Journal of Neurology (2024) 271:3426–3438
https://doi.org/10.1007/s00415-023-11997-4

ORIGINAL COMMUNICATION

Machine learning models help differentiate between causes of recurrent spontaneous vertigo

Chao Wang[1,2] · Allison S. Young[2] · Chahat Raj[3] · Andrew P. Bradshaw[1] · Benjamin Nham[1,2] · Sally M. Rosengren[1,2] · Zeljka Calic[4,5] · David Burke[1,2] · G. Michael Halmagyi[1,2] · Gnana K. Bharathy[3] · Mukesh Prasad[3] · Miriam S. Welgampola[1,2]

Received: 27 May 2023 / Revised: 10 September 2023 / Accepted: 13 September 2023 / Published online: 23 March 2024

Abstract

Background Vestibular migraine (VM) and Menière's disease (MD) are two common causes of recurrent spontaneous vertigo. Using history, video-nystagmography and audiovestibular tests, we developed machine learning models to separate these two disorders.

Methods We recruited patients with VM or MD from a neurology outpatient facility. One hundred features from six "feature subsets": history, acute video-nystagmography and four laboratory tests (video head impulse test, vestibular-evoked myogenic potentials, caloric testing and audiogram) were used. We applied ten machine learning algorithms to develop classification models. Modelling was performed using three "tiers" of data availability to simulate three clinical settings. "Tier 1" used all available data to simulate the neuro-otology clinic, "Tier 2" used only history, audiogram and caloric test data, representing the general neurology clinic, and "Tier 3" used history alone as occurs in primary care. Model performance was evaluated using tenfold cross-validation.

Results Data from 160 patients with VM and 114 with MD were used for model development. All models effectively separated the two disorders for all three tiers, with accuracies of 85.77–97.81%. The best performing algorithms (AdaBoost and Random Forest) yielded accuracies of 97.81% (95% CI 95.24–99.60), 94.53% (91.09–99.52%) and 92.34% (92.28–96.76%) for tiers 1, 2 and 3. The best feature subset combination was history, acute video-nystagmography, video head impulse test and caloric testing, and the best single feature subset was history.

Conclusions Machine learning models can accurately differentiate between VM and MD and are promising tools to assist diagnosis by medical practitioners with diverse levels of expertise and resources.

Keywords Artificial intelligence · Menière's disease · Vestibular migraine

Introduction

Vertigo is a false sensation of movement caused by disorders affecting the inner ear balance organs and their connections with the central nervous system. It is common, disabling and treatable, yet is undertreated worldwide. Recurrent vertigo that occurs at rest without provocation, also known as "recurrent spontaneous vertigo", is most often caused by one of two disorders: vestibular migraine (VM) or Menière's disease (MD) [8, 23]. VM is diagnosed based on association with migraine headaches and migraine-related symptoms such as photophobia, phonophobia, visual aura and motion sensitivity [17], while MD, which is attributed to excessive fluid accumulation in the endolymph compartment of the

✉ Miriam S. Welgampola
miriam@icn.usyd.edu.au

1 Institute of Clinical Neurosciences, Royal Prince Alfred Hospital, Sydney, Australia

2 Central Clinical School, University of Sydney, Sydney, Australia

3 School of Computer Science, Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, Australia

4 Department of Neurophysiology, Liverpool Hospital, Sydney, Australia

5 South Western Sydney Clinical School, University of New South Wales, Sydney, Australia

inner ear, is associated with fluctuating hearing loss, tinnitus and fullness affecting one ear [18].

As VM and MD are treated differently, the correct diagnosis is essential for optimal management. There is no definitive diagnostic test for either condition, so neuro-otologists separate VM and MD by seeking clues in the history, physical examination and laboratory tests of hearing and balance. For example, patients with MD report spinning vertigo that lasts 20 min to 12 h [18], whereas episodes of VM can last seconds to days [14, 24, 44]. Migraine symptoms such as headache and/or photophobia accompany vertigo in 95% of VM [19, 24], although they also occur in 29–45% of MD patients during their vertigo attacks [12, 24, 31]. Aural symptoms (hearing loss, tinnitus and fullness) commonly occur with vertigo in MD (51–83%) [19, 24], yet are also reported by 15–54% of patients with VM [14, 24, 44]. Examination of an asymptomatic patient is frequently unremarkable for both VM [29] and MD [43]. However, when assessed during vertigo attacks, patients with MD may demonstrate diagnostic eye movement abnormalities: typically, high-velocity spontaneous horizontal nystagmus, which may also spontaneously change direction within 12 h of vertigo onset [41]. In contrast, VM may demonstrate low-velocity nystagmus of diverse directions or no nystagmus [39]. Tests that assess the inner ear balance organs may demonstrate abnormalities in distinctive patterns that help identify the cause of vertigo. It is now possible to assess the function of all five vestibular end-organs (the three semicircular canals which sense rotation and the two otolith organs which sense linear acceleration) using laboratory tests [5, 20, 21, 35]. The caloric test, a method of comparing the integrity of the horizontal semicircular canals, shows abnormal results in up to 75% of patients with MD [24, 37] but only 25% of patients with VM [2, 14]. The cervical and ocular vestibular-evoked myogenic potentials (cVEMPs and oVEMPs) are tests of otolith function. In patients with MD, 38–45% of cVEMPs and 32–65% of oVEMPs are abnormal [13, 43], whereas both tests usually show normal results in VM [14, 37]. An audiogram demonstrating low-frequency hearing loss is very suggestive of MD [12], whereas patients with VM usually do not have hearing loss [2, 24]. The current diagnostic workflow used by neuro-otologists to distinguish between causes of recurrent spontaneous vertigo is summarised in Fig. 1.

We hypothesised that clinical information used by experts (structured history, nystagmus characteristics and vestibular function tests) could be used to train a machine learning algorithm to perform the classification task of a neuro-otologist. Previous investigators have used decision tree and neural network techniques to develop models for identifying VM or MD [11]. Their neural network models performed best, with accuracies of 98.4% for isolating VM from other vestibular disorders and 98.0% for isolating MD. However, these high accuracies likely reflect the fact that their models were designed to identify either VM or MD from a pool of several causes of dizziness, some of which have very different characteristics, rather than the harder task of distinguishing between two conditions which both often present as recurrent spontaneous vertigo and share clinical features.

In the present study, we limited our patients to just VM and MD, which can be difficult to separate clinically. The aim of this study was to develop and validate a machine learning model that could use the same information that is used by a neuro-otologist to classify recurrent spontaneous vertigo into VM or MD with a level of accuracy similar to an expert. The clinical diagnosis made by an experienced neuro-otologist with access to a full suite of laboratory tests was taken as the "gold standard" against which the performance of the developed machine learning models was compared. We also explored the performance of models developed using limited datasets simulating what would be available to general neurologists/otolaryngologists and primary care physicians. Our vision was to develop tools of assisted diagnosis usable by all healthcare practitioners who encounter patients with recurrent spontaneous vertigo.

Materials and methods

Participants

Adult patients who were seen in the neuro-otology outpatient clinic at Royal Prince Alfred Hospital from August 2014 to September 2021 for recurrent vertigo were consecutively recruited with informed consent. Patients were included if at initial or subsequent review they met Bárány Society diagnostic criteria for confirmed or probable VM [17] or MD [18] as determined by an experienced neuro-otologist (authors MW, GMH) and when ≥ 50% of the vestibular tests were undertaken. Patients who met criteria for both diagnoses were excluded. Data obtained during assessment across 100 variables from the categories of history, acute videonystagmography (VNG), and four laboratory tests were used for model development. The variables are detailed in Supplementary Table 1.

