Contents lists available at ScienceDirect

Intelligence-Based Medicine

journal homepage: www.sciencedirect.com/journal/intelligence-based-medicine

Feasibility of deep learning to predict tinnitus patient outcomes

Katherine S. Adcock [a], Gabriel Byczynski [a], Emma Meade [b], Sook Ling Leong [b], Richard Gault [c], Hubert Lim [b,d], Sven Vanneste [a,*]

[a] *Trinity Institute of Neuroscience, Trinity College Dublin, Ireland*
[b] *Neuromod Devices Limited, Dublin, D08 R2YP, Ireland*
[c] *School of Electronics, Electrical Engineering and Computer Science, Queen's University, Belfast, UK*
[d] *Department of Biomedical Engineering, University of Minnesota, Minneapolis, MN, 55455, USA*

ARTICLE INFO

Keywords:
Deep learning
Tinnitus
Personalized medicine

ABSTRACT

Advances in machine and deep learning techniques provide a novel approach in understanding complex patterns within large datasets, leading to an implementation of personalized medicine approaches to support clinical decision making. Results from recent clinical trials (TENT-A1 and TENT-A2 studies; clinicaltrials.gov: NCT02669069 and NCT03530306) support that a novel bimodal neuromodulation approach could be a breakthrough treatment for patients with tinnitus, which adversely affects 10–15 % of the population. Given the heterogeneity of symptoms, it is important to identify whether treatment has an optimal effect on specific subgroups of tinnitus patients. The current study is a first look at the feasibility of using deep learning modelling on patient reported data to predict treatment outcomes in individuals with tinnitus, and highlights what features are most beneficial for clinical decision making.

1. Introduction

Recent advancement in machine learning techniques have served as a catalyst in understanding complex patterns within large datasets, leading to a rise in implementation of personalized medicine approaches to support clinical decision making [1–4]. General machine and deep learning models are data-driven computational approaches that help identify complex patterns, hidden correlations, and non-linear relationships in datasets for which traditional exploratory and inferential statistical analyses struggle to accurately measure [2,3]. Deep learning approaches, inspired by the biological neural communication networks in the brain, can extract and process information from given data to make individual predictions on new data [2,3]. Furthermore, these approaches can provide a better understanding to complex heterogeneous disorders, encompassing multi-factorial conditions through the stratification of patients into subgroups with similar features.

Personalized medicine has leveraged artificially intelligent algorithms in drug development, disease characteristics, and therapeutic effects for several diseases such as diabetes, cancer, heart disease, chronic inflammatory disease, and psychiatric disease [5–12]. Recent studies have highlighted personalized medicine towards tinnitus management, in which regression or classification modelling could improve clinical decision making [13,14]. Tinnitus is a highly heterogeneous disorder, in which various components are reported amongst patients including the etiology, phenotype, and comorbidities [15]. The heterogenous nature of the disorder likely affects treatment options and response. Machine learning approaches such as deep learning gives researchers the opportunity to improve clinical diagnostic, prognostic, and treatment decision making through individualized, tailored predictions [5,6,9,11,16,17].

In the current study, we sought to test the feasibility of using deep learning to predict tinnitus patient outcomes following treatment. To do so, we used data from two recent clinical trials (TENT-A1 study, clinicaltrials.gov: NCT02669069; TENT-A2 study, clinicaltrials.gov: NCT03530306) [18,19] to train our deep neural network model to predict tinnitus severity during or after bimodal stimulation via a novel neuromodulation approach to treatment. We generated an initial deep learning model to predict interim Tinnitus Handicap Inventory (THI) scores after 6 weeks of treatment using various features within the database [18,19] representing patient characteristics, tinnitus

[*] Corresponding author. Lab for Clinical and Integrative Neuroscience, Institute for Neuroscience, School of Psychology, Trinity College Dublin, College Green, Dublin 2, Ireland.

E-mail address: sven.vanneste@tcd.ie (S. Vanneste).

URL: http://www.lab-clint.org (S. Vanneste).

https://doi.org/10.1016/j.ibmed.2024.100141
Received 1 December 2023; Received in revised form 4 March 2024; Accepted 13 May 2024
Available online 18 May 2024

characteristics, hearing loss, and psychoacoustic measures. As a benchmark, we compared the performance of our deep learning model to a range of commonly utilized machine learning algorithms.

Various tinnitus related measures served as input features for the models however only some features may be informative in predicting tinnitus severity. Moreover, some features may even be distracting in making predictions. Feature reduction is a common tool to optimize deep learning models and reduce model complexity, in which only those features deemed most relevant are included as input whilst those that are not important or redundant are excluded. Knowing which features are critical to accurate decision making can optimize the patient and clinician's time and resources they will have to collect. To determine if feature reduction methods impact model performance, we compared model performance when feature reduction was and was not applied. We then further test the clinical applicability of the model, to determine if we can predict the final THI score, by classifying whether or not participants responded to treatment. Together, these findings will test the feasibility and establish the initial framework of using deep learning to predict tinnitus patient outcomes following treatment.

2. Methods

2.1. Participants

Data used for training and testing our initial deep learning models were obtained from two previous studies [18,19]. The TENT-A1 [18] and TENT-A2 [19] were randomized, double-blind, parallel-arm, clinical trials with a 12 week treatment period and a 12 month follow-up period investigating the safety and efficacy of various bimodal auditory and somatosensory stimulation treatments for tinnitus. Comprehensive description of the TENT-A1 [18] and A2 [19] cohorts and the intervention designs have been previously described. Briefly, in TENT-A1, 326 tinnitus participants enrolled at two separate sites (Dublin, Ireland and Regensburg, Germany) were randomized to one of three treatment arms. In TENT-A2, 191 tinnitus participants enrolled at a single-site in Dublin, Ireland were randomized to one of four treatment arms. Sixty-six participants from TENT-A1 and 19 participants from TENT-A2 were excluded from analyses due to no interim Tinnitus Handicap Inventory (THI) scores, final THI scores, or treatment hours, resulting in a total of 432 participants (i.e., TENT A-1 n = 260, TENT-A2 n = 172).

2.2. Variable selection

Together, our datasets included 25 variables: input variables representing patient characteristics, tinnitus characteristics, hearing loss, and psychoacoustic measures at baseline, and output variables representing tinnitus severity following treatment. The following discussion summaries these variables and describes their characteristics.

