BMJ Open

Construction of a Risk Prediction Model for Occupational Noise Induced Hearing Loss Based on Routine Blood and Biochemical Indicators

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2024-097249
Article Type:	Original research
Date Submitted by the Author:	28-Nov-2024
Complete List of Authors:	Wang, Dianpeng; Shenzhen Prevention and Treatment Center for Occupational Diseases, Li, Caiping; Southern Medical University, Department of Toxicology shi, liuwei; Jilin University Chen, Linlin; Southern Medical University Lin, Dafeng; Shenzhen Prevention and Treatment Center for Occupational Diseases yang, xiangli; Shenzhen Prevention and Treatment Center for Occupational Diseases Li, peimao; Shenzhen Prevention and Treatment Center for Occupational Diseases Zhang, Wen; Shenzhen Prevention and Treatment Center for Occupational Diseases Feng, wenting; Shenzhen Prevention and Treatment Center for Occupational Diseases Guo, Yan; Shenzhen Prevention and Treatment Center for Occupational Diseases Zhou, Liang; Southern Medical University, Department of Toxicology Zhang, naixing; Shenzhen Prevention and Treatment Center for Occupational Diseases
Keywords:	Machine Learning, Audiology < OTOLARYNGOLOGY, Blood bank & transfusion medicine < HAEMATOLOGY, Risk Factors

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Construction of a Risk Prediction Model for Occupational Noise Induced Hearing Loss Based on Routine Blood and Biochemical Indicators

Caiping Li, MS[1], Liuwei Shi, MS[3], Linlin Chen, MS[1], Dafeng Lin, PhD[2], Xiangli Yang, BS[2], Peimao Li, MS[2], Wen Zhang, MS[2], Wenting Feng, MS[2], Yan Guo, PhD[2], Liang Zhou, PhD[1*], Naixing Zhang, PhD[2*], Dianpeng Wang, MS[1,2*]

[1]Department of Toxicology, School of Public Health, Southern Medical University, Guangzhou 510515, China

[2]Medical laboratory, Shenzhen Prevention and Treatment Center for Occupational Diseases, Shenzhen 518020, China

[3]School of Public Health, Jilin University, Changchun 130012, China

[*]**Correspondence:** Dianpeng Wang and Naixing Zhang, Medical laboratory, Shenzhen Prevention and Treatment Center for Occupational Diseases, 2019 Buxin Rd., Luohu District, Shenzhen 518020, China. E-mail address: szpcr@126.com and zhanghealth@126.com

Liang Zhou, Department of Toxicology, School of Public Health, Southern Medical University, No.1023-1063 Shatai South Rd., Baiyun District, Guangzhou 510515, China. E-mail address: zhzliang@smu.edu.cn

C L and L S contributed equally to this work.

word count: 4360

Keywords Noise-induced hearing loss, risk prediction, machine learning, blood routine indicators, biochemical indicators.

ABSTRACT

Objectives Occupational noise-induced hearing loss (ONIHL) represents a prevalent occupational health condition, traditionally necessitating multiple pure-tone audiometry assessments. We have developed and validated a machine learning model leveraging routine hematological and biochemical parameters, thereby offering novel insights into the risk prediction of ONIHL.

Design, setting and participants This study analyzed data from 3,311 noise-exposed workers in Shenzhen, including 163 ONIHL cases, with the dataset divided into D1 (2,868 samples, 107 ONIHL cases) and D2 (443 samples, 56 ONIHL cases). The inclusion criteria were formulated based on the GBZ49-2014 Diagnosis of Occupational Noise-Induced Hearing Loss. Model training was performed using D1, and model validation was conducted using D2. Routine blood and biochemical indicators were extracted from the case data, and a range of machine learning algorithms including extreme gradient boosting (XGBoost) were employed to construct predictive models. The model underwent refinement to identify the most representative variables, and Decision Curve Analysis (DCA) was conducted to evaluate the net benefit of the model across various threshold levels.

Primary outcome measures Model creation dataset and validation datasets: ONIHL.

Results The prediction model, developed using XGBoost, demonstrated exceptional performance, achieving an area under the curve (AUC) of 0.934, a sensitivity of 0.909,

and a specificity of 0.875 on the validation dataset. On the test dataset, the model achieved an AUC of 0.886. After implementing feature selection, the model was refined to include only 16 features, while maintaining strong performance on a newly acquired independent dataset, with an AUC of 0.852, a balanced accuracy of 0.782, a sensitivity of 0.814, and a specificity of 0.750.The analysis of feature importance revealed that serum albumin (ALB), coefficient of variation in red cell distribution width (RDW-CV), lymphocyte percentage (LYMPHP), monocyte count (MOC), and standard deviation in red cell distribution width (RDW-SD) are critical factors for risk stratification in patients with ONIHL.

Conclusion The analysis of feature importance identified ALB, the coefficient of variation in RDW-CV, LYMPHP, MOC, and the standard deviation in RDW-SD as pivotal factors for risk stratification in patients with ONIHL. The machine learning model, utilizing XGBoost, effectively distinguishes ONIHL patients among individuals exposed to noise, thereby facilitating early diagnosis and intervention.

Article Summary

Strengths and limitations of this study

The model predicts ONIHL using routine blood and biochemical indicators, eliminating the need for audiometric tests or direct noise exposure data.

Simplifies the diagnostic process, reducing time, costs, and manpower requirements.

Provides an accessible and efficient alternative for early screening and prevention of noise-induced hearing loss.

The study is limited to the Shenzhen population, and the model's generalizability to other groups and settings remains uncertain.

The positive-to-negative sample ratio exceeds 1:20, mirroring real-world conditions but limiting predictive accuracy.

INTRODUCTION

ONIHL is characterized as a progressive sensorineural hearing impairment predominantly attributed to damage of the hair cells within the inner ear, consequent to prolonged exposure to high-intensity noise environments[1]. As reported by the World Health Organization (WHO), approximately 10% of the global workforce is impacted by elevated noise levels, with occupational noise exposure accounting for 7% to 21% of hearing loss among workers[2]. A national occupational research agenda says that ONIHL has the highest prevalence of occupational diseases in the United States[3]. About 22 million U.S. workers are currently exposed to hazardous occupational noise[4]. This incidence is notably higher in developing countries[5]. As the largest developing nation, China has witnessed an increasing trend in the incidence of occupational ONIHL in recent years. The prevalence of ONIHL has been reported to be over 20% among noise-exposed workers in China[6]. Such hearing loss can result in communication challenges, social isolation, loneliness, and depression, thereby adversely impacting patients' quality of life and leading to indirect economic losses for society[7].

Currently, pure-tone audiometry is regarded as the gold standard for diagnosing ONIHL. However, its reliance on costly audiological equipment and the necessity for highly trained professionals restrict its practicality for large-scale ONIHL screening among noise-exposed occupational groups[8]. Consequently, there is a pressing need to develop a practical and user-friendly screening tool specifically designed for ONIHL

patients to prevent the advancement to clinically significant ONIHL. Numerous instances of ONIHL are characterized by an initial deterioration in high-frequency hearing, which gradually progresses to impairments in low-frequency or speech frequency hearing[9]. The early identification of individuals at high risk is essential for effective prevention and intervention strategies.

Consequently, the development of predictive models to screen high-risk populations for further evaluation represents a viable alternative approach. The growing volume of data has facilitated the application of machine learning techniques in the context of ONIHL. At present, a variety of methodologies employing either traditional statistical analysis or machine learning techniques are utilized to predict the risk of ONIHL. These methodologies frequently necessitate substantial human resources and present challenges in manual definition[10]. The integration of machine learning (ML) within the field of audiology has demonstrated potential, particularly in its capacity to effectively analyze nonlinear relationships within data, such as forecasting hearing thresholds for individuals exposed to specific risk factors[11]. Abdollahi[12] constructed eight ML models to forecast sensorineural hearing loss following radiotherapy and chemotherapy, with five of these models demonstrating accuracy and precision exceeding 70%. Comparable levels of accuracy have been reported in other investigations employing ML models to predict sudden sensorineural hearing loss (SSNHL) and ototoxic hearing loss[13,14]. Additionally, various studies have documented accuracy rates between 0.64 and 0.99 when utilizing diverse ML

algorithms and input parameters to predict ONIHL risk factors[15–19]. Among the diverse array of machine learning techniques, SVM models, RF models, and XGBoost models have demonstrated superior performance in classification tasks[9].

Established risk factors for ONIHL encompass age, medical history (including conditions such as hypertension and diabetes), history of noise exposure, and behavioral factors such as smoking and physical activity[20–22]. Furthermore, several biomarkers associated with inflammation, including elevated levels of white blood cells (WBC), neutrophils (NE), monocytes (MO), and lymphocytes (LY), alongside metabolic parameters such as low-density lipoprotein (LDL) and high-density lipoprotein (HDL), are recognized as risk indicators for hearing loss[23]. The chronic alterations in the inflammatory state that occur with aging, a phenomenon known as inflammaging, may contribute to or expedite long-term auditory system damage[24]. Red cell distribution width (RDW), a parameter traditionally utilized for the classification of anemia, has recently been identified as being associated with inflammation and microcirculatory disorders[25]. HDL and LDL have been reported to influence blood supply, thereby potentially affecting sudden sensorineural hearing loss[23]. While numerous studies have explored the relationship between hearing loss and various blood inflammatory and metabolic parameters, there is a paucity of research employing these parameters to predict ONIHL.

It is noteworthy that individuals exposed to occupational noise are subject to annual medical evaluations, which routinely include blood tests comprising both

standard and biochemical analyses[14]. Physicians often extract limited information from these routine blood test results. In light of this, our study seeks to comprehensively leverage routine hematological and biochemical indicators, in conjunction with machine learning methodologies, to construct a risk prediction model for ONIHL. The objective is to facilitate early detection and intervention for ONIHL utilizing data from standard medical examinations.

METHODS

Data Collection and Processing

The medical examination data was obtained from the Shenzhen Prevention and Treatment Center for Occupational Diseases from January 2023 to July 2024. The data was divided into two parts in chronological order, named D1 and D2. The first step involved data cleaning, removing samples with erroneous or abnormal values. The inclusion criteria were formulated based on the GBZ49-2014 *Diagnosis of Occupational Noise-Induced Hearing Loss*: (1) Noise exposure duration ≥ 3 years; (2) Bilateral high-frequency (3000 Hz, 4000 Hz, 6000 Hz) average hearing threshold ≥ 40 dB. Exclusion criteria included pseudohypacusis, exaggerated hearing impairment, drug-induced hearing loss, traumatic hearing loss, infectious hearing loss, hereditary hearing loss, Ménière's disease, sudden deafness, acoustic neuroma, and auditory neuropathy. We divided the samples into two groups: the occupational noise-induced hearing loss group and the noise-exposed normal hearing group. After preprocessing, a

total of 3,311 samples were retained, with D1 and D2 consisting of 2,868 and 443 samples, respectively. Among them, there were 107 and 56 cases of noise-induced hearing loss, representing the positive samples. We then applied random sampling to split D1 into a training set and a test set at a 7:3 ratio. D2 was used as an independent test set.

All datasets included the following variables: sex, age, total protein (TP), albumin (ALB), glucose (GLU), cholesterol (CHO), triglycerides (TG), high-density lipoprotein (HDL), low-density lipoprotein (LDL), total bilirubin (TBIL), direct bilirubin (DBIL), indirect bilirubin (IBIL), alanine aminotransferase (ALT), aspartate aminotransferase (AST), blood urea nitrogen (BUN), serum creatinine (Scr), uric acid (UA), globulin (GLB), hemoglobin (Hb), red blood cell count (RBC), hematocrit (HCT), mean corpuscular volume (MCV), mean corpuscular hemoglobin (MCH), mean corpuscular hemoglobin concentration (MCHC), white blood cell count (WBC), eosinophil count (EOC), basophil count (BAC), lymphocyte count (LYMPHC), monocyte count (MOC), platelet count (PLT), neutrophil count (GRANC), eosinophil percentage (EOP), basophil percentage (BAP), red cell distribution width (CV), mean platelet volume (MPV), platelet distribution width (PDW), plateletcrit (PCT), neutrophil percentage (GRANP), lymphocyte percentage (LYMPHP), monocyte percentage (MOP), red cell distribution width (SD), platelet-to-HDL ratio (PLT/HDL), glucose-to-HDL ratio (GLU/HDL), platelet-to-lymphocyte ratio (PLT/LYMPHC), albumin-to-globulin ratio (A/G), neutrophil-to-lymphocyte ratio (S/L), triglyceride-glucose index (TyG), and

estimated glomerular filtration rate (eGFR) (The calculation formulas for TyG and eGFR are detailed in Additional file 1).

In light of the pronounced class imbalance present across all datasets, we employed oversampling of the positive instances within the training set utilizing the `ovun.sample()` function from the ROSE package. This function randomly replicates samples from the minority class, thereby equalizing the number of positive and negative samples in the training set and achieving a balanced class distribution[26]. This approach effectively increases the sample size of the minority class, mitigating the effects of class imbalance during model training. All datasets underwent Z-score normalization, utilizing the mean and standard deviation derived from the training set data.

Framework

Employing occupational health examination data, we introduce an integrated framework for the identification of patients with noise-induced hearing loss, as illustrated in Figure 1. Initially, we preprocessed two datasets, designated as D1 and D2. Dataset D1 was partitioned into training and validation subsets in a 7:3 ratio, while dataset D2 served as an independent test set for the evaluation of the final model. Due to the class imbalance present in the dataset, we employed an oversampling technique on the training set. Subsequently, we utilized a comprehensive array of machine learning algorithms, including XGBoost, Logistic Regression (LR), Random Forest

(RF), Support Vector Machine (SVM), and K-Nearest Neighbor (KNN), to construct predictive models. We then applied feature selection methods to the most optimal predictor among the five to enhance the tool's feasibility. The performance of the refined model was evaluated using an independent test set. we conducted a feature importance analysis to identify variables correlated with the incidence of noise-induced hearing loss. Additionally, we optimized the model to select the most representative variables and employed DCA to evaluate the net benefit of the model across various threshold levels.

Model Construction

In order to construct predictive models, we employed five machine learning algorithms: LR, RF, SVM, KNN, and XGBoost. LR is a form of linear regression that utilizes the Sigmoid function to convert outputs into probabilities for classification purposes[27]. RF comprises an ensemble of independently trained decision trees, with the ultimate prediction being derived through a voting mechanism among these trees, thereby mitigating the risk of overfitting[28]. SVM algorithm classifies samples by identifying an optimal hyperplane within the feature space, and it is capable of managing nonlinearly separable data[29]. KNN algorithm, an instance-based learning method, classifies samples according to the proximity of their k nearest neighbors, making it particularly suitable for small datasets and straightforward to implement[30]. XGBoost is an ensemble method based on decision trees that enhances model performance through a gradient boosting framework. It constructs decision trees in an iterative

manner to minimize model error, demonstrating particular efficacy in handling large-scale, high-dimensional datasets due to its robust generalization capabilities and computational efficiency[31]. In this study, models including LR, RF, SVM, KNN, and XGBoost were implemented using the R programming language. The 'caret' package facilitated model training and evaluation, while the 'xgboost' package was specifically employed for the XGBoost model. For the KNN model, various values of k were tested, with the optimal performance observed at k=5.

Model Evaluation

To evaluate model performance, considering the class imbalance in the validation and test sets, we used the following metrics to comprehensively assess model performance: sensitivity, specificity, balanced accuracy, AUC, PR-AUC, F1-score, and precision. These metrics are defined as follows:

$$\text{Sensitivity} = \text{Recall} = \text{TRP} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \text{TNR} = \frac{TN}{TN + FP}$$

$$\text{Balanced Accuracy} = \frac{TPR + TNR}{2}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{F1 score} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

The performance of all models was assessed using the 'pROC' package in R to calculate AUC and PR-AUC values. We also used five-fold cross-validation

(implemented with the 'caret' package) to prevent overfitting. Furthermore, the XGBoost model was fine-tuned by adjusting hyperparameters such as the learning rate, tree depth, and the number of trees.

TP, that is, true positive, is the number of cases of noise-induced hearing loss. *FP*, false positive, denotes the number of normal subjects incorrectly predicted as having ONIHL. *TN*, True Negative, indicates the number of healthy subjects correctly classified as normal. *FN*, False Negative, refers to the number of cases with ONIHL incorrectly classified as normal. And all above metrics range from 0 to 1.

Feature Selection and Feature Importance Analysis

Despite the relatively high performance of the prediction model utilizing 48 features, there remains the possibility of redundant information or noise features that could adversely affect the decision-making process. To enhance the effective utilization of features and streamline the model, we employed a combination of manual selection, Principal Component Analysis (PCA), and Maximum Relevance Minimum Redundancy (mRMR) methods to extract essential features for the final model[32,33]. In the manual selection process, we initially identified features that exhibited significant differences between positive and negative samples. To improve the stability of the predictive model, we eliminated features that contributed to significant collinearity[32]. As a result, 16 features were retained. To ensure consistency, the number of feature subsets was also fixed at 16 during the application of PCA and mRMR analysis.

Furthermore, feature selection was conducted on the training set to mitigate the risk of overfitting. The analysis of feature importance facilitates the interpretation of the predictive model and aids in identifying the features most closely associated with ONIHL. In this context, each feature's significance is quantified by the corresponding weight coefficients within the XGBoost model.

RESULTS

We initially gathered occupational health examination data from the Shenzhen Occupational Disease Prevention and Control Institute for the period spanning 2023 to 2024, resulting in 3,311 noise-exposed samples that met our inclusion criteria. Of these, 107 participants were diagnosed with ONIHL. Table 1 provides a detailed description of the characteristics of both noise-exposed individuals and ONIHL patients. The five most prominent features exhibiting significant differences between the ONIHL and non-ONIHL samples include RDW-CV, PDW, HCT, LYMPHP and RDW-SD (Additional file 1: Figure S1).

