Application of Machine Learning in the Diagnosis of Vestibular Disease

Tram Anh Do
University of Toyama
Hiromasa Takakura
University of Toyama
Masatsugu Asai (masai@med.u-toyama.ac.jp)
University of Toyama
Naoko Ueda
University of Toyama
Hideo Shojaku
University of Toyama

Article

Keywords: equilibrium function tests, machine learning (artificial intelligence), scikit-learn, GridSearchCV, peripheral vestibular disease, non-peripheral vestibular disease

Posted Date: June 27th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1725101/v1

Abstract

Machine learning is expected as a potential aid to support human decision-making in disease prediction. In this study, we determined the utility of various machine learning algorithms in classifying peripheral vestibular (PV) and non-PV diseases through equilibrium function test results. The 1009 patients who had undergone our standardized neuro-otological examinations were recruited. We adopted five supervised machine learning algorithms (Random Forest, AdaBoost, Gradient Boosting, Support Vector Machine, and Logistic Regression). After preprocessing, tuning the best hyperparameters using GridSearchCV, and obtaining the final evaluation using scikit-learn in the testing set, the prediction capability was evaluated by various diagnostic test measures, namely accuracy, F1-score, area under the receiver operating characteristic curve, precision, recall, and Matthews correlation coefficient (MCC). All five algorithms yielded relatively good results with the accuracy of each machine learning algorithm ranging from 76–79%, with the best being the Support Vector Machine classifier. In cases where the predictions of the five models were consistent, the accuracy of the PV diagnosis results was improved with a probability of 83%, whereas it increased to 85% for the non-PV diagnosis results. Increasing the number of patients and optimizing classification methods are warranted to obtain the highest diagnostic accuracy.

Introduction

Dizziness and vertigo are frequently symptoms triggering patients to visit a physician. However, determining their cause is complicated because of the wide range of diseases with which they are associated. Peripheral vestibular (PV) system dysfunction is one of the most common causes of vertigo[1] such as benign paroxysmal positional vertigo, vestibular neuritis, and Meniere's disease[2]. The diagnostic criteria for PV disease are mainly based on the patient's history and a systematic clinical examination of the vestibular, ocular motor, and cerebellar systems according to the clinically oriented diagnostic criteria of the Bárány Society[3].

In our department, various equilibrium examinations, including the caloric test and optokinetic nystagmus test, have been used as diagnostic methods. In addition, imaging tests such as brain magnetic resonance imaging (MRI), brain computed tomography (CT), temporal bone CT, and cervical vascular echo are performed as needed in almost all patients. Emergency patients with vertigo/dizziness are not included in the subjects who underwent the examinations mentioned above. Our equilibrium examinations have been performed on outpatients who visited our department including patients who were referred from other hospitals and patients who were admitted to our department due to vertigo/dizziness. In addition, the patients after admission to various departments such as emergency, neurology, and neurosurgery in our hospital are also included. These patients often have difficult medical conditions resulting from undiagnosed, ineffective treatment and long-term persistence of symptoms. Therefore, accurate diagnosis of these patients and the differentiation of patients with PV disease from others are important tasks for us.

Machine learning (ML) has been developing rapidly in recent years and is being used in many aspects of medicine, especially in radiology, robotic surgical systems, and disease diagnosis[4–6]. Since the 1980s, most computer-based algorithms in medicine, called “expert systems”, have been used to simulate the steps and decision-making processes in the specific field of otolaryngology[7,8]. Recently, ML has been studied as a useful software method to assist medical decision-making for vestibular dysfunction[9–13]. These studies showed that ML is becoming a potential solution to help physicians most effectively access and use a huge amount of information to make accurate diagnoses. However, the patients we must diagnose in our daily practice vary in time from disease onset and in their degrees of symptoms. The most important performance we desire for ML is the ability to distinguish between PV disease and non-PV disease in patients who are difficult to diagnose. This result influences the choice of both the treatment and the next test to be performed.

The purpose of the present study was to evaluate the ML models created from various learning algorithms for binary classification between PV disease and non-PV disease using the datasets generated from our equilibrium examinations. Furthermore, a method to improve the performance of the ML model as much as possible was also devised.

Materials And Methods

This study was approved by the regional ethical standards committee of the Faculty of Medicine at the University of Toyama (Approval number: R2019003), and all experiments were performed in accordance with the relevant guidelines and regulations.

Subjects

The data of 1009 patients who underwent equilibrium examinations in our department from the Department of Otolaryngology, Head and Neck Surgery, University of Toyama, for the 10 years from 2009 to 2019 were retrieved. The number of patients with PV was 497 and that with non-PV disease was 512 (611 males and 398 females; mean age, 55.6 years). PV disease and non-PV disease were diagnosed according to the International Classification of Vestibular Disorders of the Bárány Society[14] and the guidelines of the Japan Society for Equilibrium Research[15] (Table 1). Among them, small acoustic neuromas corresponding to Koos[16] grade I or II were classified into the PV disease group. The patients who were confirmed to have unilateral PV dysfunction but could not be diagnosed as having an established clinical entity were classified into the PV disease group as having Inner ear disorder. Patients who were judged to have normal PV function and but in whom central nervous system disease was ruled out by neurological examinations and brain MRI/magnetic resonance angiography (MRA) or brain CT were classified into the non-PV disease group as having Dizziness syndromes of unclear etiology. However, even if brain MRI/MRA and brain CT did not show any abnormalities, patients who showed normal vestibular function but showed abnormalities in the Optokinetic nystagmus test and Eye tracking test were classified into the non-PV disease group as having Central balance disorder. These patients

often showed downbeat nystagmus, failure of fixation suppression, and abnormal eye movement. Although the cause of persistent postural-perceptual dizziness may exist in the PV system, these symptoms are thought to be modified by other factors. For this reason, patients with these symptoms were classified into the non-PV disease group.

Table 1
Clinical diagnosis of patients.

