· Intelligent Ophthalmology ·

Automatic diagnosis of extraocular muscle palsy based on machine learning and diplopia images

Xiao–Lu Jin, Xue–Mei Li, Tie–Juan Liu, Ling–Yun Zhou

Ocular Motility Disorder Treatment Centre, the First Affiliated Hospital of Harbin Medical University, Harbin 150001, Heilongjiang Province, China

Correspondence to: Ling-Yun Zhou. Ocular Motility Disorder Treatment Centre, the First Affiliated Hospital of Harbin Medical University, 199 Dongdazhi Street, Nangang District, Harbin 150001, Heilongjiang Province, China. no1zhly@163.com

Received: 2024-04-22 Accepted: 2025-01-25

Abstract

● **AIM:** To develop different machine learning models to train and test diplopia images and data generated by the computerized diplopia test.

● **METHODS:** Diplopia images and data generated by computerized diplopia tests, along with patient medical records, were retrospectively collected from 3244 cases. Diagnostic models were constructed using logistic regression (LR), decision tree (DT), support vector machine (SVM), extreme gradient boosting (XGBoost), and deep learning (DL) algorithms. A total of 2757 diplopia images were randomly selected as training data, while the test dataset contained 487 diplopia images. The optimal diagnostic model was evaluated using test set accuracy, confusion matrix, and precision-recall curve (P-R curve).

● **RESULTS:** The test set accuracy of the LR, SVM, DT, XGBoost, DL (64 categories), and DL (6 binary classifications) algorithms was 0.762, 0.811, 0.818, 0.812, 0.858 and 0.858, respectively. The accuracy in the training set was 0.785, 0.815, 0.998, 0.965, 0.968, and 0.967, respectively. The weighted precision of LR, SVM, DT, XGBoost, DL (64 categories), and DL (6 binary classifications) algorithms was 0.74, 0.77, 0.83, 0.80, 0.85, and 0.85, respectively; weighted recall was 0.76, 0.81, 0.82, 0.81, 0.86, and 0.86, respectively; weighted F1 score was 0.74, 0.79, 0.82, 0.80, 0.85, and 0.85, respectively.

● **CONCLUSION:** In this study, the 7 machine learning algorithms all achieve automatic diagnosis of extraocular muscle palsy. The DL (64 categories) and DL (6 binary classifications) algorithms have a significant advantage over other machine learning algorithms regarding diagnostic accuracy on the test set, with a high level of consistency with clinical diagnoses made by physicians. Therefore, it can be used as a reference for diagnosis.

● **KEYWORDS:** machine learning; extraocular muscle paralysis; automatic diagnosis; diplopia images

DOI:10.18240/ijo.2025.05.01

Citation: Jin XL, Li XM, Liu TJ, Zhou LY. Automatic diagnosis of extraocular muscle palsy based on machine learning and diplopia images. *Int J Ophthalmol* 2025;18(5):757-764

INTRODUCTION

Extraocular muscle palsy is caused by vascular diseases, cranial nerve injuries, infections, and other factors affecting the ocular motor nerve system or extraocular muscles themselves, resulting in complete or partial dysfunction of ≥1 extraocular muscle. This can lead to diplopia, restricted eye movement, and strabismus[1-3]. Hess screen test is a common clinical examination method used to distinguish and diagnose extraocular muscle palsy by breaking binocular visual fusion[4-5]. The computerized diplopia test, a computer-automated detection device based on the traditional Hess screen principle, showed accurate and reliable results in the clinic[6-7]. The diplopia image is the plot generated by the computerized diplopia test, reflecting the image perceived by the patient during the gaze test[6] (Figure 1). However, the shapes of diplopia images can be complex and vary depending on which extraocular muscles are paralyzed. Therefore, a professionally trained doctor need to manually interpret diplopia images to diagnose the paralyzed extraocular muscles. However, this requirement for specialized expertise and manual interpretation is not conducive to the widespread adoption of this method in clinical practice.

In recent years, artificial intelligence has been widely used in ophthalmic diagnosis[8-9]. Deep learning (DL) models achieved automatic diagnosis of diseases such as glaucoma[10-11], cataracts[12], strabismus[13-14], ptosis[15], and acanthamoeba keratitis[16], including automatic classification of glaucoma severity[11].

A previous study[17] utilized a support vector machine (SVM) to construct a model for diagnosing specific extraocular muscle palsy. A total of 229 patients underwent testing and diagnosis through this system, with 156 patients achieving consistency between the diagnostic results based on the SVM and clinical diagnoses established by physicians, resulting in an accuracy rate of 68.12%. However, that study had a small sample size and could only predict single extraocular muscle palsy with the highest probability and a relatively low accuracy rate. Currently, we have established a large sample diplopia image database, which provides a data basis for machine learning, after accumulating clinical information about extraocular muscle palsy patients for a long time. These factors have led us to consider whether we can use the existing large amount of data to train machine learning models to establish the accurate, efficient, and simple-to-operate computer-aided diagnosis of paralyzed extraocular muscles.