1. Patient Characteristics

Age and sex were included as input variables to reflect patient demographics. As part of the eligibility criteria, only participants between the ages of 18–70 years were recruited in TENT-A1 and -A2.

2. Tinnitus Characteristics

Tinnitus severity, loudness, annoyance, duration, location and tinnitus type were utilized to characterize the patient's tinnitus. Severity of tinnitus was assessed using the 25-item Tinnitus Handicap Inventory (THI) [20] at three time points. Responses were rated as 'Yes' (4-points), 'Sometimes' (2-points) or 'No' (0-points). Baseline THI was the average THI score at screening and enrolment. Interim THI was assessed 6 weeks after treatment, and final THI was assessed 12 weeks after treatment. Participants were also asked to describe the loudness and annoyance of their tinnitus by respectively rating on a 10-point visual analogue scale (VAS) (0, 'not loud/annoyed at all' and 10, 'extremely loud/annoyed') to the questions 'Can you rate the loudness of your tinnitus on a scale of 0–10 right now' and 'Can you rate the annoyance of your tinnitus on a scale of 0–10 right now'. The duration (years) and location of tinnitus was also assessed. In TENT-A1, participants had experienced tinnitus for more than 3 months and less than 5 years. In TENT-A2, the upper boundary for tinnitus duration during recruitment was increased to 10 years [21,22]. Tinnitus location was categorically described as either in the left ear, right ear, both ears but left is worse, and both ears but right is worse. Tinnitus type was categorically described as pure tone, hissing, constant noise, or a combination of. Participants also described their tinnitus as either constant or fluctuating.

3. Hearing Loss

The variable Hearing Loss (HL) was derived from an average dbHL in bands 0.25–8 kHz in both the left and right ears. The degree of High Frequency Hearing Loss (HF-HL) was generated from an average dbHL in the bands from 1 to 8 kHz in both ears. In both TENT-A1 and -A2, participants with conductive hearing loss (i.e., >40 dbHL in at least one measurement frequency in the range of 0.25–1.00 kHz or has >80 dBHL in at least one measurement frequency in the range of 2.0–8.0 kHz) were excluded [21,22]. Hearing loss asymmetry was calculated by subtracting total hearing loss on the right ear from the left ear.

4. Psychoacoustic Measures

Loudness Discomfort Level (LDL), Minimal Masking Level (MML) dBSL, and Tinnitus Loudness Matching (TLM) dBSL were utilized to characterize psychoacoustic measures. MML dBSL was used to estimate the lowest level of broadband noise needed to minimally mask the participant's tinnitus. TLM dBSL was utilized to assess the stimulus that is equal in loudness to the participant's tinnitus. Loudness Discomfort Level (LDL) assessed auditory hypersentivity. LDL dBSL was generated using the LDL averaged at 500 Hz in the right and left ears. Participants in both TENT-A1 and TENT-A2 were recruited to have a wide-band noise MML measurement between 20- and 80-dB hearing level. A detailed description of psychoacoustic assessment methods can be found in previous publications [21,22].

2.3. Data pre-processing

Categorical variables were encoded into numerical labels for each category. Baseline and interim THI scores were categorized into 5 groups: slight (0–16 points), mild (18–36 points), moderate (38–56), severe (58–76), and catastrophic (78–100) [23]. All input values were scaled from 0 to 1 using a min-max scaling technique to normalize the range of input data. The dataset was randomly split into training, validation, and testing sets, in which 80 % of the data was used for training (64 %) and validation (16 %) of the model, and 20 % of the data was used to test the model.

In our regression-based models, either interim or final THI scores were used as primary outcomes. In our classification-based models, change in THI was considered the primary outcome, calculated by subtracting the baseline THI score from the final THI score. If the change in THI decreased by 7 points or more [24], subjects were considered a significant responder to treatment (n = 338). If the change in THI was less than 7 points [24], subjects were considered a non-responder (n = 94). To address class imbalance, Synthetic Minority Over-Sampling TEchnique (SMOTE) to generate synthetic data for our training dataset. SMOTE uses the k-nearest neighbour algorithm to oversample the minority class, and thereby produces synthetic samples of non-responder data to balance the dataset [25].

2.4. Model building and training

We trained four sequential deep learning models, and five machine learning models to benchmark our deep learning models (Table 1). Deep learning models were implemented using Keras with a TensorFlow backend [26] and machine learning models were implemented with Scikit-Learn. Model 1 and the corresponding benchmark models aimed to predict interim THI score, and utilized 22 input variables including sex, age, tinnitus duration, tinnitus location, tinnitus type, THI, LDL (right ear, left ear, and worst case), VAS for annoyance and loudness, MML dBSL, TLM dBSL, high frequency hearing loss (left ear and right ear), total hearing loss (left ear and right ear), and hearing loss asymmetry. We compared performance to five machine learning models: ElasticNet, a linear Support Vector Machine (SVM), Random Forest, Decision Tree, and a Dummy Regression. For our Dummy Regression, we utilized the DummyRegressor function in SciKit-Learn, which is useful as a simple baseline to compare with other regressors.

We used SHAP (Shapley Additive Explanations) as a feature reduction method, in which input features were ranked by the highest Shapley values which evaluate an importance value for a particular prediction [27]. We then reduced the number of input variables to include the 5 features with the highest SHAP values to simplify the model. Model 2 utilized baseline THI group, VAS for annoyance, MML dBSL, tinnitus location, and high frequency hearing loss (left side) as input variables to predict interim THI score. Model 3 aimed to predict final THI score, and utilized baseline THI group, interim THI group, VAS for annoyance, MML dBSL, tinnitus location, and high frequency hearing loss (left side) as input variables. Moreover, Model 2 aimed to predict short term predictions (i.e., interim THI) whereas Model 3 aimed to predict long-term predictions (i.e., final THI). Using these same input variables as Model 3, Model 4 aimed to classify between responders and non-responders to treatment.

Each regression neural network (Models 1–3) consisted of two hidden layers with a width of 50 nodes and 25 nodes, and an output layer of 1 node. We applied a rectified linear (ReLU) activation function on the hidden layers, a linear activation function on the output layer, and used mean squared error as the loss function. Our classification model (Model 4) consisted of two hidden layers with a width of 50 nodes and 25 nodes, and an output layer of 2 nodes. We applied a rectified linear (ReLU) activation function on the hidden layers, a sigmoid activation function on the output layer, and sparse categorical cross entropy as the loss function. All models used Adam optimization with 0.001 for learning rate. Model training consisted of 100 epochs and a batch size of 8. We performed a repeated random sub-sampling cross validation to validate model performance, in which model performance was evaluated over 50 iterations. Here, the database was randomly split into a training, validation, and testing dataset for every iteration.