Peripheral vestibular disease (n = 497)	**Number**
Acoustic tumor (Koos I/II)	18
Benign paroxysmal positioning vertigo	44
Bilateral vestibulopathy	2
Cholesteatoma/chronic otitis media	29
Delayed endolymphatic hydrops	19
Facial nerve paralysis/Hunt syndrome	13
Inner ear disorder	156
Meniere's disease	120
Perilymphatic fistula	4
Sudden deafness with vertigo	44
Vestibular neuritis	48
Non-peripheral vestibular disease (n = 512)	**Number**
Brain infarction/bleeding	22
Brain tumor	20
Central balance disorder	87
Congenital nystagmus	8
Disembarkment syndrome	3
Dizziness syndromes of unclear etiology	32
Head injury	9
Hemodynamic orthostatic dizziness/vertigo	159
Migraine	7
Other central nervous system disease (n < 5)	31
Parkinsonism	5
Persistent postural-perceptual dizziness	23
Psychogenic dizziness	33
Spinocerebellar degeneration	24
Vertebrobasilar insufficiency/Vertebral basilar artery stenosis	49

All patients underwent our standardized neuro-otological examinations, listed from number (No.) 1 to No. 16 in Table 2. There are two types of features, Continuous features and Categorical features. In the former, the numerical values are used as they were. In the latter, an integer from 0 to 3 was assigned. The total number of features comprised 14 Continuous features and four Categorical features. Examination Nos. 1 to 14 were performed as routine examinations, and examination Nos. 15 and 16 were added as required. In the caloric test (No. 6), we used air calorization with the injection of airflow at 24°C and 50°C (6 L/min) for 60 seconds with eyes closed. The maximal slow phase velocity (MSPV), canal paresis percentage (CP%), and directional preponderance percentage (DP%) of the caloric nystagmus were recorded after each irrigation, and the CP% and DP% were calculated based on Jonkees' formula[17] using MSPV. In our department, if the CP is ≥ 20%, the ear with the lower response is assumed to have unilateral vestibular hypofunction, indicating an abnormal caloric reflex. Bilateral vestibular hypofunction as evaluated by MSPV was < 6°/s in each ear after caloric stimulation[18]. Failure of the fixation suppression test (No. 7) started at 80 seconds from the start of air calorization for 10 seconds. The patient with eyes open stared at the optotype[19,20]. The pendular sinusoidal rotation test (No. 8) was performed with rotation of the chair at 0.l Hz, amplitude 240°, maximum velocity of 75.4°/s, with the patient's eyes closed[21]. In the Eye tracking test (No. 9), the patient gazed at and pursued an optotype lamp (viewing angle 20 degrees, frequency 0.3 Hz) that moves left and right[22]. The waveform of electronystagmography (ENG) is judged by inspection. In the Optokinetic nystagmus test (No. 10), 12 striations are projected onto a hemispherical drum. The rotation of striations to the clockwise (CW) direction starts at 1°/s until a velocity of 100°/s is reached. After that, rotation to the counter-clockwise (CCW) direction starts[23]. Stabilometry (No. 11) was performed according to the Japanese standard[24]. The Mann test (No. 12) is performed under tandem standing with the eyes open and closed for 30 seconds then the position of the front and back legs is reversed[25]. In the Fukuda stepping test (No. 13), the patient stands upright with eyes closed and with arms extended forward, and steps for 50 steps[26,27]. In the Schellong test (No. 14), blood pressure is measured twice in a lying position and 3 times: immediately after standing and at 5 and 10 minutes later[28]. The Galvanic Body Sway test (GBST) (No. 15) evaluates body sway response induced by 0.2 mA and 0.4 mA electrical stimulation applied to the retroauricular area. Bipolar rectangular current stimulation for 3 seconds is repeated 10 times alternately left and right for the patient standing on the stabilometer with both feet closed together[29]. The stimulus conditions of the cervical vestibular evoked myogenic potential test (cVEMP) (No. 16) were a click sound of 0.1 msec, a frequency of 5 Hz, and a sound pressure level of 105 dB. The reaction waveform was added 200 times[30].

Table 2
Equilibrium examinations and each feature's name.

Examinations	Feature names	Types of features			
		0	1	2	3
1. Spontaneous nystagmus test	SpontanNy	No nystagmus	Nystagmus to the right	Nystagmus to the left	Other
2. Nystagmus during neck torsion to the right or left	NeckTor_R NeckTor_L				
3. Nystagmus during neck compression to the right or left carotid sinus	NeckComp_R NeckComp_L				
4. Positional nystagmus of 6 head positions	PositionalNy1- PositionalNy6				
	PositionalNum	**Number of head positions nystagmus appeared (0 to 6)**			
5. Positioning nystagmus of 6 head positions	PositioningNy1- PositioningNy6	No nystagmus	Nystagmus to the right	Nystagmus to the left	Other
	PositioningNum	**Number of head positions nystagmus appeared (0 to 6)**			
6. Bithermal caloric test	Caloric_CP% Caloric_DP%	**Caloric_CP%, Caloric_DP%**			
	Caloric_CP	No CP	Suspected	Unilateral CP	Bilateral CP
	Caloric_DP	No DP	Suspected	DP	
7. Failure of fixation suppression test (FFS)	FFS	Suppression ≥ 60% in 4 calorizations	Suppression < 60% in 1–2 calorizations	Suppression < 60% in 2–4 calorizations	
8. Pendular sinusoidal rotation test (PRST)	PSRT_R PSRT_L	**PSRT gain in rotation to the right or the left**			
	PSRT_DP%	**PSRT_DP%**			
9. Eye tracking test (ETT)	ETT	Smooth	Saccadic	Ataxic	

Examinations	Feature names	Types of features			
		0	1	2	3
10. Optokinetic nystagmus (OKN) test	OKN_CW OKN_CCW	**MSPV in CW rotation, CCW rotation**			
	OKN_DP%	**OKN_DP%**			
11. Stabilometry	Envelop area_Op Envelop area_Cl	**Enveloped area (cm^2) with eyes open and eyes closed**			
	Sway Length_Op Sway Length_Cl	**Sway length (mm) with eyes open and eyes closed**			
	Romberg_Area Romberg_Length	**Romberg ratio of enveloped area and sway length**			
12. Mann test	Mann	Standing ≥ 30 sec	Fall < 30 sec		
13. Fukuda stepping test	Stepping	RA < 45 deg and TA < 45 deg and TD < 1m	RA ≥ 45 deg or TA ≥ 45 deg or TD ≥ 1m		
14. Schellong test	Schellong	Decreased SBP < 20 mmHg and PP narrowing < 20 mmHg	Decreased SBP ≥ 20 mmHg or PP narrowing ≥ 20 mmHg		
15. Galvanic body sway test (GBST)	GBST_R, GBST_L	Normal response to 0.2 mA or 0.4 mA	Suspected	No response to 0.2 mA and 0.4 mA	
16. Cervical vestibular evoked myogenic potentials (cVEMP)	cVEMP_R, cVEMP_L	0.5 ≤ L/R ≤ 2.0	L/R < 0.5 or L/R > 2.0		

Examinations	Feature names	Types of features			
		0	1	2	3
Caloric_CP% = ((24°L + 50°L) - (24°R + 50°R)) * 100 / (24°L + 24°R + 50°L + 50°R)					
Caloric_DP% = ((50°L + 24°R) - (50°R + 24°L)) *100 / (50°L + 24°R + 24°L + 50°R)					
PSRT_Gain = MSPV/75.4 (°/s) in rotation to the right/to the left					
PSRT_DP% = (L□MSPV - R□MSPV) * 100 / (L□MSPV + R□MSPV)					
OKN_DP% = (CW□MSPV - CCW□MSPV) * 100 / (CW□MSPV + CCW□MSPV)					
Romberg_Area = Area (eyes closed) / Area (eyes open)					
Romberg_Length = Length (eyes closed) / Length (eyes open)					
Abbreviations: Ny, Nystagmus; PositionalNy1-PositionalNy6, Positional nystagmus in each of 6 head positions; PositioningNy1- PositioningNy6, Positioning nystagmus of 6 head positions; R, right; L, left; CP, canal paresis; DP, directional preponderance; MSPV, maximal slow phase velocity (°/s); CW or CCW, clockwise or counterclockwise; Op, open; Cl, close; RA, rotation angle; TA, transition angle; TD, transition distance; SBP, systolic blood pressure; PP, pulse pressure; L/R, L/R ratio of amplitude.					

