Occupational Noise Exposure and Hearing Health Disparities: A Predictive Modeling Approach

Frank Boahen (300461469), Chikezie Ndimkora (300458935) and Oleksandr (Alex) Bogdan (300455675)

Ottawa University

MIA5126: Group 15

Abstract—Hearing loss is a widespread and under-addressed public health issue, particularly among workers exposed to occupational noise. This project analyzes a large-scale dataset from the Centers for Disease Control and Prevention (CDC), encompassing over 42 million audiometric tests conducted between 1981 and 2010 across 791 industries. Our objective was to explore the correlation between industry-specific noise exposure and hearing loss, while also identifying demographic patterns based on age and gender. Using machine learning models, including XGBoost and Random Forest, we trained predictive models to assess hearing risk and simulate the impact of enhanced preventive measures. Key findings show that hearing loss is significantly higher in high-noise industries and among older male workers, particularly those in the courier and manufacturing sectors. The XGBoost model achieved an R^2 of 0.93 and RMSE of 2.45 dB, indicating strong predictive capability. Our results can inform targeted occupational safety policies and lay the foundation for further economic and behavioral analysis of hearing protection strategies.

Keywords—*Random Forest, XGBoost, Noise-induced hearing loss, Machine learning*

1. Introduction

Noise-induced hearing loss (NIHL) remains one of the most common occupational illnesses globally, despite advances in hearing protection technologies and regulatory frameworks. The National Institute on Deafness and Other Communication Disorders (NIDCD) reports that approximately 1 in 8 Americans suffer from some form of hearing loss, with occupational exposure being a major contributing factor [6]. Although the Occupational Safety and Health Administration (OSHA) [3] mandates hearing conservation programs at workplaces that exceed certain decibel thresholds, there is limited visibility into how risk varies by industry, age, and gender. This lack of granular data and insight inhibits the development of targeted safety measures. Hearing loss tends to progress gradually and is often irreversible, yet many workers do not receive regular screening or education. This project addresses this gap by analyzing a comprehensive CDC dataset and leveraging predictive modeling techniques to identify high-risk groups. Our goal is to support evidence-based interventions that reduce hearing loss through better-targeted regulations and resource allocation.

2. Literature Review

Numerous studies have investigated the causes and effects of occupational hearing loss, but few have combined large-scale audiometric data with demographic and machine learning analyses. Rabinowitz et al. [2] provide a comprehensive overview of hearing degradation in industrial workers, highlighting the role of cumulative exposure and poor compliance with protective measures. Their work underscores the need for better enforcement of existing guidelines but does not differentiate risks by demographic groups. Choi et al. [4] propose machine learning as a valuable tool for predicting hearing outcomes based on individual exposure histories and frequency-specific hearing thresholds. Their research validates the potential of AI in clinical audiology but relies on relatively small, localized datasets. The CDC's dataset offers a unique opportunity to scale this work nationally. OSHA has released periodic surveillance reports showing the distribution of hearing loss by industry sector [3]. However, these analyses are largely descriptive and do not incorporate predictive modeling. Moreover, they often treat age and gender as control variables rather than central components of risk analysis. Our project fills this gap by explicitly modeling the interaction of industry type, age group, and gender to identify vulnerable populations.

3. Methodology

Our approach consisted of four key phases: data acquisition and preparation, feature engineering, model development, and evaluation.

3.1. Data Collection and Preparation

We retrieved the “Trends in Worker Hearing Loss” dataset from the CDC’s public repository, which includes over 42 million rows of hearing test data collected between 1981 and 2010 [5]. Each record corresponds to an individual test and includes:

- Demographic data (age group, gender).
- Industry identifier (NAICS code).
- Audiometric results at frequencies from 500 Hz to 8000 Hz.

After removing incomplete or inconsistent entries, we retained approximately 95.5% of the original dataset. We mapped coded age groups into readable brackets (e.g., Group 1 = 15–25 years), categorized hearing thresholds using OSHA guidelines, and standardized industry classifications for interpretability.

3.2. Feature Engineering

To enable meaningful modeling, we derived several new variables from raw inputs (See fig. 1). Pure-tone average (PTA) is a calculation used to estimate hearing impairment for speech understanding by averaging the hearing thresholds at specific frequencies, typically 500, 1000, 3000 and 4000 Hz [1].

- **Left and Right Ear Averages**: Calculated as the mean of thresholds at each frequency band.
- **High-Frequency Loss Indicator**: Averaged thresholds at 4000–8000 Hz.
- **Overall Hearing Loss Score**: Combined metric using both ears across all frequencies.
- **Industry Noise Profile**: Grouped NAICS codes by noise exposure risk (low, moderate, high).

