

<https://doi.org/10.1038/s41746-025-01880-z>

Clinical decision support for vestibular diagnosis: large-scale machine learning with lived experience coaching



Cecilia A. Callejas Pastor^{1,2}, Hyun Tae Ryu¹, Jung Sook Joo¹, Yunseo Ku^{2,3}✉ & Myung-Whan Suh¹✉

Diagnosing vestibular disorders remains challenging due to complex symptoms and extensive history-taking required. While machine learning approaches have shown promise in medical diagnostics, their application to vestibular disorder classification has been limited. We developed a CatBoost machine learning model to classify six common vestibular disorders using a retrospective dataset of patients. The model incorporates 50 clinical features, selected through a hybrid approach combining algorithmic methods (RFE-SVM and SKB score) and expert clinical knowledge. We designed the system to achieve high sensitivity for common vestibular disorders (BPPV and VM) and high specificity for conditions requiring intensive interventions (MD and HOD) or careful differential diagnosis (PPPD and VEST) to minimize unnecessary invasive treatments. When applied to test data, reaches 88.4% accuracy, with 60.9% correct classifications, 27.5% partially correct, and 11.6% incorrect classifications. Results suggest that machine learning can support clinical decision-making in vestibular disorder diagnosis when combining algorithmic capabilities with clinical expertise.

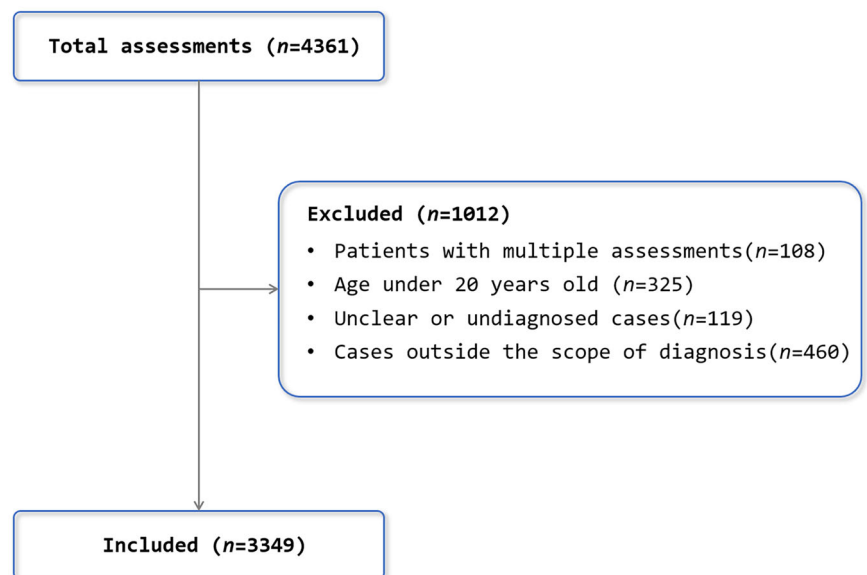
Diagnosing vestibular disorders presents a significant clinical challenge due to the extensive history-taking requirements and the complexity of organizing patient information into recognizable diagnostic patterns¹. The Barany Society's International Classification of Vestibular Disorders (ICVD) has established evidence-based consensus definitions for vestibular diseases, providing a standardized framework for diagnosis^{2,3}. While some vestibular disorders require physical examination or lab test results for the confirmative diagnosis, others are only based on the patients' descriptions on the vestibular symptoms. For instance, vestibular migraine (VM)⁴, probable Menière's disease (MD)⁵, probable benign paroxysmal positional vertigo (BPPV)⁶, vestibular paroxysmia⁷, mal de débarquement syndrome⁸, probable hemodynamic orthostatic dizziness (HOD)⁹, and persistent postural-perceptual dizziness (PPPD)¹⁰ can be diagnosed primarily based on patient-reported symptoms^{11,12}. Collecting and organizing the history of patients' symptoms may be very laborious, but is the most important first step in patient care with vestibular problems.

Clinicians must navigate numerous questions covering multiple disorders defined by the ICVD, potentially supplemented by additional inquiries specific to conditions such as migraine. For example, at least 4 core features (vertigo, hearing loss, ear fullness, and tinnitus) must be reviewed and meet the diagnostic criteria to make a diagnosis of MD. If the same

patient had concomitant headache, this number increases to 10 when including migraine-specific criteria. To rule out PPPD, an additional 7 key criteria must be evaluated, resulting in 21 features that need to be assessed for a comprehensive evaluation. An experienced vestibular specialist has a hierarchy of algorithms on whether to weigh some findings and to discard others to reach a rational conclusion. This thinking process requires not only memorization of all the diagnostic criteria of ICVD but also repeated training on organizing the complex information into a structured combination of features that makes the most sense. Machine Learning (ML) may have an advantage in simulating or enhancing this diagnostic thinking process. Considering the long training period of a vestibular specialist, this process may help to train beginners and save time for experienced specialists.

Recent advances in ML have demonstrated significant potential in improving the diagnosis of vestibular disorders. For instance, Vivar et al.¹³ developed the multivariable analyses and ML pipeline tool, which has shown promising results in classifying a range of vestibular conditions, including functional dizziness, MD, and vestibular migraine, using clinical datasets from the a large clinical patient registry ($N = 540$). This tool applies multiple ML algorithms to accurately classify complex vestibular disorders, with classification accuracies ranging from 50% to 92.5%, depending on the

¹Department of Otorhinolaryngology-Head and Neck Surgery, Seoul National University Hospital, Seoul, Republic of Korea. ²Department of Biomedical Engineering, College of Medicine, Chungnam National University, Daejeon, Republic of Korea. ³Medical Device Research Center, Department of Biomedical Research Institute, Chungnam National University Hospital, Daejeon, Republic of Korea. ✉e-mail: yunseo.ku@cnu.ac.kr; drmung@naver.com

Fig. 1 | Flow diagram of patients.

task. Similarly, Ahmadi et al.¹⁴ highlighted the utility of ML in differentiating between central and peripheral vestibular disorders, with models outperforming traditional diagnostic scores like HINTS (Head Impulse, gaze-evoked Nystagmus, Test of Skew) in acute settings. Wang et al.¹⁵ further demonstrated the power of ML in differentiating between vestibular migraine and MD, achieving high classification accuracies across different clinical settings. Agrimi et al.¹⁶ also emphasized the potential of artificial intelligence in enhancing diagnosis and differentiation of vestibular disorders, improving clinical decision-making, and ultimately optimizing patient management. Additionally, Strobl et al.¹⁷ focused on simplifying the diagnostic process using accessible data, such as patient history, which can be particularly helpful in primary care settings where diagnostic resources may be limited. These efforts underscore the growing importance of ML as a tool for not only enhancing diagnostic accuracy but also for streamlining the diagnostic process across different clinical environments. However, previous studies for vestibular diagnostics are often limited by the scope of disorders they addressed (2–5 disorders at most), and the number of patients analyzed (40–1066 cases)^{13–17}. There remains a pressing need for a larger-scale model capable of differentiating between a broader range of vestibular disorders using accessible clinical data, such as patient history and symptom reports.