History

A standardised history generated 33 variables including disease length (years), vertigo trigger (spontaneous, positional or both), quality of vertigo (rotatory or non-rotatory, with or without imbalance), duration of shortest and longest vertigo attacks (seconds, minutes, hours, days), associated auditory symptoms (tinnitus, aural fullness, subjective hearing loss, fluctuations in hearing), associated migraine-related symptoms (headache, photophobia, phonophobia, visual aura) and cardiovascular risk factors (hyperlipidaemia,

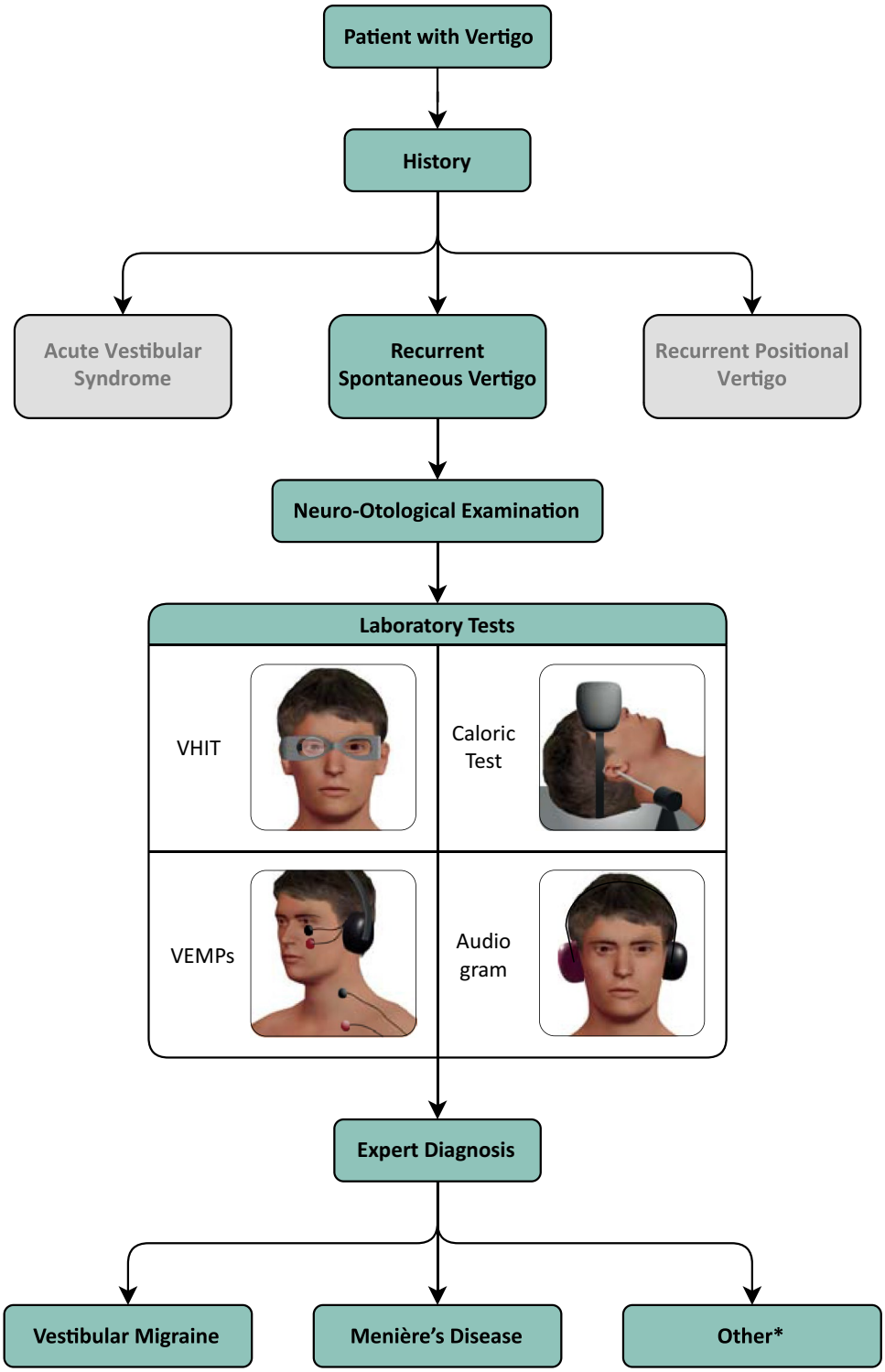


Fig. 1 Current diagnostic workflow used by neuro-otologists for patients with recurrent spontaneous vertigo. From the history, a medical practitioner can distinguish recurrent spontaneous vertigo from the other vertigo presentations of acute vestibular syndrome and recurrent positional vertigo. Video-nystagmography and laboratory tests are then performed, following which the most likely diagnosis is determined. *Other: this includes rarer causes of recurrent spontaneous vertigo such as posterior circulation ischaemia, autoimmune inner ear disease and vestibular paroxysmia. *VEMPs* vestibular-evoked myogenic potentials, *VHIT* video head impulse test

hypertension, diabetes, atrial fibrillation, family history of vascular disease).

Video-nystagmography

Miniature portable video glasses (Neuromed Electronics, Sydney, Australia) were used by patients to self-record any nystagmus present during a vertigo attack; this technique of home-recording was recently developed and validated

[41]. Nystagmus was recorded in the upright, supine and side-lying positions. Variables included nystagmus direction (horizontal, vertical or absent), whether the nystagmus changed direction spontaneously and the nystagmus slow-phase velocity (SPV) in degrees per second. When multiple video records were made, the fastest SPV was used.

Laboratory tests of inner ear hearing and vestibular function

All patients were offered all four tests chosen by neuro-otologists seeking to assess the hearing and vestibular organs: the caloric test and video head impulse test (VHIT) for horizontal semicircular canal function [21], vestibular-evoked myogenic potentials (VEMPs) to assess the otolith organs (oVEMP for the utricle and cVEMP for the saccule) [5, 35] and pure-tone audiogram to test cochlear function. The laboratory tests were performed when patients were asymptomatic. Figure 2a illustrates the end-organ assessed by each test, together with an exemplative test result.

VHITs were performed on the horizontal canals using ICS Impulse USB goggles (Natus, CA, USA) and analysed with custom software using previously described methods [42, 43]. Variables generated from VHIT included the gain of the vestibulo-ocular reflex, saccade frequency (%), total saccade displacement (degrees), first saccade displacement (degrees), first saccade peak velocity (degrees per second), first saccade onset time (ms) and first saccade duration (ms).

Both oVEMPs and cVEMPs were recorded using a Natus Medelec Synergy device (version 20.0, CA, USA) in response to both air-conducted (AC) clicks and bone-conducted (BC) forehead taps using previously described methods [42, 43]. For AC and BC oVEMPs, data were collected on reflex peak-to-peak amplitude (μV), n1 latency (ms) and the asymmetry ratio between left and right ears as calculated using Jongkees' formula as shown in Eq. (1). For AC and BC cVEMPs, data were collected on corrected reflex amplitude (ratio of peak-to-peak amplitude to baseline sternocleidomastoid activation), p13 latency (ms) and the asymmetry ratio.

$$\text{Asymmetry ratio} = \frac{\text{Left ear amplitude} - \text{right ear amplitude}}{\text{Left ear amplitude} + \text{right ear amplitude}} \times 100. \quad (1)$$

Pure tone audiometry was performed with air-conduction and bone-conduction transducers. Those with an air–bone gap of ≥ 15 dB HL (decibels hearing level) at any frequency were excluded. Air-conduction thresholds were recorded for the 250 Hz, 500 Hz, 1000 Hz, 2000 Hz, 4000 Hz, 6000 Hz and 8000 Hz frequencies. Bithermal caloric testing was performed using widely used methods with cold (30°) and hot (44°) water for 25–40 s. Nystagmus SPV in response to cold and warm water stimulation and left–right asymmetry (absolute canal paresis and absolute directional preponderance) was calculated using Jongkees' formula [15].