Table 1 Statistical Characteristics of Noise-Exposed Hearing Normal Individuals and ONIHL Patients

Characteristics	**Control**	**Case**	***p***
N	2761	107	
Sex:			<0.001**
female	25 (0.91%)	21 (19.6%)	
male	2736 (99.1%)	86 (80.4%)	
Age, year	38.5 ±7.65	43.5 ± 7.17	<0.001**

TP, g/L	72.7 ± 3.91	68.1 ± 4.69	<0.001**
ALB, g/L	46.7 ± 2.59	42.7 ± 2.95	<0.001**
GLU, mmol/L	5.23 ± 0.63	5.42 ± 0.86	0.023*
CHO, mmol/L	4.89 ± 0.86	4.90 ± 1.02	0.930
TG, mmol/L	1.64 ± 1.21	1.97 ± 1.28	0.011*
LDL, mmol/L	1.32 ± 0.26	1.31 ± 0.31	0.717
HDL, mmol/L	3.03 ± 0.62	3.02 ± 0.77	0.833
TBIL, μmol/L	16.5 ± 6.26	15.7 ± 6.02	0.188
DBIL, μmol/L	3.02 ± 1.24	2.80 ± 1.16	0.061
IBIL, μmol/L	13.4 ± 5.20	12.6 ± 4.90	0.098
ALT, U/L	29.5 ± 22.1	27.2 ± 21.8	0.294
AST, U/L	24.7 ± 9.79	27.5 ± 42.3	0.503
BUN, mmol/L	5.04 ± 1.18	5.06 ± 5.07	0.966
Scr, μmol/L	85.1 ± 11.2	76.7 ± 16.9	<0.001**
UA, μmol/L	401 ± 79.2	482 ± 893	0.351
GLB, g/L	26.0 ± 3.36	25.4 ± 3.37	0.086
Hb, g/L	154 ± 10.1	144 ± 17.9	<0.001**
RBC, $\times 10^{12}$/L	5.02 ± 0.37	4.84 ± 0.49	<0.001**
HCT, L/L	0.45 ± 0.03	0.42 ± 0.04	<0.001**
MCV, fL	89.6 ± 4.77	87.6 ± 8.47	0.015*
MCH, pg	30.8 ± 1.86	29.9 ± 3.47	0.007**
MCHC, g/L	344 ± 6.82	341 ± 12.3	0.013*
WBC, $\times 10^9$/L	6.09 ± 1.53	6.81 ± 1.61	<0.001**
EOC, $\times 10^9$/L	0.16 ± 0.14	0.21 ± 0.14	<0.001**
BAC, $\times 10^9$/L	0.03 ± 0.02	0.03 ± 0.02	0.019*
LYMPHC, $\times 10^9$/L	2.10 ± 0.57	2.12 ± 0.53	0.615

MOC, $\times 10^9$/L	0.37 ± 0.12	0.44 ± 0.13	<0.001**
PLT, $\times 10^9$/L	241 ± 50.4	252 ± 63.3	0.064
GRANC, $\times 10^9$/L	3.44 ± 1.15	4.00 ± 1.15	<0.001**
EOP, %	2.59 ± 1.91	3.07 ± 1.72	0.005**
BAP, %	0.51 ± 0.27	0.38 ± 0.23	<0.001**
RDW-CV, %	12.9 ± 0.65	13.9 ± 1.54	<0.001**
MPV, fL	10.2 ± 1.05	10.2 ± 1.07	0.758
PDW, fL	16.2 ± 0.35	14.9 ± 1.97	<0.001**
PCT, %	0.24 ± 0.04	0.26 ± 0.06	0.041*
GRANP, %	55.8 ± 7.74	58.4 ± 6.22	<0.001**
LYMPHP, %	35.0 ± 7.32	31.5 ± 5.71	<0.001**
MOP, %	6.09 ± 1.41	6.46 ± 1.35	0.006**
RDW-SD, %	41.9 ± 1.96	44.1 ± 3.91	<0.001**
PLT/HDL	189 ± 55.0	202 ± 65.5	0.046*
GLU/HDL	4.11 ± 0.98	4.38 ± 1.32	0.038*
PLT/LYMPHC	122 ± 38.0	124 ± 37.6	0.571
A/G	1.83 ± 0.28	3.27 ± 16.3	0.361
S/L	1.02 ± 0.43	2.09 ± 10.4	0.289
TyG	8.67 ± 0.56	8.87 ± 0.63	0.001**
eGFR, mL/(min×1.73m²)	91.8 ± 3.84	92.3 ± 5.18	0.349

Notes: Data are presented as N (%) or Mean ± SE. P-values were based on chi-square tests (χ^2-test) or t-test. *p<0.05, **p<0.01.

Performance Comparison of the Five Machine Learning Methods

Table 2 and Additional file 1: Figure S2 A and B show that the XGBoost algorithm has

the highest AUC (0.934) and PR-AUC (0.754), as well as high Recall (0.908) and Precision (0.995). The F1-Score (0.950) is only second to that of SVM, making its overall performance excellent. The LR algorithm also performs well, with AUC (0.923) and F1-Score (0.921), and a high Balanced Accuracy (0.896), indicating an outstanding overall performance. To maximize the identification of ONIHL patients, the XGBoost algorithm was ultimately selected to further build the prediction model.

Table 2 Performance Comparison of the Five Machine Learning Methods

	AUC	PR-AUC	Recall	Precision	Balanced accuracy	F1-sore
Logistic Regression (LR)	0.923	0.683	0.855	0.997	0.896	0.921
Random Forest (RF)	0.687	0.397	0.999	0.976	0.687	0.988
Support Vector Machine (SVM)	0.761	0.418	0.992	0.982	0.761	0.988
k-Nearest Neighbors (KNN)	0.784	0.309	0.973	0.984	0.784	0.979
XGBoost	0.934	0.754	0.908	0.995	0.892	0.950

The values of AUC, PR-AUC, Recall, Precision, Balanced Accuracy, and F1-Score range from 0 to 1, with higher values indicating better performance. LR stands for Logistic Regression, RF for Random Forest, SVM for Support Vector Machine, KNN for K-nearest Neighbors, and XGBoost for Extreme Gradient Boosting. AUC represents the area under the receiver operating characteristic curve, and PR-AUC represents the area under the precision-recall curve.

The results of the five-fold cross-validation on the training set show an AUC of 0.999, PR-AUC of 0.999, Sensitivity of 0.994, and Balanced Accuracy of 0.997. Additionally, the XGBoost model demonstrates reliable performance on the test set (AUC = 0.886, PR-AUC = 0.648), as shown in Additional file 1: Figure 2 A and B.

Feature Selection for the Final Model

Several pairs of features were observed to have high correlations, such as MCH and CV, and GRANP and LYMPHP, which may introduce redundant information and affect the model's decision-making and stability (Additional file 1: Figure S3 for related heat maps). Therefore, we used manual selection, PCA, and mRMR methods to identify the optimal features. As a result, 16 features were used to reconstruct the XGBoost model from each of manual selection, PCA, and mRMR (Table S1). The AUC for manual selection was 0.883, with a PR-AUC of 0.520, Sensitivity of 0.92, Specificity of 0.75, and Balanced Accuracy of 0.835. For PCA, the AUC was 0.885, with a PR-AUC of 0.435, Sensitivity of 0.734, Specificity of 0.781, and Balanced Accuracy of 0.758. For mRMR, the AUC was 0.941, with a PR-AUC of 0.731, Sensitivity of 0.886, Specificity of 0.844, and Balanced Accuracy of 0.865 (Figure 3 A, B, C, D). The model constructed with 16 features from mRMR on the validation set showed a slight improvement compared to the model built with 48 features (Figure 3 E). We further evaluated the model using an independent test set (D2). In the test set, the model built with manual selection had an AUC of 0.844 and PR-AUC of 0.543, Sensitivity of 0.711, Specificity of 0.875, and Balanced Accuracy of 0.793. The PCA model had an AUC of

0.85 and PR-AUC of 0.513, Sensitivity of 0.646, Specificity of 0.875, and Balanced Accuracy of 0.760. The mRMR model had an AUC of 0.852 and PR-AUC of 0.584, Sensitivity of 0.814, Specificity of 0.75, and Balanced Accuracy of 0.782. The mRMR model showed the best performance in the test set evaluation (Figure 3 F).

Feature Importance Ranking

To investigate which features contribute the most to the risk of ONIHL, we first used mRMR to select 16 important features and then built an XGBoost model based on these features. Subsequently, we ranked these features according to their weights in the XGBoost model, as shown in Figure 4. And the feature importance of the predictors based on PCA and manual curation is shown in Additional file 1: Figure S4, S5. The results indicated that the top five features, in order of importance, were ALB, CV, LYMPHP, MOC, and SD. Further comparisons between the ONIHL and normal samples revealed significant differences in CV, PDW, HCT, LYMPHP, and SD. These findings are highly consistent with the top-ranked results in the XGBoost model, indicating a strong correlation between these indicators (ALB, CV, LYMPHP, MOC, and SD) and ONIHL.

DCA Decision Curve Analysis

The DCA decision curve analysis evaluates the net benefit of three models at different thresholds (Figure 5). The mRMR model performed best across the entire threshold range, showing high and stable net benefit, indicating its higher clinical value in

decision-making. If no intervention (None) or intervention for all (All) was applied, their net benefit was lower than that of the mRMR model over a broad range, demonstrating the advantages of the decision model.

DISCUSSION

ONIHL represents a significant global public health concern[34,35]. Despite its complexity, ONIHL is a preventable condition[36,37]. The Occupational Safety and Health Administration (OSHA) requires the implementation of hearing conservation programs for workers exposed to noise levels of 85 decibels or higher, with the objective of safeguarding auditory health in noisy occupational environments[38]. Consequently, the development of a risk screening tool for ONIHL is crucial as a primary strategy for screening and prevention among workers exposed to occupational noise. In this study, we employed five ML algorithms utilizing hematological test results to construct an ONIHL risk screening model. The models demonstrated AUC values exceeding 0.85, with accuracy and sensitivity surpassing 0.75 in both validation and independent test datasets. These results suggest that ML models are capable of accurately identifying ONIHL patients within the population of noise-exposed workers.

In an evaluation of model performance on the validation set, the XGBoost model exhibited superior efficacy compared to all other algorithms assessed, achieving an AUC of 0.924 and a PR-AUC of 0.754. The precision, specificity, F-score, and balanced accuracy metrics for the XGBoost model all exceeded 0.8 on the validation

set. Furthermore, the XGBoost model maintained consistent performance on the test set, with an AUC of 0.886 and a PR-AUC of 0.648. XGBoost is recognized as a machine learning technique that efficiently and flexibly manages missing data and integrates weak predictive models into a robust predictive framework[39]. As an open-source package, XGBoost has gained significant recognition in various machine learning and data mining competitions. For example, in 2015, 17 out of the 29 winning solutions featured on Kaggle's blog utilized XGBoost, and all of the top 10 winning teams in the 2015 KDD Cup also incorporated XGBoost into their solutions[40]. Owing to its superior accuracy and performance, XGBoost-based machine learning algorithms are increasingly highlighted as competitive alternatives to traditional regression analysis and are employed in predicting adverse clinical outcomes. Furthermore, our findings indicate that the predictive efficacy of the XGBoost model surpasses that of LR, RF, SVM, and KNN. This aligns with previous research demonstrating that traditional logistic regression frequently exhibits comparatively lower AUC values in ROC curve analyses, alongside higher prediction errors and inferior performance relative to more contemporary methodologies[41,42].

Traditional hearing loss diagnosis often relies on audiometric tests, such as pure-tone audiometry, which require specialized equipment and trained personnel, thus increasing the time, cost, and resources involved in diagnosis. Additionally, many published machine learning models, despite showing potential in predicting the impact of noise exposure on hearing, often still depend on individual hearing test data or direct

measurements of noise exposure levels to build accurate prediction models. This dependency adds complexity to their practical application. In contrast, our model operates independently of hearing assessments and direct noise exposure measurements, focusing instead on analyzing routine blood and biochemical indicators for prediction. This approach significantly reduces reliance on specialized equipment and data, thereby saving manpower and material resources in the diagnostic process. It provides an efficient and convenient alternative that holds promise for screening and early prevention of noise-induced hearing loss.

Nonetheless, it is crucial to acknowledge the limitations inherent in this study. Firstly, ONIHL is affected by a multitude of factors, and our study sample is confined to the population of Shenzhen. Consequently, the model's performance in other demographic groups remains uncertain. Therefore, further validation is required to assess the model's adaptability and generalizability in larger and more diverse populations, ensuring its efficacy across various occupations and environmental contexts. Secondly, the accuracy of the current method is relatively low, primarily due to the imbalance in the distribution between patients with occupational ONIHL and noise-exposed individuals with normal hearing. The ratio of positive to negative samples exceeds 1:20, which undoubtedly mirrors the real-world scenario where ONIHL cases are infrequent among noise-exposed workers.

Consequently, the early identification and intervention of risk factors identified in our model could have substantial implications for the prevention of ONIHL among

workers exposed to noise. The risk factors contributing to the development of ONIHL are varied. We have developed a risk assessment model for ONIHL utilizing clinical data and routine physical examination indicators, employing a machine learning algorithm. This approach contrasts with most existing methods for predicting ONIHL risk, which predominantly depend on variables such as age, sex, medical history (including conditions like hypertension and diabetes), history of noise exposure, and behavioral factors such as smoking and physical activity[15,43,44]. For instance, prior research has developed risk models for workers exposed to noise, yielding favorable predictive outcomes. These models primarily incorporate risk factors such as industry type, duration of noise exposure, and median peak intensity, which contrast with the physical examination indicators utilized in our study[19]. Yi Wang[9] formulated a machine learning-based risk assessment model for high-frequency hearing loss employing routine physical examination data, attaining AUC of 0.868. This model, however, was principally designed for community residents and incorporated risk factors including 13 blood test indicators, demographic characteristics, disease-related features, behavioral factors, environmental exposure, and auditory cognitive factors, which differ from the population of noise-exposed workers in our study. Our model offers a more comprehensive approach than previous research by integrating a wide range of biochemical and Routine Blood indicators to assess the risk of ONIHL from multiple dimensions. Unlike models that rely on hearing assessments and direct noise exposure measurements, our model focuses on routine blood and biochemical

indicators, reducing the need for specialized equipment and resources. This makes it a more efficient, cost-effective alternative for early detection and prevention of ONIHL, offering personalized risk assessments without the reliance on extensive testing.

Routine blood tests administered at occupational disease prevention clinics are typically conducted on an annual basis. Based on these tests, the application of these indicators can enhance early screening and provide warnings for prevalent occupational diseases. In our study, the developed model demonstrates the significance of hematological test data in screening for ONIHL. This includes variables such as age, sex, inflammatory and immune markers (e.g., ALB, WBC, LYMPHP, MOC, and GRANC), as well as oxidative stress and metabolic markers (e.g., Scr and TP), among others. Five of these indicators exhibit a substantial impact on the ONIHL screening model: ALB, CV, LYMPHP, MOC, and SD. Reduced ALB levels, elevated LYMPHP, and increased MOC are associated with an elevated risk of ONIHL. Empirical evidence suggests that diminished serum albumin concentrations correlate with a heightened risk of SSNHL[45]. Serum albumin may contribute to cochlear function maintenance or serve as an indicator of vascular conditions influencing auditory health. LYMPHP and MOC, as immune markers, frequently rise during inflammatory responses. Current research suggests that immune and inflammatory processes in the inner ear subsequent to noise exposure may be intricately connected to hearing impairment. Noise-induced cochlear damage initiates an immune response, resulting in the heightened activity of white blood cells, lymphocytes, monocytes, and granulocytes. These immune cells

migrate to the site of cochlear injury, instigating an inflammatory response and releasing cytokines, including tumor necrosis factor-alpha (TNF-α) and interleukins (IL-6), which have been demonstrated to exacerbate cochlear damage and contribute to the deterioration of auditory function[46,47].

Furthermore, the oxidative stress induced by noise and the consequent inflammatory response can impair the microcirculation within the inner ear, leading to additional damage. A Mendelian randomization study indicated a significant association between lymphocyte count and susceptibility to SSNHL[48]. CV and SD are identified as risk factors for noise-induced hearing loss. The CV and SD of RDW serve as primary metrics for assessing RDW. These metrics reflect the extent of variability in red blood cell volume, with higher RDW values generally linked to various health conditions, such as chronic inflammation and anemia[49]. Chronic inflammation has the potential to impair the auditory system, thereby elevating the risk of ONIHL. There exists a positive correlation between CV, SD, and average hearing threshold, underscoring the importance of identifying inflammatory conditions for the screening of workers susceptible to chronic inflammation and ONIHL[50]. Furthermore, elevated RDW values may suggest variability in the morphology and function of red blood cells, potentially affecting oxygen transport and utilization, which in turn could influence inner ear health[51]. Additionally, our model suggests that Scr and TP metabolic markers are associated with ONIHL. The interconnections among Scr, TP, oxidative stress, and antioxidant capacity have been investigated in numerous

studies[52,53]. Oxidative stress is considered a potential pathological mechanism underlying ONIHL, as it can result in damage to inner ear cells through the production of free radicals and other reactive oxygen species (ROS). Metabolic markers, including serum creatinine and total protein, may function as indirect indicators of an individual's oxidative stress levels and antioxidant capacity. Additionally, prior research has identified age and male gender as risk factors for hearing loss[54]. Consequently, the early identification and intervention of risk factors identified in our model could have substantial implications for the prevention of ONIHL among workers exposed to noise.

CONCLUSION

In this study, we developed five machine learning models to construct a risk screening model for ONIHL, with the XGBoost-based model demonstrating superior performance. By integrating biochemical and hematological indicators with machine learning techniques, this model effectively identifies individuals at high risk for ONIHL. This approach not only introduces a novel tool for the early screening of hearing loss but also lays the groundwork for the development of personalized intervention strategies. In the future, the integration of additional biological data is anticipated to further augment the model's predictive capabilities. Furthermore, this model holds potential for extension to forecast risks associated with other occupational or chronic diseases, thereby offering substantial support for the maintenance and enhancement of public health.

Abbreviations

ONIHL Occupational noise-induced hearing loss

XGBoost Extreme gradient boosting

LR Logistic regression

RF Random forest

SVM Support vector machines

KNN K-nearest neighbors

DCA Decision curve analysis

ALB Serum albumin

RDW-CV Coefficient of variation in red cell distribution width

LYMPHP Lymphocyte percentage

MOC Monocyte count

RDW-SD Standard deviation in red cell distribution width

ML Machine learning

SSNHL Sudden sensorineural hearing loss

WBC White blood cells

NE Neutrophils

MO Monocytes

LY Lymphocytes

LDL Low-density lipoprotein

HDL High-density lipoprotein

RDW Red cell distribution width

TP Total protein

ALB Albumin

GLU Glucose

CHO Cholesterol

TG Triglycerides

TBIL Total bilirubin

DBIL Direct bilirubin

IBIL Indirect bilirubin

ALT Alanine aminotransferase

AST Aspartate aminotransferase

BUN Blood urea nitrogen

Scr Serum creatinine

UA Uric acid

GLB Globulin

Hb Hemoglobin

RBC Red blood cell count

HCT Hematocrit

MCV Mean corpuscular volume

MCH Mean corpuscular hemoglobin

MCHC Mean corpuscular hemoglobin concentration

EOC Eosinophil count

BAC Basophil count

LYMPHC Lymphocyte count

MOC Monocyte count

PLT Platelet count

GRANC Neutrophil count

EOP Eosinophil percentage

BAP Basophil percentage

MPV Mean platelet volume

PDW Platelet distribution width

PCT Plateletcrit

GRANP Neutrophil percentage

MOP Monocyte percentage

PLT/HDL Platelet-to-HDL ratio

GLU/HDL Glucose-to-HDL ratio

PLT/LYMPHC Platelet-to-lymphocyte ratio

A/G Albumin-to-globulin ratio

S/L Neutrophil-to-lymphocyte ratio

TyG Triglyceride-glucose index

eGFR Estimated glomerular filtration rate

PCA Principal component analysis

mRMR Maximum relevance minimum redundancy

OSHA The Occupational Safety and Health Administration

TNF-α Tumor necrosis factor-alpha

IL-6 Interleukin-6

ROS Reactive oxygen species

Acknowledgment

We thank Shenzhen Prevention and Treatment Center for Occupational Diseases for the approval of the ethical clearance. We also extend our warm gratitude to the different hospital stakeholders and participants for their valuable contribution during data collection.

Author contributions

The authors made substantial contributions to the acquisition, analysis, and

interpretation of the data and the drafting and revision of the manuscript. All authors also approved the final version of the paper and agreed to be accountable for all aspects of the work. Caiping Li and Dianpeng Wang: Writing – original draft, Investigation, Data curation, Conceptualization. Caiping Li, Liuwei Shi and Linlin Chen: Methodology, Data curation. Dafeng Lin: Data curation. Xiangli Yang and Liang Zhou, Investigation. Peimao Li: Validation, Investigation. Wen Zhang: Validation. Yan Guo and Naixing Zhang: Supervision, Project administration, Conceptualization. Dafeng Lin: Writing – original draft, Supervision, Project administration, Formal analysis, Conceptualization.

Funding

This work was supported by Science and Technology Planning Project of Shenzhen Municipality (No.KCXFZ20201221173602007, No.JCYJ20220531091211026), Shenzhen Fund for Guangdong Provincial High- level Clinical Key Specialties (No.SZGSP015) and Scientific research project of Shenzhen Prevention and Treatment Center for Occupational Diseases (NO:SZF-PY-2023-008).

