Hearing Research 469 (2026) 109473

Contents lists available at ScienceDirect

Hearing Research

journal homepage: www.elsevier.com/locate/heares

Towards precision medicine for otology and neurotology: Machine learning applications and challenges

Katherine Adcock [a,1], Elva Arulchelvan [a,1], Nathan Shields [a,1], Sven Vanneste [a,b,*]

[a] Lab for Clinical and Integrative Neuroscience, Trinity Institute for Neuroscience, School of Psychology, Trinity College Dublin, Ireland
[b] Global Brain Health Institute of Neuroscience, Trinity College Dublin, Dublin, Ireland

A R T I C L E I N F O

Keywords:
Otology
Tinnitus
Precision medicine
Artificial intelligence
Machine learning

A B S T R A C T

Advances in artificial intelligence, particularly machine learning and deep learning, in conjunction with the rise of personalised medicine, can facilitate tailored decision-making for diagnoses, prognoses, and treatment responses based on individual patient data. The multifaceted nature of symptoms and disorders in (neuro)otology, with their diverse aetiologies and subjective characteristics, makes this field an ideal candidate for computational personalised medicine. This narrative review critically synthesises applications of machine learning and deep learning in otology and neurotology published between 2013 and 2025. Relevant studies were identified through targeted searches of PubMed, Scopus, and Google Scholar using combinations of terms related to artificial intelligence, tinnitus, cochlear implants, and otologic or neurotologic disorders. Only peer-reviewed articles focusing on human applications of machine learning or deep learning in these fields were included, excluding theoretical papers or animal studies. Recent breakthroughs, such as the Whisper speech recognition model for cochlear implant simulations and large language models for refining tinnitus subgroup identification and therapy predictions, underscore the transformative potential of AI in improving clinical outcomes. This review is distinct in its emphasis on these emerging technologies and their integration into multimodal datasets, combining imaging, audiometric data, and patient-reported outcomes to refine diagnosis and treatment approaches. However, challenges including the lack of standardisation, limited generalisability of models, and the need for improved frameworks for multimodal data integration impede rigorous and reproducible implementation, topics that are critically explored in this review. Here, we explore the applications of machine learning, deep learning, and large language models in tinnitus, cochlear implants, and (neuro)tology, providing a critical analysis of recent advancements, persistent challenges, and recommendations for future research. By addressing these challenges and implementing recommended strategies, this review outlines a pathway for integrating cutting-edge artificial intelligence tools into clinical practice, underscoring their immense potential to revolutionise precision medicine in otology and neurotology and improve patient outcomes.

1. Introduction

The rapid advancement of computational methods has led to a rise in personalised medicine aiming to improve patient outcomes. Personalised medicine is a tailored approach to predict diagnoses, prognoses, or response to treatment based on individual patient data (Papadakis et al., 2019; Peng et al., 2021; Zhang et al., 2019). The growing availability of healthcare databases allow researchers to make personalised predictions based on common features such as genetics, lifestyle and environmental factors, socio-demographics, or disease phenotypes, therefore facilitating tailored patient management. Personalised medicine can use the power of computational and statistical approaches to identify complex patterns and hidden correlations in health care datasets for which traditional exploratory and inferential statistical analyses are unable to detect (Peng et al., 2021). Recent advancements, such as the Whisper speech recognition model for cochlear implant simulations and the application of large language models (LLMs) for tinnitus subgroup identification, highlight the transformative potential of computational methods in otology and neurotology (Sinha and Azadpour, 2024). These tools exemplify how state-of-the-art machine learning models can

* Corresponding author at: 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).
[1] These three authors contributed equally to this work.

https://doi.org/10.1016/j.heares.2025.109473
Received 4 July 2025; Received in revised form 23 October 2025; Accepted 13 November 2025
Available online 13 November 2025
0378-5955/

uncover patterns in diverse and complex datasets, offering new avenues for personalised therapeutic approaches. The adoption of computational approaches such as machine learning (ML) and deep learning (DL) holds promise to improve diagnostic and prognostic tools, stratification of disease subtypes, and targeted treatment plans.

Significant progress continues to be made in many medical fields, with machine learning being leveraged to enable personalised medicine approaches from drug development to disease characteristics, and therapeutic effects for several diseases including diabetes, cancer, heart disease, chronic inflammatory disease, and psychiatric disease (Peng

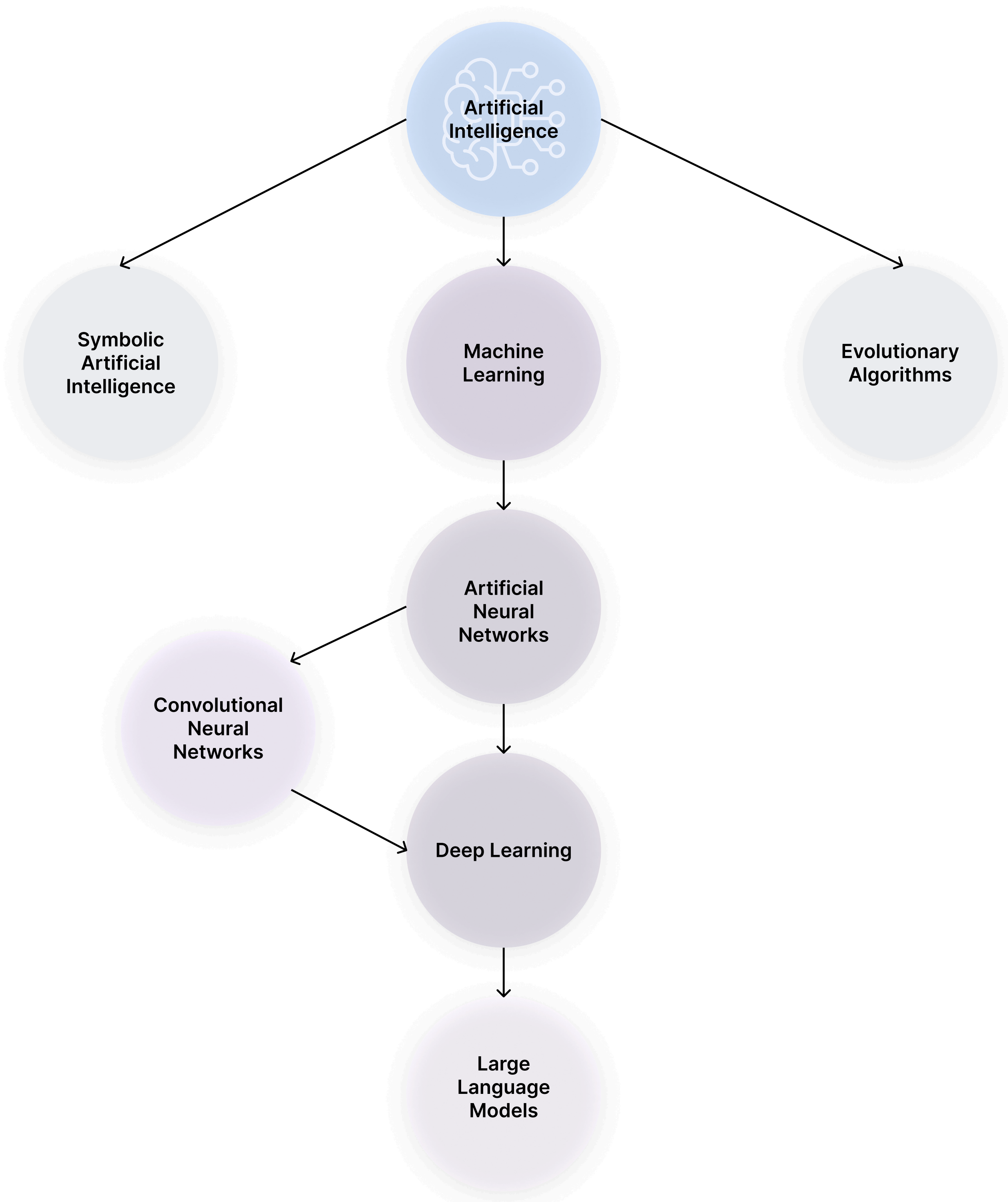

Fig. 1. Representation of deep learning as a subset of machine learning. In this figure, deep learning is visualised as a subset of machine learning, which is a form of artificial intelligence. Machine learning methods will typically require structured data sets alongside human input to extract important features, and ultimately enable pattern identification to form predictions. Deep learning, a subset of machine learning, is distinct from conventional machine learning methods in that it requires larger data sets but can process unstructured data and make predictions without human instruction or manual feature extraction.

et al., 2021; Ali et al., 2020; Antman and Loscalzo, 2016; Bzdok and Meyer-Lindenberg, 2018; Esteva et al., 2017; Kim and Shin, 2013; Wei et al., 2009). In cases of complex heterogeneous disease, deep learning (a subset of machine learning, see Fig. 1 for overview) can be a particularly effective strategy for individualised patient management and personalised medicine, given its capacity to process and analyse complex, unstructured, multi-dimensional data. In particular, the integration of multimodal datasets—combining imaging, audiometric, and patient-reported data—has become a focus of recent research. For example, the work by Essaid et al. (2024) demonstrates how multimodal artificial intelligence (AI) approaches can enhance surgical planning and auditory outcomes in cochlear implant research, setting a new standard for precision medicine in neurotology. As such, it is not surprising that recent studies have highlighted how utilising computational techniques could enable personalised medicine in specific otology and neurotology management and to improve clinical decision making (Cardon et al., 2022; Simoes et al., 2019). Furthermore, advanced models such as Whisper and LLMs underline the importance of large, high-quality datasets in driving innovation. These models demonstrate how AI-driven tools can address challenges like subgroup heterogeneity in tinnitus and improve auditory simulation accuracy in cochlear implants, further bridging the gap between research and clinical practice (Sinha and Azadpour, 2024). Machine learning approaches (including both traditional and deep learning methods), in conjunction with growing healthcare databases, give researchers the opportunity to improve clinical diagnostic, prognostic, and treatment decision-making through individualised, tailored predictions to an individual's multimodal data—encompassing genetic markers, imaging results, audiometric profiles, and patient-reported outcomes. These tailored predictions can include estimating a patient's risk of developing a specific auditory or vestibular disorder, forecasting disease progression (for example, hearing deterioration in Meniere's disease), identifying the most effective therapeutic approach (such as the optimal neuromodulation or behavioural therapy protocol for tinnitus), and anticipating treatment response or potential side effects based on each patient's unique clinical, genetic, and lifestyle characteristics. This personalised predictive capacity exemplifies how machine learning can enable precision medicine approaches in otology and neurotology (Peng et al., 2021; Zhang et al., 2019; Ali et al., 2020; Esteva et al., 2017; Collin et al., 2022; Doborjeh et al., 2023).

This perspective will discuss the distinctions between traditional machine learning and deep learning approaches, explore their applications in otology and neurotology—including tinnitus and cochlear implants—and provide actionable recommendations to address challenges such as data standardisation, interpretability, and ethical considerations, ensuring sustained progress in the field. A narrative review format was chosen to allow for critical synthesis and integration of emerging multidisciplinary literature spanning clinical otology, computational neuroscience, and artificial intelligence.

2. Brief overview of machine and deep learning

Machine learning algorithms, a subset of AI, can identify complex patterns and trends within a data set through learning from data. AI encompasses various computational techniques, including rule-based systems and robotics, alongside machine learning and deep learning. For instance, early medical diagnostics used expert systems based on programmed rules rather than learning from data, whereas machine learning and deep learning leverage data-driven approaches for improved adaptability and accuracy (Rahman et al., 2024).

Conventional machine learning algorithms can be supervised or unsupervised, with the main distinction between the two being whether the data are labelled or unlabelled, respectively (Papadakis et al., 2019; Rajoub, 2020). Labelled data refers to datasets where each input is paired with a known output, tag, or class, with the help of human annotators. For example, in a supervised machine learning task, a labelled dataset might consist of otoscopic images of ears (inputs) annotated with diagnoses like otitis media or a healthy ear (outputs).

