Dean&Francis
ISSN 2959-6157

Disease Diagnosis Based on Machine Learning

Yehao Huang

Institute of Information Technology, Jilin Agricultural University, Changchun, China
knccqmjh@pgu.edu.pl

Abstract:

In recent years, disease diagnosis based on machine learning gradually become a hot topic. With the advancement of technology and medical standards, machine learning has also developed rapidly. This paper reviews the research background of disease diagnosis based on machine learning, the principles of machine learning, its applications in real life, and the challenges. First, the current situation of medical diagnosis is analyzed, including the limitations of traditional diagnosis methods and the advantages of machine diagnosis, Second, the key principles, processes, and core algorithms of machine learning are expounded. Then, the application of disease diagnosis based on machine learning in diseases such as breast cancer, Parkinson's disease, and osteoporosis is introduced in detail, and the performance of research methods and algorithms in the model is analyzed in combination with actual cases. Finally, the challenges faced by machine learning in the field of disease diagnosis and the prospects for the future are summarized. For example, there are challenges such as data quality, algorithm selection, and model interpretation. It looks forward to the development directions such as data privacy protection and sharing in the future, as well as technological innovation through multidisciplinary integration, providing references for relevant research.

Keywords: Machine learning; Disease diagnosis; Parkinson; Osteoporosis; Breast cancer

1. Introduction

In the current context, a series of factors, including the uneven distribution of medical equipment, the limitations of technical diagnosis, and individual differences among patients, have made it extremely difficult to achieve high - precision and rapid diagnosis. Moreover, most traditional medical diagnosis methods rely on doctors' personal experience. This situation is highly subjective. As the condition worsens, doctors have to deal with more medical data, which significantly reduces the accuracy of diagnosis. In contrast, machine learning can process a large amount of medical data simultaneously, analyze and

integrate various types of data, making diagnosis more efficient and the results more accurate. At the same time, it can also formulate more appropriate treatment plans for special patients. This promotes disease diagnosis based on machine learning to play a crucial role in the future. Nowadays, there are different diagnostic models and relevant algorithms for different diseases, such as decision trees, logistic regression, deep learning neural networks, support vector machines, etc. They have made great contributions to the progress of machine diagnosis. However, there are still some research gaps. In some studies, the data sample size is not large enough, which makes the research models unable to be widely used in diagnosis. There are also deficiencies in the interpretability of the models. The exploration of the practicality of different algorithms in different diseases and how to optimize them is not in - depth enough. Some studies fail to highlight their own advantages by comparing with other studies. G. Harshitha et al. proposed using computer vision technology to detect cotton diseases in response to the problem of yield decline caused by cotton diseases. They used pictures of healthy and infected cotton to train a deep learning model, enabling it to classify the state of cotton. The accuracy of this model reached 97.13% [1]. Second, Park et al. proposed a machine model for disease diagnosis based on laboratory tests to predict diseases. The relevant research results were successfully published in Scientific Reports, and the corresponding learning model was also successfully developed [2]. Then, PATEL, et al. proposed a disease diagnosis system based on machine learning, conducted research on building a disease diagnosis system with machine learning, and successfully published the results in Journal of Pharmaceutical Research International. The second part of this article mainly introduces relevant algorithms for disease diagnosis based on machine learning, and the third part mainly describes the application of machine learning in disease diagnosis [3]. Author: Poudel, S. The paper explores machine learning in disease diagnosis, comparing algorithms to find the best, with 20 machine learning methods tested on the Pima Indians Diabetes Dataset via tools like Autogluon; results show most exceeded the 65% baseline accuracy, with the best around 77% [4].

2 Machine Learning Related Models and Algorithms

2.1 Logistic Regression

Logistic regression is a statistical learning method used to solve binary classification problems. Although the word “regression” is in its name, it is used for classification tasks. The core principle is to map the output (continuous value) of linear regression to a probability value between 0 and 1 through the Sigmoid function, so as to judge the possibility of a sample belonging to a certain category. The Sigmoid function is in the form of $S(z)=1/(1+e(-z))$, where z is the result of linear regression $(z=w_1x_1+w_2x_2+...+w_nx_n+b)$, w is the weight, and b is the bias. During training, logistic regression optimizes the parameters through the maximum likelihood estimation method to maximize the predicted probability of the model for the training data. During prediction, 0.5 is usually used as the threshold. If the probability is greater than 0.5, it is classified into one category; if it is less than 0.5, it is classified into the other category. It is widely used in scenarios such as credit scoring and disease prediction. Its advantages are that the model is simple, highly interpretable, and can output classification probabilities; its disadvantages are that it is difficult to handle non-linear relationships and is sensitive to outliers.