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethics Approval and Consent to Participate

This study was approved by the Ethics Committee of Shenzhen Prevention and Treatment Center for Occupational Diseases (Approval Number: LL2020-34, Date: 14th December 2020). All methods were carried out in accordance with relevant ethical guidelines and regulations.

Patient consent for publication

Not applicable.

Data availability

Data will be made available on request.

Patient and Public Involvement

Patients or the public were not involved in the design, conduct, reporting, or dissemination plans of our research.

REFERENCES

1 Ding T, Yan A, Liu K. What is noise-induced hearing loss? *Br J Hosp Med*. 2019;80:525–9.

2 Nelson DI, Nelson RY, Concha-Barrientos M, *et al.* The global burden of occupational noise-induced hearing loss. *Am J Ind Med*. 2005;48:446–58.

3 Themann C, Suter A, Stephenson M. National Research Agenda for the Prevention of Occupational Hearing Loss—Part 1. *Semin Hear*. 2013;34:145–207.

4 Themann CL, Masterson EA. Occupational noise exposure: A review of its effects, epidemiology, and impact with recommendations for reducing its burden. *J Acoust Soc Am*. 2019;146:3879.

5 D T-V, A A, Gp R. What can we learn from adult cochlear implant recipients with single-sided deafness who became elective non-users? *Cochlear implants international*. 2020;21.

6 Li YH, Jiao J, Yu SF. Research status of influencing factors of noise-induced hearing loss. *Chin J Occup Dis*. 2014;32:469–73.

7 Vlaming MSMG, MacKinnon RC, Jansen M, *et al.* Automated screening for high-frequency hearing loss. *Ear Hear*. 2014;35:667–79.

8 Cunningham LL, Tucci DL. Hearing Loss in Adults. *N Engl J Med*. 2017;377:2465–73.

9 Wang Y, Yao X, Wang D, *et al.* A machine learning screening model for identifying the risk of high-frequency hearing impairment in a general population. *BMC Public Health*. 2024;24:1160.

10 Rm M, Rc M. Objective auditory brainstem response classification using machine learning. *International journal of audiology*. 2019;58.

11 Chang Y-S, Park H, Hong SH, *et al.* Predicting cochlear dead regions in patients with hearing loss through a machine learning-based approach: A preliminary study. *PLoS One*. 2019;14:e0217790.

12 Abdollahi H, Mostafaei S, Cheraghi S, *et al.* Cochlea CT radiomics predicts chemoradiotherapy induced sensorineural hearing loss in head and neck cancer patients: A machine learning and multi-variable modelling study. *Phys Med*. 2018;45:192–7.

13 Tomiazzi JS, Pereira DR, Judai MA, *et al.* Performance of machine-learning algorithms to pattern recognition and classification of hearing impairment in Brazilian farmers exposed to pesticide and/or cigarette smoke. *Environ Sci Pollut Res Int*. 2019;26:6481–91.

14 D B, J Y, J M, *et al.* Predicting the hearing outcome in sudden sensorineural hearing loss via machine learning models. *Clinical otolaryngology : official journal of ENT-UK ; official journal of Netherlands Society for Oto-Rhino-Laryngology & Cervico-Facial*

Surgery. 2018;43.

15 Aliabadi M, Farhadian M, Darvishi E. Prediction of hearing loss among the noise-exposed workers in a steel factory using artificial intelligence approach. *Int Arch Occup Environ Health*. 2015;88:779–87.

16 Farhadian M, Aliabadi M, Darvishi E. Empirical estimation of the grades of hearing impairment among industrial workers based on new artificial neural networks and classical regression methods. *Indian J Occup Environ Med*. 2015;19:84–9.

17 Ys K, Yh C, Oj K, *et al.* The Risk Rating System for Noise-induced Hearing Loss in Korean Manufacturing Sites Based on the 2009 Survey on Work Environments. *Safety and health at work*. 2011;2.

18 Nawi NM, Rehman MZ, Ghazali MI. Noise-induced hearing loss prediction in Malaysian industrial workers using gradient descent with adaptive momentum algorithm. *International Review on Computers and Software*. 2011;6:740–8.

19 Y Z, J L, M Z, *et al.* Machine Learning Models for the Hearing Impairment Prediction in Workers Exposed to Complex Industrial Noise: A Pilot Study. *Ear and hearing*. 2019;40.

20 Li P, Pang K, Zhang R, *et al.* Prevalence and risk factors of hearing loss among the middle-aged and older population in China: a systematic review and meta-analysis. *Eur Arch Otorhinolaryngol*. 2023;280:4723–37.

21 Tsimpida Dialechti, Kontopantelis Evangelos, Ashcroft Darren, *et al.* Socioeconomic and lifestyle factors associated with hearing loss in older adults: a cross-sectional study of the English Longitudinal Study of Ageing (ELSA). *BMJ open*. 2019;9.

22 Baiduc Rachael R, Sun Joshua W, Berry Caitlin M, *et al.* Relationship of cardiovascular disease risk and hearing loss in a clinical population. *Scientific reports*. 2023;13.

23 Jung Da Jung, Do Jun Young, Cho Kyu Hyang, *et al.* Association between triglyceride/high-density lipoprotein ratio and hearing impairment in a Korean population. *Postgraduate medicine*. 2017;129.

24 Verschuur Carl Anton, Dowell Aphra, Syddall Holly Emma, *et al.* Markers of inflammatory status are associated with hearing threshold in older people: findings from the Hertfordshire Ageing Study. *Age and ageing*. 2012;41.

25 Nonoyama Hiroshi, Tanigawa Tohru, Shibata Rei, *et al.* Red blood cell distribution width predicts prognosis in idiopathic sudden sensorineural hearing loss. *Acta oto-laryngologica*. 2016;136.

26 Lunardon N, Menardi G, Torelli N. ROSE: A Package for Binary Imbalanced Learning. *The R Journal*. 2014;6:79–89.

27 Dina Mohamed Ahmed Samir Elkahwagy, Caroline Joseph Kiriacos. Logistic regression and other statistical tools in diagnostic biomarker studies. *Clinical & translational oncology : official publication of the Federation of Spanish Oncology*

Societies and of the National Cancer Institute of Mexico. 2024;26.

28 Schauberger G, Klug SJ, Berger M. Random forests for the analysis of matched case-control studies. *BMC Bioinformatics*. 2024;25:253.

29 Valkenborg D, Rousseau A-J, Geubbelmans M, *et al.* Support vector machines. *Am J Orthod Dentofacial Orthop*. 2023;164:754–7.

30 Garcia-Carretero R, Vigil-Medina L, Mora-Jimenez I, *et al.* Use of a K-nearest neighbors model to predict the development of type 2 diabetes within 2 years in an obese, hypertensive population. *Med Biol Eng Comput*. 2020;58:991–1002.

31 Bridgelall Raj, Tolliver Denver D. Railroad accident analysis using extreme gradient boosting. *Accident Analysis and Prevention*. 2021;156.

32 Peng Hanchuan, Long Fuhui, Ding Chris. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE transactions on pattern analysis and machine intelligence*. 2005;27.

33 Xia Zhiming, Chen Yang, Xu Chen. Multiview PCA: A Methodology of Feature Extraction and Dimension Reduction for High-Order Data. *IEEE transactions on cybernetics*. Published Online First: 2021vo PP.

34 Nelson Deborah Imel, Nelson Robert Y, Concha-Barrientos Marisol, *et al.* The global burden of occupational noise-induced hearing loss. *American journal of industrial medicine*. 2005;48.

35 Mariola Śliwińska-Kowalska, Kamil Zaborowski. WHO Environmental Noise Guidelines for the European Region: A Systematic Review on Environmental Noise and Permanent Hearing Loss and Tinnitus. *International Journal of Environmental Research and Public Health*. 2017;14.

36 Seixas Noah S, Neitzel Rick, Stover Bert, *et al.* A multi-component intervention to promote hearing protector use among construction workers. *International journal of audiology*. Published Online First: 2011.

37 Amjad-Sardrudi Hossein, Dormohammadi Ali, Golmohammadi Rostam, *et al.* Effect of noise exposure on occupational injuries: a cross-sectional study. *Journal of research in health sciences*. 2012;12.

38 Park Sungwon, Johnson Michael D, Hong OiSaeng. Analysis of Occupational Safety and Health Administration (OSHA) noise standard violations over 50 years: 1972 to 2019. *American journal of industrial medicine*. 2020;63.

39 Kuo-Ching Yuan, Lung-Wen Tsai, Ko-Han Lee, *et al.* The development an artificial intelligence algorithm for early sepsis diagnosis in the intensive care unit. *International Journal of Medical Informatics*. 2020;141.

40 Tianqi Chen, Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. *CoRR*. Published Online First: 2016vo abs.

41 Xiao Jing, Ding Ruifeng, Xu Xiulin, *et al.* Comparison and development of machine

learning tools in the prediction of chronic kidney disease progression. *Journal of translational medicine*. 2019;17.

42 Li Y-M, Li Z-L, Chen F, *et al.* A LASSO-derived risk model for long-term mortality in Chinese patients with acute coronary syndrome. *J Transl Med*. 2020;18:157.

43 Chen F, Cao Z, Grais EM, *et al.* Contributions and limitations of using machine learning to predict noise-induced hearing loss. *Int Arch Occup Environ Health*. 2021;94:1097–111.

44 Sun R, Shang W, Cao Y, *et al.* A risk model and nomogram for high-frequency hearing loss in noise-exposed workers. *BMC Public Health*. 2021;21:747.

45 Zheng Zhong, Liu Chengqi, Shen Ying, *et al.* Serum Albumin Levels as a Potential Marker for the Predictive and Prognostic Factor in Sudden Sensorineural Hearing Loss: A Prospective Cohort Study. *Frontiers in Neurology*. 2021;12.

46 Keithley EM. Cochlear Inflammation Associated with Noise-Exposure. In: Ramkumar V, Rybak LP, eds. *Inflammatory Mechanisms in Mediating Hearing Loss*. Cham: Springer International Publishing 2018:91–114.

47 Hu B hua, Zhang C. Immune System and Macrophage Activation in the Cochlea: Implication for Therapeutic Intervention. In: Pucheu S, Radziwon KE, Salvi R, eds. *New Therapies to Prevent or Cure Auditory Disorders*. Cham: Springer International Publishing 2020:113–34.

48 Chen Jialei, Wu Chao, He Jing, *et al.* Causal associations of thyroid function and sudden sensorineural hearing loss: a bidirectional and multivariable Mendelian randomization study. *Frontiers in Neurology*. 2023;14.

49 Jung Da Jung, Yoo Myung Hoon, Lee Kyu-Yup. Red cell distribution width is associated with hearing impairment in chronic kidney disease population: a retrospective cross-sectional study. *European archives of oto-rhino-laryngology : official journal of the European Federation of Oto-Rhino-Laryngological Societies (EUFOS) : affiliated with the German Society for Oto-Rhino-Laryngology - Head and Neck Surgery*. 2020;277.

50 Natarajan N, Batts S, Stankovic KM. Noise-Induced Hearing Loss. *Journal of Clinical Medicine*. 2023;12:2347.

51 Shi X. Cochlear Vascular Pathology and Hearing Loss. In: Ramkumar V, Rybak LP, eds. *Inflammatory Mechanisms in Mediating Hearing Loss*. Cham: Springer International Publishing 2018:61–90.

52 José Pedraza-Chaverrí, Diana Barrera, Omar N Medina-Campos, *et al.* Time course study of oxidative and nitrosative stress and antioxidant enzymes in K2Cr2O7-induced nephrotoxicity. *BMC Nephrology*. 2005;6.

53 Galiniak Sabina, Biesiadecki Marek, Moło ń Mateusz, *et al.* Serum Oxidative and Nitrosative Stress Markers in Clear Cell Renal Cell Carcinoma. *Cancers*. 2023;15.

54 Chou Chiu-Fang, Beckles Gloria L A, Zhang Xinzhi, *et al.* Association of Socioeconomic Position With Sensory Impairment Among US Working-Aged Adults. *American journal of public health*. 2015;105.

LEGENDS FOR FIGURES

Fig. 1 A Combined Framework for Identifying ONIHL Patients.

Fig. 2 Performance of the Prediction Model on the Validation Set of Dataset D1 and the Test Set of Dataset D2. A: ROC Curve; B: Precision-Recall Curve

Fig. 3 Feature Selection for the Final Model Using PCA, Manual Selection, and mRMR. A and B: ROC Curves for Models Constructed with PCA, Manual Selection, and mRMR Features on the Validation Set of Dataset D1 and the Test Set of Dataset D2; C and D: Precision-Recall Curves for the Above Models; D: Comparison of Sensitivity, Specificity, and Balanced Accuracy of Models Before and After Feature Selection on the Test Set; F: Comparison of Sensitivity, Specificity, and Balanced Accuracy of Models Using the Selected Features on the Independent Test Set D2.

Fig. 4 Feature Importance Ranking for the Model Built Using Features Selected by mRMR.

Fig. 5 DCA Decision Curves for Models Built Using Three Different Feature Selection Methods.

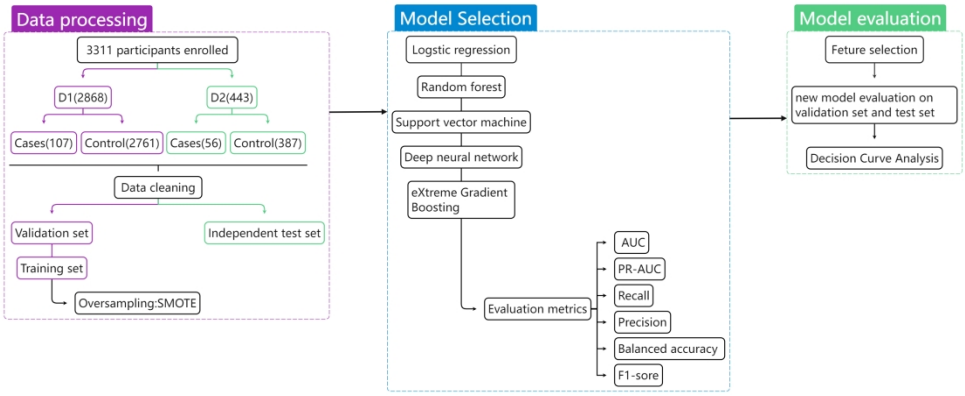

Fig. 1 A Combined Framework for Identifying ONIHL Patients.

283x124mm (236 x 236 DPI)

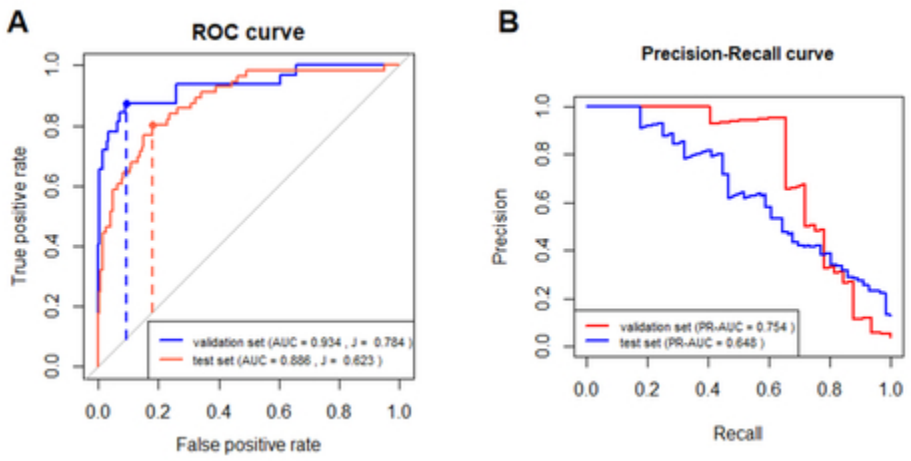

Fig. 2 Performance of the Prediction Model on the Validation Set of Dataset D1 and the Test Set of Dataset D2. A: ROC Curve; B: Precision-Recall Curve

41x19mm (300 x 300 DPI)

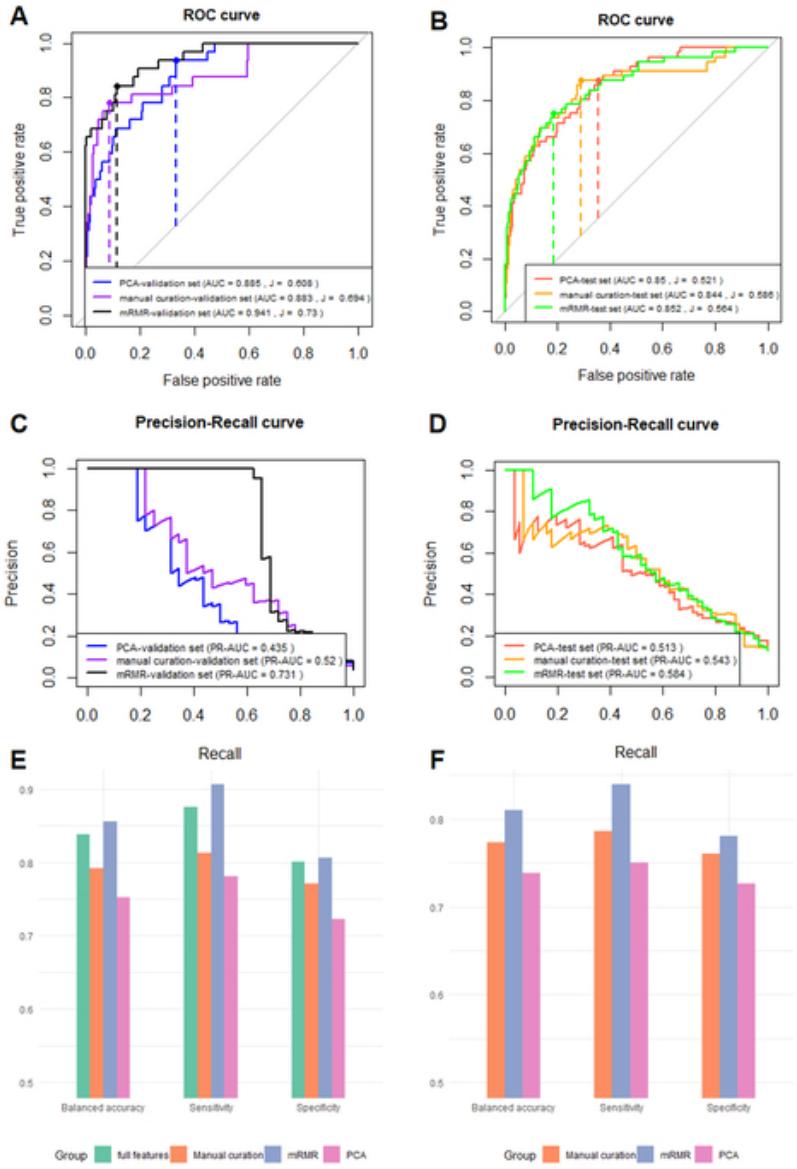

Fig. 3 Feature Selection for the Final Model Using PCA, Manual Selection, and mRMR. A and B: ROC Curves for Models Constructed with PCA, Manual Selection, and mRMR Features on the Validation Set of Dataset D1 and the Test Set of Dataset D2; C and D: Precision-Recall Curves for the Above Models; D: Comparison of Sensitivity, Specificity, and Balanced Accuracy of Models Before and After Feature Selection on the Test Set; F: Comparison of Sensitivity, Specificity, and Balanced Accuracy of Models Using the Selected Features on the Independent Test Set D2.

