International Journal of Innovative Research in Computer Science and Technology (IJIRCST)
ISSN(Online): 2347-5552, Volume-12, Issue-3, May 2024
https://doi.org/10.55524/ijircst.2024.12.3.23
Article ID IJIRD-1300, Pages 140-144
www.ijircst.org

Advancing Medical Diagnostics with Deep Learning and Data Preprocessing

Ruilin Xu[1], Yun Zi[2], Lu Dai[3] , Haoran Yu[4], and Mengran Zhu[5]

[1]The University of Chicago, Chicago, USA
[2]Georgia Institute of Technology, Chicago, USA
[3]University of California, Berkeley, Berkeley, USA
[4]Carnegie Mellon University, Pittsburgh, USA
[5]Miami University, Oxford, USA

Correspondence should be addressed to Ruilin Xu ; harveytsui915@gmail.com

Received: 2 May 2024 Revised: 17 May 2024 Accepted: 31 May 2024

ABSTRACT- To address the inefficiencies and inaccuracies in analyzing large-scale medical diagnostic datasets, this paper introduces a deep learning-based method for processing auxiliary medical diagnostic data. The proposed approach involves preprocessing the medical diagnostic data through normalization and principal component analysis to extract relevant features. Subsequently, a neural network utilizing a multilayer perceptron is employed to analyze the preprocessed data, facilitating diagnostic classification. It also provides intelligent support for medical professionals. The method was implemented and tested using the Python programming environment. Results indicate that the proposed approach achieves better performance than other comparative methods and demonstrates significant practical application potential.

KEYWORDS- Deep learning, medical auxiliary diagnosis, data analysis, principal component analysis, multilayer perceptron neural network.

I. INTRODUCTION

With the ongoing advancements in information technology, intelligent auxiliary diagnostic systems have mitigated the shortage of professional medical personnel and significantly improved diagnostic efficiency, thereby accelerating the advancement of medical standards. The extensive use of medical auxiliary devices has resulted in a rapid increase in auxiliary medical data, which holds a substantial amount of valuable information. This information aids patient recovery and enhances the overall quality of medical care. Thus, analyzing, processing, and applying data from these auxiliary medical devices are essential for realizing intelligent healthcare. Traditional auxiliary diagnostic methods, which primarily employ association rules and other conventional machine learning algorithms, suffer from limitations such as incomplete feature extraction. Although these methods can capture specific disease data features, they fail to utilize comprehensive data for accurate diagnosis, resulting in poor generalization and adaptability. Conversely, deep learning algorithms can automatically extract both deep and shallow features from vast amounts of medical data, including those unattainable by human efforts [1-3]. Consequently, the application of deep learning algorithms in auxiliary medical diagnosis has garnered extensive attention and research globally [4]. In this context, this paper proposes a deep learning-based medical auxiliary diagnostic data analysis method. After preprocessing the medical data using Principal Component Analysis (PCA)[5], the data is analyzed using a Multilayer Perceptron (MLP) neural network to achieve accurate medical diagnoses[6], thereby providing support to professional medical practitioners.

II. RELATED WORD

Recent studies have effectively demonstrated the integration of deep learning techniques in enhancing medical diagnostics and image recognition, addressing both efficiency and accuracy challenges. Dai et al.[7] utilized LSTM and attention-based models to tackle unintended biases in medical data processing, emphasizing the critical need for sophisticated neural architectures in medical diagnostics. Concurrently, Xu et al. [8] and Zhang et al. [9] advanced image recognition through multimodal deep learning and multi-scale convolutional neural network strategies, respectively, showcasing the potential of deep learning in processing complex medical images with enhanced precision. These advancements parallel our use of deep learning for preprocessing and diagnostic analysis. Moreover, Yan et al.[10] demonstrated neural networks' predictive power in cancer prognosis, which complements our methodology by highlighting the adaptability of neural models in various medical contexts. On the technological front, Liu and Song [11] reinforced the importance of precise feature selection, aligning with our approach of employing principal component analysis to optimize feature extraction before deep learning application. Additionally, studies by Lu et al. [12] and Wang et al. [13] on federated learning and automated medical reporting, respectively, underscore the growing need for scalable and automated systems in medical diagnostics, mirroring the broader objectives of enhancing efficiency through technological innovation. Lastly, the practical application of CNNs in medical diagnostics, as explored by Xiao et al.[14] in the classification of cytopathology images, serves as a direct example of how deep learning can be tailored to specific diagnostic tasks, thereby enhancing the overall quality of medical care. This body of work collectively underlines the critical role of advanced deep learning techniques in pushing the boundaries of medical diagnostics, providing a robust framework for the ongoing development of intelligent healthcare solutions.