2.2 Support Vector Machine

Support Vector Machine, abbreviated as SVM, is a supervised learning model mainly used for classification problems and regression tasks. Its core idea is to find the optimal hyperplane in the feature space for classification. The core principle is to search for the hyperplane that maximizes the margin by calculating the distance (i.e., the margin) from sample points to the hyperplane. The sample points closest to the hyperplane are called “support vectors”, which determine the position of the hyperplane and are crucial to the model. The core principle is to calculate the distance (margin) from sample points to the hyperplane and find the hyperplane that maximizes the margin. The sample points closest to the hyperplane are called “support vectors”, which determine the position of the hyperplane and are the key to the model. When dealing with non - linear problems, when the data is linearly inseparable, SVM uses kernel functions such as polynomial kernels and Gaussian kernels to map low - dimensional features to a high - dimensional space, making the data linearly separable in the high - dimensional space and avoiding the complexity of direct high - dimensional calculations. One advantage is its strong generalization ability, especially its good performance in small-sample scenarios. it controls the model complexity through regularization parameters, effectively alleviating overfitting; it has good adaptability to high - dimensional data. It is widely used in fields such as image recognition, text classification, and bioinformatics, for example, spam filtering and handwritten digit recognition. SVM combines theo-

ISSN 2959-6157

retical rigor and practical effects and is an important basic model in the field of machine learning.

2.3 Decision Trees and Random Forests

A decision tree is a supervised learning algorithm based on a tree-like structure. Its core principle is to simulate human decision-making logic for classification or regression. It builds a model by recursively partitioning data: starting from the root node, the data is split according to features each time. The criteria for selecting the optimal partitioning feature include information gain (used in the ID3 algorithm), information gain ratio, or Gini index, etc., until the data categories in the child nodes tend to be consistent or the stopping conditions are met. Its advantages lie in its extremely high interpretability. The tree structure is intuitive and easy to understand, clearly showing the decision path. The training process is fast, and it has relatively low requirements for data preprocessing. However, the disadvantages are also obvious. It is prone to overfitting the training data, resulting in poor generalization ability. The model may be too complex and sensitive to noise. Usually, optimization is required through pruning, such as pre - pruning and post - pruning. It is suitable for scenarios where clear decision - making logic is needed, such as formulating credit approval rules. A random forest is a typical application of the Bagging idea in ensemble learning, composed of multiple decision trees. Its construction process introduces double randomness: first, multiple different training sets are generated from the original data through bootstrap sampling; second, only a part of the features is randomly selected for partitioning during the training of each tree. The final result is determined by "voting" or taking the mean of multiple trees. This design effectively reduces the risk of overfitting. Its stability and generalization ability far exceed those of a single decision tree. It can handle high - dimensional data and is not sensitive to noise. However, the model complexity is relatively high, the training time is longer, and the interpretability is weaker than that of a single decision tree, making it difficult to intuitively show the decision - making basis. It is widely used in scenarios with high - precision requirements, such as medical diagnosis and auxiliary analysis of image recognition.

2.4 Deep Learning and Neural Networks

Deep learning is a branch of machine learning. Its core is to simulate the information - processing mode of the human brain through multi - layer non - linear neural networks and automatically learn complex features in data. Different from traditional machine learning, which relies on manual feature extraction, deep learning can learn abstract representations layer by layer from raw data such as images, texts, and sounds. For example, it can identify edges and textures from pixel values and then progress to higher - level object contours and categories. Its key characteristic is the deep architecture, which usually consists of an input layer, multiple hidden layers, and an output layer. The more hidden layers there are, the stronger the model's ability to capture complex patterns. Common models include convolutional neural networks (e.g., CNN), which are good at image processing; recurrent neural networks (e.g., RNN), which are suitable for sequential data; and Transformer, which is widely used in natural language processing. Deep learning has achieved breakthrough results in fields such as computer vision, speech recognition, and natural language processing. However, it requires a large amount of data and computing resources, has a high model complexity, and relatively weak interpretability. A neural network is a mathematical model inspired by the biological nervous system. It consists of a large number of artificial neuron nodes forming a network structure through connection weights. The basic unit is the neuron, which receives input and outputs a signal through an activation function such as Sigmoid or ReLU, simulating the "excitation" or "inhibition" state of biological neurons. The simplest is the single - layer neural network, which is used for linear classification, while the multi - layer neural network can handle non - linear problems. The network adjusts the connection weights through the back - propagation algorithm to minimize the prediction error and achieve learning of data patterns. Neural networks are the foundation of deep learning. In the early days, their development was slow due to limited computing power. With the improvement of computing power and algorithm optimization, they gradually evolved into deep neural networks, becoming the core tool for handling complex tasks. They have the characteristics of adaptive learning and parallel processing, but the structural design and parameter tuning have a significant impact on performance.

