Iowa Initiative for Artificial Intelligence
Final Report

Project title:	Understanding the role of genetics in acquired hearing loss and tinnitus using artificial intelligence
Principal Investigator:	Ishan Sunilkumar Bhatt, Ph.D., CCC-A (Audiogenomics, Associate Professor, University of Iowa)
Prepared by (IIAI):	Avinash Mudireddy
Other investigators:	**co-I**: Raquel Dias, Ph.D. (Bioinformatics, Assistant Professor, University of Florida)
Date:	06/22/2024

Were specific aims fulfilled:	Yes
Readiness for extramural proposal?	Yes
If yes … Planned submission date	June 2024 (already submitted), revision in June 2025
Funding agency	NIH/NIDCD, CDC/NIOSH
Grant mechanism	R01
If no … Why not? What went wrong?	

Brief summary of accomplished results:

Research report:

Aims (provided by PI):

Specific Aims: About 430 million people across the globe suffer from disabling hearing loss that requires audiological services. It is estimated that over 10% of the global population (>700 million) will have disabling hearing loss by 2050[1]. Tinnitus, the phantom perception of sound without an external sound source, is a prevalent hearing condition that often accompanies acquired hearing loss. Almost 15% of the world's population experience some form of tinnitus, and about 20% of them struggle with debilitating tinnitus[2]. People with hearing loss and tinnitus often experience communication difficulties, social isolation, cognitive impairment, depression, and insomnia. Aging, hearing loss, noise exposure, and ototoxic agents are known risk factors for hearing loss and tinnitus[3]. There is no cure for acquired hearing loss and tinnitus. *There is a pressing need to identify molecular mechanisms underlying tinnitus and hearing loss for developing novel prophylactics and therapeutics*.

Heritability studies estimated ~40-70% variability in acquired hearing loss and tinnitus could be attributed to genetic variability[4]. Recent studies showed that the polygenic risk scores (**PRS**) derived from standard genome-wide association studies (**GWAS**) could account for only about 5% of the variability in these phenotypes[5,6]. The standard statistical methods employed by large-scale GWAS are largely inefficient at interrogating the influence of rare variants, gene-gene, and gene-environment interactions, collectively reducing the predictive utility of genomics and hindering the enormous potential applications of genomics in clinical audiology, otolaryngology, and other relevant biomedical areas[7]. The emergence of Artificial Intelligence (**AI**) in medicine has led to significant advances toward implementing personalized medicine. Due to its exceptional performance when utilizing complex big data, novel AI techniques represent significant potential for revolutionizing genomics in clinical research and practice[8].

This project will utilize new AI-powered algorithms and analytical techniques to shed light on the underlying association between genetics, environment, and hearing phenotypes (hearing loss and

tinnitus). **Our central hypothesis** is that AI-powered methods could significantly improve prediction of tinnitus and hearing loss compared to standard PRS. **In this preliminary study**, we plan to test the central hypothesis using the UK Biobank database (Total N>500,000, NTinnitus = 77605, NHearing Loss = 151677). Here we choose to limit our analysis to chromosome 4 because creating a deep-learning model for the entire genome is not computationally feasible (included in NIDCD/NIH R01), and chromosome 4 showed significant associations with hearing loss and tinnitus in past GWAS[5,6], which could allow us to obtain a proof-of-concept. We developed a deep-learning model for chromosome 4 using the Haplotype Reference Consortium database with high resolution whole genome sequencing data (N=30,000). It will be used to obtain a compressed representation of the genome for the UK Biobank database. The compressed representation will be used to conduct AI-powered GWAS, and for calculating AI-powered PRS. In addition, we will use deep-learning models on chromosome 4 data to predict tinnitus and hearing loss. **Our short-term goal** is to compare the receiver-operating curves (**ROC**) obtained with standard PRS, AI-powered PRS, and deep-learning-based models for predicting tinnitus and hearing loss.