This study designed, developed, and evaluated multiple machine learning algorithms for diagnosing extraocular muscle paralysis using digital electronic strabismus detection. Through this method design, all algorithms could simultaneously predict multiple paralyzed muscles at once. Then, the most suitable machine learning algorithm model was selected for application in clinical diagnosis.

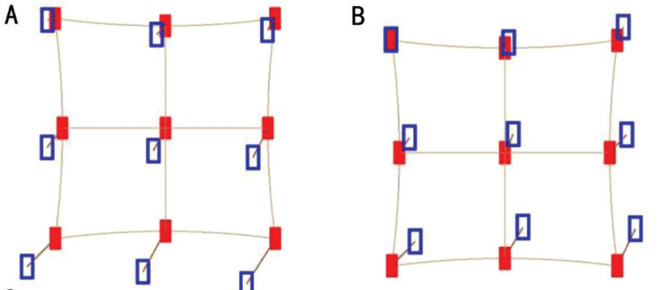
PARTICIPANTS AND METHODS

Ethical Approval This study was approved by the Ethics Committee of the First Affiliated Hospital of Harbin Medical University (No.2024JS97) and followed the principles outlined in the Declaration of Helsinki.

Data Collection Diplopia images and data generated by computerized diplopia tests, along with patient medical records, were retrospectively collected from 3244 cases. The inclusion criteria for patients in this study were 1) patients diagnosed with extraocular muscle palsy; 2) healthy volunteers; and 3) those who have completed binocular diplopia imaging tests and have complete images. The exclusion criteria comprised patients with restrictive strabismus (including cases of orbital wall fractures, hematomas, or thyroid-related orbitopathy), glaucoma, color blindness, high myopia, or diplopia occurring during monocular fixation.

Data Preprocessing Recoding and preprocessing of variable data were performed since these are an important step in machine learning model training. Assigning values to data sequentially allows the model to train more easily, significantly reducing the runtime of machine learning and improving the efficiency of model evaluation.

Processing of raw diplopia image data Taking the right-eye image of the original strabismus image (Figure 1) as an example, a coordinate system was established with the central red dot as the origin, the horizontal direction as the X-axis, and the vertical direction as the Y-axis. Based on the strabismus image detection standard test angle of 20°[6] and the deviation angle in the table, the coordinates of the 9 red and 9 blue dots corresponding to the left upper, upper, right upper, left, straight ahead, right, left lower, lower, and right lower positions could be obtained (Figure 2).



C

Direction		1	2	3	4	5	6	7	8	9
Eye		Upper left	Upper	Upper right	Left	Direct vision	Right	Lower left	Lower	Lower right
Right eye	α	-1	-2	-1	-3	-2	-3	-4	-4	-6
	β	0	-2	-2	-4	-5	-6	-4	-7	-8
Left eye	α	0	1	1	2	2	3	4	4	4
	β	0	1	2	3	4	3	4	6	7

Figure 1 Diplopia image A: Right-eye figure; B: Left-eye figure; C: The horizontal and vertical deviation angles of both eyes at 9 gaze points.

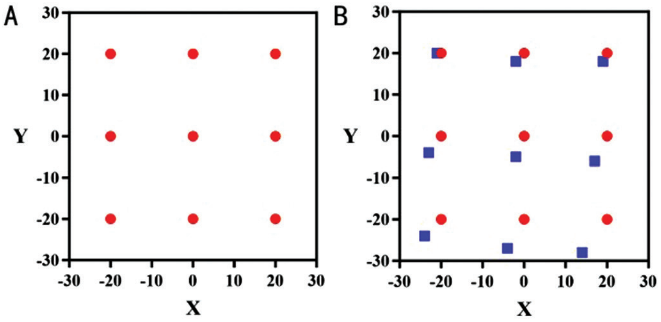


Figure 2 Processing of strabismus image data A: Based on the strabismus image detection standard test angle of 20°, the coordinates of the 9 red dots corresponding to the left upper, upper, right upper, left, straight ahead, right, left lower, lower, and right lower positions are (-20, 20), (0, 20), (20, 20), (-20, 0), (0, 0), (20, 0), (-20, -20), (0, -20), and (20, -20), respectively; B: Based on the deviation angle in the table, the coordinates of the 9 blue dots corresponding to the left upper, upper, right upper, left, straight ahead, right, left lower, lower, and right lower positions are (-21, 20), (-2, 18), (19, 18), (-23, -4), (-2, -5), (17, -6), (-24, -24), (-4, -27), and (14, -28), respectively.