Machine learning modelling

We employed an iterative process for developing and evaluating the machine learning models, starting from problem formulation, through to data acquisition, pre-processing, modelling, validation and evaluation of the proof-of-concept models.

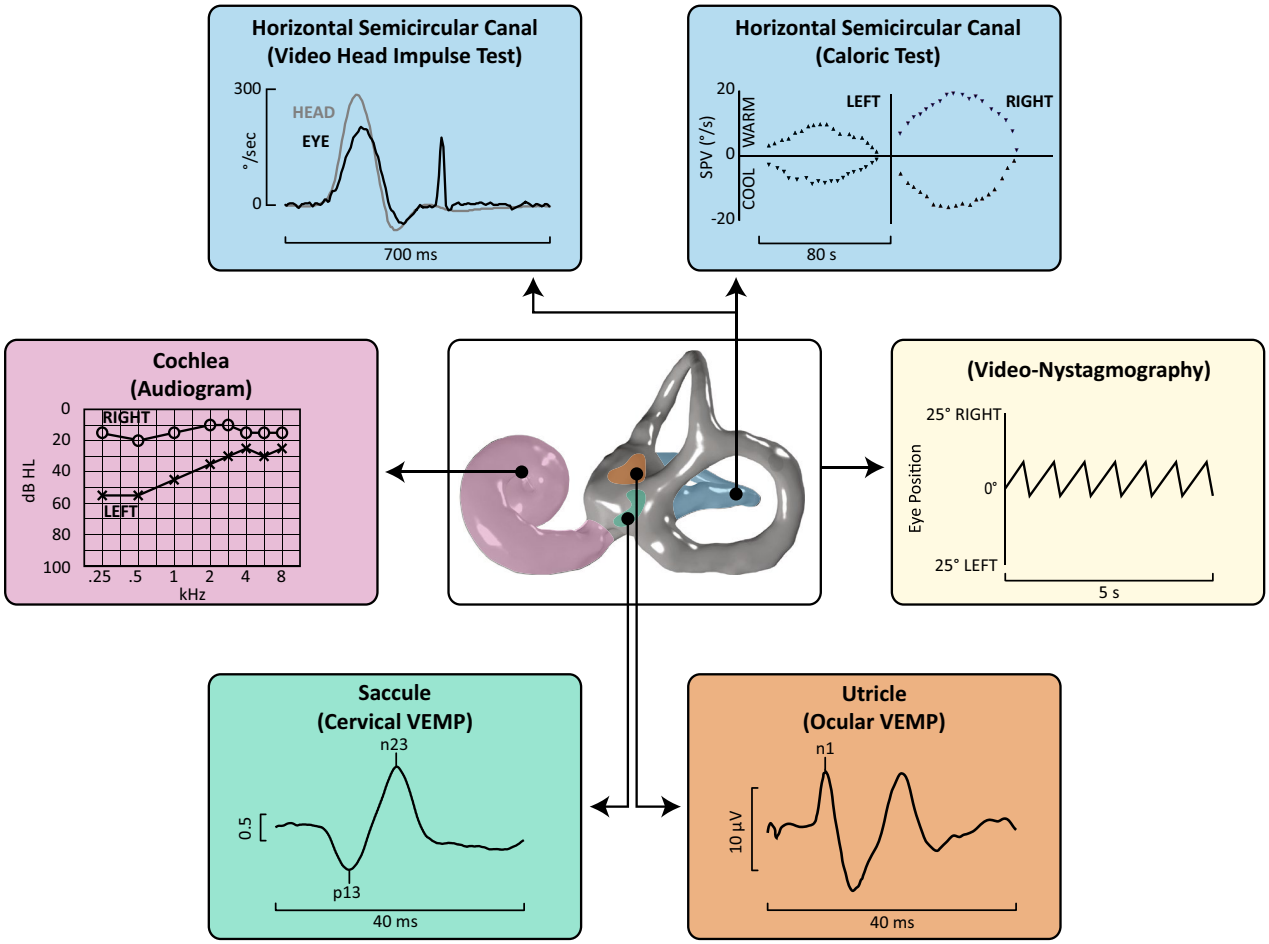
Problem formulation

We formulated the problem as determining the disease class, namely VM or MD, using machine learning techniques on all the variables (termed "features" in machine learning) arranged into six categories ("feature subsets"): history, VNG, VHIT, VEMPs, audiogram and caloric testing. We applied machine learning to three combinations, or "tiers", of feature subsets. The first combination ("Tier 1") included all the feature subsets and simulated the data obtainable in a neuro-otology clinic. However, as not all tests are available to non-expert specialists (general neurologists or otolaryngologists) or primary care physicians, we also used two restricted feature subset combinations. One of these ("Tier 2") simulated the non-expert specialist's clinic by limiting the available feature subsets to history as well as the widely available tests of audiogram and caloric testing. The other ("Tier 3") used only features from the history, which is available to all healthcare practitioners including the primary care setting. The feature subsets used in each tier are summarised in Fig. 2b.

Model development, validation and evaluation

Model development, validation and evaluation were performed using the scikit learn Python machine learning library version 0.24.2 [27]. For each of the three simulated clinical settings, we developed ten models by applying each of ten machine learning classification algorithms (XGBoost [4], logistic regression [40], K-nearest neighbour [28], decision tree [30], random forest [3], passive aggressive classifier [7], support vector machine [6], multilayer perceptron [34], gradient boosting classifier [10] and AdaBoost classifier [36]) to the corresponding data tier. For model validation, we employed tenfold stratified cross-validation. This method splits the data into ten subsets, uses nine for model training and one for testing, and then repeats the process nine more times by setting aside a different subset as the testing set each time. This technique provides more robust and reliable results than testing against a single dataset for validation. We used the same seed value of 42 to allow for reproducibility and comparison between algorithms. Metrics of

a



b

Tier	Clinical Setting	History	Audiogram	Caloric Test	VNG	VHIT	VEMPs
Tier 1	Neuro-Otology Clinic						
Tier 2	General Neurology Clinic						
Tier 3	Primary Care						

Fig. 2 Tests of the hearing and vestibular end-organs of the inner ear used for model development. **a** The coloured areas of the central inner ear diagram represent the end-organs as labelled with their corresponding test(s) in brackets. A sample result for each test is shown. The horizontal semicircular canal is assessed by the horizontal video head impulse test and the caloric test. Video-nystagmography can be done by the patient at home during a vertigo attack and also does not necessarily correlate with a specific end-organ. The saccule and utricle are the otolith organs and are tested by the cervical VEMP and ocular VEMP, respectively. For these, muscle activity is recorded in response to an acoustic or vibratory stimulus, and the amplitude and latencies of the responses are calculated. **b** This image shows which tests were used for model development in each of the three different tiers of data availability that simulated different clinical settings. *dB HL* decibels hearing level, *SPV* slow-phase velocity, *VEMP* vestibular-evoked myogenic potential, *VHIT* video head impulse test, *VNG* video-nystagmography

model performance were obtained from the cross-validation results as per the statistical analysis section below. Hyperparameter tuning was not carried out since satisfactory results were achieved using base algorithm parameters. See Supplementary Methods for further details of the machine learning methodology including pre-processing.

