Prediction of Disease Progression of ALS based on Machine Learning

Shuzhe Zhang

Dalian, Houston International Institute, Dalian Maritime University, 116033, China
zsz2003113@dlmu.edu.cn

Abstract: Based on machine learning methods, this paper explores the potential of decision time and Multilayer Perceptron (MLP) in predicting trends in Amyotrophic Lateral Sclerosis(ALS), and introduces a prediction model based on decision trees and multi-layer perceptrons (MLPS). The paper first analyzes the pathological features of ALS as a neurodegenerative disease, its complex pathogenesis, and the data scarcity and treatment challenges facing the medical community. Then, it describes in detail the construction process of the proposed predictive model, including data source acquisition, feature engineering processing, model training, and evaluation methods, as well as the adoption of weighted processing method and the application of decision trees and multilayer perceptron (MLP) in prediction and diagnosis.The results show that there is a certain degree of error in the prediction of random number groups, but with the increase of sample size, the prediction accuracy is gradually improved. Further discussion highlighted new advances in current ALS research, such as genetic and environmental factors' influence and the application of neuroimaging techniques. Finally, the paper summarizes the findings of the study. It points out that as the accumulation of medical data increases, the accuracy of the predictive model will be further improved to provide more accurate support for the management and treatment of ALS. The significance of this study lies in providing a new approach to ALS prediction and a more precise tool for the medical community to better understand and respond to this serious disease.

Keywords: ALS, MLP, Decision-tree.

1 Introduction

Amyotrophic Lateral Sclerosis (ALS) , also known as motor neuron disease, is mainly a chronic progressive neurodegenerative disease. Because it mainly involves the motor system of the whole body, such as motor cells and motor conduction tracts, patients mainly have motor involvement, manifested as muscle weakness and atrophy. It usually starts from one limb, starting from the hand, the weakness of one hand, muscle atrophy, and then gradually progresses to the upper arm, and then gradually involves the lower limbs, so the patient gradually loses movement ability. However, the complex pathogenesis of ALS has not been fully elucidated. The pathogenesis of neurodegenerative diseases is usually very complex, involving multiple genetic,

Y. Wang (ed.), *Proceedings of the 2024 2nd International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2024)*, Advances in Computer Science Research 115,
https://doi.org/10.2991/978-94-6463-540-9_6

environmental and biochemical factors. At the same time, clinical brain specimens are scarce [1]. Such diseases are difficult to obtain as tumor specimens, and scholars have difficulty obtaining neurodegenerative brain tissue specimens. The second is the lack of effective treatment. At present, almost all neurodegenerative diseases have no method to alleviate the progression of the disease, and the treatment is mainly to alleviate the symptoms.

With the development of machine learning techniques, biomedical signals are used to process and predict neuromuscular diseases [2], from convolutional neural networks to detect brain damage [3], from finger tapping tests (FTT) and hand strength tests [4] to speech perturbation for ALS [5].

There is an opportunity to use these decision trees and MLPS to predict ALS trends. Machine learning can effectively process large amounts of complex data and discover potential patterns and associations, which can provide more accurate tools for predicting ALS. Some scholars have used convolutional neural network(CNN) to analyze images and help ALS patients use eye tracking to achieve interaction.

This paper explores the potential of machine learning methods to predict ALS trends and introduces a machine learning-based prediction model (MLP) that can analyze multiple latent factors to predict ALS incidence, prevalence, and other related factors [6]. First, this article describe the machine learning prediction model in detail, including data collection, feature engineering, model training, and evaluation. Finally, this article will show the predictive effects of the model and discuss its potential application in clinical practice.

The main contribution of this paper is to use machine learning techniques to provide a new approach to the prediction of ALS, providing the medical community with more precise and accurate tools for better management and treatment of this deadly disease. By combining the existing public data and symptoms of ALS patients, decision trees and vector machines can be used to evaluate better and predict the conditions of ALS patients with different symptoms.