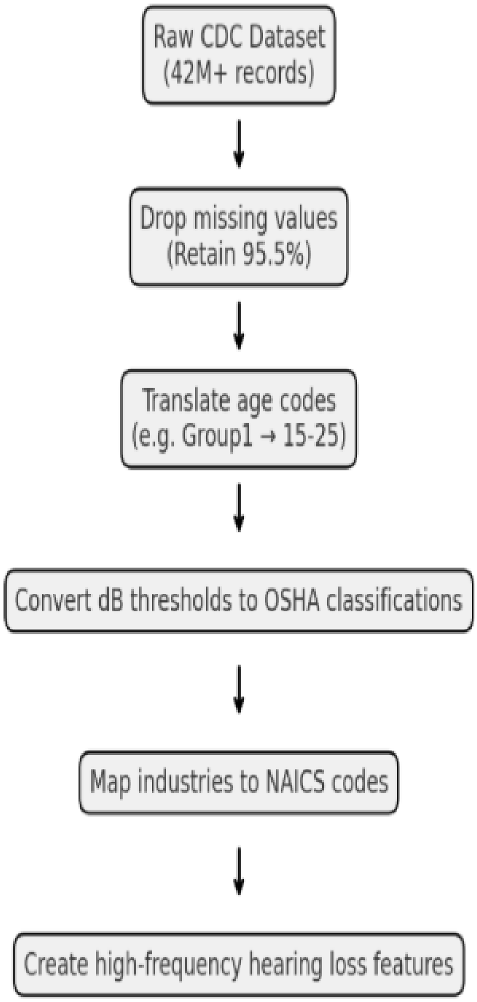

Raw CDC Dataset
(42M+ records)
Drop missing values
(Retain 95.5%)
Translate age codes
(e.g. Group1 → 15-25)
Convert dB thresholds to OSHA classifications
Map industries to NAICS codes
Create high-frequency hearing loss features

Figure 1. data cleaning pipeline

- Pure-tone average (PTA):

$$\frac{Left\ frequencies\ + Right\ frequencies}{number\ of\ frequencies}$$

- Converted decibel thresholds into hearing loss classifications using OSHA standards (PTA > 25 dB).
- Mapped industry description to industry categories.
- One-hot encoding was applied to categorical data and StandardScaler to numerical data.

3.3. Health intervention Modelling

Beyond identifying high-risk populations, a key goal of this project was to estimate the potential benefits of implementing targeted workplace interventions. To this end, we conducted a series of simulations using our trained predictive models to forecast how different levels of intervention could affect future hearing loss outcomes. The feature set includes age group, gender, region, initial pure-tone audiometry

(PTA) scores, industry category, and days between hearing tests. The target variable is the final PTA test score of workers in various industry sectors..

We defined intervention scenarios by adjusting hearing loss risk features—particularly high-frequency threshold averages—to reflect reductions in exposure. These simulated reductions modeled the effect of implementing protective strategies such as:

- Mandatory use of hearing protection (e.g., earmuffs, earplugs)
- Routine audiometric screenings for early detection
- Real-time noise exposure tracking and hazard alerts
- Engineering controls to reduce decibel levels in industrial environments

The predictive models—XGBoost and Random Forest—were then used to re-calculate hearing loss scores under adjusted input values reflecting 10%, 20%, and up to 50% reductions in exposure impact.

4. Model

Two machine learning models were selected based on their proven ability to handle tabular, high-dimensional data:

4.1. XGBoost

XGBoost is a gradient-boosting algorithm that performs well on structured datasets with heterogeneous features. Its ability to handle multicollinearity and rank feature importance made it ideal for our analysis.

Hyperparameters (tuned via GridSearchCV):

- $learning\_rate = 0.1$
- $max\_depth = 6$
- $n\_estimators = 200$
- $subsample = 0.8$

4.2. Random Forest

As a comparative baseline, we used Random Forest—an ensemble model that aggregates decision trees to reduce overfitting. While slightly less accurate, it offered a transparent benchmark for XGBoost's performance.

Hyperparameters:

- $n\_estimators = 100$
- $max\_depth = 10$
- $min\_samples\_split = 5$

Both models were trained on an 80/20 train-test split, with age group, gender, industry, and hearing thresholds as input features and overall hearing loss (dB) as the target variable.

5. Performance Evaluation

To evaluate model performance, we used root mean squared error (RMSE) and coefficient of determination (R^2). Results were as follows (table 1):

Table 1. Model performanc

Model	**RMSE**	R^2
XGB Regressor	2.45	0.93
Random Forest Regressor	4.12	0.83

These results show that XGBoost provided highly accurate predictions and outperformed Random Forest in every tested configuration. The feature importance rankings from XGBoost indicated that **age group**, **industry**, and **high-frequency hearing loss** were the strongest predictors of overall hearing degradation (See fig. 2).

