talenta PUBLISHER

Jurnal Sistem Teknik Industri
Journal homepage: https://talenta.usu.ac.id/jsti

Prediction of Occupational Health Risk Using the Random Forest Machine Learning Model in a Metal Casting Workplace: A Case Study at CV. Karya Yudita Baroqah

Nadilah Sary[1], Nismah Panjaitan[*1], Abdul Rahim Matondang[1], Ukurta Tarigan[1]
[1]*Industrial Engineering Study Program, Faculty of Engineering, Universitas Sumatera Utara, Medan, 20155, Indonesia*
[*]Corresponding Author: nismah.panjaitan@usu.ac.id

ARTICLE INFO

Article history:
Received 8 December 2025
Revised 11 May 2026
Accepted 11 May 2026
Available online 8 June 2026

E-ISSN: 2527-9408
P-ISSN: 1411-5247

How to cite:
N. Sary, N. Panjaitan, A.R. Matondang, and U. Tarigan, "Prediction of occupational health risk using the random forest machine learning model in a metal casting workplace: a case study at CV. Karya Yudita Baroqah," *J. Sist. Tek. Ind.*, vol. 28, no. 2, pp. 69–79, June 2026.

ABSTRACT

Occupational health risks from chronic exposure to noise and airborne particulate matter remain a major concern in metal casting workplaces, especially in small-scale foundries with limited controls. The parameters measured in this study include noise exposure (Leq, dBA), particulate matter concentrations ($PM_{2.5}$ and PM_{10}), and workers' health symptoms. Field measurements at CV. Karya Yudita Baroqah showed exceedances of regulatory limits: noise levels in Molding and Finishing reached 89–93 dBA, and $PM_{2.5}$ and PM_{10} concentrations reached 72–80 µg/m³ and 155–174 µg/m³, surpassing recommended thresholds. These conditions indicate that workers are consistently exposed to hazardous environments that may lead to cumulative health impairments. This study aims to predict occupational health risk using a two-stage Random Forest model integrating environmental exposure data and workers' symptoms. Stage-1 classified environmental risk levels with 99% accuracy, while Stage-2 predicted symptom-based health risk categories with 71% accuracy. PM_{10} and $PM_{2.5}$ were the strongest predictors, followed by noise intensity. The model demonstrates reliable performance and captures individual variability that traditional threshold-based assessments often overlook. The findings highlight that a combined machine-learning and HRA approach provides a practical, data-driven tool for early detection of high-risk workers and supports targeted interventions in metal casting workplaces.

Keyword: *Occupational Health Risk, Random Forest, Machine Learning, Noise Exposure, Particulate Matter*

ABSTRAK

Risiko kesehatan kerja akibat paparan kebisingan dan partikulat udara merupakan masalah penting pada lingkungan pengecoran logam, terutama pada industri kecil dengan keterbatasan pengendalian. Parameter yang diukur dalam penelitian ini meliputi paparan kebisingan (Leq, dBA), konsentrasi partikulat udara ($PM_{2.5}$ dan PM_{10}), serta gejala kesehatan yang dilaporkan oleh pekerja. Pengukuran di CV. Karya Yudita Baroqah menunjukkan pelampauan ambang batas: kebisingan pada stasiun pencetakan dan pengepakan mencapai 89–93 dBA, dan kadar $PM_{2.5}$ serta PM_{10} mencapai 72–80 µg/m³ dan 155–174 µg/m³. Kondisi ini menunjukkan bahwa pekerja terpapar lingkungan berbahaya secara konsisten dan berpotensi mengalami gangguan kesehatan jangka panjang. Penelitian ini bertujuan memprediksi risiko kesehatan kerja menggunakan model random forest dua tahap yang menggabungkan paparan lingkungan dan gejala pekerja. Tahap-1 memprediksi tingkat risiko lingkungan dengan akurasi 99%, sedangkan Tahap-2 memprediksi kategori risiko kesehatan berbasis gejala dengan akurasi 71%. PM_{10} dan $PM_{2.5}$ menjadi prediktor paling dominan, diikuti kebisingan. Model ini menunjukkan kinerja yang stabil serta mampu menangkap variasi respons individu yang sering tidak terlihat pada penilaian berbasis ambang batas. Temuan ini menunjukkan bahwa pendekatan machine learning dan health risk assessment dapat menjadi alat prediktif yang efektif untuk deteksi dini pekerja berisiko tinggi dan mendukung perencanaan pengendalian yang lebih tepat sasaran.

http://doi.org/10.32734/register.v27i1.idarticle

Keyword: Risiko Kesehatan Kerja, Random Forest, Machine Learning, Kebisingan, Debu Partikulat

1. Introduction

A healthy and safe working environment, commonly referred to occupational safety and health, is the discipline that deals with the prevention of worked relate injuries and illnesses, as well as the protection and promotion of employees' health [1]. Occupational health hazards refer to workplace conditions that can cause adverse health effects due to physical, chemical, biological, or ergonomic exposures. In industrial environments such as metal casting, these hazards are primarily associated with physical agents, including excessive noise and airborne particulate matter. Noise exposure can lead to noise-induced hearing loss and physiological stress, while particulate matter can penetrate deep into the respiratory system and contribute to chronic respiratory and cardiovascular diseases. These hazards often occur simultaneously and interact with each other, increasing the overall risk to workers' health beyond the impact of single exposures [2].

