Industrial Accident Analysis and Predictive Models for Workplace Hazard Prevention

Rao Eswara Veera Raghava* and Pandey Ashutosh*
Department of Civil Engineering, Kalinga University, Raipur, INDIA
*ku.eswaraveeraraghavarao@kalingauniversity.ac.in; ku.ashutoshpandey@kalingauniversity.ac.in

Abstract

Accidents are a constant problem in numerous enterprises and they greatly affect workers and project results. This study aims to find out what caused these crashes, focusing on how worker-related, environmental and managerial factors all interact with each other. The goal is to find the main factors affecting Workplace Hazard Prevention (WHP) and make a model that can predict the future to lower risks. This study uses an ensemble machine learning (EML) approach to show Industrial Accident Analysis and Predictive Models for Workplace Hazard Prevention (IAA-PM-WHP). An analysis is conducted on a publicly accessible collection of 65,518 workplace injury reports from the Occupational Safety and Health Administration (OSHA), using four distinct ML models.

This study suggested a way to build a model that takes into account three important factors: "type of damage," "kind of event," and "harmed organ." The EML model integrates predictions from four fundamental ML methodologies via soft voting. Among classic ML models, the RF method had the greatest accuracy (0.89), indicating robust overall prediction power. The EML method outperformed all models, attaining the greatest accuracy (0.92), precision (0.99), recall (0.899), F1-score (0.94) and AUC (0.92).

Keywords: Workplace Hazard, Ensemble Machine Learning, Accident, Industries, Prediction.

Introduction

Several industrial safety management systems have been implemented and enhanced in recent decades, yet workplace safety remains precarious and inadequate. Specifically, several disciplines and tasks are executed concurrently and cohesively in the construction sector, accompanied by many hazardous elements. Consequently, safety management within the construction sector is challenging due to the intricate nature of many operations and the participation of diverse stakeholders. Furthermore, most tasks are executed by people; hence, methods for predicting workplace incidents by straightforward correlations and subsequently implementing safety measures to avert them, are constrained. Consequently, comprehensive research has been undertaken in recent decades to enhance the safety performance of building sites[10]. Many investigators have performed analytical investigations across diverse domains using historical accident information; nonetheless, some limitations have been identified in the study of industrial accident information[7]. The personal and subjective views of the individual compiling the occupational incidents report are evident in the data; consequently, it is challenging to analyze and represent the features of occupational incident information in the construction sector, which is generated without a systematic procedure and encompasses numerous variables and standards.

The framework of industrial accident analysis comprises of a combination of mixed elements, such as numerical and category text depictions and missing information. The multitude of variable types and the structure of various categories complicate the interpretation of findings from data components, allowing for very limited correlations between characteristics[4]. The following inferences may be derived from the current study findings. The WHP contains several variables and values, complicating data processing, characteristic reflection and correlation interpretation.

Nevertheless, if the variables are too diminished, their attributes are forfeited, making the derivation of significant conclusions impossible. Consequently, the categories and value ranges for appropriate variables must be defined to use data including additional accident data effectively. Additionally, it is essential to build a procedure that effectively captures patterns in industrial accidents, along with a predictive system that learns from historical incidents to mitigate the risk of future occurrences.

Review of Literature

It is common for accidents to happen at work in the building industry, which is a high-risk field. A study that was just released, looks into how ML and analysis forecasts can be used to make the workplace safer. Cavalcanti et al[3] did a lot of research on ML technologies in WHP and stressed the need for more studies in this area. Goldberg[8] discussed how building information modelling (BIM) could help make building workers safer and suggested real-world uses such as safety training and risk evaluation. Gao et al[7] used machine learning and the five major psychological theories to build a model that can predict how building workers will behave regarding safety and find the workers who are most likely to do something dangerous.

Fargnoli et al[6], the authors improved their study by using ML to predict the safety effects of driverless buildings. This made a big difference in how well they could predict injuries. These studies show that machine learning and predictive models might be able to make workplaces safer. Using a database from the Department of Labor and

Employment of the Korean Republic, researchers made a prediction model that used machine learning to determine how likely fatal accidents will happen on building sites.

The study found that the RF method had the most accurate predictions, with the month of the event and the number of jobs being important factors. A different study project looked at how to predict what would happen in crashes in China by using eight different methods to look at 16 important factors. The study focused on "Kind of accident" and "Accident reporting and management" as important factors. Naive Bayes (NB) and Logistic Regression (LR) had the best F1 scores on the raw database.