Using the above methods, we developed proof-of-concept models which can be further refined into tools

to assist medical practitioners. Figure 3 summarises the workflow we used to design our models.

Statistical analysis

Model performance was evaluated using the metrics ("performance metrics") of accuracy, precision, sensitivity, weighted F1-score, specificity and the area under the curve (AUC) from the receiver operating characteristic (ROC) curve. For calculation purposes, VM was defined as the positive diagnosis. Our two classes of VM and MD had relatively balanced representation in the dataset, and there was no preference between the two classes, meaning that false negatives and false positives were equally important. Thus, we employed accuracy as the primary metric to evaluate model performance. However, as the classes were not perfectly balanced (42% to 58%), we also considered the weighted F1-score as an alternative metric which takes in account this class imbalance. The mean value and 95% confidence interval for accuracy and the mean values for the other performance metrics were calculated by treating the ten iterations generated by cross-validation as the sample set.

Results

Patient characteristics

We collected data from 274 patients, of whom 160 (58.4%) had VM and 114 (41.6%) had MD. Select characteristics of each patient group are shown in Table 1, and the full list of characteristics is available in Supplementary Table 2. For both diagnoses, we did not make a distinction between patients with probable versus confirmed disease.

Machine learning model performance

Model performance for each tier

Tier 1 (the neuro-otology clinic) was simulated by models which used all the feature subsets (history, VNG and all four laboratory tests). Most algorithms generated models with accuracies above 95% (Table 2). The best performing model used AdaBoost and achieved 97.81% accuracy.

In Tier 2 (the non-expert specialist clinic), the models were developed only using the feature subsets of history and the two widely available tests (audiogram and caloric testing). Random forest was the best performing algorithm with 94.53% accuracy.

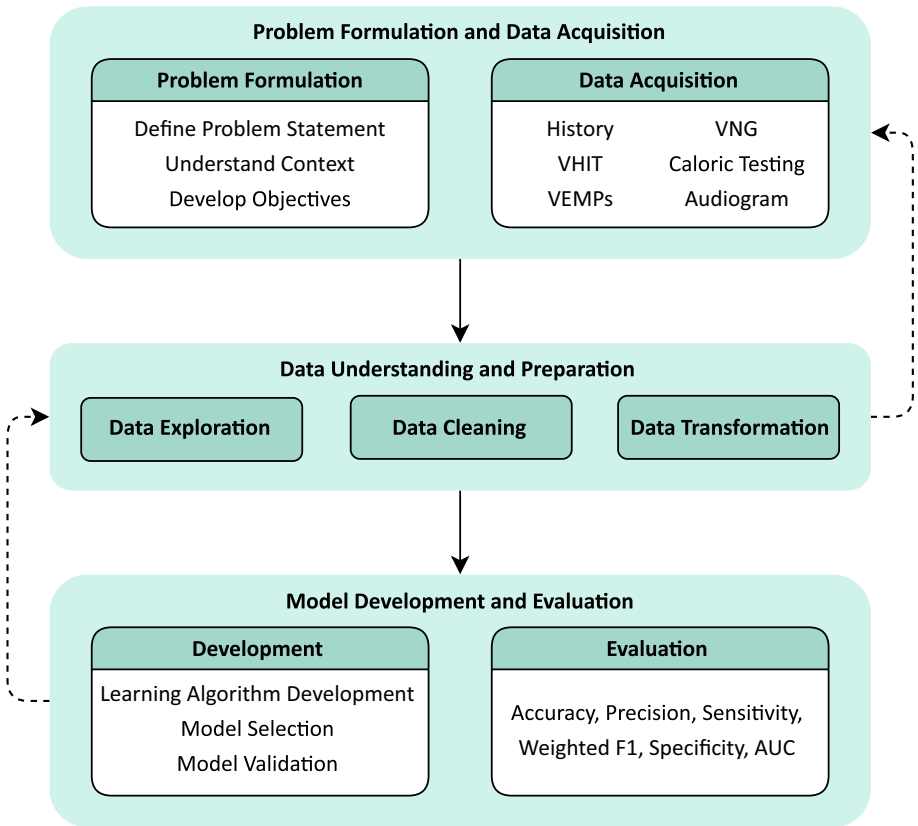


Fig. 3 Workflow of the Machine Learning Methodology. This figure illustrates the process we used to develop our machine learning models for differentiating vestibular migraine from Menière's disease. The dotted arrows indicate that this was an iterative process. *AUC* area under the curve, *VEMPs* vestibular-evoked myogenic potentials, *VHIT* video head impulse test, *VNG* video-nystagmography

Table 1 Select characteristics of patients

	Menière's disease ($n=114$)	Vestibular migraine ($n=160$)
History		
Sex, female, *n* (%)	55 (48.2)	108 (67.5)
Age at clinical presentation, years, median (IQR)	61.0 (47.5–70.0)	48.0 (38.0–59.0)
Disease duration, years, median (IQR)	3.0 (1.0–7.0)	3.0 (1.0–9.3)
Vertigo trigger, *n* (%)		
Spontaneous	85 (74.6)	70 (43.8)
Positional	0 (0)	11 (6.9)
Both	29 (25.4)	79 (49.4)
Longest duration of attacks, *n* (%)		
Seconds	0 (0)	4 (2.9)
Minutes	4 (3.7)	17 (12.2)
Hours	88 (80.7)	34 (24.5)
Days	17 (15.6)	59 (42.4)
Weeks	0 (0)	10 (7.2)
Constant	0 (0)	15 (10.8)
Tinnitus, *n* (%)		
Unilateral	93 (89.4)	26 (18.2)
Bilateral	5 (4.8)	41 (28.7)
Aural fullness, *n* (%)		
Unilateral	77 (81.9)	18 (12.6)
Bilateral	3 (3.2)	22 (15.4)
Subjective hearing loss, *n* (%)		
Unilateral	85 (90.4)	7 (5.4)
Bilateral	6 (6.4)	13 (10.1)
Headache, *n* (%)	50 (56.8)	123 (83.1)
Video-Nystagmography		
Spontaneous nystagmus, *n* (%)		
Horizontal	66 (94.3)	53 (47.3)
Vertical	4 (5.7)	29 (25.9)
Direction-changing spontaneous nystagmus, *n* (%)		
Occurring at < 12 h	27 (38.6)	5 (4.5)
Occurring at > 12 h	15 (21.4)	0 (0)
Spontaneous slow-phase velocity, °/s, median (IQR)	34.0 (21.7–54.8)	3.0 (0.0–7.0)
Video Head Impulse Test		
Right horizontal canal gain (raw), median (IQR)	1.00 (0.95–1.05)	0.99 (0.93–1.06)
Left horizontal canal gain (raw), median (IQR)	0.93 (0.87–0.99)	0.91 (0.87–0.99)
Vestibular-evoked myogenic potentials		
Air-conducted cVEMP asymmetry, median (IQR)[a]	20.5 (9.0–100.0)	12.2 (5.0–20.2)
Bone-conducted oVEMP asymmetry, median (IQR)[a]	13.8 (7.6–21.0)	8.3 (4.4–13.1)
Audiogram		
Affected Ear 250 Hz threshold, dB HL, median (IQR)[b]	55 (40–65)	–
Unaffected Ear 250 Hz threshold, dB HL, median (IQR)[b]	15 (10–20)	–
Right ear 250 Hz threshold, dB HL, median (IQR)	–	10 (5–15)
Left ear 250 Hz threshold, dB HL, median (IQR)	–	10 (5–15)
Affected ear 500 Hz threshold, dB HL, median (IQR)[b]	55 (40–60)	–
Unaffected ear 500 Hz threshold, dB HL, median (IQR)[b]	13 (10–20)	–
Right ear 500 Hz threshold, dB HL, median (IQR)	–	10 (5–15)
Left ear 500 Hz threshold, dB HL, median (IQR)	–	10 (5–15)