2.5. Model evaluation

Model evaluation and analysis was performed using Python. We first evaluated deep leaning model performance to 5 machine learning models: ElasticNet, SVM, Random Forest. Decision Tree, and a Dummy Regressor model as a benchmark [28,29]. The Dummy Regressor always predict the mean value of the training data regardless of the input given, acting as reference model trained deliberately to aim for the centre of the data and not learn the wider distribution. Therefore, if a model performs similarly to the Dummy Regressor, that would indicate the model hasn't learned the data distribution or the data is very tightly distributed. Regression models (Models 1, 2, and 3) were evaluated using R^2, mean squared error (MSE), and mean absolute error (MAE) on the test dataset. Feature importance was interpreted using DeepSHAP to approximate mean absolute SHAP values, and individual feature SHAP values for each prediction [27,30].

The classification model (Model 4) was evaluated using prediction accuracy, receiver operating statistic (ROC), precision, and recall. The test set prediction accuracies measure the model's ability to accurately classify data in the test dataset following training, illustrating the model's generalizability to data it has not seen yet. Accuracies were calculated as the percentage of correct predictions (true positives + true negatives divided by the total sum of predictions). The ROC curve plots the true positive rate (TPR, i.e., recall) to the false positive rate (FPR) at various thresholds [31]. Area under the curve was calculated for the ROC curve (i.e., AUC-ROC), a commonly reported performance metric representing the model's capability of distinguishing between classes [32]. Here, we also report the area under the curve for the Precision-Recall curve (i.e., AUC-PRC), which is a more suitable metric for imbalanced datasets compared to ROC [33]. The Precision-Recall curve plots the TPR (i.e., recall) to positive predictive values (i.e., precision) for different probability thresholds.

We performed a repeated, random sub-sampling, cross validation to validate model performance over 50 iterations. All measures were recorded for each iteration. Graphical representations illustrate the data distribution over the course of 50 iterations. Mean and standard deviation were also reported for each measure, averaging the measures across 50 iterations of each model, as seen in Table 2.

3. Results

We first investigated feasibility of using deep learning to predict tinnitus severity after 6 weeks of treatment. Here, we used our 22 baseline variables to train the models to predict interim THI score. Our initial deep learning model, Model 1, outperformed the machine learning models, indicated by an increase in variance explained (R^2), and decreased MSE and MAE compared to the other models (Fig. 1, Model 1: $R^2 = 0.43$, MSE = 161.12, MAE = 9.99). As each of the models were evaluated following a repeated, random sub-sampling, cross validation, the box-and-whisker diagrams in Fig. 1 illustrate how each model performed across all the repetitions for the relevant metric.

Various baseline variables representing patient characteristics, tinnitus characteristics, hearing loss, and psychoacoustic measures were used to train the models, however not all features may be important for predicting tinnitus severity. We used SHAP as a data-driven feature reduction method, in which feature importance was determined post-hoc for our deep learning model. Baseline THI category was identified as most important, with an average absolute SHAP value of 8.09 (Fig. 2).

Table 1
Description of deep learning model input and outputs. Benchmark machine learning models utilized the same input and output as Model 1.

Label	Type	Input	Output
DL Model 1	Regression	**22 variables:** sex, age, tinnitus duration, tinnitus location, tinnitus type (1–5), baseline THI subgroup, LDL (right ear, left ear, and worst case), VAS for annoyance and loudness, MML dBSL, TLM dBSL, high frequency hearing loss (left ear and right ear), total hearing loss (left ear and right ear), and hearing loss asymmetry	Interim THI score
DL Model 2	Regression	**5 variables:** Baseline THI subgroup, VAS for annoyance, MML dBSL, tinnitus location, high frequency hearing loss (left)	Interim THI score
DL Model 3	Regression	**6 variables:** Baseline THI subgroup, Interim THI subgroup, VAS for annoyance, MML dBSL, tinnitus location, high frequency hearing loss (left)	Final THI score
DL Model 4	Classification	**6 variables:** Baseline THI subgroup, Interim THI subgroup, VAS for annoyance, MML dBSL, tinnitus location, high frequency hearing loss (left)	Responder or Non-responder

Table 2
Model performance metrics for our Regression and Classification models. The table outlines the model name, the number of input variables (see Table 1 for variable names), the output variable, and the corresponding performance metrics for regression or classification models. Performance metrics are presented as mean ± standard deviation for 50 iterations.

Regression Models

Model	Input (# vars)	Output	R^2	MSE	MAE
ElasticNet	22	Interim THI score	0.07 ± 0.02	269.45 ± 35.24	13.02 ± 0.97
SVC - R	22	Interim THI score	−0.07 ± 0.19	307.54 ± 47.65	13.79 ± 1.09
Random Forest	22	Interim THI score	0.08 ± 0.16	264.83 ± 47.35	12.71 ± 1.12
Decision Tree	22	Interim THI score	−0.05 ± 0.20	302.88 ± 59.59	13.65 ± 1.32
Dummy	22	Interim THI score	−0.01 ± 0.02	294.04 ± 37.03	13.64 ± 0.98
DL Model 1	22	Interim THI score	0.43 ± 0.09	161.12 ± 24.11	9.99 ± 0.75
DL Model 2	5	Interim THI score	0.45 ± 0.08	156.48 ± 20.30	9.83 ± 0.71
DL Model 3	6	Final THI score	0.52 ± 0.97	128.99 ± 25.32	8.65 ± 0.77

Classification Model

Group	Input (# vars)	Output	Accuracy	AUC-PRC	AUC-ROC	F1
DL model 4	6	Responder or Non-responder	71.71 ± 4.49 %	0.86 ± 0.05	0.64 ± 0.07	0.82 ± 0.03

