovations.

3. Method

In this study, we used a long short-term memory network (LSTM) to predict diseases in electronic health records. LSTM is a special recurrent neural network (RNN) whose structural design enables it to effectively capture long-term dependencies in time series data. Compared with traditional RNN, LSTM can solve the gradient vanishing and gradient exploding problems faced by traditional RNN in long sequence data by introducing forget gates, input gates, and output gates. The core idea of the model is to store important information through memory cells (cell states), and to decide which information to keep and which information to discard through a gating mechanism, thereby effectively capturing the temporal characteristics in the data. Its network architecture is shown in Figure 1.

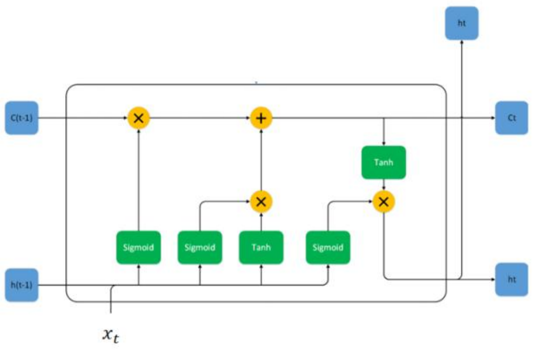

Figure 1. Network architecture diagram

The basic structure of LSTM includes three main gating mechanisms: input gate, forget gate and output gate. At each time step, the input gate determines how much of the current input information is passed to the cell state; the forget gate determines how much of the memory of the previous moment is forgotten; and the output gate determines how much information in the current cell state is output for calculation in the next time step. Specifically, the calculation process of LSTM can be expressed by the following formulas:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

Among them, h_{t-1} is the hidden state of the previous time step, x_t is the input of the current time step, W_i and b_i are the weight and bias of the input gate respectively, and σ is the Sigmoid activation function.

Next, the forget gate (f) controls how much of the memory from the previous moment is forgotten:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

Among them, W_f and b_f are the weight and bias of the forget gate respectively, and σ is also the Sigmoid function.

Then, update the cell state (c):

$$c'_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot c'_t$$

Among them, c'_t is the candidate memory unit, c_t is the cell state at the current time step, and $\tanh$ is the hyperbolic tangent activation function.

In practical applications, LSTM can effectively capture long-term dependencies in time series data through this gating mechanism. For the electronic health record dataset, we regard various medical information of patients (such as diagnosis records, medication records, laboratory test results, etc.) as input features and train and predict them through the LSTM model. In the data processing process, the original data is first preprocessed, including normalization, missing value filling, and time series partitioning. In order to improve the generalization ability of the model, we use the dropout technology to prevent overfitting.

During the model training process, we use the Adam optimization algorithm to optimize the parameters of LSTM. The Adam algorithm combines the momentum method and adaptive learning rate to make the training process more stable and converge faster. The goal of optimization is to minimize the loss function so that the model can achieve better prediction results on electronic health record data.

4. Experiment

4.1. Datasets

In this study, the electronic health record dataset used comes from the public MIMIC-III (Medical Information Mart for Intensive Care) database. MIMIC-III is a large clinical database provided by Bethesda Hospital in Boston, USA, which contains detailed medical records of more than 40,000 patients in the intensive care unit (ICU). The dataset includes various types of medical information, such as patient demographic characteristics, laboratory test results, drug use records, diagnostic codes, and treatment records in the ICU. All patient information is anonymized to ensure data privacy and security, suitable for various clinical studies and machine learning modeling.

The data in the MIMIC-III database is organized in chronological order, providing detailed time series information for each patient. This makes it particularly suitable for training models such as long short-term memory networks (LSTM) for disease prediction and health trend analysis. For example, the database contains physiological data, laboratory test results, drug treatment records, etc. of patients during hospitalization, which can help the model capture the dynamic changes of patients' conditions and thus more accurately predict the occurrence of diseases. The dataset also contains the final results such as the mortality rate and metastasis of the patients, which provides label information for the disease prediction model and can help the model learn the correct prediction target during the training process.

In order to predict the disease, this study extracted disease-related time series data from the MIMIC-III database, including the basic information of the patients, clinical examination results, and treatment records. In the data preprocessing stage, the original data was first cleaned, missing values were filled, and normalized to ensure that the data can adapt to the training of the LSTM model. We selected some common diseases, such as heart disease and diabetes, as research objects, and annotated the dataset with the diagnostic codes of these diseases. Finally, the dataset was divided into a training set and a test set to ensure the model's effective evaluation and generalization ability.