To address these limitations, we adopted a hybrid feature selection method combining algorithmic analysis with clinical expertise to assess six common vestibular disorders: MD, BPPV, vestibulopathy (VEST), HOD, VM, and PPPD. This hybrid approach has shown promising results in various medical fields, such as primary progressive aphasia diagnosis or diabetes prediction^{18,19}, suggesting its potential for improving clinical decision-making in vestibular disorders.

A total of 4361 patients presenting with dizziness symptoms were initially enrolled at Seoul National University Hospital (SNUH) between June 2012 and May 2022. This study was reviewed and approved by the Institutional Review Board (IRB No. 2206-008-1328) of SNUH, Seoul, South Korea. All procedures performed in this study were in accordance with the ethical standards of the institutional research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all participants prior to their involvement in the study.

Following the application of exclusion criteria, 1012 cases were removed from the initial cohort. Exclusions were based on several predetermined criteria: (i) duplicate assessments from the same patient, in which case only the most recent evaluation was retained to ensure data independence; ii)

patients under 20 years of age were excluded as vestibular disorders in children and adolescents have distinct diagnostic criteria and clinical presentations compared to adults²⁰; (iii) cases with unclear or inconclusive diagnoses that could not be definitively classified, including those with significant missing data that precluded accurate diagnostic assessment; and (iv) patients whose final diagnoses fell outside the predefined diagnostic categories of interest for this study such as Sudden Sensorineural Hearing Loss (SSNHL), vestibular paroxysmia, or cervical dizziness. As shown in Fig. 1, the final analytical sample comprised 3349 participants (69.9% female; $n = 2340$), with a mean age of 56.42 years ($SD = 14.66$).

Vestibular specialists at SNUH conducted standardized medical assessments using a comprehensive 145-item history taking, based on the ICVD^{1,2}. The assessment protocol systematically evaluated patients' symptoms related to dizziness and headache, along with other clinical parameters necessary for diagnostic determination. Following the medical evaluation, vestibular specialists provided either single or dual diagnoses for each patient's dizziness symptoms, which served as the reference standard for this study. All assessments were conducted during in-person visits to ensure standardized data collection and diagnostic accuracy.

Results

Dataset and feature selection

The data distribution for each class in the dataset is visualized in Fig. 2.

From the two feature selection algorithms utilized (RFE-SVM and SKB score) and the clinical analysis performed by experts, Table 1 displays the 50 selected features. Among these, 30 features were selected purely based on the algorithmic results, and 20 features were chosen as a result of the expert clinical analysis were included.

All of the selected features were subsequently used in the ML process to train and evaluate the predictive models.

Model performance and selection

On the methods section, we described the three ML models employed in this study: CatBoost, Decision Trees, and XGBoost.

Table 2 summarizes the results obtained by the three ML models on the validation and testing set.

CatBoost was selected despite Random Forest's higher validation accuracy (98% vs 93%) due to better generalization on unseen data. CatBoost results are bolded in Table 2. Random Forest's larger accuracy drop (98% to 85%) indicated overfitting compared to CatBoost's stable performance (93% to 88%).

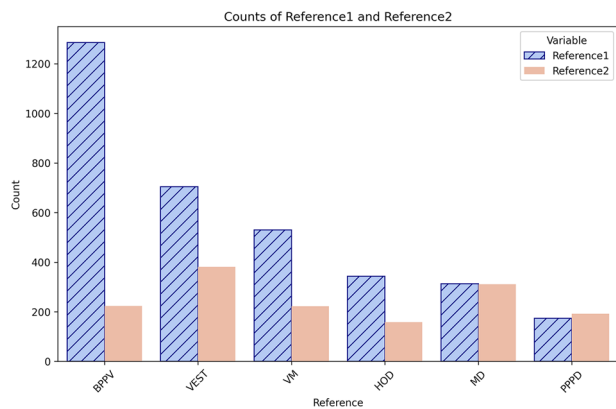


Fig. 2 | Barplot for counts of ref. 1 (First impression by vestibular specialists) and ref. 2 (second impression by vestibular specialists). BPPV benign paroxysmal positional vertigo, VEST vestibulopathy, VM vestibular migraine, HOD hemodynamic orthostatic dizziness, MD Meniere's disease, PPPD persistent postural-perceptual dizziness.

Classification results

The ML model, when applied to the unseen test session, yielded the following results: out of a total of 670 instances (100%), the model correctly predicted 408 instances (60.90%), partially correctly predicted 184 instances (27.46%), and incorrectly predicted 78 instances (11.64%).

The model's performance varied across the different classes, as shown in Fig. 3. For BPPV, the model achieved a moderate accuracy of 0.77, with a notably high sensitivity of 0.81, though specificity was lower at 0.75. The HOD class showed strong overall accuracy of 0.91 and excellent specificity of 0.97, though sensitivity was limited at 0.33. MD similarly demonstrated high accuracy of 0.91 and strong specificity of 0.96, while sensitivity remained moderate at 0.44. PPPD achieved the highest overall accuracy of 0.95 and near-perfect specificity of 0.99, though sensitivity was notably low at 0.09. The VEST class showed good performance with an accuracy of 0.82 and strong specificity of 0.90, while maintaining moderate sensitivity at 0.52. VM demonstrated well-balanced performance with an accuracy of 0.86, accompanied by good sensitivity of 0.70 and strong specificity of 0.89.

The inclusion of 'Partially Correct' classifications reflects clinical reality where differential diagnosis commonly involves multiple considerations, making these predictions clinically valuable for supporting diagnostic decision-making rather than creating diagnostic uncertainty.