Globally, occupational health problems remain a significant burden. According to the International Labour Organization an estimated 1.8 million work related deaths occur annually in the Asia Pacific region, accounting for nearly two thirds of all occupational fatalities worldwide. This indicates that Asia remains the most affected region in terms of occupational mortality. At the global level, more than 2.78 million people die each year from occupational diseases and injuries, underscoring the severe burden of worked relate health problems worldwide [3].

In Indonesia, data from the Ministry of Health reveal that the number of occupational disease cases remains significantly high. Between 2011 and 2014, the recorded cases fluctuated from 57,292 in 2011, 60,322 in 2012, 97,144 in 2013, and 40,694 in 2014, indicating persistent challenges in occupational disease prevention and monitoring across industrial sectors [3].

Occupational diseases are defined as chronic or acute health disorders that develop due to prolonged exposure to workplace hazards such as noise**,** dust, vibration, chemicals, and ergonomic strain. These exposures can lead to conditions including respiratory impairment, hearing loss, fatigue, and reduced work performance [4]. The World Health Organization (WHO) emphasizes that most of these health outcomes are preventable through proper environmental control and early risk detection [5]. Effective regulation not only protects workers from physical injury, but also mitigates the risk of chronic diseases associated with exposure to hazardous work environments [6].

In Indonesia, industrial environments particularly small and medium scale enterprises (SMEs) such as metal casting workshops often face serious occupational health challenges due to limited monitoring systems and poor engineering controls. Measurements in similar small manufacturing facilities have shown noise exposure exceeding the 85 dBA threshold and particulate concentrations ($PM_{2.5}$ and PM_{10}) surpassing the permissible exposure limits of 65 µg/m³ and 150 µg/m³, respectively, as stated in Minister of Manpower Regulation No. 5/2018. High levels of $PM_{2.5}$ (particles smaller than 2.5 micrometers) are particularly concerning because they can penetrate deeply into the alveolar region of the lungs, increasing the risk of chronic bronchitis, asthma, and cardiovascular disease [7]. Studies by Nyanza et al. (2024) and Agbehadji et al. (2025) further confirmed that prolonged exposure to airborne particulate matter and noise in low resource industries significantly elevates workers' respiratory and cardiovascular risks [8][9].

Measurements conducted at CV. Karya Yudita Baroqah align with these findings. Field data show that noise levels in the Molding and Finishing stations ranged from 89–93 dBA, clearly exceeding regulatory limits. Airborne particulate concentrations also surpassed recommended thresholds, with $PM_{2.5}$ reaching 72–80 µg/m³ and PM_{10} reaching 155–174 µg/m³. These exceedances indicate that workers are consistently exposed to hazardous environmental conditions, increasing their occupational health risk [10].

Health risk assessment (HRA) approaches, although widely used for compliance purposes, are generally based on deterministic risk equations and fixed exposure thresholds. These methods assume simplified and linear relationships between exposure and health effects, which do not fully reflect real industrial conditions. In practice, occupational exposure is characterized by variability in intensity, duration, and combination of hazards across different tasks and workers. For example, noise and airborne particulate matter often occur simultaneously, creating combined and non-linear effects that cannot be adequately captured by single-factor models [11]. Furthermore, deterministic approaches have limitations in representing uncertainty and individual physiological variability, which play an important role in determining health outcomes under similar exposure conditions [12]. As highlighted in recent studies, there is a growing need for data-driven approaches that can better represent uncertainty and integrate multiple exposure variables in occupational risk assessment [13].

In this context, machine learning (ML) has emerged as a promising tool in occupational health research due to its ability to handle large, multidimensional datasets and identify complex, non-linear relationships between exposure and health outcomes. ML models can learn directly from empirical data without relying on strict assumptions, enabling more flexible and accurate prediction of occupational risks [14]. Among various machine learning techniques, the random forest (RF) algorithm has gained particular attention due to its robustness, ability to reduce overfitting, and capability to evaluate feature importance. RF constructs multiple decision trees and aggregates their results, allowing it to capture complex interactions among multiple exposure variables while maintaining model stability and interpretability [15].

Despite these advantages, the application of ML in occupational health remains limited in integrating environmental exposure measurements and workers' health symptoms, especially in small-scale metal casting industries where exposure levels are high and monitoring systems are limited [16][17].

This study integrates the principles of health risk assessment (HRA) with the predictive power of random forest machine algorithm to address the limitations of conventional deterministic risk assessment methods in metal casting environments. Specifically, this study aims to quantify environmental exposure levels of noise, $PM_{2.5}$, and PM_{10} across different workstations, evaluate occupational health risks using the HRA approach based on measured exposure data, and develop a two-stage random forest model to predict environmental risk levels and workers' health risk categories (low, moderate, or high) by integrating exposure variables and symptom data. [18]. This approach addresses the limitations of deterministic HRA models by incorporating variability, individual differences, and the multi exposure nature of real workplaces. The outcomes are expected to provide a scientifically grounded, low cost, and explainable decision support framework for early identification of high risks workstations and for guiding targeted engineering and administrative controls in small and medium scale foundries [19].