Table 1 (continued)

	Menière's disease ($n=114$)	Vestibular migraine ($n=160$)
Caloric testing		
Caloric canal paresis, %, median (IQR)[a]	39 (21–59)	11 (6–21)

See Supplementary Table 1 for variable descriptions, and Supplementary Table 2 for the full list of characteristics

cVEMP cervical vestibular-evoked myogenic potentials, *dB HL* decibels hearing level, *oVEMP* ocular vestibular-evoked myogenic potentials

[a]Absolute values used to calculate these summary statisticsn

[b]Affected and unaffected ear values shown for illustrative purposes only; affected ear was not labelled for model development

Table 2 Performance metrics of machine learning models for differentiating between vestibular migraine and Menière's disease

Algorithm Used	Accuracy	Precision	Sensitivity	Weighted F1	Specificity	AUC
Tier 1: All data used for model development						
XGB	96.72 (95.24–99.60)	97.30	94.74	96.71	98.13	96.43
LR	97.08 (95.41–98.74)	97.32	95.61	97.08	98.13	96.87
KNN	92.70 (92.25–98.28)	91.96	90.35	92.69	94.38	92.36
DT	89.42 (82.62–95.56)	90.48	83.33	89.34	93.75	88.54
RF	96.35 (92.57–100.00)	96.43	94.74	96.35	97.50	96.12
PAC	97.45 (95.39–98.73)	97.35	96.49	97.44	98.13	97.31
SVM	95.62 (94.57–98.10)	98.11	91.23	95.59	98.75	94.99
MLP	95.62 (95.96–99.65)	93.22	96.49	95.63	95.00	94.99
GBC	91.61 (92.44–98.83)	95.05	84.21	91.52	96.88	89.79
ABC	97.81 (95.24–99.60)	97.37	97.37	97.81	98.13	97.75
Tier 2: Only history, audiogram and caloric testing used for model development						
XGB	93.43 (91.90–97.97)	92.86	91.23	93.42	95.00	93.11
LR	93.43 (92.38–97.43)	92.11	92.11	93.43	94.38	93.24
KNN	89.42 (89.65–96.41)	86.32	88.60	89.43	90.00	89.30
DT	89.78 (84.16–93.25)	89.09	85.96	89.75	92.50	89.23
RF	94.53 (91.09–99.52)	93.81	92.98	94.52	95.63	94.30
PAC	92.34 (90.73–96.86)	91.15	90.35	92.33	93.75	92.05
SVM	93.43 (91.14–97.19)	92.86	91.23	93.42	95.00	93.11
MLP	91.24 (92.31–96.76)	89.47	89.47	91.24	92.50	93.11
GBC	91.97 (89.61–96.64)	93.40	86.84	91.92	95.63	91.67
ABC	93.43 (90.09–96.86)	92.11	92.11	93.43	94.38	93.24
Tier 3: Only history used for model development						
XGB	90.51 (90.72–96.16)	90.00	86.84	90.48	93.13	89.98
LR	90.15 (90.76–96.14)	89.91	85.96	90.11	93.13	89.54
KNN	85.77 (85.95–94.23)	84.40	80.70	85.72	89.38	85.04
DT	89.78 (85.05–92.41)	89.09	85.96	89.75	92.50	89.23
RF	92.34 (92.28–96.76)	94.29	86.84	92.28	96.25	91.55
PAC	90.88 (89.52–93.74)	90.09	87.72	90.86	93.13	90.42
SVM	91.97 (91.98–96.36)	91.82	88.60	91.95	94.38	91.49
MLP	89.05 (89.90–95.49)	88.18	85.09	89.02	91.88	91.49
GBC	89.42 (90.56–94.84)	89.72	84.21	89.36	93.13	89.11
ABC	89.42 (89.45–93.76)	88.99	85.09	89.38	92.50	88.79

Values are %, with 95% confidence interval in parentheses for accuracy

AUC area under the curve, *XGB* XGBoost, *LR* logistic regression, *KNN* K-nearest neighbour, *DT* decision tree, *RF* random forest, *PAC* passive aggressive classifier, *SVM* support vector machine, *MLP* multilayer perceptron, *GBC* gradient boosting classifier, *ABC* AdaBoost classifier

Models in Tier 3 (the primary care setting) used only features from the history, which is available to all healthcare practitioners. In this setting, the best performer was random forest with 92.34% accuracy.

Table 2 shows the complete performance metrics for all models in all three tiers. Accuracies ranged from 85.77 to 97.81%, indicating that all algorithms performed well across all tiers. Although accuracy was our primary metric, the models performed consistently well across the full range of performance metrics. The algorithms which achieved the highest accuracy in each tier also had the highest weighted F1-score and AUC values in that tier, indicative of robust model performance.

Figure 4 shows the ROC curve and corresponding confusion matrix of the model generated by the top-performing algorithm for each of the three clinical settings. The ROC curves shown were the mean ROC curves calculated from tenfold stratified cross-validation. All three mean ROCs passed close to the upper left corner, which indicated that the machine learning models were robust and were close in performance to an ideal classifier. The best ROC curve in Fig. 4 was obtained by the model which used AdaBoost on all the features in the dataset.

When comparing the algorithms across all three simulated clinical settings, we see that the top algorithms (AdaBoost and random forest) did well across all six performance metrics in all tiers.

Comparison of history, VNG and vestibular function tests in isolation

We also assessed how each of our six feature subsets performed individually at classifying VM and MD. We did this by taking AdaBoost as the preferred algorithm and compared the accuracies of models developed using AdaBoost on datasets limited to the features from that subset. The history was the best performing feature subset for distinguishing between VM and MD with an accuracy of 89.42%, followed by the audiogram (accuracy 87.23%). Next, acute VNG and caloric testing were similarly useful with accuracies of 68.61% and 67.15%, respectively. VEMPs and VHIT performed less well with accuracies of 59.85% and 51.82%, respectively.