Strabismus image diagnosis results and processing Two physicians with at least 5y of professional experience performed diagnosis and annotation of the diplopia images in a double-blind manner. If their diagnostic results were consistent, they were considered the standard result. If their diagnostic results were inconsistent, a third, more senior expert would diagnose the inconsistent diplopia images. All experimental data were anonymized before the study. After diagnosis and verification by three professional doctors, normal extraocular muscles (including synergistic muscles with overaction and competitor muscles with underaction) were labeled as "0", while abnormal extraocular muscles (including partially or

Table 1 Processing of diagnosis results

Name	Superior rectus	Inferior rectus	Medial rectus	Inferior oblique	Superior oblique	Lateral rectus
Patient 1	0	0	0	0	1	0

0: Normal; 1: Abnormal.

completely paralyzed extraocular muscles) were labeled as “1”. Using Figure 1 as an example, the diagnostic result was right superior oblique muscle paresis, which could be marked as shown in Table 1. Following this method, all diplopia images meeting the inclusion criteria would be processed sequentially for both eyes, and the data would be entered into an Excel spreadsheet to create a database.

When using machine learning to model and analyze eye muscle palsy, we encountered a problem: previous research could only determine which muscle was more likely to be damaged but could not simultaneously judge the palsy of multiple eye muscle. Although DL algorithms can solve the problem of simultaneously predicting 6 muscles by designing a fully connected layer output, followed by a rectified linear unit activation function, and using a loss function that can handle binary cross-entropy, such as binary cross-entropy with logits loss, other machine learning algorithms cannot be directly applied. To solve this problem, we arranged the output results of the labels according to the order of superior rectus, inferior rectus, medial rectus, inferior oblique, superior oblique, and lateral rectus. Furthermore, the output result was represented as a binary code, such as 000010 (Table 1). This binary code can be interpreted as paralysis of the superior oblique muscle and can be translated into decimal form as 2. By using the numbers 0–63, it is possible to uniquely represent which eye muscles are paralyzed. The model only needs to perform 64-class classification on the input data. This method is also applicable to other machine learning classification models.

The coordinate values of the 9 standardized blue points from all included diplopia images are used as input data for the machine learning model, with the corresponding diagnostic results serving as the output labels. The model extracts feature patterns from the input data to establish the relationship between the input data and the labels. After training, the model can generate predicted labels based on the input data from the test set.

Data Set Partitioning The data set of strabismus images for each diagnostic type was randomly divided into training and testing sets in an 85:15 ratio. The training data set was used to train the models, whereas the testing data set was used to evaluate the performance of the models.

Machine Learning and Algorithm Application Logistic regression (LR), also known as logistic odds regression, is a generalized linear regression analysis model commonly used for binary classification. It can be used for multi-classification by replacing the nonlinear mapping function

$$Sigmoid(x) = \frac{1}{1 + e^{-x}}$$

with the

$$Softmax(x) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}$$

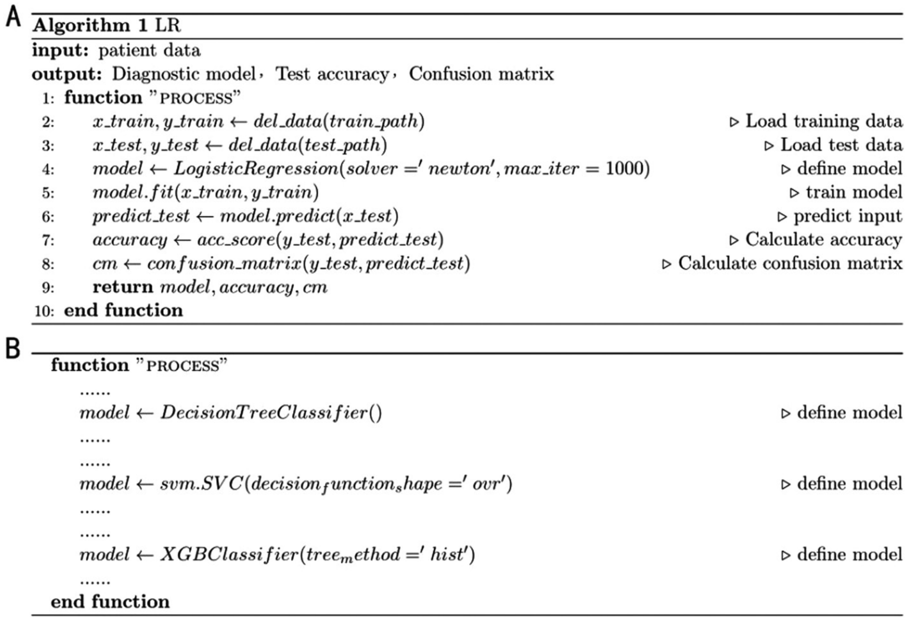
function. This study used Python (3.10.8) as the programming language, applying data processing libraries, such as Numpy (1.22.3) and Panda (1.3.5), to preprocess the coordinate data of 9 points. Additional, the labels were processed into decimal numbers ranging from 0 to 63. The LR function of scikit-learn (1.2.1) was used to perform LR, and the Newton method iterative algorithm was applied for training. Then, the trained model was used to diagnose and test the test data. Figure 3A shows the detailed processing flow in the pseudocode. Decision trees (DT), SVM, extreme gradient boosting (XGBoost) and other traditional machine learning algorithms were applied as described above, with corresponding modifications made to data processing and algorithm function implementation (Figure 3B).

