



OPEN

## Comparison between logistic regression and machine learning algorithms on prediction of noise-induced hearing loss and investigation of SNP loci

Jie Lu<sup>1</sup>, Xinhao Lu<sup>3,4</sup>, Yixiao Wang<sup>5</sup>, Hengdong Zhang<sup>2</sup>, Lei Han<sup>2</sup>, Baoli Zhu<sup>1,2,6,7</sup>✉ & Boshen Wang<sup>1,2</sup>✉

To compare the comprehensive performance of conventional logistic regression (LR) and seven machine learning (ML) algorithms in Noise-Induced Hearing Loss (NIHL) prediction, and to investigate the single nucleotide polymorphism (SNP) loci significantly associated with the occurrence and progression of NIHL. A total of 1,338 noise-exposed workers from 52 enterprises in Jiangsu Province were included in this study. 88 SNP loci involving multiple genes related to noise exposure and hearing loss were detected. LR and multiple ML algorithms were employed to establish the NIHL prediction model with accuracy, recall, precision, F-score, R<sup>2</sup> and AUC as performance indicators. Compared to conventional LR, the evaluated ML models Generalized Regression Neural Network (GRNN), Probabilistic Neural Network (PNN), Genetic Algorithm-Random Forests (GA-RF) demonstrate superior performance and were considered to be the optimal models for processing large-scale SNP loci dataset. The SNP loci screened by these models are pivotal in the process of NIHL prediction, which further improves the prediction accuracy of the model. These findings open new possibilities for accurate prediction of NIHL based on SNP locus screening in the future, and provide a more scientific basis for decision-making in occupational health management.

**Keywords** NIHL, Machine learning, Logistic regression, SNP loci

Noise-Induced Hearing Loss (NIHL) is a common sensory-induced hearing impairment caused by long-term exposure of workers to high intensity noise<sup>1,2</sup>. Approximately 16% of disabling cases of adult hearing loss worldwide can be attributed to occupational noise exposure<sup>3</sup>. It is a complex multi-factorial disease resulting from the combined effects of genetic, environmental and life behavior factors<sup>4,5</sup>. Numerous animal experiments have confirmed the role of genetic factors in NIHL susceptibility<sup>6,7</sup>. There is growing evidence that significant differences in susceptibility to NIHL exist between individuals<sup>8</sup>. Based on epidemiological studies of noise-exposed populations, susceptibility associations have been found between NIHL and single nucleotide polymorphisms (SNPs) in several genes, including HDAC2, SOD2, and STAT3<sup>9-11</sup>. Therefore, in-depth mining of the potential information in SNP loci data pertaining to genetic susceptibility to NIHL is the key in accurately predicting the occurrence and progression of NIHL, which has significant practical value for the early prevention, accurate diagnosis and timely treatment of NIHL.

With the rise of data science and artificial intelligence, various disease risk prediction models have gained widespread use<sup>12</sup>. Logistic regression (LR), a generalized linear model, is usually the primary choice for predicting binary classification outcomes (e.g. the presence or absence of disease)<sup>13</sup>. In recent years, it has widely been applied to explore the susceptibility associations between SNP loci and diseases<sup>14,15</sup>. However,

<sup>1</sup>Key Laboratory of Environmental Medicine Engineering of Ministry of Education, Southeast University, Nanjing 210009, China. <sup>2</sup>Institute of Occupational Disease Prevention, Jiangsu Provincial Center for Disease Prevention and Control (Jiangsu Academy of Preventive Medicine), No. 172 Jiangsu Road, Nanjing 210000, Jiangsu, China. <sup>3</sup>School of Cyber Science and Engineering, Southeast University, Nanjing 211189, China. <sup>4</sup>State Grid Jiangsu Electric Power Co., Ltd. Nanjing Power Supply Branch, Nanjing 210012, China. <sup>5</sup>School of Public Health, Nanjing Medical University, Nanjing 211166, Jiangsu, China. <sup>6</sup>Jiangsu Preventive Medical Association, Nanjing 210000, Jiangsu, China. <sup>7</sup>Nanjing Medical University, Nanjing 211166, Jiangsu, China. ✉email: zhulcd@ sina.com; doudouwbs@126.com

when used to predict NIHL, LR shows limitations in genetic information mining, with accuracy, recall, and precision often unsatisfactory. The effectiveness and reliability of applying the SNP loci selected by LR for NIHL prediction remains to be validated. In contrast, Machine learning (ML) algorithms, as an essential branch of artificial intelligence, have demonstrated superior capability in predicting acute kidney injury<sup>16</sup>, breast cancer<sup>17</sup>, hypertension<sup>18</sup> and other diseases due to their excellent performance and efficiency. They have become potential substitute to LR and other conventional statistical methods, such as neural networks, random forests, decision trees, etc<sup>19–21</sup>. Nowadays, ML has made remarkable progress in both the theory and application of neural networks, among which Probabilistic Neural Network (PNN), Generalized Regression Neural Network (GRNN) standing out as representative models. Compared with other ML algorithms, both of them show stronger data adaptability due to the use of hyperparameter optimization, especially when dealing with nonlinear and complex datasets. Hyperparameters are manually set before training, and unlike parameters that are automatically adjusted during training (like the weight of the neural network), they directly control the training process, thus affecting the training efficiency and final performance of the model. Minimizing the prediction error by manually setting appropriate hyperparameters (e.g. learning rate, regularization strength and kernel function type) before training starts can effectively improve model accuracy. Currently, PNN and GRNN have been widely used to learn and perform medical image data discrimination, predict disease survival, and analyze clinical decisions<sup>22,23</sup>.

Since the fundamental differences in model construction, processing data capability and complexity among various algorithms, the execution efficiency of using different models on the same datasets may vary significantly. According to our knowledge, there is no study that systematically compares and analyzes LR with different ML algorithms to clarify the applicability of each algorithm in NIHL risk assessment and early warning. Therefore, this study performed a comprehensive analysis and comparison of the model performance of LR and seven different ML algorithms in NIHL prediction. We hope to identify more accurate prediction models for NIHL, which can be applied in the early screening of susceptible individuals during pre-employment medical examinations and the early screening of high-risk individuals already working in noisy environments, to prevent the occurrence and further progression of NIHL.

## Materials and methods

### Study population

This study initially screened 1,490 workers exposed to occupational noise from 52 noise-exposed enterprises covered by the Occupational Disease Hazard Surveillance System of Jiangsu Province, following the inclusion and exclusion criteria outlined below.

Inclusion criteria: (1) Chinese Han workers; (2) A history of occupational noise exposure  $\geq 3$  years; (3) Complete occupational health surveillance materials; (4) The levels of occupational hazards (heavy metals, organic solvents, CO, high temperature and vibration, etc.) that may affect NIHL except noise in the work environment are below the requirements of occupational exposure limits (OELs).

Exclusion criteria: (1) A clear family history of hereditary deafness or a current medical history of diseases that could affect hearing; (2) A history of head trauma or blast deafness; (3) Have taken or currently taking ototoxic drugs (e.g., quinolones, aspirin, aminoglycosides, etc.).

During the health check-up, the study population completed an occupational health questionnaire under the guidance of trained and assessed investigators or on their own. The questionnaires mainly included gender, age, smoking habits, alcohol consumption, medication use, occupational history medical history.

The noise exposure intensity measurement data in the working environments of these 52 enterprises, employees' previous noise exposure records, occupational health physical examination data, and SNP genotyping data were all derived from the database of Jiangsu Provincial Center for Disease Control and Prevention.

The study protocol has been reviewed and approved by the Ethics Committee of Jiangsu Provincial Center for Disease Control and Prevention. All research was performed in accordance with relevant guidelines and regulations and in accordance with the Declaration of Helsinki. All the participants are informed about the study, and they have all signed the informed consent form.

### Noise exposure intensity measurement

The noise exposure levels in the work environment are measured according to the "Measurement of Physical Factors in the Workplace, Part 8: Noise" national standard (GBZ/T 189.8–2007). Noise exposure measurements were conducted three times a year at selected workplaces using a Quest Noise Pro-DL multifunctional personal noise dosimeter (Quest, USA). Prior to each measurement, the equipment was calibrated and the results were converted to an 8-hour equivalent continuous A-weighted sound pressure (LEX, 8 h) to represent the noise exposure intensity.

### Pure-tone audiometry and the definition of NIHL

According to the provisions of Chinese Diagnosis of Occupational Noise Deafness(GBZ 49-2014), all the study population had to be detached from the noise environment for at least 48 h before undergoing PTA. The formal test was conducted in an anechoic chamber with good soundproofing effect (background noise value  $< 25$ dB (A)), an experienced occupational doctor used an audiometer to measure the hearing threshold of both ears of the study population at a total of 6 frequencies: 0.5, 1.0, 2.0, 3.0, 4.0 and 6.0 kHz. All hearing threshold measurement results were adjusted for age and gender in accordance with the "Acoustics-Statistical Distribution of Hearing Threshold and Age and Gender". Participants exhibiting an average hearing threshold  $> 25$  dB(A) at high frequencies (3.0, 4.0, and 6.0 kHz) in one or both ears were assigned to the case group, and the control was frequency matched for age, gender, smoking habit, alcohol consumption and other factors.

## Blood sample collection and DNA extraction

A vacuum Ethylene Diamine Tetraacetic Acid (EDTA) anticoagulation blood collection tube was used to collect 5 mL venous blood from each participant for genomic DNA extraction.

DNA extraction kit provided by Tiangen Biotechnology Co., Ltd. (Beijing, China) was used to extract genomic DNA from blood samples according to the instructions and preserved at -80°C for later use.

## SNP selection, quality control, and genotyping

### Selection

By consulting the Thousand Genomes Database (<http://www.1000genomes.org/>) and the National Center for Biotechnology Information (NCBI) dbSNP database (<https://www.ncbi.nlm.nih.gov/snp/>) to screen suitable SNP loci, screening criteria as illustrated below:

- (1) The SNP loci frequently reported in both Chinese and English literature over the past decade as being associated with NIHL.
- (2) The minor allele frequency (MAF) corresponding to locus > 0.05.
- (3) The linkage disequilibrium (LD) value between any two loci is  $r^2 > 0.8$ .

### Quality control

The SNP loci screened according to the above criteria were firstly processed by TASSEL 5.0<sup>24</sup> software, including missing data processing, genotype filtering and data format conversion, to ensure data quality and compatibility. Then the pLINK v1.07<sup>25</sup> with the command line option “-indep-pairwise” was used to prune the SNP loci. Across the entire genome, the LD between all SNPs pairs in the window is calculated by sliding forward with 50 consecutive SNPs as the window size and 10 SNPs as the step size. If the  $r^2$  value between any two SNP loci exceeds 0.5, one of them is marked as redundant and removed from the dataset. In addition, we also compared each SNP site with the known SNP loci in the HapMap3 database<sup>26</sup> to further verify the effectiveness and reliability of the SNP screening.

*Note:* The “-indep-pairwise” command option refers to the process of using a sliding window approach, calculating the LD value between each pair of SNPs (typically using  $r^2$  as the measure), and pruning redundant SNP loci according to the set threshold; The HapMap3 database is known for their rigorous quality control (e.g., MAF  $\geq 5\%$ , genotype leak detection rate  $> 95\%$ , Hardy-Weinberg equilibrium  $p > 1 \times 10^{-6}$ , etc.).

