Vol. 9 (2025): Proceedings of the 2025 International Conference on Artificial Intelligence, Modern Engineering and Environmental Sustainability
(IC-AIMEES 2025)
E-ISSN: 3105-0522 P-ISSN：3105-0514
DOI: https://doi.org/10.70267/ ic-aimees.202503

Machine Learning in Medical Diagnosis: From Image Analysis to EHRs

Xin Wang*

Beijing Jiaotong University, Computer Science, Beijing, 100044, China

Corresponding author: Xin Wang

Abstract

This article researched the widespread application of machine learning in medical diagnosis in depth, hoping to solve many of the current challenges in the medical field by improving the accuracy and efficiency of diagnosis. This paper expounds the important role of machine learning in disease prediction and patient risk stratification from three aspects: medical image analysis, electronic health record (EHR) processing and time series data analysis. Through algorithms such as supervised learning, unsupervised learning, and reinforcement learning, machine learning can extract valuable information from massive medical data to help doctors make more accurate diagnostic decisions. In the paper, the application of convolutional neural network (CNN), bidirectional encoder representation (BERT) and long short-term memory network (LSTM) in medical diagnosis is studied. Specifically, CNN performed well in medical image analysis and was able to quickly identify diseased areas; BERT extracts key information from unstructured text through natural language processing to support disease prediction and drug recommendations. LSTM specializes in processing time series data and can predict a patient’s disease risk. In addition, the successful cases of these models in data analysis such as pneumonia detection and heart disease prediction are also introduced. This shows their great potential in the field of medicine. Finally, the paper discusses the challenges of data privacy, model interpretability and algorithm optimization, and looks forward to future research directions such as federated learning, multimodal data fusion and real-time diagnosis, providing theoretical support and practical guidance for the intelligent transformation of healthcare industry.

Keywords

machine learning, medical diagnosis, electronic health records (EHR), deep learning models

1. Introduction

The field of medical diagnostics is facing increasing challenges, including the explosive growth of medical data and the unequal distribution of medical resources. There is an urgent need to improve the accuracy and efficiency of diagnosis. At the same time, machine learning technology has made breakthroughs in image recognition, gene sequence prediction and other fields. It offers unprecedented opportunities for medical diagnosis.

In today’s rapidly evolving medical field, diagnostic accuracy and efficiency are key factors in safeguarding patient health. However, medical diagnosis faces many serious challenges (Obermeyer and Emanuel, 2016). On the one hand, the early symptoms of the disease are often complex and changeable, and it is difficult to guarantee the accuracy of diagnosis. On the other hand, the uneven allocation of medical resources, especially the scarcity of professional doctors, makes it difficult to improve the efficiency of diagnosis. In addition, patients wait too long, affecting the overall quality of healthcare services (Organization, 2016).

The rapid development of machine learning technology has brought a new dawn to the field of medical diagnosis. Machine learning technology can process a wide variety of medical data sets to help doctors make medical diagnoses efficiently. Machine learning can quickly analyze large amounts of medical data to provide accurate evidence for the diagnosis of diseases. For example, (Esteva et al., 2017) demonstrated that through deep learning analysis of medical images, machine learning models can quickly identify diseased sites and assist physicians in early diagnosis. In the mining of electronic medical record data, machine learning can identify potential disease risk factors to support a doctor's medical diagnosis. In addition, machine learning can also automate decision-making and reduce the interference of human factors. This can improve the objectivity of diagnosis, thereby relieving the pressure of insufficient physician resources to a certain extent.

This article will comprehensively explore how machine learning algorithms can improve the shortcomings of medical diagnosis by improving accuracy and efficiency, and analyze the challenges and future directions.

2. Overview of Machine Learning Applications in Medical Diagnosis

Each algorithm of machine learning has its own unique advantages, enabling tasks ranging from precise disease prediction to personalized treatment optimization (Jiang et al., 2017).

2.1 Machine Learning Algorithms and Their Applicable Scenarios

The application of machine learning algorithms in medical diagnosis can be divided into the following three categories:

Supervised learning: Supervised learning relies on labeling data to learn and to make predictions or categorizations about new data. Algorithms such as "support vector machine (SVM), Random Forest and convolutional neural network (CNN)" are commonly used in supervised learning. According to Litjens et al. (2017), the advantage of supervised learning is that it can provide highly accurate predictions, but it requires a number of labeled data.