Mining, which has safety problems similar to building, has used modeling to make predictions. We used machine learning methods like DT and artificial neural networks (ANN) to guess what would happen in workplace accidents and how many people would miss work. Narrative information gives us more information than organized information[14]. A different study[13] suggested a way to predict how well safety measures will work before building jobs start. It used a DT method with the k-Nearest Neighbors (k-NN) algorithm to find the most important factors for predicting security results such as the number of safety staff, their training, their commitment to following the rules and management commitment.

A careful review of the literature on machine learning uses in building security found trends and gaps in the field[2]. It was found that figuring out how bad a building accident is, has been studied the most. Logistic regression (LR) was used as a standard model. Scientists tested whether machine learning techniques could predict how bad accidents would be in different areas of agriculture and the models they created were very accurate[11]. This study showed how important it is for safety scientists to think about observational accident data in a mathematical way. ML has shown promise in improving building safety, with several studies showing how well different systems can predict what will happen in accidents. As this study will explain, these changes can greatly improve safety rules and lower the number of accidents in the building industry.

Material and Methods

Data Collection: The database utilized for this investigation was sourced from serious accident reports from the OSHA[1].

Preparation of information: The categories of information in this database are nature of damage, kind of incident, harmed organ, type of industry, resource type, treatment status and amputation occurrence. The Industry Classification System (ICS) categorizes industries into 20 distinct sectors: farming, forestry, mineral extraction and construction. The type of damage encompasses 10 categories that delineate the physical aspects of the injury such as superficial wounds, severe injuries and different ailments. The afflicted bodily component comprises to eight groups. Included are the trunk and both of the lower extremities. The incident or exposure delineated how the harm was sustained.