It is important to clarify how machine learning differs from, yet complements, traditional statistical approaches. While traditional statistics typically test predefined hypotheses and evaluate relationships between variables using measures such as p-values or confidence intervals, machine learning excels at identifying complex, multidimensional relationships in large datasets without requiring an explicit hypothesis (Leo, 2001). These approaches are not mutually exclusive—statistical testing remains critical for validating model outputs, assessing significance, and ensuring that observed associations are robust across subgroups. In practice, traditional statistics and ML are complementary: statistical methods provide interpretability and inferential rigor, whereas ML offers scalability and predictive strength for nonlinear, high-dimensional data.

This allows the model to learn from the examples and make accurate predictions on new, unseen data, as demonstrated in recent deep-learning systems developed to classify middle-ear conditions from otoscopic images (Dubois et al., 2024; Myburgh et al., 2018). Common supervised learning approaches utilise classification or regression algorithms to investigate relationships within labelled datasets to predict either a categorical classification or continuous outcome score, respectively (Rajoub, 2020). These algorithms are useful for understanding commonalities between predetermined subgroups or to estimate an outcome score for patients. Other popular supervised learning algorithms include linear models, support vector machine, and tree-based models (Lo Vercio et al., 2020; MacEachern and Forkert, 2021). Although these algorithms are most commonly applied in supervised contexts, certain variants—such as one-class support vector machines for anomaly detection or clustering-based decision trees—can also be adapted for unsupervised or semi-supervised learning tasks.

Popular unsupervised machine learning approaches include clustering and dimensionality-reduction algorithms such as K-means clustering, hierarchical clustering, and principal component analysis (PCA), which can uncover the inherent structure of unlabelled data or stratify samples into homogeneous groups based on shared characteristics (Lo Vercio et al., 2020; MacEachern and Forkert, 2021). These algorithms are particularly useful for exploratory data analysis, for instance, identifying latent subgroups within heterogeneous clinical datasets. While K-nearest neighbour (K-NN) is commonly employed as a supervised method for classification or regression, these unsupervised approaches serve a complementary role by revealing patterns and relationships without requiring labelled outputs. Whereas traditional and unsupervised learning approaches are primarily applied to structured datasets such as imaging or audiometric data, recent advances in artificial intelligence have enabled the analysis of unstructured information, including clinical narratives and radiology reports, through large language models.

Additionally, LLMs, a recent breakthrough in AI, demonstrate the ability to analyse unstructured data such as clinical notes and offer personalised recommendations in otology and neuro-otology, broadening the potential applications of AI in these fields. In neurotology, LLMs have shown promise in tasks like classifying tinnitus subgroups, analysing clinical notes for patient stratification, and refining radiological diagnostics, paving the way for personalised care (Moura et al., 2024).

LLMs, such as GPT-4, GPT-5 and Google T5, represent a significant advancement in AI, particularly in their ability to process and analyse unstructured data like clinical notes, radiology reports, and patient histories (Zahid et al., 2024; Wang and Zhang, 2024; Handler et al., 2025). These models excel at uncovering hidden patterns in textual data, offering personalised recommendations and insights. For example, in neurotology, LLMs have been employed to classify patient subgroups in tinnitus research and to standardise radiology reports, improving diagnostic accuracy for conditions such as otosclerosis and otitis media (Tang et al., 2024). This adaptability to diverse data types positions

LLMs as a transformative tool in precision medicine.

Deep learning, a subset of machine learning, utilises multi-layered algorithms to extract and process information from data, enabling predictions or classifications on new data (Papadakis et al., 2019; Peng et al., 2021). While often inspired by the structure of biological neural networks in the brain, not all implementations of deep learning architectures explicitly mimic brain functionality. For instance, convolutional neural networks (CNNs) are loosely modelled on the visual cortex and excel in processing spatial data like images, whereas other architectures, such as transformers and recurrent neural networks (RNNs), are designed with different computational goals in mind (Lo Vercio et al., 2020; MacEachern and Forkert, 2021). These models use artificial neurons and weights, where the weights adjust the strength of connections between neurons based on their importance, allowing the network to learn patterns during training.

In otology, CNNs have been used to analyse otoscopic images, achieving diagnostic accuracies exceeding 90 % for conditions like otitis media, outperforming general practitioners (Ellis et al., 2022). Similarly, clustering methods have identified tinnitus patient subgroups based on coping mechanisms or distress levels, providing clinicians with actionable insights to tailor therapeutic approaches (Lo Vercio et al., 2020; MacEachern and Forkert, 2021; Makar, 2021). These examples highlight the ability of machine learning to address the heterogeneity of neurotologic disorders, supporting improved patient outcomes through precision medicine.

An important distinction between deep learning and conventional machine learning models is that deep learning methods automatically determine important features for the model, bypassing the manual feature extraction step required by traditional approaches (see Fig. 2). The term "deep" refers to the presence of multiple hidden layers or parameters in a neural network. Modern deep learning architectures often consist of tens or hundreds of layers and millions (or even billions) of parameters, such as weights and biases, optimised during training to achieve high performance. These layers process data hierarchically, with each layer extracting increasingly abstract features. For example, a CNN designed for image recognition might include layers for detecting edges, textures, and shapes, culminating in high-level object identification. The ability of deep learning models to handle unstructured data and capture complex relationships makes them highly suitable for studying heterogeneous disorders in medicine. Consequently, deep learning has become a transformative tool in artificial intelligence, frequently outperforming traditional machine learning algorithms (Lo Vercio et al., 2020; MacEachern and Forkert, 2021). However, the choice of whether to use a traditional machine learning algorithm or deep learning model will depend on the problem's specific characteristics and available data. For example, given that deep learning models (such as neural networks) can analyse a more complex set of, and significantly larger number of features, they require a much larger training set of data in order to ensure high accuracy (Lo Vercio et al., 2020). Nevertheless, often in real-world clinical situations, this is not always feasible, a point that is explored further in the 'challenges' section of this paper. As such, both traditional machine learning and deep learning approaches can provide a better understanding of complex heterogeneous disorders, encompassing multi-factorial conditions, and are ideal in predicting disease or treatment outcome, or stratification of disease subtypes (Peng et al., 2021) See Fig. 2 for overview.

In relation to model evaluation and performance, it is best practice for an algorithm to be 'trained' on a large proportion of a dataset, with

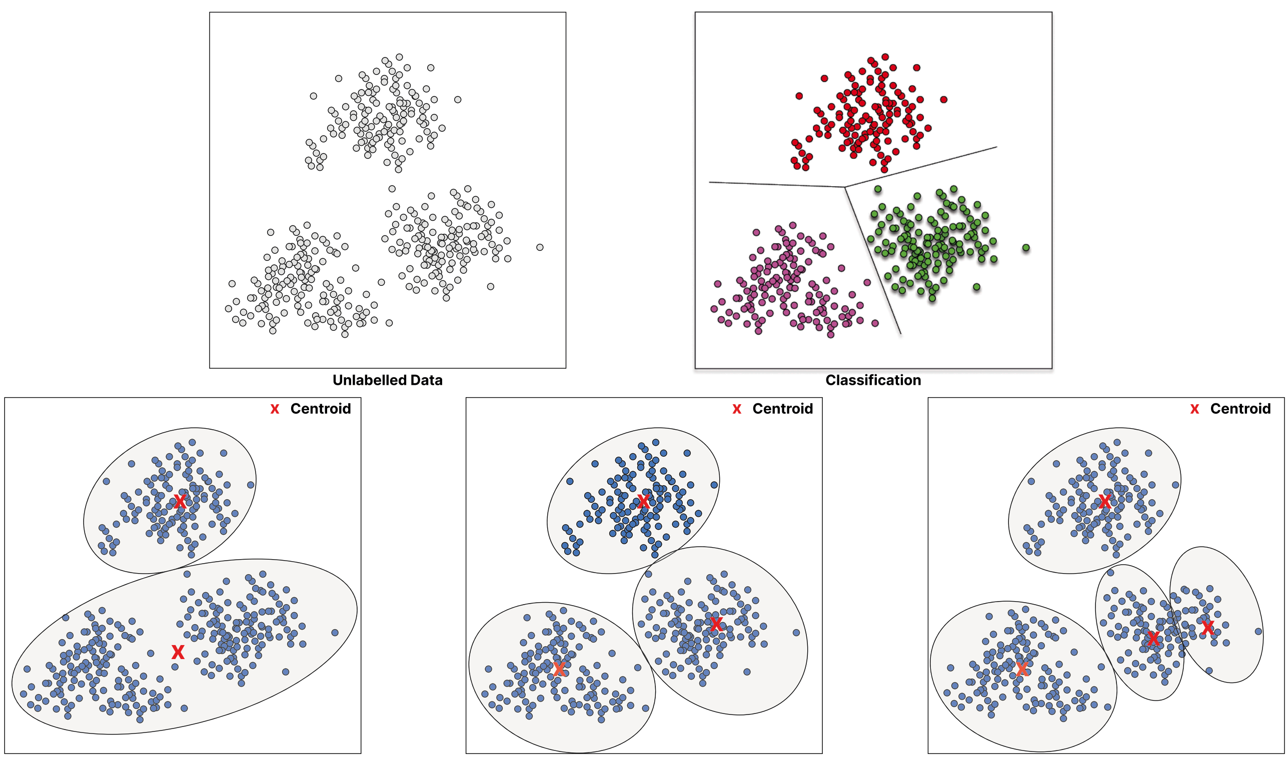

Fig. 2. Exemplar workflows for utilisation of conventional machine learning or deep learning approaches for precision medicine. Both traditional machine learning techniques and deep learning methods can be leveraged with a common goal of precision medicine in otology in mind. In this figure, exemplar steps in machine learning workflows are outlined to visually represent how certain stages in this process differ for traditional machine learning approaches vs deep learning methods. For example, selection of which specific machine learning approach is appropriate will depend in part on what data is available (both the type and size of the data set) and whether manual or automatic feature extraction is required. See the data input, feature extraction and model development steps in the exemplar workflows. In any machine learning workflow, model evaluation should include a cross-validation step across a diversified data set. Similarly, testing on an external data set, with a number of performance metrics, is essential for model validation and to ensure clinical validity in the field of otology and neurotology. With continued development and refinement of a model, computational insights can be leveraged to facilitate tailored treatment decisions and precision medicine.

the remaining smaller proportion split into a 'test' and 'validation' set (Ellis et al., 2022). Using these sets, the model's performance can be validated through approaches such as nested k-fold cross-validation, which involves dividing the data into multiple subsets (folds) and iteratively training and testing the model to ensure robust evaluation (Ellis et al., 2022; Maleki et al., 2020). This process reduces overfitting and provides a reliable estimate of model performance on unseen data. Overfitting occurs when a model is overly complex, capturing noise or random fluctuations in the training data rather than the true underlying relationships. This leads to excellent performance on the training set but poor generalisation to new, unseen data. Conversely, underfitting arises when a model is too simple to capture meaningful patterns, resulting in poor performance on both the training and test data. The best-performing model can be determined from assessing its performance on the validation and test data sets through various performance measures (Maleki et al., 2020). Finally, an unbiased evaluation of the final model fit can be examined using an external data set (see Fig. 2) (Ellis et al., 2022).

Machine learning models are typically trained on a large portion of a dataset, with smaller portions reserved for testing and validation (Ellis et al., 2022). Nested k-fold cross-validation is commonly employed to reduce overfitting and ensure robust performance evaluation on unseen data (Ellis et al., 2022; Maleki et al., 2020). In supervised learning, tasks are generally divided into classification and regression problems. Classification models aim to predict discrete categories or classes—for example, distinguishing between patients with otitis media and healthy controls—while regression models predict continuous numerical outcomes, such as estimating hearing threshold levels or tinnitus severity scores. Performance is assessed using metrics tailored to the task at hand. For classification models, metrics like accuracy, area under the receiver operator characteristic curve (AUROC), precision, recall, and specificity are commonly used (Peng et al., 2021; Hanley and McNeil, 1982; Saito and Rehmsmeier, 2015). In cases of regression tasks, metrics such as root mean squared error (RMSE) and mean absolute error (MAE) measure how well predictions align with actual values (Lo Vercio et al., 2020).

For datasets with imbalanced classes, accuracy alone can be misleading. In such cases, metrics like AUROC or precision-recall curves (PRC) are preferred as they better account for performance in rare conditions, such as specific neurotologic disorders (Zahid et al., 2024; Ellis et al., 2022). Oversampling and undersampling techniques can mitigate imbalances, ensuring more reliable predictions (Ellis et al., 2022). However, debates persist regarding the optimal metric for imbalanced datasets, with some recommending AUROC and others favouring PRC, depending on the context.