2. Methodology

This study adopts a quantitative survey design to investigate the relationship between environmental exposures (noise, $PM_{2.5}$, and PM_{10}) and occupational health risks among workers in the metal casting industry. The survey approach is used to collect cross-sectional data on workers' health symptoms through structured questionnaires, while environmental exposure data are obtained through direct field measurements. In addition to the survey component, this study integrates Health Risk Assessment (HRA) and machine learning analysis to provide a more comprehensive evaluation of occupational health risks. The measured exposure data are first analyzed using the HRA framework to determine risk levels. Subsequently, a two-stage random forest model is developed to predict environmental risk and workers' health risk categories by combining exposure and symptom data. This integrated approach enables the study to go beyond conventional survey analysis by incorporating data-driven modeling, allowing a more accurate representation of complex exposure conditions in real industrial environments [20].

The population for this study consists of workers at CV. Karya Yudita Baroqah, a metal casting company located in Medan, Indonesia. The sample will include 20 workers who represent different job roles within the factory, including casting, molding, storage, and finishing stations. A purposive sampling technique will be employed to select workers who have been exposed to high levels of noise and airborne particulate matter during their daily tasks. The total sample size of 20 was chosen based on the nature of the industry and the homogeneous working conditions across roles, making it representative of the environment in the studied factory [21].

The data will be collected through a combination of environmental exposure measurements and self-reported health symptom data.

1. Environmental exposure measurement: Noise (dBA) will be measured using a 4-in-1 environmental meter and particulate matter ($PM_{2.5}$ and PM_{10}) will be measured using an air quality monitor over a month period at three different times each day (10:00, 13:00, and 16:00). The measurements will be recorded at each workstation to assess variability in exposure throughout the working day [22].

2. Health symptom data were collected using a structured questionnaire developed based on relevant occupational health literature and adjusted to the study context. The questionnaire includes symptoms such as fatigue, shortness of breath, cough, dizziness, eye irritation, and other related indicators. Each symptom was recorded using a binary or scaled response to represent the presence and severity of symptoms. This approach ensures that the collected data are relevant to the exposure conditions and suitable for further analysis using machine learning techniques [23].

In this study, noise exposure in the metal casting industry will be measured using the LEQ (equivalent continuous sound level) method, which is a standard approach in occupational noise assessment. LEQ is a widely accepted standard for measuring the continuous equivalent level of fluctuating sound over a specified period, typically used in occupational noise studies. LEQ represents the constant sound level that, if it were to occur continuously over the same period, would have the same energy as the fluctuating noise levels actually present during that time [24].

The LEQ can be calculated using the following formula:

$$\mathrm{LEQ} = 10 \times \log_{10}\left(\frac{1}{\mathrm{T}}\int_{0}^{\mathrm{T}} \mathrm{p(t)}^{2}\mathrm{dt}\right) \tag{1}$$

The LEQ method provides a more accurate representation of cumulative noise exposure over a period of time, accounting for both short term peaks and long term exposure. This method is particularly useful in environments with fluctuating noise levels, such as metal casting workshops, where workers are exposed to various levels of noise depending on the specific task and machinery being used [25].

Before proceeding to the machine learning analysis, the Health Risk Assessment (HRA) framework will be used to calculate the Hazard Quotient (HQ), which is commonly employed in occupational health studies to evaluate the degree of risk posed by environmental exposure. The HQ is calculated using the following formula [26]:

$$\mathrm{HQ} = \frac{\text{Exposure Concentration}}{\text{Reference Concentration (RfC)}} \tag{2}$$

After calculating hazard quotients and classifying environmental risk levels, the random forest (RF) algorithm is applied in Stage 2 to predict workers' health symptom categories. Environmental exposure data (noise, $PM_{2.5}$, PM_{10}) and workers' symptom data are integrated and preprocessed before being divided into training and testing datasets. The RF model is trained to learn the relationship between exposure variables and health outcomes, and its performance is evaluated using accuracy, precision, recall, and F1-score. The output of this stage is the classification of workers' health risk categories (low, moderate, or high), implemented using python in Google Colab with the Scikit-learn library [27].

The random forest model is well suited for this study due to its robustness in handling complex datasets and its ability to handle multiple input features (such as exposure data for noise, $PM_{2.5}$, and PM_{10}) and predict outcomes (such as workers' health symptoms). The model will be trained using training data and tested on unseen data to validate its predictive accuracy [28].

The dataset is divided into a training set (80%) and a test set (20%). The RF model is trained on the training dataset and evaluated on the test dataset. Performance metrics include accuracy, precision, recall, and F1 score to assess the model's performance. Additionally, cross validation is conducted to evaluate the model's generalization ability[29].

Furthermore, feature importance analysis will be performed to identify the exposure variables (noise, $PM_{2.5}$, PM_{10}) that most influence health outcomes. This will help determine which environmental factors should be prioritized in occupational health management strategies[30].

3. Result and Discussion

In this study, noise exposure and particulate matter ($PM_{2.5}$, PM_{10}) levels were measured across four workstations in a metal casting factory: storage, casting, molding, and finishing. The data was collected over a 25-day observation period at three different times during the workday (10:00, 13:00, and 16:00) providing sufficient temporal variation for reliable machine learning analysis.

3.1. Noise Exposure

Noise exposure was measured using a 4-in-1 environmental meter, and the results were expressed in decibels (dBA) using the equivalent continuous sound level (Leq) method. The measurements were conducted at each workstation to capture variations in noise levels throughout the working day.