In DL applications, the PyTorch (1.13.1) DL training framework was used. A fully connected neural network with 3 hidden layers was constructed based on the input data of 9 points (18 dimensions), with a 64-class classifier and 6 binary classifiers (Figure 4A) in the prediction head. The batch size for each training session was 128, and the SGD optimizer was used for gradient optimization. Furthermore, the 9-point coordinates were formed into images similar to those in Figure 2B, and convolutional neural networks were used to classify the generated images (Figure 4B). The above-mentioned algorithms were trained on a system with an Intel Xeon 8-core processor, a Tesla T4 GPU, and 64GB of memory running the Ubuntu 18.04 operating system.

Model Evaluation The classification performance of different models on the test set was evaluated by calculating accuracy, weighted precision, weighted recall, and weighted F1-score, as well as by plotting the confusion matrix and the P-R curve.

RESULTS

Data Collection and Artificial Diagnosis Results of Fundus Images A total of 3244 data entries were included in this study, encompassing 16 types of manually annotated results. Table 2 shows the baseline characteristics of patients in the training and test sets. Mean age, gender, or diagnostic distribution showed no significant differences between the two groups.



A

```
Algorithm 1 LR
input: patient data
output: Diagnostic model, Test accuracy, Confusion matrix
function "PROCESS"
    $x\_train, y\_train \leftarrow del\_data(train\_path)$                    ▷ Load training data
    $x\_test, y\_test \leftarrow del\_data(test\_path)$                       ▷ Load test data
    $model \leftarrow LogisticRegression(solver =' newton', max\_iter = 1000)$ ▷ define model
    $model.fit(x\_train, y\_train)$                                        ▷ train model
    $predict\_test \leftarrow model.predict(x\_test)$                        ▷ predict input
    $accuracy \leftarrow acc\_score(y\_test, predict\_test)$                  ▷ Calculate accuracy
    $cm \leftarrow confusion\_matrix(y\_test, predict\_test)$                 ▷ Calculate confusion matrix
    return $model, accuracy, cm$
end function
```

B

```
function "PROCESS"
    ......
    $model \leftarrow DecisionTreeClassifier()$                            ▷ define model
    ......
    ......
    $model \leftarrow svm.SVC(decision_function_shape =' ovr')$            ▷ define model
    ......
    ......
    $model \leftarrow XGBClassifier(tree_method =' hist')$                 ▷ define model
    ......
end function
```

Figure 3 Processing flow A: Pseudocode for logistic regression; B: Pseudocode for decision tree, support vector machine and extreme gradient boosting algorithms.

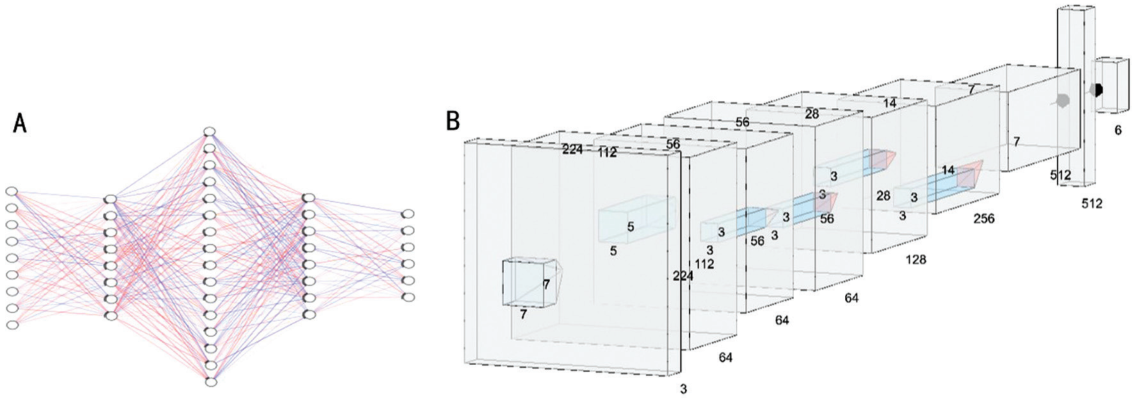


Figure 4 Structure diagram of fully connected and convolutional neural networks A: Fully connected neural network, with input as 9-point coordinates and output as 6 binary classifiers; B: Convolutional neural network, with input as a 224×224 image and output as 6 binary classifiers.

