Risk factors associated with work-related musculoskeletal disorders among dumper operators: A machine learning approach

Mohith Bekal Kar *, Mangalpady Aruna, Bijay Mihir Kunar

Department of Mining Engineering, NITK Surathkal, India

ARTICLE INFO

Keywords:
- Work-related musculoskeletal disorders
- Machine learning techniques
- Dumper operators
- Epidemiological study

ABSTRACT

Aims: This study aimed to determine the risk factors associated with work-related musculoskeletal disorders (WRMSDs) among dumper operators working in Indian iron ore mines.

Methods: A total of 246 dumper truck operators meeting inclusion and exclusion criteria were chosen for data collection. A self-report custom and the standard Nordic questionnaire were used for collecting data about risk factors and WRMSDs. The data were pre-processed and analyzed using machine learning (ML) algorithms (such as logistic regression (LR), support vector machines (SVM), decision trees (DT), gradient boosting machine (GBM) and random forest (RF)).

Results: RF model was found to outperform the other algorithms with high accuracy (0.71), precision (0.75), recall (0.78), F1 score (0.76), and area under the receiver operating characteristic curve (0.82). The mean rank of the risk factors showed that age is the most critical parameter, followed by awkward posture, experience in mines, job demand, alcohol consumption, smoking cigarettes, work design, and marriage status.

Conclusion: Overall, the study provides valuable insights into the risk factors associated with WRMSDs among dumper operators and suggests that measures should be taken to address these risk factors to prevent WRMSDs in the dumper operator population.

1. Introduction

Work-related musculoskeletal disorders (WRMSDs) are a significant occupational health problem affecting workers worldwide. The prevalence of WRMSDs is increasing rapidly in many countries (Bureau of Labor Statistics, 2020). The study conducted by the Indian Council of Medical Research (ICMR) showed that 65% of the workers in the construction and mining industry suffer from WRMSDs. The economic burden of WRMSDs is significant, with direct and indirect costs estimated to be billions of dollars annually.

Many researchers have developed statistical models to determine the relationship between risk factors and WRMSDs. These methods implicitly assumed that each risk factor has a linear association with the outcomes. The complex relationships between nonlinear interaction factors might be oversimplified, potentially losing related information. Further, when the number of variables increased, the hypothesis testing method became complicated. But, in contrast, machine learning (ML) algorithms can learn the nonlinear interactions iteratively and handle many variables.

Given the increasing amount of health data generated, ML algorithms in epidemiology studies have gained popularity in recent years. Epidemiology is concerned with understanding the distribution and determinants of disease in populations. Therefore, the ML technique is well-suited for the epidemiology study. Several ML algorithms, such as logistic regression (LR), support vector machines (SVM), decision trees (DT), gradient boosting machine (GBM), and random forest (RF), have been used in epidemiology studies. These ML algorithms can help predict disease occurrence, identify risk factors, and inform public health interventions.

So far, no researchers have investigated the risk factors associated with WRMSDs in the dumper operator population using the ML approach. Therefore, this study will bridge this gap by using a range of ML algorithms for predicting the risk of WRMSD among dumper operators working in Indian iron ore mines.

* Corresponding author.
E-mail addresses: mohithbekalkar.217mn500@nitk.edu.in (M.B. Kar), arunm@nitk.edu.in (M. Aruna), bkunar@nitk.edu.in (B.M. Kunar).

https://doi.org/10.1016/j.cegh.2023.101438
Received 13 June 2023; Received in revised form 27 September 2023; Accepted 11 October 2023
Available online 16 October 2023
2213-3984/© 2023 The Author(s). Published by Elsevier B.V. on behalf of INDIACLEN. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
2. Methods

2.1. Background of the mines

The iron ore mine considered in this study covers an area of 62 Ha with a production capacity of 6 Mt per year. The mine was working in two shifts of 8 h each. The ore was transported from the pit to the dumping point using dumpers of 30 tons capacity. The distance between the loading and unloading point was about 1 km. The dumper operators performed an average of 11–12 cycles per shift (one cycle comprises loaded, full capacity travel, unloading, and empty travel).

Land utilized for infrastructure and road facilities of case study mines is 25.61 Ha. At the time of the visit, mine had two dead dumps, two active dumps, and one settling pond. The mine used eleven excavators, nine wheel loaders, 152 dumpers, six dozers, and four water tankers to extract ore from the mine site.