41x59mm (300 x 300 DPI)

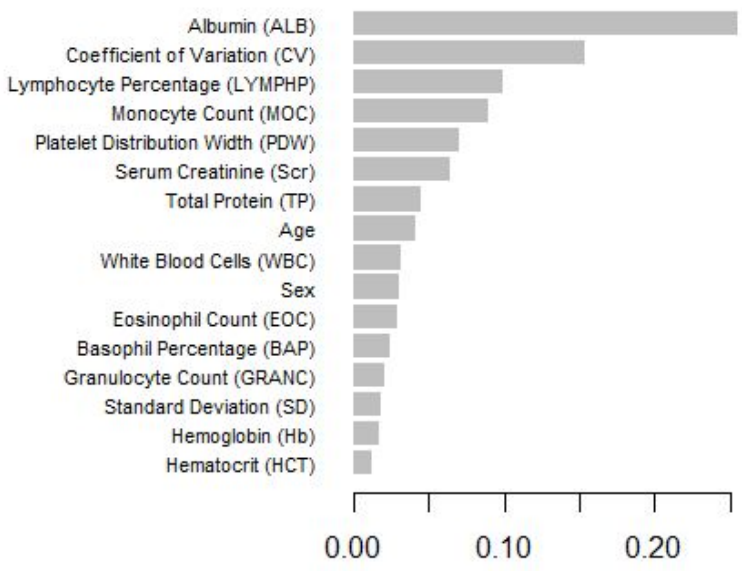

Fig. 4 Feature Importance Ranking for the Model Built Using Features Selected by mRMR.

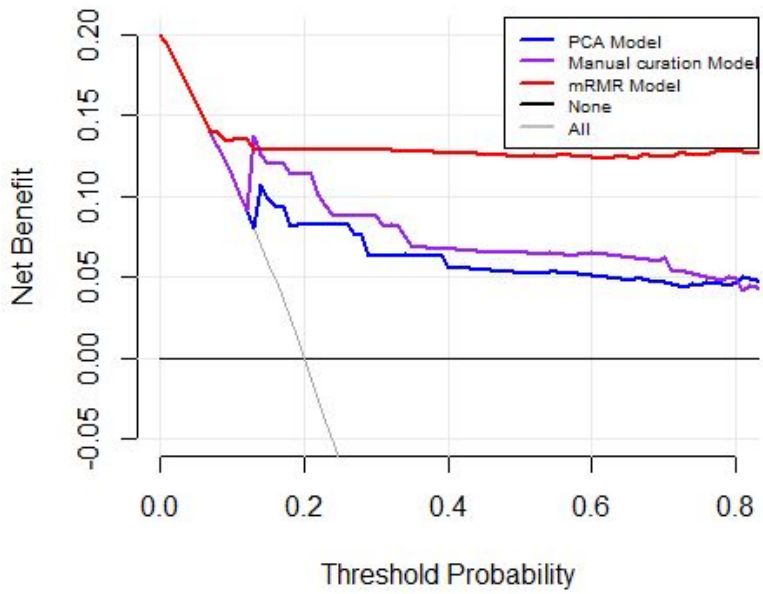

Fig. 5 DCA Decision Curves for Models Built Using Three Different Feature Selection Methods.

Table S1 Feature Subsets for the Final Model Obtained Through PCA, Manual Selection, and mRMR

Manual curation	PCA	mRMR	Rank
Albumin (ALB)	Serum Creatinine (Scr)	Albumin (ALB)	1
Age	Eosinophil Count (EOC)	Coefficient of Variation (CV)	2
Estimated Glomerular Filtration Rate (eGFR)	Hemoglobin (Hb)	Lymphocyte Percentage (LYMPHP)	3
Hemoglobin (Hb)	Albumin-Globulin Ratio (A/G)	Monocyte Count (MOC)	4
Triglyceride-Glucose Index (TyG)	Triglyceride-Glucose Index (TyG)	Platelet Distribution Width (PDW)	5
Eosinophil Count (EOC)	Basophil Percentage (BAP)	Serum Creatinine (Scr)	6
Granulocyte Count (GRANC)	Globulin (GLB)	Total Protein (TP)	7
Total Bilirubin (TBIL)	Granulocyte Percentage (GRANP)	Age	8
Albumin-Globulin Ratio (A/G)	Monocyte Percentage (MOP)	White Blood Cells (WBC)	9
White Blood Cells (WBC)	Indirect Bilirubin (IBIL)	Sex	10
Platelet Count (PLT)	Red Blood Cells (RBC)	Eosinophil Count (EOC)	11

Red Blood Cells (RBC)	Blood Urea Nitrogen (BUN)	Basophil Percentage(BAP)	12
Basophil Count (BAC)	Alanine Aminotransferase (ALT)	Granulocyte Count (GRANC)	13
Mean Corpuscular Hemoglobin (MCH)	Glucose to High-Density Lipoprotein Ratio (GLU/HDL)	Standard Deviation (SD)	14
Segmented Neutrophil to Lymphocyte Ratio (S/L)	Cholesterol (CHO)	Hemoglobin (Hb)	15
Sex	Glucose (GLU)	Hematocrit (HCT)	16

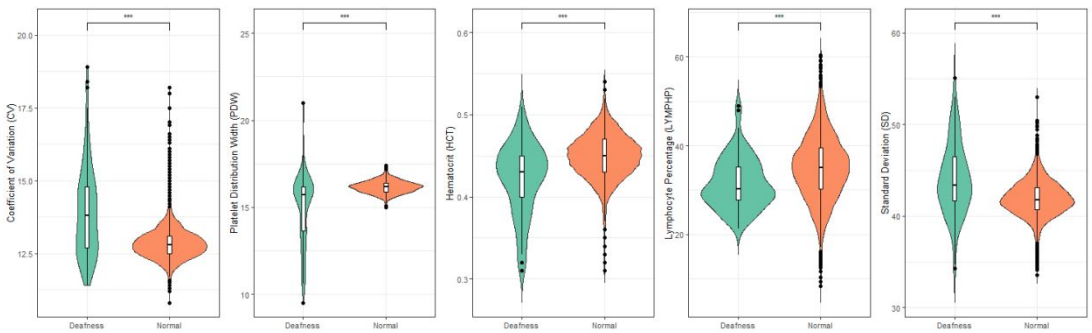

Fig. S1 The Five Features with the Greatest Differences Between ONIHL Patients and Noise-Exposed Individuals with Normal Hearing.

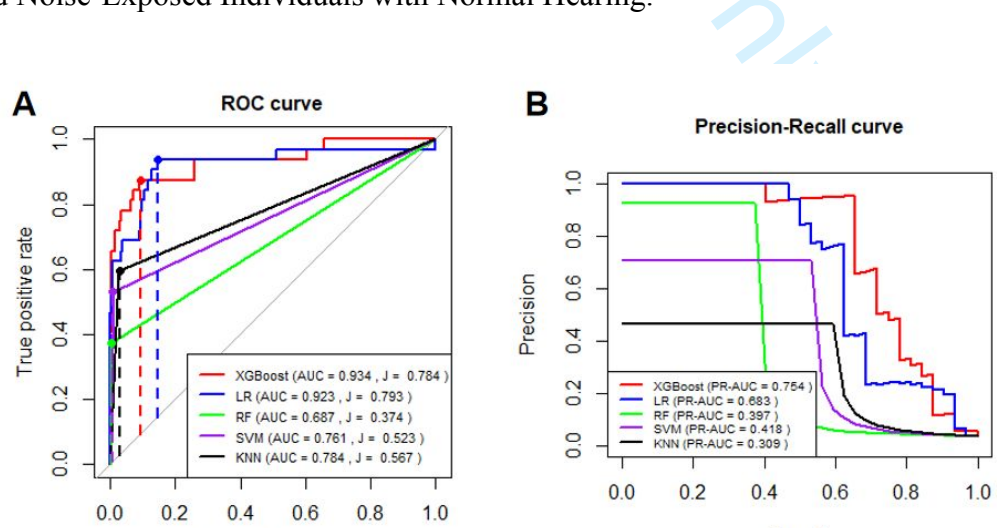

Fig. S2 Performance of the Five Prediction Models in the Validation Set of Dataset D1. A: ROC Curve; B: Precision-Recall Curve

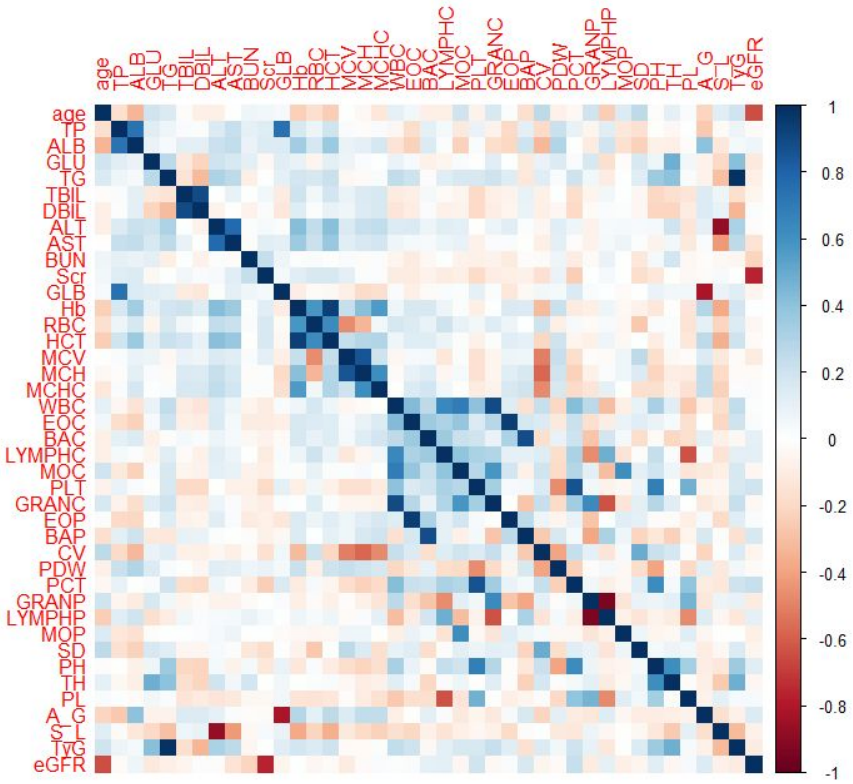

Fig. S3 Heat map of relationship between full variables.

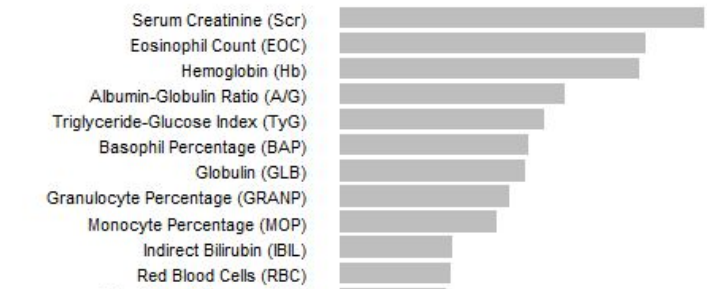

Fig. S4 feature ranking of PCA screening variables.

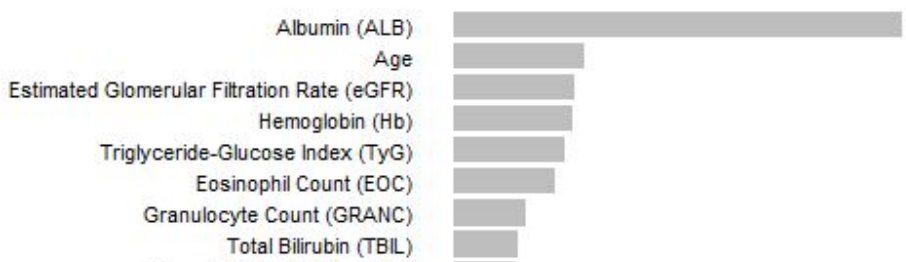

Fig. S5 feature ranking of manual screening variables.

1. TyG Index (Triglyceride-Glucose Index)

The triglyceride-glucose (TyG) index is an established marker for evaluating insulin resistance, commonly used in assessing metabolic syndrome and diabetes risk. It is calculated using the following formula:

$$TyG = \ln\left[\frac{\mathrm{TG}(mg/dL) \times GLU(mg/dL)}{2}\right]$$

where:

- **TG** represents triglyceride levels (mg/dL),
- **GLU** denotes fasting glucose levels (mg/dL), and
- **ln** is the natural logarithm.

2. eGFR Estimation Formula (Specific to Guangzhou, China Population)

To improve the accuracy of estimated glomerular filtration rate (eGFR) for the Guangzhou population in China, a modified formula has been developed based on local demographic and clinical data:

$$\mathrm{eGFR}\left(mL/(min \times 1.73m^2)\right) = 106 \times \left(\frac{88.4}{Scr(\,mg/dL\,)}\right)^{0.203} \times 0.996^{Age\ (\mathrm{year})}$$

where:

- **Scr** represents serum creatinine concentration (mg/dL), and
- **Age** is the individual's age in years.

The constant **88.4** is employed to convert creatinine units from μmol/L to mg/dL for international standardization.

Contextual Relevance of the Formulas

The **TyG index** serves as an indirect measure of insulin resistance and is useful in predicting metabolic health outcomes. In contrast, the **eGFR formula** provides an estimate of kidney function, tailored to the Chinese population, and is instrumental in identifying and monitoring renal health.

BMJ Open

Construction of a Risk Prediction Model for Occupational Noise-Induced Hearing Loss Using Routine Blood and Biochemical Indicators in Shenzhen, China: A Predictive Modeling Study

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2024-097249.R1
Article Type:	Original research
Date Submitted by the Author:	26-Mar-2025
Complete List of Authors:	Wang, Dianpeng; Southern Medical University, School of Public Health; Shenzhen Prevention and Treatment Center for Occupational Diseases Li, Caiping; Southern Medical University, School of Public Health shi, liuwei; Jilin University Chen, Linlin; Southern Medical University, School of Public Health Lin, Dafeng; Shenzhen Prevention and Treatment Center for Occupational Diseases yang, xiangli; Shenzhen Prevention and Treatment Center for Occupational Diseases Li, peimao; Shenzhen Prevention and Treatment Center for Occupational Diseases Zhang, Wen; Shenzhen Prevention and Treatment Center for Occupational Diseases Feng, wenting; Shenzhen Prevention and Treatment Center for Occupational Diseases Guo, Yan; Shenzhen Prevention and Treatment Center for Occupational Diseases Zhou, Liang; Southern Medical University, School of Public Health Zhang, naixing; Shenzhen Prevention and Treatment Center for Occupational Diseases
Primary Subject Heading:	Occupational and environmental medicine
Secondary Subject Heading:	Public health
Keywords:	Machine Learning, Audiology < OTOLARYNGOLOGY, Blood bank & transfusion medicine < HAEMATOLOGY, Risk Factors

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Construction of a Risk Prediction Model for Occupational Noise-Induced Hearing Loss Using Routine Blood and Biochemical Indicators in Shenzhen, China: A Predictive Modeling Study

Caiping Li, MS[1], Liuwei Shi, MS[3], Linlin Chen, MS[1], Dafeng Lin, PhD[2], Xiangli Yang, BS[2], Peimao Li, MS[2], Wen Zhang, MS[2], Wenting Feng, MS[2], Yan Guo, PhD[2], Liang Zhou, PhD[1*], Naixing Zhang, PhD[2*], Dianpeng Wang, MS[1,2*]

[1]Department of Toxicology, School of Public Health, Southern Medical University, Guangzhou 510515, China

[2]Medical laboratory, Shenzhen Prevention and Treatment Center for Occupational Diseases, Shenzhen 518020, China

[3]School of Public Health, Jilin University, Changchun 130012, China

[*]**Correspondence:** [1] Dianpeng Wang and Naixing Zhang, Medical laboratory, Shenzhen Prevention and Treatment Center for Occupational Diseases, 2019 Buxin Rd., Luohu District, Shenzhen 518020, China. E-mail address: szpcr@126.com and zhanghealth@126.com

Liang Zhou, Department of Toxicology, School of Public Health, Southern Medical University, No.1023-1063 Shatai South Rd., Baiyun District, Guangzhou 510515, China. E-mail address: zhzliang@smu.edu.cn

C L and L S contributed equally to this work.

word count: 5443

Keywords Noise-induced hearing loss, risk prediction, machine learning, blood routine indicators, biochemical indicators.

ABSTRACT

Objectives Occupational noise-induced hearing loss (ONIHL) represents a prevalent occupational health condition, traditionally necessitating multiple pure-tone audiometry assessments. We have developed and validated a machine learning model leveraging routine hematological and biochemical parameters, thereby offering novel insights into the risk prediction of ONIHL.

Design, setting and participants This study analyzed data from 3,297 noise-exposed workers in Shenzhen, including 160 ONIHL cases, with the dataset divided into D1 (2,868 samples, 107 ONIHL cases) and D2 (429 samples, 53 ONIHL cases). The inclusion criteria were formulated based on the GBZ49-2014 Diagnosis of Occupational Noise-Induced Hearing Loss. Model training was performed using D1, and model validation was conducted using D2. Routine blood and biochemical indicators were extracted from the case data, and a range of machine learning algorithms including extreme gradient boosting (XGBoost) were employed to construct predictive models. The model underwent refinement to identify the most representative variables, and Decision Curve Analysis (DCA) was conducted to evaluate the net benefit of the model across various threshold levels.

Primary outcome measures Model creation dataset and validation datasets: ONIHL.

Results The prediction model, developed using XGBoost, demonstrated exceptional performance, achieving an area under the curve (AUC) of 0.942, a sensitivity of 0.875, and a specificity of 0.936 on the validation dataset. On the test dataset, the model

achieved an AUC of 0.990. After implementing feature selection, the model was refined to include only 16 features, while maintaining strong performance on a newly acquired independent dataset, with an AUC of 0.872, a balanced accuracy of 0.798, a sensitivity of 0.755, and a specificity of 0.840.The analysis of feature importance revealed that serum albumin (ALB), platelet distribution width (PDW), coefficient of variation in red cell distribution width (RDW-CV), serum creatinine (Scr) and lymphocyte percentage (LYMPHP) are critical factors for risk stratification in patients with ONIHL.

Conclusion The analysis of feature importance identified ALB, PDW, RDW-CV, Scr and LYMPHP as pivotal factors for risk stratification in patients with ONIHL. The machine learning model, utilizing XGBoost, effectively distinguishes ONIHL patients among individuals exposed to noise, thereby facilitating early diagnosis and intervention.

Strengths and limitations of this study

The model predicts ONIHL using routine blood and biochemical indicators, eliminating the need for audiometric tests or direct noise exposure data.

Simplifies the diagnostic process, reducing time, costs, and manpower requirements.

Provides an accessible and efficient alternative for early screening and prevention of ONIHL.

The study is limited to the Shenzhen population, and the model's generalizability to other groups and settings remains uncertain.

The positive-to-negative sample ratio exceeds 1:20, mirroring real-world conditions but limiting predictive accuracy; future integration of additional biomarkers, such as DNA methylation, may improve performance.

INTRODUCTION

Occupational noise-induced hearing loss (ONIHL) is characterized as a progressive sensorineural hearing impairment predominantly attributed to damage of the hair cells within the inner ear, consequent to prolonged exposure to high-intensity noise environments[1]. As reported by the World Health Organization (WHO), approximately 10% of the global workforce is impacted by elevated noise levels, with occupational noise exposure accounting for 7% to 21% of hearing loss among workers[2]. A national occupational research agenda says that ONIHL has the highest prevalence of occupational diseases in the United States[3]. About 22 million U.S. workers are currently exposed to hazardous occupational noise[4]. This incidence is notably higher in developing countries[5]. As the largest developing nation, China has witnessed an increasing trend in the incidence of occupational ONIHL in recent years. The prevalence of ONIHL has been reported to be over 20% among noise-exposed workers in China[6]. Such hearing loss can result in communication challenges, social isolation, loneliness, and depression, thereby adversely impacting patients' quality of life and leading to indirect economic losses for society[7]. However, despite being a major global public health issue, early screening methods for ONIHL remain limited.