### Genotyping

The genotyping of SNP loci in this study was entrusted to Shanghai Biowing Applied Biotechnology Company (<http://www.biowing.com.cn/>) utilizing multiplex PCR and next-generation sequencing technology<sup>27</sup>.

## Statistical analysis

All data were processed and analyzed employing SPSS 27.0 software. Among them, the continuous variables (age, noise exposure levels, etc.) did not satisfy normal distribution with median and interquartile range M (P25,P75) and the Mann-Whitney U test was performed for comparative analysis; Categorical variables (like age, gender, smoking habit, and drinking consumption) were compared using Pearson's  $\chi^2$  test. The statistical significance was defined as  $P < 0.05$ . The genotypes of the 88 SNP loci, which were finally coded as 0, 1, and 2 to respectively represent wild type, heterozygous type, and mutant type, respectively, to indicate the number of alleles at each SNP locus. Additionally, goodness-of-fit chi-square test was used to verify whether the gene frequency distribution of each SNP locus in the whole population complied with Hardy-Weinberg law of genetic equilibrium ( $P$  values  $> 0.05$ ).

## Models and model building strategies

### Models

#### ① LR (Logistics Regression)

Logistic Regression (LR) is a widely used statistical modeling technique for binary classification problems. The core idea is to model the probability of the event of interest for a binary response variable as a function of covariates. The modelling is done using the logit function. The formula can be expressed as follows<sup>28</sup>:

$$\text{logit}(P) = \ln \left( \frac{P}{1 - P} \right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \quad (1)$$

Among them:  $P$  represents the probability of the dependent variable  $Y$  being 1, given the independent variables  $x_1, x_2, \dots, x_p$ .  $x_1, x_2, \dots, x_p$  represent the genotype values corresponding to the single nucleotide polymorphisms (SNPs).  $\beta_0$  is the intercept term.  $\beta_0, \beta_1, \dots, \beta_p$  are the regression coefficients corresponding to the independent variables  $x_1, x_2, \dots, x_p$ .

#### ② DT (Decision Tree)

Decision tree (DT) is a common classification and regression model, which is mainly based on the principle of tree structure in probability theory, information theory (especially concepts such as information gain or Gini impurity) and graph theory, and constructs a tree structure by splitting the data into several branches through recursive partitioning, so as to be able to efficiently classify the target variable homogeneous, in which each internal node represents a test condition of a feature attribute, each branch represents a result of the test, and

each leaf node corresponds to a category label (classification task). During the construction process, the decision tree algorithm tries to select the best features to segment the data to maximize criteria such as information gain (ID3 algorithm), Gini impurity (CART algorithm), or information gain rate (C4.5 algorithm), which enhances the classification accuracy of the model.

To effectively predict NIHL based on SNP genotype data, this study utilizes information entropy and information gain to determine the most informative SNP features for classification. Information entropy quantifies the impurity or uncertainty in a dataset, while information gain measures the reduction in entropy when the dataset is split based on a given attribute. These metrics help in selecting the most significant SNPs contributing to NIHL risk prediction<sup>29</sup>.

The entropy of a dataset is  $S$  given by:

$$info(S) = - \sum_{i=1}^n \frac{freq(C_i, S)}{|S|} \log_2 \left( \frac{freq(C_i, S)}{|S|} \right) \quad (2)$$

where  $S$  represents the dataset containing all samples (genotype and NIHL phenotype data for all individuals);  $C_i$  denotes the  $i$ -th class in the dataset (e.g., NIHL cases and controls);  $|S|$  is the total number of samples in  $S$ ;  $freq(C_i, S)$  is the frequency of class  $C_i$  in dataset  $S$ .

When a SNP locus  $X$  is introduced and splitting  $S$  based on an attribute, it results in subsets  $S_j$ . The entropy of the partitioned dataset is given by:

$$info_X(S) = \sum_{j=1}^m \frac{|S_j|}{|S|} info(S_j) \quad (3)$$

Here,  $S_j$  is a subset of  $S$  after splitting by attribute  $x$ .  $|S_j|$  is the number of samples in subset  $S_j$ .  $m$  is the number of subsets created after the split.

Equation (4) can be obtained from Eqs. (2) and (3) for information gain measurement:

$$gain(X) = info(S) - info_X(S) \quad (4)$$

### ③ GBDT (Gradient Boosting Decision Tree)

The Gradient Boosting Decision Tree (GBDT) model, an ensemble tree-based approach, has become widely used for regression tasks. Unlike traditional single-tree methods such as M5Tree or Random Forest, GBDT builds a complex tree by training on data weighted differently, which helps to reduce bias. The GBDT algorithm's predictive function, denoted as  $F(x)$ , is formulated as follows:

$$\left\{ F(x; \{\beta_m, \mathbf{a}_m\}_1^M) = \sum_{m=1}^M \beta_m h(x; \mathbf{a}_m) \right. \quad (5)$$

$$\left\{ h(x; \{b_j, R_j\}_1^J) = \sum_{j=1}^J b_j 1(x \in R_j) \right. \quad (6)$$

$$\left\{ F_m(x) = F_{m-1}(x) + \sum_{j=1}^J \gamma_{jm} 1(x \in R_{jm}) \right. \quad (7)$$

Each individual tree partitions the input space into  $j$  distinct segments, denoted as  $R_{1m}, R_{2m}, \dots, R_{jm}$ , with a constant value  $\gamma_{jm}$  assigned to each region  $R_{jm}$ . Here,  $\mathbf{a}_m$  denotes the average value corresponding to the terminal nodes and indicates the points at which the variables of each decision tree are split.  $b_j$  is the predicted value of the leaf node  $R_j$ , representing the fitted output of all samples in the region (usually the mean of the samples in the region). Furthermore,  $\beta_m$  represents the weights attributed to the nodes within each tree. For an exhaustive explanation of this model, please refer to the work of Friedman<sup>30</sup>.

### ④ KNN (K-Nearest Neighbor)

The algorithm flow of K-Nearest Neighbor (KNN) is outlined as follows<sup>31</sup>: Initially, K cluster centers are randomly assigned, and sample points to be classified are then grouped into respective classes based on the principle of nearest proximity. Following this using the average method to recalculate the centroid of each class, to establish the new clustering center. This process continues iteratively until the distance between each sample point and its assigned cluster center is minimized.

$$T = (x_1, y_1), (x_2, y_2), \dots, (x_N, y_N) \quad (8)$$

$T$  represents the training dataset containing  $N$  data points, each consisting of a pair  $(\mathbf{x}_i, y_i)$ .  $\mathbf{x}_i$  denotes the feature vector of each sample, representing the SNP genotype data of an individual.  $y$  is the class of instances.  $i$  is the constant, sequence numbers 1, 2, 3...  $N$ . Based on the measured distance, the  $k$  points in  $T$  (sample) that are nearest to the classified object are found, and the region encompassing all  $k$  points is denoted as  $N_k(\mathbf{x})$ . The class of  $\mathbf{x}$  within  $N_k(\mathbf{x})$  is determined based on the classification decision (minority follows majority voting principle):

$$\hat{y} = \arg \max_{y \in Y} \sum_{(\mathbf{x}_i, y_i) \in N_k(\mathbf{x})} \mathbb{I}(y_i = y), \quad i = 1, 2, \dots, N \quad (9)$$

$Y$  denotes the set of all possible category labels (1 for NIHL cases and 0 for healthy controls). The  $\mathbb{I}$  is the indicator function, where  $\mathbb{I}(y_i = y) = 1$  if  $y_i = y$ , and  $\mathbb{I}(y_i = y) = 0$  otherwise. A special case of the  $k$ -nearest neighbor method occurs when  $k = 1$ , which simplifies to the nearest neighbor algorithm. In the nearest neighbour method, the class of the input instance  $\mathbf{x}$  is determined by the class of the nearest point in the dataset to  $\mathbf{x}$ .

#### ⑤ XGBoost (eXtreme Gradient Boosting)

eXtreme Gradient Boosting (XGBoost) is an ensemble learning model composed of multiple CART (classification regression) tree combinations, which generally have strong generalization ability, can effectively avoid high degree of fitting, enabling large-scale parallel computation.

The XGBoost model<sup>32</sup> can be represented as:

$$\hat{y}_i = \varphi(\mathbf{x}_i) = \sum_{k=1}^K f_k(\mathbf{x}_i), \quad f_k \in F \quad (10)$$

where  $\hat{y}_i$  is the final prediction result, and  $\varphi(\mathbf{x}_i)$  is the prediction score of sample, and  $K$  is the total number of trees, and  $f_k$  denotes the specific first  $k$  CART tree.  $F$  represents the functional space of all possible regression trees (e.g., CART trees).

As shown in the above equation, the XGBoost model is a forward iterative model, which needs to solve for the  $t$ -th tree according to the objective function and the model containing the first  $t-1$  tree variables, i.e., the process entails identifying an optimal set of parameters to minimize the objective function. The objective function of the XGBoost model compose of two sections: the first one is the damage function, which quantifies the discrepancy between the predicted values and the true values of the model; the second part is the regularization term, which constrains the model complexity and thus helps to prevent overfitting to a certain extent. The objective function is expressed as:

$$Obj^{(t)} = \sum_{i=1}^n l\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{j=1}^t \Omega(f_j) \quad (11)$$

$$= \sum_{i=1}^n \left[ l\left(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)\right) \right] + \Omega(f_t) + \text{constant} \quad (12)$$

For the regularization term,  $\gamma$  and  $\lambda$  are two regularization coefficients used to balance and adjust the model's complexity and generalization capacity.

$$\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T \omega_j^2 \quad (13)$$

Here,  $T$  denotes the number of leaf nodes in each decision tree, while  $\omega$  is the number of leaf nodes in the first  $j$  vector of scores on the first leaf node. Since  $T$  and  $\omega$  are known, so we can consider the regularization terms involving the first  $t-1$  trees as constants, which will have no effect on the model optimization.

The objective function and gain function after the expansion optimization using second order Taylor series are:

$$Obj(\mathbf{x}) = -\frac{1}{2} \sum_{j=1}^T \frac{G_j^2}{H_j + \lambda} + \gamma T \quad (14)$$

$$\text{Gain} = \frac{1}{2} \left[ \frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma \quad (15)$$

#### ⑥ GA-RF (Genetic Algorithm-Random Forests)

Genetic Algorithm (GA) is classified as a heuristic-based search algorithm, and its core is to simulate the survival of the fittest and the genetic variation of individual genes in nature, so as to explore the optimal solution in the problem domain. Within this framework, each individual in the population represents a potential solution in the solution space. Through continuous selection, crossover, and mutation, an adaptive search is carried out to identify the optimal solution to the problem.

Random Forest (RF) is an integrated learning method built upon categorical regression trees, its core principle is Bagging (Bootstrap aggregating) technique. Some samples are randomly extracted from the overall dataset multiple times to form multiple sample subsets, and each subset is used to train an independent decision tree, which achieves effective modeling and prediction of complex data by constructing multiple decision trees and integrating their predictions to ultimately determine the final output.