Unsupervised learning: Unsupervised learning algorithms are used to process unlabeled data, often for clustering and dimensionality reduction. In medical diagnosis, unsupervised learning can be used to discover underlying structures in patient populations or to identify abnormal data points (Miotto et al., 2018). Commonly used algorithms include "K-means clustering" and "autoencoder". The advantage of unsupervised learning is that there is no need to label the data, but the results are less interpretable.

Reinforcement learning: Reinforcement learning models can gradually learn optimal treatment strategies through continuous interaction with the environment. In medical diagnosis, it can dynamically adjust the dosage of drugs according to real-time changes in the patient's condition, avoiding overtreatment or under-treatment. Reinforcement learning can deal with dynamic environment, but the training process is complex and requires a lot of computing resources.

2.2 Application Scenarios and Successful Cases

Deep learning models have demonstrated exceptional success in medical image analysis. A deep learning model developed by Google Health has shown impressive performance in breast cancer screening, achieving a 5.7% reduction in misdiagnosis rates and a 1.2% decrease in false positive rates (McKinney et al., 2020).

Machine learning models, by analyzing electrocardiogram (ECG) data, have the capability to predict the risk of heart disease with high precision. A research team from Stanford University has created a model that can detect arrhythmias from ECG data with an impressive accuracy rate of 95% ((Rajpurkar et al., 2017).

Genomic data analysis: The integration of machine learning into genomic data analysis has resulted in extraordinary achievements. For instance, researchers using deep learning models to analyze genomic data have successfully predicted a patient's response to a specific cancer drug with 85% accuracy (Ching et al., 2018).

Deep learning excels in medical imaging. Recent studies have shown that these models exceed the accuracy of human experts in analyzing medical images. Specifically, a study published in Nature revealed that deep learning models attained an accuracy rate of 94.5% in lung cancer screening, compared to 93.1% accuracy for human experts (Ardila et al., 2019).

Recent years have witnessed remarkable advancements in genomic data analysis, driven by the powerful application of deep learning techniques. A study published in Cell, for instance, demonstrated that a deep learning model could predict cancer patient survival rates from genomic data with an impressive 89% accuracy (Khera et al., 2018).

3. Application of Machine Learning in Medical Image Analysis

Here, we will show the pneumonia detection based on **CNN** algorithm through the automatic detection of lung X-ray images.

3.1 CNN Model Preparation

3.1.1 Object Definition

The goal: to detect pneumonia automatically by X-ray images of the lungs.

Input: Lung X-ray image (PNG or DICOM format).

Output: Binary classification results (pneumonia/normal).

3.1.2 Data Pre-Processing

Using a publicly available dataset of chest X-ray images, such as CheXpert. The process of data pre-processing is shown in Figure 1.

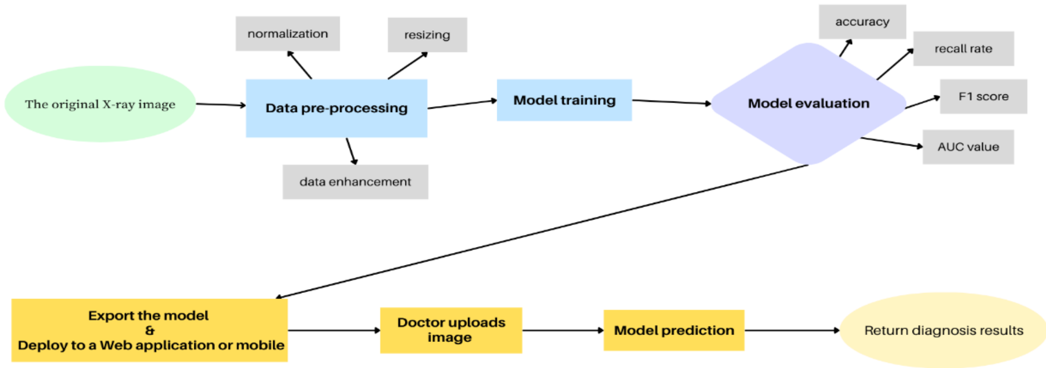
Figure1: Data preprocessing flow chart

3.2 CNN Model Establishment

3.2.1 Model Architecture

In the pneumonia X-ray image classification model based on deep learning, the input layer receives RGB images with a size of 224×224×3. Subsequently, as shown in Figure 2, the pre-trained convolutional neural networks such as ResNet50 are utilized for feature extraction, allowing the deep convolutional structure to fully learn the visual features of the images. On this basis, the network introduces 1-2 fully connected layers to further fuse and abstract the extracted high-level features. Finally, the output layer employs the Sigmoid activation function to map the results into binary classification outputs (pneumonia/normal). The model training uses binary cross-entropy as the loss function and selects the Adam optimizer for parameter updates, with a learning rate of 0.0001 to balance training stability and convergence efficiency.