3 The Application of Machine Learning in Disease Diagnosis

3.1 Parkinson's

Machine learning has been widely applied in the diagnosis of Parkinson's disease, covering various aspects such as early disease diagnosis, condition assessment, and postoperative treatment optimization. In terms of early diagnosis and screening of the disease, machine learning plays an important role by analyzing information such as

language movements and patients' imaging pictures. It can identify subtle changes that are difficult for humans to detect, for example, using algorithms to detect changes in speech rate, intonation, and the degree of voice tremor when patients speak, which greatly improves the accuracy of early diagnosis of Parkinson's disease. The relevant work of researchers such as Qianrong Xie, Yue Chen, and Yimei Hu has also provided research support for this field, further promoting the application and development of machine learning in the early identification of Parkinson's disease. In disease monitoring, machine learning has realized wearability. With the help of sensors, it continuously collects patients' daily movement data, such as walking movements and the degree of body shaking, for real-time assessment, which is more convenient and timely compared with traditional assessment methods [5]. In treatment, machine learning helps formulate personalized treatment plans. Wearable sensors for deep brain stimulation, for instance, can extract more detailed information from the depth of the brain, gain an in-depth understanding of the patient's physical state, thereby optimizing treatment plans and improving efficiency. Similarly, in the study published by Almohaimeed, M. in 2025, Enhancing Prediction of Osteoporosis Using Supervised and Unsupervised Learning: New Approach to Disease Subtyping, the method of enhancing osteoporosis prediction and classifying disease subtypes using supervised and unsupervised learning, although targeting osteoporosis, the ideas embodied in it regarding machine learning in disease prediction and personalized analysis also provide useful references for the formulation of personalized plans in the treatment of Parkinson's disease, that is, to achieve more targeted medical intervention through accurate data analysis [6]. These applications not only improve the efficiency of diagnosis and treatment of Parkinson's disease but also play a positive role in promoting the development of the medical field.

3.2 Breast Cancer

Machine learning has become an important direction in breast cancer diagnosis. In early screening, it combines clinical data and images to build risk evaluation models for early intervention. Methods like random forests, which use regression to fill missing data after random loss, show strong performance—even with 50% data missing, diagnostic accuracy remains as high as 96.85%, ideal for early stages with limited data. In risk assessment, machine learning aids in automatic identification of cancer cell states through slice analysis, integrating multi-dimensional data like images for comprehensive evaluation. For postoperative treatment, models such as logistic regression and random forests achieve up to 81.8% accuracy, optimizing treatment plans and resource allocation, and providing molecular-level references for prevention, diagnosis, and prognosis. Similar to its applications in early detection of Parkinson's Disease using deep learning and machine learning [7], and in distinguishing deep brain stimulation parameter configurations for Parkinson's treatment via machine learning with wearable sensor data [8], machine learning demonstrates versatile value in medical diagnostics and intervention.

3.3 Osteoporosis

Now, the application of machine learning in the field of osteoporosis has become a broad consensus. Using machine learning to diagnose osteoporosis can significantly improve the accuracy of diagnosis, buying valuable time for patients to intervene in the disease. In the early detection and screening of the disease, bone mineral density is first measured. Machine learning can not only focus on traditional T-values but also detect other values, significantly enhancing the accuracy of osteoporosis detection. Meanwhile, it can integrate a large amount of patient data such as age, eating habits, and medical history for multi-dimensional consideration to ensure diagnostic accuracy.

In disease assessment, radiomics feature extraction methods are often used. With the help of image-assisted analysis, changes in bone microstructure are identified to determine the degree of osteoporosis and the risk of fractures, which is more efficient than doctors simply reading X-rays. Machine learning can also combine bone mineral density, bone quality, fracture risk, etc., to predict the probability of osteoporosis through models and provide specific treatment plans. In postoperative treatment after diagnosis, machine learning can track osteoporosis patients in real-time, timely obtain their real-time data, evaluate them, and then guide rehabilitation plans. Just like the random forest method for breast cancer tumor diagnosis and the application of machine learning in breast cancer survival analysis [9, 10], these applications of machine learning in osteoporosis diagnosis and treatment also show great value, promoting the progress of osteoporosis diagnosis and treatment technologies.