By addressing these challenges and leveraging advanced AI tools such as LLMs and deep learning architectures, researchers can refine diagnostic accuracy and enhance treatment personalisation for complex disorders like tinnitus and cochlear pathologies. For example, confusion matrices, which highlight true positives, false positives, true negatives, and false negatives, offer a comprehensive view of model performance for classification tasks. These tools and metrics are critical for developing reliable models in neurotology and otology.

Firstly, the subsequent sections review recent successful applications of both conventional machine learning and deep learning approaches in tinnitus and cochlear implants specifically, and neurotology and otology more generally, as well as the challenges that need to be addressed in the field to maximise the generalisability and success of these models moving forward. Some of these successful applications showcase the potential of machine learning to tackle complex clinical challenges. For example, CNNs have been used to analyse otoscopic images, achieving a diagnostic accuracy of over 90 % for conditions like otitis media, outperforming general practitioners (Livingstone and Chau, 2020). Similarly, in tinnitus research, unsupervised clustering methods have identified patient subgroups based on coping mechanisms and distress levels, informing more tailored therapeutic approaches (Makar, 2021). In cochlear implant studies, deep learning models have enhanced speech recognition by optimising sound processing algorithms, significantly improving auditory outcomes for patients in noisy environments (Cha et al., 2019). These examples highlight the transformative potential of machine learning in addressing specific clinical problems, as discussed further in subsequent sections.

The potential of machine learning and deep learning to transform neurotology and otology lies in their ability to integrate diverse data types, from imaging to clinical records, and uncover patterns that guide diagnosis and treatment. The following sections explore recent applications of these methods, including how clustering, CNNs, and LLMs have been leveraged to address clinical challenges in tinnitus and cochlear implants.

2.1. Applications in neurotology–otology

Examining hearing and vestibular disorders poses a significant challenge due to the inaccessibility of the inner ear for analysis in patients and the complex, multifactorial nature of several of these disorders (Crowson et al., 2020a). However, machine learning techniques have become indispensable tools in the domains of neurotology and otology, offering a wide range of functions that significantly impact research, diagnosis, and treatment in these fields. For example, conventional machine learning as well as deep learning can play a crucial role in image analysis, particularly in the interpretation of radiological and histological images. In otology, deep learning enhances the precise detection of anatomical abnormalities, tumours, or pathologies within the ear, enabling more accurate diagnoses and surgical planning. For example, a recent review highlighted how machine learning poses an opportunity for enabling tele-otoscopy in paediatric otology (Ezzibdeh et al., 2022). In one paper highlighted in the review (Livingstone and Chau, 2020), a machine learning-based multi-label classifier using Google AutoML Vision and a dataset of 1366 otoscopic images annotated with 14 otologic diagnoses was developed. The model achieved an average precision (positive predictive value, the proportion of true positive predictions among all positive predictions) of 90.9 %, sensitivity (the proportion of true positives identified among all actual positives) of 86.1 %, and an overall accuracy of 88.7 % when evaluated on an 89-image test set (Livingstone and Chau, 2020). Diagnoses such as cerumen impaction, tympanostomy tube placement, and serous otitis media showed the highest performance, with precision and sensitivity above 90 %. However, conditions like cholesteatoma and otomycosis demonstrated significantly lower performance due to limited training data. By comparison, physician accuracy averaged 58.9 % across paediatricians, emergency medicine, otolaryngologists, and family medicine specialists, highlighting the algorithm's superiority in diagnostic consistency. Accuracy was defined as the proportion of correct predictions to total predictions, and these results underscore the model's potential to improve diagnostic reliability in otology, particularly for underserved areas.

A common condition in otology is otitis media, which is a common middle-ear infection with multifactorial pathogenesis, that often affects children or those in developing countries, can lead to hearing loss, and sometimes requires antibiotic treatment or surgery (Lesica et al., 2021; Myburgh et al., 2016; Rovers et al., 2004). It is traditionally diagnosed by a physician's review of an otoscopy. However, despite its prevalence, correct diagnoses for otitis media can range from 73 % for specialists to 50 % for non-specialists (Lesica et al., 2021; Shaik et al., 2023). This highlights the potential for machine learning technologies to supplement existing clinical situations, especially in low-to-middle income countries where it could facilitate telemedicine (Lesica et al., 2021). In fact, evidence of machine learning's potential application in such space has already been demonstrated in specific studies.

One such study explored the application the applicability of such technologies by training a decision tree model on 489 high-quality, pre-assessed images to classify tympanic membrane conditions into five categories: obstructing wax or foreign bodies (O/W), normal tympanic membrane (n-TM), acute otitis media (AOM), otitis media with effusion (OME), and chronic suppurative otitis media with perforation (CSOM). Among these, the model's accuracy varied by diagnosis, achieving sensitivities between 79 and 82 %, with AOM having the lowest specificity (92 %) and O/W the highest (97 %). Misclassifications occurred, such as O/W being mistaken for CSOM in 12.5 % of cases, often due to overlapping visual features like yellow discoloration. The study participants included two otologists, each with over 35 years of experience, and the dataset incorporated images from both paediatric and adult patients, predominantly outside the United States (Myburgh et al., 2016). Ultimately, the model achieved a mean accuracy of 80.6 %, outperforming the average diagnostic rates of general practitioners and aligning closely with those of specialists.

Subsequent research by the same group compared performance of a decision tree classifier to a neural network on classifying images based on the same 5 pathologies, with these models trained on 80 % of 389 images and tested on 20 % (Myburgh et al., 2018). In that study, the neural network outperformed the traditional decision tree model with an average accuracy of 89.84 % vs 81.58 %, respectively (Myburgh et al., 2018). However, the efficacy of deep learning for image analysis in otoscopy was further demonstrated with increased accuracy. In a recent paper, a neural network was trained on otoendoscopic images to identify 6 types of common ear disease, with an average prediction accuracy of 93.73 % (Cha et al., 2019). Notably, this model was trained on a much larger data set. 10,544 images of ear drums and auditory canals were sourced from an otorhinolaryngology department in an outpatient hospital clinic. 80 % of these were used as a training data set, and the remaining 20 % as a test set. While the model wasn't validated on an external data set, the authors note that the images consisted of a variety of clinical acquisition conditions (with differing composition, white balance, light reflection or colour arrangements, location, rotation, angle etc) which reflect real-world clinical situations and should support replication (Cha et al., 2019).

While these results are encouraging, reported accuracies above 90 % should be interpreted with caution. The model by Cha et al. (2019) was not validated using an independent external dataset, which limits the ability to generalise its performance to other hospital settings or imaging systems (Cha et al., 2019). High accuracy values in such studies may partly result from similarities in data distribution between training and test sets, leading to optimistic performance estimates. This can occur through subtle data leakage or overfitting to dataset-specific characteristics such as lighting or otoscope type. Consequently, while these models demonstrate strong potential under controlled conditions, their real-world generalisability remains unproven. Future studies should therefore prioritise external validation and cross-institutional testing to ensure robustness and clinical reliability.

Thus, both conventional machine learning and deep learning approaches can offer an advantage in enabling remote patient monitoring and telemedicine consultations, making healthcare more accessible for individuals with ear-related conditions (Shaik et al., 2023). Patients can receive real-time guidance and adjustments to their treatment plans, enhancing their overall care experience. While both the conventional machine and deep learning methods above offered greater accuracy than average clinician accuracy rates, the deep learning techniques offered slightly superior accuracy rates. However, it's worth highlighting how these models required significantly larger data sets to be trained on, which is a point that will be discussed further in the challenges section. For neurotology, deep learning can also facilitate the evaluation of neuroimaging data, aiding in the localisation of lesions and abnormalities in the auditory and vestibular pathways. For example, an initial study in animals using neural networks trained to identify normal morphology of Reissner`s membrane on optical coherence tomography images cross-sectional at the level of the cochlea found already histopathologically-confirmed endolymphatic hydrops with a 100 % specificity and 83 % sensitivity (Liu et al., 2017a).

Machine learning algorithms excel at processing and analysing large datasets of patient information, audiometric data, and medical imaging (Kufel et al., 2023). In neurotology and otology, these technologies can accurately diagnose various disorders, such as Meniere's disease, tinnitus, and vestibular disorders. By recognising patterns and anomalies in data, machine learning aids in the early identification and classification of conditions, allowing for timely intervention and tailored treatment plans. For example, an essential aspect of understanding the cause of hearing impairments is classifying hearing loss phenotypes (Dubno et al., 2013). Since accessing the cochlea without disrupting its function is not feasible, audiograms serve as the primary method for categorising hearing loss phenotypes, despite their limitations in capturing the full complexity of hearing loss. Machine learning has been leveraged to use data from human audiogram patterns to categorise adult hearing loss types with a high degree of accuracy. Specifically, a supervised classifier model was developed to correlate human audiogram patterns with animal hearing loss models (Dubno et al., 2013). This approach, while constrained by the limited information in audiograms, enabled the model to distinguish adult hearing loss threshold shifts from those of gerbil models exposed to various conditions (e.g., age-related and furosemide-induced hearing loss). Upon validating the model against human age-related hearing loss audiogram patterns, the authors reported a classification accuracy exceeding 89 % using three distinct machine-learning algorithms. Although promising, such results should be interpreted with caution, as the reliance on audiogram data highlights the need for complementary diagnostic tools to fully characterise hearing loss phenotypes.

Various machine learning models can be leveraged to predict disease progression and outcomes based on patient-specific variables. In neurotology, for instance, these models can predict the likelihood of sensorineural hearing loss worsening over time. A recent study investigated this, and compared the role of neural networks, support vector machines, multilayer perceptrons, and logistic regression models to predict the recovery or nonrecovery of adults who had experienced sudden sensorineural hearing loss (Bing et al., 2018). They incorporated 149 potential predictor variables, including demographics, audiogram data, comorbidities, and laboratory results, from a sample of 1220 patients (643 men and 577 women) into their models. Neural networks demonstrated superior performance compared to logistic regression, achieving an overall accuracy of 77.6 % and an AUROC of 0.84 when all features were included, outperforming logistic regression (67.5 % accuracy, AUROC 0.73). The AUROC measures a model's ability to distinguish between classes, with higher values indicating better overall performance across all classification thresholds (Hanley and McNeil, 1982). Their feature identification method emphasised previously established predictors of hearing recovery, such as audiogram hearing level, audiogram configuration, symptom chronology, age, and concurrent vestibular symptoms, as being the most predictive factors. Interestingly, while the neural network excelled with larger feature sets, simpler models like logistic regression and multilayer perceptrons performed best when limited to just three variables, achieving AUROCs of 0.80 and 0.81, respectively.

Using 149 features on 1220 patients increases the risk of overfitting. The higher AUC of neural networks may result from overfitting to the training data if techniques like regularisation and dropout are not carefully applied. The similar performance of logistic regression with fewer features suggests that it may be a more practical option in clinical applications due to its simplicity and interpretability.

These findings highlight the utility of machine learning in stratifying SSHL outcomes and tailoring treatment strategies. Such predictions enable healthcare providers to make informed decisions and optimise patient care strategies, making precision medicine feasible. In line with this, machine learning techniques can help develop personalised rehabilitation plans for patients with vestibular or auditory disorders (Lesica et al., 2021). By analysing patient data and monitoring progress, these systems can adapt therapy regimens to optimise recovery and improve the patient's quality of life. This can be further facilitated by deep learning assisted drug discovery, where models analyse molecular interactions and identify potential therapeutic targets for conditions affecting the auditory and vestibular systems (Nag et al., 2022). This accelerates the development of new treatments for neurotological and otologic diseases.

In summary, deep learning or traditional machine learning approaches are transforming neurotology and otology by enhancing diagnosis accuracy, personalising treatment approaches, advancing research, and ultimately improving the lives of individuals with auditory and vestibular disorders (Crowson et al., 2020a; Crowson and Chan, 2020; Crowson et al., 2020b; Crowson et al., 2020c; Baguley et al., 2013; Makar, 2021; Langguth et al., 2019; Zenner et al., 2017; Cederroth et al., 2019; Langguth et al., 2013; Genitsaridi et al., 2020; Niemann et al., 2020a; Tyler et al., 2008; Côté et al., 2019). These technologies continue to evolve and offer new possibilities for better understanding and addressing complex issues within these specialised medical fields.