Figure2: The whole process of pneumonia prediction model based on CNN



3.2.2 Model Training

An 80:20 split was applied to the dataset, separating it into training and validation sets, and the batch size was set to either 32 or 64. The number of training rounds is controlled at 50-80, adjusted according to the performance of the verification set.

3.2.3 Model Evaluation

(1) Confusion Matrix

As a foundational evaluation metric, the confusion matrix provides an intuitive visualization of the discrepancies between a classification model and true labels. In the case of binary classification problems, the confusion matrix is shown in Table 1:

Table1: Confusion matrix

Conditions	Predicted pneumonia	Predicted normal
Actual pneumonia	TP	FN
Actual normal	FP	TN

(2) Accuracy

Accuracy is calculated as the fraction of correctly predicted instances by the model over the entire set of samples. It can be calculated using the following formula:

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \tag{1}$$

(3) Precision

Accuracy was determined by dividing the number of pneumonia cases accurately classified by the model by the total number of true pneumonia cases. It is calculated using the formula below:

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

(4) Recall

The recall rate measures the fraction of actual pneumonia cases correctly classified by the model. It is calculated as below:

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

(5) F1-Score

The F1 score serves as a balanced evaluation metric, combining precision and recall through their harmonic average:

$$\text{F1} - \text{Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

(6) ROC curve and AUC values

The ROC curve, also known as the "Receiver Operating Characteristic" curve, plots the False Positive Rate (FPR) on its X-axis. AUC is calculated by

$$FPR = \frac{FP}{FP+TN} \tag{5}$$

On the ROC curve, the Y-axis denotes the True Positive Rate (TPR), which is determined by:

$$TPR = \frac{TP}{TP+FN} \tag{6}$$

3.3 Result

Figure3: ROC curve and confusion matrix when the amount of medical image is set to 100

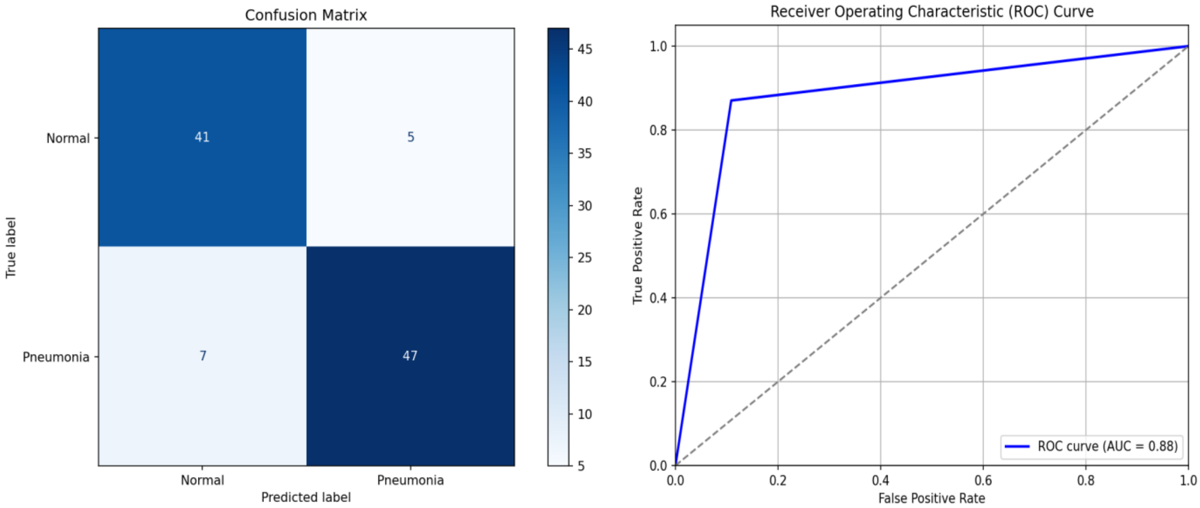
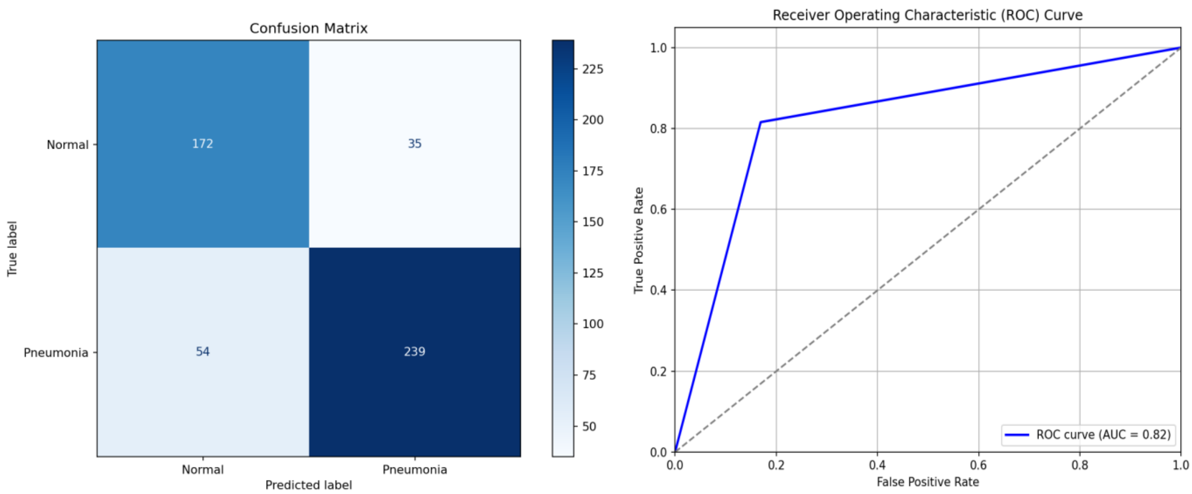