Steps in the machine learning classification method

In the present research, we used “classification” in supervised ML, which aims to predict categories of a new observation based on a training set of data whose category is known[31]. The program was created on the Google Colaboratory using Python version (v) 3.7.12, scikit-learn[32] v1.0.2, NumPy v1.21.5, SciPy v1.4.1, Pandas v1.3.5, and Matplotlib v3.2.2. Five well-known algorithms, Random Forest (RF), AdaBoost (AB), Gradient Boosting (GB), Support Vector Machine (SVM), and Logistic Regression (LR), were adopted. They have been used in a large number of treatises and specialized books based on an already established theory[33–39]. The steps in classification are as follows.

Import the data

From the results of the 1009 patients, we created a data CSV file consisting of 44 features and target categories (PV = 0, non-PV = 1). After importing the CSV data into the program, the data were pre-processed to ensure the accuracy of future predictions[40].

Split the data

The pre-processed dataset was randomly divided into 75% training data (n = 756) and 25% testing data (n = 253), as shown in Fig. 1.

ML and predictions

ML was performed to create the best model using the training data. In the learning process, various parameters in the algorithm are automatically adjusted. However, some parameters need to be determined by a human to achieve the best prediction[41]. These are called hyperparameters, which can be determined using “GridsearchCV” in Scikit-learn[32]. We applied this to the training data and determined the best hyperparameters. The best models created by the best hyperparameters were applied to the testing data as shown in Fig. 1 to create the final evaluation output.

Test measures

In the binary classification, one of the predicted two groups was called the negative group (N), and the other was called the positive group (P). We defined a PV disease as (N) and a non-PV disease as (P). The confusion matrix is commonly used to evaluate the diagnostic ability of classifiers. In Table 3, the basic framework of the confusion matrix[32] displays the predicted number of subjects into four categories by each model: TP (true positive), FP (false positive), FN (false negative), and TN (true negative).

Table 3
Basic framework of the confusion matrix.

Confusion matrix		**Predicted class**	
		Negative (class 0)	**Positive (class 1)**
Real class	Negative (class 0)	Number of True Negatives	Number of False Positives
	Positive (class 1)	Number of False Negatives	Number of True Positives

Class 0 = peripheral vestibular disease; class 1 = non-peripheral vestibular disease.

The six test measures used for evaluating the predictive performance of ML are shown below. The range of values of the first five measures is displayed as numerical values from 0 to 1, whereas the Matthews correlation coefficient (MCC) is displayed as numerical values from − 1 to 1. The larger this value, the higher the prediction performance.

1. Accuracy $= \frac{(\mathrm{TP}+\mathrm{TN})}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}}$
2. Precision $= \frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FP})}$
3. Recall (also known as sensitivity) $= \frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FN})}$
4. Area under the receiver operating characteristic curve (AUC-ROC)
5. F1- score $= \frac{2 \times \mathrm{precision} \times \mathrm{recall}}{\mathrm{precision}+\mathrm{recall}}$

6. $$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(\mathrm{TP}+\mathrm{FP}) \times (\mathrm{TP}+\mathrm{FN}) \times (\mathrm{TN}+\mathrm{FP}) \times (TN+FN)}}$$

The receiver operating characteristic (ROC) curve presents the points of the recall plotted against the false-positive rate (FPR = FP/(TN + FP)) at different classification thresholds in the graph (Fig. 2). The AUC-ROC is a numerical representation of the proportion of the graph that falls under the ROC curve[42,43]. The accuracy, AUC-ROC, and the MCC present the prediction performance with one numerical value. Among them, accuracy is fundamental and the most frequently used measure, but it has been pointed out that its predictive performance for an imbalanced dataset is unreliable[44]. In contrast, the MCC is superior to accuracy or AUC-ROC for imbalanced data[45,46]. The precision, recall, and F1-score evaluate the prediction performance for the positive classes. However, replacement of the positive class from non-PV to PV is possible in Scikit-learn. Thus, these three measures were calculated for both PV and non-PV. The F1-score was calculated as a harmonic mean of precision and recall and helps to consider both metrics[47].

Statistical analysis

The Mann-Whitney U test was used for statistical evaluation of precision, recall, and F1-score between PV and non-PV. BellCurve for Excel v3.21 (Social Survey Research Information Co., Ltd., Japan) was used for the analysis, and $P < 0.05$ was considered statistically significant.

Results

We created five models of classifiers for binary classification using training data (n = 756) and applied them to testing data (n = 253), which were composed of PV (n = 123) and non-PV (n = 130) data. The prediction performances of the different models are summarized in Table 4 with the six evaluation measures. The best results among the five models were 0.79 for SVM in accuracy, 0.87 for LR in AUC-ROC, and 0.57 for SVM in MCC. Although SVM showed the best prediction performance, there was not so much difference between the five models. The AUC-ROC is one of the most-used metrics in evaluating the performance of binary classifiers. Based on the comparison of the ROC curves, which was conducted between the five ML models in Fig. 2, high predictive results from 0.85 to 0.87 were shown. All models showed coverage of a large area with a high AUC value and similar ROC curves representing the superiority and good results of all of the classifiers. The predicted performance for each of PV and non-PV was confirmed by precision, recall, and the F1-score. The best precision was 0.78 by SVM for PV, and 0.80 by SVM and LR for non-PV. The best recall was 0.80 by LR for PV and 0.78 by GB and SVM for non-PV. The best F1-score was 0.78 by SVM and LR for PV, and 0.79 by SVM for non-PV. Apart from the superiority of the individual model, the average of precision of non-PV showed a higher value than that of PV with 0.79 for non-PV and 0.76 for PV ($P < 0.05$). The average of recall was 0.78 for PV and 0.76 for non-PV (no significant difference). The average for the F1-score was 0.77 for PV and 0.78 for non-PV (no significant difference).

Table 4
Performance of different machine learning algorithms.