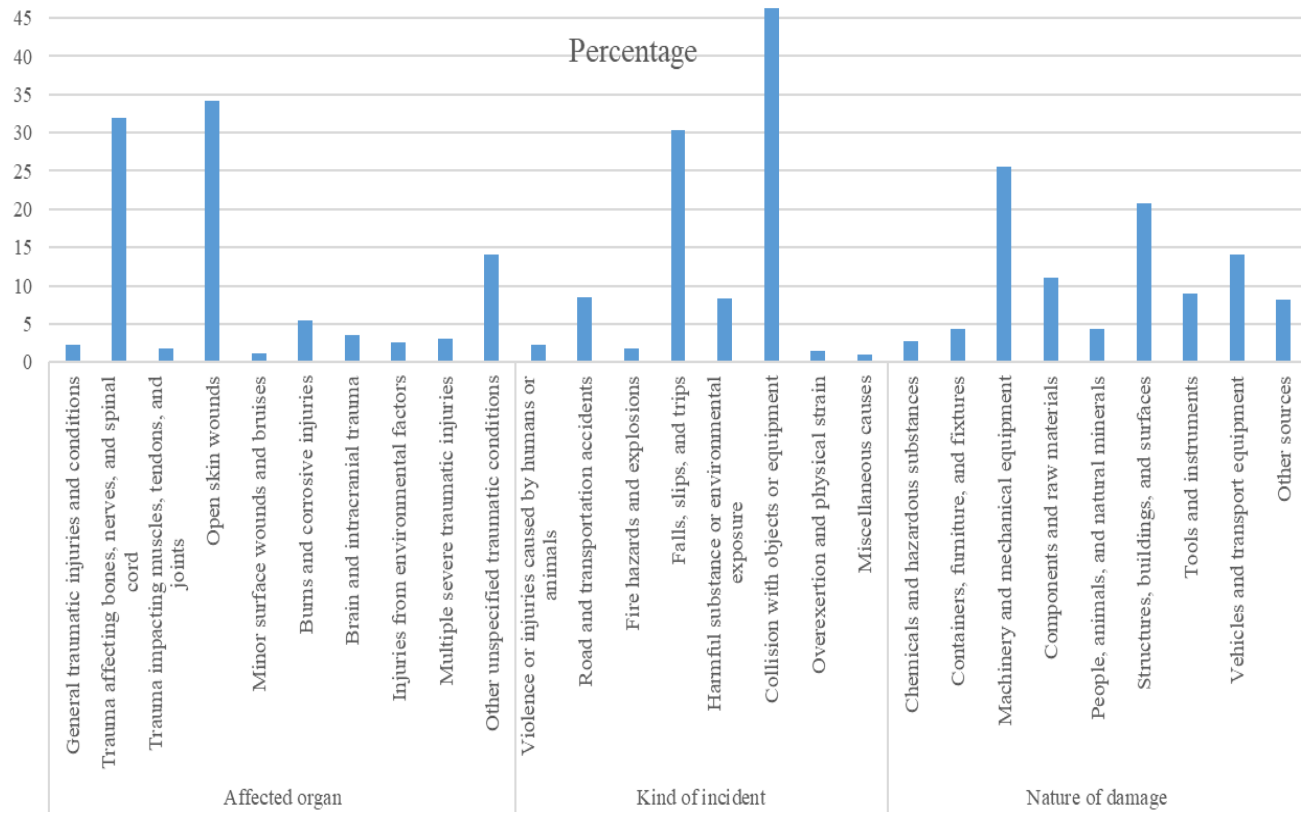

Fig. 1: Information distributions of the used parameters in workplace accidents from 2018 to 2023

There are eight kinds of incidents including falls, slips, trips and encounters with hazardous chemicals. Finally, nine sources delineate what contributed to the injury including tools, instruments and machines. Simultaneously, columns about ID number, dates, employers' locations, town, state and location were omitted. Columns such as examination and additional information were eliminated since most items were marked as of 'no value.' Additionally, all records containing empty columns were removed. The narrative column is omitted since this research focuses only on organized information.

This research uses just the highest labels: For instance, a primary classification for this incident type involves contact with items and equipment, subdividing into subcategories, such as needle-stick and struck by materials or equipment. To mitigate the limited coverage for each criterion[9], only the highest-ranking labels are considered for further examination. The non-classifiable category about the impacted body, incident type and source category is reclassified as 'Other(s).' Sixty-five thousand five hundred eighteen organized data points were used as inputs to forecast the seriousness of occupational injuries. Fig. 1 displays the information distributions of the used parameters, illustrating the proportion of damaged organs in workplace accidents from 2018 to 2023.

The architecture of the proposed IAA-PM-WHP using the EML model

Pre-processing: Preliminary data processing is a crucial phase in creating ML models. Missing values in the acquired data might distort the performance predictions of the models. This research eliminated 301 rows (0.5%) with missing columns and used the standard scaling algorithm for data normalization.

Database Splitting: The data collection is divided into two subsets: (i) the learning dataset and (ii) the testing dataset. This research used a 75:25 ratio, with 75% of the data allocated to the training data set and 25% designated as the testing dataset. The 75:25 ratio is often used in ML classification experiments since it is thought to provide high accuracy and mitigate overfitting. Figure 2 depicts the architecture of the suggested technique.