The results show that the finishing workstation recorded the highest noise levels, ranging from 92.4 to 93.5 dBA, while the molding workstation ranged from 89.2 to 90.0 dBA, both consistently exceeding the occupational exposure limit of 85 dBA set by Permenaker No. 5/2018. The finishing process involves grinding, cutting, and surface smoothing using mechanical tools, which generate high-intensity continuous noise. The molding process includes sand preparation and mold shaping, which also produce significant noise due to equipment operation.

Casting activities showed fluctuating noise levels between 86–88 dBA, indicating intermittent exposure above the threshold. In contrast, the storage area maintained lower noise levels (76–78 dBA), remaining within acceptable limits. Casting activities involve pouring molten metal and handling materials, resulting in moderate and fluctuating noise levels. Meanwhile, storage activities mainly consist of material handling and organization, which generate relatively low noise exposure.

In contrast, the storage maintained noise levels that were well below the 85 dBA threshold, ranging from 76.4–77.8 dBA. These levels suggest a low risk environment in terms of noise related health hazards, indicating that the workers in this area were less likely to experience hearing impairment or stress from noise exposure. Casting, while often hovering near the threshold at 85.9–87.7 dBA, represents a borderline risk zone, where workers are exposed to noise levels that may not immediately cause hearing damage but could lead to discomfort or long term issues if sustained [31].

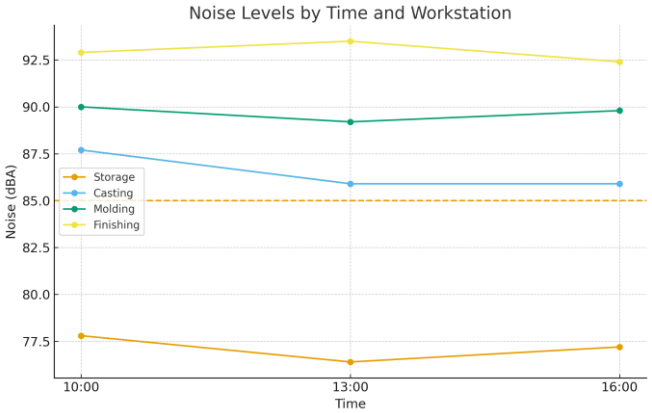


Figure 1. Noise Levels

Figure 1 visualizes this pattern clearly. Noise levels in Molding and Finishing consistently appear above the regulatory threshold throughout all measurement times. This reinforces the classification of these areas as priority targets for noise control interventions such as machine enclosure, damping materials, and redesign of workflow processes. The dashed line marks the 85 dBA exposure limit. Storage remains well below the limit across the day (≈76–78 dBA). Casting hovers around the threshold slightly above it at 10:00 (≈88 dBA) and near it at 13:00–16:00 (≈86 dBA). Molding is consistently high (≈89–90 dBA at all times), while Finishing records the highest levels (≈92–93 dBA) with a midday peak. The small diurnal spread (generally ≤1–2 dB per station) indicates stable process related noise rather than time of day effects. These results prioritize finishing

> molding > casting >> storage for engineering controls (enclosures, silencing, isolation), supported by administrative measures (task rotation, enforced rest breaks) and consistent hearing protection use.

3.2. Particulate Matter ($PM_{2.5}$ and PM_{10})

The average concentrations of $PM_{2.5}$ and PM_{10} measured at each workstation using an air quality monitor. The measurements were conducted over a 25-day observation period to capture variations in particulate exposure.

The results show that $PM_{2.5}$ levels were highest in finishing (80 µg/m³) and molding (72 µg/m³), significantly exceeding the 65 µg/m³ limit set by Indonesian regulations. PM_{10} levels reached 174 µg/m³ in Finishing and 155 µg/m³ in molding, surpassing the 150 µg/m³ permissible limit.

These results are consistent with the dust-generating activities performed in these workstations. The finishing process involves mechanical abrasion, grinding, and cleaning operations that produce fine particulate emissions, while the molding process involves sand handling and material preparation, which generate both fine ($PM_{2.5}$) and coarse (PM_{10}) particles. The extended observation period strengthens the reliability of the data by capturing daily variations in exposure, making the results more representative for further analysis. $PM_{2.5}$ exposure is particularly critical because these particles can penetrate deeply into the alveolar region of the lungs, leading to inflammation, reduced lung function, increased asthma risk, and potential cardiovascular effects [22].

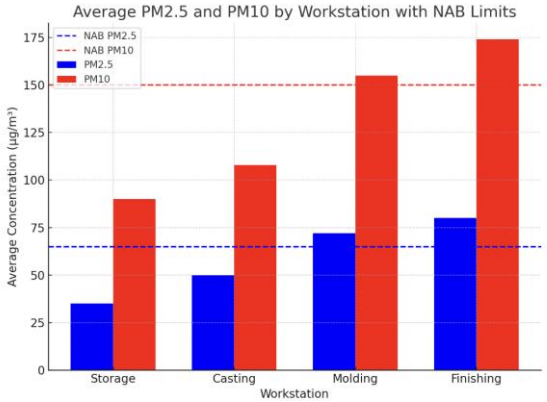


Figure 2. Average PM2.5 and PM10 by Workstation.