Discussion

In this study, we developed an ML model for the classification of vestibular disorders using a large-scale clinical dataset. The model was trained on a diverse set of features and evaluated on an unseen test set to assess its performance and generalizability. The accuracy of our developed model in diagnosing vestibular disorders was generally high (>85%) and notably high (>90%) for MD, PPPD, and HOD. Our results showed promise to help clinicians as a screening and decision support tool in clinical settings rather than as a definitive diagnostic instrument. There can be at least 3 user scenarios for our developed model. First, the system may be helpful in saving time for vestibular specialists. As history taking of 145 questions is time-consuming and laborious, the classification system can perform simple repetitive tasks that are required for routine patients who visit the dizzy clinic. The questions and answers can be presented and collected using a kiosk or smart device, such as an iPad. The preliminary classification may help the specialist to save time in making the diagnosis and allocate more resources in treating the patient. Our ongoing pilot usability study demonstrates approximately over 55% reduction in completion time compared to the traditional comprehensive vestibular assessments, which typically require 17 minutes. While our 50-feature questionnaire represents a significant reduction from the original 145-item assessment, we acknowledge it still poses a considerable burden in busy clinical settings.

Future iterations should incorporate adaptive questioning algorithms that dynamically select the most informative questions based on initial patient responses. Such algorithms could potentially reduce the questionnaire to 20–30 questions per patient while maintaining diagnostic accuracy, further improving clinical feasibility and patient acceptance. Second, the system may be recommended for clinicians who are not specialized in vestibular disorders: for instance, primary care physicians and doctors at the emergency rooms. These healthcare providers encounter a good volume of dizzy patients but are not specialized in all 6 vestibular disorders. In certain situations, they may feel unconfident about their diagnosis and hope to request some guidance. Our classification system can provide information if their diagnostic practice is generally in line with the practice of a vestibular specialist. The healthcare delivery system will be more efficient, since the primary care physician does not have to refer the patient to a specialist if their impression corresponds to the system's output. Third, this system can be used by patients to see which specialists they should see first. While BPPV, VEST, MD are mostly managed in the department of Otolaryngology, VM is managed in Neurology, PPPD in Psychiatry, and HOD in Internal Medicine. By this classification system, dizzy patients may be allocated to the correct department in the initial stage, instead of visiting all 4 departments.

To mitigate potential anchoring bias and preserve essential clinician-patient interactions, we recommend specific implementation protocols. The system presents the top two most likely diagnoses with quantified probability scores rather than single definitive predictions, encouraging differential diagnosis consideration. The tool's primary value lies in systematically organizing complex symptom data, allowing clinicians to allocate more time to therapeutic discussions, physical examination, and patient education.

The accuracy of the CatBoost model reached 88.4% on the unseen test set, correctly identifying 408 cases (60.9%) and partially correctly identifying 184 cases (27.5%). These findings highlight the model's ability to accurately classify vestibular disorders in the majority of cases. A comparison table with former publications is shown in Table 3.

Our model selection process prioritized generalization capability over validation performance metrics. While Random Forest achieved higher validation accuracy (98%), its substantial performance degradation on unseen data (85%) and shift toward increased reliance on partially correct predictions indicated overfitting. CatBoost demonstrated superior stability with consistent prediction quality across validation and test datasets, making it more suitable for clinical deployment where reliable performance on new patient cases is essential.

The model's performance demonstrated distinct patterns across different vestibular disorders, which align with both clinical priorities and diagnostic complexities. Among the disorders with strong performance, BPPV and VM stood out with particularly robust results. BPPV achieved high sensitivity (0.81) and VM demonstrated balanced performance (sensitivity 0.70, specificity 0.89). The strong performance in VM can be attributed to its diagnostic criteria being based solely on patient symptoms, without requiring additional laboratory results or physical examination findings. This is particularly valuable since BPPV and VM represent the most common causes of dizziness in outpatient settings and are typically managed with conservative treatments.

On the other hand, conditions requiring more invasive interventions or specific treatment pathways showed a distinctly conservative diagnostic pattern. MD demonstrated very high specificity (0.96) but lower sensitivity (0.44), which can be attributed to several key factors: (1) the requirement of hearing tests for definitive diagnosis, (2) symptom fluctuation over time and lifestyle patterns, and (3) the presence of concomitant disorders such as VM and VEST that can make clinical presentation atypical. HOD showed a similar pattern (specificity 0.97, sensitivity 0.33), with its lower sensitivity partially attributable to the limited number of HOD instances in our dataset.

The most conservative performance was observed in PPPD, VEST, and HOD, which reflected both system limitations and intentional design choices. PPPD showed extremely high specificity (0.99) but low sensitivity (0.09), while VEST showed intermediate performance (sensitivity 0.52,