Figure 4: ROC curve and confusion matrix when the amount of medical image is set to 500



According to the subsequent calculation (Figure 3,4), it can be obtained:

The accuracy rate is 88%, the Precision rate is about 90.4%, the recall rate is about 87.0%, and the F1-Score is about 88.6%. This indicates that the model did well in predicting Pneumonia.

3.4 The Comparison of Multiple Algorithms

The comparison of multiple algorithms is shown in Table 2.

Table2: The comparison of multiple algorithms

Deep learning architecture	Advantages	Disadvantages	Applicable tasks	Manifestations of pneumonia
CNN	The structure is simple and the calculation efficiency is high.	Feature extraction ability is limited, it is difficult to capture global information.	Small scale sorting task.	Performance is mediocre, with an AUC value of about 0.75-0.80.
ResNet	The depth can be extended, and the feature extraction ability is strong.	This process is computationally intensive and necessitates the use of large datasets.	Large-scale classification task.	Excellent performance, AUC value of about 0.82-0.85.
U-Net	It is suitable for segmentation tasks and can accurately locate the lesion area.	Pixel-level annotation data is required, and the calculation cost is high.	Medical image segmentation.	The segmentation accuracy is high, and the Dice coefficient is above 0.85.
DenseNet	The feature reuse ability is strong, and the performance is better than ResNet.	High computational cost.	Large-scale classification task.	Excellent performance, AUC value of about 0.83-0.87.

4. Application of Machine Learning to Electronic Health Record Analysis

Electronic health records (EHRs) contain a large amount of patient information. Machine learning techniques, particularly natural language processing (NLP) and time series analysis, can extract valuable insights from electronic medical records. Recent studies have shown the effectiveness of advanced models, including BERT and LSTM, in managing the complexity and heterogeneity of EHR data (Rajpurkar et al., 2017, Topol, 2019). Below, we will explore the application of BERT to NLP tasks and the application of LSTM to the analysis of time series of electronic medical records.

4.1 Application of Natural Language Processing (NLP) in EHR

4.1.1 BERT Model

As shown in Figure 5, at the heart of BERT is the encoder part of Transformer. BERT can simultaneously extract and encode the contextual information for each word within the text, using the mask Language Model (MLM).

BERT’s input comprises a sequence of words, with each word represented as a vector. Input representations include:

Token Embeddings: Map each word to a fixed dimensional vector.

Position Embeddings: which encode the positional information of each word within the sequence.