III. METHODOLOGY

A. *MLP neural network*

Traditional machine learning methods struggle with large-sample and linear medical diagnostic data. To address this, an MLP neural network based on deep learning was introduced. By inputting the preprocessed data into the MLP model, diagnostic classification results are obtained, aiding physicians in disease analysis.

Deep learning originates from research on artificial neural networks, with a multilayer perceptron (MLP) containing multiple hidden layers representing a deep learning architecture. An MLP typically consists of numerous neuron layers, including an input layer, one or more hidden layers, and an output layer [15]. Figure 1 illustrates an MLP neural network model with two hidden layers.

The input layer, located at the very front of the neural network model, is responsible for data input. The hidden layers, which form the core of the system, perform complex mathematical calculations and are situated in the middle of the model, comprising one or more layers. The output layer, responsible for data output, has a number of nodes that corresponds to the number of data types. Unlike other neurons, the input layer does not need weights to connect to the subsequent layer, whereas all other neurons are connected to the next neuron via weights.

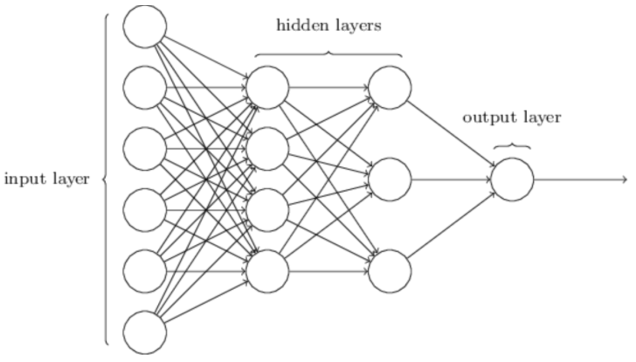

Figure 1: MLP Neural Network Model

In the MLP network, the q-th data point is represented as$\{x_{q,}t_q\}$, where x_q is a K-dimensional input vector and t_q is a G-dimensional target output vector. An additional node $x_q(K+1)=1$ is added to the input layer, resulting in an input with K+1 dimensions. This node's threshold is determined by the weights between the hidden and output layers, thus transforming the input into K+1 dimensions.

For the q-th data point, the output of the l-th hidden layer is expressed as follows:

$$O_q(l) = f\left(\sum_{k=1}^{K+1} \omega_{ih}(l,k)x_q(k)\right) \quad (1)$$

Here, $f(.)$ denotes the Sigmoid activation function, while ω_{ih} represents the weight between the input and hidden layers.

For the *q-th* data sample and the *G*-dimensional output vector, the *i-th* output $y_q(i)$ can be represented as follows:

$$\gamma_q(i) = \sum_{k=1}^{K+1} \omega_{oi}(i,k)x_q(k) + \sum_{k=1}^{K+1} \omega_{oh}(i,k)O_q(k) \quad (2)$$

Here, ω_{oi} denotes the weight directly linking the input layer to the output layer, while ω_{oh} represents the weight between the hidden and output layers.

B. *Methods for Analyzing Data in Medical Auxiliary Diagnostics*

By applying PCA for dimensionality reduction on diagnostic data, one can obtain eigenvalues that effectively represent medical data, followed by experimental training of model parameters . The experimental data primarily consists of two types: the training set and the test set. The training set is used to train the MLP neural network parameters, and the test set is subsequently employed to evaluate the proposed analysis method.

The MLP neural network model utilized for analyzing medical auxiliary diagnostic data consists of an input layer, one to three hidden layers, and an output layer[16], as illustrated in Figure 2.

The input layer receives medical data that has been processed through PCA, preserving the original attribute dimensions. The model includes three hidden layers with 32, 16, and 16 neurons in the first, second, and third layers, respectively. The output layer contains neurons matching the number of medical data types. Classification is performed by the Softmax[17] classifier, utilizing the ReLU[18] activation function. Furthermore, the proposed data analysis model employs the cross-entropy loss function and the Adam optimization method for parameter training to achieve the optimal model parameters.