Currently, pure-tone audiometry is regarded as the gold standard for diagnosing ONIHL[8]. However, its reliance on costly audiological equipment and the necessity for highly trained professionals restrict its practicality for large-scale ONIHL screening among noise-exposed occupational groups[9]. Additionally, PTA relies on subjective

auditory feedback and may be influenced by individual auditory adaptation. Consequently, there is a pressing need to develop a practical and user-friendly screening tool specifically designed for ONIHL patients to prevent the advancement to clinically significant ONIHL. Numerous instances of ONIHL are characterized by an initial deterioration in high-frequency hearing, which gradually progresses to impairments in low-frequency or speech frequency hearing[10]. The early identification of individuals at high risk is essential for effective prevention and intervention strategies.

Consequently, the development of predictive models to screen high-risk populations for further evaluation represents a viable alternative approach. The growing volume of data has facilitated the application of machine learning techniques in the context of ONIHL. At present, a variety of methodologies employing either traditional statistical analysis or machine learning techniques are utilized to predict the risk of ONIHL. These methodologies frequently necessitate substantial human resources and present challenges in manual definition[11]. The integration of machine learning (ML) within the field of audiology has demonstrated potential, particularly in its capacity to effectively analyze nonlinear relationships within data, such as forecasting hearing thresholds for individuals exposed to specific risk factors[12]. Abdollahi[13] constructed eight ML models to forecast sensorineural hearing loss following radiotherapy and chemotherapy, with five of these models demonstrating accuracy and precision exceeding 70%. Comparable levels of accuracy have been reported in other investigations employing ML models to predict sudden sensorineural hearing loss (SSNHL) and ototoxic hearing loss[14,15]. Additionally, various studies have

documented accuracy rates between 0.64 and 0.99 when utilizing diverse ML algorithms and input parameters to predict ONIHL risk factors[16–20]. Among the diverse array of machine learning techniques, SVM models, RF models, and XGBoost models have demonstrated superior performance in classification tasks[10]. Although these studies demonstrate that machine learning (ML) can effectively predict various types of hearing loss, most existing models primarily rely on audiometric data rather than non-invasive biomarkers.

Established risk factors for ONIHL encompass age, medical history (including conditions such as hypertension and diabetes), history of noise exposure, tinnitus and behavioral factors such as smoking and physical activity[21–24]. Furthermore, several biomarkers associated with inflammation, including elevated levels of white blood cells (WBC), neutrophils (NE), monocytes (MO), and lymphocytes (LY), alongside metabolic parameters such as low-density lipoprotein (LDL) and high-density lipoprotein (HDL), are recognized as risk indicators for hearing loss[25]. The chronic alterations in the inflammatory state that occur with aging, a phenomenon known as inflammaging, may contribute to or expedite long-term auditory system damage[26]. Red cell distribution width (RDW), a parameter traditionally utilized for the classification of anemia, has recently been identified as being associated with inflammation and microcirculatory disorders[27]. HDL and LDL have been reported to influence blood supply, thereby potentially affecting sudden sensorineural hearing loss[25]. While numerous studies have explored the relationship between hearing loss and various blood inflammatory and metabolic parameters, there is a paucity of research

employing these parameters to predict ONIHL.

It is noteworthy that individuals exposed to occupational noise are subject to annual medical evaluations, which routinely include blood tests comprising both standard and biochemical analyses[15]. Physicians often extract limited information from these routine blood test results. In light of this, our study seeks to comprehensively leverage routine hematological and biochemical indicators, in conjunction with machine learning methodologies, to construct a risk prediction model for ONIHL. The objective is to facilitate early detection and intervention for ONIHL utilizing data from standard medical examinations.

METHODS

Data Collection and Processing

The medical examination data was obtained from the Shenzhen Prevention and Treatment Center for Occupational Diseases from January 2023 to July 2024. The data was divided into two parts in chronological order, named D1 and D2. The first step involved data cleaning, removing samples with erroneous or abnormal values. The inclusion criteria were formulated based on the GBZ49-2014 *Diagnosis of Occupational Noise-Induced Hearing Loss*: (1) Noise exposure duration ≥ 3 years; (2) Bilateral high-frequency (3000 Hz, 4000 Hz, 6000 Hz) average hearing threshold ≥ 40 dB. Exclusion criteria included pseudohypacusis, exaggerated hearing impairment, drug-induced hearing loss, traumatic hearing loss, infectious hearing loss, hereditary hearing loss, Ménière's disease, sudden deafness, acoustic neuroma, and auditory

neuropathy. We divided the samples into two groups: the occupational noise-induced hearing loss group and the noise-exposed normal hearing group. After preprocessing, a total of 3,297 samples were retained, with D1 and D2 consisting of 2,868 and 429 samples, respectively. Among them, there were 107 and 53 cases of noise-induced hearing loss, representing the positive samples. We then applied random sampling to split D1 into a training set and a test set at a 7:3 ratio. D2 was used as an independent test set.

All datasets included the following variables: sex, age, total protein (TP), albumin (ALB), glucose (GLU), cholesterol (CHO), triglycerides (TG), high-density lipoprotein (HDL), low-density lipoprotein (LDL), total bilirubin (TBIL), direct bilirubin (DBIL), indirect bilirubin (IBIL), alanine aminotransferase (ALT), aspartate aminotransferase (AST), blood urea nitrogen (BUN), serum creatinine (Scr), uric acid (UA), globulin (GLB), hemoglobin (Hb), red blood cell count (RBC), hematocrit (HCT), mean corpuscular volume (MCV), mean corpuscular hemoglobin (MCH), mean corpuscular hemoglobin concentration (MCHC), white blood cell count (WBC), eosinophil count (EOC), basophil count (BAC), lymphocyte count (LYMPHC), monocyte count (MOC), platelet count (PLT), neutrophil count (GRANC), eosinophil percentage (EOP), basophil percentage (BAP), red cell distribution width (RDW-CV), mean platelet volume (MPV), platelet distribution width (PDW), plateletcrit (PCT), neutrophil percentage (GRANP), lymphocyte percentage (LYMPHP), monocyte percentage (MOP), red cell distribution width (RDW-SD), platelet-to-HDL ratio (PLT/HDL), glucose-to-HDL ratio (GLU/HDL), platelet-to-lymphocyte ratio (PLT/LYMPHC),

albumin-to-globulin ratio (A/G), neutrophil-to-lymphocyte ratio (S/L), triglyceride-glucose index (TyG), and estimated glomerular filtration rate (eGFR) (The calculation formulas for TyG and eGFR are detailed in Additional file 1).

In light of the pronounced class imbalance present across all datasets, we employed oversampling of the positive instances within the training set utilizing the `ovun.sample()` function from the ROSE package. This function randomly replicates samples from the minority class, thereby equalizing the number of positive and negative samples in the training set and achieving a balanced class distribution[28]. This approach effectively increases the sample size of the minority class, mitigating the effects of class imbalance during model training. All datasets underwent Z-score normalization, utilizing the mean and standard deviation derived from the training set data.

Framework

Employing occupational health examination data, we introduce an integrated framework for the identification of patients with noise-induced hearing loss, as illustrated in Figure 1. Initially, we preprocessed two datasets, designated as D1 and D2. Dataset D1 was partitioned into training and validation subsets in a 7:3 ratio, while dataset D2 served as an independent test set for the evaluation of the final model. Due to the class imbalance present in the dataset, we employed an oversampling technique on the training set. Subsequently, we utilized a comprehensive array of machine learning algorithms, including XGBoost, Logistic Regression (LR), Random Forest

(RF), Support Vector Machine (SVM), and K-Nearest Neighbor (KNN), to construct predictive models. We then applied feature selection methods to the most optimal predictor among the five to enhance the tool's feasibility. The performance of the refined model was evaluated using an independent test set. we conducted a feature importance analysis to identify variables correlated with the incidence of noise-induced hearing loss. Additionally, we optimized the model to select the most representative variables and employed DCA to evaluate the net benefit of the model across various threshold levels.

Model Construction

In order to construct predictive models, we employed five machine learning algorithms: LR, RF, SVM, KNN, and XGBoost. LR is a form of linear regression that utilizes the Sigmoid function to convert outputs into probabilities for classification purposes[29]. RF comprises an ensemble of independently trained decision trees, with the ultimate prediction being derived through a voting mechanism among these trees, thereby mitigating the risk of overfitting[30]. SVM algorithm classifies samples by identifying an optimal hyperplane within the feature space, and it is capable of managing nonlinearly separable data[31]. KNN algorithm, an instance-based learning method, classifies samples according to the proximity of their k nearest neighbors, making it particularly suitable for small datasets and straightforward to implement[32]. XGBoost is an ensemble method based on decision trees that enhances model performance through a gradient boosting framework. It constructs decision trees in an iterative

manner to minimize model error, demonstrating particular efficacy in handling large-scale, high-dimensional datasets due to its robust generalization capabilities and computational efficiency[33]. All models were developed in R (v4.3.1) using a standardized 5-fold cross-validation framework, with performance evaluated by the Area Under the ROC Curve (AUC). Hyperparameter optimization was performed via grid search to maximize validation AUC, supported by a heatmap illustrating key parameter interactions in XGBoost (Additional file 1: Figure S1) and boxplots comparing cross-validation stability across models (Additional file 1: Figure S2). For XGBoost, critical parameters included tree depth (max_depth), learning rate (eta), and subsampling ratios, optimized to max_depth=7, eta=0.1, and subsampling ratios of 0.6. RF, SVM and KNN employed targeted tuning strategies—such as feature subset selection, regularization balancing, and dynamic neighbor selection—while LR utilized L2 regularization. To ensure reproducibility, data splitting and randomization were controlled by a global seed (set.seed(123)), with parallel processing (4 threads) accelerating computations. And we chose the model with the best performance on validation set for further optimization.

Model Evaluation

To evaluate model performance, considering the class imbalance in the validation and test sets, we used the following metrics to comprehensively assess model performance: sensitivity, specificity, balanced accuracy, AUC, PR-AUC, F1-score, and precision. These metrics are defined as follows:

$$\text{Sensitivity} = \text{Recall} = \text{TPR} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \text{TNR} = \frac{TN}{TN + FP}$$

$$\text{Balanced Accuracy} = \frac{TPR + TNR}{2}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{F1 score} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

The performance of all models was assessed using the 'pROC' package in R to calculate AUC and PR-AUC values.

TP, that is, true positive, is the number of cases of noise-induced hearing loss. *FP*, false positive, denotes the number of normal subjects incorrectly predicted as having ONIHL. *TN*, True Negative, indicates the number of healthy subjects correctly classified as normal. *FN*, False Negative, refers to the number of cases with ONIHL incorrectly classified as normal. And all above metrics range from 0 to 1.

Feature Selection and Feature Importance Analysis

Despite the relatively high performance of the prediction model utilizing 48 features, there remains the possibility of redundant information or noise features that could adversely affect the decision-making process. To enhance the effective utilization of features and streamline the model, we employed a combination of manual curation, Principal Component Analysis (PCA), and Maximum Relevance Minimum Redundancy (mRMR) methods to extract essential features for the final model[34]. In the manual curation process, we initially identified features that exhibited significant

differences between positive and negative samples. To improve the stability of the predictive model, we eliminated features that contributed to significant collinearity[35]. As a result, 16 features were retained. To ensure consistency, the number of feature subsets was also fixed at 16 during the application of PCA and mRMR analysis. PCA selected principal components based on cumulative explained variance, retaining those accounting for up to 80% of the variance to balance dimensionality reduction and information preservation. Meanwhile, mRMR leveraged mutual information to maximize feature relevance while minimizing redundancy, ensuring an optimal feature subset. Furthermore, feature selection was conducted on the training set to mitigate the risk of overfitting. The analysis of feature importance facilitates the interpretation of the predictive model and aids in identifying the features most closely associated with ONIHL. In this context, Feature importance was assessed using XGBoost's weight coefficients, with Gain (SHAP-based Importance) highlighting features that maximally improve model performance while minimizing redundancy.

RESULTS

We initially gathered occupational health examination data from the Shenzhen Occupational Disease Prevention and Control Institute for the period spanning 2023 to 2024, with subgroup D1 comprising 2,868 noise-exposed workers. Of these, 107 participants were diagnosed with ONIHL. Table 1 provides a detailed description of the characteristics of both noise-exposed individuals and ONIHL patients. The five most prominent features exhibiting significant differences between the ONIHL and non-

ONIHL samples include ALB, TP, Age, RDW-CV, PDW (Additional file 1: Figure S3).

Table 1 Statistical Characteristics of Noise-Exposed Hearing Normal Individuals and ONIHL Patients

Characteristics	**Control**	**Case**	***p***
N	2761	107	
Sex:			<0.001**
female	25 (0.91%)	21 (19.6%)	
male	2736 (99.1%)	86 (80.4%)	
Age, year	38.5 ±7.65	43.5 ± 7.17	<0.001**
TP, g/L	72.7 ± 3.91	68.1 ± 4.69	<0.001**
ALB, g/L	46.7 ± 2.59	42.7 ± 2.95	<0.001**
GLU, mmol/L	5.23 ± 0.63	5.42 ± 0.86	0.023*
CHO, mmol/L	4.89 ± 0.86	4.90 ± 1.02	0.930
TG, mmol/L	1.64 ± 1.21	1.97 ± 1.28	0.011*
LDL, mmol/L	1.32 ± 0.26	1.31 ± 0.31	0.717
HDL, mmol/L	3.03 ± 0.62	3.02 ± 0.77	0.833
TBIL, μmol/L	16.5 ± 6.26	15.7 ± 6.02	0.188
DBIL, μmol/L	3.02 ± 1.24	2.80 ± 1.16	0.061
IBIL, μmol/L	13.4 ± 5.20	12.6 ± 4.90	0.098
ALT, U/L	29.5 ± 22.1	27.2 ± 21.8	0.294
AST, U/L	24.7 ± 9.79	27.5 ± 42.3	0.503
BUN, mmol/L	5.04 ± 1.18	5.06 ± 5.07	0.966
Scr, μmol/L	85.1 ± 11.2	76.7 ± 16.9	<0.001**
UA, μmol/L	401 ± 79.2	482 ± 893	0.351
GLB, g/L	26.0 ± 3.36	25.4 ± 3.37	0.086
Hb, g/L	154 ± 10.1	144 ± 17.9	<0.001**
RBC, $\times 10^{12}$/L	5.02 ± 0.37	4.84 ± 0.49	<0.001**
HCT, L/L	0.45 ± 0.03	0.42 ± 0.04	<0.001**
MCV, fL	89.6 ± 4.77	87.6 ± 8.47	0.015*
MCH, pg	30.8 ± 1.86	29.9 ± 3.47	0.007**
MCHC, g/L	344 ± 6.82	341 ± 12.3	0.013*
WBC, $\times 10^9$/L	6.09 ± 1.53	6.81 ± 1.61	<0.001**
EOC, $\times 10^9$/L	0.16 ± 0.14	0.21 ± 0.14	<0.001**
BAC, $\times 10^9$/L	0.03 ± 0.02	0.03 ± 0.02	0.019*
LYMPHC, $\times 10^9$/L	2.10 ± 0.57	2.12 ± 0.53	0.615
MOC, $\times 10^9$/L	0.37 ± 0.12	0.44 ± 0.13	<0.001**
PLT, $\times 10^9$/L	241 ± 50.4	252 ± 63.3	0.064
GRANC, $\times 10^9$/L	3.44 ± 1.15	4.00 ± 1.15	<0.001**
EOP, %	2.59 ± 1.91	3.07 ± 1.72	0.005**
BAP, %	0.51 ± 0.27	0.38 ± 0.23	<0.001**

RDW-CV, %	12.9 ± 0.65	13.9 ± 1.54	<0.001**
MPV, fL	10.2 ± 1.05	10.2 ± 1.07	0.758
PDW, fL	16.2 ± 0.35	14.9 ± 1.97	<0.001**
PCT, %	0.24 ± 0.04	0.26 ± 0.06	0.041*
GRANP, %	55.8 ± 7.74	58.4 ± 6.22	<0.001**
LYMPHP, %	35.0 ± 7.32	31.5 ± 5.71	<0.001**
MOP, %	6.09 ± 1.41	6.46 ± 1.35	0.006**
RDW-SD, %	41.9 ± 1.96	44.1 ± 3.91	<0.001**
PLT/HDL	189 ± 55.0	202 ± 65.5	0.046*
GLU/HDL	4.11 ± 0.98	4.38 ± 1.32	0.038*
PLT/LYMPHC	122 ± 38.0	124 ± 37.6	0.571
A/G	1.83 ± 0.28	3.27 ± 16.3	0.361
S/L	1.02 ± 0.43	2.09 ± 10.4	0.289
TyG	8.67 ± 0.56	8.87 ± 0.63	0.001**
eGFR, mL/(min×1.73m²)	91.8 ± 3.84	92.3 ± 5.18	0.349

Notes: Data are presented as N (%) or Mean ± SE. P-values were based on chi-square tests (χ^2-test) or t-test. *p<0.05, **p<0.01.

Performance Comparison of the Five Machine Learning Methods

Table 2 and Additional file 1: Figure S4 A and B show that the XGBoost algorithm has the highest AUC (0.942) and PR-AUC (0.791), as well as high Recall (0.875) and Balanced accuracy (0.905). The F1-Score is only second to that of RF, making its overall performance excellent. The RF algorithm also performs well, with AUC (0.921) and PR-AUC (0.690), and a high Balanced Accuracy (0.872), indicating an outstanding overall performance.To maximize the identification of ONIHL patients, the XGBoost algorithm was ultimately selected to further build the prediction model.

Table 2 Performance Comparison of the Five Machine Learning Methods

	AUC	PR-AUC	Recall	Precision	Balanced accuracy	F1-sore
Logistic Regression	0.923	0.683	0.938	0.200	0.896	0.330

(LR)						
Random Forest (RF)	0.921	0.690	0.781	0.446	0.872	0.568
Support Vector Machine (SVM)	0.797	0.369	0.750	0.140	0.786	0.235
k-Nearest Neighbors (KNN)	0.829	0.521	0.688	0.333	0.817	0.449
XGBoost	0.942	0.791	0.875	0.346	0.905	0.496

The values of AUC, PR-AUC, Recall, Precision, Balanced Accuracy, and F1-Score range from 0 to 1, with higher values indicating better performance. LR stands for Logistic Regression, RF for Random Forest, SVM for Support Vector Machine, KNN for K-nearest Neighbors, and XGBoost for Extreme Gradient Boosting. AUC represents the area under the receiver operating characteristic curve, and PR-AUC represents the area under the precision-recall curve.

The results of the five-fold cross-validation on the training set show an AUC of 0.999, Sensitivity of 0.995, and Balanced Accuracy of 0.998. Additionally, the XGBoost model demonstrates reliable performance on the test set (AUC = 0.900, PR-AUC = 0.648), as shown in Figure 2 A and B.