Model Evaluation Results Table 3 demonstrates the accuracies of training and test sets for each machine learning model. The LR model demonstrated relatively low accuracy on both training and test sets. The DT and XGBoost models exhibited high accuracy on the training set but performed poorly on the test set, indicating significant overfitting. The DL models for 64-class classification and binary classification achieved high accuracy on both training and test sets, with minimal differences in the diagnostic performance between the two. Additionally, the convolutional neural network (CNN) model showed very low accuracy on both datasets, probably due to the difficulty in learning effective diagnostic rules from image point information compared to explicit coordinate point information.

A confusion matrix is a standard format for evaluating the accuracy of a model. Accordingly, Figure 5 shows the confusion matrices of 6 machine learning models on the test set. According to the method of diagnosing multiple eye extrinsic muscle disorders simultaneously, the labels of 974 test set samples contained 14 categories. Different machine learning models output different numbers of categories; hence, the number of categories in the confusion matrices of different machine learning algorithms was not consistent.

Weighted precision, weighted recall, and weighted F1 score are indicators evaluating the quality of a machine learning model. Table 4 shows the average precision (AP), recall, and F1 scores of the 6 machine learning models on the test set, with the DL-based algorithm having the optimal technical indicators.

Based on the confusion matrices of the test set for each model, we could calculate the precision and recall and plot the P-R curve (Figure 6). The area under the P-R curve is called AP. Generally, the better the classifier, the higher the value of the AP. According to the diagnostic accuracy of the test set and the AP value in the P-R curve, the DL-based model (64 categories) had the best diagnostic effect.

Table 2 Statistics and baseline characteristics of patients in training and test sets

Parameters	Training set	Test set	*P*
Image quantity (pieces)	2757	487	-
Average age, y (mean±SD)	55.39±0.29	55.60±0.57	0.362[a]
Number of males (%)	61.50	58.10	0.160[b]
Diagnostic distribution, % (*n*)			0.018[b]
Normal	49.24 (2702)	48.46 (472)	
Lateral rectus muscle palsy	23.76 (1304)	24.03 (234)	
Oculomotor nerve palsy	11.28 (619)	9.14 (89)	
Superior oblique muscle palsy	6.43 (353)	9.04 (88)	
Inferior rectus muscle palsy	2.75 (151)	1.95 (19)	
Superior rectus muscle palsy	2.15 (118)	2.16 (21)	
Medial rectus muscle palsy	2.06 (113)	2.46 (24)	
Oculomotor nerve and lateral rectus muscle palsy	1.29 (71)	1.64 (16)	
Medial rectus and lateral rectus muscle palsy	0.38 (21)	0.41 (4)	
Superior rectus and lateral rectus muscle palsy	0.27 (15)	0.31 (3)	
Lateral rectus and inferior rectus muscle palsy	0.18 (10)	0.10 (1)	
Inferior oblique muscle palsy	0.11 (6)	0	
Superior rectus and inferior rectus muscle palsy	0.06 (3)	0.10 (1)	
Superior rectus, inferior rectus, and lateral rectus muscle palsy	0.04 (2)	0	
Superior oblique and lateral rectus muscle palsy	0	0.10 (1)	
Inferior rectus and medial rectus muscle palsy	0	0.10 (1)	

[a]Mann-Whitney *U* test; [b]Chi-square test.

Table 3 Accuracy of different models

Algorithm types	Training parameters	Training set accuracy	Test set accuracy
LR	max_iter=1000	0.785	0.762
SVM	-	0.815	0.811
DT	-	0.998[a]	0.818
XGBoost	-	0.965	0.812
DL (64 categories)	epoch=1000	0.968	0.858[b]
DL (6 binary classifications)	epoch=1000	0.967	0.858[b]
DL (CNN)	epoch=1000	0.590	0.588

[a]Optimal metric of training set; [b]Optimal metric of test set. LR: Logistic regression; SVM: Support vector machine; DT: Decision tree; XGBoost: Extreme gradient boosting; DL: Deep learning; CNN: Convolutional neural network.

Table 4 Average precision, recall, and F1 scores of different models on the test set

Model	Weighted average		
	Precision	Recall	F1 scores
LR	0.74	0.76	0.74
SVM	0.77	0.81	0.79
DT	0.83	0.82	0.82
XGBoost	0.80	0.81	0.80
DL (64 categories)	0.85[a]	0.86[b]	0.85[c]
DL (6 binary classifications)	0.85[a]	0.86[b]	0.85[c]

[a]Optimal metric of weighted precision; [b]Optimal metric of weighted recall; [c]Optimal metric of weighted F1 scores. LR: Logistic regression; SVM: Support vector machine; DT: Decision tree; XGBoost: Extreme gradient boosting; DL: Deep learning.