2.2. Applications in tinnitus

One common otologic condition that affects approximately 15 % of individuals globally, places considerable burden on health-care systems and economies and which is well-placed to benefit from machine learning applications is tinnitus (Baguley et al., 2013). Tinnitus is a condition where an individual perceives an auditory sound despite the absence of an external sound source. However, not only can the manifestation of its phenotype differ, but so too can its aetiology and comorbidities (Simoes et al., 2019). Often, individuals with tinnitus perceive a ringing in the ears, although for others the auditory percept can be a buzzing, whistling, chirping, or hissing sound, among others (Baguley et al., 2013). Additionally, various aetiologies and subsequently various pathologies have been reported in several studies (Makar, 2021). For example, the aetiology or set of causes of tinnitus can range from hearing loss to persistent loud noise exposure, whiplash, or psychological or ototoxicity factors (Simoes et al., 2019; Makar, 2021; Makar, 2021). Alongside this, differing pathophysiologies can include cochlear, auditory or vestibular nerve damage or central nervous system anomalies (Baguley et al., 2013; Makar, 2021; Langguth et al., 2019; Zenner et al., 2017). Additionally, tinnitus can occur in conjunction with other comorbidities such as depression, insomnia, or anxiety, that likely exacerbate degradation in quality of life (Cederroth et al., 2019; Langguth et al., 2013). Thus, the heterogeneous nature of the disorder, partially mediated by of the underlying heterogeneous aetiology and pathophysiology, likely affects treatment options and response, adding another dimension to tinnitus patients (Simoes et al., 2019; Langguth et al., 2019; Zenner et al., 2017; Cederroth et al., 2019). Further insight into the heterogeneity of tinnitus and its contribution to the responsiveness to treatment may assist in identifying individual patients or patient subgroups who are best suited for treatment.

Treatment approaches for tinnitus are similarly diverse and depend on the underlying cause. They include hearing aids to manage tinnitus associated with hearing loss, sound therapy to mask or desensitise auditory perception, pharmacological interventions for associated anxiety or depression, and psychological therapies like cognitive behavioural therapy (CBT) to address emotional distress (Makar, 2021; Cederroth et al., 2019). More advanced treatments include transcranial magnetic stimulation (TMS), which targets specific brain areas to modulate neural activity, and direct electrical stimulation. Diagnostic advancements, such as electroencephalography (EEG), provide insights into abnormal neural oscillations that may underlie tinnitus, which could guide treatment selection (Langguth et al., 2019; Zenner et al., 2017).

Machine learning holds promise in refining both diagnostics and treatment. For instance, machine learning algorithms can analyse EEG data to identify neural biomarkers associated with tinnitus subtypes, enabling personalised therapeutic strategies. In treatment, machine learning can optimise TMS protocols by identifying stimulation parameters tailored to individual patients, potentially improving outcomes. Similarly, machine learning can facilitate the identification of patient subgroups most likely to respond to specific interventions by analysing patterns in large datasets, paving the way for precision medicine in tinnitus care. Further exploration of these applications may transform the management of this heterogeneous and debilitating condition.

Tinnitus exhibits different characteristics among individuals, and this is postulated to be mediated by the underlying pathophysiology and the way the patient reacts to the perceived tinnitus. This can consist of variations in the sound perceived, laterality of sound perception, presence of hearing loss, or the extent of distress experienced by the patient (Simoes et al., 2019; Cederroth et al., 2019; Genitsaridi et al., 2020). In line with this, recent progress in machine learning using cluster analysis yielded results in which coping attitude and tinnitus distress appear to play a dominant role in the stratification of tinnitus patients (Niemann et al., 2020a; Tyler et al., 2008). Cluster analysis, an unsupervised machine learning approach, is a popular technique used for assembling or grouping data into clusters based on similar attributes (Peng et al., 2021; Collin et al., 2022). In the context of tinnitus research, cluster analysis has been utilised in previous studies to define tinnitus subgroups, as well as subgroups to predict treatment response (Genitsaridi et al., 2020; Niemann et al., 2020a; Tyler et al., 2008; Côté et al., 2019; Kloostra et al., 2019; van den Berge et al., 2017). These studies highlight the potential of clustering techniques in uncovering meaningful subtypes within complex conditions like tinnitus, and recent advancements in LLMs have further expanded this potential by enabling the analysis of both textual and numerical datasets.

Building on the foundational work of clustering and subgroup identification, these advancements in LLMs have opened new avenues for analysing complex datasets in tinnitus research. A study by Jeong et al. (2024) applied GPT-2 embeddings to cluster patients based on CBT session transcripts, revealing distinct patient subgroups with varying therapy engagement and outcomes. Using unsupervised learning approaches like Density-Based Spatial Clustering of Applications with Noise (DBSCAN), one of the most commonly used and cited clustering algorithms (Jeong et al., 2024), this analysis identified patterns that highlight the importance of addressing maladaptive thoughts in reducing tinnitus severity. Additionally, LLMs like Google T5 and Flan-T5 were employed to predict post-therapy Tinnitus Handicap Inventory (THI) scores, achieving strong predictive performance on augmented datasets (Jeong et al., 2024). These models provide a promising pathway for identifying meaningful subgroups and tailoring interventions, even when faced with small datasets, through robust augmentation techniques.

However, the study also highlighted challenges, such as limited generalisability due to small dataset sizes and potential noise introduced during data augmentation. Despite these limitations, the integration of LLMs for text-based analysis underscores their ability to address the dimensionality of tinnitus heterogeneity by identifying subphenotypes and improving predictions for treatment outcomes (Jeong et al., 2024). These findings complement the existing clustering and predictive modelling approaches in tinnitus research and pave the way for precision medicine applications.

In addition to text-based analyses, recent advancements in AI have explored the integration of imaging data with clinical variables to refine diagnostic precision for tinnitus and related conditions. Tang et al. (2024) explored the application of AI, including LLMs, such as GPT-4 and ChatGLM-6B, in CT-based diagnostics for ear diseases, providing insights that are also relevant to tinnitus. Their study highlighted how LLMs can process and interpret CT imaging data alongside clinical records to enhance diagnostic accuracy and efficiency. By standardising radiology reports, LLMs reduced inconsistencies in human interpretations and identified subtle radiological markers for conditions like otosclerosis and middle-ear anomalies, which are often comorbid with tinnitus (Tang et al., 2024).

Tang et al. (2024) proposed that combining CT data with clinical variables, such as hearing loss severity or tinnitus distress levels, could enable predictive models to refine subgroup identification and tailor treatment strategies. This integration demonstrates the broader utility of LLMs in bridging imaging diagnostics and clinical decision-making. However, challenges remain, including the need for high-quality annotated datasets and the variability in imaging protocols across institutions (Tang et al., 2024). Despite these limitations, the study underscores the potential for LLMs to standardise and optimise diagnostic workflows, paving the way for more personalised and efficient tinnitus care.

These advancements in large language models complement traditional clustering approaches, which have been widely utilised to define meaningful tinnitus subgroups. For instance, Niemann et al., (2020b) reported four chronic tinnitus phenotypes following cluster analysis: (1) *avoidant*, characterised by few affective or psychosomatic symptoms (also has the least health burden); (2) *psychosomatic*, in which patients reported high tinnitus burden with high affective impairments and somatoform complaints; (3) *somatic*, characterised by high somatic symptoms, and (4) *distress*, characterised by increased distress scores. This work underscores the diversity of tinnitus experiences and the role of psychosomatic and somatic factors in subgroup distinctions. The *psychosomatic* group was similarly reported by Tyler and colleagues as the *constant distressing tinnitus* group (Tyler et al., 2008). Mirroring Niemann and colleagues' finding, this cluster-analysis based group was likewise characterised by high affective impairments with loud, distressing tinnitus, with these patients reporting a high tinnitus burden (Tyler et al., 2008; Niemann et al., 2020b). Tyler and colleagues also reported three additional cluster subgroups including: (1) variable tinnitus that is worse in noise, (2) copers whose tinnitus is not influenced by touch, and (3) copers whose tinnitus is worse in quiet environments (Tyler et al., 2008). Overall, coping attitude and tinnitus distress appear to play major roles in the distinction between phenotype in both studies, however, it's important to note that the effect of background noise also has a dramatic distinction despite it not being reported in both studies.

Van den Berge and colleagues speculate that distinct treatments could be applied to distinct cluster subgroups (van den Berge et al., 2017). However, in their study using cluster analysis, the authors found no substantial cluster structure, suggesting a continuum rather than subgroups with this particular tinnitus cohort (van den Berge et al., 2017). Despite no distinct clusters being identified, this study observed that 'stress' and 'loud sounds' have relatively high discriminative value in tinnitus patients (van den Berge et al., 2017).

These findings are both conflicting and yet harmonious, which could be due to a number of limitations. There is currently no standard protocol for profiling tinnitus, which can be detrimental to finding an effective treatment. The differences in cluster phenotypes and strengths also likely contribute to various responses to treatment. Moreover, in the case of van den Burge and colleagues' study, where results suggested a continuum, it may be necessary to use a number of treatments to find the optimum for each individual patient (van den Berge et al., 2017). Furthermore, the distinction between whether patients prefer noise or silence can determine how well patients will respond to sound therapy, in which those that prefer sound may respond more positively (van den Berge et al., 2017). This is further confirmed through an exploratory analysis in which participants who reported responding positively to sounds more frequently reported a positive response to treatments with an acoustic component (Simoes et al., 2019). Altogether, these studies highlight the dimensionality of tinnitus heterogeneity and differential treatment response, in which patients respond better if the treatment is tailored to their individual characteristics or clinical profiles (Simoes et al., 2019; Côté et al., 2019; Kloostra et al., 2019; Fisher and Boswell, 2016; Searchfield et al., 2017). This emphasises the potential benefits precision medicine can offer for these patients, if the challenges to applying machine learning approaches in this field can be overcome.

To address varied treatment responses in tinnitus patients, recent studies have utilised supervised machine learning approaches to predict patient outcomes (Cardon et al., 2022; Niemann et al., 2020a; Rodrigo et al., 2021). Machine learning algorithms have predicted patient outcomes after various treatments including cochlear implantation, physiotherapy, and behavioural or neuromodulation therapy with varying success (Cardon et al., 2022; Niemann et al., 2020a; Kloostra et al., 2019; Rodrigo et al., 2021). However, both random forest and decision tree machine learning models have shown promising results in predicting treatment outcomes (Cardon et al., 2022; Niemann et al., 2020a; Rodrigo et al., 2021). Most recently, a study tested the feasibility of random forest classification to predict response to neuromodulation treatment (Cardon et al., 2022). While the random forest classifier predicted treatment response with high accuracy, the Receiver Operating Characteristic (ROC) indicated that the model did not perform significantly better than chance when applied to a new dataset (Cardon et al., 2022). A separate study utilised a deep learning model to predict sound therapy outcomes using tinnitus patients' Electroencephalogram (EEG) data prior to treatment, resulting in up to 99 % prediction accuracy (Doborjeh et al., 2023). The reported 99 % accuracy should be interpreted with caution. Such extraordinary results may stem from unintended overlap between training and test data (data leakage) or overfitting to noise within the training dataset. It is therefore essential that the methodology, particularly data splitting and preprocessing strategies, be carefully examined to validate the credibility and generalisability of these findings.

Various studies implement feature importance to determine predictor variables for treatment response, with different important features identified. Baseline Tinnitus Handicap Inventory (THI) and Edinburgh Handedness Inventory (EHI) scores were found to be of highest importance when developing their random forest model (Cardon et al., 2022) to predict response to transcranial direct current stimulation. Tinnitus impairment, depression severity, and sleep problems were important features to predict response to multi-modal therapy in a gradient boosted tree model (Niemann et al., 2020a). Education level, employment type, hearing aid usage, and tinnitus maskability were important features to predict response to behavioural therapy in a gradient boosted tree model and classification and regression tree models (Rodrigo et al., 2021). These differences in feature importance could be attributed to a number of reasons, including using different variables, as well as how different treatments may rely on different features. A recent review revealed that of 61 variables that have been used for subgrouping, tinnitus severity, age, hearing ability, and depressive symptoms were among the most popular and most important variables (Genitsaridi et al., 2020). To illustrate this, we categorised the 61 variables identified into subdomains (Table 1). The lack in standardisation can be detrimental to finding the right treatment, which leads to patient disappointment as well as increased costs (Cardon et al., 2022). Overall, the large variability in tinnitus features suggests that different treatment modalities may have optimal effect for a specific subgroup of patients. As research progresses to include more deep learning and machine learning, certain challenges should be addressed to improve effectiveness.