Table 1 | Selected features and description

N	Feature	Description
Algorithm Selection		
1	F_Hx_head_move	First attack: Aggravated by head movement
2	F_Hx_MD_excessive_exercise	First attack: Aggravated by physical overexertion
3	F_Hx_MD_no_sleep_yesterday	First attack: Aggravated by poor sleep quality
4	F_Relieving_good_mood	First attack: Relief by good mood
5	F_Relieving_good_sleep	First attack: Relief by good sleep
6	F_Relieving_head_body_relax	First attack: Relief by fixing head/body
7	F_Relieving_watching_liked_drama	First attack: Relief by relaxing activity
8	Hx_recurrent_headache	History of recurring headaches
9	Hx_recurrent_headache_aggravation_FHx	Family history of headache
10	Hx_recurrent_headache_aggravation_GI_aura	Headache with gastrointestinal aura
11	Hx_recurrent_headache_aggravation_HZ_Dz	Headache with dizziness
12	Hx_recurrent_headache_aggravation_sensory_aura	Headache with sensory aura
13	Hx_recurrent_headache_aggravation_visual_aura	Headache with visual aura
14	Hx_recurrent_headache_location	Location of headaches
15	Hx_recurrent_headache_nausea_diet	Nausea during headaches
16	Hx_recurrent_headache_nausea_smell	Headache triggered by smells
17	L_AS_Earfullness_aggravation_now	Latest attack: Concomitant ear fullness
18	L_Hx_MD_excessive_exercise	Latest attack: Aggravated by physical overexertion
19	L_Hx_MD_no_sleep_yesterday	Latest attack: Aggravated by poor sleep quality
20	L_Hx_MD_stress	Latest attack: Aggravated by mental stress
21	L_Relieving_good_mood	Latest attack: Relief by good mood
22	L_Relieving_good_sleep	Latest attack: Relief by good sleep
23	L_Relieving_head_body_relax	Latest attack: Relief by fixing head/body
24	L_Relieving_watching_liked_drama	Latest attack: Relief by relaxing activity
25	Nature_Vertigo	Characteristics of vertigo episodes
26	PhysExam_Saccade	Presence of catch up saccades
27	PhysExam_Saccade_severity	Obviousness of catch up saccades
28	Recent_stressful_event_claustrophobia	Recent claustrophobia
29	Recent_stressful_event_depression	Recent depressive disorder
30	Recent_stressful_event_fear	Recent anxiety disorder
Clinical Selection		
31	Desc_DurationFirstAttack	First attack: Duration of vertigo/dizziness
32	Desc_DurationHeadache	Duration of headache
33	Desc_DurationLastAttack	Latest attack: Duration of vertigo/dizziness
34	Desc_FrequencyHeadache	Frequency of headaches
35	Desc_LocationHeadache	Specific location of headache
36	Desc_PercentageHeadache	Percentage of dizziness accompanied by headaches
37	F_AS_Earfullness	First attack: Concomitant ear fullness
38	F_AS_headache	First attack: Concomitant headache
39	F_Hx_lay_down	First attack: Aggravated by lying down
40	F_Hx_turn_laying	First attack: Aggravated by rolling over
41	Hx_recurrent_headache_aggravation_physical_activity	Headache aggravated by physical activity
42	Hx_recurrent_headache_intensity	Intensity of recurring headaches
43	Hx_recurrent_headache_motion_sickness	History of motion sickness
44	Hx_recurrent_headache_nature_pulsating	Pulsating nature of headaches
45	Hx_recurrent_headache_phonophobia	Sound sensitivity during headaches
46	Hx_recurrent_headache_photophobia	Light sensitivity during headaches
47	L_AS_hearing_fluctuation_aggravation_now	Latest attack: Concomitant hearing fluctuation
48	L_AS_tinnitus_aggravation_now	Latest attack: Concomitant tinnitus
49	L_AS_visual_aura	Latest attack: Concomitant visual aura
50	Nature_Presyncope	Characteristics of near-fainting episodes

specificity 0.90). This pattern aligns with our system’s intentional design of focusing on organic vestibular disorders that are clinically relevant and treatable, deliberately avoiding the categorization of ambiguous cases into PPPD or HOD. Considering the functional nature of PPPD and its potential overlap with other vestibular disorders, our approach was particularly exclusive to avoid categorizing non-typical patients into this diagnosis without careful consideration. While this may result in delayed treatment of additional functional dizziness, this is clinically less problematic than misdiagnosing organic vestibular disorders that require specific interventions.

During the feature analysis, we discovered the critical importance of balancing algorithmic feature selection with clinical expertise. While our machine learning approach initially identified key features through mathematical criteria, we intentionally incorporated additional features based on vestibular specialists’ domain knowledge, even when their statistical significance was not prominently highlighted by the algorithms. For instance,

symptoms related to blood pressure changes and detailed characteristics of initial episodes were retained despite showing relatively lower algorithmic importance scores. When clinically-guided features were omitted, the model’s accuracy dropped to 80% during validation with unseen data, highlighting the necessity of integrating domain expertise into feature selection. This finding aligns with previous research suggesting that the combination of data-driven insights and expert knowledge leads to more robust predictive models in vestibular disorder classification.³

For clinical deployment, we suggest flexible implementation strategies adaptable to different healthcare settings and user expertise levels. Based on our data collection experience using tablets within hospital settings, the questionnaire can be administered through patient self-completion in waiting areas using plain language interfaces or through healthcare provider-guided structured interviews. Our ongoing usability assessment will further evaluate optimal deployment approaches and address practical considerations such as patient assistance needs and electronic health record integration before clinical implementation.

There are several limitations to this study. First, this was a retrospective, single-center study, which may limit its generalizability. Second, the model’s performance may be affected by the imbalance in the dataset, with some vestibular disorders being more prevalent than others. Third, only 6 disorders were included, while there can be as many as 25 distinct vestibular disorders defined by ICVD³. Fourth, the reference diagnosis was based on a few physicians’ impressions rather than a consensus of several vestibular specialists. Fifth, as this study was conducted at a single center in Korea, generalizability may be limited by cultural and healthcare system differences. Cultural variations in symptom reporting patterns, linguistic expressions of dizziness symptoms, and healthcare-seeking behaviors may affect model performance across different populations. To address this limitation, we are developing a tablet-based application for easy deployment across external institutions to facilitate multi-center validation studies. Future validation should include multinational datasets with consideration of the cultural adaptation of questionnaire items and healthcare system contexts in different countries.

The model’s performance could be further enhanced through several data-driven approaches, including dataset expansion by incorporating data from multiple medical centers to increase geographic and demographic diversity, and the integration of external vestibular disorder datasets from validated medical databases. This could be complemented by implementing

Table 2 | Results obtained from the ML models

Model	Validation Set	Unseen Test Set
Catboost	Total: 536	Total: 670
	Correct: 375	Correct: 408
	Part Correct: 124	Part Correct: 184
	Incorrect: 37	Incorrect: 78
	Accuracy: 0.93	Accuracy: 0.88
Random Forest	Total: 536	Total: 670
	Correct: 449	Correct: 401
	Part Correct: 76	Part Correct: 171
	Incorrect: 11	Incorrect: 98
	Accuracy: 0.98	Accuracy: 0.85
XGBoost	Total: 536	Total: 670
	Correct: 394	Correct: 406
	Part Correct: 114	Part Correct: 183
	Incorrect: 28	Incorrect: 81
	Accuracy: 0.95	Accuracy: 0.88