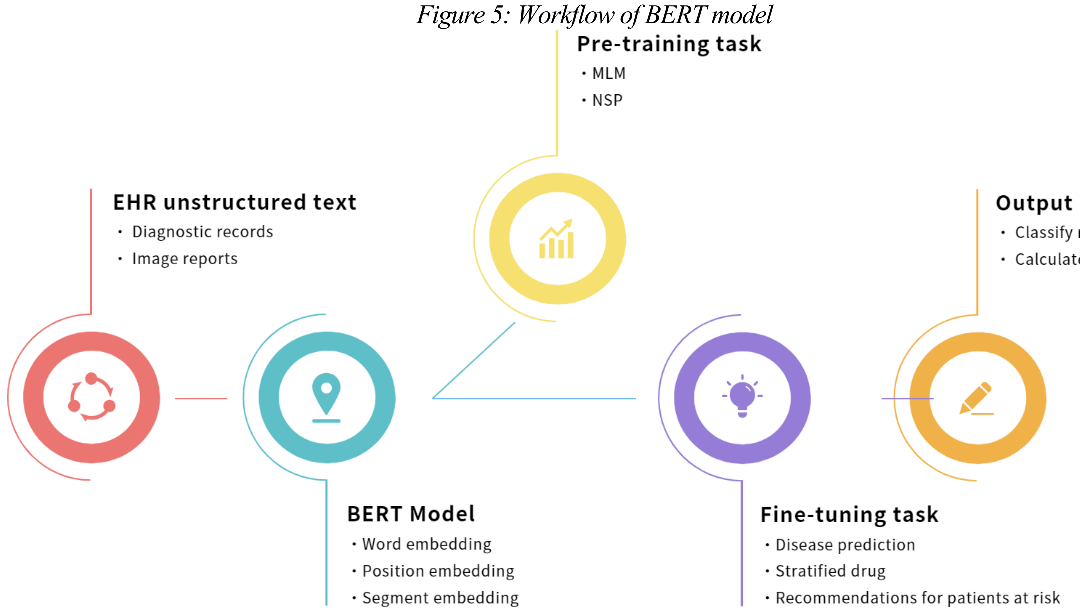
Segment Embeddings: which are utilized to differentiate between distinct sentences, such as diagnostic records and image reports in Electronic Health Records (EHRs).

BERT’s focus mechanism, self-attention mechanism, is calculated as:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{7}$$

where Q, K, V are the query, key and value matrices respectively, and $\sqrt{d_k}$ is the key dimension.

Figure 5: Workflow of BERT model



4.1.2 Result

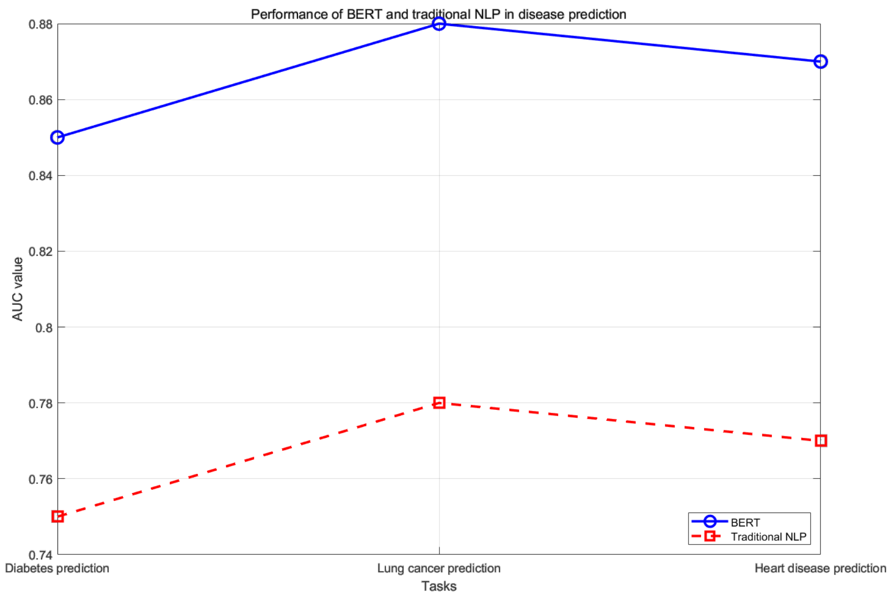
BERT's proficiency in managing various tasks within the Electronic Health Record (EHR) (Table 3):

Table 3: BERT's performance in handling different tasks in the HER

Tasks	Input data	Model output	Performance indicators
Disease prediction	diagnosis records, imaging reports	classification results (pneumonia/normal)	AUC: 0.85
Risk-stratified	patient history and medication record	risk score (low/medium/high)	AUC: 0.87
Drug recommendation	Medication records, allergy history	List of recommended drugs	Accuracy: 0.89

According to the study, the performance of BERT and traditional NLP in disease prediction can be visualized in Figure 6. BERT's AUC values for all tasks are higher than traditional NLP methods, indicating its superiority in EHR unstructured text analysis.