The integration of machine learning and advanced AI technologies, such as LLMs, offers transformative potential for managing the complexity of tinnitus. By addressing the heterogeneity of tinnitus through clustering, predictive modelling, and imaging-based diagnostics, these approaches provide clinicians with powerful tools to identify subphenotypes and tailor treatments to individual patient profiles (Lo Vercio et al., 2020; MacEachern and Forkert, 2021; Tang et al., 2024; Jeong et al., 2024). Text-based LLMs, as demonstrated by Jeong et al. (2024) enable nuanced subgroup identification and therapy prediction, while imaging-focused applications, like those explored by Tang et al. (2024) bridge clinical and radiological data for personalised diagnostics. Traditional clustering studies, such as those by Niemann et al. (2020a) and Tyler et al. (2008) further underscore the importance of subgroup identification in guiding effective treatments. Collectively,

Table 1
Review revealed that of 61 variables that have been used for subgrouping, tinnitus severity, age, hearing ability, and depressive symptoms were among the most popular and most important variables.

	Category	Variable	Included in previous studies	Significant	Classification Model	Future Recommendation
1	Impact & Reactions	Overall Severity	Yes	Yes	Yes	Standardise measurement criteria
2		Impact on Emotion & Mental Health	Yes	Yes	Yes	Develop psychological assessment tools
3		Awareness	Yes	No	No	Further study needed on patient awareness
4		Impact on Sleep	Yes	Yes	No	Investigate links to sleep disorders
5		Impact on Concentration	Yes	Yes	No	Standardise cognitive assessment tools
6		Impact on Hearing	Yes	Yes	Yes	Improve audiometric testing methods
7		Intrusiveness & Ability to Ignore	Yes	Yes	Yes	Further research on habituation strategies
8		Ability to Cope	Yes	Yes	Yes	Study coping mechanisms in different patient groups
9		Impact on Situations	No	No	No	Limited research; explore environmental influences
10		Impact on Somatic Sensations	Yes	No	No	Investigate somatic tinnitus mechanisms
11	Ear & Hearing Function	Hearing Ability	Yes	Yes	Yes	Improve objective testing methods
12		Problems with Sounds	Yes	Yes	No	Research noise sensitivity in tinnitus and its role in auditory processing disorders
13	Demographic & Socio-economic Characteristics	Age	Yes	Yes	Yes	Consider age-specific models
14		Sex	Yes	No	No	Examine gender differences in tinnitus
15		Employment & Occupation	No	No	No	Limited data; assess impact on work performance
16		Education	No	No	No	Explore correlation with treatment adherence
17		Hearing-related Family History	No	No	No	Investigate genetic predisposition
18		Marital Status	No	No	No	Limited research; explore social factors
19		Economic Status	No	No	No	Study financial burden of tinnitus treatments
20	Mental Health	Depressive Symptoms	Yes	Yes	Yes	Integrate psychiatric screening
21		Symptoms of Anxiety	Yes	Yes	Yes	Investigate treatment impact
22		Stress-related Symptoms	Yes	Yes	No	Explore stress reduction interventions
23		Personality & Coping Strategies	Yes	No	No	Further study needed
24		Various Psychological Conditions	Yes	Yes	No	Examine comorbid psychiatric disorders
25		Mood	Yes	Yes	Yes	Study links between mood disorders & tinnitus
26	Modulating Factors	External Sound Effect	Yes	Yes	No	Investigate sound enrichment therapies
27		Psychological Factors Effect	Yes	Yes	No	Study cognitive-behavioural interventions
28		Sleep Effect	Yes	Yes	No	Explore links to insomnia treatments
29		Medication Effect	Yes	No	No	Standardise pharmacological impact assessment
30	Perceptual Characteristics	Loudness	Yes	Yes	Yes	Use consistent measurement tools
31		Localisation	Yes	No	No	Study relationship with underlying pathology
32		Quality	Yes	No	No	Standardise sound perception testing
33		Varying Perception	Yes	No	No	Research fluctuating symptoms
34		Pitch	Yes	Yes	No	Further explore neural mechanisms
35		Rhythmicity	No	No	No	Limited research; potential link to brain oscillations
36		Presence Pattern	No	No	No	Study episodic vs. continuous tinnitus
37	Onset Related Characteristics	Duration	Yes	Yes	Yes	Standardise longitudinal tracking
38		Age at Onset	Yes	Yes	Yes	Consider age-specific interventions
39		Temporal Relationship with Other Incidence	No	No	No	Further research needed on triggering events
40		Gradual or Abrupt Onset	No	No	No	Further research into onset mechanisms
41		Emergence Time of Day	No	No	No	Study circadian influences on tinnitus
42	Other Symptoms & Conditions	Various Pain Problems	Yes	Yes	No	Investigate links between tinnitus and chronic pain syndromes
43		Dizziness	Yes	Yes	Yes	Explore vestibular involvement
44		Quality of Life	Yes	Yes	Yes	Develop holistic assessment tools
45		Mandible Problems	Yes	Yes	No	Investigate temporomandibular joint (TMJ) dysfunction as a tinnitus subtype
46		Headache	Yes	Yes	No	Study links between migraines and tinnitus
47		Neck Problems	Yes	Yes	No	Explore cervical spine contributions to tinnitus
48		Various Somatic Symptoms	Yes	Yes	No	Identify patterns of somatic tinnitus
49		Various Symptoms & Conditions	Yes	Yes	No	Standardise classification of somatic influences on tinnitus
50	Modulating Factors	Somatic Manoeuvres Effect	Yes	Yes	No	Study diagnostic potential of somatic modulation tests
51	Healthcare & Treatments	Psychiatric Treatments	Yes	Yes	No	Assess effectiveness of therapy-based interventions
52	Brain Anatomy & Function	Activation Patterns	Yes	Yes	No	Identify biomarkers for tinnitus subtypes
53	Healthcare & Treatments for Tinnitus	Clinician Consultation	Yes	Yes	No	Improve patient referral pathways

(*continued on next page*)

Table 1 (*continued*)

	Category	Variable	Included in previous studies	Significant	Classification Model	Future Recommendation
54		Preceding Tinnitus Treatments	Yes	Yes	No	Evaluate long-term treatment efficacy
55		Treatment Response	Yes	Yes	Yes	Investigate precision medicine approaches
56	Clinical Subtypes	Somatosensory Tinnitus	Yes	Yes	Yes	Develop targeted treatments for somatic tinnitus
57	Brain Anatomy & Function	Connectivity	Yes	Yes	No	Explore neural network disruptions in tinnitus
58		Structural Features	Yes	Yes	No	Investigate anatomical correlates of tinnitus severity
59	Associations with Other Conditions	Tinnitus Localisation & Hearing Loss Profile	Yes	Yes	Yes	Improve diagnostic precision for tinnitus-related hearing loss
60	Genetic Profile	Single Nucleotide Polymorphisms	No	No	No	Expand research on genetic links
61	Lifestyle & Exposures	Physical Activity	No	No	No	Explore impact of exercise on tinnitus severity

these studies exemplify how machine learning can refine precision medicine in tinnitus care, potentially improving treatment efficacy, reducing healthcare costs, and enhancing patient quality of life.

2.3. Applications in cochlear implants

The above sections have reviewed how both the diagnoses and treatment of neurotologic and otologic conditions including tinnitus can stand to benefit from machine learning applications. However, within the field of otology, there is a large scope for how machine learning models can be applied to improve the efficacy of cochlear implants as a resolution for hearing loss. Cochlear implants have revolutionised the lives of individuals with severe hearing impairment by providing them with a means to partially restore their hearing (O'Donoghue, 2013). These intricate electronic devices work by directly stimulating the auditory nerve, bypassing damaged hair cells in the inner ear. Cochlear implants have made significant progress over the years, with many cochlear implant patients displaying meaningful benefits in recognising speech and understanding spoken language following implantation (Pisoni et al., 2017). Despite this, there remains a significant number of patients that achieve poor outcomes (Pisoni et al., 2017). However, understanding and explaining the reasons for poor outcomes following implantation is a very challenging research problem. The field of cochlear implants is now embracing the power of machine learning and deep learning to enhance their performance and usability in various ways.

In neurotology, for instance, these models can assess the potential success of cochlear implantation or predict the likelihood of hearing loss worsening over time. Such predictions enable healthcare providers to make informed decisions and optimise patient care strategies. Deep learning algorithms can also be applied to hearing aids and cochlear implants to adapt and personalise sound amplification based on the individual's hearing profile and environmental factors. This technology enhances the user's auditory experience and improves speech recognition in noisy environments, benefiting individuals with hearing impairment. For example, commercial hearing devices now boast 'on-board deep neural networks' that are trained on over 10 million 'real-life' sounds, to enable enhanced speech understanding and reduced sustained listening effort (Murmu Nielsen and Ng, 2022). It's important to clarify that these findings were only reported in the company's whitepaper, and accuracy rates for the device were not published in a peer-reviewed paper. However, they highlight how deep learning models such as neural networks require extremely large data sets in order to have clinically meaningful or commercial impact.

While traditional methods for evaluating cochlear implant performance rely heavily on human participants and are constrained by their inherent variability and cost, emerging deep learning technologies have begun to reshape this landscape. A recent study by Sinha and Azadpour (2024) explored the potential of advanced deep learning models, specifically the Whisper speech recognition model, to evaluate speech information in cochlear implant (CI) simulations. This model demonstrated human-like performance patterns when tested with vocoder-processed speech under various parameters, such as the number of channels, envelope dynamics, and signal-to-noise ratios. However, the success of simulation models such as Whisper remains fundamentally limited to processing acoustic and psychoacoustic parameters. These models cannot replicate the higher-level cognitive mechanisms—such as attention, contextual understanding, neural plasticity, and learning—that shape a real cochlear implant user's perceptual experience. Consequently, findings derived from such simulations should be corroborated by studies involving human participants before being generalised to clinical applications.

Despite these inherent limitations, the study still provides valuable insight into how deep learning models can approximate aspects of auditory perception under controlled conditions.

Notably, Whisper excelled in simulating how CI users perceive speech in quiet conditions but also revealed challenges in noisy environments, reflecting limitations observed in real-world CI outcomes. Unlike traditional approaches requiring extensive human participant testing, Whisper achieved large-scale evaluations rapidly and cost-effectively, eliminating variability introduced by human factors such as attention or fatigue. The study's findings emphasise the applicability of deep learning models in uncovering the physiological and psychophysical barriers to optimal CI outcomes, such as channel interaction and the inability to process fine temporal details (Sinha and Azadpour, 2024). These results not only highlight the promise of deep learning-based simulations in advancing CI research but also underline the need for future models to integrate auditory-specific mechanisms, which may further bridge the gap between simulations and actual patient experiences. These advancements illustrate the potential for deep learning models to revolutionise cochlear implant research. Beyond simulation-based insights, machine learning continues to drive innovation in CI performance optimisation, from sound processing to surgical planning.

As summarised in a recent structured review, by Crowson et al. (2020d), of publications from 2013 to 2018, there has been an increase in the year-over-year applications of machine learning technologies involved speech/signal processing optimisation (17; 43.6 % of articles), automated evoked potential measurement (6; 15.4 %), postoperative performance/efficacy prediction (5; 12.8 %), and surgical anatomy location prediction (3; 7.7 %), and 2 (5.1 %) in each of robotics, electrode placement performance, and biomaterials performance (Crowson et al., 2020d). Machine learning algorithms can be employed to improve the processing of sound signals received by cochlear implants (Hajiaghababa et al., 2018). This includes the suppression of background noise, enhancement of speech signals, and the ability to focus on specific sound sources. Deep learning models, such as CNNs and recurrent neural networks, have shown promise in enhancing the real-time processing of auditory information, resulting in better speech perception for implant users (Drakopoulos et al., 2022). Adding to this, in a recent review, Wang and Chen (2018) summarised how deep learning neural networks have advanced machine learning applications in the fields of speech separation tasks including monaural speech enhancement, speech dereverberation and speaker separation. Building on these developments, recent reviews have provided a broader perspective on the role of AI in transforming cochlear implant technology and addressing patient-specific challenges.