Feature Selection for the Final Model

Several pairs of features were observed to have high correlations, such as MCH and RDW-CV, and GRANP and LYMPHP, which may introduce redundant information and affect the model's decision-making and stability (Additional file 1: Figure S5 for related heat maps). Therefore, we used manual curation, PCA, and mRMR methods to identify the optimal features. As a result, 16 features were used to reconstruct the XGBoost model from each of manual curation, PCA, and mRMR (Table S1). The PCA and mRMR feature selection methods identified seven shared features, with three of the

top five selected features overlapping: ALB, RDW-CV, and Scr. The model constructed using 16 features selected by mRMR and PCA showed a slight improvement on the validation set compared to the model built with all 48 features. Specifically, the PCA model achieved an AUC of 0.957 and a PR-AUC of 0.741, while the mRMR model had an AUC of 0.957 and a PR-AUC of 0.720. In contrast, the model based on manual feature selection exhibited a decline in performance, with an AUC of 0.919 and a PR-AUC of 0.540 (Figure 3A, B). Similarly, the models constructed using mRMR and PCA demonstrated improvements in sensitivity and balanced accuracy on the validation set, with maximum increases of 29.2% and 3%, respectively. In comparison, the model based on manual curation showed suboptimal performance across all evaluation metrics (Figure 3C). We further evaluated the models using an independent test set (D2). In this test set, the manual curation model achieved an AUC of 0.830, the PCA model had an AUC of 0.837, and the mRMR model outperformed both with an AUC of 0.872 (Figure 3D). The PR-AUC values were 0.524, 0.540, and 0.594 for the manual curation, PCA, and mRMR models, respectively, with mRMR again demonstrating the best performance (Figure 3E). Regarding sensitivity, specificity, and balanced accuracy, the mRMR model exhibited the highest performance in the test set evaluation (Figure 3F). Notably, the lowest sensitivity observed for the mRMR model on D2 was 75.5%, while all specificity scores remained above 78.0%. Overall, the model demonstrated strong performance on the independent test set, indicating that the selected core features are sufficient for detecting noise-induced hearing loss among noise-exposed workers.

Feature Importance Ranking

To investigate which features contribute the most to the risk of ONIHL, we first used mRMR to select 16 important features and then built an XGBoost model based on these features. Subsequently, we ranked these features according to their weights in the XGBoost model, as shown in Figure 4. And the feature importance of the predictors based on PCA and manual curation is shown in Additional file 1: Figure S6, S7. The results indicated that the top five features, in order of importance, were ALB, PDW, RDW-CV, Scr, LYMPHP. Further comparisons between the ONIHL and normal samples revealed significant differences in ALB, TP, Age, RDW-CV and PDW. These findings are highly consistent with the top-ranked results in the XGBoost model, indicating a strong correlation between these indicators (ALB, PDW, RDW-CV, Scr and LYMPHP) and ONIHL.

DCA Decision Curve Analysis

Decision curve analysis (Figure 5) revealed distinct net benefit patterns across threshold probabilities for the three models. The mRMR model demonstrated superior performance, achieving the highest net benefit over a broad threshold range (0.0–0.6), with particularly pronounced advantages at lower thresholds (0.0–0.2). In contrast, the PCA model exhibited competitive efficacy within the moderate threshold interval (0.2–0.4). Notably, both the manual curation model and the "All/None" strategy underperformed: manual curation yielded consistently lower net benefits across all thresholds, and "All/None" resulted in negative net benefits at thresholds below 0.3, indicating clinical impracticality. These findings support a threshold-adaptive selection

strategy: prioritizing mRMR for thresholds ⩽0.4 (maximizing robustness) and PCA for 0.2–0.4 thresholds (balancing accuracy and efficiency). This approach optimizes clinical utility by aligning model strengths with context-specific decision risks.

DISCUSSION

ONIHL represents a significant global public health concern[2,36]. Despite its complexity, ONIHL is a preventable condition[37,38]. The Occupational Safety and Health Administration (OSHA) requires the implementation of hearing conservation programs for workers exposed to noise levels of 85 decibels or higher, with the objective of safeguarding auditory health in noisy occupational environments[39]. Consequently, the development of a risk screening tool for ONIHL is crucial as a primary strategy for screening and prevention among workers exposed to occupational noise. In this study, we employed five ML algorithms utilizing hematological test results to construct an ONIHL risk screening model. The models demonstrated AUC values exceeding 0.85, with accuracy and sensitivity surpassing 0.75 in both validation and independent test datasets. These results suggest that ML models are capable of accurately identifying ONIHL patients within the population of noise-exposed workers.

In an evaluation of model performance on the validation set, the XGBoost model exhibited superior efficacy compared to all other algorithms assessed, achieving an AUC of 0.942 and a PR-AUC of 0.791. The precision, specificity, F-score, and balanced accuracy metrics for the XGBoost model all exceeded 0.8 on the validation set. Furthermore, the XGBoost model maintained consistent performance on the test set,

with an AUC of 0.900 and a PR-AUC of 0.648. XGBoost is recognized as a machine learning technique that efficiently and flexibly manages missing data and integrates weak predictive models into a robust predictive framework[40]. As an open-source package, XGBoost has gained significant recognition in various machine learning and data mining competitions. For example, in 2015, 17 out of the 29 winning solutions featured on Kaggle's blog utilized XGBoost, and all of the top 10 winning teams in the 2015 KDD Cup also incorporated XGBoost into their solutions[41]. In neurology, XGBoost achieved AUC values of 0.950 (mortality) and 0.958 (functional outcomes) in aneurysmal subarachnoid hemorrhage patients, outperforming logistic regression[42]. XGBoost has been applied to predict 5-year survival in elderly intrahepatic cholangiocarcinoma patients (AUC=0.713, SEER database) and Type 2 diabetes risk (accuracy=89.09%, AUC=0.9182 in Beijing residents)[43,44]. These advancements highlight XGBoost's utility in high-dimensional clinical datasets with interpretable feature insights. Furthermore, our findings indicate that the predictive efficacy of the XGBoost model surpasses that of LR, RF, SVM, and KNN. This aligns with previous research demonstrating that traditional logistic regression frequently exhibits comparatively lower AUC values in ROC curve analyses, alongside higher prediction errors and inferior performance relative to more contemporary methodologies[45,46].

Screening for ONIHL is important, and various methods have been explored for this purpose. Otoacoustic emissions (OAE) testing, particularly distortion product otoacoustic emissions (DPOAE), is a sensitive tool for detecting early cochlear damage

before significant hearing threshold shifts appear in pure-tone audiometry (PTA)[47]. It can identify subtle outer hair cell dysfunction in noise-exposed individuals, making it valuable for early intervention and monitoring[48]. Auditory brainstem response (ABR) testing, another physiological method, assesses neural integrity and can detect hidden hearing loss even when audiometric thresholds remain normal[49]. Despite their advantages, the large-scale application of OAE and ABR in occupational screening is limited by high costs, equipment availability, and the need for trained operators. In contrast, our model predicts ONIHL risk solely from routine blood and biochemical indicators, eliminating the need for specialized audiometric assessments or noise exposure data. By analyzing markers linked to inflammation, oxidative stress, and immune response, it provides a cost-effective, scalable alternative for early screening. Integrating this approach with existing methods like OAE or ABR could further enhance ONIHL risk assessment, enabling earlier interventions before irreversible damage occurs.

Consequently, the early identification and intervention of risk factors identified in our model could have substantial implications for the prevention of ONIHL among workers exposed to noise. The risk factors contributing to the development of ONIHL are varied. We have developed a risk assessment model for ONIHL utilizing clinical data and routine physical examination indicators, employing a machine learning algorithm. This approach contrasts with most existing methods for predicting ONIHL risk, which predominantly depend on variables such as age, sex, medical history (including conditions like hypertension and diabetes), history of noise exposure, and

behavioral factors such as smoking and physical activity[16,50,51]. For instance, prior research has developed risk models for workers exposed to noise, yielding favorable predictive outcomes. These models primarily incorporate risk factors such as industry type, duration of noise exposure, and median peak intensity, which contrast with the physical examination indicators utilized in our study[20]. Yi Wang[10] formulated a machine learning-based risk assessment model for high-frequency hearing loss employing routine physical examination data, attaining AUC of 0.868. This model, however, was principally designed for community residents and incorporated risk factors including 13 blood test indicators, demographic characteristics, disease-related features, behavioral factors, environmental exposure, and auditory cognitive factors, which differ from the population of noise-exposed workers in our study. Our model offers a more comprehensive approach than previous research by integrating a wide range of biochemical and Routine Blood indicators to assess the risk of ONIHL from multiple dimensions. Unlike models that rely on hearing assessments and direct noise exposure measurements, our model focuses on routine blood and biochemical indicators, reducing the need for specialized equipment and resources. This makes it a more efficient, cost-effective alternative for early detection and prevention of ONIHL, offering personalized risk assessments without the reliance on extensive testing.

Routine blood tests administered at occupational disease prevention clinics are typically conducted on an annual basis. Based on these tests, the application of these indicators can enhance early screening and provide warnings for prevalent occupational diseases. In our study, the developed model demonstrates the significance of

hematological test data in screening for ONIHL. This includes variables such as age, sex, inflammatory and immune markers (e.g., WBC, LYMPHP, MOC, BAP, EOC and GRANC), as well as oxidative stress and metabolic markers (e.g., ALB, Scr, RDW-CV and RDW-SD). Noise exposure influences hematological parameters through complex immunoinflammatory pathways, which may both reflect and exacerbate cochlear damage. Studies have shown that ONIHL is closely associated with systemic immune and inflammatory responses, with WBC serving as a key inflammatory marker linked to ONIHL. A study analyzing health examination data from 3,508 noise-exposed workers found that WBC levels were significantly higher in the NIHL group compared to those with normal hearing[52]. This suggests that noise exposure may trigger chronic inflammatory responses in the body. At the cellular level, noise activates resident macrophages in the cochlea, triggering the release of pro-inflammatory cytokines such as IL-1β and TNF-α[54]. This leads to increased permeability of the blood-labyrinth barrier, facilitating the infiltration of systemic immune cells—including MOC, GRANC, and adaptive immune lymphocytes—into the inner ear[55]. This immune influx amplifies local inflammation, creating a microenvironment that promotes sensory cell apoptosis and spiral ganglion degeneration. Notably, our study identified significant elevations in WBC, MOC, and GRANC, aligning with these immunological responses. Chronic noise stress further disrupts systemic immune homeostasis, as demonstrated in animal models[55]. Prolonged exposure induces immunosuppressive changes, including a decrease in LYMPHP and a reduced CD4+/CD8+ T cell ratio, which may impair anti-inflammatory responses and regenerative capacity. This

systemic immune imbalance is consistent with our findings of decreased LYMPHP in noise-exposed individuals. Additionally, our findings indicate that increased EOC levels may also serve as a risk factor for ONIHL. EOC may play a pathogenic role in noise-induced inner ear vasculitis, a process increasingly recognized as a critical mediator of sensorineural damage. Elevated EOC levels have been associated with SSNHL, suggesting their potential as prognostic indicators in inflammatory hearing disorders[56].

Oxidative stress is a key mechanism underlying ONIHL[57]. RDW-CV/SD is a crucial marker of red blood cell oxidative damage, with increased RDW levels indicating decreased membrane stability[58,59]. This instability shares a common pathological basis with noise-induced hair cell apoptosis. A positive correlation between RDW parameters (CV and SD) and the average hearing threshold further highlights the role of oxidative stress in ONIHL[60]. These findings underscore the need to identify inflammatory conditions when screening workers at risk for chronic inflammation and ONIHL. ALB, a critical antioxidant protein, plays a protective role in maintaining blood-labyrinth barrier integrity[61]. Low ALB levels may weaken antioxidant defenses and increase susceptibility to hearing loss. Notably, ALB levels in patients with SSNHL were significantly lower than those in the control group ($p < 0.001$)[62]. Furthermore, studies have shown a positive correlation between reduced eGFR and hearing loss, suggesting that impaired kidney function may contribute to cochlear microcirculatory dysfunction via inflammation-mediated mechanisms[63]. Since Scr is a key component in calculating eGFR, our observed reduction in Scr could

reflect a broader metabolic or physiological shift rather than direct renal impairment alone. PDW, an indicator of platelet activation, may also play a role in ONIHL by promoting microvascular inflammation. Research has demonstrated a significant association between PDW and the severity of SSNHL[64]. This finding is consistent with our study, as we also observed an association between PDW and ONIHL. Given that PDW is an indicator of platelet activation, its potential role in promoting microvascular inflammation may contribute to the pathophysiological mechanisms underlying ONIHL. Our results further support the notion that vascular and inflammatory responses play a crucial role in noise-induced cochlear damage. Age and male gender have been identified as risk factors for hearing loss[65]. Leveraging artificial intelligence and big data analysis, hematological parameters can serve as predictive markers for ONIHL. A machine learning model based on XGBoost integrates inflammatory, oxidative stress, and metabolic-related indicators to enhance risk assessment. Feature importance analysis highlights ALB, PDW, RDW-CV, Scr, and LYMPHP as key predictors of ONIHL, reinforcing their potential role in early detection and risk stratification.

Although hematological indicators provide a low-cost and accessible approach for ONIHL prediction, their specificity remains limited, necessitating integration with objective auditory assessments such as ABR and OAE to enhance predictive accuracy. Additionally, the generalizability of our model requires further validation, as it is currently based on a Shenzhen population and may not fully represent other demographic and occupational groups. The model's precision and F1 score are also

relatively low, primarily due to the severe class imbalance, with ONIHL cases being far less frequent than noise-exposed individuals with normal hearing. Despite these limitations, future studies can address these challenges by conducting large-scale, multi-center validations, employing advanced data-balancing techniques, and incorporating multi-omics data—such as metabolomics and transcriptomics—to unravel the molecular mechanisms linking inflammation, oxidative stress, and immune dysregulation in ONIHL. Such advancements will not only optimize predictive models but also facilitate their clinical application in occupational health screening and early intervention, ultimately improving hearing loss prevention strategies.

CONCLUSION

In this study, we developed five machine learning models to construct a risk screening model for ONIHL, with the XGBoost-based model demonstrating superior performance. By integrating biochemical and hematological indicators with machine learning techniques, this model effectively identifies individuals at high risk for ONIHL. This approach not only introduces a novel tool for the early screening of hearing loss but also lays the groundwork for the development of personalized intervention strategies. In the future, the integration of additional biological data is anticipated to further augment the model's predictive capabilities. Furthermore, this model holds potential for extension to forecast risks associated with other occupational or chronic diseases, thereby offering substantial support for the maintenance and enhancement of public health.

Abbreviations

ONIHL Occupational noise-induced hearing loss
XGBoost Extreme gradient boosting
LR Logistic regression
RF Random forest
SVM Support vector machines
KNN K-nearest neighbors
DCA Decision curve analysis
ALB Serum albumin
RDW-CV Coefficient of variation in red cell distribution width
LYMPHP Lymphocyte percentage
MOC Monocyte count
RDW-SD Standard deviation in red cell distribution width
ML Machine learning
SSNHL Sudden sensorineural hearing loss
WBC White blood cells
NE Neutrophils
MO Monocytes
LY Lymphocytes
LDL Low-density lipoprotein
HDL High-density lipoprotein
RDW Red cell distribution width
TP Total protein
ALB Albumin
GLU Glucose
CHO Cholesterol
TG Triglycerides
TBIL Total bilirubin
DBIL Direct bilirubin
IBIL Indirect bilirubin
ALT Alanine aminotransferase
AST Aspartate aminotransferase
BUN Blood urea nitrogen
Scr Serum creatinine
UA Uric acid
GLB Globulin
Hb Hemoglobin
RBC Red blood cell count
HCT Hematocrit
MCV Mean corpuscular volume
MCH Mean corpuscular hemoglobin
MCHC Mean corpuscular hemoglobin concentration
EOC Eosinophil count
BAC Basophil count
LYMPHC Lymphocyte count

MOC Monocyte count
PLT Platelet count
GRANC Neutrophil count
EOP Eosinophil percentage
BAP Basophil percentage
MPV Mean platelet volume
PDW Platelet distribution width
PCT Plateletcrit
GRANP Neutrophil percentage
MOP Monocyte percentage
PLT/HDL Platelet-to-HDL ratio
GLU/HDL Glucose-to-HDL ratio
PLT/LYMPHC Platelet-to-lymphocyte ratio
A/G Albumin-to-globulin ratio
S/L Neutrophil-to-lymphocyte ratio
TyG Triglyceride-glucose index
eGFR Estimated glomerular filtration rate
PCA Principal component analysis
mRMR Maximum relevance minimum redundancy
OSHA The Occupational Safety and Health Administration
TNF-α Tumor necrosis factor-alpha
IL-6 Interleukin-6
ROS Reactive oxygen species
OAE Otoacoustic emissions
ABR Auditory brainstem response

Acknowledgment

We thank Shenzhen Prevention and Treatment Center for Occupational Diseases for the approval of the ethical clearance. We also extend our warm gratitude to the different hospital stakeholders and participants for their valuable contribution during data collection.

Author contributions

The authors made substantial contributions to the acquisition, analysis, and interpretation of the data and the drafting and revision of the manuscript. All authors

also approved the final version of the paper and agreed to be accountable for all aspects of the work. Caiping Li and Dianpeng Wang: Writing – original draft, Investigation, Data curation, Conceptualization. Caiping Li, Liuwei Shi and Linlin Chen: Methodology, Data curation. Dafeng Lin: Data curation. Xiangli Yang and Liang Zhou, Investigation. Peimao Li: Validation, Investigation. Wen Zhang: Validation. Yan Guo and Naixing Zhang: Supervision, Project administration, Conceptualization. Dafeng Lin: Writing – original draft, Supervision, Project administration, Formal analysis, Conceptualization. Caiping Li is the guarantor of this manuscript.

Funding

This work was supported by Science and Technology Planning Project of Shenzhen Municipality (No.KCXFZ20201221173602007, No.JCYJ20220531091211026), Shenzhen Fund for Guangdong Provincial High- level Clinical Key Specialties (No.SZGSP015).

Disclaimer

The content of this study is solely the responsibility of the authors and does not necessarily represent the official views of the Science and Technology Planning Project of Shenzhen Municipality or the Shenzhen Fund for Guangdong Provincial High-Level Clinical Key Specialties.

Competing interests

None declared.

Ethics Approval and Consent to Participate

This study was approved by the Ethics Committee of Shenzhen Prevention and

Treatment Center for Occupational Diseases (Approval Number: LL2020-34, Date: 14th December 2020). All methods were carried out in accordance with relevant ethical guidelines and regulations.

Patient consent for publication

Not applicable.

Data availability

Data are available upon reasonable request. Original data collected within this study is not publicly available, as it might contain sensitive information. De-identified data can be shared based on a reasonable request by sending an email to szpcr@126.com.

Patient and Public Involvement

Patients or the public were not involved in the design, conduct, reporting, or dissemination plans of our research.

REFERENCES

1 Ding T, Yan A, Liu K. What is noise-induced hearing loss? *Br J Hosp Med*. 2019;80:525–9. doi: 10.12968/hmed.2019.80.9.525

2 Nelson DI, Nelson RY, Concha-Barrientos M, *et al.* The global burden of occupational noise-induced hearing loss. *Am J Ind Med*. 2005;48:446–58. doi: 10.1002/ajim.20223

3 Themann C, Suter A, Stephenson M. National Research Agenda for the Prevention of Occupational Hearing Loss—Part 1. *Semin Hear*. 2013;34:145–207. doi: 10.1055/s-0033-1349351

4 Themann CL, Masterson EA. Occupational noise exposure: A review of its effects, epidemiology, and impact with recommendations for reducing its burden. *J Acoust Soc Am*. 2019;146:3879. doi: 10.1121/1.5134465

5 D T-V, A A, Gp R. What can we learn from adult cochlear implant recipients with single-sided deafness who became elective non-users? *Cochlear implants international*. 2020;21. doi: 10.1080/14670100.2020.1733746

6 Li YH, Jiao J, Yu SF. Research status of influencing factors of noise-induced hearing loss. *Chin J Occup Dis*. 2014;32:469–73.