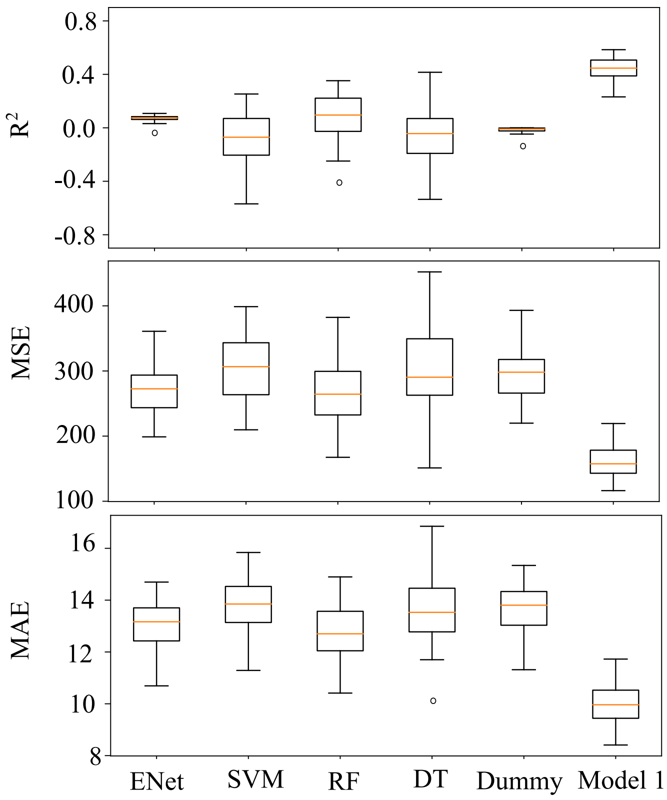

Fig. 1. Deep learning model outperforms machine learning algorithms to predict interim THI. We compared model performance, quantified by variance explained (R^2), mean squared error (MSE), and mean absolute error (MAE), across 5 machine learning models and one deep learning model. Deep learning Model 1 has greater R^2, and reduced MSE and MAE compared to machine learning models ElasticNet (ENet), linear Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), and a Dummy Regression model (Dummy). Box plots illustrate the data distribution over the course of 50 iterations.

In addition to baseline THI, VAS for annoyance, MML dBSL, tinnitus location, and high frequency hearing loss (left side) emerged as the top 5 important features to predict interim THI score. These results suggest that baseline tinnitus severity is highly influential in predicting tinnitus severity following 6 weeks of treatment. Feature importance for individual predictions of Model 1 is visualized in Fig. 2. High baseline THI group (i.e., severe and catastrophic) corresponds with higher SHAP values. In other words, high baseline THI group corresponds with a higher predicted interim THI score.

Knowing which features are critical to accurate decision making can optimize the patient and clinician's time and resources they will have to collect. To determine if feature reduction methods impact model performance, we compared model performance after only including the top 5 important features. Feature reduction increased variance explained (R^2), and decreased MSE and MAE (Fig. 3, Model 2: $R^2 = 0.45$, MSE = 156.48, MAE = 9.83) compared to our original model with 22 features (Model 1).

We next wanted to test the feasibility of predicting tinnitus severity following 12 weeks of treatment if we knew the subject's interim score (Model 3). We therefore used the same input features in addition to interim THI category (Table 1). Model 3 outperformed the other regression neural networks, as indicated by a higher variance explained and decreased MSE and MAE compared to the other models (Fig. 3, Model 3: $R^2 = 0.52$, MSE = 128.99, MAE = 8.65). Interim THI category was identified as the most important feature with an average absolute SHAP value of 9.25, followed by baseline THI category with an average absolute SHAP value of 1.59. Feature importance for individual predictions of Model 3 is visualized in Fig. 4. High interim THI group (i.e., severe and catastrophic) again corresponds with a higher SHAP values. This may suggest that a high interim THI score may indicate a higher final THI value. Exemplar model predictions are visualized in Fig. 5, in which there are clear distinctions between interim THI category and the predicted final THI score (Fig. 5).

We then further test the clinical applicability of the model, in which we sought to determine if we can train a model to classify participants into responder and non-responders to 12 weeks of treatment (Model 4). We used the same input variables as Model 3 to train Model 4, to predict pre-determined subgroups in which subjects were considered a significant responder to treatment if their final THI score decreased by 7 points or more from their baseline THI score. Model 4 was able to classify responders and non-responders with an average accuracy of 72 %, AUC-PRC of 0.86, AUC-ROC of 0.66, and f1 of 0.81. To further understand the interpretability of Model 4, we again use SHAP as a method to determine feature importance for individual predictions of each class, visualized in Fig. 6. Low interim THI groups (i.e., slight or mild) and high baseline THI groups (i.e., severe or catastrophic) have higher SHAP values corresponding to our responder class. Moreover, high interim THI groups (i.e., severe or catastrophic) and low baseline THI groups (i.e., mild) have higher SHAP values corresponding to our non-responder class.

4. Discussion

The advancement of machine and deep learning models has led to a rise in implementation of personalized medicine approaches, in which