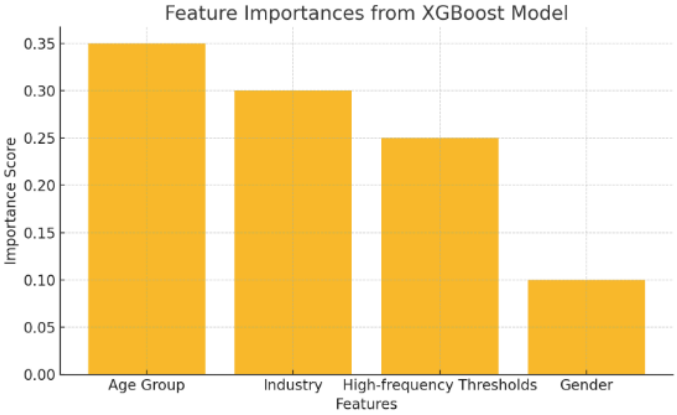
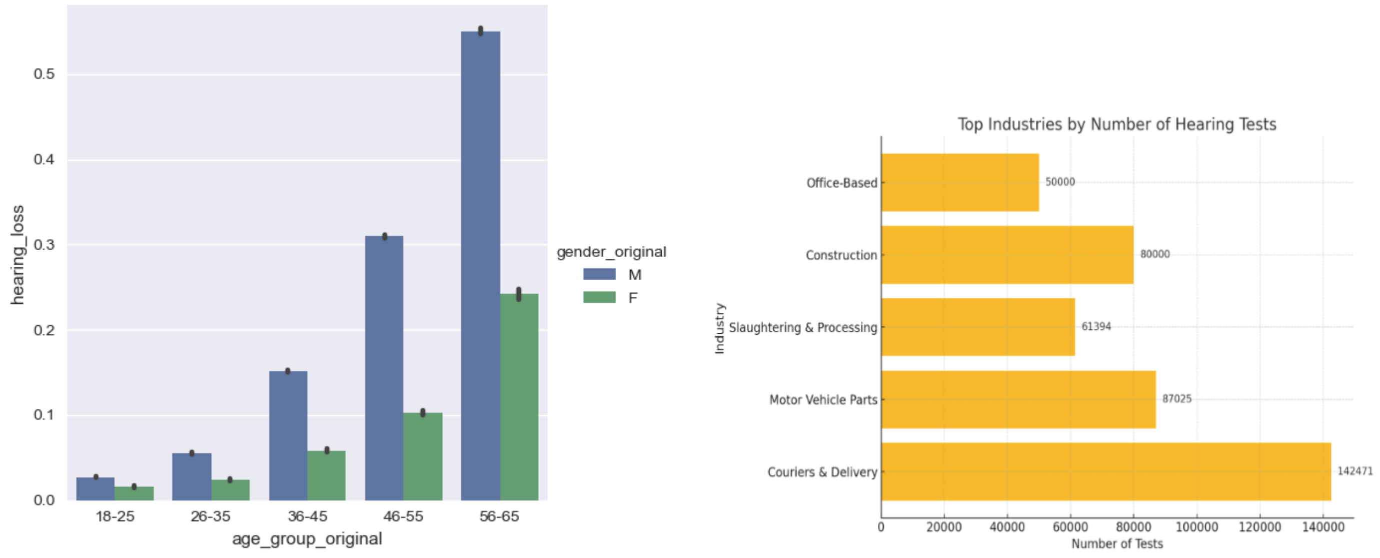


Figure 2

6. Results and Discussion

The analysis revealed:

- Males are three times more likely than females to experience hearing loss.
- Highest risk group: Males aged 50–59 in Courier/Delivery Services with approximately 9.42% average loss (See fig. 3a)



(a) Hearing Loss Prevalence by Age and Gender.

(b) Top Industries by Number of Hearing Tests).