DISCUSSION

This study collected and organized data on diplopia images, deviation angle, and corresponding manual diagnosis. Different machine learning algorithm types were applied to analyze the data, achieving simultaneous diagnosis of multiple paralyzed eye muscles. Comparing the performance of these different algorithms on training and testing sets, we found that the DL algorithm based on a full CNN (including 64-class and 6-class binary classification) had the best performance and achieved high accuracy (with a testing set accuracy of 0.858 for both). Confusion matrices and P-R curves indicated that the DL diagnostic model had the best stability and consistency in diagnosing eye muscle paralysis.

Artificial intelligence, particularly algorithms represented by CNNs in target detection[18], image segmentation[19], and facial recognition[20], has been applied in the diagnosis of extraocular muscle-related diseases. Some studies used CNNs to segment extraocular muscles and measure muscle

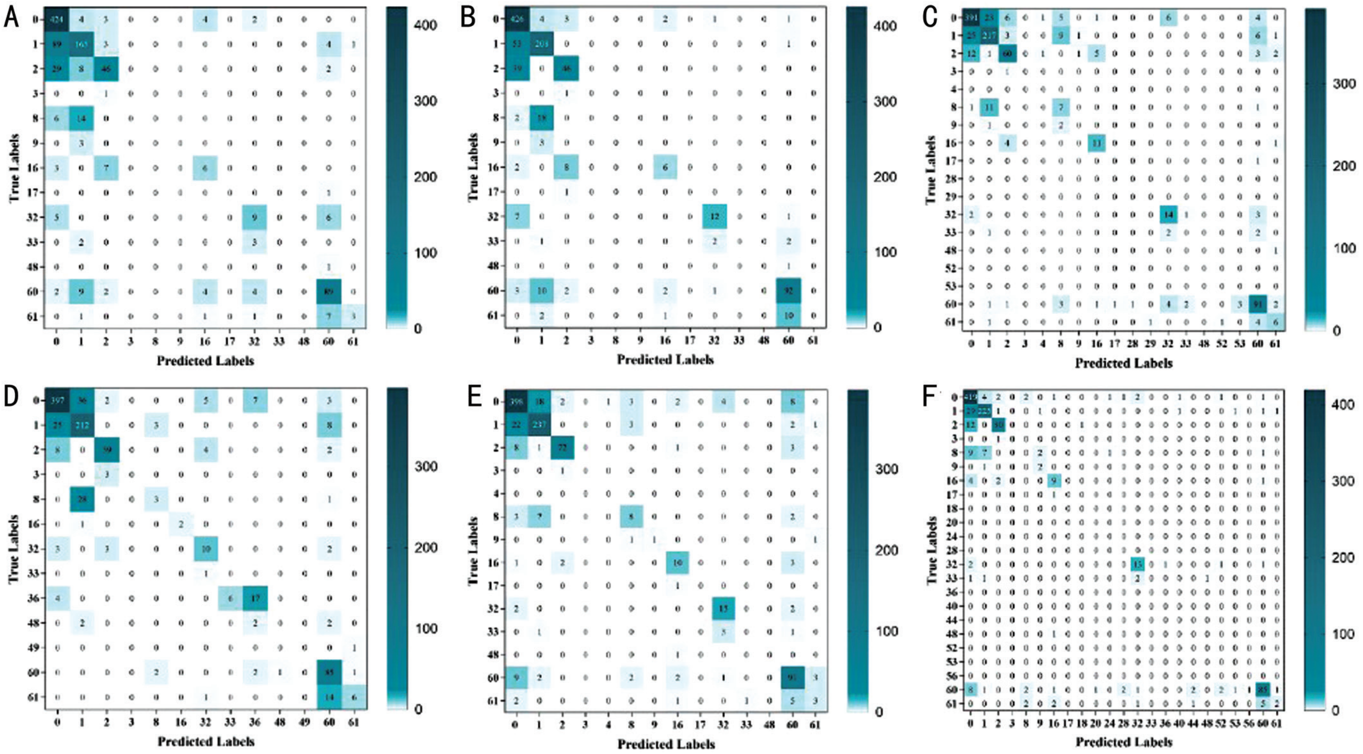


Figure 5 Confusion matrices The horizontal axis represents the classification labels predicted by the model, and the vertical axis represents the true classification labels. The values in the matrix represent the number of samples in which the model classified a specific true label as a certain predicted label, and the values on the diagonal represent the number of correctly classified samples. A: Logistic regression confusion matrix; B: Support vector machine confusion matrix; C: Decision tree confusion matrix; D: Extreme gradient boosting confusion matrix; E: Deep learning (64 categories) confusion matrix; F: Deep learning (6 binary classifications) confusion matrix.