A recent review by Essaid et al. (2024) provide a comprehensive analysis of AI applications for cochlear implants, offering critical insights into the advancements, challenges, and future perspectives in this field. The paper highlights key areas where AI has revolutionised cochlear implant technology, including speech enhancement, electrode localisation, imaging segmentation, and adaptive sound processing. For instance, CNNs have been employed to optimise auditory scene recognition, improve segmentation of cochlear structures, and reduce metal artifacts in CT imaging, enabling more precise surgical planning (Tang et al., 2024). Additionally, the paper addresses the role of advanced reinforcement learning algorithms in localising critical cochlear regions, further enhancing intraoperative precision. However, Essaid et al. (2024) emphasise that significant challenges remain, particularly in integrating multimodal data—such as audiometric, anatomical, and patient-reported outcomes—into machine learning models to address patient-specific variability. This limitation underscores a persistent gap between the capabilities of AI algorithms and the complex etiologies of hearing loss in cochlear implant candidates (Essaid et al., 2024). By categorising methodologies, datasets, and evaluation metrics, the paper provides a roadmap for researchers to bridge these gaps and highlights opportunities to leverage emerging techniques, such as generative adversarial networks (GANs) for image reconstruction and transformers for speech enhancement, to drive further innovation in cochlear implant technology (Essaid et al., 2024). While these advancements focus on algorithmic and multimodal integration challenges, progress in diagnostic imaging has also contributed significantly to improving surgical planning and outcomes for cochlear implants.

Recent advancements in AI for ear imaging, such as those outlined by Tang et al. (2024), demonstrate considerable progress in diagnostic precision and segmentation techniques for cochlear structures. Ultra-high-resolution CT (U-HRCT) combined with deep learning-based segmentation has enabled precise mapping of intricate ear anatomy, aiding surgical planning and anomaly detection. However, these applications remain primarily focused on structural diagnostics and fail to incorporate the multifactorial nature of disease etiology. For example, while these tools excel at detecting malformations or optimising cochlear duct measurements, they often overlook the broader implications of auditory neuropathy, vestibular dysfunction, or cognitive decline—factors that heavily influence cochlear implant outcomes (Tang et al., 2024). Additionally, the narrow scope of training datasets limits these models' applicability to diverse patient profiles, creating a mismatch between ML capabilities and real-world clinical needs. These gaps highlight the need for AI models that integrate functional, anatomical, and clinical data to better address the complexities of individual patient care. In addition to advancements in imaging, machine learning has shown promise in addressing real-world challenges in cochlear implant programming and user experience.

One of the challenges in cochlear implant fitting is still creating individualised sound maps (also known as cochlear implant programming) for each user (Helpard et al., 2021a; Helpard et al., 2021b). However, machine learning can analyse a user's auditory responses and preferences to optimise the stimulation parameters, such as electrode current levels and frequency mapping. These personalised sound maps can significantly improve the user's hearing experience, making it more natural and comfortable. Cochlear implant users often experience feedback or whistling sounds when the microphone picks up sounds emitted by the implant. Machine learning algorithms can detect and mitigate feedback in real-time, ensuring that users can comfortably wear and use their implants without discomfort or interruption (Wijewickrema et al., 2022). Deep learning models can adapt stimulation strategies based on the user's listening environment, movement, and auditory feedback. For instance, they can switch between different stimulation modes (e.g., noise reduction or directional microphones) to optimise sound perception in various situations, such as noisy environments or while on the move (Kokkinakis et al., 2012).

Machine learning can play a role in predicting the health and longevity of cochlear implant components, which is a significant concern for patients and healthcare providers (Pelosi et al., 2013; Sunde et al., 2013). Device failure can lead to sudden loss of hearing, negatively impacting a patient's quality of life and requiring urgent medical intervention or replacement surgeries, which are costly and resource-intensive By analysing data from the implant's sensors and performance logs, algorithms can detect potential issues before they become critical, allowing for proactive maintenance, and reducing the risk of device failures (Crowson et al., 2020b). For example, early detection of declining battery performance or electrode degradation can allow for timely interventions, extending the device's lifespan and maintaining its functionality.

Deep learning can be employed to develop applications that continuously monitor a user's auditory progress, providing feedback and suggestions for auditory training and rehabilitation (Tyler et al., 2008). These applications can adapt to the user's specific needs and track improvements over time, thereby enhancing the effectiveness of auditory therapy. Not only can these algorithms optimise cochlear implant efficacy, but they can also be leveraged to identify which patients might stand to benefit the most from implantation. A recent study already demonstrated that current models could support clinical decision making, highlighting that subsets of individuals can be identified that have a 94 % chance of improving word recognition scores by at least 10 % points after implantation, which is likely to be clinically meaningful (Shafieibavani et al., 2021). Thus, both conventional machine learning as well as deep learning approaches enable researchers to analyse vast datasets generated by cochlear implants. This data can be used to gain insights into the auditory system's functioning, further refine implant technology, and improve the overall understanding of hearing restoration. By prioritising predictive maintenance and patient-specific optimisation, these algorithms can reduce device failure rates and significantly improve patient outcomes.

As outlined above, considerable effort has been directed toward augmenting signal processing and automating postoperative mapping using machine learning algorithms. Other promising applications include augmenting cochlear implant surgery mechanics and personalised medicine approaches for boosting cochlear implant patient performance 30, 44, 45]. The integration of machine learning and deep learning techniques into cochlear implant technology holds great promise for enhancing the auditory experiences of implant users. These advancements are not only improving sound processing and customisation but also contributing to the ongoing development of more effective and user-friendly cochlear implants, ultimately improving the quality of life for individuals with hearing impairment. Nevertheless, several challenges exist not only with regards to the application of machine learning or deep learning technologies to cochlear implants, but also to tinnitus and the fields of otology and neurotology more generally (Essaid et al., 2024; Rahman et al., 2024; Ching et al., 2018).

3. Challenges

Despite the recent popularity of machine and deep learning in personalised medicine in these fields, significant challenges still exist that need to be addressed.

3.1. Interpretability

Machine learning and deep learning models are often referred to as black-box models, though this characterisation applies more strongly to deep learning. While traditional machine learning models (e.g., decision trees or linear regression) can offer some level of interpretability, deep learning models are inherently less transparent, making it difficult to understand the decision-making process behind their predictions (Carvalho et al., 2019). Recent research has increased focus on developing methods to increase interpretability of machine and deep learning models (Carvalho et al., 2019; Figueroa Barraza et al., 2021). Determining what factors drive a model's predictions could give greater insight into understanding its performance, however there remains a lack of consensus on the best way to resolve this interpretability problem (Carvalho et al., 2019).

On the other hand, some machine learning algorithms are inherently easier to interpret (Carvalho et al., 2019; Figueroa Barraza et al., 2021). In fact, logistic regression, linear regression, and decision tree models are commonly referred to as interpretable models. However, as models increase in complexity, they often decrease in interpretability, a conundrum referred to as the accuracy/interpretability trade off (Figueroa Barraza et al., 2021). For example, while deep learning may outperform machine learning algorithms, they are often more difficult to interpret. LLMs, which are being explored for text-based tinnitus subgroup identification and therapy predictions, share these interpretability issues. Their complex architectures make it difficult to trace which input features or patterns drive their recommendations, further complicating clinical adoption.

In deep learning, feature selection or feature importance are common methods to determine the influence of an individual feature in the model's prediction output. Through perturbation-based approaches, individual input features are occluded, and subsequent model performance is compared to that of the original input (Ching et al., 2018). The input feature with the greatest difference in model performance is likely the most important feature. While this will allow you to identify influential variables that may affect model performance, this method can be time consuming and computationally expensive (Singh et al., 2020).

In addition to these traditional perturbation-based methods, advanced model-agnostic interpretability techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) have gained prominence in recent years (Moulaei et al., 2024). These approaches provide local explanations by quantifying the contribution of individual features to specific predictions (Raptis et al., 2024), thereby enhancing transparency and clinician trust in complex "black-box" models (Sadeghi et al., 2024). Incorporating SHAP and LIME into otological and neurotological research could facilitate a deeper understanding of model decision-making and support their safe, evidence-based translation into clinical practice (Frasca et al., 2024).

3.2. Data

Machine and deep learning approaches are data driven, in which the model is provided with input data, and the model is then fitted to the training dataset to find patterns within the data. Therefore, the data that is used to train the model is an essential component to the quality of the model. It is critical to confirm that the model is not over or under fitted to the data. Overfitting can occur when a model is too complex, in which it may capture noise or random effects within the training data (Lo Vercio et al., 2020). Therefore, the model may 'overfit' to the input data in which the model adapts too well to the data its training on and is unable to generalise to other data sets (MacEachern and Forkert, 2021). This is evidenced by performing extremely well on training data, but performing poorly on novel data, or a test dataset. Conversely, under-fitting occurs when a model is too simple, and unable to learn patterns in the data. This is evidenced by a poor performance on training and novel data.

An important factor that can contribute to over-fitting, and limit a model's extrapolability, is how context-specific they can be. Clinical studies often have strict inclusion and exclusion criteria. This can result in data that is specific to a subgroup of patients, where the model is then only trained to make predictions for this specific subgroup of patients. The consequence of this is that the model fails to successfully replicate and generalise to the rest of the patient population and risks its ability to apply to real world data. The integration of multimodal datasets, such as combining imaging, audiometric data, and patient-reported outcomes, shows promise in improving predictive accuracy (Essaid et al., 2024). However, these complex datasets present additional challenges, including missing data, alignment across data types, and the need for harmonised data collection standards (Essaid et al., 2024). Thus, it is critical to ensure that the data equally represents the patient population. Additionally, incomplete trial data can be an issue, especially in interventional studies due to factors such as attrition (Collin et al., 2022). With attrition bias, individuals who report no positive changes in the outcome measure may be more likely to drop out of the study. The result of this is a greater risk of data sets containing imbalanced data for responders and non-responders. Likewise, in the case of class imbalance in classification tasks, additional actions need to take place to balance the dataset in order to avoid bias (Lo Vercio et al., 2020).

Furthermore, a small sample size can be a contributing factor to overfitting, and impeding model generalisation. Within a data set, the greater the number of features to be assessed, the greater the sample size/ number must be in order to maintain model robustness (Plant and Barton, 2020). This is especially relevant for deep learning models which can analyse a more complex set of features, and subsequently require a much larger training set of data to ensure high accuracy (Lo Vercio et al., 2020). Thus, this is an important consideration when it comes to deciding whether a traditional machine learning approach or deep learning model is more suitable for a specific research question. Recent advancements, such as Whisper for cochlear implant simulations and LLMs for tinnitus, emphasise the need for large, annotated datasets (Sinha and Azadpour, 2024). The lack of such resources limits the applicability of these advanced models in clinical settings. As evidenced throughout this paper, deep learning models such as neural networks were shown to outperform traditional machine learning techniques with regards predicting diagnoses from images or improving speech recognition (Myburgh et al., 2018; Myburgh et al., 2016; Cha et al., 2019; Bing et al., 2018). However, typically these neural networks relied on data sets with tens of thousands of points, as opposed to conventional machine learning models which reported training on data sets with only hundreds of points (Myburgh et al., 2018; Myburgh et al., 2016; Cha et al., 2019; Bing et al., 2018).

Adding to the challenges, using many features in the dataset can make it more difficult to identify patterns that group patients effectively, impeding model performance, generalisability, and interpretation robustness (Plant and Barton, 2020). Moreover, often in real-world clinical situations, obtaining large standardised sample sizes are not always feasible (Plant and Barton, 2020). In these cases, feature selection can be an important tool to mitigate this, whereby only relevant features are focused on, while excluding those that are not important or are redundant for analysis (Plant and Barton, 2020). However, if applications of machine learning and deep learning in the field are to be maximised, the issue of over-fitting must be considered; standardisation in data collection must be a priority for researchers and clinicians alike, and data sets for all machine learning models must aim to include information on responders and non-responders, a diversity of patients and non-patients, and a balance of classes for classification tasks.