By combining GA and RF, an innovative integrated learning framework Genetic Algorithm-Random Forests (GA-RF) is constructed<sup>33</sup>. This framework cleverly integrates the global optimization search capability of GA with the robust prediction performance of RF, and intelligently encodes and evolutionarily optimizes the parameters and inter-model weights of RF through GA without being restricted by the continuity or derivability of the objective function, thus greatly simplifying the complexity of the model parameter tuning<sup>34</sup>. The GA-RF model not only inherits the global exploration flexibility and efficiency of GA, but also ensures the accuracy and generalization ability of RF prediction results, which ultimately forms an efficient and robust combinatorial prediction model. The operation flow of the GA-RF algorithm is shown in Supplementary Fig. S1.

#### ⑦ PNN (Probabilistic Neural Networks)

Probabilistic Neural Networks (PNN), a forward neural network based on Bayesian classification rules and the Parzen window method, comprises four fully-interconnected layers: input, pattern, summation, and output<sup>35</sup>. The pattern layer's activation is realized through an exponential function.

Operating on probability theory, PNN classifies input vectors by calculating their Probability Density Functions (PDFs) for different classes. It leverages statistical class center vectors and variance information for efficient sample classification. Each input vector is assumed to be independently and randomly drawn from a class, thus having a corresponding PDF. By computing the PDFs of all relevant classes, PNN determines the category of an input vector. Its performance hinges on two factors: the number of neurons in the pattern layer and the choice of an appropriate activation function<sup>36</sup>.

PNN's classification mechanism relies on specific functional expressions. For each neuron in the pattern layer, the following activation function measures the relationship crucial for PDF calculation:

$$\varphi(\mathbf{x}) = \exp\left(-\frac{(\omega_i - \mathbf{x})^T(\omega_i - \mathbf{x})}{2\sigma^2}\right) \quad (16)$$

This formula describes how the input vector  $\mathbf{x}$  interacts with the weight-related term  $\omega_i$ , scaled by the smoothing parameter  $\sigma$ . It is vital for subsequent probability-based computations, enabling PNN to classify input vectors by evaluating functional values. Here,  $\omega_i$  represents the information weight from the input layer (with  $\mathbf{x}$  as the input vector), and  $\sigma$ , which depends on the input data, is the smoothing parameter.

The PDF estimate for class A is given by<sup>37</sup>:

$$f_A(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n} \frac{1}{m_A} \sum_{i=1}^{m_A} \exp\left[-\frac{(\mathbf{x} - \mathbf{x}_{Ai})^T(\mathbf{x} - \mathbf{x}_{Ai})}{2\sigma^2}\right] \quad (17)$$

where  $\mathbf{x}_{Ai}$  is the i-th training pattern in class A,  $n$  is the dimension of the input vector,  $m_A$  is the number of training patterns in class A, and  $\sigma$  corresponds to the standard deviation in a Gaussian distribution. The network's decision-boundary nonlinearity can be adjusted by modifying  $\sigma$ . A large  $\sigma$  makes the decision boundary close to a hyperplane, while a small  $\sigma$  approaching zero results in a highly nonlinear decision surface similar to that of a nearest-neighbor classifier. The PNN structure is shown in Supplementary Fig. S2.

#### ⑧ GRNN (Generalized Regression Neural Networks).

Generalized Regression Neural Networks (GRNN), first introduced by D.F. Specht in 1991, is a powerful and widely-applicable neural network model. It consists of an input layer, a pattern layer, a summation layer, and an output layer<sup>38</sup>. GRNN offers several advantages, including fast learning, good consistency, and the ability to achieve optimal regression for large-scale samples. GRNN is based on probabilistic regression analysis theory, typically using Parzen window estimation to construct the Probability Density Function (PDF) from observed data samples.

Assume that  $\mathbf{x}$  is a random vector variable and  $y$  is a random scalar variable. Let  $\mathbf{X}$  and  $y$  be the measurements, and  $f(\mathbf{X}, y)$  be their known continuous joint PDFs. The conditional expectation of  $y$  given  $\mathbf{X}$  is<sup>39</sup>:

$$E(y|\mathbf{X}) = \frac{\int_{-\infty}^{\infty} y f(\mathbf{X}, y) dy}{\int_{-\infty}^{\infty} f(\mathbf{X}, y) dy} \quad (18)$$

Here,  $y$  is the output of GRNN prediction.  $\mathbf{X}$  is the input vector composed of  $n$  predictor variables  $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$ .  $E(y|\mathbf{X})$  represents the expected value of the output  $y$  given the input vector  $\mathbf{X}$ , and  $f(\mathbf{X}, y)$  is the joint probability density function of  $\mathbf{X}$  and  $y$ .

The regression estimate  $\hat{Y}(\mathbf{x})$  is computed as:

$$\hat{Y}(\mathbf{x}) = \frac{\sum_{i=1}^n Y^i \exp\left(-\frac{D_i^2}{2\sigma^2}\right)}{\sum_{i=1}^n \exp\left(-\frac{D_i^2}{2\sigma^2}\right)} \quad (19)$$

The squared distance  $D_i^2$  is defined as:

$$D_i^2 = (\mathbf{x} - \mathbf{x}^i)^T * (\mathbf{x} - \mathbf{x}^i) \quad (20)$$

The variable  $\sigma$  is the smoothing parameter. A larger  $\sigma$  can smooth out noisy data, while a smaller  $\sigma$  allows the estimated regression surface to exhibit the desired nonlinearity, enabling  $\hat{Y}$  to approach the actual observations.

Based on non-parametric kernel regression, GRNN uses sample data as the posterior probability verification condition for non-parametric estimation. Its key advantage is the convenient setting of network parameters. By adjusting the smoothing factor in the kernel function, one can easily optimize the network's performance. This simplifies network training and learning. GRNN operates by estimating input-output relationships through PDFs, demonstrating strong nonlinear mapping capabilities, fast learning speed, and good prediction performance even with limited sample data. Additionally, it can effectively handle unstable data. The structure of GRNN is shown in Supplementary Fig. S3.

#### Model Building strategies

To better compare the model performance of conventional LR with different ML algorithms in NIHL prediction and to identify SNP loci that significantly impact the occurrence and progression of NIHL, we selected LR along with five classical ML algorithms and two hyperparameter-optimized ML algorithms, as outlined in Supplementary Table S1, to construct NIHL prediction models, and adopted the following model construction strategies.

Firstly, the conventional LR was used to conduct univariate analysis for all SNP loci eventually included in the study, and some non-significant SNP loci were eliminated in advance, and the significance level  $\alpha = 0.10$  was set. If the  $P$  value was below 0.10, this SNP locus was included in the multifactorial LR model to evaluate the combined impact of multiple loci on the occurrence and progression of NIHL. Confounding factors such as gender, age, smoking habit and drinking consumption were adjusted by multifactorial LR, and the significance level  $\alpha = 0.05$  was set. If the  $P$  value was below 0.05, this SNP locus was an independent factor associated with the occurrence and development of NIHL, otherwise, it was considered to be non-significantly associated. Finally, SNP loci statistically associated with the occurrence and development of NIHL were selected as candidate pathogenic SNP loci to be verified.

Subsequently, five classical ML algorithms (DT, GBDT, KNN, XGBoost, GA-RF) and two hyperparameter-optimized ML algorithms (PNN and GRNN) were applied to the SNP loci screened by LR to evaluate the accuracy and effectiveness of applying these loci for NIHL prediction. Meanwhile, PNN and GRNN were used for feature extraction of all SNP loci, and the SNP loci with the top 10 ranked feature importance were selected for modeling and analysis using each of the above algorithms.

Finally, LR and seven ML algorithms were used for pattern recognition and modeling of all SNP loci, aiming at more accurate prediction of NIHL and further evaluating the performance of each algorithmic model when dealing with large-scale SNP locus datasets. In this process, DT, GBDT, KNN, XGBoost, and GA-RF employed 10-fold cross-validation model perform parameter tuning. The obtained dataset was randomly divided into 10 equal-sized subsets, where 7 subsamples were served as the training set and the remaining 3 subsamples were used as the test set. The cross-validation process was repeated 10 times. For PNN and GRNN, not only normal hearing and abnormal hearing signals are considered, but also the running speed of the algorithm and the dynamic updating of the features are considered. Additionally, in order not to lose generality, a random method was employed to generate training sets and test sets. Among the 1338 samples in each category, 1170 samples (585 per category) were randomly selected as the training set and the remaining 234 samples (117 per category) were used as the test set. During each cross-validation, the number of normal and pathological cases is ensured to be equal. What is more critical is that reasonable hyperparameters should be manually set before training begins to obtain the best predictive performance.

In this study, SPSS 27.0 software was used to establish the LR model, which was accepted as statistically significant at the 0.05 level, with all tests being two-tailed. All ML models were developed and implemented using MATLAB 9.0 (R2016a).

#### Models performance evaluation

In an effort to evaluate the prediction performance of the model, the accuracy, recall, precision, F-score,  $R^2$  and AUC were selected as the performance indicators to comprehensively discriminate and compare the model performance of LR and seven ML algorithm models.

For LR, we can calculate true positive (TP), false positive (FP), true negative (TN) or false negative (FN) according to the number of cases and controls defined by the monaural or binaural high-frequency average hearing threshold. These values are then used to construct the classification matrix, as detailed in Supplementary Tables S2 and S3.

Based on the above classification matrix, we calculated the accuracy, which is expressed as the ratio of correctly predicted cases (for both NIHL and non-NIHL) to the total number of subjects<sup>40</sup>. Besides the correct classification of noise-exposed workers, we evaluated the precision and recall.

Precision refers to the proportion of predicted NIHL cases that are truly NIHL, while recall is the proportion of actual NIHL cases that were correctly identified. precision and recall are respectively given by

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

We also employed the F-score as a comprehensive performance indicator to assess the effectiveness of each algorithm. A higher F-score, closer to 1, indicates better performance. The F-score is defined as

$$F = \frac{2 \times P \times R}{P + R} = \frac{2 \times TP}{\text{Total number of samples} + TP - TN}$$

$R^2$  refers to the predictive or explanatory power of the independent variable (genotype coding at the SNP loci) to the dependent variable (whether NIHL is present or not). In logistic regression, Nagelkerke  $R^2$  is usually computed as an approximate estimate since the traditional coefficient of determination  $R^2$  is not applicable to binary classification problems<sup>41,42</sup>.

$$R^2_{\text{Nagelkerke}} = 1 - \left( \frac{L_0}{L_{\hat{\beta}}} \right)^{\frac{2}{n}}$$

Among them:  $L_0$  is the log-likelihood value (null model) of the baseline model that contains only intercept terms.  $L_{\hat{\beta}}$  is the log-likelihood value of a regression model with independent variables.  $n$  is the sample size.

The area under the curve (AUC) is defined as the integral of the receiver operating characteristic (ROC) curve, which quantifies the model's ability to distinguish between positive and negative classes across all possible classification thresholds. It is calculated as the area under the plot of the true positive rate (TPR) versus the false positive rate (FPR), where:

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{TN + FP}$$

For ML, all performance indexes selected in this study were calculated using MATLAB 9.0 (R2016a), with the classification threshold for the models set at 0.5.