Best performing combinations

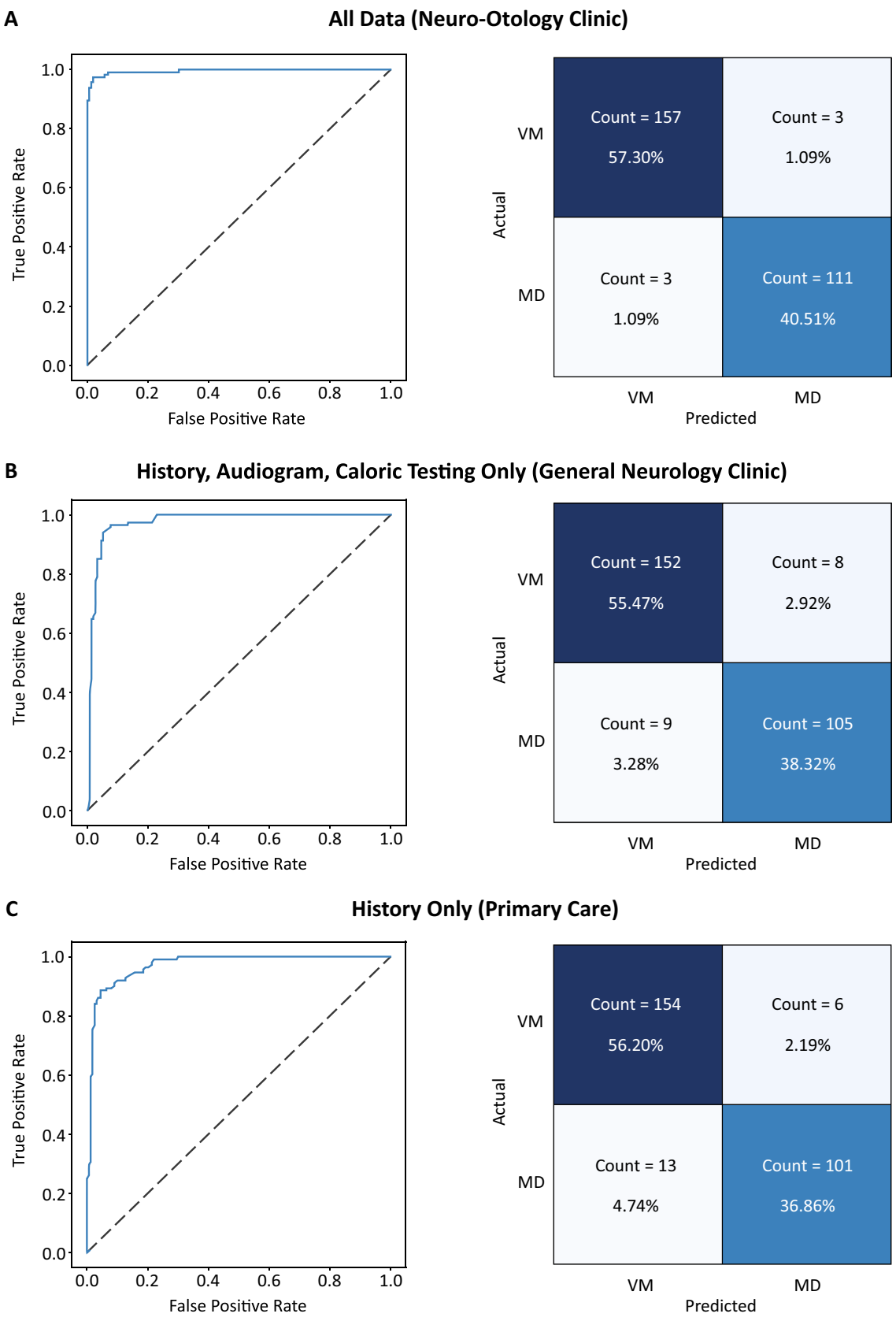
We used AdaBoost to develop models using every combination of the six feature subsets (history, VNG, VHIT, VEMPs, audiogram and caloric testing). In total, there were 63 (i.e. 2^6 minus 1) combinations, and the performance of these models is shown in Supplementary Table 3. Interestingly, one combination (history, VNG, VHIT and caloric testing) had a slightly higher accuracy (98.18%) than the model which used all the data (97.81%), and another combination (history and VNG) had the same accuracy. Sixteen combinations had an accuracy of 95% or more, and history and VNG were common to all of these combinations. When VNG was not available, then the best combination which used no more than two laboratory tests (in addition to the history) was history, audiogram and caloric testing (93.80%). Most combinations (36 out of 63) reached an accuracy of 90% or more.

Discussion

In this study, we developed machine learning models for classifying patients with recurrent vertigo as either VM or MD based on information from history, VNG and four laboratory tests. In real-world clinical practice, a neuro-otologist may have access to all the laboratory tests, but the non-expert specialist may only be able to access some of these tests, and the primary care physician may rely on the history alone. To simulate these conditions, we applied machine learning techniques to limited datasets. For the neuro-otologist setting, the top model which used all the features performed excellently with 97.81% accuracy and unsurprisingly was the best performing model overall. For the non-expert specialist setting, limiting the features to history, audiogram and caloric testing still produced a model which performed very well with 94.53% accuracy. Even when only features from history were used, the top model still achieved an accuracy of 92.34%. These results hold great promise for an accurate classification tool for recurrent spontaneous vertigo usable by all healthcare practitioners, including the primary care and rural settings where VNG and laboratory tests are often not available. Such a tool would also be valuable for clinicians who lack experience applying the diagnostic criteria or who do not have all the information that is required.

Comparison with earlier studies

Unlike other studies of machine learning in vertigo syndromes [16], we limited our study to two similar conditions that often present with recurrent spontaneous vertigo. This syndrome is straightforward to identify even for non-specialist doctors and allows the diagnosis to be narrowed to the two common culprits of VM and MD. An earlier study [11] also applied machine learning techniques to identify VM and MD as causes of recurrent spontaneous vertigo. They used data from DizzyReg, a registry of information from patients with dizziness, including demographics, history, physical assessment, test results and treatment. Models using boosted decision trees achieved accuracies of 84.5% and 93.3% for identifying VM and MD, respectively, while deep neural network models reached superior accuracies of 98.4% and 98.0%. However, their models were developed


A
All Data (Neuro-Otology Clinic)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Count = 157
57.30%
Count = 3
1.09%
Count = 3
1.09%
Count = 111
40.51%
Actual
VM
MD
Predicted
B
History, Audiogram, Caloric Testing Only (General Neurology Clinic)
Count = 152
55.47%
Count = 8
2.92%
Count = 9
3.28%
Count = 105
38.32%
C
History Only (Primary Care)
Count = 154
56.20%
Count = 6
2.19%
Count = 13
4.74%
Count = 101
36.86%

◂**Fig. 4** Receiver operating characteristic (ROC) curve and confusion matrices of the best performing classification model in each simulated clinical setting. **a** ROC curve and confusion matrix of model generated by AdaBoost using features from all data, simulating the neuro-otology clinic. **b** ROC curve and confusion matrix of model generated by random forest using only features from history, audiogram and caloric testing, simulating the non-expert specialist clinic. **c** ROC curve and confusion matrix of model generated by random forest using only features from history, simulating primary care. *MD* Menière's disease, *VM* vestibular migraine

only to identify either VM or MD from all the other diagnoses in the registry, which included conditions such as benign paroxysmal positional vertigo and vestibular failure that do not present as recurrent spontaneous vertigo and hence may have been easier to distinguish. Another study using the EMBalance decision support system [9] took a similar approach of developing binary classification models to identify each of 12 differential diagnoses (including VM and MD) from all other diagnoses in a dataset of patients with balance disorders. Like our study, in addition to using all the variables for model development (simulating the "expert"), they also used a limited dataset to mimic the primary care setting. This dataset included information from the history and excluded laboratory test results, but in contrast to our study also utilised findings from bedside examination, which untrained primary care physicians may not be able to confidently perform. Models using their best performing algorithm (AdaBoost) achieved accuracies of 92.1% (expert) and 89.8% (primary care) for identifying MD and an accuracy of 82.9% (both expert and primary care models) for VM. The results of both these studies do not reflect their models' ability to separate VM and MD, and this was not assessed. In contrast, our models differentiated only between VM and MD which share several overlapping features and where diagnostic confusion is likely to be the greatest. The greater difficulty of separating conditions with similar presentations is highlighted by a subsequent study using the DizzyReg registry [38]. This study used machine learning models to distinguish only between four causes of recurrent vertigo (VM, MD, benign paroxysmal positional vertigo and vestibular paroxysmia). Although multi-class classification is more challenging, it is still notable that the highest accuracy achieved by the many algorithms trialled was only 54.3%.