3.3. Standardisation

There is a lack of standardisation in reporting model development, evaluation, and validation. Deep learning networks, for instance, involve multiple interchangeable hyperparameters (e.g., learning rate, batch size, and architecture depth) to improve model performance. However, these hyperparameters are not consistently reported, and the rationale for why specific parameters are chosen is scarcely reported, making reproducibility difficult (Collin et al., 2022). The architecture of the network, including the number of layers and how large the layers are, optimisers, and loss function are all various ways to alter the network. Model parameters are often chosen based on prior assumptions, which may bias the outcome. There are statistical methods to select which hyperparameters work best with a given model and data (Liu et al., 2017b), however these methods choose the hyperparameters that perform best with that specific model. While these methods are beneficial in building high performing models, it risks reducing generalisation for clinical applicability. For text-based models like LLMs, establishing standardised formats for clinical notes and patient-reported data is critical to improving their applicability and generalisability across institutions. If a model's algorithm and hyperparameters are determined by those which give the best performance, then these parameters are highly dependent on the data that the model trains on. These strategies will only result in high performing models specific to the data set it is given, further reducing its generalisability to a clinical setting.

3.4. Ethics

Lastly, machine and deep learning models used to predict patient outcomes often require access to sensitive personal data. While certain patient identifiers, such as name or ID, may not be essential for training algorithms, other attributes—such as gender, ethnicity, age, sex, and comorbidities—are crucial for accurate predictions and subgroup analyses. Therefore, the use of personal data must be approached cautiously, with robust safeguards to ensure compliance with data protection regulations, such as GDPR. In-house machine learning algorithms or off-network systems can utilise non-anonymised data for training purposes, provided they adhere to strict ethical and legal frameworks to protect patient privacy. Additionally, the lack of interpretability of machine learning predictions poses challenges in explaining how a model reaches its conclusions, underscoring the need for transparency. This is particularly important in clinical decision-making, where liability concerns arise, especially in cases of malfunction (Peng et al., 2021; Collin et al., 2022). Establishing clear guidelines for the ethical use of sensitive data and ensuring model interpretability will be essential to fostering trust and accountability in machine learning applications. Bias in training data is a significant ethical concern, particularly for LLMs, which may reflect or amplify biases present in the datasets they are trained on (Ayoub et al., 2024; Dai et al., 2024; Omar et al., 2025). For instance, if an LLM is primarily trained on medical literature and patient records predominantly from a specific demographic, it may develop biases in diagnosing or recommending treatments for underrepresented groups, particularly if their symptom presentation or disease progression differs (Zack et al., 2023). This could manifest as disparities in tinnitus subgroup identification or in predicting treatment responses for cochlear implant patients across different ethnic or socioeconomic backgrounds. Furthermore, biases can emerge from historical medical data that may contain societal prejudices or unequal access to care, leading LLMs to perpetuate or even amplify these existing health inequities (Yang et al., 2024). Ensuring equitable representation of patient demographics is critical for fair clinical predictions.

4. Recommendations for future studies in otology and neurotology

Many challenges need to be overcome for clinical use of machine learning in the future. With personalised medicine in its early stages, it's important to facilitate rigorous and reproducible approaches. There is extensive research delineating the clusters or sub phenotypes within specific populations (Genitsaridi et al., 2020; Tyler et al., 2008; Côté et al., 2019; Kloostra et al., 2019; van den Berge et al., 2017; Niemann et al., 2020b). Genitsaridi and colleagues composed a comprehensive review, in which they discuss different approaches to hypothesis and data driven clustering approaches (Genitsaridi et al., 2020; Genitsaridi, 2021). While data driven may be less biased than hypothesis testing, data driven results are highly dependent on the methodology. In the case of tinnitus specifically, it is difficult to standardise the research given the heterogeneity of the condition. LLMs such as GPT-4 and T5 offer novel opportunities for addressing this heterogeneity by identifying meaningful subgroups based on textual clinical data and therapy outcomes (Zahid et al., 2024; Wang and Zhang, 2024). However, the successful integration of LLMs requires harmonised textual datasets, rigorous validation, and strategies to mitigate biases inherent in model training. Future research should explore how LLM-driven analyses can complement traditional clustering approaches to refine tinnitus subphenotyping and personalise therapeutic interventions. However, there are action items that can be taken to initialise this process. Standardised procedure guidelines from data collection, data processing, model training, validation, and implementation are essential to move forward (Peng et al., 2021).

Diversified patient populations in conjunction with harmonised data collection would be ideal to develop and validate a high performing model for personalised medicine. The sampling population should arguably encompass a variety of, for example, tinnitus characteristics and severity, in addition to diverse demographics using various recruitment strategies (Probst et al., 2017). Moreover, standardisation of data and quality of data should be addressed prior to data collection. Of 61 variables that have been used for subgrouping; tinnitus severity, age, hearing ability, and depressive symptoms were among the most popular and most important variables (Genitsaridi et al., 2020). Future studies should consider using such variables for clinical applications of deep learning approaches. Recent advancements, such as the Whisper model for cochlear implant simulations (Sinha and Azadpour, 2024), underscore the importance of domain-specific model development. Whisper demonstrates that even state-of-the-art deep learning models can excel in controlled scenarios but may struggle in noisy or real-world conditions (Sinha and Azadpour, 2024). Future studies should focus on fine-tuning models for diverse auditory environments and patient-specific variability to enhance their clinical impact.

Cross-validation across a diversified data set is another critical component to validating a model. To further improve model generalisability, integrating multimodal datasets—such as combining imaging data, audiometric variables, and patient-reported outcomes—should become a priority. The recent work of Essaid et al. (2024), which highlights how multimodal AI systems can enhance cochlear implant surgical planning and speech processing, serves as a compelling example. Developing frameworks to harmonise and process such data will be essential for maximising their utility in clinical settings. Additionally, in classification tasks, balanced datasets will avoid skewed predictions. Cross-validation should also be conducted on separate cohort, to ensure generalisability. If the model performs similarly, and reveals similar important features, then it is likely a good model. Furthermore, testing on an external dataset is essential for model validation, especially for clinical validity (Marquand et al., 2016). Building on these principles, testing frameworks must also account for the specific requirements of advanced models like Whisper and LLMs. For example, evaluating Whisper's ability to simulate diverse patient scenarios and incorporating text-based LLM predictions into cross-validation pipelines will ensure robust assessments (Sinha and Azadpour, 2024). Additionally, metrics beyond accuracy—such as interpretability and fairness—should be emphasised to promote clinical relevance and equity.

In conclusion, we need reproducible data, combined statistical approaches, and testing on an external dataset. A repeatedly well-performing model, across a number of metrics, and on an external dataset, would indicate a validated, high performing model with the potential to generalise clinically. Ultimately, the integration of cutting-edge technologies like Whisper and LLMs, alongside multimodal datasets, represents the next frontier in otology and neurotology research (Sinha and Azadpour, 2024; Essaid et al., 2024). By prioritising standardisation, dataset diversity, and innovative validation strategies, future studies can harness the full potential of these tools to transform clinical care and enable precision medicine.

5. Conclusion

Machine and deep learning approaches continue to drive significant advancements in precision medicine, offering transformative potential for otology and neurotology. These approaches excel in identifying complex patterns within large datasets, enabling more tailored decision-making for diagnoses, prognoses, and treatments. Recent breakthroughs, such as the Whisper model for cochlear implant simulations and large language models (LLMs) for tinnitus subgroup identification, highlight the growing capabilities of advanced AI in addressing clinical challenges. Whisper's ability to simulate patient responses in auditory environments and LLMs' capacity for analysing unstructured clinical data represent critical steps forward in integrating cutting-edge technologies into clinical practice.

Despite these advancements, challenges remain. Data harmonisation, model generalisability, and multimodal integration must be prioritised to overcome limitations in current AI applications. For example, the heterogeneity of tinnitus subgroups underscores the need for standardised data collection and validation frameworks, while the adoption of multimodal approaches in cochlear implant research, as demonstrated by Essaid et al. (2024), points to the necessity of unifying diverse data sources for better clinical outcomes. Ethical considerations, including addressing bias and ensuring patient privacy, remain central to fostering trust in these technologies.

With rigorous validation, standardised protocols, and continued innovation, machine and deep learning have the potential to revolutionise personalised medicine in otology and neurotology. By addressing persistent challenges and leveraging emerging technologies, researchers and clinicians can unlock AI's full potential to enhance patient care and improve quality of life for individuals with auditory and vestibular disorders.

CRediT authorship contribution statement

Katherine Adcock: Writing – original draft. **Elva Arulchelvan:** Writing – original draft. **Nathan Shields:** Writing – original draft. **Sven Vanneste:** Writing – original draft.