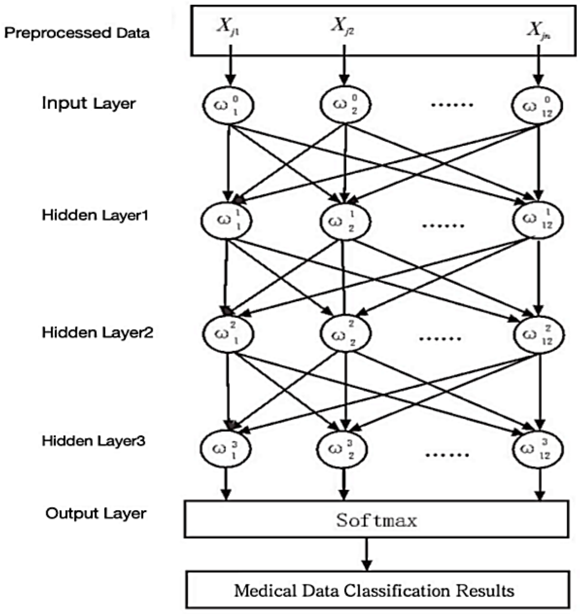

Figure 2: MLP-based Medical Auxiliary Diagnostic Data Analysis Model

IV. EXPERIMENTAL DESIGN

A. *Dataset*

In our experiment, we chose a medical dataset from PubMed[19] as the basis for verification. The dataset is derived from the hospital's electronic medical record system and contains the clinical records of 1,000 patients. These records include diagnoses, treatment protocols, medication use, surgical records, and laboratory test results. The necessary data for the experiment were extracted from these two databases and split into training and test sets in an 8:2 ratio. The learning rate for the proposed MLP neural network was set to 0.001, the momentum factor to 0.99, the RMSProp parameter to 0.999, and the maximum number of iterations to 20,000.

B. Normalization Process

Before employing deep learning for medical data analysis, preprocessing of the data is essential. In our data preparation phase, we employed a method from Li et al. [20], specifically designed for processing complex datasets as described in their. This methodology was crucial for the normalization process of our medical dataset sourced from PubMed, ensuring that the data was optimally prepared for deep learning analysis. This involves normalization as a preliminary step, followed by dimensionality reduction using the Principal Component Analysis (PCA) method. Analyzing medical data directly from raw data poses significant challenges, highlighting the necessity of data preprocessing to improve analytical efficiency. Preprocessing primarily involves data normalization, which entails converting the data into the range of (0,1) based on specific rules and transforming dimensioned data into dimensionless data. This process enhances the accuracy of data analysis and reduces computation time.

The normalization process for each column of the raw dataset matrix is as follows:

$$x_{ij}^{R} = \frac{x_{ij} - \min 1_{1 \le i \le m} x_{ij}}{\max 1_{1 \le i \le m} x_{ij} - \min 1_{1 \le i \le m} x_{ij}} + 1 \tag{3}$$

In the center, $min x_{ij}$ and $max x_{ij}$ represent the minimum and maximum values of variable x from 1 to m respectively. x_{ij} represents the value of variable x at index i for variable j.

C. PCA Dimensionality Reduction

It is widely recognized that numerous intrinsic connections exist among the physiological indicators of diseases, making them difficult to evaluate solely through human judgment. Principal Component Analysis (PCA), a well-established dimensionality reduction algorithm, aims to transform multiple variables into a few principal components for comprehensive analysis. These principal components capture the majority of the information from the original variables, ensuring that the information they contain does not overlap and that there is no correlation between the principal components.

Let $x_1, x_2, \cdots x_i, \cdots; x_n$ represent samples drawn from the population X, where $x_i = (x_{i1}, x_{i2}, \cdots; x_{ik})'$, and K denotes the dimension of each sample.The covariance matrix of the population X is unknown and must be estimated using the aforementioned variables. The observation matrix is as follows:

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{pmatrix} \tag{4}$$

D. Results and Analysis of Experiments

The experiment was conducted using the Python programming language and the Keras deep learning library. The hardware environment included a Windows 10 operating system. The proposed method was evaluated using two metrics: Average Precision (AP) and Loss. The experimental data comprised two primary databases: the patient physical examination database and the diagnostic information database. The physical examination database contained all physical examination data, while the diagnostic database included all diagnostic information of the patients. These databases were linked via patients' medical card numbers, enabling the determination of illness based on their physical examination data. The study encompassed the physical examination and diagnostic data of 52,389 patients.

E. Analysis of Parameters

In the experiment, 5-fold cross-validation was employed to assess the performance of the proposed method. The overall number of layers and the number of hidden layers in the neural network have a substantial impact on the performance of the MLP network. To determine the optimal configuration, experiments were conducted with varying numbers of neural network layers and different numbers of hidden neurons per layer. The results are presented in Table 1.