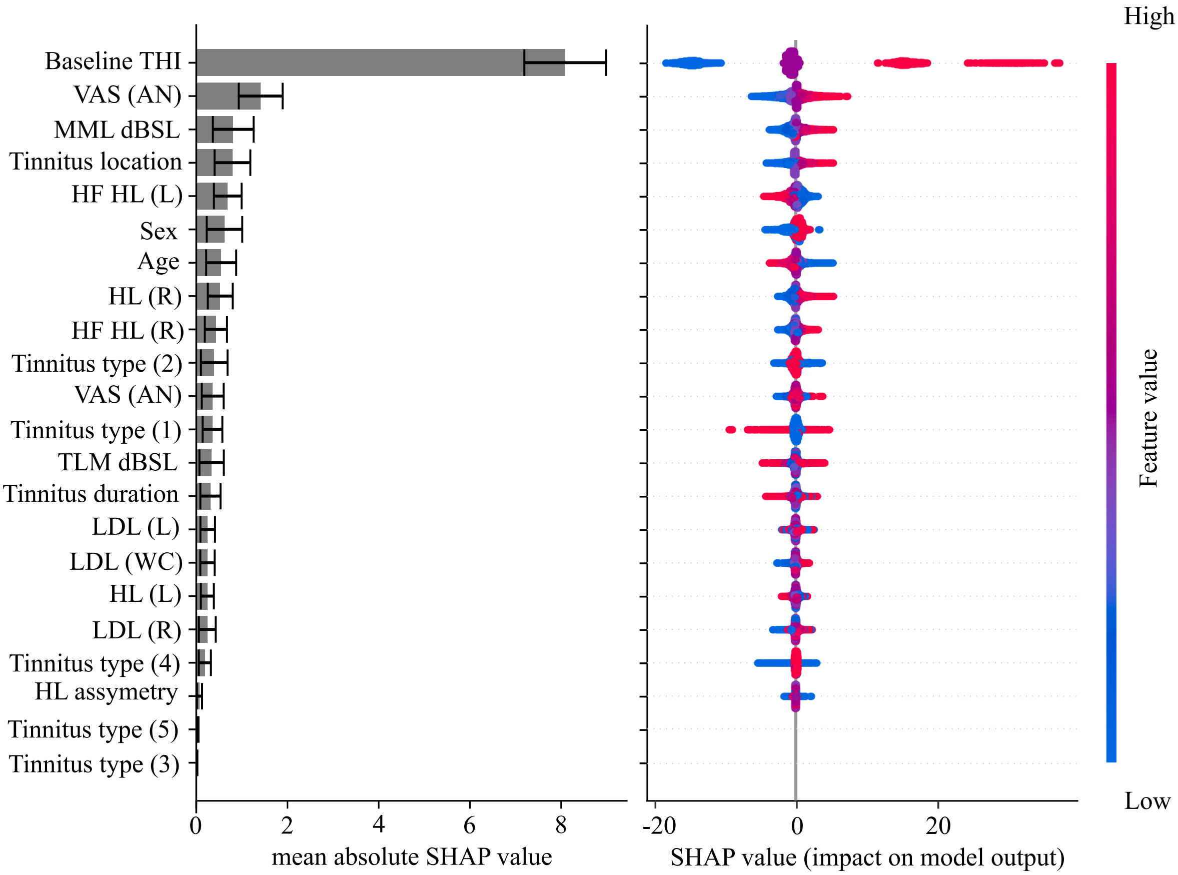

Fig. 2. Feature importance for Model 1. (A) Average top features across 50 iterations. Baseline THI group, VAS (annoyance), MML dbsl, tinnitus location, and high frequency hearing loss (left) are the top 5 features. (B) Summary plot of all predictions across 50 model iterations combing feature importance with feature effects. Each point corresponds to the Shapley value (x-axis) for a feature within a given instance, and the color of the point represents the feature value from low to high. Points are jittered in the y-axis. Features are ordered according to their importance.

computational techniques support clinical decision making to improve patient outcomes [1,2,4]. A supervised deep learning approach can extract and process information from given data to make individual predictions on new data [2]. In the current study, we tested the feasibility of using deep learning to predict tinnitus patient outcomes following treatment. We illustrated that deep learning models outperform general machine learning models, and that feature reduction slightly improves deep learning performance. Furthermore, we conclude that baseline and interim THI group emerge as the top features across both regression and classification deep learning models.

Our results suggest that interim THI could be a primary feature towards responder or non-responder classification of treatment. Furthermore, this may suggest that response to 6 weeks of treatment could be indicative to response to 12 weeks of treatment. Our feature importance methods highlighted the top features needed to predict treatment response, including THI, VAS for annoyance, MML dBSL, tinnitus location and high frequency hearing loss, in which these features are the most beneficial for AI based decision-making with the given data. While these results hold great promise for clinical applicability, further research is needed to validate prior to real world implementation.

Given that our findings put forth several measures of tinnitus including THI, annoyance, locations, loudness, and hearing loss as important features for determining treatment response, there is reasonable basis to postulate how the implementation of this, or a similar model, may look in a clinical setting. Our belief is that models developed to predict treatment response will assist clinicians in deciding on treatment approaches with patients. In such a scenario, clinicians can use gathered information from a patient's appointments, and supply information to a model such those described here. This model will, using previously trained data, provide a prediction as to whether or not the patient in question would respond to the proposed treatment. In one theoretical scenario, a classification model, such as the DL model 4 we produce here, could be implemented as part of a tinnitus patient appointment where the approach of treatment is being discussed. In either a web-based or internally-supported application, the patient's demographic and tinnitus information will be provided by the patient, and suggested treatment inputted by the clinician. The application will then return a prediction as to whether the patient is likely to respond to the suggested treatment, allowing the clinician to then decide whether or not to pursue this treatment approach, or consider an alternative. Indeed, while the focus of this study was to consider the treatment response prediction of a specific treatment (i.e., bimodal stimulation), as stimulation parameters are further explored, a future model may not only predict whether a patient will respond, but may also suggest alternative treatment approaches based on similar patient profiles that responded to different treatment approaches, and indeed similar methods have been considered and tested in other disorders [41].

Fluency with machine learning may also create a barrier to the implementation of AI-guided clinical decision making, and indeed it is impractical that all clinicians wishing to utilize a treatment decision-making aid such as a classification model should rigorously study the history of machine learning in a clinical setting. Thus, the interpretation of the model, and the mechanism(s) by which it came to the decision, are critical factors when considering implementing AI. For example, when a model makes a classification decision (e.g., Responder or Non-responder), this decision is most often made using a probability. Thus, an effective and transparent model may provide the probability of the patient responding, as opposed to a simple classification category. This

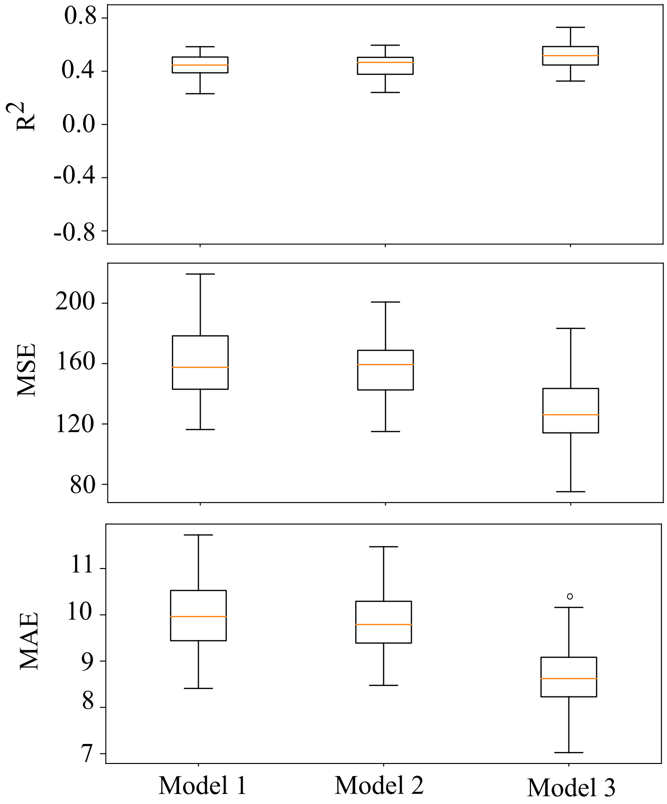

Fig. 3. Model 3 outperforms the other deep learning regression models. We compared model performance, quantified by variance explained (R^2), mean squared error (MSE), and mean absolute error (MAE), across 3 deep learning regression models. Model 3 has greater R^2, and reduced MSE and MAE compared to Models 1 and 2. Box plots illustrate the data distribution over the course of 50 iterations.