ML classifiers	Accuracy	AUC-ROC	MCC	Precision		Recall		F1-score	
				PV	non-PV	PV	non-PV	PV	non-PV
RF	0.77	0.85	0.55	0.76	0.79	0.78	0.77	0.77	0.78
AB	0.76	0.84	0.52	0.74	0.78	0.78	0.74	0.76	0.76
GB	0.76	0.86	0.52	0.76	0.76	0.74	**0.78**	0.75	0.77
SVM	**0.79**	0.85	**0.57**	**0.78**	**0.80**	0.79	**0.78**	**0.78**	**0.79**
LR	0.78	**0.87**	0.56	0.76	**0.80**	**0.80**	0.75	**0.78**	0.78
Average	0.77	0.85	0.54	0.76*	0.79*	0.78n.s.	0.76 n.s.	0.77 n.s.	0.78 n.s.
SD	0.01	0.01	0.02	0.01	0.02	0.02	0.02	0.01	0.01

Bold letter indicates best score in each measure.

Abbreviations: ML, machine learning; AUC-ROC, area under the receiver operating characteristic curve; MCC, Matthews correlation coefficient; PV, peripheral vestibular disease; non-PV, non-peripheral vestibular disease; RF, Random Forest; AB, AdaBoost; GB, Gradient Boosting; SVM, Support Vector Machine; LR, Logistic Regression; SD, standard deviation. *, $P < 0.05$ between PV and non-PV, n.s., no significant difference between PV and non-PV.

An index showing how much each of the features contributed to making a prediction can be calculated by the property of “feature_importances” for several models in scikit-learn. The “feature_importances” ranking indicates which features may be most relevant or least relevant to the research objective. The RF method is the most common method in feature importance selection and rankings[48]. In our research, feature importance of RF, AB, and GB was ranked based on the selected frequency of a variable as a decision node of decision trees. We used all of these classifiers to rank variable importance according to their discriminative performance. The selected top 10 from 44 features are presented in Fig. 3. Each feature is ranked with a numerical value ranging between 0 and 1, where 0 means “not used at all” and 1 means “perfectly predicts the target”. The higher the value, the more important it is. Within the features for evaluating vestibular function, the features of the caloric test (Caloric_CP, Caloric_CP%) were ranked the highest in all three models. This confirms that CP in the caloric test is a parameter that plays an important role in classifying PV disease and non-PV disease. As for the features related to the stabilometry test, the Romberg ratio of sway length (Romberg_Length) was included in the top 10 as high as Caloric_CP in ABC. The others, such as Envelop Area_Op, Envelop Area_Cl, Sway Length_Op, Sway Length_Cl, and Romberg_Area, were present in the top ranking of the three models. Within the features for assessing cerebellar and brainstem function, two features of the Optokinetic nystagmus test (OKN_CW, OKN_CCW) were included in RFC. The other features of the eye tracking test (ETT), Schellong test

(Schellong), and pendular sinusoidal rotation test (PRST_R, PRST_L) were present in the top ranking of the three models.

All five models were applied to evaluate the accuracy of 25% testing data (253 data cases) of PV and non-PV in Fig. 4. The number of models with matching predictions are shown in the six columns, which are marked with labels PV 0 to PV 5 and non-PV 0 to non-PV 5. In the graph, PV 5/Non-PV 0 means that all five classifiers predicted PV and no model predicted non-PV. PV 0/Non-PV 5 means that all five models predicted non-PV and no model predicted PV. The other labels indicate that the predictions were different depending on the model. In the first column, among the 104 patients predicted to have PV by all five models, 86 patients truly had PV and 18 patients had non-PV, which is equivalent to 83% accuracy. Similarly in the last column, among the 100 patients predicted to have non-PV by all five models, 85 patients truly had non-PV, and 15 patients had PV, which is equivalent to 85% accuracy. These percentages were higher than the accuracy of the models individually (Table 4).

Discussion

In the present study, we evaluated the ability of ML models created from five algorithms to discriminate between PV disease and non-PV disease. These five algorithms were the commonly used RF, AB, GB, SVM, and LR methods and suggest the potential for supporting the prediction of vestibular disease diagnosis. Furthermore, our approach of combining all five ML classifier models was expected to support the prediction performance of each model individually.

All five models presented relatively good results by tuning the algorithms and choosing the best parameters using GridSearchCV. Among the five models, the results of SVM in Table 4 seemed to be superior to those of the other models. Varpa et al. applied one-vs-one and one-vs-all classifiers in the k-nearest neighbor method and SVM in the classification of vertigo data and reported that using multiple binary classifiers improves the classification accuracies of disease classes compared to one multi-class classifier[49]. Masankaran et al.[50] used four classifier models (RFC, SVC, k-nearest neighbor, and Naïve Bayes) with the Dizziness Handicap Inventory questionnaire to distinguish benign paroxysmal positioning vertigo types with a best accuracy of 73.91%. Priesol et al.[11] applied five classifier models (DT, RF, LR, AB, and SVM) and reported an overall accuracy of 76%. Compared to these reports, the performance of our best classifier had a higher accuracy of 79%. To further improve performance, we devised a method by combining all five models in the prediction data (Fig. 4). As a result, when the predictions of the five models matched in PV, the correct answer rate was 83%, and when they matched in non-PV, the correct answer rate was 85%. This result was superior to the accuracy of SVM alone. However, when PV and non-PV predictions were presented simultaneously, the accuracy of SVM was superior. Therefore, the combination of SVM together with our new ML approach has the potential to diagnose PV disease and distinguish it from non-PV disease.

For otolaryngologists, it is important to reliably detect PV disease in patients with chaotic symptoms of vertigo/dizziness. However, the non-PV group included various diseases of cerebral etiology such as brain

tumor, brain infarction, spinocerebellar degeneration, vertebrobasilar insufficiency, and others for which a delayed diagnosis might lead to life-threatening consequences. Thus, ML should have a high predictive ability not only for PV diseases but also for non-PV diseases. This balance of predictive performance can be evaluated using precision, recall, and the F1-score. The F1-score is a measure that can comprehensively evaluate precision and recall. As shown in Table 4, the precision average of the five models was better in non-PV than PV. However, the F1-score averages of the five models were 0.77 for PV and 0.78 for non-PV. This result means that our models function well for both groups. Furthermore, the F1-scores of SVM were the best with 0.78 for PV and 0.79 for non-PV. Thus, SVM appears to be a useful classifier for discriminating both disease groups.

Our dataset was established based on the clinical data of patients who were diagnosed by our 16 different types of equilibrium function tests, whereas previous studies usually used the most commonly performed vestibular tests such as the caloric test and vestibulo-ocular reflex derived from the rotation test[11] or used head impulse, gaze-evoked nystagmus, or test of skew for differentiation of vestibular stroke and peripheral acute vestibular syndrome[9]. In Fig. 3, features related to the caloric test were the most important features, but the optokinetic nystagmus test, eye tracking test, Schellong test, pendular sinusoidal rotation test, and stabilometry also ranked in the top 10. Thus, the combination of multiple kinds of equilibrium examinations might help to increase the variety of features and improve the quality of the training dataset for ML. However, not all features in our dataset have equal importance. Determining which features yield the most predictive power is another crucial step in the model-building process.