Our specific aims are as follows:

Aim 1: To investigate the efficacy of standard and AI-powered PRS for acquired hearing loss. We will use the UK Biobank database (N>500,000) to conduct GWAS. The database includes the outcome variable for hearing loss (NHearing Loss = 151677, categorical outcome), non-genetic predictors (e.g., age, sex, ethnicity, noise, and music exposures), and genetic predictors (>10 million genetic markers). We will utilize REGENIE to conduct standard and AI-powered GWAS for chromosome 4. A deep learning model on the UK Biobank (Training: 70%, Testing: 30%) will be developed. ROC area-under-the-curve (AUC) will be used to compare the efficacy of standard PRS, AI-powered PRS, and deep-learning-based models for predicting the hearing loss phenotype.

Aim 2: To investigate the efficacy of standard and AI-powered PRS for tinnitus. We will use the UK Biobank database to conduct GWAS for tinnitus. The database includes the outcome variable of tinnitus (NTinnitus = 77605, categorical outcome). The methods described above (Aim 1) will be used to evaluate the efficacy of standard PRS, AI-powered PRS, and deep-learning based models for predicting the tinnitus phenotype.We will employ an explainable AI technique, SHapley Additive exPlanaions (SHAP), to study how predictor features in the AI models affect the outcome.

<u>Preliminary data:</u> We conducted a standard GWAS for tinnitus and hearing loss using the UK Biobank database (Bhatt et al., 2022)[9]. We identified significant associations between tinnitus and genetic variants in proximity to *GPM6A* (chromosome 4), a gene associated with neuropsychiatric conditions. Nineteen independent loci reached suggestive significance. The study identified 27 loci associated with hearing loss (2 loci on chromosome 4). Our studies showed that chromosome 4 is involved in hearing loss and tinnitus.

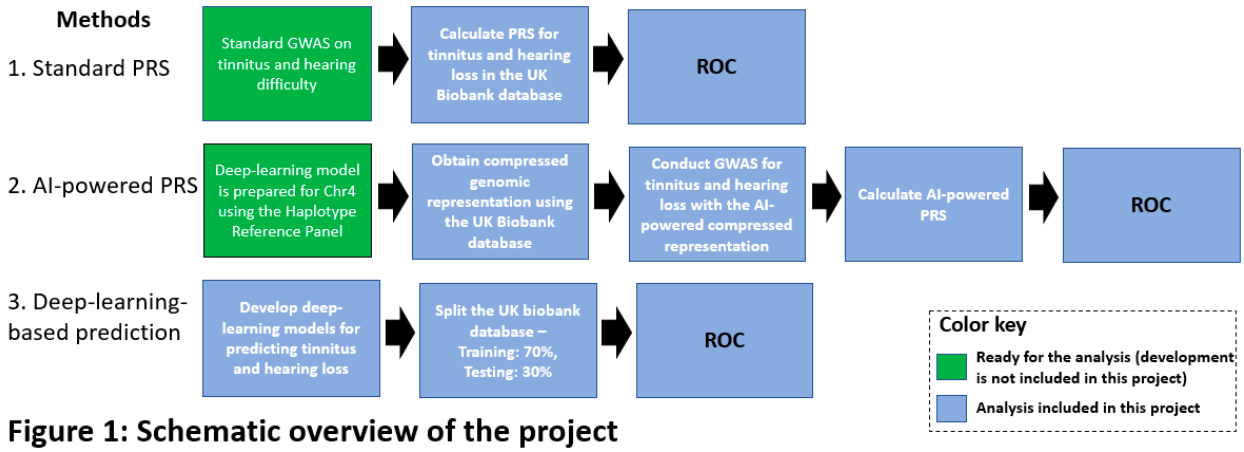


Figure 1: Schematic overview of the project

Timeline of updated goals:

- **Aim 1: Initial Chromosome Focus and Shift to SNP Filtering and Dimensionality Reduction**
  - **Aim 1.1:** Start by analyzing chromosome 4 to identify key SNPs associated with hearing loss and tinnitus.
  - **Aim 1.2:** Shift analysis from chromosome 4 to chromosome 22 due to performance challenges and assess outcomes.
  - **Aim 1.3:** Expand the analysis to cover all chromosomes, recognizing the increased complexity due to the high dimensionality introduced by millions of SNPs. Apply genome-wide suggestive significance thresholds (based on -log p-values) to filter SNPs, aiming to mitigate the curse of dimensionality.
  - **Aim 1.4:** Evaluate the impact of SNP filtering on model performance, noting any incremental improvements.
- **Aim 2: Transition to Polygenic Risk Scores and Environment Variables**
  - **Aim 2.1:** Test the effectiveness of using Polygenic Risk Scores (PRS) combined with environmental variables as model inputs.
  - **Aim 2.2:** Identify successful approaches where PRS + Environment variables demonstrate improved performance.
  - **Aim 2.3:** Test the performance of the model on a new "R21_database" of younger population
- **Aim 3: Incorporation of Phecodex Values**
  - **Aim 3.1:** Evaluate whether adding Phecodex values to the PRS + Environment variables enhances predictive accuracy.
  - **Aim 3.2:** Analyze the performance improvements and identify the optimal model configuration for genetic and environmental factors.

AI/ML Approach:

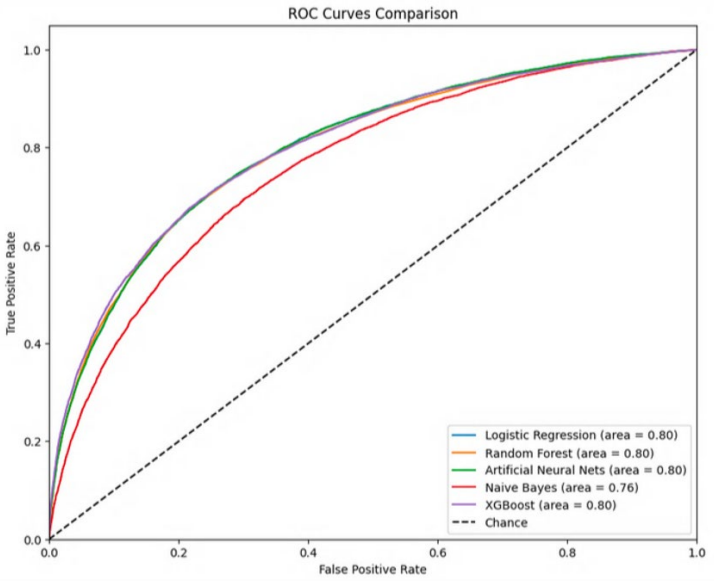
Aim 1: Dimensionality Reduction and SNP Filtering

- **Models Used:** Two primary models for SNP analysis:
  - **Model 1:** Transformer-based architecture, with token and position embeddings, batch normalization, and concatenation layers, totaling approximately 8.88 million parameters.
  - **Model 2 (Best Model):** Capsule-based model with Conv2D, PrimaryCaps, and Capsule QKV Attention layers, totaling approximately 7.06 million parameters.
- **Performance:** Model 2 achieved the best performance, yielding an AUROC of 0.636 with a threshold of 6.5.