will also allow clinicians to weight the input of the model. For example, both a 90 % and 54 % 'Responder' prediction output will be classed as 'Responder', although one is notably more extreme than the other. Furthermore, SHAP values in the models above indicate which features the model is using the most to produce predictions, and thus these 'inner-workings' may improve the clinicians understanding of how the model is coming to the conclusion. For example, if a model such as ours is rating the baseline THI score as an important feature, and the clinician is unable to provide this value, the ultimate treatment response prediction might be considered unreliable compared to a prediction made using baseline THI. Fundamental understanding of these aspects, which can easily be integrated into applications using warning systems (e.g., warning the clinician that they are missing key clinical information) or user manuals, improving the transparency of the model while also informing the clinician of how the decision is being made. Ultimately, computer-based applications which reduce the complexity of a deep learning model to a simple input-output interface, and with the addition of necessary staff training and warning systems, are low-impact and implementable addition to clinical appointments, and can likely be integrated into existing workflows without significant time or resource allocation with proper interface development and piloting. The results of this study have also arguably given sufficient support for its use, such that the next steps will be to implement a piloting of a model in a tinnitus clinic where clinicians can engage with, and review, the decisions of the

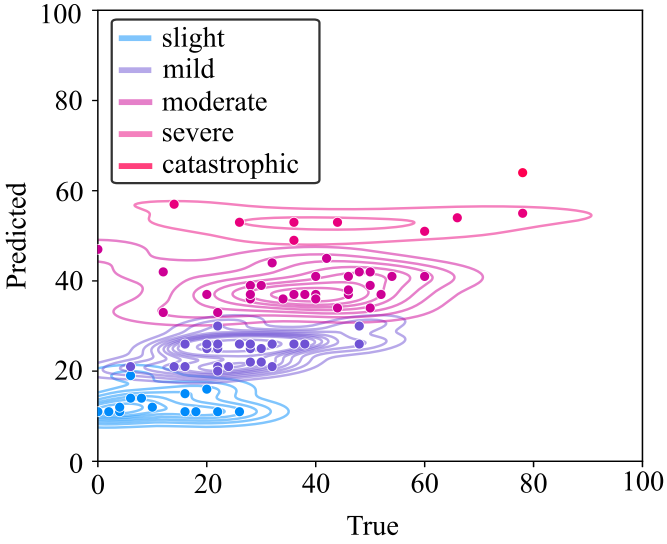

Fig. 5. Example of predicted final THI vs true THI in a single iteration of Model 3. Individual points are the predicted final THI score, with their true final THI values on the x-axis. The color of the points correspond to the THI subgroup of the individual at the interim timepoint, 6 weeks after starting treatment. Lines represent the kernel density estimate to visualize the distribution of points. Predictions of final THI score are largely dependent on interim THI group.

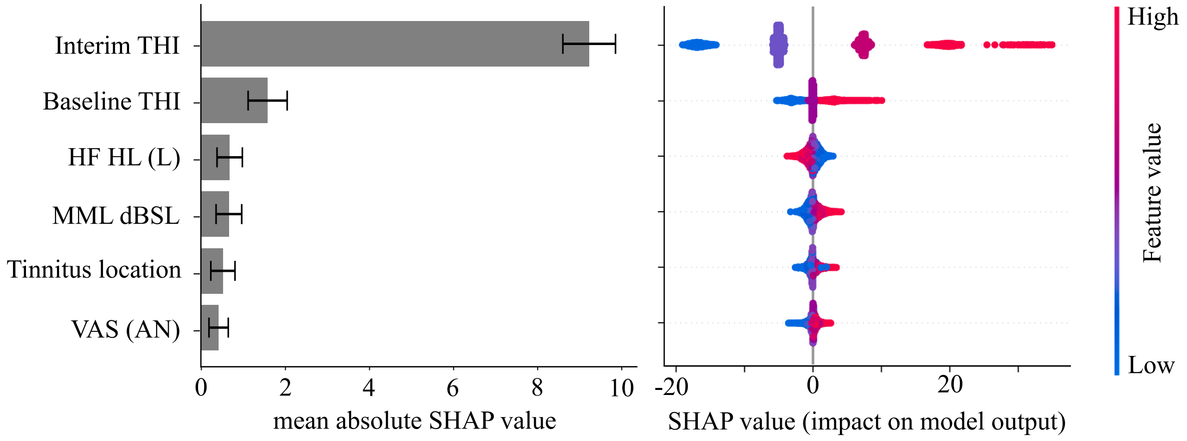

Fig. 4. Feature importance for Model 3. (A) Average top features across 50 iterations for Model 3. Interim THI group and Baseline THI group are the most important features. (B) Summary plot of all predictions across 50 model iterations combing feature importance with feature effects. Each point corresponds to the Shapley value (x-axis) for a feature within a given instance, and the color of the point represents the feature value from low to high. Points are jittered in the y-axis. Features are ordered according to their importance.