7 Vlaming MSMG, MacKinnon RC, Jansen M, *et al.* Automated screening for high-frequency hearing loss. *Ear Hear*. 2014;35:667–79. doi: 10.1097/AUD.0000000000000073

8 Bhatt IS, Washnik N, Torkamani A. Suprathreshold Auditory Measures for Detecting Early-Stage Noise-Induced Hearing Loss in Young Adults. *J Am Acad Audiol*. 2022;33:185–95. doi: 10.1055/s-0041-1740362

9 Cunningham LL, Tucci DL. Hearing Loss in Adults. *N Engl J Med*. 2017;377:2465–73. doi: 10.1056/NEJMra1616601

10 Wang Y, Yao X, Wang D, *et al.* A machine learning screening model for identifying the risk of high-frequency hearing impairment in a general population. *BMC Public Health*. 2024;24:1160. doi: 10.1186/s12889-024-18636-1

11 Rm M, Rc M. Objective auditory brainstem response classification using machine learning. *International journal of audiology*. 2019;58. doi: 10.1080/14992027.2018.1551633

12 Chang Y-S, Park H, Hong SH, *et al.* Predicting cochlear dead regions in patients with hearing loss through a machine learning-based approach: A preliminary study. *PLoS One*. 2019;14:e0217790. doi: 10.1371/journal.pone.0217790

13 Abdollahi H, Mostafaei S, Cheraghi S, *et al.* Cochlea CT radiomics predicts chemoradiotherapy induced sensorineural hearing loss in head and neck cancer patients: A machine learning and multi-variable modelling study. *Phys Med*. 2018;45:192–7. doi: 10.1016/j.ejmp.2017.10.008

14 Tomiazzi JS, Pereira DR, Judai MA, *et al.* Performance of machine-learning algorithms

to pattern recognition and classification of hearing impairment in Brazilian farmers exposed to pesticide and/or cigarette smoke. *Environ Sci Pollut Res Int*. 2019;26:6481–91. doi: 10.1007/s11356-018-04106-w

15 D B, J Y, J M, *et al.* Predicting the hearing outcome in sudden sensorineural hearing loss via machine learning models. *Clinical otolaryngology : official journal of ENT-UK ; official journal of Netherlands Society for Oto-Rhino-Laryngology & Cervico-Facial Surgery*. 2018;43. doi: 10.1111/coa.13068

16 Aliabadi M, Farhadian M, Darvishi E. Prediction of hearing loss among the noise-exposed workers in a steel factory using artificial intelligence approach. *Int Arch Occup Environ Health*. 2015;88:779–87. doi: 10.1007/s00420-014-1004-z

17 Farhadian M, Aliabadi M, Darvishi E. Empirical estimation of the grades of hearing impairment among industrial workers based on new artificial neural networks and classical regression methods. *Indian J Occup Environ Med*. 2015;19:84–9. doi: 10.4103/0019-5278.165337

18 Ys K, Yh C, Oj K, *et al.* The Risk Rating System for Noise-induced Hearing Loss in Korean Manufacturing Sites Based on the 2009 Survey on Work Environments. *Safety and health at work*. 2011;2. doi: 10.5491/SHAW.2011.2.4.336

19 Nawi NM, Rehman MZ, Ghazali MI. Noise-induced hearing loss prediction in Malaysian industrial workers using gradient descent with adaptive momentum algorithm. *International Review on Computers and Software*. 2011;6:740–8.

20 Y Z, J L, M Z, *et al.* Machine Learning Models for the Hearing Impairment Prediction in Workers Exposed to Complex Industrial Noise: A Pilot Study. *Ear and hearing*. 2019;40. doi: 10.1097/AUD.0000000000000649

21 Li P, Pang K, Zhang R, *et al.* Prevalence and risk factors of hearing loss among the middle-aged and older population in China: a systematic review and meta-analysis. *Eur Arch Otorhinolaryngol*. 2023;280:4723–37. doi: 10.1007/s00405-023-08109-3

22 Tsimpida D, Kontopantelis E, Ashcroft D, *et al.* Socioeconomic and lifestyle factors associated with hearing loss in older adults: a cross-sectional study of the English Longitudinal Study of Ageing (ELSA). *BMJ Open*. 2019;9:e031030. doi: 10.1136/bmjopen-2019-031030

23 Baiduc RR, Sun JW, Berry CM, *et al.* Relationship of cardiovascular disease risk and hearing loss in a clinical population. *Sci Rep*. 2023;13:1642. doi: 10.1038/s41598-023-28599-9

24 Soylemez E, Avci I, Yildirim E, *et al.* Predicting noise-induced hearing loss with machine learning: the influence of tinnitus as a predictive factor. *J Laryngol Otol*. 2024;138:1030–5. doi: 10.1017/S002221512400094X

25 Jung Da Jung, Do Jun Young, Cho Kyu Hyang, *et al.* Association between triglyceride/high-density lipoprotein ratio and hearing impairment in a Korean population. *Postgraduate medicine*. 2017;129. doi: 10.1080/00325481.2017.1381538

26 Verschuur CA, Dowell A, Syddall HE, *et al.* Markers of inflammatory status are associated with hearing threshold in older people: findings from the Hertfordshire ageing study. *Age and Ageing*. 2012;41:92–7. doi: 10.1093/ageing/afr140

27 Nonoyama H, Tanigawa T, Shibata R, *et al.* Red blood cell distribution width predicts prognosis in idiopathic sudden sensorineural hearing loss. *Acta Oto-Laryngologica*. 2016;136:1137–40. doi: 10.1080/00016489.2016.1195919

28 Lunardon N, Menardi G, Torelli N. ROSE: A Package for Binary Imbalanced Learning. *The R Journal*. 2014;6:79–89.

29 Elkahwagy DMAS, Kiriacos CJ, Mansour M. Logistic regression and other statistical tools in diagnostic biomarker studies. *Clin Transl Oncol*. 2024;26:2172–80. doi: 10.1007/s12094-024-03413-8

30 Schauberger G, Klug SJ, Berger M. Random forests for the analysis of matched case-control studies. *BMC Bioinformatics*. 2024;25:253. doi: 10.1186/s12859-024-05877-5

31 Valkenborg D, Rousseau A-J, Geubbelmans M, *et al.* Support vector machines. *Am J Orthod Dentofacial Orthop*. 2023;164:754–7. doi: 10.1016/j.ajodo.2023.08.003

32 Garcia-Carretero R, Vigil-Medina L, Mora-Jimenez I, *et al.* Use of a K-nearest neighbors model to predict the development of type 2 diabetes within 2 years in an obese, hypertensive population. *Med Biol Eng Comput*. 2020;58:991–1002. doi: 10.1007/s11517-020-02132-w

33 Bridgelall R, Tolliver DD. Railroad accident analysis using extreme gradient boosting. *Accident Analysis & Prevention*. 2021;156:106126. doi: 10.1016/j.aap.2021.106126

34 Xia Zhiming, Chen Yang, Xu Chen. Multiview PCA: A Methodology of Feature Extraction and Dimension Reduction for High-Order Data. *IEEE transactions on cybernetics*. Published Online First: 2021vo PP.

35 Peng Hanchuan, Long Fuhui, Ding Chris. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE transactions on pattern analysis and machine intelligence*. 2005;27.

36 Mariola Śliwińska-Kowalska, Kamil Zaborowski. WHO Environmental Noise Guidelines for the European Region: A Systematic Review on Environmental Noise and Permanent Hearing Loss and Tinnitus. *International Journal of Environmental Research and Public Health*. 2017;14. doi: 10.3390/ijerph14101139

37 Seixas NS, Neitzel R, Stover B, *et al.* A multi-component intervention to promote hearing protector use among construction workers. *Int J Audiol*. 2011;50:S46–56. doi: 10.3109/14992027.2010.525754

38 Amjad-Sardrudi Hossein, Dormohammadi Ali, Golmohammadi Rostam, *et al.* Effect of noise exposure on occupational injuries: a cross-sectional study. *Journal of research in health sciences*. 2012;12.

39 Park S, Johnson MD, Hong O. Analysis of Occupational Safety and Health

Administration (OSHA) noise standard violations over 50 years: 1972 to 2019. *American J Industrial Med*. 2020;63:616–23. doi: 10.1002/ajim.23116

40 Yuan K-C, Tsai L-W, Lee K-H, *et al.* The development an artificial intelligence algorithm for early sepsis diagnosis in the intensive care unit. *International Journal of Medical Informatics*. 2020;141:104176. doi: 10.1016/j.ijmedinf.2020.104176

41 Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco California USA: ACM 2016:785–94.

42 Wang R, Zhang J, Shan B, *et al.* XGBoost Machine Learning Algorithm for Prediction of Outcome in Aneurysmal Subarachnoid Hemorrhage. *Neuropsychiatr Dis Treat*. 2022;18:659–67. doi: 10.2147/NDT.S349956

43 Xu Q, Lu X. Development and validation of an XGBoost model to predict 5-year survival in elderly patients with intrahepatic cholangiocarcinoma after surgery: a SEER-based study. *Journal of Gastrointestinal Oncology*. 2022;13. doi: 10.21037/jgo-22-1238

44 Wang L, Wang X, Chen A, *et al.* Prediction of Type 2 Diabetes Risk and Its Effect Evaluation Based on the XGBoost Model. *Healthcare*. 2020;8:247. doi: 10.3390/healthcare8030247

45 Xiao J, Ding R, Xu X, *et al.* Comparison and development of machine learning tools in

the prediction of chronic kidney disease progression. *J Transl Med*. 2019;17:119. doi: 10.1186/s12967-019-1860-0

46 Li Y-M, Li Z-L, Chen F, *et al.* A LASSO-derived risk model for long-term mortality in Chinese patients with acute coronary syndrome. *J Transl Med*. 2020;18:157. doi: 10.1186/s12967-020-02319-7

47 Soylemez E, and Mujdeci B. Evaluation of auditory disability and cochlear functions in industrial workers exposed to occupational noise. *Hearing, Balance and Communication*. 2021;19:16–20. doi: 10.1080/21695717.2020.1727214

48 Kapoor N, Mani KV, Shukla M. Distortion Product Oto-Acoustic Emission: A Superior Tool for Hearing Assessment Than Pure Tone Audiometry. *Noise Health*. 2019;21:164–8. doi: 10.4103/nah.NAH_37_19

49 Mehraei G, Gallardo AP, Shinn-Cunningham BG, *et al.* Auditory brainstem response latency in forward masking, a marker of sensory deficits in listeners with normal hearing thresholds. *Hearing Research*. 2017;346:34–44. doi: 10.1016/j.heares.2017.01.016

50 Chen F, Cao Z, Grais EM, *et al.* Contributions and limitations of using machine learning to predict noise-induced hearing loss. *Int Arch Occup Environ Health*. 2021;94:1097–111. doi: 10.1007/s00420-020-01648-w

51 Sun R, Shang W, Cao Y, *et al.* A risk model and nomogram for high-frequency hearing

loss in noise-exposed workers. *BMC Public Health*. 2021;21:747. doi: 10.1186/s12889-021-10730-y

52 Yingjun L, Weisen Z, Hao Z, *et al.* Study on the correlation between noise-induced hearing loss and white blood cell count. *International Medicine and Health Guidance News*. 2023;29:115–8. doi: 10.3760/cma.j.issn.1007-1245.2023.01.025

53 Frye MD, Ryan AF, Kurabi A. Inflammation associated with noise-induced hearing loss. *J Acoust Soc Am*. 2019;146:4020–32. doi: 10.1121/1.5132545

54 Rai V, Wood MB, Feng H, *et al.* The immune response after noise damage in the cochlea is characterized by a heterogeneous mix of adaptive and innate immune cells. *Sci Rep*. 2020;10:15167. doi: 10.1038/s41598-020-72181-6

55 Aguas AP, Esaguy N, Grande N, *et al.* Effect low frequency noise exposure on BALB/c mice splenic lymphocytes. *Aviat Space Environ Med*. 1999;70:A128-131.

56 Zhiwei W, Yuewen L, Xiaohui D, *et al.* Analysis of related factors between sudden sensorineural hearing loss and serum indices base on artificial intelligence and big data. *Lin Chuang Er Bi Yan Hou Tou Jing Wai Ke Za Zhi*. 2020;34:977–80. doi: 10.13201/j.issn.2096-7993.2020.11.004

57 Zhou Y, Fang C, Yuan L, *et al.* Redox homeostasis dysregulation in noise-induced hearing loss: oxidative stress and antioxidant treatment. *J Otolaryngol Head Neck Surg*. 2023;52:78. doi: 10.1186/s40463-023-00686-x

58 Shi X. Cochlear Vascular Pathology and Hearing Loss. In: Ramkumar V, Rybak LP, eds. *Inflammatory Mechanisms in Mediating Hearing Loss*. Cham: Springer International Publishing 2018:61–90.

59 Jung DJ, Yoo MH, Lee K-Y. Red cell distribution width is associated with hearing impairment in chronic kidney disease population: a retrospective cross-sectional study. *Eur Arch Otorhinolaryngol*. 2020;277:1925–30. doi: 10.1007/s00405-020-05912-0

60 Natarajan N, Batts S, Stankovic KM. Noise-Induced Hearing Loss. *Journal of Clinical Medicine*. 2023;12:2347. doi: 10.3390/jcm12062347

61 Belinskaia DA, Voronina PA, Shmurak VI, *et al.* Serum Albumin in Health and Disease: Esterase, Antioxidant, Transporting and Signaling Properties. *International Journal of Molecular Sciences*. 2021;22:10318. doi: 10.3390/ijms221910318

62 Zheng Z, Liu C, Shen Y, *et al.* Serum Albumin Levels as a Potential Marker for the Predictive and Prognostic Factor in Sudden Sensorineural Hearing Loss: A Prospective Cohort Study. *Front Neurol*. 2021;12:747561. doi: 10.3389/fneur.2021.747561

63 Liu W, Meng Q, Wang Y, *et al.* The association between reduced kidney function and hearing loss: a cross-sectional study. *BMC Nephrol*. 2020;21:145. doi: 10.1186/s12882-020-01810-z

64 Mirvakili A, Dadgarnia MH, Baradaranfar MH, *et al.* Role of Platelet Parameters on

Sudden Sensorineural Hearing Loss: A Case-Control Study in Iran. *PLoS ONE*. 2016;11:e0148149. doi: 10.1371/journal.pone.0148149

65 Chou C-F, Beckles GLA, Zhang X, *et al.* Association of Socioeconomic Position With Sensory Impairment Among US Working-Aged Adults. *Am J Public Health*. 2015;105:1262–8. doi: 10.2105/AJPH.2014.302475

LEGENDS FOR FIGURES

Fig. 1 A combined framework for identifying ONIHL patients.

Fig. 2 Performance of the prediction model on the validation set of dataset D1 and the test set of dataset D2. A ROC curves. B Precision-recall curves.

Fig. 3 Feature selection for the final model using PCA, manual curation, and mRMR. A ROC curves for models constructed with PCA-, manual curation-, and mRMR-selected features on the validation set of dataset D1. B Precision-recall curves of the above models. C Comparison of sensitivity, specificity, and balanced accuracy on the validation set of dataset between the model constructed before and after feature selection. D ROC curve of models using selected features on the test set of dataset D2. E Precision-recall curve of above models. F Comparison of other metrics on the independent test set D2.

Fig. 4 Feature importance ranking for the model built using features selected by mRMR.

Fig. 5 DCA decision curves for models built using three different feature selection methods.

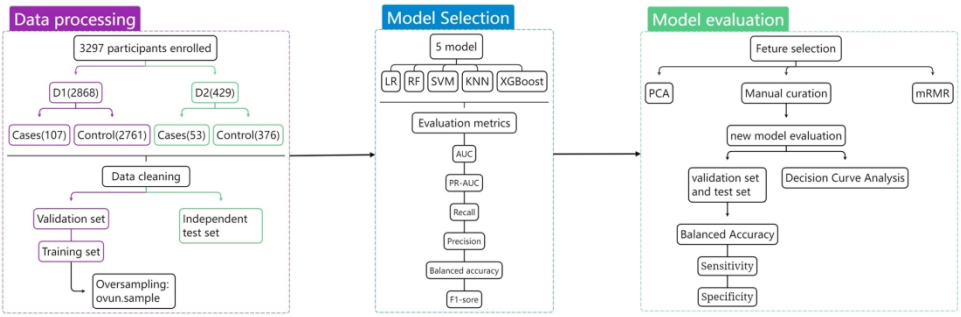

Fig. 1 A combined framework for identifying ONIHL patients.

169x55mm (300 x 300 DPI)

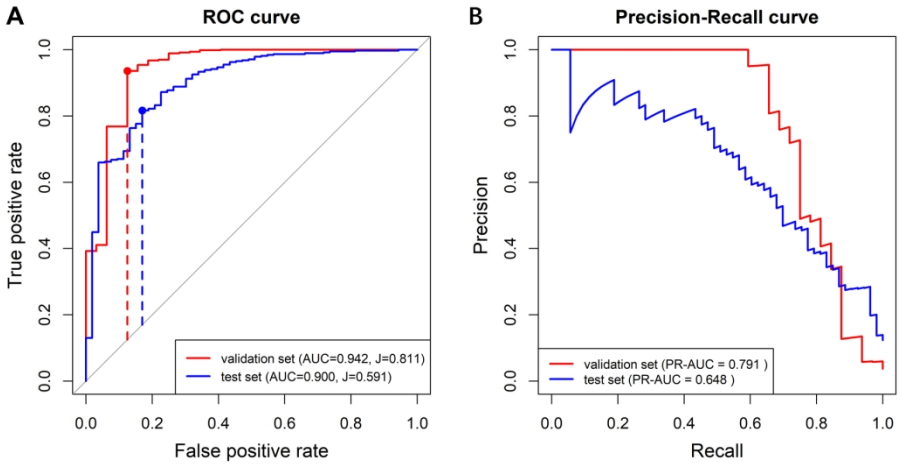

Fig. 2 Performance of the prediction model on the validation set of dataset D1 and the test set of dataset D2. A ROC curves. B Precision-recall curves.

169x95mm (300 x 300 DPI)

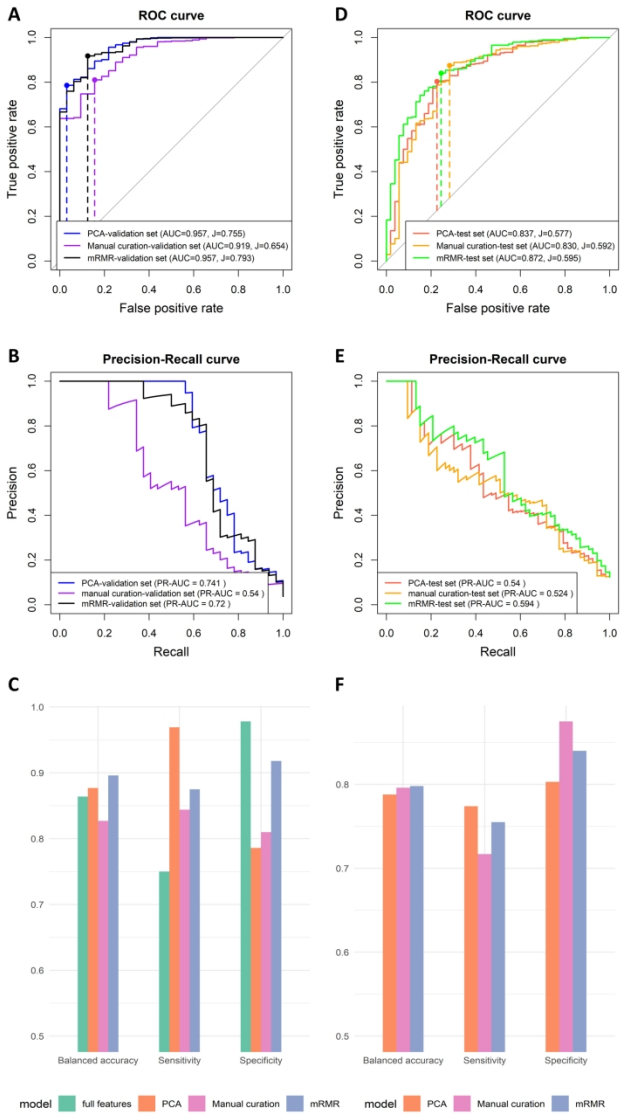

Fig. 3 Feature selection for the final model using PCA, manual curation, and mRMR. A ROC curves for models constructed with PCA-, manual curation-, and mRMR- selected features on the validation set of dataset D1. B Precision-recall curves of the above models. C Comparison of sensitivity, specificity, and balanced accuracy on the validation set of dataset between the model constructed before and after feature selection. D ROC curve of models using selected features on the test set of dataset D2. E Precision-recall curve of above models. F Comparison of other metrics on the independent test set D2.