## Results

### General characteristics of study population

Table 1 presents the general characteristics of the study population. According to the inclusion and exclusion criteria of the study population and combined with the results of pure tone audiometry (PTA), a total of 1138 noise-exposed workers were finally included in this study as the study population, comprising 753 individuals in the case group and 585 in the control group. The case group and the control group were comparable with regard to age, gender, smoking and drinking consumption etc., with no statistically significant differences ( $P > 0.05$ ). However, the differences in years of noise exposure, noise exposure levels and High-frequency hearing threshold were statistically significant ( $P < 0.05$ ). In particular, the high frequency hearing threshold of 33(28,42) in the case group was significantly higher than that of 15(12,19) in the control group, which was approximately 2.2 times higher.

### Basic information and the results of univariate and multivariate logistic regression analysis of the selected SNP loci

Based on the strict screening criteria and quality control process, a total of 88 SNP loci in 40 genes, including VEGFA, FOXM1 and AKT1, were included in this study, of which 72 (81.82%) SNP loci were confirmed to overlap with HapMap3 SNPs, and the remaining 16 SNP loci (18.18%) were not recorded in HapMap3. The basic information and the results of the univariate and multivariate LR analysis of 88 SNP loci are listed in Table 2. Univariate analysis revealed that the genotype distributions of 12 SNP loci (rs7895833, rs177918, rs8102445, rs195432, rs195434, rs2447867, rs2494732, rs2498786, rs1134648, rs2304277, rs10507486, rs2594972) were statistically different between the case and control groups ( $P < 0.05$ ). After adjusting for age, gender, smoking and drinking consumption by multivariate logistics, only 8 SNP loci (rs7895833, rs177918, rs195432, rs195434, rs2447867, rs11334648, rs2304277, rs2594972) had statistically significant differences in genotype distribution between the case and control groups ( $P < 0.05$ ). After the diagnosis of multicollinearity, the tolerance of each

Variable	Cases (n = 753)		Controls (n = 585)		P
	n	%	n	%	
Age (years)					0.127 <sup>a</sup>
M(P25,P75)	41(34,45)		40(34,45)		0.052 <sup>b</sup>
≤ 35	218	28.95	191	32.60	
35–45	347	46.08	273	46.70	
> 45	188	24.97	121	20.70	
Gender					0.376 <sup>a</sup>
Male	708	94.00	543	92.80	
Female	45	6.00	42	7.20	
Smoking					0.056 <sup>a</sup>
Now	421	55.91	289	49.40	
Ever	36	4.78	35	6.00	
Never	296	39.31	261	44.60	
Drinking					0.395 <sup>a</sup>
Now	251	33.30	185	31.60	
Ever	92	12.30	86	14.70	
Never	410	54.40	314	53.70	
Noise exposure time (years)					0.004 <sup>b</sup>
M(P25,P75)	17 (10,25)		16(10,24)		
Noise Exposure levels[dB(A)]					0.001 <sup>b</sup>
M(P25,P75)	87(82,94)		92(84,95)		
Threshold [dB(A)]					< 0.001 <sup>a</sup>
M(P25,P75)	33(28,42)		15(12,19)		< 0.001 <sup>b</sup>
≤ 26	82	10.89	585	100.00	
> 26	671	89.11	0	0.00	

**Table 1.** Basic information for noise-exposed workers. <sup>a</sup>Two-sided  $\chi^2$  test. <sup>b</sup>Two-sided Wilcoxon signed rank sum test.

SNP locus  $> 0.1$ , and the Variance Inflation Factor (VIF)  $< 2$ , indicating that there was no multicollinearity among the SNP loci. After the feature extraction of 88 loci by PNN and GRNN, we selected the top 10 SNP loci (rs12582464, rs2295080, rs195420, rs309184, rs7536272, rs13534, rs41275750, rs7204003, rs12049646, rs706713) with the highest feature importance. However, these SNP loci did not exhibit statistical significance in both univariate and multivariate analyses. Therefore, subsequent model establishment and validation centered on these 8, 10 and 88 loci.

### Model performance comparison

#### Performance comparison between conventional LR and 5 classical ML algorithms

The accuracy, recall(R), precision(P), F-score, R<sup>2</sup> and AUC of conventional LR and five classical ML algorithms during training under different SNP loci datasets are displayed in Table 3, Supplementary Table S4, Supplementary Table S5 and visualized in Fig. 1, Supplementary Figs. S4, S5. Using LR for a pointwise screening of all 88 loci, the model's accuracy, recall, precision, F-score, R<sup>2</sup> and AUC were 62.67%, 80.83%, 64.44%, 0.716, 0.698, and 0.704, respectively, with the overall performance did not reach the expected effect. Furthermore, using these five ML algorithms to model and validate the 8 SNP loci screened by LR, the performance indicators of these models also showed poor or even lower than LR (Supplementary Table S4 and Supplementary Fig. S4), suggesting that the effectiveness and reliability of applying SNP loci screened by LR regression to NIHL prediction remains to be discussed. In light of this, we performed additional feature extraction using PNN and GRNN, and applied LR with five machine learning algorithms to the extracted 10 SNP loci for modelling analysis. The results show that the performance of models constructed using each machine learning algorithms is generally improved compared to models constructed based on 8 SNP loci, and several algorithms outperform LR. Especially, the AUC of GA-RF improves from 0.524 to 0.628, an improvement of about 10% (Supplementary Table S5 and Supplementary Fig. S5). Moreover, the performance indicators of LR were similar to those of the model built using the 8 SNP loci.

Based on the above findings, we applied five ML algorithms to directly pattern identify and model all 88 SNP loci. The results, as shown in Table 3; Fig. 1, demonstrate significant improvements in the performance indicators of several algorithms, with a clear distinction in model performance when compared to LR. Although the accuracy of DT and GBDT has been improved, it is still lower than that of LR at only 60%. However, the accuracy of GA-RF, XGBoost and KNN is higher than that of LR, especially GA-RF, which has been greatly improved, with an accuracy has reached 84.40%, followed by XGBoost and KNN, which are 71.10% and 68.90% respectively, indicating the applicability of ML in NIHL prediction.

Gene	SNP	Allele	Chromosome	Functional consequence	MAF <sup>a</sup>	P for HWE <sup>b</sup>	P value <sup>c</sup>	P value <sup>d</sup>
		A1/A2						
VEGFA	rs10434*	A/G	Chromosome 6:43,785,473	3 prime UTR variant	0.48	1	0.716	-
	rs3024994*	T/C	Chromosome 6:43,775,770	Non coding transcript exon variant	0.18	0.86	0.818	-
	rs3024997*	A/G	Chromosome 6:43,777,370	Non coding transcript exon variant	0.48	1	0.427	-
	rs3025021	T/C	Chromosome 6:43,781,426	Non coding transcript exon variant	0.4	0.08	0.853	-
	rs3025039	T/C	Chromosome 6:43,784,799	3 prime UTR variant	0.34	0.88	0.947	-
	rs699947*	A/C	Chromosome 6:43,768,652	Intergenic variant	0.5	0.75	0.705	-
	rs833068*	A/G	Chromosome 6:43,774,790	Non coding transcript exon variant	0.49	0.99	0.426	-
MTHFR	rs17037396*	T/C	Chromosome 1:11,801,990	Intron variant	0.2	1	0.246	-
	rs1801131*	T/G	Chromosome 1:11,794,419	Missense variant	0.47	0.98	0.419	-
	rs1801133*	A/G	Chromosome 1:11,796,321	Missense variant	0.5	0.77	0.305	-
	rs3737966	T/C	Chromosome 1:11,787,702	3 prime UTR variant	0.49	0.83	0.501	-
	rs4846048*	A/G	Chromosome 1:11,786,195	3 prime UTR variant	0.49	0.38	0.622	-
	rs4846049*	T/G	Chromosome 1:11,790,308	3 prime UTR variant	0.5	0.89	0.448	-
SIRT1	rs10997868*	A/C	Chromosome 10:67,905,202	Intron variant	0.5	0.9	0.424	-
	rs12049646*	T/C	Chromosome 10:67,864,226	Intergenic variant	0.36	0.77	0.973	0.861
	rs12778366*	T/C	Chromosome 10:67,883,321	Intergenic variant	0.19	0.87	0.286	-
	rs1885472*	C/G	Chromosome 10:67,895,054	Intron variant	0.47	0.3	0.543	-
	rs3758391*	T/C	Chromosome 10:67,883,584	Intergenic variant	0.47	0.81	0.531	-
	rs7895833*	A/G	Chromosome 10:67,863,299	Intergenic variant	0.49	0.77	0.076	0.025
UBE2I	rs12925270*	A/C	Chromosome 16:1,325,936	3 prime UTR variant	0.29	0.32	0.623	-
	rs2281226*	A/C	Chromosome 16:1,314,766	Missense variant	0.49	0.7	0.655	-
	rs7204003	A/G	Chromosome 16:1,327,362	Synonymous variant	0.5	0.87	0.467	0.211
	rs761059*	A/G	Chromosome 16:1,324,523	Intron variant	0.47	0.92	0.241	-
TP73-AS1	rs1181865*	T/C	Chromosome 1:3,736,862	Non coding transcript exon variant	0.47	0.78	0.893	-
	rs3737589*	A/G	Chromosome 1:3,746,281	Splice donor region variant	0.49	0.97	0.372	-
	rs7515164*	A/C	Chromosome 1:3,738,183	Non coding transcript exon variant	0.42	0.96	0.537	-
	rs9800	C/G	Chromosome 1:3,736,258	Non coding transcript exon variant	0.49	0.99	0.409	-
SAE1	rs10424953*	A/G	Chromosome 19:47,202,026	Intron variant	0.49	0.98	0.128	-
	rs177918*	T/G	Chromosome 19:47,131,056	Synonymous variant	0.22	0.06	0.008	0.004
	rs309184*	T/G	Chromosome 19:47,133,570	Intron variant	0.49	0.24	0.704	0.711
	rs8102445	T/G	Chromosome 19:47,134,351	Intron variant	0.49	0.24	0.05	0.534
RNF8	rs195420*	C/G	Chromosome 6:37,354,031	5 prime UTR variant	0.48	1	0.191	0.115
	rs195432*	A/C	Chromosome 6:37,390,246	Intron variant	0.5	0.9	0.074	0.013
	rs195434*	T/C	Chromosome 6:37,392,781	3 prime UTR variant	0.48	0.7	0.095	0.029
	rs2284922*	A/G	Chromosome 6:37,381,257	Synonymous variant	0.49	1	0.853	-
ITGA1	rs1531545*	T/C	Chromosome 5:52,897,456	Splice region variant	0.5	0.16	0.595	-
	rs1979398*	A/G	Chromosome 5:52,898,497	Intron variant	0.5	0.13	0.565	-
	rs2432143*	T/C	Chromosome 5:52,835,481	Intron variant	0.22	0.36	0.265	-
	rs2447867*	T/C	Chromosome 5:52,861,540	Missense variant	0.46	1	0.016	0.033
TAB2	rs652921*	A/G	Chromosome 6:149,409,710	Synonymous variant	0.5	0.06	0.161	-
	rs7896*	C/G	Chromosome 6:149,410,340	3 prime UTR variant	0.34	0.94	0.597	-
	rs9485372*	A/G	Chromosome 6:149,287,738	Intron variant	0.48	0.96	0.532	-
PIK3R1	rs1550805*	T/C	Chromosome 5:68,287,979	Intron variant	0.2	0.74	0.774	-
	rs3730089*	A/G	Chromosome 5:68,292,320	Missense variant	0.47	0.93	0.846	-
	rs706713*	T/C	Chromosome 5:68,226,894	Synonymous variant	0.5	0.99	0.359	0.624
AKT1	rs2494732	T/C	Chromosome 14:104,772,855	Intron variant	0.5	0.99	0.044	0.221
	rs2494752*	A/G	Chromosome 14:104,797,271	Regulatory region variant	0.44	0.75	0.401	-
	rs2498786	C/G	Chromosome 14:104,796,031	Regulatory region variant	0.49	0.91	0.014	0.094
FOXM1	rs12582464	T/C	Chromosome 12:2,877,437	Intron variant	0.5	0.37	0.479	0.371
	rs2072360*	T/C	Chromosome 12:2,864,449	Synonymous variant	0.42	0.36	0.214	-
	rs2302257*	C/G	Chromosome 12:2,859,822	Intron variant	0.28	0.92	0.468	-
P21	rs3829964	T/C	Chromosome 6:36,676,721	intron variant	0.5	0.98	0.453	-
	rs762624	A/C	Chromosome 6:36,677,811	Non coding transcript exon variant	0.5	0.4	0.843	-
ATG5	rs510432*	T/C	Chromosome 6:106,326,155	TF binding site	0.49	0.74	0.895	-
	rs803360*	C/G	Chromosome 6:106,318,254	Intron variant	0.49	0.79	0.809	-