Aim 2: Transition to Polygenic Risk Scores and Environment Variables

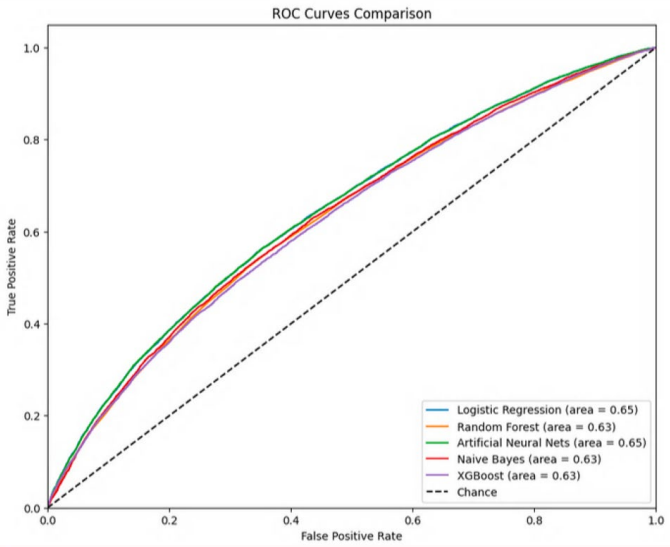
For Hearing Loss

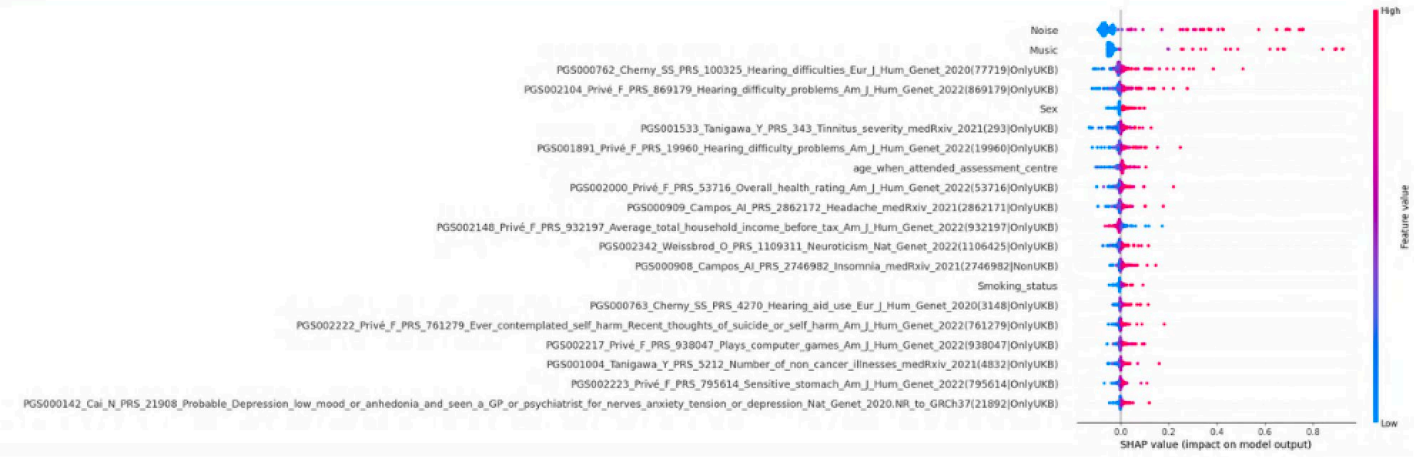
- **Models Used:** PRS models combined with environmental data, tested on diverse machine learning algorithms. (Logistic Regression, Random Forest, Naïve Bayes, XB boost, and Multi-layer Perceptron(ANN)
- **Outcome:** The PRS + Environmental variables model showed improved predictive accuracy over SNP-based models.
- **Best Model Metrics:** Multi-layer Perceptron(ANN) achieved the best performance, yielding an AUROC of 0.80



For Tinnitus

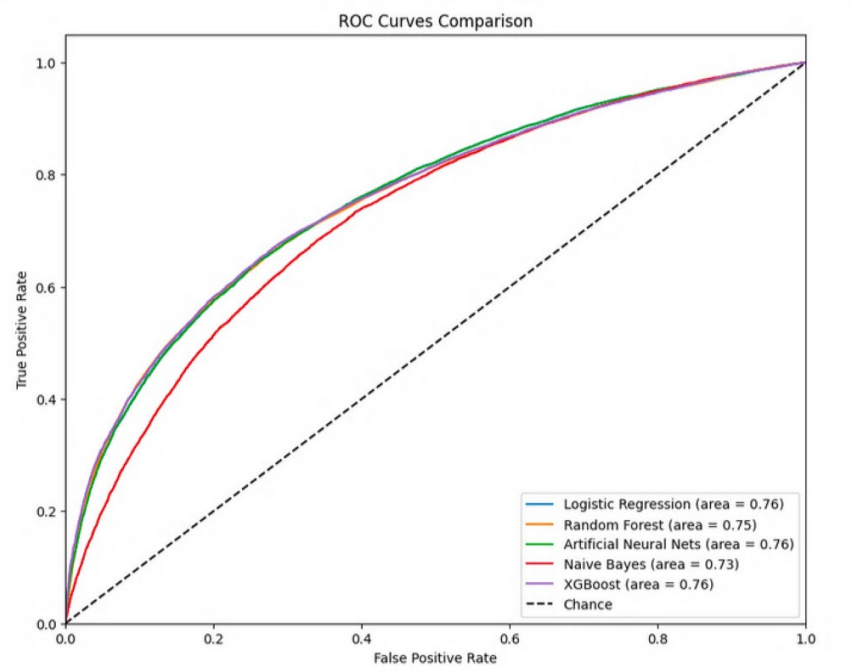
- **Models Used:** PRS models combined with environmental data, tested on diverse machine learning algorithms. (Logistic Regression, Random Forest, Naïve Bayes, XB boost, and Multi-layer Perceptron(ANN)
- **Outcome:** The PRS + Environmental variables model showed improved predictive accuracy over SNP-based models.
- **Best Model Metrics:** Multi-layer Perceptron(ANN) achieved the best performance, yielding an AUROC of 0.65