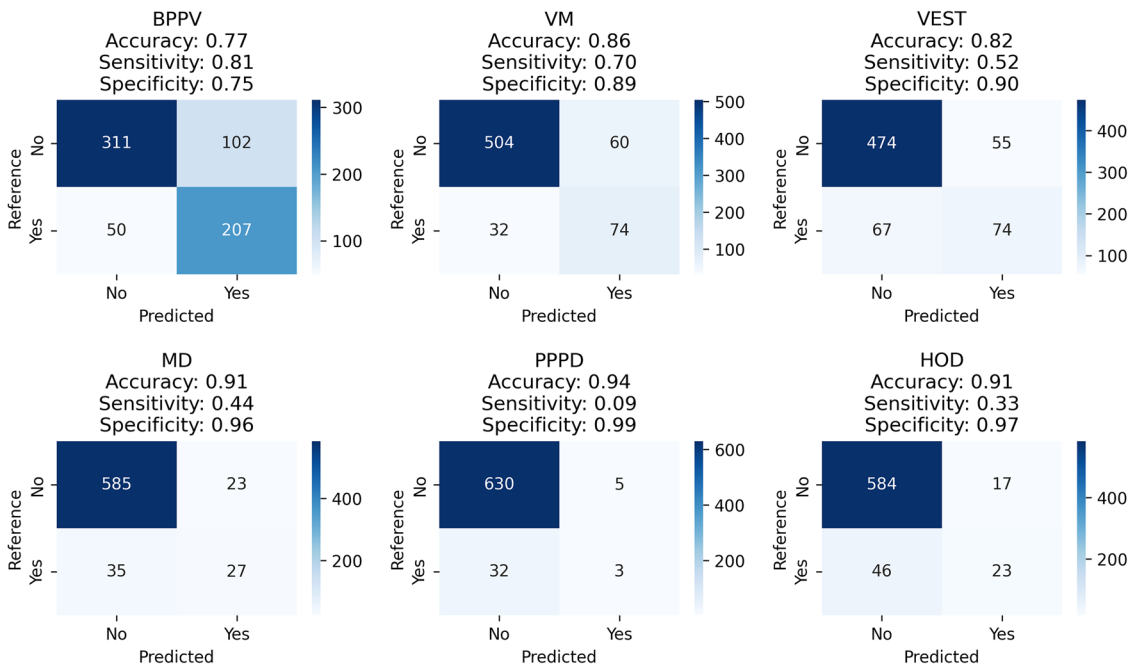


Fig. 3 | Detailed confusion matrices displaying diagnostic accuracy for each vestibular disorder, including key performance metrics.

Table 3 | Comparison of recent studies on vestibular disorder classification

Study	Objective	Machine Learning Methods	Key Findings	Classification
Vivar et al. (2021)	Development of Base-ml tool for classifying vestibular disorders	Multiple algorithms including Random Forest and deep learning models	Base-ml achieved accuracy up to 92.5% in classifying vestibular disorders	MD, BPPV, VM, PPPD
Ahmadi et al. (2020)	Differentiation between central and peripheral vestibular disorders	Linear and non-linear methods including Random Forest and deep learning models	Machine learning outperformed traditional diagnostic scores in acute vestibular syndromes	Stroke, VEST
Wang et al. (2024)	Differentiation between vestibular migraine and Menière's disease	Multiple algorithms including AdaBoost and Random Forest	Achieved 85.77–97.81 % accuracy in differentiating vestibular migraine and Menière's disease	MD, VM
Strobl et al. (2021)	Identification of key variables for classifying vestibular disorders based on patient history	Classification trees for identifying key variables	Key variables identified for distinguishing vestibular disorders with almost 50% accuracy	MD, VM, BPPV, PPPD, vestibular paroxysmia
Raponi et al. (2024)	Use of AI for enhancing diagnosis and differentiation of vestibular disorders	Machine learning algorithms applied to anamnestic questionnaire data	AI models showed high accuracy in supporting diagnosis and differentiation of vestibular disorders	MD, BPPV, AVD, VM
This Study	AI-based classifier for 6 common vestibular disorders	CatBoost machine learning model	Achieved overall accuracy of 88.4%	MD, BPPV, VEST, HOD, VM, PPPD

MD Menière's Disease, BPPV Benign Paroxysmal Positional Vertigo, VM Vestibular Migraine, PPPD - Persistent Postural-Perceptual Dizziness, VEST Vestibular Neuritis/Labyrinthitis, HOD Hemodynamic Orthostatic Dizziness, AVD Acute Vestibular Deficit, AI Artificial Intelligence.

appropriate data augmentation strategies specifically designed for medical data, particularly for addressing class imbalance in less common vestibular conditions.

Overall, the developed ML model demonstrates the potential to support healthcare professionals in the diagnosis of vestibular disorders, achieving an overall accuracy of 88.4% on unseen test data with particularly strong performance in identifying BPPV (sensitivity 0.81) and VM (accuracy 0.86) while maintaining high specificity for conditions requiring distinct treatment approaches. By providing accurate predictions, the model can serve as a valuable tool to aid in the decision-making process, ultimately leading to more efficient and effective patient care. However, it is crucial to recognize that the model is intended to complement, rather than replace, the expertise of medical professionals. This work represents a significant step toward streamlining the diagnostic process for vestibular disorders while maintaining clinical relevance and diagnostic accuracy, contributing to the evolving landscape of vestibular diagnostics with a potentially scalable and efficient solution for patients presenting with complex dizziness symptoms. Further prospective validation and refinement of the model are necessary before its deployment in clinical settings.

Methods

The overall development process, as illustrated in Fig. 4, encompasses several key stages: data processing, feature evaluation, ML model development, and performance evaluation. All processes were developed using Python (Python Software Foundation, Delaware).

The feature selection process consisted of determining the most relevant fields of the clinical record questionnaire. The analysis of the 145 fields was performed in two parts. First, two feature ranking algorithms were implemented to evaluate the importance of the features with respect to the diagnostic outcomes. Second, two vestibular specialists selected the features from the clinical record questionnaire that they considered most relevant when making a diagnosis.