Figure 3

- Office-based sectors had significantly lower risk (55.4% less) compared to high-noise sectors (See fig. 3b and fig. 4)

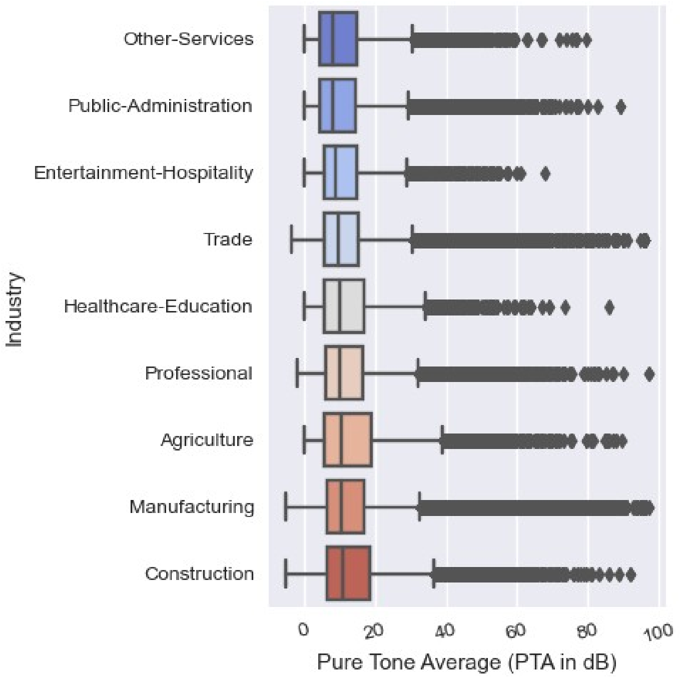


Figure 4. Hearing Loss (PTA) by Industry sector.

- The forecasted results show that targeted interventions can reduce future hearing loss rates. Both XGBoost and Random Forest models demonstrate a consistent decline in predicted hearing loss rates as intervention levels increase from 10% to 50% (See fig. 5). These findings underscore the importance of implementing protective strategies, such as mandatory hearing protection, regular screenings, and noise exposure monitoring, particularly for older male workers in high-risk, high-noise occupations.

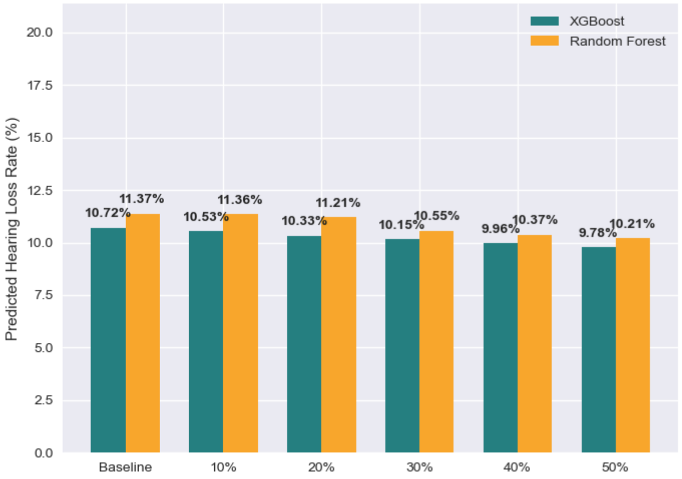


Figure 5. Forecasted Hearing Loss Rate by Intervention Scenario

7. Summary and Conclusion

Our analysis confirms that occupational noise exposure is a major determinant of hearing loss, with disproportionate effects on older males in high-noise sectors. The most affected group—males aged between 50 and 59 in the courier and express delivery industry—showed hearing loss rates exceeding 9%.

Machine learning models, especially XGBoost, proved effective in predicting hearing outcomes and simulating the effect of preventive interventions. This study provides a scalable methodology for identifying high-risk groups and prioritizing hearing conservation resources.

Future studies could integrate economic impact modeling, wearable sensor data, and international comparisons to further enhance our understanding of occupational hearing health disparities.

References

[1] M. W. Yellin, J. Jerger, and R. C. Fifer, "Norms for identifying mild hearing loss in adults. ear and hearing", *Nature review*, vol. 10, no. 5, pp. 302–306, 1989. [Online]. Available: https://doi.org/10.1097/00003446-198910000-00008.

[2] P. Rabinowitz, "Noise-induced hearing loss", *American family physician*, vol. 61, pp. 2749–56, 2759, Jun. 2000.

[3] Occupational Safety and Health Administration (OSHA), *Hearing conservation*, 2002. [Online]. Available: https://www.osha.gov/sites/default/files/publications/osha3074.pdf.

[4] Choi, Y. H., et al., "Artificial intelligence for detecting hearing loss patterns in occupational settings.", *Computers in Biology and Medicine*, Jun. 2021.

[5] Centers for Disease Control and Prevention (CDC), *Hearing loss dataset*, 2024. [Online]. Available: https://data.cdc.gov/d/c294-dri5.

[6] National Institute on Deafness and Other Communication Disorders (NIDCD), *Noise-induced hearing loss*, 2024. [Online]. Available: https://www.nidcd.nih.gov/health/noise-induced-hearing-loss.