Figure 2 shows these exposures in visual form, making it clear that finishing and molding have substantially higher particulate concentrations compared to storage and casting. This spatial pattern aligns with previous findings in both Indonesian and international foundries. This bar chart displays the average concentrations of $PM_{2.5}$ and PM_{10} across the four workstations. The blue bars represent $PM_{2.5}$ levels, and the red bars represent PM_{10} levels. The dashed lines indicate the permissible exposure limits: $PM_{2.5}$ = 65 µg/m³ and PM_{10} = 150 µg/m³. As illustrated, Finishing has the highest concentrations of both $PM_{2.5}$ (80 µg/m³) and PM_{10} (174 µg/m³), exceeding both limits, while Molding also exceeds the $PM_{2.5}$ limit and comes close to the PM_{10} limit. In contrast, Storage maintains low levels of both $PM_{2.5}$ and PM_{10}, making it a low risk area.

3.3. Health Risk Assessment (HRA)

Health risk assessment (HRA) plays a crucial role in evaluating the potential health impacts of environmental exposures in occupational settings. In the context of this study, HRA is used to assess the risks associated with noise exposure and particulate matter ($PM_{2.5}$ and PM_{10}) in the metal casting workplace. By calculating the hazard quotient (HQ), we can quantify the severity of environmental exposures and classify them into risk categories low, moderate, or high based on their potential to cause health impairments. This assessment provides a structured framework for understanding how exposure levels relate to the likelihood of developing respiratory, auditory, and cardiovascular disorders [2].

Table 3 summarizes the HRA for each workstation based on the average noise levels (dBA) and particulate matter concentrations ($PM_{2.5}$ and PM_{10}). It also includes the HQ for dust exposure and the associated risk category. These HQ outcomes reflect the combined effect of both inhalation and acoustic exposures. An HQ value > 1 indicates exposure exceeding reference safety levels, associated with increased risk of respiratory

and auditory disorders. The results from finishing and molding match the exceedances observed in the exposure data, confirming that these areas pose the highest occupational health risk [26].

Table 1. HRA Summary by Workstation

Workstation	Noise_avg(dBA)	PM2.5_avg	PM10_avg	HQ_Dust	Risk
Storage	77.1	35	90	0.54	Moderate
Casting	86.5	50	108	0.77	Moderate
Molding	89.7	72	155	1.11	High
Finishing	92.9	80	174	1.23	High

3.4. Workers' Symptom Distribution

In addition to environmental exposure measurements, assessing workers' health symptoms provides valuable insights into the physiological effects of occupational hazards. Symptoms such as fatigue, cough, shortness of breath, and eye irritation are common indicators of the body's response to prolonged exposure to noise and airborne particulate matter. These symptoms not only reflect the immediate discomfort caused by exposure but also serve as early warning signs of more severe health conditions [13].

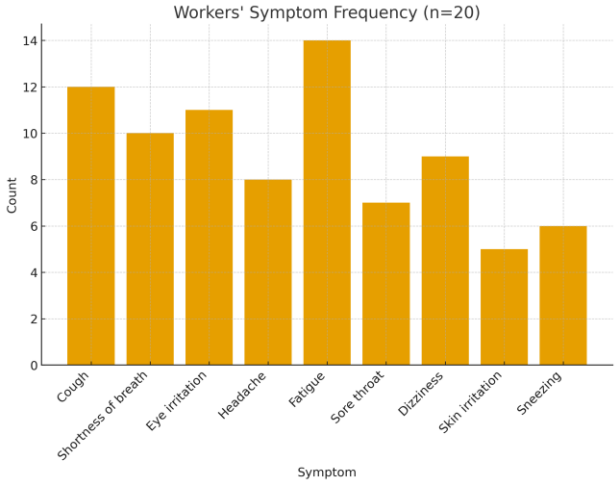


Figure 3. Workers' Symptom Frequency

Figure 3 illustrates the frequency of health symptoms reported by workers. Respiratory/irritative complaints dominate: fatigue (14), cough (12), eye irritation (11), and shortness of breath (10) lead the totals, followed by dizziness (9) and headache (8). Less frequent are sore throat (7), sneezing (6), and skin irritation (5). The pattern aligns with elevated dust and noise exposure at molding/finishing, indicating priority for ventilation, housekeeping, and consistent respirator/hearing protection use.

3.5. Machine Learning: Random Forest Model for Health Risk Prediction

Machine learning (ML) has increasingly been applied in occupational health risk prediction due to its ability to handle large and complex dataset. In this study, the random forest (RF) algorithm was employed to classify environmental health risks based on noise and particulate exposure consists of 500 observations, obtained from 20 workers over a 25-day observation period. Each observation represents the combination of environmental exposure measurements and corresponding workers' health symptoms recorded on the same day. This resulted in a dataset that captures variations in exposure across time, location, and working conditions. This method is particularly useful for modeling multi exposure data, where interactions between noise, particulate matter, and individual health responses are complex and non linear [29]

Table 2. Confusion Matrix for HRA Prediction Model

Actual → Predicted	Low	Moderate	High
Low	10	0	0
Moderate	0	4	0
High	0	0	86

Table 4 presents the results of the. stage-1 random forest model, which predicts environmental risk based solely on exposure measurements. The accuracy of this model is critical, as it lays the groundwork for classifying workers into appropriate risk categories based on environmental conditions. The confusion matrix above shows the actual vs. predicted health risk categories based on the stage-1 random forest model. The model achieved a perfect classification (99% accuracy) in identifying high risk workers, with 86 workers accurately classified as high risk. There were no misclassifications in the low or moderate categories. This suggests that the stage-1 model performed exceptionally well in categorizing environmental risk based on exposure data (noise and particulate matter), which directly correlates with the HRA thresholds for occupational exposure.