Figure 6: The performance of this model in disease prediction compared with the traditional model



4.1.3 Specific Applications and Positive Effects

Huang et al. (2020) used BERT model to extract diagnostic records from EHR to predict patients' diabetes risk. The study shows that the AUC value of BERT model reaches 0.85, which is better than the traditional NLP method. BERT excelled at disease prediction tasks, capturing key information in text.

Alsentzer et al. (2019) used BERT model to extract medical record information from EHR for patient risk stratification. The study showed that the BERT model achieved an AUC value of 0.87 in the cardiovascular disease risk stratification task. BERT can effectively extract key information from medical records to support patient risk stratification.

Zhang et al. (2019) used BERT model to extract medication records from EHR and recommend appropriate drugs. The study showed that the BERT model achieved 0.89 accuracy in the drug recommendation task. BERT can effectively extract key information from medication records to support drug recommendation.

BERT model has powerful context modeling capabilities and can extract useful information from the unstructured text of electronic medical records. It is suitable for the processing of complex text in electronic medical records. For example, BERT could understand that "high blood pressure" in "patient has a history of high blood pressure" was a disease name. BERT can be pre-trained and then fine-tuned to suit a specific task. BERT can handle multiple tasks simultaneously, such as disease prediction and patient risk stratification.

4.2 Time Series Analysis

4.2.1 LSTM Model

The goal of the model (Figure 7) is to use time series data from electronic medical records, such as heart rate, blood pressure and blood sugar, to predict a patient's risk of heart disease.

Input layer: Accepts input (24-hour heart rate data) of the shape (seq_length, 1).

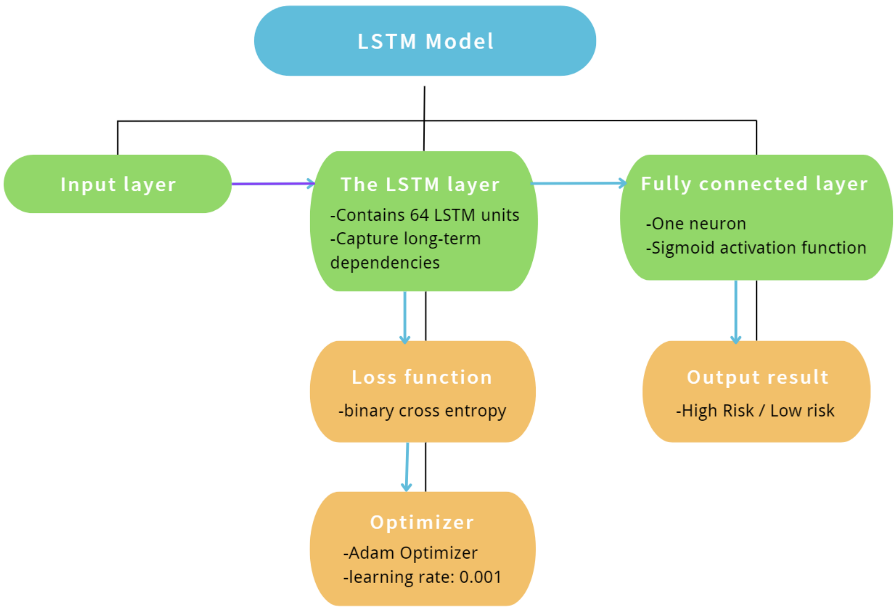
LSTM layer: 64 units that effectively model long-term dependencies within data.

Fully connected layer: 1 neuron, employing the Sigmoid activation function to produce binary classification outcomes.

Loss function: Binary Crossentropy.

Optimizer: Adam optimizer, learning rate set to 0.001.