2 Data and methods

2.1 Data Sources

A series of motor neuron diseases represented by ALS do not have a very large patient population, so the prediction of ALS mainly focuses on the refined classification and processing of a small amount of data. This paper uses data from the national ALS registered website (https://www.als.org/about-us) for public data. Based on these data, this paper also chooses to call some existing articles in medicine, computer science or statistics-related fields related to the content of the data call.

The medical repository's data providers include three large national administrative databases from the United States, the Centers for Medicare and Medicaid Services, the Veterans Health Administration, and the Center for Veterans Benefits. In addition, secure public portals are also among the data providers that allow ALS patients to

register in the registry to identify other cases not recorded in the national administrative database.

The data content selected in this paper is the number of ALS patients in the sample. For different ALS patients, there are certain differences between various data under the same variable, but this does not accurately reflect the data differences between relevant variables. Therefore, the preprocessing method of weighting the data is used in this paper so that differences are observed in the data of the same group.

2.2 Methods

Decision trees In the process of predicting the prevalence of ALS in individuals, this paper uses the differences in race, gender and age to predict the prevalence of ALS. The data used in this paper are evaluated from three perspectives, namely gender, race and age. Similarly, the information given by each group corresponds to different prevalence rates. So this paper uses a random array with three parameters to represent different classifications. The data use 1 for male and 0 for female in the first parameter. This article use 0 for white and 1 for yellow in the second parameter. In the third parameter, 0 means less than 55,1 means between 56 and 65, and 2 means more than 65. This is done by giving the prevalence rates represented by different characteristics and generating random arrays to represent different people, forming an array with three variables. When this paper uses other randomly generated arrays to analyze the data, other processing methods and judgments can be generated for the whole diagnosis process by predicting the results of the existing data.

MLP. MLP is also used in this paper to predict ALS patients. For MLP, the way of processing data is different from that of decision tree. MLP is a neural network-based model consisting of a multi-layer network structure composed of multiple neurons. Each neuron receives the input data and performs a nonlinear transformation through the activation function, and then passes the output to the next layer of neurons. The MLP optimizes the model parameters through the backward propagation algorithm to minimize the loss function, thus realizing the classification or regression of the data. The first step is data preparation. First, this paper randomly generated the gender, race, and age information of 10 people with the given prevalence and stored them in a DataFrame. The categorical variables are then converted to numbers using LabelEncoder.

Next, the data preparation is conducted. The dataset consists of feature data (gender, race, and age) and label data (0 or 1 indicating the presence or absence of the disease). The train_test_split function is employed to partition the dataset into training and test sets, with the test set comprising 20% of the total data.The MLP model is then built, creating an instance of the MLPClassifier, which is a hidden layer collector classifier that contains one layer by default, and then training the MLP model with the training set.

The prediction is made on the test set, and the trained model is used to predict the test set to obtain the prediction result. Then, the prevalence was calculated. In this paper,

two functions are defined to calculate the prevalence rate for each individual and convert the predicted results to the actual prevalence rate.

At the same time, referring to the idea of L.N.Mumry for the prediction and classification of neuromuscular diseases [7], this paper compares and analyzes the results of the two prediction models.