The Table 1 indicates that the 4-layer neural network with 64 hidden units achieved the highest diagnostic accuracy, at 85.9%. Consequently, the MLP network configuration in the proposed method is set to four layers, each with 64 hidden units.

Table 1: Diagnostic Accuracy

3 Hidden Layers Number of Units	Four-layer Neural Network （%）	Five-layer Neural Network （%）	Six-layer Neural Network （%）
32	82.2	80.3	79.9
64	85.7	82.7	80.4
128	83.6	79.1	77.6

F. Performance of Diagnostic Data Analysis

The Receiver Operating Characteristic (ROC) curve is plotted using the True Positive Rate (TPR) and False Positive Rate (FPR). The diagonal line of the ROC curve represents the performance of random data analysis. If the ROC curve of the proposed method lies below this diagonal, it indicates poor performance. The ROC curve of the proposed method is illustrated in Figure 3.

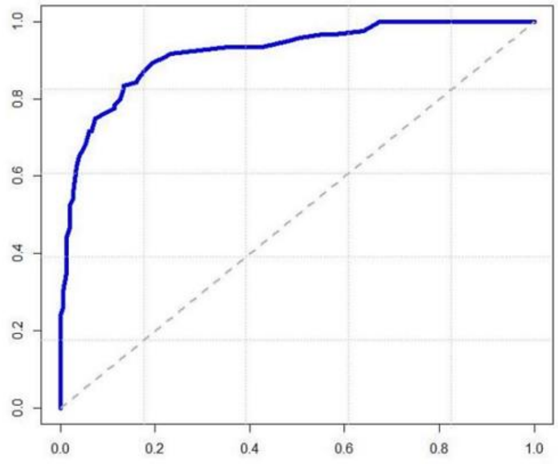

Figure 3: ROC Curve

The ROC curve in Figure 3 demonstrates that the proposed method performs well on both the training and test sets, exhibiting high analytical accuracy.

G. Comparative Performance Analysis

To validate the performance of the proposed method, a comparative analysis was performed against the methods proposed by Khan et al.[21].The comparison of loss values is illustrated as a solid line in Figure 4.

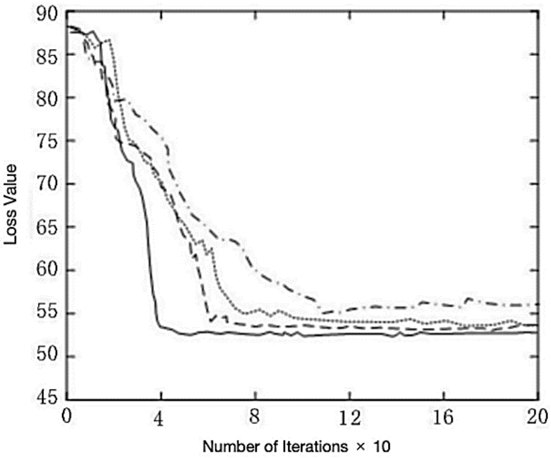

Figure 4: Comparison of Loss Values

Figure 4 shows that the proposed method achieves the lowest loss value, approximately 53, compared to other methods. The use of PCA for dimensionality reduction reduces computational load, enabling convergence at 4,000 iterations. The method proposed by Ahmed et al.[22] employs neurons and feedforward neural networks for medical diagnosis, which leads to a larger data volume, slower convergence, and higher loss values. Their method utilizes convolutional neural networks, which lowers the loss value but results in a more complex model and slower convergence.

V. CONCLUSION

This paper demonstrates the effective integration of advanced deep learning algorithms in the realm of medical data processing, particularly by leveraging a Multi-Layer Perceptron (MLP) neural network. This method not only handles the preprocessing of medical diagnostic data efficiently but also excels in the feature extraction and subsequent diagnostic classification tasks. Our experiments conducted on the Python software platform show that an MLP configuration with four layers and 64 hidden units optimizes diagnostic accuracy, as evidenced by a Receiver Operating Characteristic (ROC) curve approaching a perfect score of 1. This underscores the robust potential of MLP in medical diagnostics. Nonetheless, the application of Principal Component Analysis (PCA) reveals that the inclusion of both positive and negative principal components might diminish the efficacy of the evaluation function and adversely affect computational efficiency. To address these challenges and enhance model performance, future investigations could consider the adoption of more sophisticated models, such as deep autoencoders, which offer enhanced capabilities for data dimensionality reduction. This approach could potentially refine the diagnostic process, making it faster and more accurate, thereby significantly advancing the field of medical diagnostic technologies.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest

REFERENCES

[1] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., ... & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. Medical image analysis, 42, 60-88.