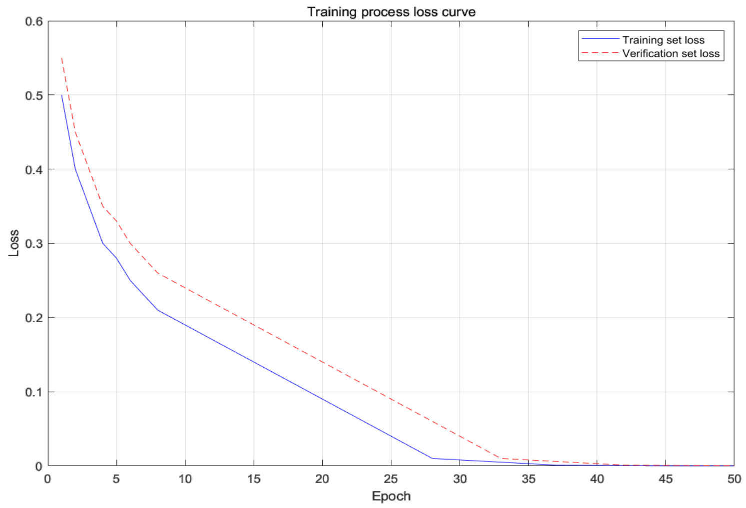
Figure 7: LSTM model flow chart



4.2.2 Result

Figure 8 is the training process loss curve:

Figure 8: Training process loss curve of LSTM model

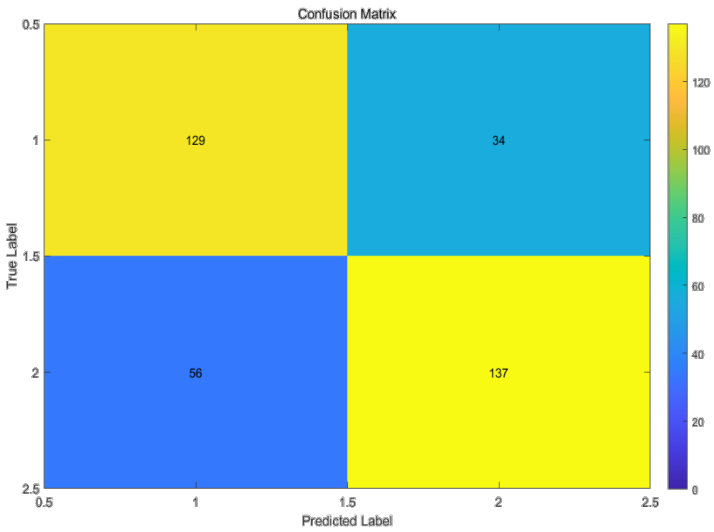


AUC (Area Under the Curve) here is:

$$AUC = \int_0^1 TPR(fpr)dfpr \tag{8}$$

Figure 9 is the confusion matrix diagram:

Figure 9: Confusion matrix diagram of LSTM model



Then,

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{266}{356} \approx 0.747$$

$$Recall = \frac{TP}{TP + FN} = \frac{129}{163} \approx 0.791$$

$$Precision = \frac{TP}{TP + FP} = \frac{129}{185} \approx 0.697$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \approx 0.716$$

The model exhibits strong overall prediction capabilities, with a moderate proficiency in recognizing positive samples. The model demonstrated a high precision rate, with the majority of its positive predictions being accurate, although further optimization is still achievable.

5. Challenges and Future Directions

Although machine learning shows substantial potential in medical diagnosis, there are still many challenges to its widespread application.

5.1 Challenges

In the process of medical information, the privacy protection of medical data is still a serious challenge. This type of data contains sensitive information, including the patient's medical history, treatment outcomes. It may also contain highly confidential content such as genomic data and lab test results. Once this information is obtained by an unauthorized third party, it may not only violate the patient's privacy rights, but also lead to a series of serious consequences. We can avoid privacy breaches as much as possible through machine learning methods such as federated learning (Rieke et al., 2020) and differential privacy (Dwork and Roth, 2014).

5.2 Future Directions

Based on the current challenges, the following three directions can be studied in depth:

Federated learning: Federated learning will become an important technology in medical diagnostics, especially when it comes to data privacy protection. It enables multiple healthcare organizations to train models without sharing sensitive data.

Multimodal data fusion: Medical diagnosis often involves multiple data types, such as medical images, genomic data, and electronic health records. Future research will focus on how to efficiently fuse multimodal data to improve diagnostic accuracy.

Real-time diagnosis and personalized treatment: With the development of edge computing and Internet of Things (IoT) technologies, machine learning models will enable real-time diagnosis and personalized treatment.