169x284mm (300 x 300 DPI)

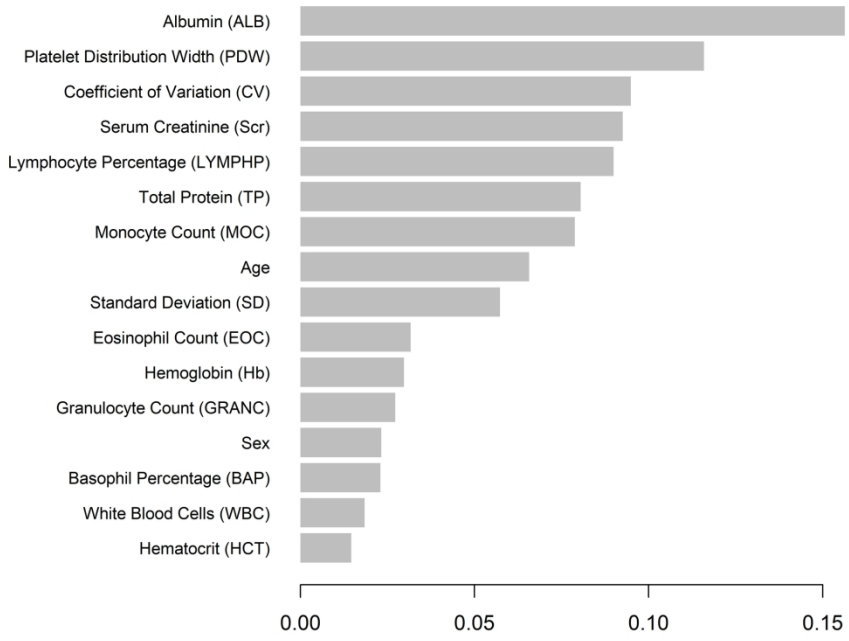

Fig. 4 Feature importance ranking for the model built using features selected by mRMR.

203x152mm (300 x 300 DPI)

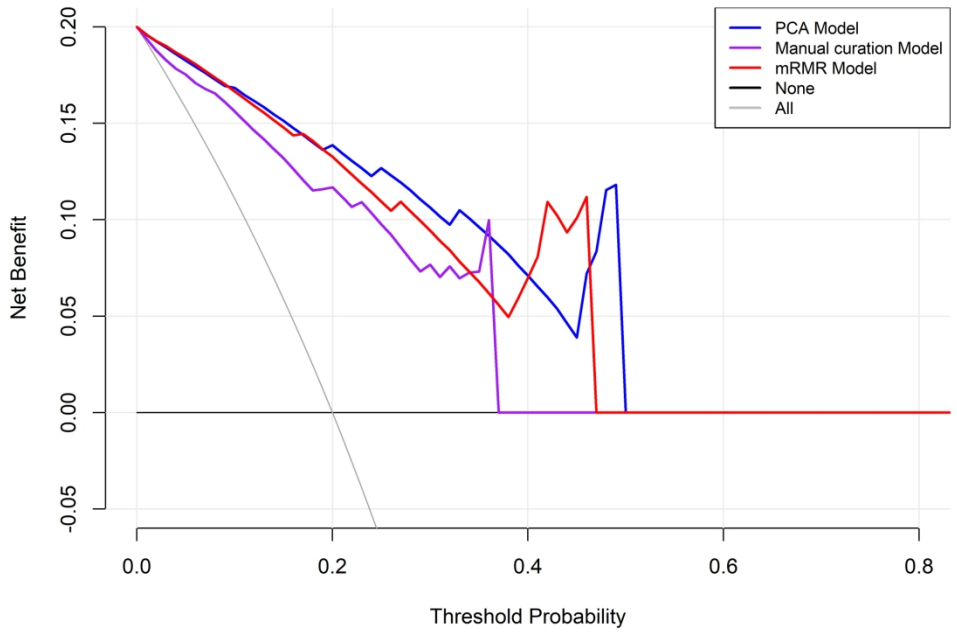

Fig. 5 DCA decision curves for models built using three different feature selection methods.

203x152mm (300 x 300 DPI)

Table S1 Feature subsets for the final model obtained through PCA, manual curation, and mRMR

Manual curation	PCA	mRMR	Rank
Albumin (ALB)	Total Protein (TP)	Albumin (ALB)	1
Age	Albumin (ALB)	Platelet Distribution Width (PDW)	2
Granulocyte Count (GRANC)	Mean Corpuscular Hemoglobin (MCHC)	Coefficient of Variation (CV)	3
Hemoglobin (Hb)	Serum Creatinine (Scr)	Serum Creatinine (Scr)	4
Estimated Glomerular Filtration Rate (eGFR)	Coefficient of Variation (CV)	Lymphocyte Percentage (LYMPHP)	5
Albumin-Globulin Ratio (A/G)	Standard Deviation (SD)	Total Protein (TP)	6
Eosinophil Count (EOC)	Platelet Distribution Width (PDW)	Monocyte Count (MOC)	7
Triglyceride-Glucose Index (TyG)	Mean Corpuscular Volume (MCV)	Age	8
White Blood Cells (WBC)	Hemoglobin (Hb)	Standard Deviation (SD)	9
Total Bilirubin (TBIL)	Mean Corpuscular Hemoglobin (MCH)	Eosinophil Count (EOC)	10
Red Blood Cells (RBC)	Red Blood Cells (RBC)	Hemoglobin (Hb)	11

Basophil Count (BAC)	Glucose/High-Density Lipoprotein (G/H)	Granulocyte Count (GRANC)	12
Segmented Neutrophil to Lymphocyte Ratio (S/L)	Plateletcrit (PCT)	Sex	13
Mean Corpuscular Hemoglobin (MCH)	Platelet/High-Density Lipoprotein (PLT/HDL)	Basophil Percentage (BAP)	14
Sex	Hematocrit (HCT)	White Blood Cells (WBC)	15
Platelet Count (PLT)	Platelet count (PLT)	Hematocrit (HCT)	16

Fig. S1 Heatmap demonstrating nonlinear interactions between max_depth and eta in XGBoost tuning.

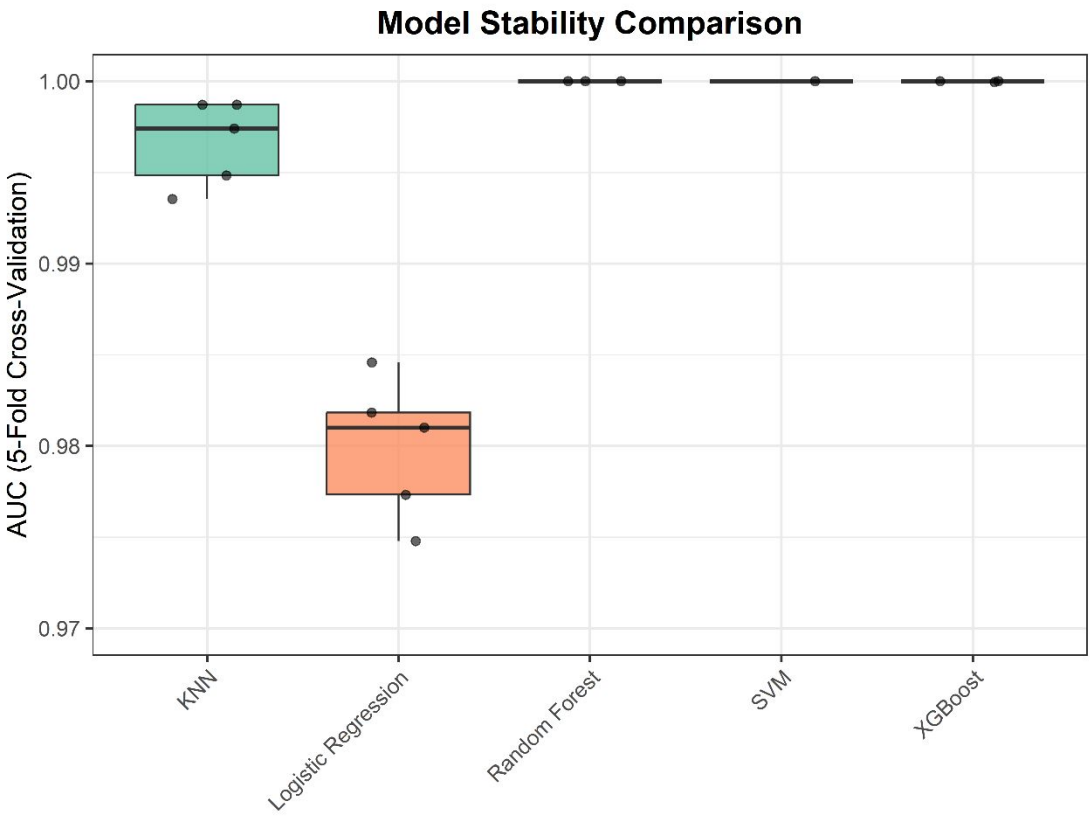

Fig. S2 Boxplots comparing cross-validation AUC distributions across models, highlighting XGBoost's superior stability.

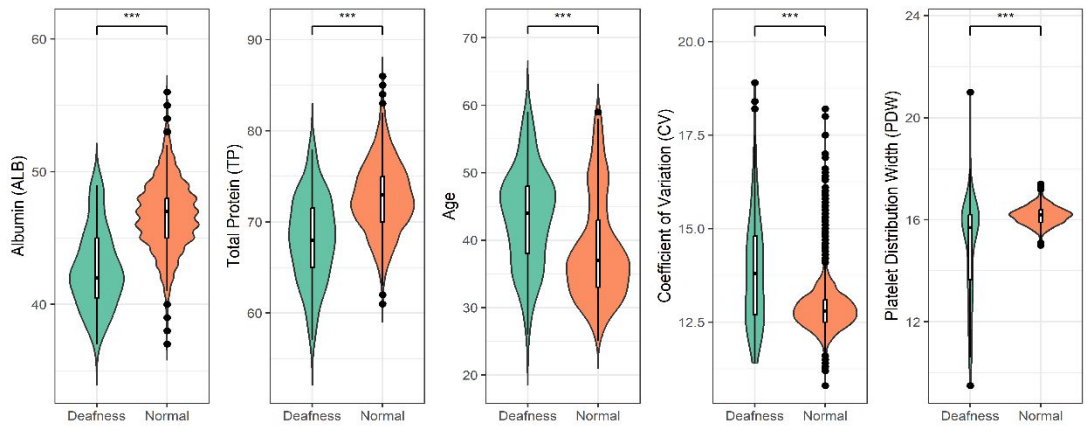

Fig. S3 The five features with the greatest differences between ONIHL patients and noise-exposed individuals with normal hearing.

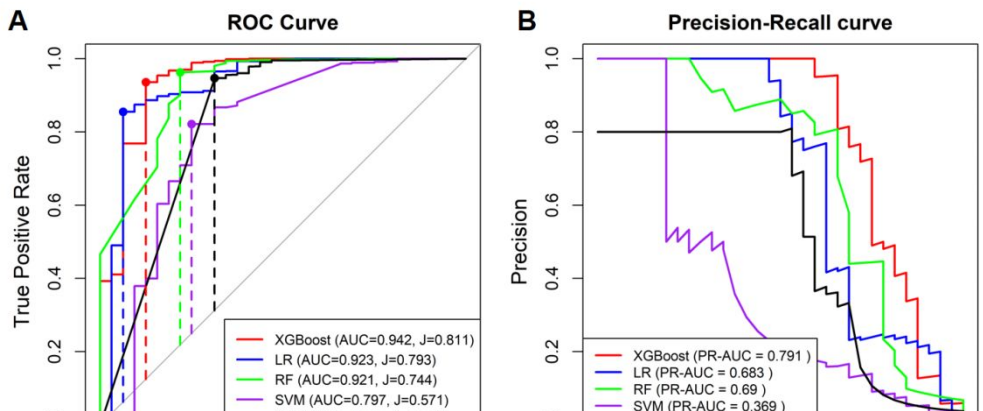

Fig. S4 Performance of the five prediction models in the validation set of dataset D1. A ROC curves. B Precision-recall curves.

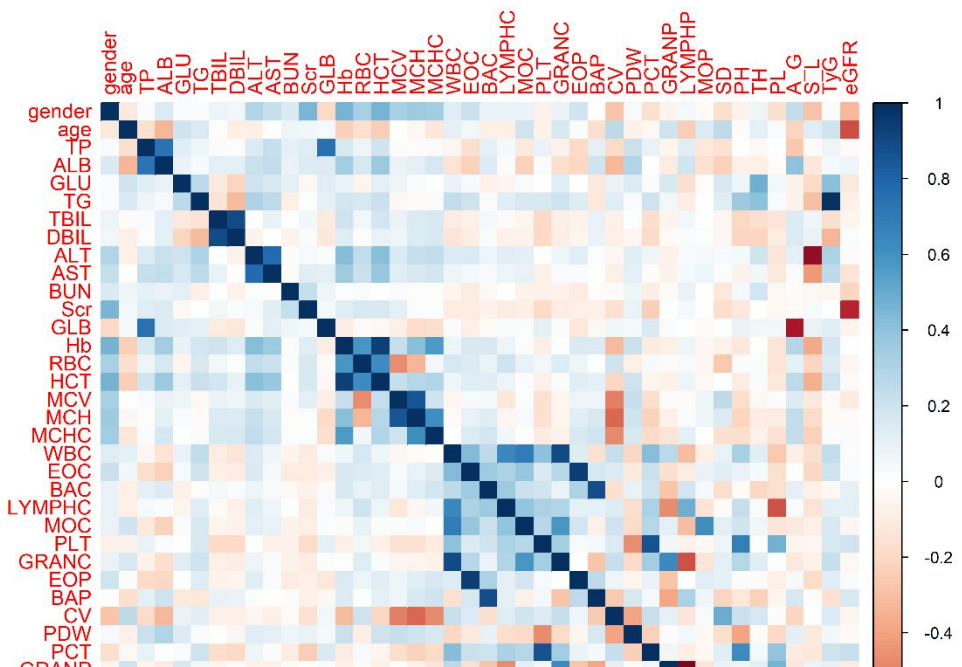

Fig. S5 Heat map of relationship between full variables.

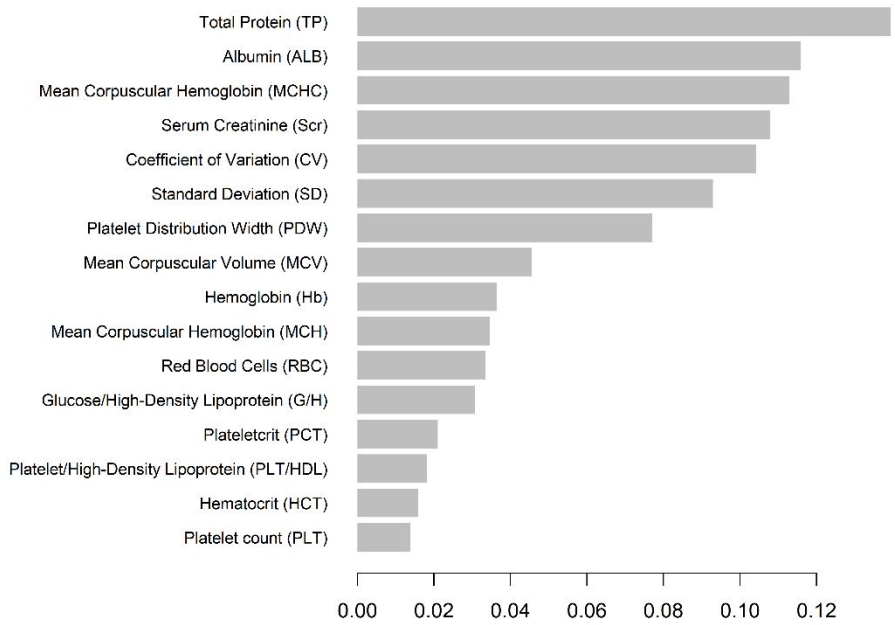

Fig. S6 Feature ranking of PCA screening variables.

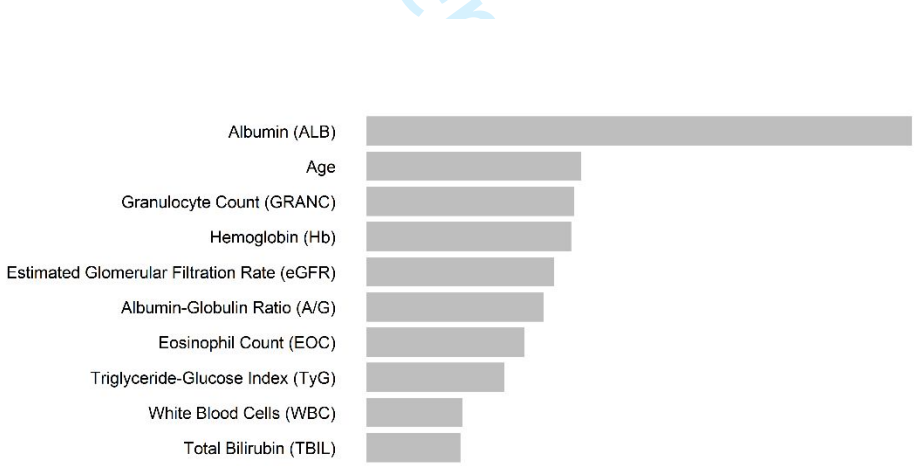

Fig. S7 Feature ranking of manual curation screening variables.

1. TyG Index (Triglyceride-Glucose Index)

The triglyceride-glucose (TyG) index is an established marker for evaluating insulin

resistance, commonly used in assessing metabolic syndrome and diabetes risk. It is calculated using the following formula:

$$TyG = \ln\left[\frac{\mathrm{TG}(mg/dL) \times GLU(mg/dL)}{2}\right]$$

where:

- **TG** represents triglyceride levels (mg/dL),
- **GLU** denotes fasting glucose levels (mg/dL), and
- **ln** is the natural logarithm.

2. eGFR Estimation Formula (Specific to Guangzhou, China Population)

To improve the accuracy of estimated glomerular filtration rate (eGFR) for the Guangzhou population in China, a modified formula has been developed based on local demographic and clinical data:

$$\mathrm{eGFR}\left(mL/(min \times 1.73m^2)\right) = 106 \times \left(\frac{88.4}{Scr(\,mg/dL\,)}\right)^{0.203} \times 0.996^{Age\ (\mathrm{year})}$$

where:

- **Scr** represents serum creatinine concentration (mg/dL), and
- **Age** is the individual's age in years.

The constant **88.4** is employed to convert creatinine units from μmol/L to mg/dL for international standardization.

Contextual Relevance of the Formulas

The **TyG index** serves as an indirect measure of insulin resistance and is useful in predicting metabolic health outcomes. In contrast, the **eGFR formula** provides an estimate of kidney function, tailored to the Chinese population, and is instrumental in identifying and monitoring renal health.