For Hearing Difficulty in Noise

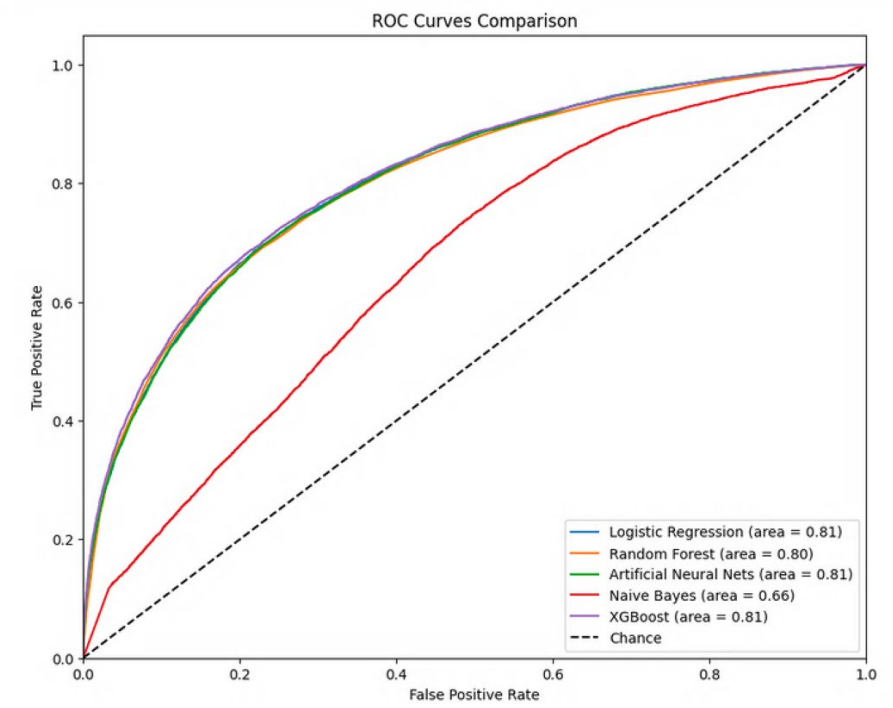
- **Models Used:** PRS models combined with environmental data, tested on diverse machine learning algorithms. (Logistic Regression, Random Forest, Naïve Bayes, XB boost, and Multi-layer Perceptron(ANN)
- **Outcome:** The PRS + Environmental variables model showed improved predictive accuracy over SNP-based models.
- **Best Model Metrics:** Multi-layer Perceptron(ANN) achieved the best performance, yielding an AUROC of 0.76