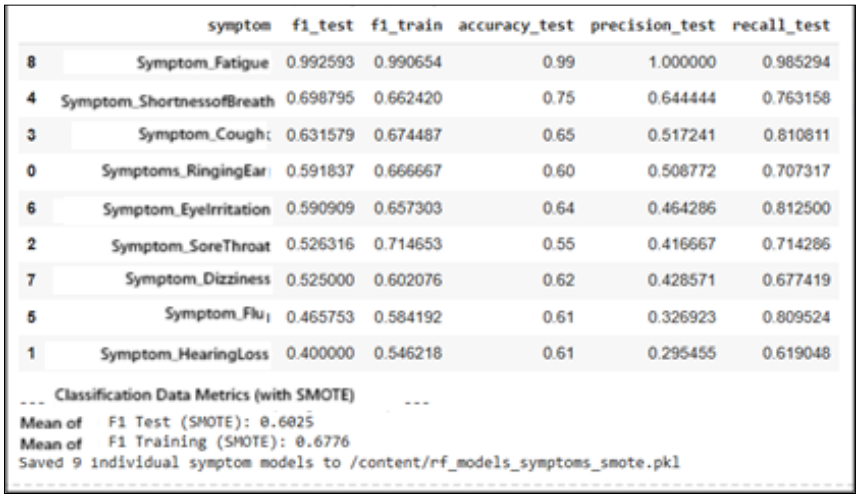
The feature importance analysis from stage-1 of the random forest model reveals the relative importance of the variables used to classify occupational health risks. Noise (dBA) emerged as the most important predictor, with an importance score of ≈0.44, followed by PM_{10} (≈0.38) and $PM_{2.5}$ (≈0.18). This aligns with the observed field exceedances in both molding and finishing, where noise and particulate exposure were the highest. The findings suggest that control efforts should first prioritize noise reduction and coarse particulate (PM_{10}) exposure, with additional measures focused on fine particles ($PM_{2.5}$). The feature importance analysis is essential for guiding risk mitigation strategies, helping prioritize interventions in high exposure areas.

Table 3. Confusion Matrix for Stage-2 Model

Actual → Predicted	Low	Moderate	High
Low	28	9	0
Moderate	2	42	4
High	0	14	1

Table 5 presents the confusion matrix for Stage-2, showing the classification results for health risk based on the combination of environmental data and symptom reports. The stage-2 confusion matrix displays the actual vs. predicted health symptom categories (low, moderate, high) based on the random forest model for health risk classification. The model's accuracy was 71%, with 28 Low risk, 42 Moderate risk, and 1 high risk worker classified correctly. Most misclassifications occurred between adjacent categories, particularly between low to moderate and moderate to high, which indicates that the symptom data used in the model has overlapping characteristics for some workers. This result suggests that while the model performs well, further refinement may be needed by adding additional covariates (age, job tenure, pre-existing conditions) to improve classification accuracy, particularly for borderline cases.

This outcome demonstrates that while the model is effective, further refinement could improve classification accuracy, particularly for workers in borderline categories. This finding is consistent with challenges reported in other occupational health studies using machine learning, where symptom overlap between categories remains a limitation[12].



	symptom	f1_test	f1_train	accuracy_test	precision_test	recall_test
8	Symptom_Fatigue	0.992593	0.990654	0.99	1.000000	0.985294
4	Symptom_ShortnessofBreath	0.698795	0.662420	0.75	0.644444	0.763158
3	Symptom_Cough	0.631579	0.674487	0.65	0.517241	0.810811
0	Symptoms_RingingEar	0.591837	0.666667	0.60	0.508772	0.707317
6	Symptom_EyeIrritation	0.590909	0.657303	0.64	0.464286	0.812500
2	Symptom_SoreThroat	0.526316	0.714653	0.55	0.416667	0.714286
7	Symptom_Dizziness	0.525000	0.602076	0.62	0.428571	0.677419
5	Symptom_Flu	0.465753	0.584192	0.61	0.326923	0.809524
1	Symptom_HearingLoss	0.400000	0.546218	0.61	0.295455	0.619048

--- Classification Data Metrics (with SMOTE) ---
Mean of F1 Test (SMOTE): 0.6025
Mean of F1 Training (SMOTE): 0.6776
Saved 9 individual symptom models to /content/rf_models_symptoms_smote.pkl

Figure 4. Classification Model Metrics with SMOTE for Health Symptoms

Figure 4 presents the classification metrics for each health symptom category predicted using the synthetic minority over sampling technique (SMOTE). The metrics include F1 scores, accuracy, precision, and recall for each symptom, both in the test dataset and during the training phase. As observed, the highest F1 score is for fatigue (0.992593), indicating that the model performs exceptionally well in classifying this symptom. The lowest F1 score is for hearing loss (0.400000), suggesting that there is room for improvement in predicting this symptom. The overall F1 test score for the model using SMOTE is 0.6025, indicating that the model achieved reasonable balance between precision and recall for multiple symptoms. The model was saved as rf_models_symptoms_smote.pkl for future use and further testing.