Algorithm analysis

For the data analysis, two feature selection algorithms were used: Recursive Feature Elimination with Support Vector Machine (RFE-SVM), which captures complex feature interactions, and Select K Best (SKB) score, which evaluates individual feature importance independently. This hybrid approach combines RFE-SVM's ability to handle non-linear relationships with SKB's computational efficiency in identifying univariate correlations^{21,22}. The RFE-SVM method is a feature selection technique that recursively eliminates features and builds models using the remaining features. It evaluates the importance of each feature by analyzing model performance, and it identifies the most significant features for predicting the target attribute²¹. In RFE-SVM, the procedure begins by training a Support Vector Machine (SVM) classifier on all available features, and then it computes a ranking criterion for each feature. This criterion is typically derived from the SVM's weight vector. The feature with the smallest ranking criterion is removed, and the process is repeated iteratively until a pre-defined number of features remain. The ranking criterion in RFE-SVM is determined by solving an optimization problem for the weight vector w and bias b , subject to the condition that all data points are correctly classified, as expressed by the following formulation:

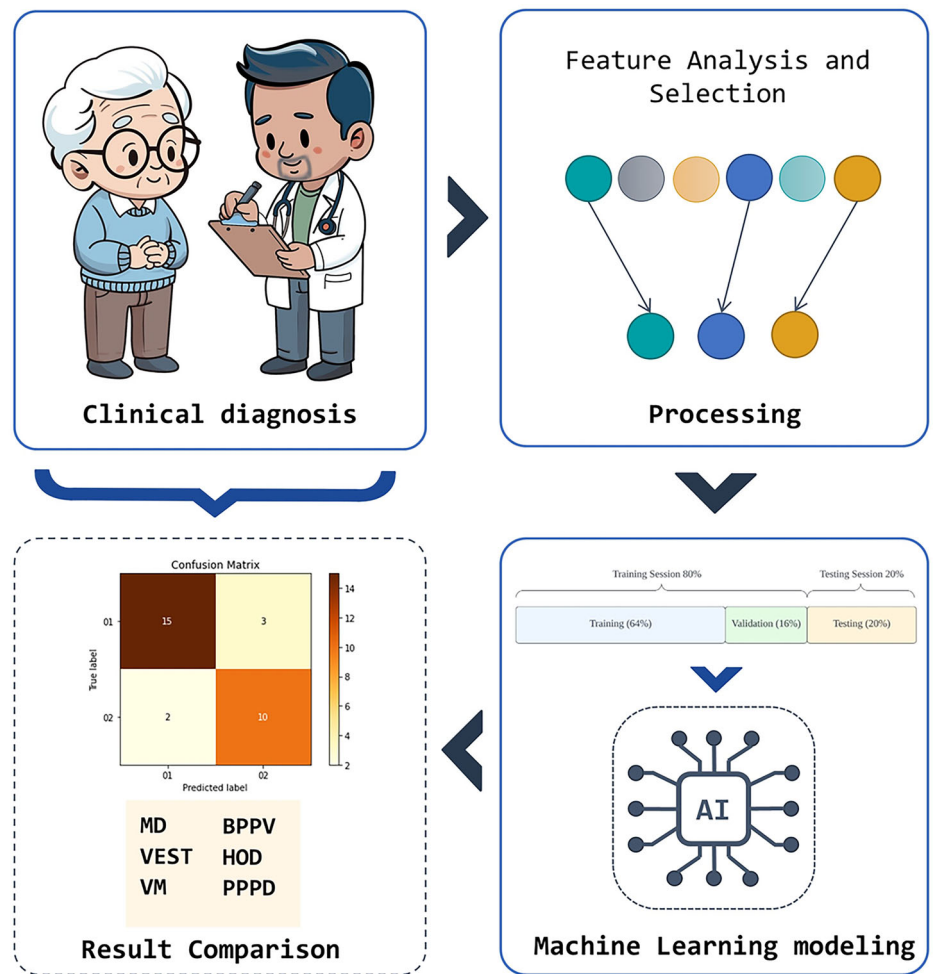
$$\frac{1}{2} ||w||^2 \text{subject to } y_i(w^T x_i + b) \geq 1 \text{ for } i = 1, 2, \dots, n$$

(1)

where w is the weight vector, x_i is the input vector for the i -th training example, y_i is the corresponding label, and n represents the total number of training examples²³. RFE-SVM has been successfully applied in medical data analysis, such as distinguishing between different neurological conditions or diagnosing vestibular disorders, where it has proven effective in identifying key clinical features that contribute to accurate classification^{24,25}.

On the other hand, SKB score, also known as the Chi-squared test, is another feature selection technique that assesses the dependency between a feature and the target variable. It computes the chi-squared statistic for each

Fig. 4 | Overall development process. Data collection through clinical diagnosis, followed by feature analysis and selection, culminating in ML modeling and comparative analysis of results.



feature and the target, helping to identify which features are most informative in relation to the outcome variable^{26,27}. Specifically, the chi-squared statistic for a feature f and the target variable y is calculated as:

$$\chi^2(f, y) = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (2)$$

where O_{ij} is the observed frequency of feature f with value i and target variable y with value j , and E_{ij} is the expected frequency of feature f with value i and target variable y with value j under the null hypothesis of independence.

The expected frequency E_{ij} is calculated as:

$$E_{ij} = \frac{\text{count}(f = i) \times \text{count}(y = j)}{n} \quad (3)$$

where $\text{count}(f = i)$ is the number of instances where feature f has value i , $\text{count}(y = j)$ is the number of instances where the target variable y has value j , and n is the total number of instances.

SKB has proven valuable in identifying the most relevant features, particularly in text classification²⁸ and biomedical data analysis²⁹, where its ability to capture important relationships between variables improves classification performance.

After applying these two algorithms, we selected the top 30 fields that were common to both methods, enabling us to identify the most important and relevant features for the classification task. This combined approach ensures that we leverage the strengths of both RFE-SVM and SKB to enhance the feature selection process.

Lived experience-based hybrid feature selection

In addition to the algorithmic feature selection, two vestibular specialists identified 20 additional features they considered crucial in making an accurate diagnosis. Their selection was informed by both their clinical experience and established frameworks such as the ICVD³. The ICVD is currently the most widely recognized system for classifying vestibular disorders, offering a standardized approach for diagnosis and management. Based on lived experience, essential features were selected within the standardized ICVD. These 20 features served as minimum clinical guidance to coach the ML model not to leave out features that are important in the real-world setting. For instance, the ML model may consider a patient's response to a question about dizziness frequency as merely one feature with equal weighting to others. However, an experienced vestibular specialist will recognize it as an important feature in making the correct diagnosis, as it provides critical insights into the severity and progression of the disorder³⁰. This highlights the necessity of integrating domain expertise into feature selection to ensure that clinically significant questionnaire-based features are not undervalued or overlooked by purely algorithmic approaches.