Another point of difference of our study from previous research (other than the aforementioned EMBalance study [9]) is our tiered approach to the data. This allowed for the development of different models that cater to doctors of varying expertise and equipment, and of immense value is the model that uses only history and so can be utilised by any healthcare practitioner, including in the primary care setting. This is important as vertigo is common (4.9% one-year prevalence in adults [25]) but patients often have limited access to neurologists and otolaryngologists, and neuro-otologists are even less available. As an example, from Australian data in 2021 [1, 22], there is approximately one primary care physician per 740 people, one neurologist or otolaryngologist per 19,000 people, and one neuro-otologist per 1,270,000 people.

The best performing model

Regarding which algorithm performed best, AdaBoost had the best accuracy in Tier 1, but random forest was the top performer in Tiers 2 and 3. However, it is also apparent that all ten machine learning algorithms performed very well across all three tiers of data availability. The fact that all the algorithms could effectively distinguish between the two conditions suggests that machine learning techniques are a highly compatible approach to this clinical problem.

Based on the results from individual feature subsets, we consider the focused history, which yielded the highest accuracy by itself, to be indispensable when classifying VM and MD. The history was also present in all the top-performing feature combinations, as was acute VNG. Furthermore, the AdaBoost model used only history and VNG achieved the same accuracy as when all the data were used. This particular result may just have been due to the characteristics of this algorithm and/or the dataset, but nevertheless these findings emphasise how helpful VNG can be for distinguishing between VM and MD. On the other hand, VHIT and VEMPs did not separate VM and MD effectively when used in isolation, although they appeared to offer more diagnostic value when taken in combination with the other tests. It is important to note that the click stimuli we use in our clinic for VEMPs are not as effective as the 500 Hz tone bursts at separating MD from normal patients [33]; the use of tone bursts may have yielded better results for VEMPs in our models. We also appreciate that most clinics do not have access to portable VNG for patients to record attacks at home when symptomatic, so based on our results, if acute VNG is not available, then we recommend audiogram and caloric testing as the optimal laboratory tests to complement the history. An interesting result was that the AdaBoost model which omitted VEMPs and audiogram and used only history, VNG, VHIT and caloric testing had a slightly higher accuracy than when all features were used. This difference was small (0.37%), and again this result may be particular to the specific algorithm and/or dataset.

Limitations

A limitation of this study is that it was conducted at a single site with a limited number of patients and would benefit from expansion to other sites in future validation studies to demonstrate robust performance of our models. In addition, our classification models did not allow for patients who

have both VM and MD, as 11–28% of patients will technically meet diagnostic criteria for both conditions [24, 32]. Our models also do not identify other causes of recurrent spontaneous vertigo, such as autoimmune inner ear disease, posterior circulation ischaemia and vestibular paroxysmia. However, these are much rarer than VM and MD [26]. One reason our models performed so well is likely to be the fact that all our patients met Bárány Society diagnostic criteria for VM [17] and MD [18], which requires multiple vertigo episodes, and thus they were likely to have a higher prevalence of the typical abnormalities. The models may not perform as well on patients with early disease and who have more limited information on history. Future iterations will build on this. Our history was also taken by neuro-otologists, and it is possible that a history taken by a less experienced clinician in a non-specialist setting may be less accurate and may not differentiate between the two conditions as effectively when our model is applied. However, the structured nature of our history reduces this variance due to the fixed list of questions and responses to choose from.

Summary

In conclusion, we have demonstrated that machine learning algorithms can effectively distinguish between VM and MD as the cause of recurrent spontaneous vertigo using data from history, acute video-nystagmography, and laboratory tests of hearing and vestibular function. Our models performed well in both expert and non-expert settings and are likely to assist most medical practitioners faced with this problem.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00415-023-11997-4.

Funding This work was supported by funding from the Garnett Passe and Rodney Williams Memorial Foundation (Grant 2021_RS_Wang).

Data availability The dataset used for this study contains patient health data and is not publicly available for privacy reasons. A deidentified version of dataset is available on reasonable request from the corresponding author M.W. The data will be shared through a data sharing agreement. With this mediated access, the data are FAIR compliant.

Declarations

Conflicts of interest The authors declare that they have no competing interests.

Ethics approval This study was approved by the Sydney Local Health District Ethics Committee (Protocol No X21-0295) and performed in accordance with the 1964 Declaration of Helsinki and its later amendments.

Informed consent Written informed consent was obtained from all participants.

References

1. Australian Bureau of Statistics (2021) Population: Census. https://www.abs.gov.au/statistics/people/population/population-census/latest-release. Accessed 14 Aug 2022
2. Baier B, Stieber N, Dieterich M (2009) Vestibular-evoked myogenic potentials in vestibular migraine. J Neurol 256:1447–1454. https://doi.org/10.1007/s00415-009-5132-4
3. Breiman L (2001) Random forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324
4. Chen T, Guestrin C (2016) XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery, San Francisco, California, USA, pp 785–794
5. Colebatch JG, Halmagyi GM, Skuse NF (1994) Myogenic potentials generated by a click-evoked vestibulocollic reflex. J Neurol Neurosurg Psychiatry 57:190–197. https://doi.org/10.1136/jnnp.57.2.190
6. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20:273–297. https://doi.org/10.1007/BF00994018
7. Crammer K, Dekel O, Keshet J, Shalev-Shwartz S, Singer Y (2006) Online passive-aggressive algorithms. J Mach Learn Res 7:551–585
8. Dieterich M, Obermann M, Celebisoy N (2016) Vestibular migraine: the most frequent entity of episodic vertigo. J Neurol 263:82–89. https://doi.org/10.1007/s00415-015-7905-2
9. Exarchos TP, Rigas G, Bibas A, Kikidis D, Nikitas C, Wuyts FL, Ihtijarevic B, Maes L, Cenciarini M, Maurer C, Macdonald N, Bamiou DE, Luxon L, Prasinos M, Spanoudakis G, Koutsouris DD, Fotiadis DI (2016) Mining balance disorders' data for the development of diagnostic decision support systems. Comput Biol Med 77:240–248. https://doi.org/10.1016/j.compbiomed.2016.08.016
10. Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29(1189–1232):1144
11. Groezinger M, Huppert D, Strobl R, Grill E (2020) Development and validation of a classification algorithm to diagnose and differentiate spontaneous episodic vertigo syndromes: results from the DizzyReg patient registry. J Neurol 267:160–167. https://doi.org/10.1007/s00415-020-10061-9
12. Gürkov R, Jerin C, Flatz W, Maxwell R (2019) Clinical manifestations of hydropic ear disease (Menière's). Eur Arch Otorhinolaryngol 276:27–40. https://doi.org/10.1007/s00405-018-5157-3
13. Huang CH, Wang SJ, Young YH (2011) Localization and prevalence of hydrops formation in Ménière's disease using a test battery. Audiol Neurootol 16:41–48. https://doi.org/10.1159/000312199
14. Inoue A, Egami N, Fujimoto C, Kinoshita M, Yamasoba T, Iwasaki S (2016) Vestibular evoked myogenic potentials in vestibular migraine: do they help differentiating from Menière's disease? Ann Otol Rhinol Laryngol 125:931–937. https://doi.org/10.1177/0003489416665192
15. Jongkees LB, Maas JP, Philipszoon AJ (1962) Clinical nystagmography. A detailed study of electro-nystagmography in 341 patients with vertigo. Pract Otorhinolaryngol (Basel) 24:65–93
16. Kabade V, Hooda R, Raj C, Awan Z, Young AS, Welgampola MS, Prasad M (2021) Machine learning techniques for differential diagnosis of vertigo and dizziness: a review. Sensors 21:7565
17. Lempert T, Olesen J, Furman J, Waterston J, Seemungal B, Carey J, Bisdorff A, Versino M, Evers S, Newman-Toker D (2012) Vestibular migraine: diagnostic criteria. J Vestib Res 22:167–172. https://doi.org/10.3233/VES-2012-0453
18. Lopez-Escamez JA, Carey J, Chung W-H, Goebel JA, Magnusson M, Mandalà M, Newman-Toker DE, Strupp M, Suzuki M,