The results of this study offer a clear answer to the research problem, which aimed to predict and evaluate occupational health risks in the metal casting industry based on environmental exposure data (noise and particulate matter). The random forest model not only successfully predicted health risks based on environmental exposure, but it also provides actionable insights for improving workplace safety. The results suggest that engineering controls, such as noise reduction technologies, ventilation improvements, and respiratory protection, should be implemented in molding and finishing to mitigate risks. Administrative measures such as task rotation and scheduled rest breaks should also be considered to reduce exposure and prevent adverse health outcomes.

4. Conclusion

This study applied two-stage machine learning model based on the random forest algorithm to predict the occupational health risk in the metal casting industry by integrating environmental exposure data and workers' health symptoms. The results demonstrated that the first-stage model, which predicted environmental risk levels from noise and particulate measurements, achieved outstanding performance with an R^2 value of 0.998 and an accuracy of 99%. The second-stage model, which classified workers' health symptoms, achieved a realistic and reliable accuracy of 71%, confirming the model's capability to capture complex relationships between exposure and physiological responses. $PM_{2.5}$ and PM_{10} were identified as the most influential variables, followed by noise intensity, indicating that respiratory exposure plays a dominant role in determining occupational health risk levels. The integration of machine learning with traditional HRA enhances predictive capacity, enabling early identification of workers at high risk and supporting preventive decision making in industrial settings. Overall, this research demonstrates that data driven risk modeling can serve as a practical, explainable, and cost effective tool for improving occupational health and safety management in small and medium-sized foundry enterprises in Indonesia.

References

[1] G. S. Hadi, L. A. Sadat, F. A. Metekohy, S. Fatimah, and E. L. Rauf, "The Role of Occupational Health and Safety Regulations in Preventing Work-related Injuries and Diseases: A Global Perspective," *The Journal of Academic Science*, vol. 2 No 1, 2025.

[2] D. Duan, P. Leng, X. Li, G. Mao, A. Wang, and D. Zhang, "Characteristics and occupational risk assessment of occupational silica-dust and noise exposure in ferrous metal foundries in Ningbo, China," Feb. 2023.

[3] M. Z. Zaman, A. Syafiuddin, A. H. Z. Fasya, and A. A. Adriansyah, "Literature Review: Jenis Penyakit Akibat Kerja, Penyebabnya Dan Mekanisme Penyebaran Dalam Industri," *Jurnal Kesehatan MAsyarakat (e-Journal)*, vol. 10 No 4, no. 57, pp. 511–517, Jul. 2022

[4] X. Chen, F. Yang, S. Cheng, and S. Yuan, "Occupational Health and Safety in China: A Systematic Analysis of Research Trends and Future Perspectives," Oct. 01, 2023, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/su151914061.

[5] C. Bartolina *et al.*, "Summary of the 2016 World Health Organization Report and 2021 Compendium on environmental diseases," *Working Paper of Public Health*, vol. 11, 2023.

[6] E. Tompa *et al.*, "A systematic literature review of the effectiveness of occupational health and safety regulatory enforcement," Nov. 01, 2016, *Wiley-Liss Inc.* doi: 10.1002/ajim.22605.

[7] M. Sahri, S. Y. Arini, F. Jannah, and M. Amin, "Occupational Exposure Assessment of Fine Particulate Matter (PM2.5) and Respirable Crystalline Silica in the Ceramic Industry of Indonesia," *Atmosphere (Basel)*, vol. 16, no. 10, p. 1125, Sep. 2025, doi: 10.3390/atmos16101125.

[8] E. C. Nyanza, S. O. Jackson, L. Magoha, P. Chilipweli, J. Joshua, and M. T. Madullu, "Perceived occupational health risks, noise and dust exposure levels among street sweepers in Mwanza City in Northern Tanzania," *PLOS Global Public Health*, vol. 4, no. 2, Feb. 2024, doi: 10.1371/journal.pgph.0002951.