ML model development

For the ML model, the complete dataset composed of 3349 records was divided into training and testing sessions. The training session comprised 80% of the data using 5-fold cross-validation with stratification during model development to ensure balanced class proportions and robust performance assessment across validation folds. The remaining 20% of the data was allocated to the testing session, serving as the unseen test data to evaluate the model's performance.

The training session was utilized to train and optimize three different ML models: CatBoost, Random Forest, and XGBoost. These models were selected due to their proven effectiveness in various classification tasks and their ability to handle complex, high-dimensional data³¹.

CatBoost is a gradient boosting algorithm that leverages ordered boosting and categorical feature encoding to improve performance³². It minimizes the following objective function:

$$L(f) = \sum_{i=1}^n l(y_i, f(x_i)) + \Omega(f) \quad (4)$$

where $l(y_i, f(x_i))$ is the loss function for the i th instance, and $\Omega(f)$ is the regularization term.

Random Forest is an ensemble learning method that constructs multiple decision trees and combines their predictions to make the final classification³³. Each decision tree is built using a random subset of features and instances, which helps to reduce overfitting and improve generalization. The final prediction is obtained by aggregating the predictions of all decision trees:

$$f(x) = \frac{1}{N} \sum_{i=1}^N f_i(x) \quad (5)$$

where $f_i(x)$ is the prediction of the i th decision tree, and N is the total number of trees in the forest.

XGBoost is another gradient boosting algorithm that optimizes the following objective function:

$$L(\Phi) = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{i=1}^n \Omega(f_i) \quad (6)$$

where $l(y_i, \hat{y}_i)$ is the loss function for the i th instance, $\hat{y}_i$ is the predicted value, and $\Omega(f_i)$ is the regularization term for the i th decision tree³⁴.

To find the best hyperparameters for each model, an optimization algorithm was implemented. Techniques such as grid search, random search, and Bayesian optimization were employed to efficiently explore the hyperparameter space and identify the most promising configurations³⁵.

We trained these models using the selected features and the curated dataset, utilizing a validation set for hyperparameter tuning to optimize their performance. The validation set allowed us to fine-tune the models' hyperparameters and assess their performance on unseen data during the training process.

To evaluate the effectiveness of the vestibular disorder classification, we categorized the results into three groups: 'Correct,' 'Partially Correct,' and 'Incorrect.' The 'Correct' category indicates that the initial prediction from the developed model aligns with the vestibular specialists' first impression. The 'Partially Correct' category denotes cases where the second most prediction from the developed model corresponds to the vestibular specialists' first impression, or the vestibular specialists' second impression coincides with either the first or second most prediction from the developed model. This classification was operationalized through a systematic rule-based algorithm applied consistently across all test cases without subjective clinical judgment, ensuring reproducible outcome assessment and eliminating potential bias. Lastly, the 'Incorrect' category denotes case where the algorithm does not match the vestibular specialist's first and second diagnoses. A similar approach has been utilized in previous studies evaluating multiple categories of outcomes^{36,37}.

A separate ongoing pilot usability study is being conducted to assess the practical feasibility of implementing our diagnostic questionnaire in clinical settings.

Metrics

To evaluate the performance of the ML models, several metrics were utilized, including the confusion matrix, accuracy, specificity, and sensitivity^{38,39}. The confusion matrix provides a tabular summary of the model's classification performance, displaying the counts of true positive (TP), true

negative (TN), false positive (FP), and false negative (FN) predictions³⁹. In our evaluation, both 'Correct' and 'Partially Correct' predictions are considered true positives (TP), as they align with the doctors' impressions either fully or partially. 'Incorrect' predictions are considered false negatives (FN) or false positives (FP), depending on the actual diagnosis. Accuracy measures the overall correctness of the model's predictions and is calculated as:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (7)$$

Specificity, also known as true negative rate, quantifies the model's ability to correctly identify negative instances and is computed as:

$$Specificity = \frac{TN}{(TN + FP)} \quad (8)$$

Sensitivity, or true positive rate, assesses the model's capability to correctly identify positive instances and is calculated as:

$$Sensitivity = \frac{TP}{(TP + FN)} \quad (9)$$

These metrics collectively provide a comprehensive evaluation of the model's performance, enabling informed decisions regarding model selection and refinement⁴⁰.

Data availability

The datasets used and analyzed during the current study available from the corresponding author on request.

Abbreviations

BPPV	Benign Paroxysmal Positional Vertigo
FN	False Negative
FP	False Positive
HINTS	Head Impulse, gaze-evoked Nystagmus, Test of Skew
HOD	Hemodynamic Orthostatic Dizziness
ICVD	International Classification of Vestibular Disorders
IRB	Institutional Review Board
MD	Menière's Disease
ML	Machine Learning
PPPD	Persistent Postural-Perceptual Dizziness
RFE-SVM	Recursive Feature Elimination with Support Vector Machine
SD	Standard Deviation
SKB	Select K Best
SSNHL	Sudden Sensorineural Hearing Loss
TN	True Negative
TP	True Positive
VEST	Vestibulopathy
VM	Vestibular Migraine
XGBoost	Extreme Gradient Boosting

Received: 3 April 2025; Accepted: 10 July 2025;

Published online: 31 July 2025

References

1. Kerber, K. A. & Baloh, R. W. The evaluation of a patient with dizziness. *Neurol. Clin. Pr.* **1**, 24–33 (2011).
2. 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).
3. Bisdorff, A. R., Staab, J. P. & Newman-Toker, D. E. Overview of the International Classification of Vestibular Disorders. *Neurol. Clin.* **33**, 541–550 (2015).