Trabalzini F, Bisdorff A (2015) Diagnostic criteria for Menière's disease. J Vestib Res 25:1–7. https://doi.org/10.3233/VES-150549
19. Lopez-Escamez JA, Dlugaiczyk J, Jacobs J, Lempert T, Teggi R, von Brevern M, Bisdorff A (2014) Accompanying symptoms overlap during attacks in Menière's disease and vestibular migraine. Front Neurol. https://doi.org/10.3389/fneur.2014.00265
20. MacDougall HG, McGarvie LA, Halmagyi GM, Curthoys IS, Weber KP (2013) The video head impulse test (vHIT) detects vertical semicircular canal dysfunction. PLoS ONE 8:e61488. https://doi.org/10.1371/journal.pone.0061488
21. MacDougall HG, Weber KP, McGarvie LA, Halmagyi GM, Curthoys IS (2009) The video head impulse test: diagnostic accuracy in peripheral vestibulopathy. Neurology 73:1134–1141. https://doi.org/10.1212/WNL.0b013e3181bacf85
22. Medical Board of Australia (2021) Medical Board of Australia Registrant Data—reporting Period: 01 October 2021 to 31 December 2021. Melbourne, Australia. https://www.ahpra.gov.au/documents/default.aspx?record=WD22%2f31646&dbid=AP&chksum=ccbe7hvigrjZ4H64z42Aaw%3d%3d. Accessed 14 Aug 2022
23. Muelleman T, Shew M, Subbarayan R, Shum A, Sykes K, Staecker H, Lin J (2017) Epidemiology of Dizzy patient population in a neurotology clinic and predictors of peripheral etiology. Otol Neurotol 38:870–875. https://doi.org/10.1097/mao.0000000000001429
24. Neff BA, Staab JP, Eggers SD, Carlson ML, Schmitt WR, Van Abel KM, Worthington DK, Beatty CW, Driscoll CL, Shepard NT (2012) Auditory and vestibular symptoms and chronic subjective dizziness in patients with Ménière's disease, vestibular migraine, and Ménière's disease with concomitant vestibular migraine. Otol Neurotol 33:1235–1244. https://doi.org/10.1097/MAO.0b013e31825d644a
25. Neuhauser HK, von Brevern M, Radtke A, Lezius F, Feldmann M, Ziese T, Lempert T (2005) Epidemiology of vestibular vertigo. A neurotologic survey of the general population 65:898–904. https://doi.org/10.1212/01.wnl.0000175987.59991.3d
26. Nham B, Reid N, Bein K, Bradshaw A, McGarvie L, Argaet E, Young A, Watson S, Halmagyi G, Black D, Welgampola M (2021) Capturing vertigo in the emergency room: three tools to double the rate of diagnosis. J Neurol. https://doi.org/10.1007/s00415-021-10627-1
27. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830
28. Peterson LE (2009) K-nearest neighbor. Scholarpedia 4:1883. https://doi.org/10.4249/scholarpedia.188
29. Polensek SH, Tusa RJ (2010) Nystagmus during attacks of vestibular migraine: an aid in diagnosis. Audiol Neurotol 15:241–246. https://doi.org/10.1159/000255440
30. Quinlan JR (1996) Learning decision tree classifiers. ACM Comput Surv 28:71–72. https://doi.org/10.1145/234313.234346
31. Radtke A, Lempert T, Gresty MA, Brookes GB, Bronstein AM, Neuhauser H (2002) Migraine and Ménière's disease: is there a link? Neurology 59:1700–1704. https://doi.org/10.1212/01.wnl.0000036903.22461.39
32. Radtke A, Neuhauser H, von Brevern M, Hottenrott T, Lempert T (2011) Vestibular migraine–validity of clinical diagnostic criteria. Cephalalgia 31:906–913. https://doi.org/10.1177/0333102411405228
33. Rauch SD, Zhou G, Kujawa SG, Guinan JJ, Herrmann BS (2004) Vestibular evoked myogenic potentials show altered tuning in patients with Ménière's disease. Otol Neurotol 25:333–338. https://doi.org/10.1097/00129492-200405000-00022
34. Rosenblatt F (1961) Principles of neurodynamics. perceptrons and the theory of brain mechanisms. In:Cornell Aeronautical Lab Inc Buffalo NY
35. Rosengren SM, McAngus Todd NP, Colebatch JG (2005) Vestibular-evoked extraocular potentials produced by stimulation with bone-conducted sound. Clin Neurophysiol 116:1938–1948. https://doi.org/10.1016/j.clinph.2005.03.019
36. Schapire RE (2013) Explaining adaboost. Empirical inference. Springer, Berlin, pp 37–52
37. Taylor RL, Zagami AS, Gibson WPR, Black DA, Watson SRD, Halmagyi MG, Welgampola MS (2012) Vestibular evoked myogenic potentials to sound and vibration: characteristics in vestibular migraine that enable separation from Menière's disease. Cephalalgia 32:213–225. https://doi.org/10.1177/0333102411434166
38. Vivar G, Strobl R, Grill E, Navab N, Zwergal A, Ahmadi SA (2021) Using base-ml to learn classification of common vestibular disorders on DizzyReg Registry data. Front Neurol 12:681140. https://doi.org/10.3389/fneur.2021.681140
39. von Brevern M, Zeise D, Neuhauser H, Clarke AH, Lempert T (2004) Acute migrainous vertigo: clinical and oculographic findings. Brain 128:365–374. https://doi.org/10.1093/brain/awh351
40. Wright RE (1995) Logistic regression. Reading and understanding multivariate statistics. American Psychological Association, Washington, DC, pp 217–244
41. Young AS, Lechner C, Bradshaw AP, MacDougall HG, Black DA, Halmagyi GM, Welgampola MS (2019) Capturing acute vertigo. A vestibular event monitor 92:e2743–e2753. https://doi.org/10.1212/wnl.0000000000007644
42. Young AS, Nham B, Bradshaw AP, Calic Z, Pogson JM, D'Souza M, Halmagyi GM, Welgampola MS (2021) Clinical, oculographic, and vestibular test characteristics of vestibular migraine. Cephalalgia 41:1039–1052. https://doi.org/10.1177/0333102421 1006042
43. Young AS, Nham B, Bradshaw AP, Calic Z, Pogson JM, Gibson WP, Halmagyi GM, Welgampola MS (2021) Clinical, oculographic and vestibular test characteristics of Ménière's disease. J Neurol. https://doi.org/10.1007/s00415-021-10699-z
44. Zhang Y, Kong Q, Chen J, Li L, Wang D, Zhou J (2016) International Classification of Headache Disorders 3rd edition beta-based field testing of vestibular migraine in China: Demographic, clinical characteristics, audiometric findings and diagnosis statues. Cephalalgia 36:240–248. https://doi.org/10.1177/0333102415587704