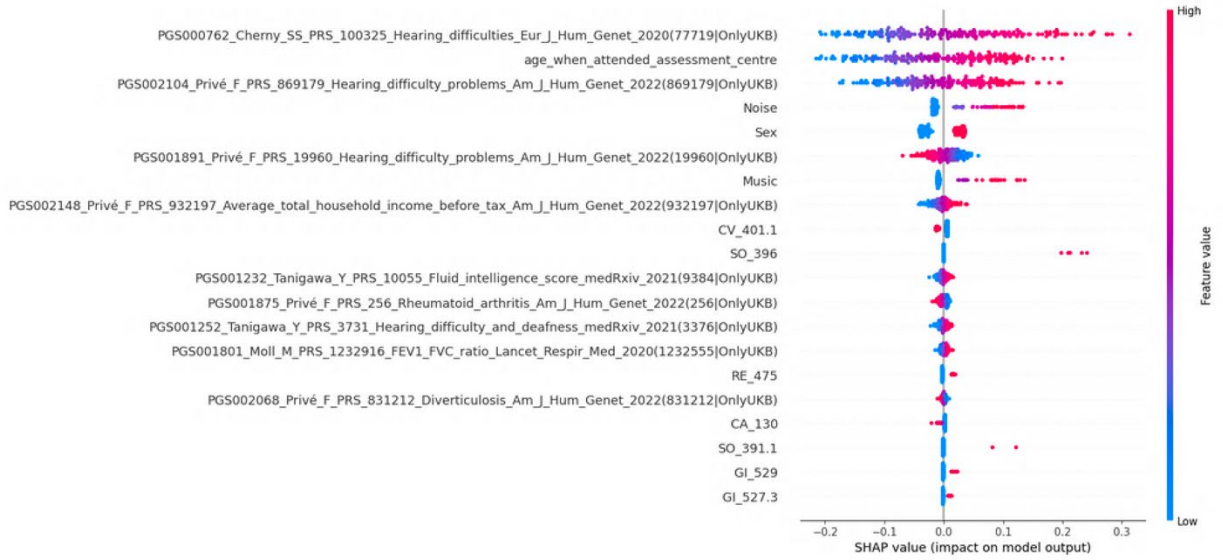


Aim 3: Incorporation of PheCODE Values

For Hearing Loss

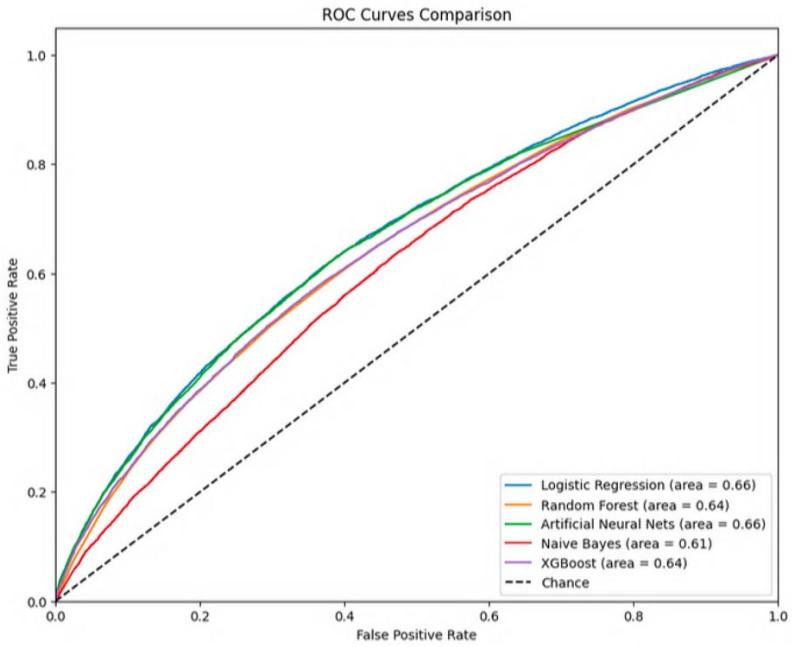
- **Models Used:** PRS + Environment variables further enhanced with PheCODE values are trained and tested on diverse machine learning algorithms. (Logistic Regression, Random Forest, Naïve Bayes, XB boost, and Multi-layer Perceptron(ANN)
- **Outcome:** The addition of PheCODE data **did not significantly increase** predictive accuracy, resulting in only a 1% gain in AUROC compared to Aim 2.
- **Best Model Metrics:** Multi-layer Perceptron(ANN) achieved the best performance, yielding an AUROC of 0.81