Continued

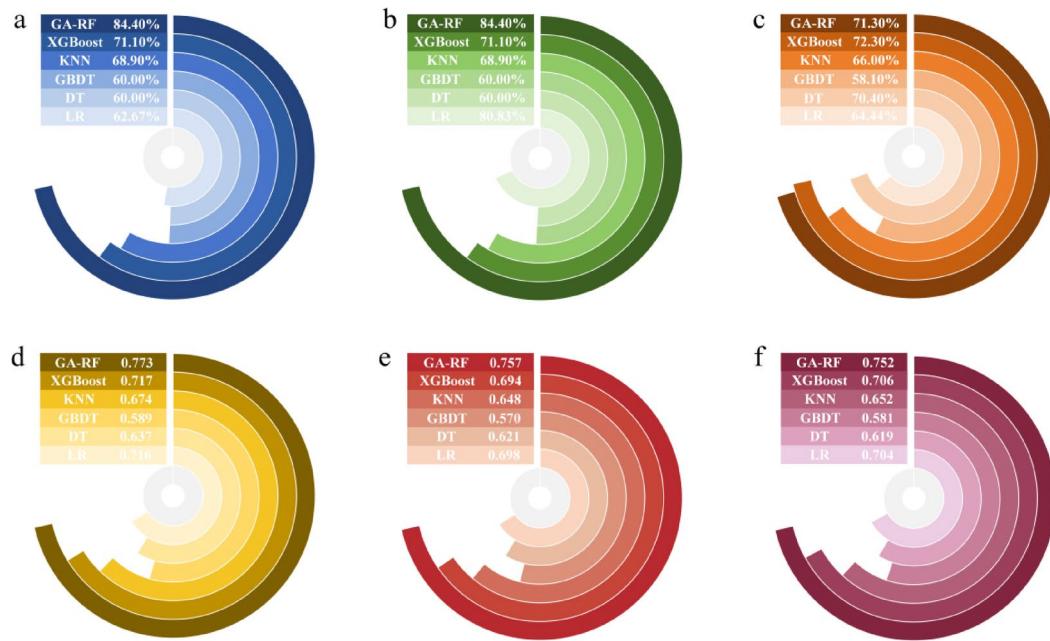
Gene	SNP	Allele	Chromosome	Functional consequence		MAF <sup>a</sup>	P for HWE <sup>b</sup>	P value <sup>c</sup>	P value <sup>d</sup>
		A1/A2							
AKT2	rs2304186*	T/G	Chromosome 19:40,233,814	3 prime UTR variant		0.5	0.39	0.359	-
	rs41275750	C/G	Chromosome 19:40,232,188	3 prime UTR variant		0.37	0.97	0.613	0.997
AIMP1	rs1134648*	C/G	Chromosome 4:106,328,087	Missense variant		0.37	0.06	0.092	0.023
	rs13534	A/G	Chromosome 4:106,348,870	3 prime UTR variant		0.31	0.85	0.687	0.705
OGG1	rs159193*	A/G	Chromosome 20:39,691,100	Intergenic variant		0.49	0.09	0.673	-
	rs2304277*	A/G	Chromosome 3:9,759,396	Non coding transcript exon variant		0.49	0.57	0.099	0.047
MTOR	rs1034528*	C/G	Chromosome 1:11,189,075	Intron variant		0.49	0.65	0.691	-
	rs2295080*	T/G	Chromosome 1:11,262,571	Regulatory region variant		0.42	0.08	0.768	0.976
HDAC2	rs3757016*	T/C	Chromosome 6:113,939,556	3 prime UTR variant		0.5	0.63	0.734	-
	rs6568819*	T/C	Chromosome 6:113,949,764	Intron variant		0.49	0.97	0.7	-
CBX4	rs1285243	T/C	Chromosome 17:79,840,307	Regulatory region variant		0.49	0.65	0.838	-
	rs4889898*	A/C	Chromosome 17:79,824,605	Intron variant		0.5	0.76	0.454	-
FOXO1	rs10507486*	A/G	Chromosome 13:40,612,364	Intron variant		0.32	0.72	0.069	0.056
	rs2701891*	T/C	Chromosome 13:40,550,515	Intron variant		0.5	0.99	0.873	-
RIPK1	rs17548629*	T/C	Chromosome 6:3,114,223	3 prime UTR variant		0.38	0.72	0.921	-
RAP1A	rs6573*	A/C	Chromosome 1:111,712,767	3 prime UTR variant		0.28	0.92	0.922	-
RANBP2	rs1478517	A/G	Chromosome 2:108,894,481	3 prime UTR variant		0.5	0.66	0.2	-
PON1	rs854552*	T/C	Chromosome 7:95,298,612	3 prime UTR variant		0.45	0.07	0.647	-
PIK3R3	rs7536272*	A/G	Chromosome 1:46,177,421	Intron variant		0.48	0.77	0.662	0.838
PIK3CA	rs7651265*	A/G	Chromosome 3:179,175,241	Intron variant		0.24	0.59	0.807	-
NOB1	rs56259873	T/C	Chromosome 16:69,755,293	TF binding site		0.2	0.96	0.314	-
MTPN	rs17168525*	A/G	Chromosome 7:135,928,514	3 prime UTR variant		0.23	0.14	0.572	-
MAPK8	rs9284*	T/G	Chromosome 10:48,435,527	3 prime UTR variant		0.5	0.18	0.939	-
MAPK1	rs13515*	T/C	Chromosome 22:21,761,597	3 prime UTR variant		0.35	0.61	0.445	-
HPGD	rs8752*	T/C	Chromosome 4:174,491,326	3 prime UTR variant		0.49	0.14	0.711	-
GAPDH	rs1136666*	C/G	Chromosome 12:6,534,825	5 prime UTR variant		0.39	0.23	0.151	-
FAS	rs2862833*	A/G	Chromosome 10:89,015,872	3 prime UTR variant		0.5	0.89	0.3	-
CDKN1A	rs1801270*	A/C	Chromosome 6:36,684,194	Missense variant		0.49	0.33	0.185	-
BCL-2	rs1564483*	T/C	Chromosome 18:63,127,421	3 prime UTR variant		0.49	1	0.283	-
ATG7	rs2594972*	A/G	Chromosome 3:11,355,943	Intron variant		0.49	0.59	0.042	0.019
UBAC2	rs2296860*	A/G	Chromosome 13:99,200,499	Non coding transcript exon variant		0.5	0.09	0.486	-
UBA2	rs7258977*	A/G	Chromosome 19:34,468,907	Intron variant		0.43	0.9	0.953	-
STAT3	rs1053004*	A/G	Chromosome 17:42,314,074	3 prime UTR variant		0.49	0.28	0.682	-

**Table 2.** Basic information and the results of univariate and multivariate logistic regression analysis of SNP loci. <sup>a</sup>Data from the NCBI dbSNP database. <sup>b</sup>P value for Hardy-Weinberg test. <sup>c</sup>Two-sided  $\chi^2$  test. <sup>d</sup>Adjusted for age, gender, smoking and drinking status. \*SNP loci are common with HapMap3 SNPs.

Number of SNP\Algorithm	Accuracy	R	P	F score	R <sup>2</sup>	AUC
88\Logistic Regression (LR)	62.67%	80.83%	64.44%	0.716	0.698	0.704
88\Decision Tree (DT)	60.00%	60.00%	70.40%	0.637	0.621	0.619
88\Gradient Boosting Decision Tree (GBDT)	60.00%	60.00%	58.10%	0.589	0.570	0.581
88\K-Nearest Neighbor (KNN)	68.90%	68.90%	66.00%	0.674	0.648	0.652
88\xeXtreme Gradient Boost (XGBoost)	71.10%	71.10%	72.30%	0.717	0.694	0.706
88\Genetic Algorithm-Random Forests (GA-RF)	84.40%	84.40%	71.30%	0.773	0.757	0.752
88\Probabilistic Neural Network (PNN)	78.64%	79.45%	78.44%	0.805	0.797	0.808
88\Generalized Regression Neural Network (GRNN)	85.36%	85.09%	84.60%	0.897	0.862	0.857

**Table 3.** Accuracy, recall, precision, F-scores, R<sup>2</sup> and AUC of LR and seven ML algorithms on 88 SNP loci datasets.

Precision and recall are variables that affect each other, and while a high level of both is a desired ideal situation, in practice it is the high precision that often leads to low recall<sup>43</sup>. Except for GA-RF, the recall rate of the other 4 ML algorithms were all lower than LR, which were 71.10%, 60.00%, 68.90% and 60.00% respectively. In turn, the precision rate of LR was only slightly better than that of GBDT's 58.10%. Notably, GA-RF has



**Fig. 1.** (a-f) Represents the comparison of accuracy, recall, precision, F-score,  $R^2$  and AUC between LR and five classical ML algorithms on 88 SNP loci dataset.

higher precision and recall than LR with 84.40% and 71.30%, respectively, indicating that it has better prediction performance.

The F-score is the harmonic mean of precision and recall. The F-scores of GA-RF and XGBoost are higher than that of LR, at 0.773 and 0.717, respectively, while the remaining three ML algorithms fail to reach the level of LR with F-scores of 0.589, 0.674, and 0.637 respectively. This relationship can be confirmed by comparing the recall and precision scores.

The coefficient of determination  $R^2$ , denoted in this study as Nagelkerke  $R^2$ , is a measure of the goodness of fit of the regression model. Among the five ML algorithms, only GA-RF demonstrated an  $R^2$  value of 0.757, which is higher than the 0.698 achieved by LR, suggesting a better overall fit and explanatory power. The  $R^2$  values of the other four algorithms were all lower than that of LR, indicating their limitations in data fitting.