This study has some important limitations, including the characteristics and size of the dataset and optimization of the models. In this study, ML was used to classify PV disease and non-PV disease, which include a wide range of diseases. Further studies using synthetic models in the classification of PV disease and a particular disease are needed to improve the diagnostic ability of ML. In addition, the number of study subjects was relatively modest, and other ML algorithms using advanced analytics techniques will be necessary to enhance the results. Furthermore, obtaining extensive testing batteries as presented here will not tailor for clinical decision making in the setting of acutely dizzy patients in an emergency condition or in an outpatient center without examination equipment. Finally, even though ML can assist in making good predictions, it does not completely replace the physician. Especially with some diseases, which require patient-physician interaction and critical thinking, the physician needs to make the final diagnosis.

Conclusion

Diagnosis in neuro-otology is mainly deductive and based on the results of various vestibular function tests, which are difficult for otolaryngologists because of the experience requirements and the time-consuming nature of the tests. The current algorithm shows the effectiveness of using five ML models as an adjunct to distinguish between PV and non-PV diseases. The adoption of ML algorithms in clinical

practice might free up physician time and enhance the accuracy and efficiency of the diagnosis and treatment of patients with PV disease.

Declarations

Acknowledgements

This work was supported by a grant from the Ministry of Health, Labor and Welfare of Japan (20FC1048).

We express our sincerely appreciation to all members of the Department of Otolaryngology, University of Toyama, for their assistance and cooperation.

Author contributions

T.A.D., study design, patient recruitment, data acquisition, analysis, and manuscript writing. H.T., study design, manuscript revision, and supervision. M.A., study design, writing machine learning programs, and manuscript revision. N.U., patient recruitment; H.S., grant application, study design, manuscript revision. All authors contributed to the article, read, and approved the submitted version.

Data availability

The datasets used and/or analyzed during the current study cannot be shared publicly so as to maintain the privacy of the individuals who participated in the study. The data will be shared on reasonable request to the corresponding author.

Competing interests

The authors declare no competing interests.

References

1. Labuguen, R. H. Initial evaluation of vertigo. Am. Fam. Physician **73**, 244–251 (2006).
2. Stern, S. D. C., Cifu, A. S. & Altkorn, D. Dizziness. in *Symptom to diagnosis: an evidence-based guide, 3e* (ed. Stern, S. D. C.) (McGraw-Hill Education, 2014).
3. Strupp, M., Feil, K. & Zwergal, A. Diagnosis and differential diagnosis of peripheral and central vestibular disorders. Laryngorhinootologie **100**, 176–183 (2021).
4. Mayo, R. C. & Leung, J. Artificial intelligence and deep learning – Radiology's next frontier? Clin. Imaging **49**, 87–88 (2018).
5. Egert, M., Steward, J. E. & Sundaram, C. P. Machine learning and artificial intelligence in surgical fields. Indian J. Surg. Oncol. **11**, 573–577 (2020).

6. Rajkomar, A., Dean, J. & Kohane, I. Machine learning in medicine. N. Engl. J. Med. **380**, 1347–1358 (2019).
7. Gavilán, C., Gallego, J. & Gavilán, J. “Carnisel”: An expert system for vestibular diagnosis. Acta Otolaryngol. **110**, 161–167 (1990).
8. Viikki, K., Kentala, E., Juhola, M. & Pyykkö, I. Decision tree induction in the diagnosis of otoneurological diseases. Med. Inform. Internet Med. **24**, 277–289 (1999).
9. Ahmadi, S. A. *et al.* Modern machine-learning can support diagnostic differentiation of central and peripheral acute vestibular disorders. J. Neurol. **267**, 143–152 (2020).
10. Kamogashira, T. *et al.* Prediction of vestibular dysfunction by applying machine learning algorithms to postural instability. Front. Neurol. **11**, 5–12 (2020).
11. Priesol, A. J., Cao, M., Brodley, C. E. & Lewis, R. F. Clinical vestibular testing assessed with machine-learning algorithms. JAMA Otolaryngol. Head Neck Surg. **141**, 364–372 (2015).
12. Juhola, M. On machine learning classification of otoneurological data. Stud. Health Technol. Inform. **136**, 211–216 (2008).
13. Walther, L. E. *et al.* Die Anwendung künstlicher neuronaler Netze bei der Auswertung posturografischer Messungen [The Use of Artificial Neural Networks in Evaluation of Posturographic Data]. Interpretation: A Journal of Bible and Theology 211–217 (2011).
14. Bisdorff, A., von Brevern, M., Lempert, T. & Newman-Toker, D. E. Classification of vestibular symptoms: Towards an international classification of vestibular disorders. J. Vestib. Res. **19**, 1–13 (2009).
15. Japan Society for Equilibrium Research. https://www.memai.jp/guide/.
16. Erickson, N. J. *et al.* Koos classification of vestibular schwannomas: a reliability study. Neurosurgery **85**, 409–414 (2019).
17. Jongkees, L. B. W., Maas, J. P. M. & Philipszoon, A. J. Clinical nystagmography. A detailed study of electro-nystagmography in 341 patients with vertigo. Pract. Otorhinolaryngol. (Basel). **24**, 65–93 (1962).
18. Strupp, M. *et al.* Bilateral vestibulopathy: Diagnostic criteria consensus document of the Classification Committee of the Bárány Society. J. Vestib. Res. **27**, 177–189 (2017).
19. Kato, I. *et al.* Caloric pattern test with special reference to failure of fixation-suppression. Acta Otolaryngol. **88**, 97–104 (1979).
20. Kato, I., Nakamura, T., Koike, Y. & Watanabe, Y. Computer analysis of fixation-suppression of caloric nystagmus. ORL J. Otorhinolaryngol. Relat. Spec. **44**, 277–287 (1982).
21. Mizukoshi, K., Kobayashi, H., Ohashi, N. & Watanabe, Y. Quantitative analysis of the human visual vestibulo-ocular reflex in sinusoidal rotation. Acta Otolaryngol. Suppl. **393**, 58–64 (1983).
22. Ohashi, N., Watanabe, Y., Kobayashi, H. & Mizukoshi, K. Quantitative comparison between saccadic and ataxic pursuits. Acta Otolaryngol. **101**, 200–206 (1986).