4. Lempert, T. et al. Vestibular migraine: diagnostic criteria. *J. Vestib. Res.* **22**, 167–172 (2012).
5. Lopez-Escamez, J. A. et al. Diagnostic criteria for Menière's disease. *J. Vestib. Res.* **25**, 1–7 (2015).
6. Von Brevern, M. et al. Benign paroxysmal positional vertigo: diagnostic criteria. *J. Vestib. Res.* **25**, 105–117 (2015).
7. Strupp, M. et al. Vestibular paroxysmia: Diagnostic criteria. *J. Vestib. Res.* **26**, 409–415 (2016).
8. Cha, Y. H. et al. Mal de débarquement syndrome diagnostic criteria: Consensus document of the Classification Committee of the Barany Society. *J. Vestib. Res.* **30**, 285–293 (2020).
9. Kim, H. A. et al. Hemodynamic orthostatic dizziness/vertigo: Diagnostic criteria. *J. Vestib. Res.* **29**, 45–56 (2019).
10. Staab, J. P. et al. Diagnostic criteria for persistent postural-perceptual dizziness (PPPD): Consensus document of the committee for the Classification of Vestibular Disorders of the Barany Society. *J. Vestib. Res.* **27**, 191–208 (2017).
11. Staab, J. P. et al. Diagnostic criteria for persistent postural-perceptual dizziness (PPPD): Consensus document of the committee for the Classification of Vestibular Disorders of the Bárány Society. *J. Vestib. Res.* **27**, 191–208 (2017).
12. Strupp, M. et al. Bilateral vestibulopathy: Diagnostic criteria Consensus document of the Classification Committee of the Bárány Society 1. *J. Vestib. Res.* **27**, 177–189 (2017).
13. Vivar, G. et al. Using Base-ml to learn classification of common vestibular disorders on DizzyReg registry data. *Front. Neurol.* **12**, 681140 (2021).
14. 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).
15. Wang, C. et al. Machine learning models help differentiate between causes of recurrent spontaneous vertigo. *J. Neurol.* **271**, 3426–3438 (2024).
16. Raponi, I. Artificial intelligence applications for diagnosis and differentiation of vestibular disorders. *Otolaryngol., Head. Neck Surg.* **10**, 1–7 (2024).
17. Strobl, R. et al. A set of eight key questions helps to classify common vestibular disorders-results from the DizzyReg Patient Registry. *Front. Neurol.* **12**, 670944 (2021).
18. Sirocchi, C., Bogliolo, A. & Montagna, S. Medical-informed machine learning: integrating prior knowledge into medical decision systems. *BMC Med. Inform. Decis. Mak.* **24**, 186 (2024).
19. Alvarez, J. D., Matias-Guiú, J. A., Cabrera-Martin, M. N., Risco-Martin, J. L. & Ayala, J. L. An application of machine learning with feature selection to improve diagnosis and classification of neurodegenerative disorders. *BMC Bioinforma.* **20**, 491 (2019).
20. van de Berg, R. et al. Vestibular Migraine of Childhood and Recurrent Vertigo of Childhood: Diagnostic criteria Consensus document of the Committee for the Classification of Vestibular Disorders of the Barany Society and the International Headache Society. *J. Vestib. Res.* **31**, 1–9 (2021).
21. Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **46**, 389–422 (2002).
22. Li, J. et al. Feature Selection. *ACM Comput. Surv.* **50**, 1–45 (2017).
23. Cortes, C. Support-Vector Networks. *Machine Learning*, (1995).
24. Richhariya, B., Tanveer, M., Rashid, A. H. & Initia, A. D. N. Diagnosis of Alzheimer's disease using universum support vector machine based recursive feature elimination (USVM-RFE). *Biomed. Signal Process. Control* **59**, 101903 (2020).
25. Yang, M. et al. Establishing a prediction model of severe acute mountain sickness using machine learning of support vector machine recursive feature elimination. *Sci. Rep.* **13**, 4633 (2023).
26. Jin, X., Xu, A., Bie, R. & Guo, P. In *Data Mining for Biomedical Applications: PAKDD 2006 Workshop, BioDM 2006, Singapore, April 9, 2006. Proceedings*. 106–115 (Springer).
27. Rupapara, V., Rustam, F., Ishaq, A., Lee, E. & Ashraf, I. Chi-square and PCA based feature selection for diabetes detection with ensemble classifier. *Intell. Autom. Soft Comput.* **36**, 1931–1949 (2023).
28. Chen, Y. T. & Chen, M. C. Using chi-square statistics to measure similarities for text categorization. *Expert Syst. Appl.* **38**, 3085–3090 (2011).
29. Yousef, M. & Voskergian, D. TextNetTopics: Text classification based word grouping as topics and topics' scoring. *Front. Genet.* **13**, 893378 (2022).
30. Zhou, G. & Brodsky, J. R. Objective vestibular testing of children with dizziness and balance complaints following sports-related concussions. *Otolaryngol. Head. Neck Surg.* **152**, 1133–1139 (2015).
31. Bischl, B. et al. Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *Wiley Interdiscip. Rev. -Data Min. Knowl. Discov.* **13**, e1484 (2023).
32. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. CatBoost: unbiased boosting with categorical features. *Advances in Neural Information Processing Systems 31 (NIPS 2018)* **31** (2018).
33. Agrawal, P. In *2024 2nd International Conference on Disruptive Technologies (ICDT)*. 11–15 (IEEE).
34. Putatunda, S. & Rama, K. In *2019 Fifteenth International Conference on Information Processing (ICINPRO)*. 1–6 (IEEE).
35. Liashchynskiy, P. & Liashchynskiy, P. Grid search, random search, genetic algorithm: a big comparison for NAS. *arXiv preprint arXiv:1912.06059* (2019).
36. Pallauta, J. D., Arteaga, P. & Garzón-Guerrero, J. A. Secondary school students' construction and interpretation of statistical tables. *Mathematics* **9**, 3197 (2021).
37. Sushil, M. et al. CORAL: expert-curated oncology reports to advance language model inference. *NEJM AI* **1**, Aldbp2300110 (2024).
38. Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. & Nielsen, H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* **16**, 412–424 (2000).
39. Cullerne Bown, W. Sensitivity and specificity versus precision and recall, and related dilemmas. *J. Classif.* **41**, 402–426 (2024).
40. Tharwat, A. Classification assessment methods. *Appl. Comput. Inform.* **17**, 168–192 (2021).

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2021R1A2C1094147).

Author contributions

Ku and Suh had full access to all data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Conceptualization: Ryu, Callejas Pastor, Ku, Suh. Data curation: Callejas Pastor, Ryu, Joo, Ku, Suh. Formal analysis: Callejas Pastor, Ku, Suh. Methodology: Callejas Pastor, Ku, Suh. Project administration: Ku, Suh. Writing – original draft: Callejas Pastor, Ku, Suh. Writing – review & editing: Callejas Pastor, Ku, Suh.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Yunseo Ku or Myung-Whan Suh.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025