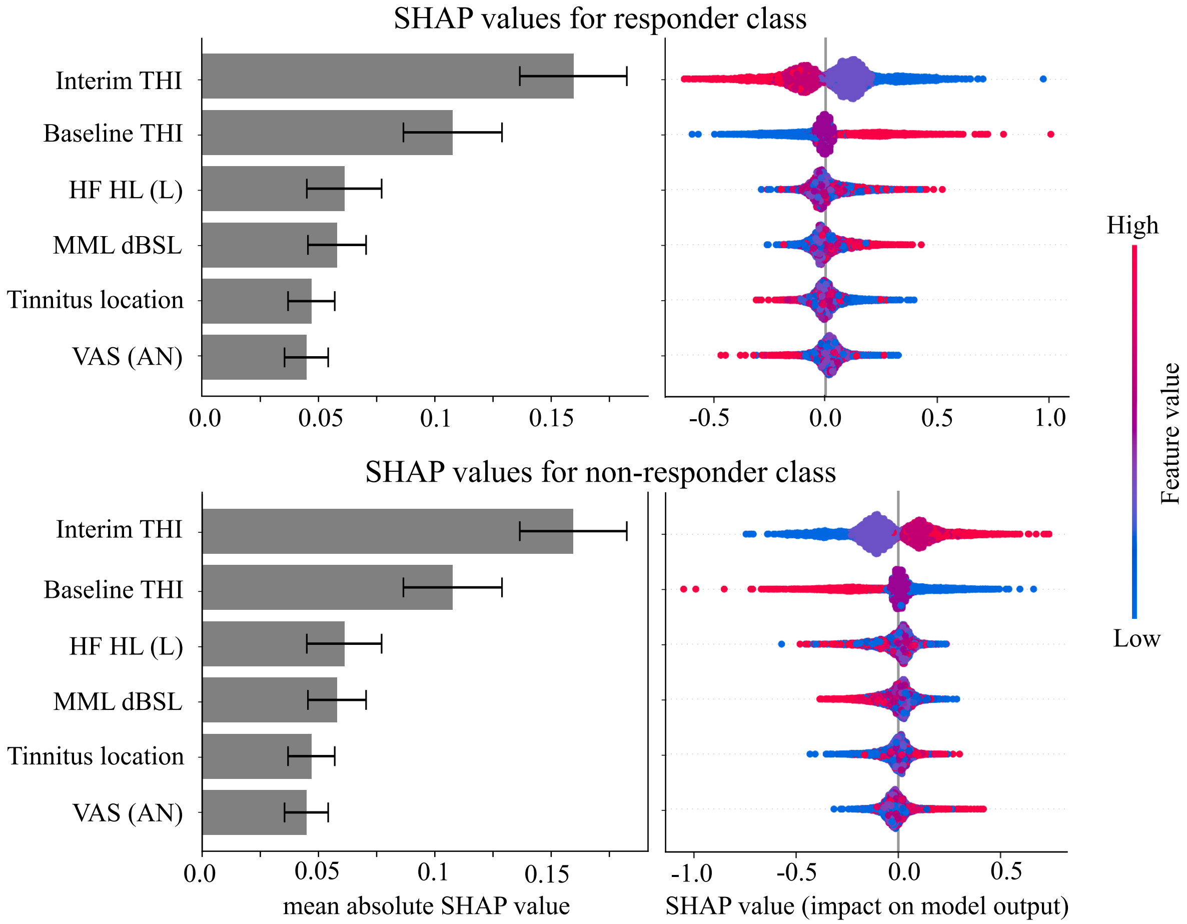

Fig. 6. Feature importance for Model 4, for the (top) responder class and (bottom) non-responder class. Interim THI group and Baseline THI group are the most important features for both responders and non-responders. (B) Summary plot of all predictions across 50 model iterations combing feature importance with feature effects. Each point corresponds to the Shapley value (x-axis) for a feature within a given instance, and the color of the point represents the feature value from low to high. Points are jittered in the y-axis. Features are ordered according to their importance.

model in a real-world clinical setting.

Limitations within this study merit consideration. Clinical features of individuals with tinnitus vary considerably, which exacerbates variable response to treatments [34]. Inclusion and exclusion criteria of studies can reduce heterogeneity by excluding specific subgroups of tinnitus. However, reducing the heterogeneity can limit the input variables, thereby limiting the scope of the model. Additionally, it's possible that factors outside of of our measures may be more insightful or indicative of a person's likelihood to respond (or not-respond) to this treatment. Furthermore, our output variable, THI score, predominantly assesses the emotional and functional impact of tinnitus, rather than the perception of tinnitus. Whether these measures are best attribute to use to evaluate treatment efficacy is yet to be understood. Recently, a study utilized EEG data from tinnitus patients to predict treatment response with a 99 % accuracy, highlighting the robustness of EEG signals as a predictor [17]. However, collecting and analysing EEG data is time consuming, and using a tailored questionnaire to predict patient outcomes could reduce time spent in the clinic, quickly informing patients regarding the probability of treatment success.

Another challenge within personalized medicine is the lack in standardized reporting of model performance [35]. A wide range of performance metrics are applied to determine model classification performance including accuracy, error rate, area under the ROC curve [2,36,37]. While ROC is the most popular to report for binary classification, studies suggest Precision-Recall curve is a better alternative for imbalanced datasets [36]. The Precision-Recall curve is ideal for imbalanced datasets in which the positive class is the minority, as it evaluates the fraction of true positives among positive predictions. In the current study, we observe an imbalance in which the positive class, the responders, is the majority, and the negative class, the non-responders, are in the minority. This can likely be attributed to the fact that individuals who perceive they are not responding to treatment may be more likely drop out of the study [38]. Therefore, Precision-Recall curve may not be the best measure to evaluate model performance. Currently, there is no single measure that can evaluate all the desirable components or accurately portray the clinical applicability of a model [36,39]. In attempt to address this, several measurements are reported instead to summarize the performance [39,40]. We also recognize that in order to test model generalizability, the model would ideally be tested on an external dataset such that the performance can be gauged using a variety of data. While this study relied on, to our knowledge, the only available clinical data investigating the effect of bimodal stimulation on tinnitus improvement, future studies or alternate implementations of deep learning models for predicting tinnitus treatment outcomes would benefit from validation with external datasets. This may also include the application of the model to different populations, ages, and treatment methods.

Through our analyses, we establish the initial framework for future development of a robust model to identify subgroups best suited for bimodal stimulation treatment [18,19]. To our knowledge, this is the first study to utilize a deep learning approach to predict patient outcomes with patient-reported outcome measures in individuals with tinnitus and highlight the important measures for model predictions. These results show great promise in using deep learning for clinical applications, in which we can use a data-driven approach to reduce the number of measures necessary to collect for treatment prediction. While deep learning modelling has the potential to translate previous clinical results into future clinical predictions, future research is needed to optimize deep learning modelling as a decision support tool for an efficient personalized medicine approach in a clinical setting.