[9] I. E. Agbehadji and I. C. Obagbuwa, "Explainable Artificial Intelligence and Machine Learning for Air Pollution Risk Assessment and Respiratory Health Outcomes: A Systematic Review," *Atmosphere (Basel)*, vol. 16, no. 10, p. 1154, Oct. 2025, doi: 10.3390/atmos16101154.
[10] B. S. Fakinle, D. O. Oke, J. A. Sonibare, A. J. Adewale, A. O. Adetoyese, and O. O. Fasuuhan, "Assessment of Factory Workers Exposure to Particulate Matter Fractions using Exceedance Factor and Pollution Standard Index," in *2024 International Conference on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG)*, IEEE, Apr. 2024, pp. 1–7. doi: 10.1109/SEB4SDG60871.2024.10630255.
[11] K. Stødle, R. Flage, S. D. Guikema, and T. Aven, "Data-driven predictive modeling in risk assessment: Challenges and directions for proper uncertainty representation," *Risk Analysis*, vol. 43, no. 12, pp. 2644–2658, Dec. 2023, doi: 10.1111/risa.14128.
[12] D. Nadler, "Machine Learning in Occupational Health and Safety: A Review of Knowledge Gaps," Nov. 06, 2024. doi: 10.20944/preprints202411.0464.v1.
[13] P. Manini, G. De Palma, and A. Mutti, "Exposure assessment at the workplace: Implications of biological variability," *Toxicol Lett*, vol. 168, no. 3, pp. 210–218, Feb. 2007, doi: 10.1016/j.toxlet.2006.09.014.
[14] K. L. Holt, "Predictive Modeling of Occupational Exposure Using Machine Learning and Environmental Sensor Data," *Journal of Exceptional Multidisciplinary Research*, vol. 2, no. 1, pp. 82–89, May 2025, doi: 10.69739/jemr.v2i1.617.
[15] N. Elsayed, S. Abd Elaleem, and M. Marie, "Improving Prediction Accuracy using Random Forest Algorithm," Jan. 2024. [Online]. Available: www.ijacsa.thesai.org
[16] S. Lee, L. Liu, R. Radwin, and J. Li, "Machine Learning in Manufacturing Ergonomics: Recent Advances, Challenges, and Opportunities," *IEEE Robot Autom Lett*, vol. 6, no. 3, pp. 5745–5752, Jul. 2021, doi: 10.1109/LRA.2021.3084881.
[17] W. Susihono and I. P. Gede Adiatmika, "Assessment of inhaled dust by workers and suspended dust for pollution control change and ergonomic intervention in metal casting industry: A cross-sectional study," *Heliyon*, vol. 6, no. 5, May 2020, doi: 10.1016/j.heliyon.2020.e04067.
[18] Q. Huang *et al.*, "Occupational health risk assessment of workplace solvents and noise in the electronics industry using three comprehensive risk assessment models," Mar. 2023.
[19] C. Mitrakas, A. Xanthopoulos, and D. Koulouriotis, "Techniques and Models for Addressing Occupational Risk Using Fuzzy Logic, Neural Networks, Machine Learning, and Genetic Algorithms: A Review and Meta-Analysis," Feb. 01, 2025, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/app15041909.
[20] Sugiyono, *Metode Penelitian Kuantitatif, Kualitatif, dan R & D*, 2nd edition. Penerbit Alfabeta, 2019.
[21] S. B. Budi, "A Study on Assessment of Noise Exposure in the Port Industry: Implications for Occupational Health and Safety," *Galore International Journal of Health Sciences and Research*, vol. 9, no. 1, pp. 19–25, Feb. 2024, doi: 10.52403/gijhsr.20240103.
[22] A. K. Guha and S. Gokhale, "Urban workers' cardiovascular health due to exposure to traffic-originated PM2.5 and noise pollution in different microenvironments," *Science of The Total Environment*, vol. 859, p. 160268, Feb. 2023, doi: 10.1016/j.scitotenv.2022.160268.
[23] M. Sadat-Mohammadi, S. Shakerian, Y. Liu, S. Asadi, and H. Jebelli, "Non-invasive physical demand assessment using wearable respiration sensor and random forest classifier," *Journal of Building Engineering*, vol. 44, p. 103279, Dec. 2021, doi: 10.1016/j.jobe.2021.103279.
[24] F. Majidi, Y. Khosravi, and kamalad-D. Abedi, "Determination of the Equivalent Continuous Sound Level (Leq) in Industrial Indoor Space Using GIS-based Noise Mapping," *Journal of Human, Environment, and Health Promotion*, vol. 5, no. 2, pp. 50–55, Jun. 2019, doi: 10.29252/jhehp.5.2.1.
[25] B. Roberts, N. S. Seixas, B. Mukherjee, and R. L. Neitzel, "Evaluating the Risk of Noise-Induced Hearing Loss Using Different Noise Measurement Criteria," *Ann Work Expo Health*, vol. 62, no. 3, pp. 295–306, Mar. 2018, doi: 10.1093/annweh/wxy001.
[26] A. U. Abidin, A. L. Munawaroh, A. Rosinta, A. T. Sulistiyani, I. Ardianta, and F. M. Iresha, "Environmental health risks and impacts of PM2.5 exposure on human health in residential areas, Bantul, Yogyakarta, Indonesia," *Toxicol Rep*, vol. 14, Jun. 2025, doi: 10.1016/j.toxrep.2025.101949.
[27] G. N. Ferrari, G. C. L. Leal, P. C. Ossani, and E. V. C. Galdamez, "Investigation of the usage of machine learning to explore the impacts of climate change on occupational health: a systematic review and research agenda," 2025, *Frontiers Media SA*. doi: 10.3389/fpubh.2025.1578558.
[28] Q. Gong, L. Xie, D. Dou, K. Wang, and G. Zhang, "A random forest model for exertional heat illness prediction in the power grid work place," in *2022 International Conference on Frontiers of*

Communications, Information System and Data Science (CISDS), IEEE, Nov. 2022, pp. 60–63. doi: 10.1109/CISDS57597.2022.00017.

[29] A. Badhoutiya, R. P. Verma, A. Shrivastava, K. Laxminarayanamma, A. L. N. Rao, and A. K. Khan, "Random Forest Classification in Healthcare Decision Support for Disease Diagnosis," in *2023 International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIIHI)*, IEEE, Dec. 2023, pp. 1–7. doi: 10.1109/ICAIIHI57871.2023.10489244.

[30] M. Rahmiani Iranshahi, M. Aliabadi, R. Golmohammadi, A. Soltanian, and M. Babamiri, "Empirical prediction model of psychophysiological responses of workers with respect to noise exposure based on random forest," *Noise and Vibration Worldwide*, vol. 53, no. 6, pp. 290–299, Jun. 2022, doi: 10.1177/09574565221093258.

[31] M. Basner *et al.*, "Auditory and non-auditory effects of noise on health," *The Lancet*, vol. 383, no. 9925, pp. 1325–1332, Apr. 2014, doi: 10.1016/S0140-6736(13)61613-X.