AUC provides a comprehensive evaluation of the model performance. The AUC of LR and different ML models are as follows: 0.704 for LR, 0.619 for DT, 0.581 for GBDT, 0.652 for KNN, 0.706 for XGBoost, and 0.752 for GA-RF. Among them, the AUC of GA-RF and XGBoost are higher than those of LR, indicating that both of them are discriminative ability is better than logistic regression and can provide higher prediction accuracy, especially in the prediction task of large-scale SNP loci dataset.

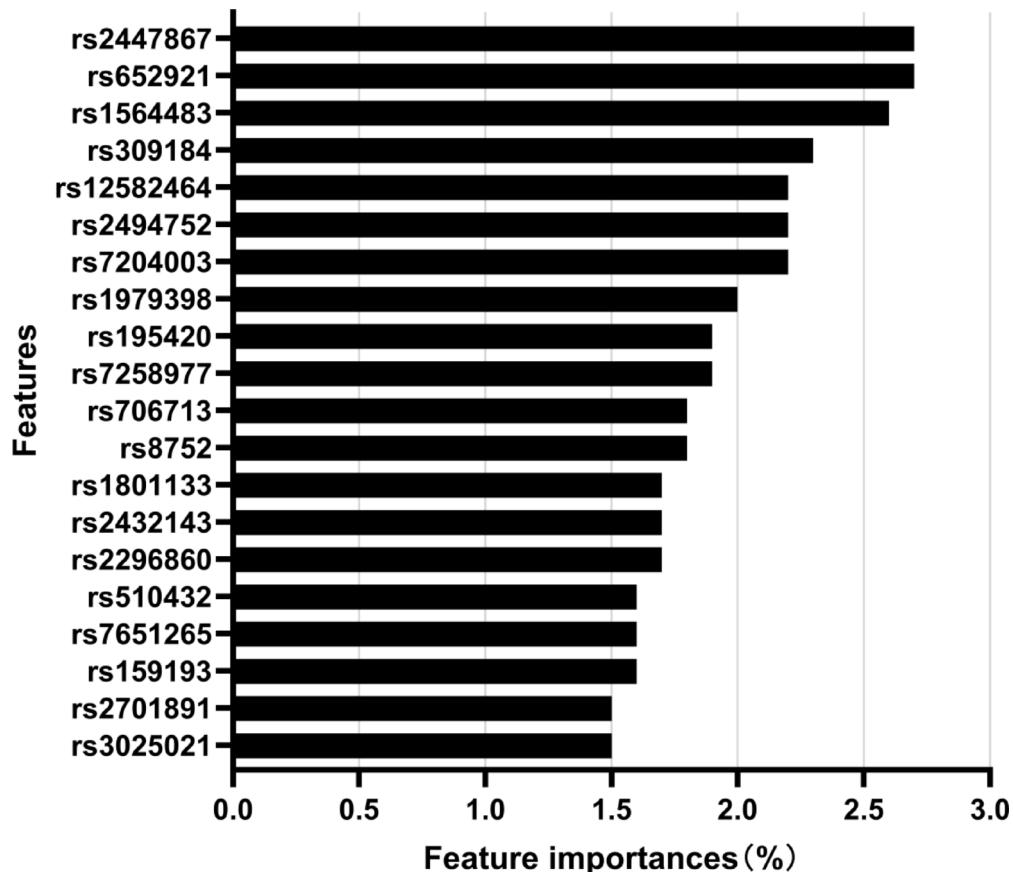
Based on the above analysis, under the 88 SNP loci dataset, GA-RF outperforms LR in all the performance indexes. Therefore, GA-RF is selected for feature filtering of 8, 10 and 88 SNP loci for further analysis. For the 8 SNP loci, rs2304277 was the most important SNP locus with a feature importance of 15.30%, as shown in Supplementary Fig. S6. For the 10 SNP loci, rs309184 was identified as the most significant, exhibiting a feature importance of 13.51%, as illustrated in Supplementary Fig. S7. For the 88 SNP loci, the top 20 SNP loci based on feature importance are outlined in Fig. 2, which are particularly effective for binary classification of NIHL data, among which the most important SNP locus is rs2447867 with a feature importance of 2.70%.

#### Performance comparison between conventional LR and two hyperparameter-optimized ML algorithms

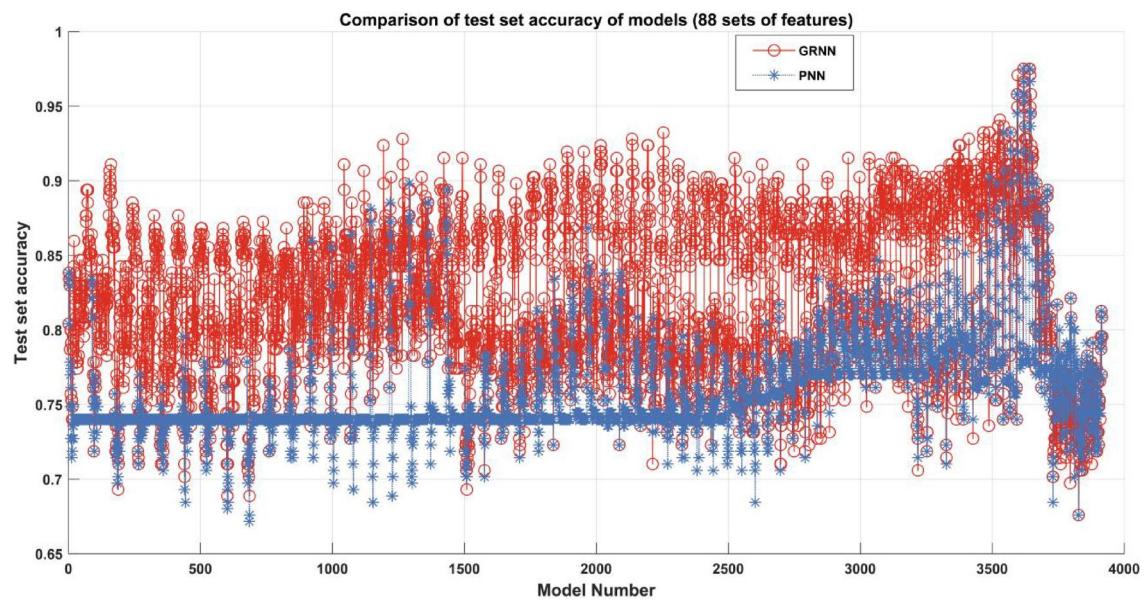
The training process and corresponding accuracy using PNN and GRNN for the 8 SNP loci screened by LR are shown in Supplementary Fig. S8, the highest accuracy of the model at this time was 63.24%, when the three SNP loci rs1134648, rs195434 and rs2304277 were trained together. The lowest accuracy of the model was 51.28%, when rs2594972 was trained alone. At this time, the limited number of model training iterations results in similar outcomes.

Training and testing of 10 SNP loci were conducted applying these two ML models, with the training process and corresponding accuracy presented in Supplementary Fig. S9. The results show that when nine SNP loci, namely rs309184, rs12582464, rs12049646, rs2295080, rs195420, rs7204003, rs7536272, rs13534, rs41275750, are selected for joint training, the model achieves the highest accuracy of 70.98%; When only the rs12049646 locus was used for training, the accuracy of the model was reduced to the lowest at 54.13%.

The process and corresponding accuracy of training and testing all 88 SNP loci using these two ML algorithms are shown in Fig. 3. It is evident from Fig. 3, GRNN has a significant advantage over PNN under the condition of large sample size. This is because one of the advantages of GRNN over PNN is that it is able to give continuous output values in the range of [0, 1], allowing it to describe experimental conditions in a more accurate way. The highest accuracy of the model was 97.50%, occurring when rs10424953, rs1181865, and rs12582464 were trained



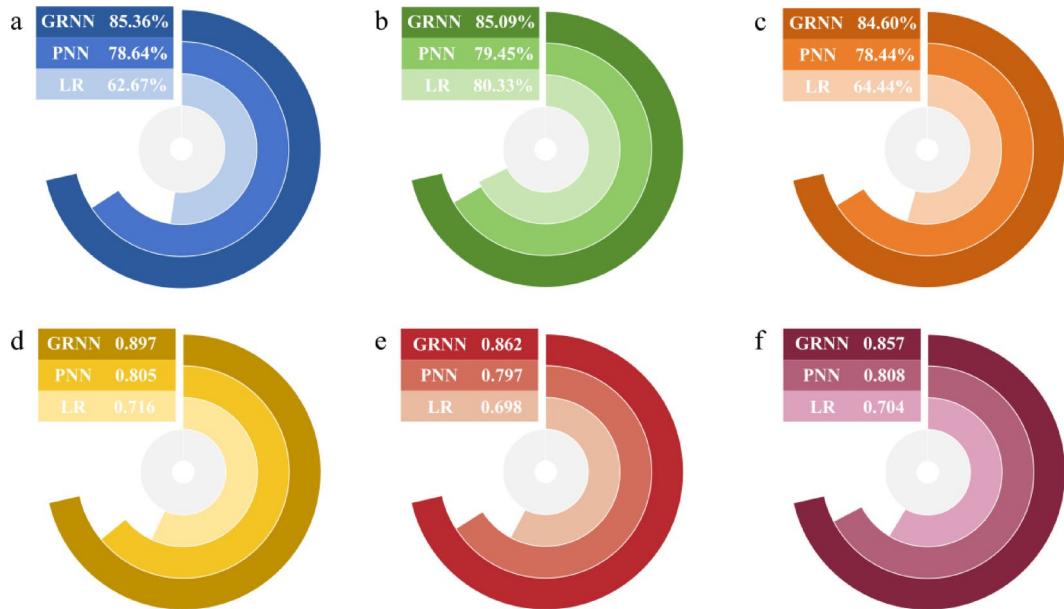
**Fig. 2.** The top 20 SNP loci ranked by feature importance among all 88 SNP in the GA-RF model.



**Fig. 3.** Comparison of GRNN and PNN on the training process of all 88 SNP loci.

together. As for the individual SNP locus, PNN and GRNN reached the consistent conclusion that rs12582464 was the most significant one among the 88 SNP loci, which is significantly associated with the occurrence and development of NIHL. At this point, Table 4 presents the top 20 SNP combinations with the highest accuracy during the PNN and GRNN training processes.