References

Ali, F., et al., 2020. A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. Inf. Fusion 63.
Antman, E.M., Loscalzo, J., 2016. Precision medicine in cardiology. Nat. Rev. Cardiol. 13 (10), 591–602.
Ayoub, N.F., et al., 2024. Inherent bias in large language models: a random sampling analysis. Mayo Clin. Proc.: Digit. Health 2 (2), 186–191.
Baguley, D., D. McFerran, and D. Hall, Tinnitus. Lancet, 2013. 382(9904): p. 1600–7.
Bing, D., et al., 2018. Predicting the hearing outcome in sudden sensorineural hearing loss via machine learning models. Clin. Otolaryngol. 43 (3), 868–874.
Bzdok, D., Meyer-Lindenberg, A., 2018. Machine learning for precision psychiatry: opportunities and challenges. Biol. Psychiatry Cogn. Neurosci. Neuroimaging 3 (3), 223–230.
Côté, C., et al., 2019. Long-term effects of a multimodal physiotherapy program on the severity of somatosensory tinnitus and identification of clinical indicators predicting favorable outcomes of the program. J. Am. Acad. Audiol. 30 (8), 720–730.
Cardon, E., et al., 2022. Random forest classification to predict response to high-definition transcranial direct current stimulation for tinnitus relief: a preliminary feasibility study. Ear Hear. 43 (6), 1816–1823.
Carvalho, D.V., E.M. Pereira, and J.S. Cardoso, Machine learning interpretability: a survey on methods and metrics. Electronics (Basel), 2019. 8(8): p. 832.
Cederroth, C.R., et al., 2019. Editorial: towards an understanding of tinnitus heterogeneity. Front. Aging Neurosci. 11.
Cha, D., et al., 2019. Automated diagnosis of ear disease using ensemble deep learning with a big otoendoscopy image database. EBioMedicine 45, 606–614.
Ching, T., et al., 2018. Opportunities and obstacles for deep learning in biology and medicine. J. R. Soc. Interface 15 (141).
Collin, C.B., et al., 2022. Computational models for clinical applications in personalized medicine—guidelines and recommendations for data integration and model validation. J. Pers. Med. 12 (2), 166.
Crowson, M.G., Chan, T.C.Y., 2020. Machine learning as a catalyst for value-based health care. J. Med. Syst. 44 (9), 139.
Crowson, M.G., et al., 2020a. A contemporary review of machine learning in otolaryngology-head and neck surgery. Laryngoscope 130 (1), 45–51.
Crowson, M.G., et al., 2020b. Predicting postoperative cochlear implant performance using supervised machine learning. Otol. Neurotol. 41 (8), e1013–e1023.
Crowson, M.G., et al., 2020c. Machine learning for pattern detection in cochlear implant FDA adverse event reports. Cochlear Implants Int. 21 (6), 313–322.
Crowson, M.G., et al., 2020d. Machine learning and cochlear implantation-a structured review of opportunities and challenges. Otol. Neurotol. 41 (1), e36–e45.
Dai, S., et al., 2024. Bias and unfairness in information retrieval systems: new challenges in the LLM era. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 6437–6447.
Doborjeh, M., et al., 2023. Prediction of tinnitus treatment outcomes based on EEG sensors and TFI score using deep learning. Sensors (Basel) 23 (2).
Drakopoulos, F., Baby, D., Verhulst, S., 2022. A convolutional neural-network framework for modelling auditory sensory cells and synapses. Commun. Biol. 4, 1–17.
Dubno, J.R., et al., 2013. Classifying human audiometric phenotypes of age-related hearing loss from animal models. J. Assoc. Res. Otolaryngol. 14 (5), 687–701.
Dubois, C., et al., 2024. Development and validation of a smartphone-based deep-learning-enabled system to detect middle-ear conditions in otoscopic images. NPJ Digit. Med. 7 (1), 162.
Ellis, R.J., Sander, R.M., Limon, A., 2022. Twelve key challenges in medical machine learning and solutions. Intell. Based Med. 6, 100068.
Essaid, B., et al., 2024. Artificial intelligence for Cochlear implants: review of strategies, challenges, and perspectives. IEEE Access 12, 119015–119038.
Esteva, A., et al., 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542 (7639), 115–118.
Ezzibdeh, R., et al., 2022. Artificial intelligence and tele-otoscopy: a window into the future of pediatric otology. Int. J. Pediatr. Otorhinolaryngol. 160, 111229.
Figueroa Barraza, J., Droguett, E.López, Martins, M.R., 2021. Towards interpretable deep learning: a feature selection framework for prognostics and health management using deep neural networks. Sensors 21 (17), 5888.
Fisher, A.J. and J.F. Boswell, Enhancing the personalization of psychotherapy with dynamic assessment and modeling. Assessment, 2016. 23(4): p. 496–506.
Frasca, M., et al., 2024. Explainable and interpretable artificial intelligence in medicine: a systematic bibliometric review. Discov. Artif. Intell. 4 (1).
Genitsaridi, E., et al., 2020. A review and a framework of variables for defining and characterizing tinnitus subphenotypes. Brain Sci. 10 (12).
Genitsaridi, E., Novel approaches for tinnitus subphenotyping: evidence synthesis, standardised assessment, and supervised and unsupervised machine learning applications. 2021.
Hajiaghababa, F., Marateb, H.R., Kermani, S., 2018. The design and validation of a hybrid digital-signal-processing plug-in for traditional cochlear implant speech processors. Comput. Methods Programs Biomed. 159, 103–109.
Handler, R., Sharma, S., Hernandez-Boussard, T., 2025. The fragile intelligence of GPT-5 in medicine. Nat. Med.
Hanley, J.A., McNeil, B.J., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143 (1), 29–36.
Helpard, L., et al., 2021a. Three-dimensional modeling and measurement of the human cochlear hook region: considerations for tonotopic mapping. Otol. Neurotol. 42 (6), e658–e665.
Helpard, L., et al., 2021b. An approach for individualized cochlear frequency mapping determined from 3D synchrotron radiation phase-contrast imaging. IEEE Trans. Biomed. Eng. 68 (12), 3602–3611.
Jeong, Y., et al., 2024. Advancing Tinnitus Therapeutics: GPT-2 driven clustering analysis of cognitive behavioral therapy sessions and Google T5-based predictive modeling for THI score assessment. IEEE Access 12, 52414–52427.
Kim, J., Shin, H., 2013. Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data. J. Am. Med. Inform. Assoc. 20 (4), 613–618.
Kloostra, F.J.J., et al., 2019. Models to predict positive and negative effects of cochlear implantation on tinnitus. Laryngoscope Investig. Otolaryngol. 4 (1), 138–142.
Kokkinakis, K., et al., 2012. Single and multiple microphone noise reduction strategies in cochlear implants. Trends Amplif. 16 (2), 102–116.
Kufel, J., et al., 2023. What is machine learning, artificial neural networks and deep learning? Examples of practical applications in medicine. Diagnostics (Basel) (15), 13.
Langguth, B., et al., 2013. Tinnitus: causes and clinical management. Lancet Neurol. 12 (9), 920–930.
Langguth, B., Elgoyhen, A.B., Cederroth, C.R., 2019. Therapeutic approaches to the treatment of tinnitus. Annu. Rev. Pharmacol. Toxicol. 59, 291–313.
Leo, B., 2001. Statistical modeling: the two cultures (with comments and a rejoinder by the author). Stat. Sci. 16 (3), 199–231.
Lesica, N.A., Mehta, N., Manjaly, J.G., Deng, L., Wilson, B.S., Zeng, F.-G., 2021. Harnessing the power of artificial intelligence to transform hearing healthcare and research. Nat. Mach. Intell. 3, 840–849.
Liu, G.S., et al., 2017a. ELHnet: a convolutional neural network for classifying cochlear endolymphatic hydrops imaged with optical coherence tomography. Biomed. Opt. Express. 8 (10), 4579–4594.
Liu, H., et al., Hierarchical representations for efficient architecture search. ArXiv, 2017b. **abs/1711.00436**.
Livingstone, D., Chau, J., 2020. Otoscopic diagnosis using computer vision: an automated machine learning approach. Laryngoscope 130 (6), 1408–1413.

Lo Vercio, L., et al., 2020. Supervised machine learning tools: a tutorial for clinicians. J. Neural Eng. 17 (6), 062001.
MacEachern, S.J., Forkert, N.D., 2021. Machine learning for precision medicine. Genome 64 (4), 416–425.
Makar, S.K., 2021. Etiology and pathophysiology of tinnitus - a systematic review. Int. Tinnitus J. 25 (1), 76–86.
Maleki, F., et al., 2020. Overview of machine learning part 1: fundamentals and classic approaches. Neuroimaging Clin. 30 (4), e17–e32.
Marquand, A.F., et al., 2016. Beyond lumping and splitting: a review of computational approaches for stratifying psychiatric disorders. Biol. Psychiatry Cogn. Neurosci. Neuroimaging 1 (5), 433–447.
Moulaei, K., et al., 2024. Explainable artificial intelligence (XAI) for predicting the need for intubation in methanol-poisoned patients: a study comparing deep and machine learning models. Sci. Rep. 14 (1), 15751.
Moura, L., et al., 2024. Implications of large language models for quality and efficiency of neurologic care: emerging issues in neurology. Neurology 102 (11), e209497.
Murmu Nielsen, R. and E.R. Ng, Reducing sustained listening effort: oticon more new evidence. *Oticon Whitepaper*. 2022, Oticon.
Myburgh, H.C., et al., 2016. Otitis media diagnosis for developing countries using tympanic membrane image-analysis. EBioMedicine 5, 156–160.
Myburgh, H.C., et al., 2018. Towards low cost automated smartphone- and cloud-based otitis media diagnosis. Biomed. Signal. Process. Control 39, 34–52.
Nag, S., et al., 2022. Deep learning tools for advancing drug discovery and development. 3. Biotech 12 (5), 110.
Niemann, U., et al., 2020a. Tinnitus-related distress after multimodal treatment can be characterized using a key subset of baseline variables. PLoS One 15 (1), e0228037.
Niemann, U., et al., 2020b. Phenotyping chronic tinnitus patients using self-report questionnaire data: cluster analysis and visual comparison. Sci. Rep. 10 (1), 16411.
O'Donoghue, G., 2013. Cochlear implants–science, serendipity, and success. N. Engl. J. Med. 369 (13), 1190–1193.
Omar, M., et al., 2025. Evaluating and addressing demographic disparities in medical large language models: a systematic review. Int. J. Equity Health 24 (1), 57.
Papadakis, G.Z., et al., 2019. Deep learning opens new horizons in personalized medicine. Biomed. Rep. 10 (4), 215–217.
Pelosi, S., et al., 2013. Cochlear implantation versus hearing amplification in patients with auditory neuropathy spectrum disorder. Otolaryngol. Head Neck Surg. 148 (5), 815–821.
Peng, J., et al., 2021. Machine learning techniques for personalised medicine approaches in immune-mediated chronic inflammatory diseases: applications and challenges. Front. Pharmacol. 12.
Pisoni, D.B., et al., 2017. Three challenges for future research on cochlear implants. World J. Otorhinolaryngol. Head Neck Surg. 3 (4), 240–254.
Plant, D., Barton, A., 2020. Machine learning in precision medicine: lessons to learn. Nat. Rev. Rheumatol. 17.
Probst, T., et al., 2017. Outpatient Tinnitus Clinic, self-help web platform, or mobile application to recruit tinnitus study samples? Front. Aging Neurosci. 9.
Rahman, A., et al., 2024. Machine learning and deep learning-based approach in smart healthcare: recent advances, applications, challenges and opportunities. In: AIMS Public Health, 11, pp. 58–109.
Rajoub, B., 2020. Chapter 3 - supervised and unsupervised learning. In: Zgallai, W. (Ed.), Biomedical Signal Processing and Artificial Intelligence in Healthcare. Academic Press, pp. 51–89. Editor.
Raptis, S., Ilioudis, C., Theodorou, K., 2024. From pixels to prognosis: unveiling radiomics models with SHAP and LIME for enhanced interpretability. Biomed. Phys. Eng. Express 10 (3).
Rodrigo, H., et al., 2021. Exploratory data mining techniques (decision tree models) for examining the impact of internet-based cognitive behavioral therapy for tinnitus: machine learning approach. J. Med. Internet Res. 23 (11), e28999.
Rovers, M.M., et al., 2004. Otitis media. Lancet 363 (9407), 465–473.
Sadeghi, Z., et al., 2024. A review of explainable artificial Intelligence in healthcare. Comput. Electr. Eng. 118.
Saito, T., Rehmsmeier, M., 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 10 (3), e0118432.
Searchfield, G.D., Durai, M., Linford, T., 2017. A State-of-the-art review: personalization of Tinnitus sound therapy. Front. Psychol. 8.
Shafieibavani, E., et al., 2021. Predictive models for cochlear implant outcomes: performance, generalizability, and the impact of cohort size. Trends Hear. 25, 23312165211066174.
Shaik, T., Tao, X., Higgins, N., Li, L., Gururajan, R., Zhou, X., Acharya, R.U., 2023. Remote patient monitoring using artificial intelligence: current state, applications, and challenges. Wires: Data Min. Knowl. Discov. 13, 1–31.
Simoes, J., et al., 2019. Toward personalized tinnitus treatment: an exploratory study based on internet crowdsensing. In: Front. Public Health, 7, p. 157.
Singh, A., Sengupta, S., Lakshminarayanan, V., 2020. Explainable deep learning models in medical image analysis. J. Imaging 6 (6), 52.
Sinha, R., Azadpour, M., 2024. Employing deep learning model to evaluate speech information in acoustic simulations of cochlear implants. Sci. Rep. 14 (1), 24056.
Sunde, J., et al., 2013. Cochlear implant failure, revision, and reimplantation. Otol. Neurotol. 34 (9), 1670–1674.
Tang, R., et al., 2024. Artificial intelligence in CT diagnosis: current status and future prospects for ear diseases. Meta-Radiol. 2 (4).
Tyler, R., et al., 2008. Identifying tinnitus subgroups with cluster analysis. Am. J. Audiol. 17 (2), S176–S184.
van den Berge, M.J.C., et al., 2017. Cluster analysis to identify possible subgroups in tinnitus patients. Front. Neurol. 8, 115.
Wang, D., Chen, J., 2018. Supervised speech separation based on deep learning: an overview. IEEE/ACM. Trans. Audio Speech Lang. Process. 26 (10), 1702–1726.
Wang, D., Zhang, S., 2024. Large language models in medical and healthcare fields: applications, advances, and challenges. Artif. Intell. Rev. 57 (11).
Wei, Z., et al., 2009. From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. PLoS Genet. 5 (10), e1000678.
Wijewickrema, S., et al., 2022. Automatic analysis of cochlear response using electrocochleography signals during cochlear implant surgery. PLoS One 17 (7), e0269187.
Yang, Y., et al., 2024. Unmasking and quantifying racial bias of large language models in medical report generation. Commun. Med. (Lond) 4 (1), 176.
Zack, T., et al., Coding inequity: assessing GPT-4's potential for perpetuating racial and gender biases in healthcare. medRxiv, 2023: p. 2023.07.13.23292577.
Zahid, I.A., et al., 2024. Unmasking large language models by means of OpenAI GPT-4 and Google AI: a deep instruction-based analysis. Intell. Syst. Appl. 23.
Zenner, H.P., et al., 2017. A multidisciplinary systematic review of the treatment for chronic idiopathic tinnitus. Eur. Arch. Otorhinolaryngol. 274 (5), 2079–2091.
Zhang, S., et al., 2019. Learning for personalized medicine: a comprehensive review from a deep learning perspective. IEEE Rev. Biomed. Eng. 12, 194–208.