6. Conclusion

This paper reviews the application of machine learning in medical diagnosis, challenges and its future development direction. By analyzing medical images, electronic health records, and time series data, machine learning models show significant advantages in disease prediction, patient risk stratification, and personalized treatment recommendations. Deep learning models, such as CNN, BERT, and LSTM, perform particularly well in medical image analysis, text processing, and time series prediction, significantly outperforming traditional methods. For example, CNN can reduce the misdiagnosis rate and false positive rate in breast cancer screening; BERT extracts critical information from electronic medical records through natural language processing to support disease prediction and drug recommendations. However, despite the remarkable progress machine learning has made in medicine, there are still many challenges to its widespread application. In complex medical scenarios, issues such as data privacy and security, lack of model interpretability, and algorithm optimization need to be addressed. Future research directions include federated learning, multimodal data fusion, and real-time diagnostics. In short, integrating machine learning into medical diagnosis holds great promise. It is expected that the intelligent transformation of the medical industry will be promoted through technological innovation and interdisciplinary cooperation in the future. Ultimately, this will improve patient outcomes and improve the overall efficiency of healthcare delivery.

References

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T. and McDermott, M. B. A., (2019). Published. Publicly available clinical BERT embeddings. *roceedings of the 2nd Clinical Natural Language Processing Workshop*, Minneapolis, Minnesota. Association for Computational Linguistic, pp. 72-78.

Ardila, D., Kiraly, A. P., Bharadwaj, S., Choi, B., Reicher, J. J., Peng, L., Tse, D., Etemadi, M., Ye, W. and Corrado, G., (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature medicine,* vol. 25**,** no. 6, pp. 954-961.

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., Ferrero, E., Agapow, P.-M., Zietz, M. and Hoffman, M. M., (2018). Opportunities and obstacles for deep learning in biology and medicine. *Journal of the royal society interface,* vol. 15**,** no. 141, p. 20170387.

Dwork, C. and Roth, A., (2014). The algorithmic foundations of differential privacy. *Foundations and trends® in theoretical computer science,* vol. 9**,** no. 3–4, pp. 211-407.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M. and Thrun, S., (2017). Dermatologist-level classification of skin cancer with deep neural networks. *nature,* vol. 542**,** no. 7639, pp. 115-118.

Huang, K., Altosaar, J. and Ranganath, R., (2020). Published. ClinicalBERT: Modeling clinical notes and predicting hospital readmission. *Proceedings of the ACM Conference on Health, Inference, and Learning*, Toronto, ON, Canada. ACM, pp. 1-9.

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H. and Wang, Y., (2017). Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology,* vol. 2**,** no. 4, pp. 230-243.

Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., Choi, S. H., Natarajan, P., Lander, E. S., Lubitz, S. A. and Ellinor, P. T., (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature genetics,* vol. 50**,** no. 9, pp. 1219-1224.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B. and Sánchez, C. I., (2017). A survey on deep learning in medical image analysis. *Medical image analysis,* vol. 42, pp. 60-88.

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S. and Darzi, A., (2020). International evaluation of an AI system for breast cancer screening. *Nature,* vol. 577**,** no. 7788, pp. 89-94.

Miotto, R., Wang, F., Wang, S., Jiang, X. and Dudley, J. T., (2018). Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics,* vol. 19**,** no. 6, pp. 1236-1246.

Obermeyer, Z. and Emanuel, E. J., (2016). Predicting the future—big data, machine learning, and clinical medicine. *The New England journal of medicine,* vol. 375**,** no. 13, pp. 1216-1219.

Organization, W. H., (2016). *Global strategy on human resources for health: workforce 2030,* Geneva, Switzerland: World Health Organization.

Rajpurkar, P., Hannun, A. Y., Haghpanahi, M., Bourn, C. and Ng, A. Y., (2017). Cardiologist-level arrhythmia detection with convolutional neural networks. *arXiv preprint arXiv:1707.01836*.

Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A. and Maier-Hein, K., (2020). The future of digital health with federated learning. *NPJ Digital Medicine,* vol. 3**,** no. 1, p. 119.

Topol, E. J., (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine,* vol. 25**,** no. 1, pp. 44-56.

Zhang, Y., Chen, Q., Yang, Z., Lin, H. and Lu, Z., (2019). BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Scientific data,* vol. 6**,** no. 1, p. 52.

Funding
This research received no external funding.

Conflicts of Interest
The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Open Access