SN	GRNN	PNN
1	rs10424953, rs1181865, rs12582464	rs1181865, rs12582464
2	rs1181865, rs12582464	rs12582464
3	rs1181865, rs12582464, rs12925270	rs12582464, rs12925270
4	rs12582464	rs1181865, rs12582464, rs12925270
5	rs12582464, rs12925270	rs12582464, rs12925270, rs159193
6	rs12582464, rs12925270, rs159193	rs10424953, rs1181865, rs12582464
7	rs854552, rs10424953, rs1181865, rs12582464	rs10424953, rs1181865, rs12582464, rs12925270, rs159193
8	rs1181865, rs12582464, rs12925270, rs159193	rs1181865, rs12582464, rs12925270, rs159193
9	rs1181865, rs12582464, rs12925270, rs159193, rs195420, rs195434, rs2281226	rs1181865, rs12582464, rs12925270, rs159193, rs195420
10	rs12582464, rs12925270, rs159193, rs195420, rs195434	rs10424953, rs1181865, rs12582464, rs12925270
11	rs10424953, rs1181865, rs12582464, rs12925270, rs159193	rs12582464, rs12925270, rs159193, rs195420
12	rs1181865, rs12582464, rs12925270, rs159193, rs195420, rs195434	rs12582464, rs12925270, rs159193, rs195420, rs195434, rs2281226
13	rs12582464, rs12925270, rs159193, rs195420, rs195434, rs2281226	rs1181865, rs12582464, rs12925270, rs159193, rs195420, rs195434
14	rs12582464, rs12925270, rs159193, rs195420	rs12582464, rs12925270, rs159193, rs195420, rs195434
15	rs1181865, rs12582464, rs12925270, rs159193, rs195420	rs6573, rs854552, rs10424953, rs1181865, rs12582464
16	rs854552, rs10424953, rs1181865, rs12582464, rs12925270	rs854552, rs10424953, rs1181865, rs12582464
17	rs6573, rs854552, rs10424953, rs1181865, rs12582464	rs2295080, rs2432143, rs2447867, rs2494732, rs2494752
18	rs1885472, rs1979398, rs2295080	rs1801131, rs1801133, rs1885472, rs1979398, rs2295080
19	rs2295080, rs2432143, rs2447867, rs2494732	rs1885472, rs1979398, rs2295080
20	rs1979398, rs2295080, rs2432143	rs1979398, rs2295080, rs2432143

**Table 4.** The top 20 models with the highest accuracy during the training process of PNN and GRNN.**Fig. 4. (a-f)** Represents the comparison of accuracy, recall, precision, F-score,  $R^2$  and AUC between LR and two hyperparameter-optimized ML algorithms on 88 SNP loci dataset.

The accuracy, recall, precision, F-score,  $R^2$  and AUC of conventional LR and two hyperparameter-optimized ML algorithms during training under different SNP loci datasets are shown in Table 3, Supplementary Tables S4, Supplementary S5 and Fig. 4, Supplementary Figs. S10, S11. Under the 8 SNP loci dataset, the performance indicators of PNN and GRNN still demonstrate poor results and are lower than those of LR itself (Fig. 4). When constructing the model using these two ML models for the 10 SNP loci dataset, the model performance was improved to a certain extent, surpassing that of the model built based on 8 SNP loci selected by LR (Supplementary Fig. S10). However, when these two models are directly applied to the 88 SNP loci dataset, all performance indicators showed significant improvements. Except for the recall rate of PNN being slightly lower than that of LR, the other performance indicators are higher than those of LR (Supplementary Fig. S11).

## Discussion

With the deepening of the global industrialization process, NIHL has gradually revealed its extensive impact and has become one of the major public health problems worldwide. According to the world health organization (WTO) estimates that there are billions of people worldwide due to exposure to hazardous levels of noise inevitably faces a risk of NIHL<sup>44,45</sup>. Due to the lack of specific and sensitive early screening indicators for NIHL, most individuals have already progressed to moderate or severe stages when the disease is confirmed by physical examination, and there is currently no effective treatment, which are mainly based on early prevention<sup>46,47</sup>. Therefore, accurate prediction of noise-exposed workers who are most at risk of developing NIHL is crucial to improving their quality of life and reducing the associated medical and socioeconomic burden. Several studies have shown that ML outperforms LR in analyzing high-dimensional genomic data, SNP loci data, or other biomarker data for disease prediction<sup>48,49</sup>. Given the advanced capabilities and flexibility of various ML algorithms, as well as their potential in complex data analysis, we believe that incorporating ML into NIHL prediction and classification aligns well with the development trend in real-time conditions.

In this study, we used TASSEL and pLINK software to perform quality control on the relevant SNP loci dataset and verified the overlap between these SNP loci and those in the HapMap3 database to ensure the effectiveness and reliability of the SNP loci, enhanced the confidence and biological significance of the findings. On this basis, we systematically analyze and compare the performance of conventional LR and seven ML algorithms in predicting NIHL across different SNP loci datasets for the first time, and cross-validated the SNP loci screening results of multiple models.

Applying LR to all of the 88 SNP loci for pointwise screening, the various performance indicators of the models performed poorly and did not meet the expected standards. In addition, when we used multiple ML algorithms to model and analyse the 8 SNP loci screened by LR, the performance of the models did not improve significantly, and the various performance indicators of each model were generally low, even lower than those of the LR models on the same dataset. This finding prompted us to re-examine the general applicability of conventional LR in NIHL prediction and the effectiveness and reliability of its screened SNP loci in NIHL prediction. We suggest that the SNP loci screened by LR that are statistically associated with the occurrence and progression of NIHL may not be the SNP loci that were significantly associated with the occurrence and progression of NIHL.

Under the 10 SNP loci dataset extracted based on PNN and GRNN, the model performance of each model was improved to different degrees compared with the models built based on the 8 loci, and several ML models outperformed the LR. Nevertheless, the comprehensive performance of each model is still fell short of the expected results, showing a certain gap compared to the ideal level.

Under the all 88 SNP loci dataset, most of the ML algorithms had higher accuracy than LR's 62.67% (except for DT and GBDT). However, it is not comprehensive to evaluate the model solely on the basis of accuracy, considering the imbalance of dataset categories due to the relatively low incidence of NIHL, a model that accurately predicts that all people will not develop NIHL also can achieve a fairly high accuracy, even if it performs poorly in predicting that NIHL will actually occur. In addition, we also expected the model to have a high recall rate because we want to minimize missed diagnoses and ensure that all potential NIHL cases are identified in a timely manner for early intervention, thereby reducing the long-term health risks associated with missed diagnoses. Although this would sacrifice accuracy, it could lead to some non-NIHL cases being incorrectly identified as positive, resulting in a waste of medical resources, such as unnecessary further examinations and treatment for individuals who do not actually have NIHL. In all models, GRNN and GA-RF had higher recall and precision rates than LR's 80.83% and 64.44%. In fact, in NIHL prediction, we would like to find a balance that maintains a high recall to minimize missed diagnoses while also maintaining a relatively high precision to reduce misdiagnoses as much as possible, and the trade-off between the two can be comprehensively evaluated by the F-score. The F-score of LR is 0.716, while those of GRNN, PNN, GA-RF and XGBoost are better than that of LR. Comparing the results of  $R^2$ , it is clear that the goodness-of-fit of the three ML models, GRNN, PNN & GA-RF, is significantly outperforms that of LR. Notably, GRNN and PNN demonstrate a greater ability to capture the complex patterns associated with the occurrence and progression of NIHL. As for AUC, GRNN and PNN also performed excellently, with XGBoost slightly outperforming LR (0.706 vs. 0.704), suggesting that these models models show enhanced comprehensive performance in predicting the occurrence and progression of NIHL.

From the above analyses, it can be argued that multiple ML algorithms outperform or at least equal to conventional LR in NIHL prediction, and its results have good consistency and reproducibility. It is noteworthy that GRNN, PNN and GA-RF exhibit better comprehensive performance across various indicators than conventional LR, which makes them the primary choice for NIHL prediction, and these can also be used as a valuable complementary method to the conventional LR. This result strongly validates our initial idea that when analyzing the association between the NIHL and SNP loci, considering the combined effects of multiple SNP loci and employing ML algorithms can more accurately reveal the underlying associations.

Compared with LR, the SNP loci screened by the ML algorithms with better performance indicators (GRNN, PNN, GA-RF) are more reliable and representative, which is consistent with the results of studies in other fields, they found that ML algorithms have achieved superior predictive effect than LR in identifying SNP associated with disease occurrence and progression<sup>50-52</sup>. By comparing the SNP loci screening results of multiple models, for all 88 SNP loci, rs12582464 located on the FOXM1 gene, rs309184 located on the SAE1 gene and rs2447867 located on the ITGA1 gene, which may be novel pathogenic loci in the NIHL population, significantly improved the accuracy of NIHL prediction, which was also reflected in the screening results of the respective models. Whereas, rs2304277 located on the gene OGG1 is more likely to be associated with the occurrence and progression of NIHL, and it contributes to the prediction accuracy of NIHL across various models.

The performance of various modeling algorithms differs across different studies<sup>53–55</sup>. The performance of LR in NIHL prediction is not as good as desired, which may be closely related to factors such as sample size, model peculiarity and dataset characteristics. For NIHL prediction based on SNP loci data, the model often contains many variables (SNP loci), while LR may be limited by computational power when processing high-dimensional data (such as thousands or tens of thousands of SNP loci), resulting in poor performance and significantly reduced model robustness—a phenomenon referred to as the “curse of dimensionality”<sup>56</sup>; ML, As a representative of modern advanced technology, can efficiently process large-scale datasets and extract critical information from them quickly and efficiently<sup>57,58</sup>. Meanwhile, LR relies on methods such as stepwise regression, forward selection or backward elimination in the process of variable selection, which may not be efficient or accurate enough when handling high-dimensional data or complex feature interactions. while ML usually has built-in feature selection or importance scoring mechanism, which can screen out the most important features for the predicted results, thus ensuring the objectivity of the results<sup>59,60</sup>. In addition, the relationship between SNP loci and disease is often complex and non-linear, and ML algorithm is not constrained by predefined mathematical relationships between dependent and independent variables, allowing for modeling arbitrarily complex nonlinear relationships and being able to take into account interactions between variables<sup>55,61</sup>; Whereas, the operation of LR needs to satisfy the linear assumption, which meaning that it assumes a linear relationship between the independent variables and the log odds ratio, and may fail to capture the complex nonlinear relationships in the data and complex interactions between variables<sup>62</sup>.

This study has certain limitations. First, the study population was limited to noise-exposed workers who underwent occupational health check-ups during a specific time period, whereas NIHL is a gradual developmental process that requires long-term observation to fully reveal its long-term associations with exposure time, noise exposure level, and high-frequency hearing threshold. Second, given the wide variety of machine learning algorithms with its own characteristics and applicable scenarios, this study may not have comprehensively evaluated all potential models in the algorithm selection process. Finally, this study is a retrospective study, which may not fully represent the target population or fulfill the needs of the study design, making it difficult to directly infer causality.

With the popularization and application of Electronic Health Records (EHRs) in healthcare systems, medical research is rapidly becoming data-driven<sup>51</sup>. Applying ML to disease prediction serves as an attractive alternative to conventional LR and can provide a tool for developing high-performance NIHL prediction models. Simultaneously, we also need to flexibly adjust the strategy according to the specific situation of the data and the peculiarity of the model, and continuously explore and optimize the parameter settings and feature engineering of the ML model to further enhance the predictive accuracy and practicability. In the future, we will continue to collect larger population samples and incorporate different risk factors for testing and evaluation to validate the results of this study, striving to make the results more objective and reduce the variability of the results, thereby enabling us to realize the accurate prediction of NIHL.

## Data availability

The datasets generated and/or analysed during the current study are available in the [NCBI] repository, [<https://dataview.ncbi.nlm.nih.gov/object/PRJNA1251789?reviewer=8aejvc3opedag6vatlnl52a1>, PRJNA1251789].