23. Watanabe, Y., Ohashi, N., Ohmura, A., Itoh, M. & Mizukoshi, K. Gain of slow-phase velocity of optokinetic nystagmus. Auris Nasus Larynx **13**, S63–S68 (1986).
24. Yamamoto, M. *et al.* Japanese standard for clinical stabilometry assessment: Current status and future directions. Auris Nasus Larynx **45**, 201–206 (2018).
25. Ito, S., Odahara, S., Hiraki, M. & Idate, M. Evaluation of imbalance of the vestibulo-spinal reflex by "The Circular Walking Test." Acta Otolaryngol. Suppl. **115**, 124–126 (1995).
26. Fukuda, T. The stepping test: Two phases of the labyrinthine reflex. Acta Otolaryngol. **50**, 95–108 (1959).
27. Cohen, H. S. A review on screening tests for vestibular disorders. J. Neurophysiol. **122**, 81–92 (2019).
28. Fanciulli, A., Campese, N. & Wenning, G. K. The Schellong test: detecting orthostatic blood pressure and heart rate changes in German-speaking countries. Clin. Auton. Res. **29**, 363–366 (2019).
29. Watanabe, Y. *et al.* Retro-labyrinthine disorders detected by galvanic body sway responses in routine equilibrium examinations. Acta Otolaryngol. Suppl. **108**, 343–348 (1989).
30. Shojaku, H., Takemori, S. & Watanabe, Y. Vestibular evoked myogenic potentials. Equilib. Res. **59**, 186–192 (2000).
31. Handelman, G. S. *et al.* eDoctor: machine learning and the future of medicine. J. Intern. Med. **284**, 603–619 (2018).
32. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011).
33. Breiman, L. Random forests. Machine Learning **45**, 5–32 (2001).
34. Friedman, J. H. Greedy function approximation: A gradient boosting machine. Ann. Stat. **29**, 1189–1232 (2001).
35. Hu, W., Member, S., Hu, W. & Maybank, S. AdaBoost-based algorithm for network intrusion detection. IEEE Trans. Syst. Man. Cybern. B Cybern. **38**, 577–583 (2008).
36. Cortes, C. & Vapnik, V. Support-vector networks. Mach. Learn. **20**, 273–297 (1995).
37. Yu, W., Liu, T., Valdez, R., Gwinn, M. & Khoury, M. J. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC Med. Inform. Decis. Mak. **10**, 16 (2010).
38. Hosmer, D. & Lemeshow, S. *Applied logistic regression, 3e* (ed. Hosmer, D.) (Wiley, 2004).
39. Colombet, I., Jaulent, M. C., Degoulet, P. & Chatellier, G. Logistic regression model: an assessment of variability of predictions. Stud. Health Technol. Inform. **84**, 1314–1318 (2001).
40. Ngiam, K. Y. & Khor, I. W. Big data and machine learning algorithms for health-care delivery. Lancet Oncol. **20**, e262–e273 (2019).
41. Müller, A. C. & Guido, S. Model evaluation and improvement in *Introduction to Machine Learning with Python, 1e* (ed. Müller, A. C.) 262–263 (O'Reilly Media, 2017).
42. Fawcett, T. An introduction to ROC analysis. Pattern Recognit. Lett. **27**, 861–874 (2006).

43. Bewick, V., Cheek, L. & Ball, J. Statistics review 13: Receiver operating characteristic curves. Crit. Care **8**, 508–512 (2004).
44. Chicco, D. & Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1-score and accuracy in binary classification evaluation. BMC Genomics **21**, 1–13 (2020).
45. Cao, C., Chicco, D. & Hoffman, M. M. The MCC-F1 curve: a performance evaluation technique for binary classification. *arXiv:2006.11278* 1–17 (2020).
46. Chicco, D., Tötsch, N. & Jurman, G. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. BioData Min. **14**, 1–22 (2021).
47. Rousseau, R. The F-measure for research priority. J. Data Inf. Sci. **3**, 1–18 (2020).
48. Genuer, R., Poggi, J. M. & Tuleau-Malot, C. Variable selection using random forests. Pattern Recognit. Lett. **31**, 2225–2236 (2010).
49. Varpa, K., Joutsijoki, H., Iltanen, K. & Juhola, M. Applying one-vs-one and one-vs-all classifiers in k-nearest neighbour method and support vector machines to an otoneurological multi-class problem. Stud. Health Technol. Inform. **169**, 579–583 (2011).
50. Masankaran, L., Viyanon, W. & Mahasittiwat, V. Classification of benign paroxysmal positioning vertigo types from Dizziness Handicap Inventory using machine learning techniques. 2018 *International Conference on Intelligent Informatics and Biomedical Sciences, ICIIBMS 2018* **3**, 209–214 (2018).