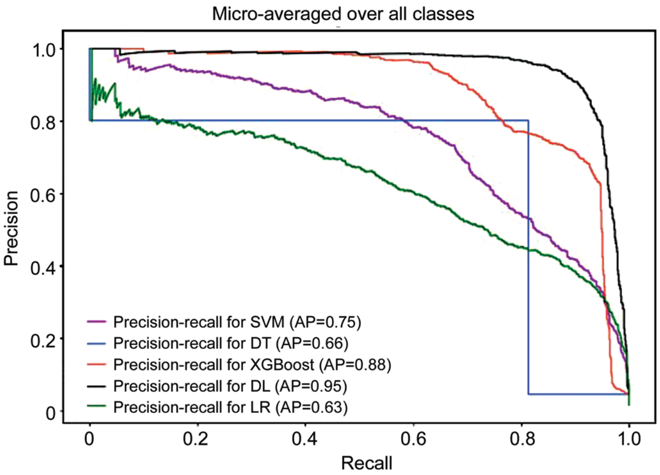


Figure 6 Precision-recall curve of 5 machine learning algorithms LR: Logistic regression; SVM: Support vector machine; DT: Decision tree; XGBoost: Extreme gradient boosting; DL: Deep learning; AP: Average precision.

size from CT images[21]. Another study employed 4 different DL frameworks to segment extraocular muscles in magnetic resonance images, achieving high accuracy[22]. A research team utilized CNN algorithms to automatically locate and segment ocular regions to evaluate the extent of inferior oblique muscle overaction[23] and measure the range of motion of 6 extraocular muscles, demonstrating a high degree of consistency with manual measurements[24]. These studies still required manual measurements after achieving image localization and segmentation, whereas our study achieved fully automated diagnosis. One study designed an algorithm based on fuzzy logic that automatically diagnoses paralytic strabismus using statistical data from Hess charts[4]. However, that study only covered 14 disease categories and could only identify the extraocular muscle with primary dysfunction. In our previous research, we used an SVM algorithm to build a diagnostic model for extraocular muscle paralysis[17], which similarly could only predict the most likely paralyzed extraocular muscle. However, simultaneous paralysis of multiple eye muscles is common in clinical diagnosis and treatment; thus, achieving simultaneous diagnosis of multiple paralyzed eye muscles is key to improving diagnostic accuracy and efficiency. In this study, different machine learning algorithms were used for 64-class classification, with output categories 0-63 corresponding to binary numbers 000000-111111. Furthermore, a fixed order of 6 eye muscles was used to achieve a simultaneous diagnosis of multiple paralyzed eye muscles.

We trained and tested the data for each eye separately, rather than using binocular data, for two main reasons. First, if we had predicted all 12 eye muscles in both eyes simultaneously, the model would have required 2^{12}=4096 output categories, which would be difficult to converge and significantly reduce accuracy given the current sample size. Second, we preprocessed the deviation angles for the left and right eyes, converting them into standard data and eliminating any differences, without affecting the final diagnosis.

Compared to previous studies, the machine learning model in this experiment achieved automatic diagnosis of extraocular muscle paralysis by learning complex patterns that are difficult for humans to detect. This reduced human subjectivity-related diagnostic errors, improved diagnostic efficiency, and achieved a certain level of accuracy. However, some limitations remain. First, the diagnostic accuracy of the best-performing DL model on the test set was only 85.8% with limited improvement. We speculate that errors in the manual diagnosis of strabismus images may be a primary reason as these errors could influence model optimization during training and decrease accuracy due to incorrect labels in the test set. Additionally, the lack of sufficient strabismus image data may impact the generalization performance of the model. Second, restrictive and paralytic cases exhibit the same patterns in diplopia images. However, this study focused solely on the diagnosis of paralytic extraocular muscles, lacking the capability to diagnose restrictive strabismus cases, which needs to be addressed in future research. Third, as a retrospective study, this research faced limitations such as incomplete patient medical records and inconsistencies in the conditions of computerized diplopia examinations (*e.g.*, lighting levels, refractive correction, and the distance between the patient and the projection screen), which may affect the reliability of the study. Moreover, with the accumulation of data and optimization of the model, we aim to continuously improve the performance of the model, provide higher-quality medical services to patients, and bring promising prospects for future diagnosis and treatment.

In conclusion, seven machine learning models were designed and implemented to enable simultaneous diagnosis of multiple paralyzed eye muscles. The DL (64 categories) and DL (6 binary classifications) algorithms achieved higher accuracy on the test set compared to other algorithms and demonstrated greater stability.

ACKNOWLEDGEMENTS

Foundations: Supported by National Natural Science Foundation of China (No.82074524); Harbin Medical University Graduate Research and Practice Innovation Project (No.YJSCX2023-50HYD).

Conflicts of Interest: **Jin XL,** None; **Li XM,** None; **Liu TJ,** None; **Zhou LY,** None.