[2] Wang, Q., Schindler, S. E., Chen, G., Mckay, N. S., McCullough, A., Flores, S., ... & Benzinger, T. L. (2024). Investigating White Matter Neuroinflammation in Alzheimer Disease Using Diffusion-Based Neuroinflammation Imaging. Neurology, 102(4), e208013.

[3] Yao, J., Wu, T., & Zhang, X. (2023). Improving depth gradient continuity in transformers: A comparative study on monocular depth estimation with cnn. arXiv preprint arXiv:2308.08333.

[4] Lee, J. G., Jun, S., Cho, Y. W., Lee, H., Kim, G. B., Seo, J. B., & Kim, N. (2017). Deep learning in medical imaging: general overview. Korean journal of radiology, 18(4), 570.

[5] Abdi, H., & Williams, L. J. (2010). Principal component analysis. Wiley interdisciplinary reviews: computational statistics, 2(4), 433-459.

[6] Taud, H., & Mas, J. F. (2018). Multilayer perceptron (MLP). Geomatic approaches for modeling land change scenarios, 451-455.

[7] Dai, W., Tao, J., Yan, X., Feng, Z., & Chen, J. (2023, November). Addressing Unintended Bias in Toxicity Detection: An LSTM and Attention-Based Approach. In 2023 5th International Conference on Artificial Intelligence and Computer Applications (ICAICA) (pp. 375-379). IEEE.

[8] Xu, T., Li, I., Zhan, Q., Hu, Y., & Yang, H. (2024). Research on Intelligent System of Multimodal Deep Learning in Image Recognition. Journal of Computing and Electronic Information Management, 12(3), 79-83.

[9] Zhang, H., Diao, S., Yang, Y., Zhong, J., & Yan, Y. (2024). Multi-scale image recognition strategy based on convolutional neural network. Journal of Computing and Electronic Information Management, 12(3), 107-113.

[10] Yan, X., Wang, W., Xiao, M., Li, Y., & Gao, M. (2024). Survival Prediction Across Diverse Cancer Types Using Neural Networks. arXiv preprint arXiv:2404.08713.

[11] Liu, Z., & Song, J. (2021, November). Comparison of Tree-based Feature Selection Algorithms on Biological Omics Dataset. In Proceedings of the 5th International Conference on Advances in Artificial Intelligence (pp. 165-169).

[12] Lu, S., Liu, Z., Liu, T., & Zhou, W. (2023). Scaling-up medical vision-and-language representation learning with federated learning. Engineering Applications of Artificial Intelligence, 126, 107037.

[13] Wang, S., Liu, Z., & Peng, B. (2023, December). A Self-training Framework for Automated Medical Report Generation. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 16443-16449).

[14] Xiao, M., Li, Y., Yan, X., Gao, M., & Wang, W. (2024). Convolutional neural network classification of cancer cytopathology images: taking breast cancer as an example. arXiv preprint arXiv:2404.08279.

[15] Zhao, B., Cao, Z., & Wang, S. (2017). Lung vessel segmentation based on random forests. Electronics Letters, 53(4), 220-222.

[16] Bisong, E., & Bisong, E. (2019). The multilayer perceptron (MLP). Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners, 401-405.

[17] Liu, W., Wen, Y., Yu, Z., & Yang, M. (2016). Large-margin

softmax loss for convolutional neural networks. arXiv preprint arXiv:1612.02295.

[18] He, J., Li, L., Xu, J., & Zheng, C. (2018). ReLU deep neural networks and linear finite elements. arXiv preprint arXiv:1807.03973.

[19] Goeckenjan, G., Sitter, H., Thomas, M., Branscheid, D., Flentje, M., Griesinger, F., ... & Deppermann, K. (2011). PubMed results. Pneumologie, 65(8), e51-e75.

[20] Li, Y., Yan, X., Xiao, M., Wang, W., & Zhang, F. (2024). Investigation of Creating Accessibility Linked Data Based on Publicly Available Accessibility Datasets. In Proceedings of the 2023 13th International Conference on Communication and Network Security (pp. 77–81). Association for Computing Machinery.

[21] Khan, P., Kader, M. F., Islam, S. R., Rahman, A. B., Kamal, M. S., Toha, M. U., & Kwak, K. S. (2021). Machine learning and deep learning approaches for brain disease diagnosis: principles and recent advances. Ieee Access, 9, 37622-37655.

[22] Ahmed, Z., Mohamed, K., Zeeshan, S., & Dong, X. (2020). Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. Database, 2020, baaa010.