4.2. Experimental Results

In order to verify the effectiveness of the LSTM model in the task of disease prediction in electronic health records, this study conducted comparative experiments with several other commonly used machine learning models. Specifically, we selected support vector machine (SVM), random forest (RF) and multi-layer perceptron (MLP) as comparison models. Support vector machine is a powerful classifier that can improve prediction accuracy by maximizing classification intervals, but its computational complexity is high when processing large-scale data. Random forest makes predictions by constructing multiple decision trees, has strong generalization ability and robustness, and is suitable for processing complex nonlinear relationships. Multi-layer perceptron is a basic deep learning model that is trained using a fully connected network structure. Although it can capture complex patterns in the data to a certain extent, it is not as good as LSTM in modeling time series data. By comparing with these models, we can fully evaluate the advantages and actual effects of LSTM in disease prediction tasks. The experimental results are shown in Table 1.

Table 1. Experimental Results

Model	MSE	RMSE	MAE
SVM	0.0423	0.2057	0.1685
RF	0.0382	0.1954	0.1542
BILSTM	0.0359	0.1895	0.1468
LSTM	0.0267	0.1635	0.1320

Judging from the experimental results, the LSTM model has shown obvious advantages in the disease prediction task of electronic health records. Compared with other models, especially in mean square error (MSE), root mean square error (RMSE) and mean absolute In terms of evaluation indicators such as error (MAE), LSTM has significantly better performance than support vector machine (SVM), random forest (RF) and bidirectional LSTM (BiLSTM). Specifically, the MSE of LSTM is 0.0267, RMSE is 0.1635, and MAE is 0.1320. In contrast, the MSE and RMSE of SVM, RF, and BiLSTM are higher than LSTM, indicating that LSTM has more advantages in accuracy and model stability.

Although support vector machine (SVM) is a powerful classification model that performs well in many machine learning tasks, it shows certain limitations when processing electronic health records, which are high-dimensional data that contain time series characteristics. sex. Judging from the experimental results, the MSE of SVM is 0.0423, the RMSE is 0.2057, and the MAE is 0.1685. These values are much higher than LSTM, indicating that SVM fails to fully capture the long-term dependencies in the data when processing time series data. Although SVM may perform well with less static data and features, its performance is not as good as LSTM when faced with complex and dynamic electronic health record data.

As an integrated learning method, Random Forest (RF) uses the voting results of multiple decision trees for classification and regression. It can effectively reduce over-fitting problems and has strong adaptability to high-dimensional data. However, although random forests perform stably in many machine learning tasks, they fail to demonstrate the powerful advantages of LSTM when dealing with data such as electronic health records that contain time series and long-term dependencies. Random forest has an MSE of 0.0382, RMSE of 0.1954, and MAE of 0.1542, which although better than SVM, is still lower than LSTM. This indicates that while random forests can capture certain data patterns, they are not specifically optimized for time series data. As a result, they are relatively ineffective at modeling time-dependent features and fail to fully leverage the temporal information inherent in electronic health records.

Bidirectional long short-term memory network (BiLSTM) is an extension of standard LSTM that improves the performance of the model by considering both forward and backward time dependencies. Bidirectional LSTM can capture more contextual information in some tasks, but in this study, the performance of BiLSTM is still inferior to unidirectional LSTM. Its MSE is 0.0359, RMSE is 0.1895, and MAE is 0.1468. Although it is better than SVM and RF on these indicators, it still fails to surpass the one-way LSTM. Bidirectional LSTM often increases model complexity, which may lead to increased computational overhead and the risk of overfitting, especially when processing large-scale electronic health record data. In addition, although BiLSTM can consider both forward and backward dependencies, for certain disease prediction tasks, one-way LSTM is more stable and efficient due to its simpler structure and lower computational cost.

The significant advantage of LSTM is its ability to effectively capture dependencies over long time spans, and is particularly suitable for processing time series data typical of electronic health records. Electronic health records usually involve medical data at multiple time points, including diagnosis, treatment, medication use, physiological indicators, etc., and the relationship between these data often has long-term dependencies. Through its unique memory unit (cell state) and gating mechanism, LSTM can remember important information over a long period of time and filter out unnecessary interference. Therefore, LSTM can better capture the changing trend of patients' health status in this task, thereby providing more accurate results in disease prediction. In addition, LSTM has strong adaptive ability and can continuously optimize network parameters through the back-propagation algorithm, avoiding manual feature selection and over-fitting problems that may occur in traditional machine learning models.

Overall, the performance of LSTM in this experiment highlights its advantages in time series data modeling. Especially in data tasks such as electronic health records that contain multiple time series features, LSTM can effectively utilize its powerful time-dependent modeling capabilities to accurately predict the occurrence of diseases. This provides a valuable reference for future deep learning applications in the medical field, especially in disease prevention, personalized treatment and health management. LSTM has a very broad application prospect. Compared with traditional machine learning models, LSTM has demonstrated a stronger ability to process time series data and is expected to play a greater role in large-scale medical data analysis and clinical decision support in the future.

5. Conclusion

This study verified the superiority of long short-term memory network (LSTM) in the task of disease prediction in electronic health records through comparative experiments. The experimental results show that LSTM can effectively capture long-term dependencies when processing electronic health data containing time series features, and provides more accurate prediction results than support vector machine (SVM), random forest (RF) and bidirectional LSTM (BiLSTM). The lower mean square error (MSE), root mean square error (RMSE) and mean absolute error (MAE) of LSTM prove its accuracy and stability in disease prediction tasks. In the future, LSTM-based models are expected to play a greater role in the medical field, helping to identify disease risks early and support the formulation of personalized treatment plans, providing strong data support for clinical decision-making.