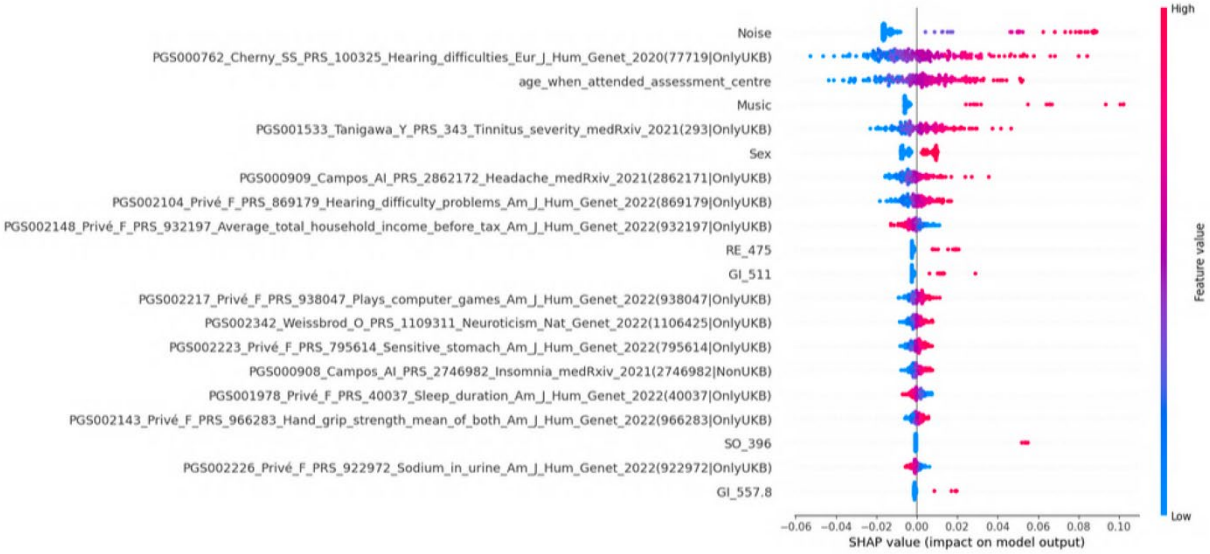




For Tinnitus

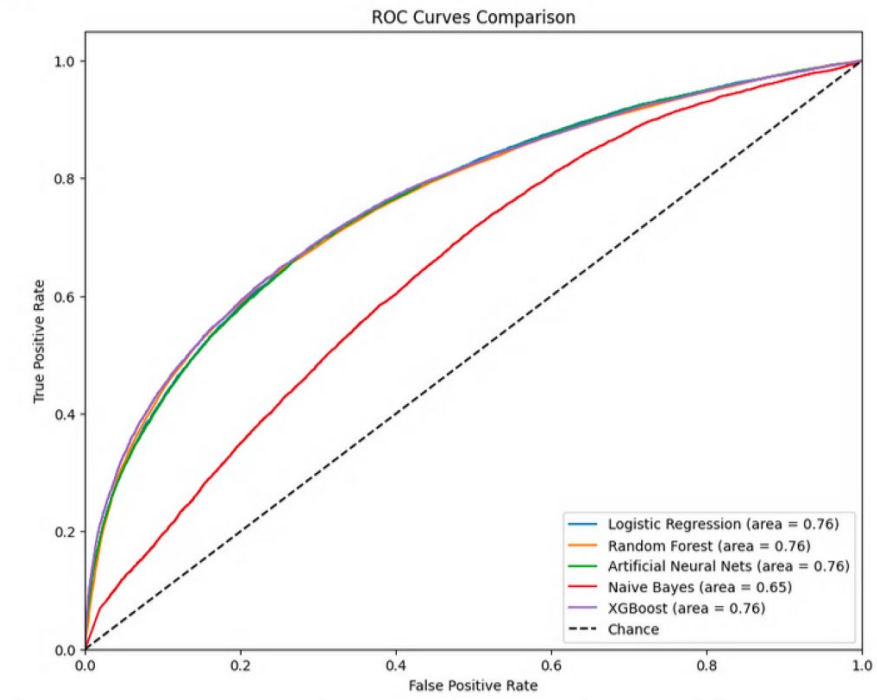
- **Models Used:** PRS + Environment variables further enhanced with PheCODE values are trained and tested on diverse machine learning algorithms. (Logistic Regression, Random Forest, Naïve Bayes, XB boost, and Multi-layer Perceptron(ANN)
- **Outcome:** The addition of PheCODE data **did not significantly increase** predictive accuracy, resulting in only a 1% gain in AUROC compared to Aim 2.
- **Best Model Metrics:** Multi-layer Perceptron(ANN) achieved the best performance, yielding an AUROC of 0.66