REFERENCES

1 Danieli L, Montali M, Remonda L, *et al*. Clinically directed neuroimaging of ophthalmoplegia. *Clin Neuroradiol* 2018;28(1):3-16.

2 Weidauer S, Hofmann C, Wagner M, *et al*. Neuroradiological and clinical features in ophthalmoplegia. *Neuroradiology* 2019;61(4):365-387.

3 Ranjan R, Singh D, Mahesh KV, *et al*. Infectious ophthalmoplegias. *J Neurol Sci* 2021;427:117504.

4 Yamin A, Khan SA, Yasin UU. Automated system of Hess screen for diagnosis of paralytic strabismus using computer aided diagnosis. *2013 IEEE International Conference on Imaging Systems and Techniques (IST)*. October 22-23, 2013, Beijing, China. IEEE, 2013:300-305.

5 Orduna-Hospital E, Maurain-Orera L, Lopez-de-la-Fuente C, *et al*. Hess Lancaster screen test with eye tracker: an objective method for the measurement of binocular gaze direction. *Life* (*Basel*) 2023;13(3):668.

6 Zhou LY, Liu TJ, Li XM, *et al*. A new interpretation and quantitative method for diplopia test: 304 cases of ocular motor nerve palsy for clinical test and verify. *Int J Ophthalmol* 2017;10(11):1768-1770.

7 Zhou LY, Liu TJ, Li XM. Adult reference values of the computerized diplopia test. *Int J Ophthalmol* 2016;9(11):1646-1650.

8 Ting DSW, Lee AY, Wong TY. An ophthalmologist's guide to deciphering studies in artificial intelligence. *Ophthalmology* 2019;126(11):1475-1479.

9 Ting DSW, Pasquale LR, Peng L, *et al*. Artificial intelligence and deep learning in ophthalmology. *Br J Ophthalmol* 2019;103(2):167-175.

10 Xue Y, Zhu J, Huang X, *et al*. A multi-feature deep learning system to enhance glaucoma severity diagnosis with high accuracy and fast speed. *J Biomed Inform* 2022;136:104233.

11 Noury E, Mannil SS, Chang RT, *et al*. Deep learning for glaucoma detection and identification of novel diagnostic areas in diverse real-world datasets. *Transl Vis Sci Technol* 2022;11(5):11.

12 Keenan TDL, Chen Q, Agrón E, *et al*. DeepLensNet: deep learning automated diagnosis and quantitative classification of cataract type and severity. *Ophthalmology* 2022;129(5):571-584.

13 Karaaslan Ş, Kobat SG, Gedikpınar M. A new method based on deep learning and image processing for detection of strabismus with the Hirschberg test. *Photodiagnosis Photodyn Ther* 2023;44:103805.

14 Zheng C, Yao Q, Lu JW, *et al*. Detection of referable horizontal strabismus in children's primary gaze photographs using deep learning. *Transl Vis Sci Technol* 2021;10(1):33.

15 Hung JY, Perera C, Chen KW, *et al*. A deep learning approach to identify blepharoptosis by convolutional neural networks. *Int J Med Inform* 2021;148:104402.

16 Lincke A, Roth J, Macedo AF, *et al*. AI-based decision-support system for diagnosing acanthamoeba keratitis using *in vivo* confocal microscopy images. *Transl Vis Sci Technol* 2023;12(11):29.

17 Guo BT. Diplopia image testing equipment for ophthalmoplegia patients and diagnostic decision support system. Harbin Engineering University. 2013.

18 Zong ZF, Song GL, Liu Y. DETRs with collaborative hybrid assignments training. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. October 1-6, 2023, Paris, France. IEEE, 2023:6725-6735.

19 Wang P, Wang SJ, Lin JY, *et al*. ONE-PEACE: exploring one general representation model toward unlimited modalities. 2023:2305.11172. https://arxiv.org/abs/2305.11172v1.

20 An X, Deng JK, Guo J, *et al*. Killing two birds with one stone: efficient and robust training of face recognition CNNs by partial FC. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 18-24, 2022, New Orleans, LA, USA. IEEE, 2022:4032-4041.

21 Shanker RRBJ, Zhang MH, Ginat DT. Semantic segmentation of extraocular muscles on computed tomography images using convolutional neural networks. *Diagnostics* (*Basel*) 2022;12(7):1553.

22 Qureshi A, Lim S, Suh SY, *et al*. Deep-learning-based segmentation of extraocular muscles from magnetic resonance images. *Bioengineering (Basel)* 2023;10(6):699.

23 Lou L, Huang X, Sun Y, *et al*. Automated photographic analysis of inferior oblique overaction based on deep learning. *Quant Imaging Med Surg* 2023;13(1):329-338.

24 Lou L, Sun Y, Huang X, *et al*. Automated measurement of ocular movements using deep learning-based image analysis. *Curr Eye Res* 2022;47(9):1346-1353.