References

[1] Liu S., J. J. Schlesinger, A. B. McCoy, et al., "New onset delirium prediction using machine learning and long short-term memory (LSTM) in electronic health record", Proceedings of the 2023 Journal of the American Medical Informatics Association, vol. 30, no. 1, pp. 120-131, 2023.

[2] Javeed A., J. S. Berglund, A. L. Dallora, et al., "Predictive power of XGBoost_BiLSTM model: a machine-learning approach for accurate sleep apnea detection using electronic health data", Proceedings of the 2023 International Journal of Computational Intelligence Systems, vol. 16, no. 1, p. 188, 2023.

[3] P. Singhal, S. Gupta, Deepak, et al., "An integrated approach for analysis of electronic health records using blockchain and deep learning", Proceedings of the 2023 Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science), vol. 16, no. 9, pp. 1-10, 2023.

[4] M. Al Olaimat, S. Bozdag, and Alzheimer's Disease Neuroimaging Initiative, "TA-RNN: an attention-based time-aware recurrent neural network architecture for electronic health records", Proceedings of the 2024 Bioinformatics, vol. 40, Supplement_1, pp. i169-i179, 2024.

[5] H. Lu and S. Uddin, "Disease prediction using graph machine learning based on electronic health data: a review of approaches and trends", Proceedings of the 2023 Healthcare, MDPI, vol. 11, no. 7, p. 1031, 2023.

[6] S. Chintala, "The application of deep learning in analysing electronic health records for improved patient outcomes", Proceedings of the 2024 Chelonian Research Foundation, vol. 19, no. 1, 2024.

[7] G. Ramkumar, J. Seetha, R. Priyadarshini, et al., "IoT-based patient monitoring system for predicting heart disease using deep learning", Proceedings of the 2023 Measurement, vol. 218, p. 113235, 2023.

[8] Z. Gao et al., "Multi-Channel Hypergraph-Enhanced Sequential Visit Prediction," in Proc. ICEDCS, Sep. 2024, pp. 421-425.

[9] T. Mei et al., "Collaborative Hypergraph Networks for Enhanced Disease Risk Assessment," in Proc. ICEDCS, Sep. 2024, pp. 416-420.

[10] Y. Yang and C. Huang, "Tree-based RAG-Agent Recommendation System: A Case Study in Medical Test Data," arXiv preprint, Jan. 2025.

[11] X. Fei et al., "A Systematic Study on the Privacy Protection Mechanism of NLP in Medical Health Records," in Proc. ICSECE, Aug. 2024, pp. 1819-1824.

[12] Y. Cang et al., "Leveraging Deep Learning Techniques for Enhanced Analysis of Medical Textual Data," in Proc. ICSECE, Aug. 2024, pp. 1259-1263.

[13] Y. Cang et al., "ALBERT-Driven Ensemble Learning for Medical Text Classification," J. Comput. Technol. Softw., vol. 3, no. 6, 2024.

[14] J. Hu et al., "Accurate Medical Named Entity Recognition Through Specialized NLP Models," arXiv preprint, Dec. 2024.

[15] Z. Zheng et al., "Named Entity Recognition: A Comparative Study of Advanced Pre-Trained Models," J. Comput. Technol. Softw., vol. 3, no. 5, 2024.

[16] Y. Liang et al., "Contextual Analysis Using Deep Learning for Sensitive Information Detection," in Proc. CIPAE, Aug. 2024, pp. 633-637.

[17] X. Yan et al., "The Synergistic Role of Deep Learning and Neural Architecture Search in Advancing Artificial Intelligence," in Proc. ICEDCS, Sep. 2024, pp. 452-456.

[18] Z. Qi et al., "Optimizing Multi-Task Learning for Enhanced Performance in Large Language Models," arXiv preprint, Dec. 2024.

[19] J. Gao et al., "Multi-Level Attention and Contrastive Learning for Enhanced Text Classification with an Optimized Transformer," arXiv preprint, Jan. 2025.

[20] Z. Zheng et al., "Fully Convolutional Neural Networks for High-Precision Medical Image Analysis," Trans. Comput. Sci. Methods, vol. 4, no. 12, 2024.

[21] M. Sui et al., "Deep Learning-Based Channel Squeeze U-Structure for Lung Nodule Detection and Segmentation," in Proc. ICBASE, Sep. 2024, pp. 634-638.

[22] W. He et al., "Deep Learning in Image Classification: Evaluating VGG19's Performance on Complex Visual Data," arXiv preprint, Dec. 2024.

[23] J. Hu et al., "Few-Shot Learning with Adaptive Weight Masking in Conditional GANs," in Proc. ICEDCS, Sep. 2024, pp. 435-439.