Figures

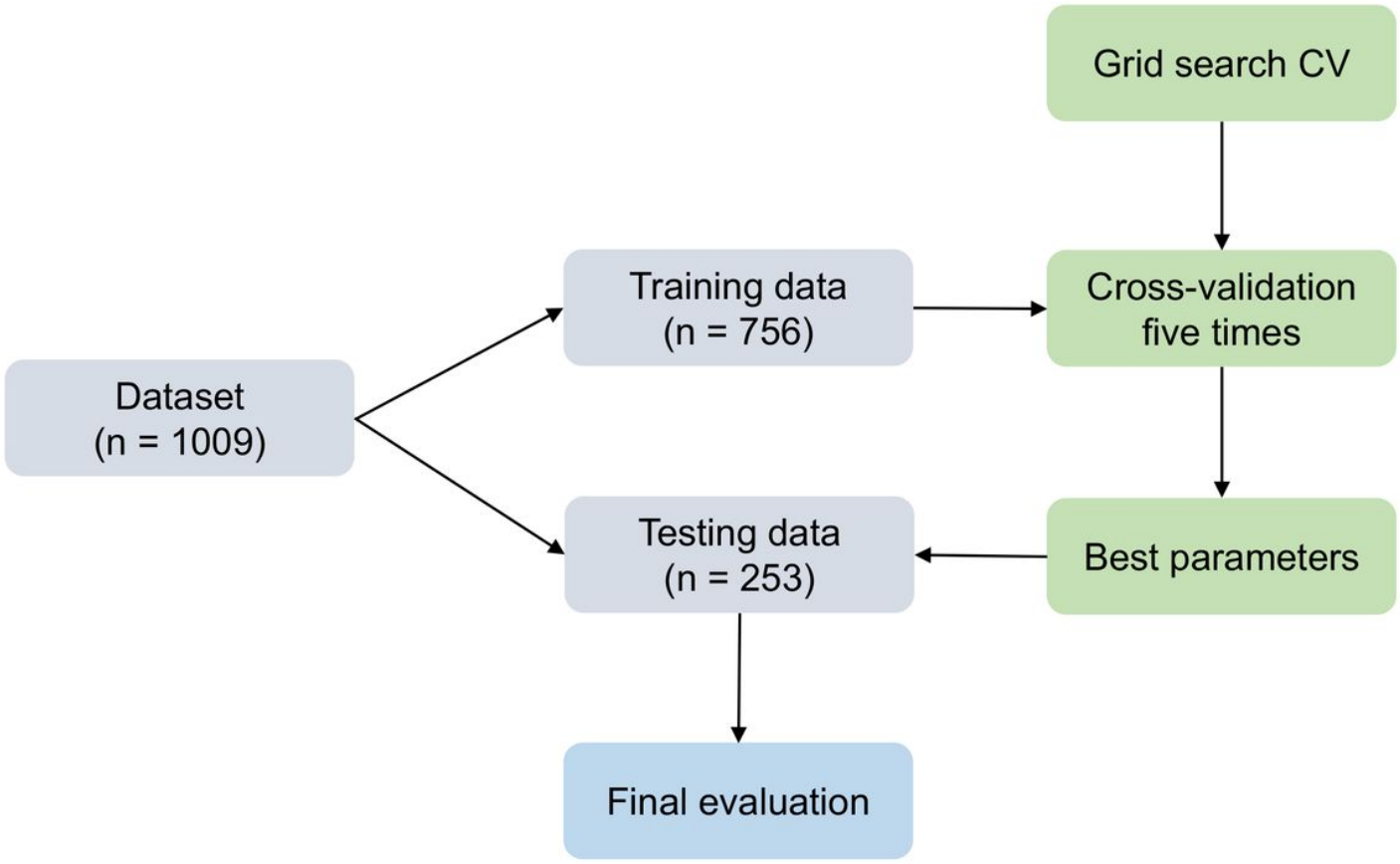

Figure 1

Overview of our machine learning process.

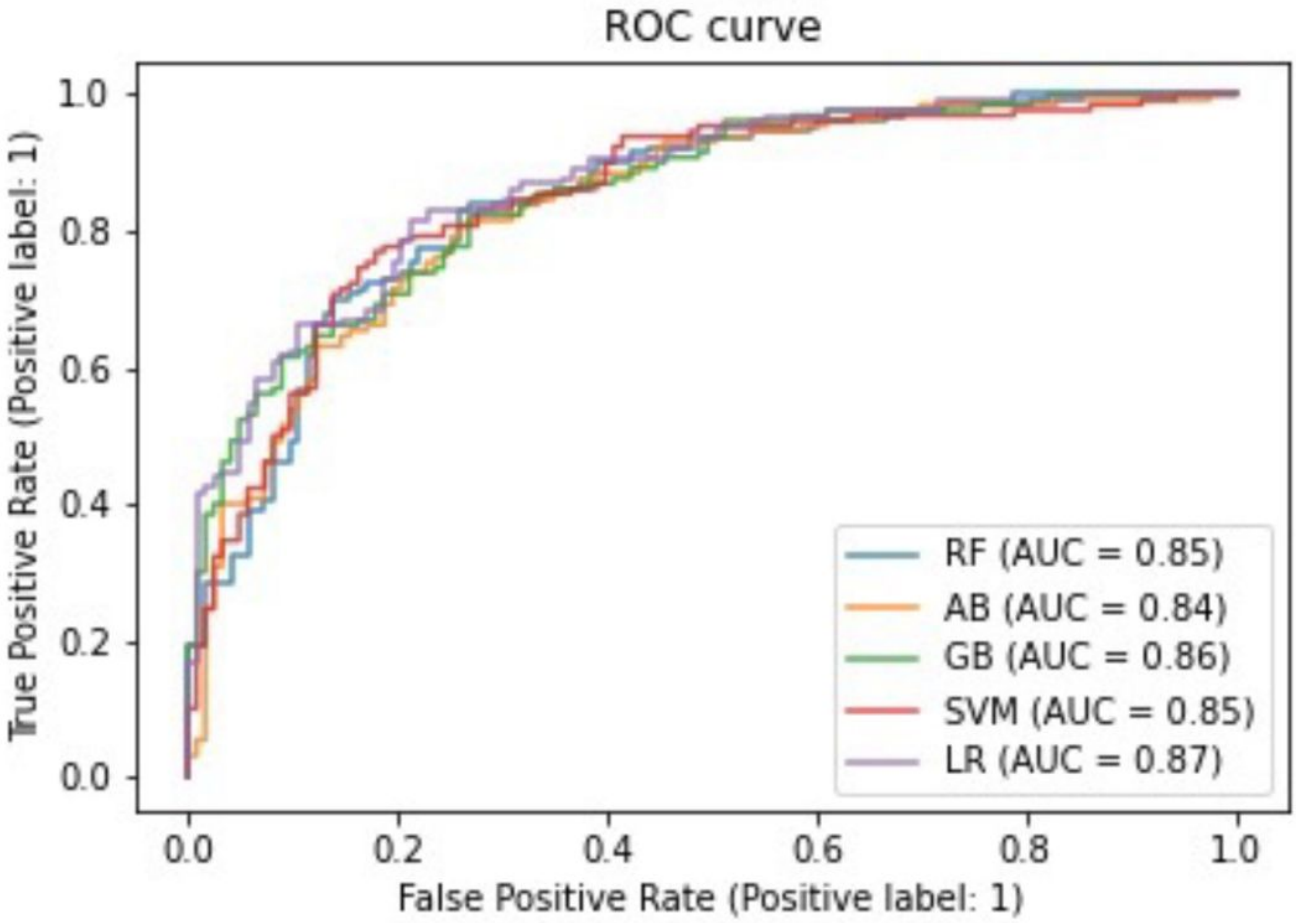

Figure 2

Comparison of ROC curves among the five machine learning models. **Abbreviations:** ROC, receiver operating characteristic; AUC, area under the curve; RF, Random Forest; AB, AdaBoost; GB, Gradient Boosting; SVM, Support Vector Machine; LR, Logistic Regression.