Received: 13 November 2024; Accepted: 24 April 2025

Published online: 02 May 2025

## References

1. Rikhotso, O., Morodi, T. J. & Masekameni, D. M. Occupational health and safety statistics as an Indicator of worker physical health in South African industry. *Int. J. Environ. Res. Public Health* **19**, 1690 (2022).
2. Lavinsky, J. et al. The genetic architecture of noise-induced hearing loss: Evidence for a gene-by-environment interaction. *G3* **6**, 3219–3228 (2016).
3. Natarajan, N., Batts, S. & Stankovic, K. M. Noise-induced hearing loss. *J. Clin. Med.* **12**, 2347 (2023).
4. Basner, M. et al. Auditory and non-auditory effects of noise on health. *Lancet* **383**, 1325–1332 (2014).
5. Wang, B. et al. Association of TagSNP in LncRNA HOTAIR with susceptibility to noise-induced hearing loss in a Chinese population. *Hear. Res.* **347**, 41–46 (2017).
6. Henderson, D., Subramaniam, M. & Boettcher, F. A. Individual susceptibility to noise-induced hearing loss: An old topic revisited. *Ear Hear.* **14**, 152–168 (1993).
7. Konings, A., Van, Laer, L. & Van, Camp, G. Genetic studies on noise-induced hearing loss: A review. *Ear Hear.* **30**, 151–159 (2009).
8. Li, X. et al. PON2 and ATP2B2 gene polymorphisms with noise-induced hearing loss. *J. Thorac. Dis.* **8**, 430–438 (2016).
9. Xu, Y. D., Li, Y. L. & Huang, W. N. Association of HDAC2 gene polymorphisms with susceptibility to Noise-Induced hearing loss. *J. Xinjiang Med. Univ.* **46**, 668–673 (2023).
10. Zheng, X. L. et al. Association of GSTP1 and SOD2 gene polymorphisms with susceptibility to noise-induced hearing loss. *J. Audiol. Speech Pathol.* **31**, 221–225 (2023).
11. Gao, D. F. Study on the Correlation and Mechanism of STAT3 Gene Polymorphisms with Susceptibility To Noise-Induced Hearing Loss. Southeast University (2021).
12. Christodoulou, E. et al. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J. Clin. Epidemiol.* **110**, 12–22 (2019).
13. Oosterhoff, J. H. F. et al. Feasibility of machine learning and logistic regression algorithms to predict outcome in orthopaedic trauma surgery. *J. Bone Jt. Surg. Am.* **104**, 544–551 (2022).
14. Xu, S. et al. Polymorphisms in the FAS gene are associated with susceptibility to noise-induced hearing loss. *Environ. Sci. Pollut. Res. Int.* **28**, 21754–21765 (2021).
15. Wan, L., Wang, B., Zhang, J., Zhu, B. & Pu, Y. Associations of genetic variation in glyceraldehyde 3-Phosphate dehydrogenase gene with Noise-Induced hearing loss in a Chinese population: A Case-Control study. *Int. J. Environ. Res. Public Health.* **17**, 2899 (2020).
16. Zhao, X. et al. Predicting renal function recovery and short-term reversibility among acute kidney injury patients in the ICU: Comparison of machine learning methods and conventional regression. *Ren. Fail.* **44**, 1326–1337 (2022).

17. Ge, T., Chen, C. Y., Ni, Y., Feng, Y. A. & Smoller, J. W. Polygenic prediction via bayesian regression and continuous shrinkage priors. *Nat. Commun.* **10**, 1776 (2019).
18. Cui, J. Hypertension Risk Prediction Based on Machine Learning Methods and Regional Single Nucleotide Polymorphisms. Xidian University (2022).
19. Beam, A. L. & Kohane, I. S. Big data and machine learning in health care. *JAMA* **319**, 1317–1318 (2018).
20. Chen, J. H. & Asch, S. M. Machine learning and prediction in Medicine—Beyond the peak of inflated expectations. *N. Engl. J. Med.* **376**, 2507–2509 (2017).
21. Rajkomar, A. et al. Scalable and accurate deep learning with electronic health records. *NPJ Digit. Med.* **1**, 18 (2018).
22. Varghese, S. A. et al. Identification of diagnostic urinary biomarkers for acute kidney injury. *J. Investig. Med.* **58**, 612–620 (2010).
23. Zhang, F. Y., Yu, Y., Zhao, Y., Yang, K. & Hu, X. H. Applications of artificial neural networks in clinical medicine. *Beijing Biomed. Eng.* **35**, 422–428 (2016).
24. Bradbury, P. J. et al. TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633–2635 (2007).
25. Purcell, S. et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
26. Duan, S., Zhang, W., Cox, N. J. & Dolan, M. E. FstSNP-HapMap3: A database of SNPs with high population differentiation for HapMap3. *Bioinformation* **3**, 139–141 (2008).
27. Chen, K. et al. A novel three-round multiplex PCR for SNP genotyping with next generation sequencing, analytical and bioanalytical chemistry. *Anal. Bioanal. Chem.* **408**, 4371–4377 (2016).
28. Hosmer, J. D. W., Lemeshow, S. & Sturdivant, R. X. *Applied Logistic Regression* (Wiley, 2013).
29. Song, Y. Y. & Lu, Y. Decision tree methods: Applications for classification and prediction. *Shanghai Arch. Psychiatry* **27**, 130–135 (2015).
30. Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* 1189–1232 (2001).
31. Mei, J. & Chen, J. M. Application of K-Nearest neighbor algorithm in diabetes prediction. *Comput. Inform. Technol.* **32**, 7–9 (2024).
32. Huang, H. R. Research on cardiovascular disease prediction based on machine learning. Master's thesis, Hubei University (2024).
33. Elyan, E. & Gaber, M. M. A genetic algorithm approach to optimising random forests applied to class engineered data. *Inf. Sci.* **384**, 220–234 (2017).
34. Shi, J. Q. & Zhang, J. H. Load forecasting based on Multi-model by stacking ensemble learning. *Chin. Soc. Electr. Eng.* **39**, 4032–4042 (2019).
35. Khelil, C. K. M., Amrouche, B., Kara, K. & Chouder, A. The impact of the ANN's choice on PV systems diagnosis quality. *Energy Conv. Manag.* **240**, 114278 (2021).
36. Ahmadipour, M., Hizam, H., Othman, M. L., Radzi, M. A. M. & Murthy, A. S. Islanding detection technique using slantlet transform and ridgelet probabilistic neural network in grid-connected photovoltaic system. *Appl. Energy* **231**, 645–659 (2018).
37. Fan, H. G., Pei, J. H. & Zhao, Y. An optimized probabilistic neural network with unit Hyperspherical crown mapping and adaptive kernel coverage. *Neurocomputing* **373**, 24–34 (2020).
38. Wang, L., Lee, T. J., Bawdeniek, J. & Eckstein, L. A data-driven approach towards the full anthropometric measurements prediction via generalized regression neural networks. *Appl. Soft Comput.* **109**, 107551 (2021).
39. Yuan, Q. Q. et al. Estimating surface soil moisture from satellite observations using a generalized regression neural network trained on sparse ground-based measurements in the continental U.S. *J. Hydrol.* **580**, 124351 (2020).
40. Miotto, R., Li, L., Kidd, B. A. & Dudley, J. T. Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Sci. Rep.* **6**, 26094 (2016).
41. Nagelkerke, N. J. D. A note on a general definition of the coefficient of determination. *Biometrika* **78**, 691–692 (1991).
42. Momin, M. M., Lee, S., Wray, N. R. & Lee, S. H. Significance tests for  $R^2$  of out-of-sample prediction using polygenic scores. *Am. J. Hum. Genet.* **110**, 349–358 (2023).
43. Zeng, J., Jiang, H. & Yang, H. *Study on Systems Biology and Clinical Medicine* (Science Press, 2017).
44. World Health Organization. Hearing loss due to recreational exposure to loud sounds: A review (2015).
45. Yang, P., Xie, H., Li, Y. & Jin, K. The effect of noise exposure on high-frequency hearing loss among Chinese workers: A meta-analysis. *Healthcare* **11**, 1079 (2023).
46. Yu, HuanXin, Y. H. Research progress of noise induced hearing loss (2014).
47. Wan, L. et al. Association between UBAC2 gene polymorphism and the risk of noise-induced hearing loss: A cross-sectional study. *Environ. Sci. Pollut. Res. Int.* **29**, 32947–32958 (2022).
48. Libbrecht, M. W. & Noble, W. S. Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* **16**, 321–332 (2015).
49. Larranaga, P. et al. Machine learning in bioinformatics. *Brief. Bioinform.* **7**, 86–112 (2006).
50. Huang, S. et al. Applications of support vector machine (SVM) learning in cancer genomics. *Cancer Genomics Proteom.* **15**, 41–51 (2018).
51. Goldstein, B. A., Hubbard, A. E., Cutler, A. & Barcellos, L. F. An application of random forests to a genome-wide association dataset: Methodological considerations & new findings. *BMC Genet.* **11**, 1–13 (2010).
52. Couronné, R., Probst, P. & Boulesteix, A. L. Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinform.* **19**, 1–14 (2018).
53. Bishara, A. et al. Postoperative delirium prediction using machine learning models and preoperative electronic health record data. *BMC Anesthesiol.* **22**, 1–12 (2022).
54. Racine, A. M. et al. Machine learning to develop and internally validate a predictive model for post-operative delirium in a prospective, observational clinical cohort study of older surgical patients. *J. Gen. Intern. Med.* **36**, 265–273 (2021).
55. Song, X., Liu, X., Liu, F. & Wang, C. Comparison of machine learning and logistic regression models in predicting acute kidney injury: A systematic review and meta-analysis. *Int. J. Med. Inf.* **151**, 104484 (2021).
56. Feng, J. Z. et al. Comparison between logistic regression and machine learning algorithms on survival prediction of traumatic brain injuries. *J. Crit. Care* **54**, 110–116 (2019).
57. Zhou, Z. H. *Machine Learning* (Springer, 2021).
58. Wang, S. & Summers, R. M. Machine learning and radiology. *Med. Image Anal.* **16**, 933–951 (2012).
59. Fan, J. & Lv, J. A selective overview of variable selection in high dimensional feature space. *Stat. Sin.* **20**, 101 (2010).
60. Lu, H. Y., Zhang, M., Liu, Y. Q. & Ma, S. P. Feature importance analysis of convolutional neural networks and an enhanced feature selection model. *J. Softw.* **28**, 2879–2890 (2017).
61. Deo, R. C. Machine learning in medicine. *Circulation* **132**, 1920–1930 (2015).
62. Tu, J. V. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *J. Clin. Epidemiol.* **49**, 1225–1231 (1996).

## Acknowledgements

We acknowledge all participants for their contributions to this study. This study was supported by Natural Science Foundation of Jiangsu Province (BK20230742), 2024 Annual Project of the National Health Commission (NHC) Capacity Building and Continuing Education Center (GWJJ2024100202), Scientific Research Project of Jiangsu Health Committee (M2022083), Jiangsu Provincial Medical Key Discipline (ZDXK202249).

## Author contributions

J.L. was responsible for data organization and analysis, and article writing. X.H.L. contributed to data analysis as well as article writing. J.L. and X.H.L. contributed equally to this article as co-first authors. B.S.W. provided article ideas and gave data analysis support. Y.X.W., H.D.Z., and L.H. reviewed and corrected the article. B.L.Z. performed the final review and gave fund support.

## Declarations

### Competing interests

The authors declare no competing interests.

### Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-025-00050-1>.

**Correspondence** and requests for materials should be addressed to B.Z. or B.W.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

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

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

© The Author(s) 2025