2.2. Study design

This cross-sectional study included 246 dumper operators selected from the case study mine who meet the inclusion (i.e., age between 18 and 56 years, at least 6 months of professional driving experience) and exclusion criteria (i.e., no history of injuries). Data were collected through a self-reported custom questionnaire (Table 1) to obtain information about age, driving experience, job demands, posture, medication usage, cigarette smoking, alcohol consumption, work design, and marital status. Additionally, the standard Nordic questionnaire was used to collect information on WRMSD issues encountered by the dumper operators. The collected data was pre-processed (i.e., data entry, data cleaning, categorical variables encoding, handling outliers, data scaling, and feature creation) and the missing data was imputed. For the categorical variables, the mode of the variable for entire dataset was determined and substituted in the place of missing values. Similarly for the continuous variables, the mean value was calculated and was used to impute the missing values. Further, one-hot encoding method was employed for coding the categorical variables. Subsequently, the data was divided into two sets, with 198 datasets (80%) allocated for training the ML model and the remaining 50 datasets (20%) reserved for testing the ML model. The ML models (such as LR, SVM, DT, GBM, and RF) were then developed and validated using training and testing datasets. The model building and data analysis was performed with the help of scikit-learn (version 1.3) Python library. Fig. 1 demonstrates the flowchart of the study design.

2.3. Ethical considerations

Approval for this study was obtained from the institutional review board (Ref. No. MIN/ED/133/2022). All methods were performed in accordance with the relevant guidelines and regulations set by the institutional review board. The participants were informed about this study, and consent was obtained from them. Confidentiality of the participant’s personal and medical information was ensured.

3. Results

This study determined the importance of the risk factors based on the model coefficient (in the case of LR and SVM) and feature importance scores (in the case of DT, GBM, and RF). Similarly, the best model for predicting the WRMSD among the dumper operators was determined based on the model’s accuracy, precision, recall score, F1 score, and area under the Receiver Operating Characteristic curve (ROC).

3.1. Logistic regression

The LR model was developed, and hyperparameter tuning was conducted using the grid search method (as shown in Table 2) to obtain a high-performance LR model. The result of LR indicated that among the risk factors, ‘experience in mines’, ‘medicine’, and ‘work design’ had a negative impact on the WRMSD (Table 3). On the other hand, the ‘age’, ‘smoking cigarettes’, ‘alcohol consumption’, ‘marriage status’, ‘awkward posture’, and ‘job demand’ positively impacted the outcome. The most significant parameters were awkward posture (0.82), alcohol consumption (0.58), medicine (−0.87), and job demand (0.47).

When the performance of the LR model was evaluated on the test dataset, it showed that the model had an accuracy of 0.64, a precision score of 0.69, a recall score of 0.66, and an F1 score of 0.68. This accuracy score indicates that the LR model correctly predicted the target class for 64% of the instances, which is a moderate level of performance. The precision score of 0.69 suggests that the model correctly predicted the target class 69% of the time when it made a positive prediction. However, the recall score of 0.66 indicates that the model could only identify 66% of the instances that belonged to the target class. The F1 score, a balanced measure of precision and recall, was 0.68, suggesting that the model’s overall performance was moderate, with a room for improvement in correctly identifying all instances of the target class.

3.2. Support vector machine

Similar to LR, in the SVM model was developed and hyperparameter was tuned using grid search method (Table 2). The most positive and significant coefficient was associated with alcohol consumption (1.15), awkward posture (1.03), and job demand (0.59), indicating that this variable has the strongest and positive influence on the WRMSDs. Interestingly, experience in mines (−0.41), medicine (−1.17), work design (−0.18), and marriage status (−0.08) were found to have a negative impact on the WRMSDs (Table 3).

The SVM model performance was also evaluated with accuracy, precision, recall, and F1 score metrics. The results showed that the model achieved an accuracy of 0.64, a precision score of 0.76, a recall score of 0.54, and an F1 score of 0.63. The relatively high precision score
indicates that the model is better at identifying true positives than avoiding false positives. However, the recall score is lower, showing that the model misses a significant number of actual positive cases. The F1 score suggested that the model’s overall performance is moderate.

### 3.3. Decision tree

In this study, the DT model was built using Gini impurity as the splitting criterion and parameters obtained after hyperparameters tuning using grid search method (Table 2). The results showed that age (0.42), experience in mines (0.21), and job demand (0.079) were the most significant risk factors associated with WRMSD (Table 3). When the model was tested using the test dataset, it achieved an accuracy of 0.63, a precision score of 0.72, a recall score of 0.77, and an F1 score of 0.74. The recall score is the highest among all the evaluation metrics, indicating that the model is good at identifying the true positives. The precision score is relatively high, indicating that the model is effective at avoiding false positives. The F1 score is a balanced measure of precision and recall, and suggests that the model’s overall performance is good.