3 Result analysis

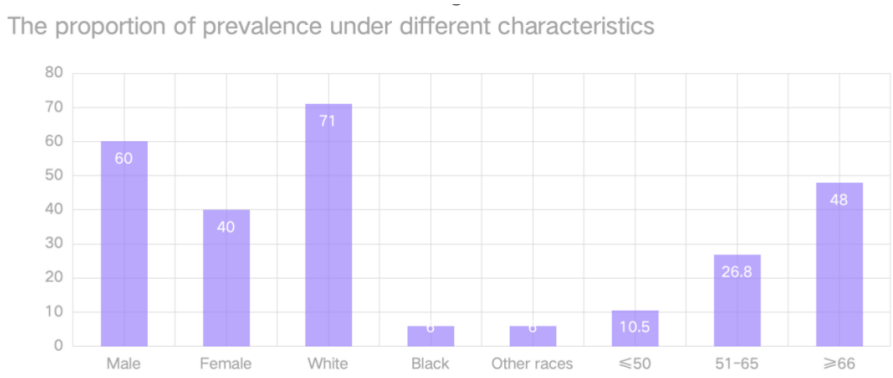

Fig.1 The proportion of prevalence under different characteristics

Output result prediction and prevalence rate: According to the prediction result and prevalence rate, output the prediction result of each individual and the corresponding disease condition. According to Fig. 1, the data under different characteristics have different prevalence rates. In this paper, the product of the prevalence rates corresponding to three characteristics is more than 0.5%, which is considered as possible disease

This part of the decision is the same as the decision tree. After making 20 MLP and decision tree predictions of random arrays, the two predictions exist 7 times and the results differ, with an error of 35%. The predictions of the MLP were less accurate than those of the decision tree, and the decision tree produced better accurate results than the MLP if only the prediction was based on the criteria of prevalence greater than 1%.

However, by changing some of the parameters, such as defining prevalence as more than 0.5%, the error of the MLP algorithm is 25%, which is lower than before. This may be because the results are assumed to be binary variables in this study, so when the proportion of one of the possible results becomes larger, it may impact the predicted results.

When the random array size is increased, the prediction error for the decision tree and MLP is 15% if the sample size is 20 and the prevalence is greater than 0.5% (Fig.2). This paper subsequently tested a sample size of 30, with an error of 0.5% for the case where the prevalence was defined as being diseased if it exceeded 0.5%. All these indicate that the final result is relatively more accurate when the sample covers a wider range.

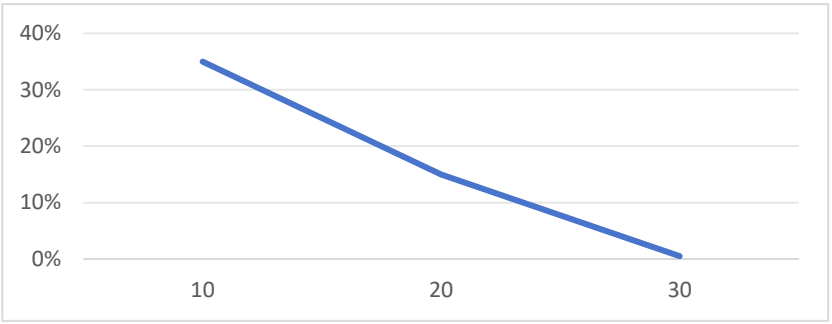

Fig. 2 MLP vs. decision tree result error (number of predictions)

4 Discussion

Through the understanding of some cutting-edge academics, the current research on ALS is referring to more and more factors. First, genetics: about 10% of ALS patients are genetically affected, due to a mutation in the TAR DNA-binding protein 43 (TARDBP) gene that causes motor neuron atrophy. If a family member has ALS patients, the risk of ALS is increased by 1%, and the most common age is 40-50 years old, most of them are men. The second is heavy metal poisoning; About 5% of ALS patients are poisoned by heavy metals, such as cadmium, lead and mercury, which affect the immune system. According to literature, the number of ALS patients in Guam's indigenous people is too large. Studies have found that the local traditional food contains the heavy metal toxin BMMA, which greatly reduces the incidence of ALS in this area after removing the cause of disease. And then some unexplained reasons:

85% of the patients are caused by unknown causes of motor neuron atrophy, which is still being studied by the academic community. Meanwhile, in the diagnostic research of ALS, the study by K.Tsarapatsani et al. also retrieved a novel approach focusing on biomarker selection, including three different gene reduction procedures. In this paper, the dataset was retrieved and differential expression analysis was performed to identify significantly different gene markers. Then, the data use filters and embedded methods to perform gene selection, maximum relevance minimum redundancy (MRMR), followed by Least Absolute shrinkage and Selection Operator (LASSO) regression. The various machine learning algorithms corresponding to a set of gene combinations are then estimated using a series of cross-validation procedures.) At the same time, there are some studies in neurology, by identifying the nerve CT of some patients, so as to get some reference data. These can increase the richness of the data to determine whether a patient is an ALS patient. For the data used in this paper, the data content of the experiment can be enriched in the future, from three data dimensions, so as to make the prediction more accurate when facing more complex information. These related studies also support the results of this paper. Can prove that with the information dimension. It also means that more data is needed to support the

training of MLP and decision tree, so that the prediction error is smaller. However, too many data types may not match the sample size [8-10].

5 Conclusion

This paper investigates methods for predicting the probability of ALS by analyzing publicly available data provided by the National ALS Registry website. Because of the background of ALS, the prevalence of ALS was analyzed from the perspective of race, gender and age. It was found that prediction models such as MLP and decision tree are highly influenced by data in prevalence prediction. With the continuous accumulation of data from medical institutions, the accuracy of prediction models will gradually improve, because more data can provide more dimensional information to better process and classify data. If insufficient information exists, short-term prediction can only be used to increase the sample size.

The study of this paper hopes that when some patients are examined for ALS, the probability of disease can be predicted through the characteristics of the patients themselves, and more appropriate disease screening methods and treatment methods can be selected to improve the confirmation and treatment of ALS.

References

1. Halbersberg D. and Lerner B.:Temporal Modeling of Deterioration Patterns and Clustering for Disease Prediction of ALS Patients, in Proceedings of the 2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA), Boca Raton, FL, USA, pp. 62-68,(2019).
2. Sayeed Ud Doulah A. B. M., Iqbal M. A. and Jumana M. A.:ALS disease detection in EMG using time-frequency method, in Proceedings of the 2012 International Conference on Informatics, Electronics & Vision (ICIEV), Dhaka, Bangladesh, pp. 648-651,(2012).
3. Hameed M. and Inan T.:Motor Imagery EEG Classification Using Algorithms and Machine Learning for ALS Disease, in Proceedings of the 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), Ankara, Turkey, pp. 1-6,(2022).
4. Bustamante P., Grandez K., Solas G. and Arrizabalaga S.:A low-cost platform for testing activities in Parkinson and ALS patients, in Proceedings of the 12th IEEE International Conference on e-Health Networking, Applications and Services, Lyon, France, 302-307,(2010).
5. Vashkevich A., Petrovsky A. and Rushkevich Y.:Bulbar ALS Detection Based on Analysis of Voice Perturbation and Vibrato, in Proceedings of the 2019 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA), Poznan, Poland, pp. 267-272, (2019).
6. Bai X., Chen B., Gao X. and Li J.:Correlation Between Diabetes and Body Composition of Based on Decision Tree and Neural Network, in Proceedings of the 2019 Chinese Control And Decision Conference (CCDC), Nanchang, China, 4992-4997,(2019).
7. Murthy G. L. N., Saii G. Phawahan, Pavani T. and Mohan J. Lalith:A machine learning based frame work for classification of neuromuscular disorders, in Proceedings of the 2022 IEEE 1st International Conference on Data, Decision and Systems (ICDDS), Bangalore,

India, 01-05, (2020).
8. Ko K. D., El-Ghazawi T., Kim D. and Morizono H.:Predicting the severity of motor neuron disease progression using electronic health record data with a cloud computing Big Data approach, in Proceedings of the 2014 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, Honolulu, HI, USA, 1-6., (2014)
9. Hadad B. and Lerner B.:Domain adaptation from clinical trials data to the tertiary care clinic – Application to ALS, in Proceedings of the 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA), Miami, FL, USA, 539-544, (2020).
10. Hameed M. and Inan T.:Motor Imagery EEG Classification Using Algorithms and Machine Learning for ALS Disease, 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), Ankara, Turkey, pp. 1-6, (2022).