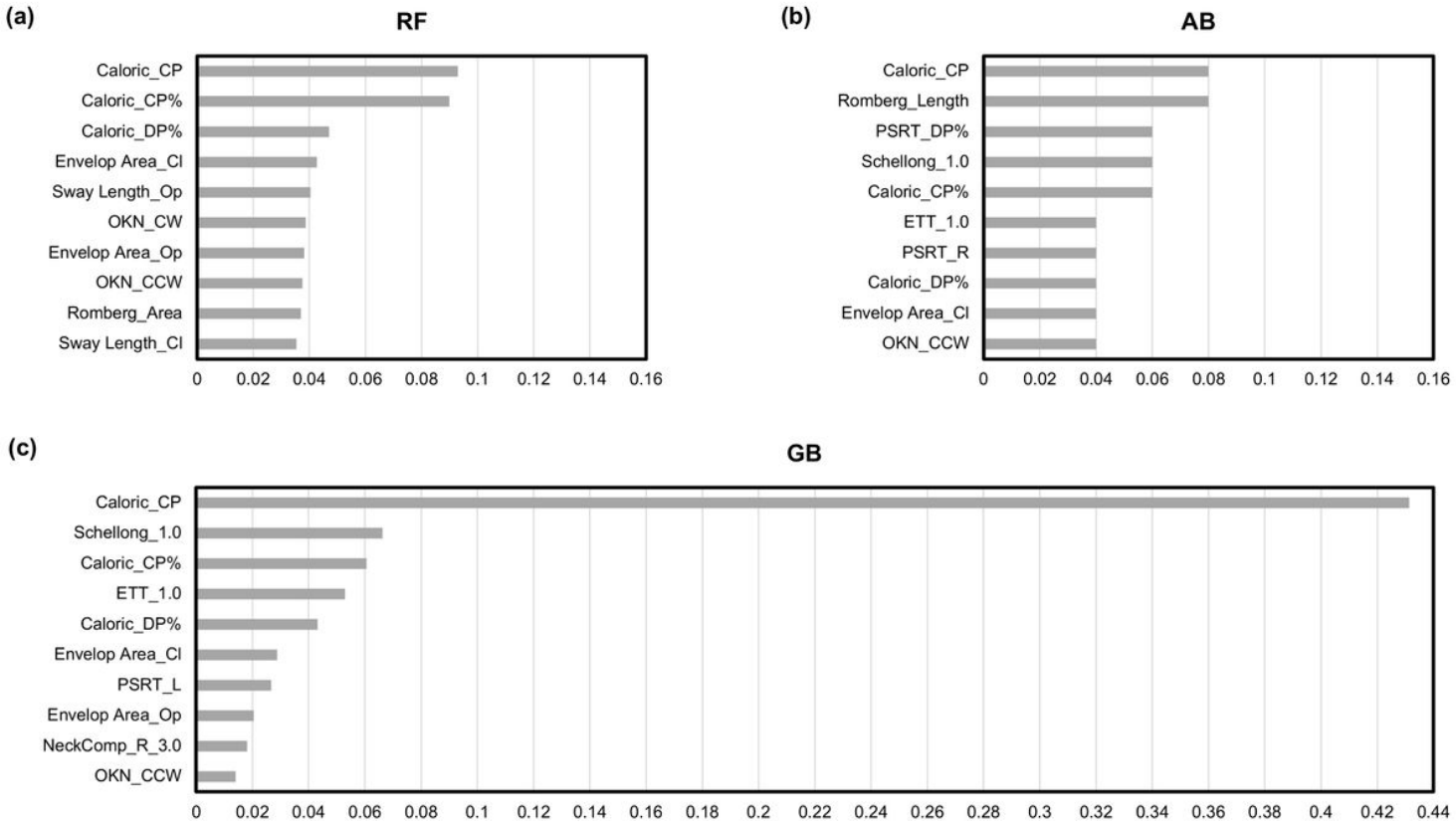

Figure 3

Top 10 most important features, ranked from high to low by classifier models. (a) Random Forest. (b) AdaBoost. (c) Gradient Boosting. **Abbreviations:** RF, Random Forest; AB, AdaBoost; GB, Gradient Boosting; CP, canal paresis in caloric test; DP, directional preponderance in caloric test; PSRT, pendular sinusoidal rotation test; R, right; L, left; Op, open; Cl, close; OKN, optokinetic; CW, clockwise; CCW, counterclockwise; FFS, failure of fixation suppression test; ETT, eye tracking test.

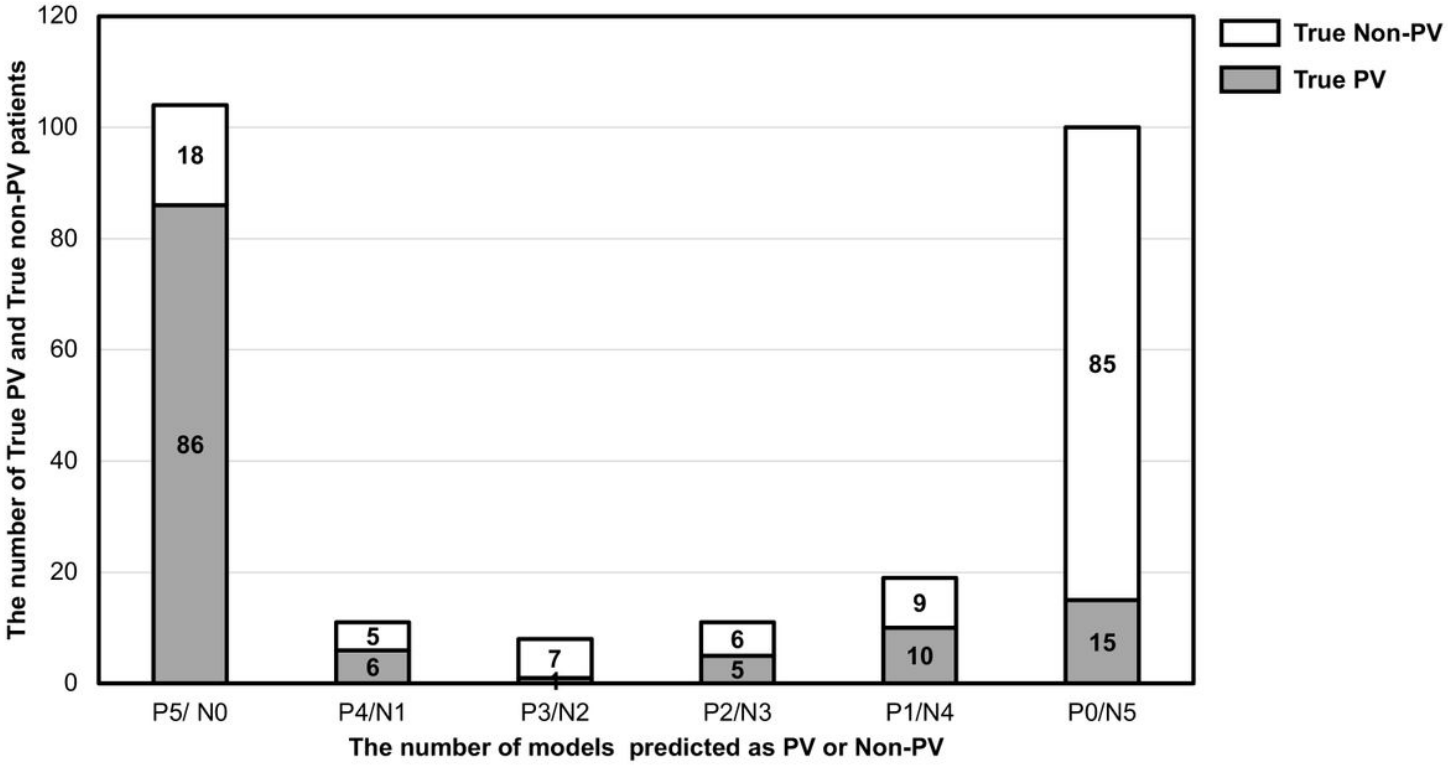

Figure 4

Evaluation of the accuracy of the five machine learning models in the testing dataset (n=252).
Abbreviations: PV, peripheral vestibular disease; Non-PV, non-peripheral vestibular disease.