### 3.4. Gradient boosting machine

In this study the GBM model was build by training the DT sequentially, with each tree correcting the errors of the combined ensemble of the existing trees. The goal is to optimize classification results through multiple iterations and address the weaknesses of the classifier. The results showed that age (0.36), experience in mines (0.27), and awkward posture (0.14) are the most prominent risk factors associated with WRMSDs. The results revealed that the GBM model had an accuracy was 0.61, and its precision score, recall score, and F1 score were 0.77, 0.72, and 0.75, respectively. The precision score was higher than the recall score, indicating that the model was better at correctly identifying true positive cases than avoiding false negative ones. However, the F1 score, which considers precision and recall, suggests that the model’s overall performance is fair.

### 3.5. Random forest

RF is an algorithm that combines bagging ensemble learning theory with a random subspace approach. RF generates many decision trees for the random data at training time. Each tree provides a classification, and the RF chooses the classification with the most votes. Similar to DT model, the RF model ranked age (0.42), work experience (0.29), and job demand (0.09) as the critical parameters that are associated with the WRMSDs. The results showed that the model had an accuracy of 0.71, a precision score of 0.75, a recall score of 0.78, and an F1 score of 0.76. The model performed well in accuracy and precision, indicating that it correctly classified a high percentage of positive samples. The recall score suggests that the model also identified a significant number of true positives, although it may have missed some positive samples. The F1 score indicated that the model’s overall performance is moderate.

### 3.6. Comparing the performance of ML algorithms

In this study, the Receiver Operating Characteristic curve (ROC) was used to compare the performance of five different ML algorithms. It was found that RF had the highest area under the curve value (0.82), followed by GBM (0.79), DT (0.76), SVM (0.73), and LR (0.69). These
results suggest that RF is the most accurate algorithm for the present dataset (Fig. 2). In addition, while comparing the performance metrics (Table 4), the RF model had relatively high metric values compared to the other models.

### 3.7. Mean rank of the risk factors

The rank of the risk factors is highly dependent on the ML model used. The mean rank of the risk factors was determined to get a generalized idea about the importance of the risk factors (Fig. 3). It showed that age is highly associated with WRMSDs, followed by awkward posture, smoking cigarettes, work design, and marriage status.

### 3.8. Reliability of the questionnaire data

The stability and consistency of the questionnaire data over time were assessed by re-administering it to a subset of the sample after a 9-month interval. A total of 20% of the participants were selected to participate in the retest. By comparing the responses at Time 1 and Time 2, the consistency of the measurements over the extended period was tested. The result indicated a strong (ranging from 0.82 to 0.91) and statistically significant correlation between the responses to the items of the custom questionnaire at Time 1 and Time 2.

### 4. Conclusions

The study aimed to determine the risk factors associated with WRMSDs among dumper operators in a mine extracting iron ore. The data were collected through a self-report custom and standard Nordic questionnaire from 246 participants. The reliability of the questionnaire data was cross-checked with the test-retest method. The correlation coefficient of questionnaire items at Time 1 and Time 2 was above 0.82. The raw data from the self-reported questionnaire was pre-processed to ensure its quality and prepared for analysis. The pre-processed data were analyzed using ML algorithms using the scikit-learn library.

The ML model performance was measured using accuracy, precision, recall, F1 score, and ROC. The study found that the RF model outperformed the other model (SVM, DT, GBM, and LR) due to relatively
The study showed that age is highly associated with WRMSD, followed by awkward posture, work experience, job demand, alcohol consumption, smoking cigarettes, work design, and marital status. The findings of this study corroborate with the results of the previous research works on various occupational groups. For a study by He et al. showed that age is a significant risk factor associated with the prevalence of WRMSDs. Similarly, another study by Sharma & Singh found that work-related factors such as job demand, job control, and work-related stress were significantly related to the prevalence of WRMSDs. In addition, this study addresses the Bradford Hill Criteria such as strength of the association, consistency of association, specificity of association, and coherence of the association. Overall, the study provided a comprehensive analysis of the data collected from the study participants. The findings of this research can be further strengthened by considering more diverse samples from different mine sites.

**Ethical considerations**

Ethical approval for this study was obtained from the institutional review board (committee set by National Institute of Technology Karnataka). All methods were performed in accordance with the relevant guidelines and regulations set by the institutional review board. The participants were informed about this study, and consent was obtained from them. Confidentiality of the participant’s personal and medical information was ensured.

**Funding source**

The authors wish to acknowledge the support of the Science and Engineering Research Board, India, for funding the research work.

**Author contributions**

All authors have equally contributed for this study.

**Conflict of interest**

The authors declare that there is no conflict of interest.

**Acknowledgement**

None.

**References**