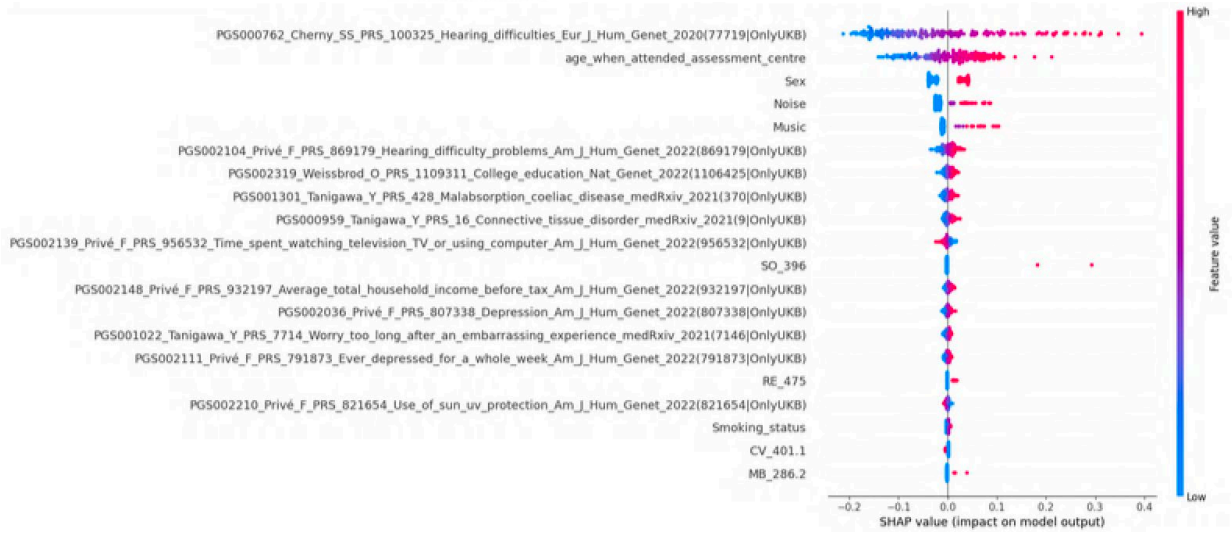




For Hearing Difficulty in Noise

- **Models Used:** PRS + Environment variables further enhanced with PheCODE values are trained and tested on diverse machine learning algorithms. (Logistic Regression, Random Forest, Naïve Bayes, XB boost, and Multi-layer Perceptron(ANN)
- **Outcome:** The addition of PheCODE data **did not increase** the predictive accuracy.
- **Best Model Metrics:** More or less all the models performed similarly. Hence, reporting Multi-layer Perceptron(ANN) yielding an AUROC of 0.76





DISCUSSION

The present study evaluated the utility of PRS across the health spectrum for predicting hearing difficulty, SIN deficits, and tinnitus. ANN emerged as one of the most efficient models for predicting hearing traits while efficiently handling multidimensional genetic and non-genetic data. We observed a wide range of PRS contributing to the prediction of ANN, which suggests that genetic predisposition to the comorbidities can influence the susceptibility to acquiring hearing traits.

Ideas/aims for future extramural projects:
We plan to extend this work by (1) applying modern AI methods, such as graphical pre-trained transformers, (2) adding gene-level predictors to improve the accuracy and interpretability of the

models, and (3) adding radiological measures that connect genes to behaviors, working as intermediate endophenotypes. We also plan to use XAI methods for stratifying genetic risk into biologically meaningful pathways and cell-types.