CRediT authorship contribution statement

Katherine S. Adcock: Writing – review & editing, Writing – original draft, Visualization, Validation, Project administration, Methodology, Formal analysis, Conceptualization. **Gabriel Byczynski:** Writing – review & editing. **Emma Meade:** Writing – review & editing, Data curation. **Sook Ling Leong:** Writing – review & editing, Data curation. **Richard Gault:** Writing – review & editing, Methodology. **Hubert Lim:** Writing – review & editing, Funding acquisition. **Sven Vanneste:** Writing – review & editing, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Sook Ling Leong reports a relationship with Neuromod Devices that includes: employment. Emma Meade reports a relationship with Neuromod Devices that includes: employment. Hubert Lim reports a relationship with Neuromod Devices that includes: employment. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Verma D, Bach K, Mork PJ. Application of machine learning methods on patient reported outcome measurements for predicting outcomes: a literature review. Informatics 2021;8.

[2] Peng J, Jury EC, Dönnes P, Ciurtin C. Machine learning techniques for personalised medicine approaches in immune-mediated chronic inflammatory diseases: applications and challenges. Front Pharmacol 2021;12.

[3] Papadakis G, et al. Deep learning opens new horizons in personalized medicine (Review) Biomed. Reports 2019;10:215–7.

[4] Collin CB, et al. Computational models for clinical applications in personalized medicine—guidelines and recommendations for data integration and model validation. J Personalized Med 2022;12:166.

[5] Zhang S, Bamakan SMH, Qu Q, Li S. Learning for personalized medicine: a comprehensive review from a deep learning perspective. IEEE Rev. Biomed. Eng. 2019;12:194–208.

[6] Ali F, et al. A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. Inf Fusion 2020;63:208–22.

[7] Antman EM, Loscalzo J. Precision medicine in cardiology. Nat Rev Cardiol 2016; 13:591–602.

[8] Bzdok D, Meyer-Lindenberg A. Machine learning for precision psychiatry: opportunities and challenges. Biol. Psychiatry Cogn. Neurosci. Neuroimaging 2018; 3:223–30.

[9] Esteva A, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115–8.

[10] Kim J, Shin H. Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data. J Am Med Inf Assoc 2013;20:613–8.

[11] Peng J, Jury EC, Dönnes P, Ciurtin C. Machine learning techniques for personalised medicine approaches in immune-mediated chronic inflammatory diseases: applications and challenges. Front Pharmacol 2021;12:1–18.

[12] Wei Z, et al. From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. PLoS Genet 2009;5.

[13] Cardon E, et al. Random forest classification to predict response to high-definition transcranial direct current stimulation for tinnitus relief: a preliminary feasibility study. Ear Hear 2022;43:1816–23.

[14] Simoes J, et al. Toward personalized tinnitus treatment: an exploratory study based on internet crowdsensing. Front Public Health 2019;7:1–10.

[15] Cederroth CR, et al. Editorial: towards an understanding of tinnitus heterogeneity. Front Aging Neurosci 2019;11.

[16] Collin CB, et al. Computational models for clinical applications in personalized medicine—guidelines and recommendations for data integration and model validation. J Personalized Med 2022;12.

[17] Doborjeh M, et al. Prediction of tinnitus treatment outcomes based on EEG sensors and TFI score using deep learning. Sensors 2023;23:902.

[18] Conlon B, et al. Bimodal neuromodulation combining sound and tongue stimulation reduces tinnitus symptoms in a large randomized clinical study. Sci Transl Med 2020;12. https://www.science.org.

[19] Conlon B, et al. Different bimodal neuromodulation settings reduce tinnitus symptoms in a large randomized trial. Sci Rep 2022;12.

[20] Newman CW, Sandridge SA, Jacobson GP. Psychometric adequacy of the Tinnitus Handicap Inventory (THI) for evaluating treatment outcome. J Am Acad Audiol 1998;9:153–60.

[21] D'Arcy S, et al. Bi-modal stimulation in the treatment of tinnitus: a study protocol for an exploratory trial to optimise stimulation parameters and patient subtyping. BMJ Open 2017;7.

[22] Conlon B, et al. Noninvasive bimodal neuromodulation for the treatment of tinnitus: protocol for a second large-scale double-blind randomized clinical trial to optimize stimulation parameters. JMIR Res. Protoc. 2019;8:1–15.

[23] McCombe A, et al. Guidelines for the grading of tinnitus severity: the results of a working group commissioned by the British Association of Otolaryngologists, Head and Neck Surgeons, 1999. Clin Otolaryngol Allied Sci 2001;26:388–93.

[24] Zeman F, et al. Tinnitus handicap inventory for evaluating treatment effects: which changes are clinically relevant? Otolaryngol Head Neck Surg 2011;145:282–7.

[25] Waqar M, et al. An efficient SMOTE-based deep learning model for heart attack prediction. Sci Program 2021;2021.

[26] Chicho BT, Bibo Sallow A. A comprehensive survey of deep learning models based on Keras framework. J. Soft Comput. Data Min. 2021;2:49–61.

[27] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. 2017-Decem 2017:4766–75.

[28] Hittmeir M, Ekelhart A, Mayer R. Utility and privacy assessments of synthetic data for regression tasks. Proc. - 2019 IEEE Int. Conf. Big Data, Big Data 2019;2019: 5763–72. https://doi.org/10.1109/BigData47090.2019.9005476.

[29] Johnson HR, et al. A machine learning approach for using the postmortem skin microbiome to estimate the postmortem interval. PLoS One 2016;11:1–23.

[30] Chen H, Lundberg SM, Lee SI. Explaining a series of models by propagating Shapley values. Nat Commun 2022;13:1–15.

[31] Jalali A, et al. Deep learning for improved risk prediction in surgical outcomes. Sci Rep 2020;10:1–13.

[32] Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recogn 1997;30:1145–59.

[33] Davis J, Goadrich M. The relationship between precision-recall and ROC curves. ACM Int. Conf. Proceeding Ser. 2006;148:233–40.

[34] Baguley D, McFerran D, Hall D. Tinnitus. Lancet 2013;382:1600–7. Elsevier B.V.

[35] Plant D, Barton A. Machine learning in precision medicine: lessons to learn. Nat Rev Rheumatol 2021;17:5–6.

[36] Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 2015;10: 1–21.

[37] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29–36.

[38] Kaplan RM, Atkins CJ. Selective attrition causes overestimates of treatment effects in studies of weight loss. Addict Behav 1987;12:297–302.

[39] Shah NH, Milstein A, Bagley PhD SC. Making machine learning models clinically useful. JAMA 2019;322:1351.

[40] Liu X, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit. Heal. 2019;1:e271–97.

[41] Dubey S, et al. Using machine learning for healthcare treatment planning. Front. Artif. Intell. 2023;6:1124182.