4 Challenges and Prospects

4.1 Challenges

In recent years, technological progress has been extremely rapid, which has promoted the development of machine learning. However, machine learning still faces very sig-

ISSN 2959-6157

nificant challenges in disease diagnosis.

Firstly, in terms of data quality, data quality is one of the important challenges that machine learning encounters in disease diagnosis. The accuracy, completeness, and consistency of data directly affect the performance and reliability of the model. Therefore, strict quality control and pre - processing of data are required. Secondly, in terms of algorithm selection, algorithm selection is another important challenge for machine learning in disease diagnosis. Different algorithms are suitable for different data types and problems, and need to be selected according to specific circumstances. In addition, the complexity and computational efficiency of the algorithm also need to be considered. Finally, there is a challenge in model interpretability. Model interpretability is an important issue for machine learning in disease diagnosis. Doctors and patients need to understand the decision - making process and basis of the model in order to better accept and use it. Therefore, the interpretability and transparency of the model need to be improved.

4.2 Prospects

In terms of data sharing and privacy protection, data sharing serves as an important foundation for the development of machine learning in the medical field, but data privacy protection is also crucial. In the future, it is necessary to find a balance between data sharing and privacy protection to promote the rational application of medical big data. In terms of technological innovation, in the future, machine learning technologies will continue to innovate and develop, such as deep learning and reinforcement learning. These technologies will provide stronger support for disease diagnosis and treatment, thereby improving the intelligent level of medical care. Finally, in terms of multidisciplinary integration, the application of machine learning in the medical field requires the integration of multiple disciplines, such as medicine, computer science, and mathematics. Through interdisciplinary cooperation, researchers can promote the in - depth application of machine learning in the medical field and make greater contributions to human health.

5 Conclusion

This article mainly introduces relevant algorithms and models of machine learning, such as logistic regression, support vector machines, decision trees, and random forests. It also summarizes and presents application cases of machine learning in disease diagnosis, enabling people to better understand the working principles of machine learning. In future development, it is hoped that a balance can be found between data sharing and privacy protection, so that people can protect their privacy while receiving treatment. At the same time, it is also hoped that machine learning can continue to develop and integrate with multiple disciplines, so as to continuously improve the intelligence and accuracy of machine diagnosis.

References

[1] Harshitha G, Kumar S. Rani S and Jain A. Cotton disease detection based on deep learning techniques, 4th Smart Cities Symposium (SCS 2021), 2021.

[2] Park D J, Park M W, Lee H. et al., Development of machine learning model for diagnostic disease prediction based on laboratory tests, Sci Rep, 2021.

[3] Patel P, Aw A N, Disease Diagnosis System Using Machine Learning, Journal of Pharmaceutical Research International, 2021.

[4] Poudel S. A Study of Disease Diagnosis Using Machine Learning, Med. Sci. Forum, 2022.

[5] Xie Q R, Chen Y, Hu Y M, Zeng F W, Wang P X, Xu L, Wu J H, Li J, Zhu J, Xiang M, Zeng F X et al, Using radiomic features of lumbar spine CT images to differentiate osteoporosis from normal bone density, <BMC Medical Imaging>, 2022.

[6] Almohaimeed M, Enhancing Prediction of Osteoporosis Using Supervised and Unsupervised Learning: New Approach to Disease Subtyping, Intelligent Information Management, 2025.

[7] Wang W, Lee J H, Harrou F Z, Sun Y, Early Detection of Parkinson's Disease Using Deep Learning and Machine Learning, IEEE Access, 2020.

[8] LeMoyne R, Mastroianni T, Whiting D. and Tomycz N, Distinction of an Assortment of Deep Brain Stimulation Parameter Configurations for Treating Parkinson's Disease Using Machine Learning with Quantification of Tremor Response through a Conformal Wearable and Wireless Inertial Sensor. Advances in Parkinson's Disease. 2020.

[9] Cai M Y, A Novel Method for Diagnosis of Breast Cancer Tumors Based on Random Forest, Journal of Biosciences and Medicines, 2023.

[10] Zhuang Z K, Decipher Clinical and Genetic Underpins of Breast Cancer Survival with Machine Learning Methods, Advances in Breast Cancer Research, 2023.