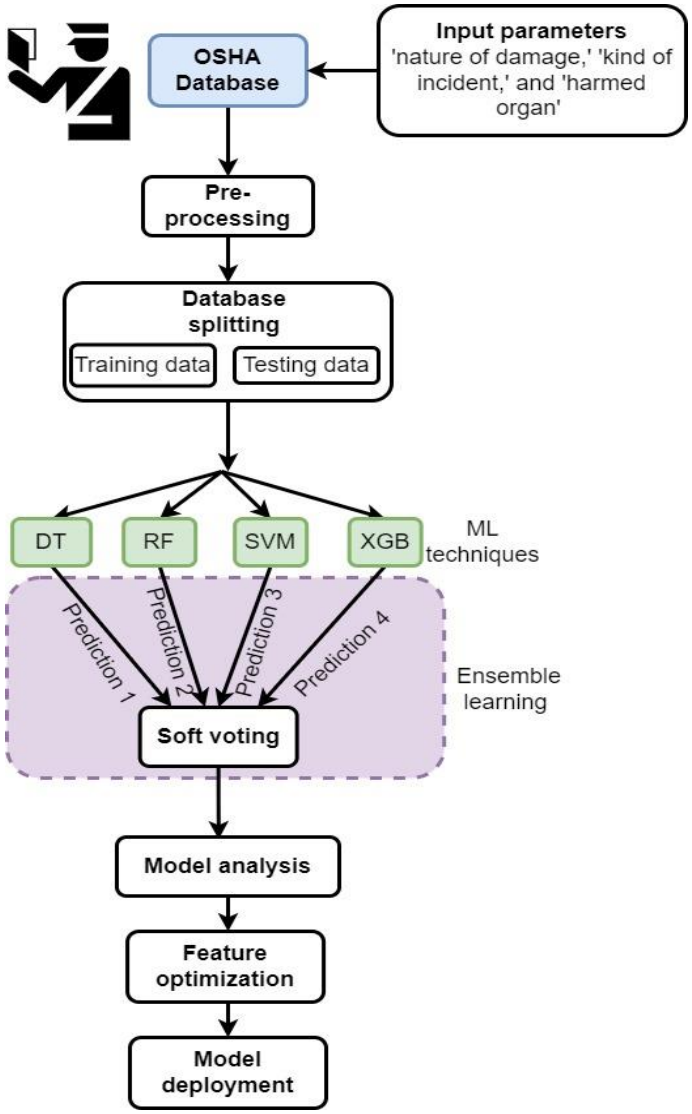

Fig. 2: Architecture of the proposed IAA-PM-WHP using EML model

Predictive Modeling: For the research, four distinct ML models have been used: SVM, DT, RF and XGB.

SVM: SVM may provide optimal transferable decision limits for data categorization. This approach transforms the initial feature space into a higher-dimensional space using a kernel function specified by the function operator. It then delineates the two categories using a hyperplane and adjusts support vectors to maximize their separation. A hyperplane is characterized as a demarcation that distinguishes the two groups. The quantity of input factors dictates the dimensions of the hyperplane in the database.

DT: The fundamental elements of a DT are: (i) the root node, which serves as the starting point of the DT model; (ii) the decision node, responsible for choice-making and branching the model into numerous paths; and (iii) the leaf node, representing the results of those choices. The categorization in DT starts with dividing the root node into the leaf node. The division persists until it attains the leaf node. The classifier identifies the feature and its associated threshold at every node to perform a split.

The database experiences a maximum reduction in entropy or impurity after splitting it. A leaf containing samples exclusively from a single class is considered optimal throughout the splitting process. In DT, the classifier analyzes the learning database to produce a tree-like decision framework, beginning with a root node and concluding with various leaves. Due to its superior understanding, DT is often used in predictive analysis of workplace accidents.

RF: RF is an EML algorithm that employs tagging as its ensemble technique and DT as its base approach, mitigating volatility and bias to enhance results. The classifier integrates multiple choice trees with a more resilient classifier, enhancing adaptability and simplifying hyperparameter tuning to mitigate overfitting concerns. In categorization tasks inside Random Forest, each tree delivers a classification or casts a 'vote.'

XGB: XGB is an efficient, adaptable and optimized ML technique derived from the gradient boosting framework. It is designed for velocity and efficacy including parallel and dispersed computing, normalization strategies and sophisticated tree-pruning algorithms to enhance precision and mitigate overfitting. XGB operates by incrementally constructing DT, with each subsequent tree rectifying its predecessors' flaws by minimizing a loss function. It accommodates categorization and regression problems and is extensively used in data science contests and practical applications including identifying scams, suggestion systems and healthcare diagnosis, owing to its exceptional prediction capability and adaptability.

EML: The EML method involves intentionally combining fundamental models to create a robust model. The ensemble method employs a synthesis of learning techniques to tackle a classification or regression problem that proves difficult for each model to resolve autonomously. EL may exceed the efficacy of an individual model. This research used soft-voting ensemble learning. We first acquired fundamental models like DT, RF, SVM and XGB by utilizing the training dataset. After the training phase, the model's efficacy has been assessed by examining its predictions against the test data. The predictions from these models provide further input to the EL, which operates as a cohesive model designed to provide the final prediction.

Feature Optimization: This step aims to evaluate and prioritize the most significant property of the workplace injury prediction model. The performance of each ML model was evaluated and the model yielding the optimal results was used to identify the significant features influencing workplace injury severity. The attribute with the highest importance score is the most crucial indicator of the model. The feature optimization phase included reconstructing the optimal performance framework using the three most significant characteristics as input parameters. The framework then performs hyperparameter optimization with the k-fold cross-validation method. K-fold cross-validation is a method used to assess the efficacy of a suggested predictive model. Fig. 3 depicts the steps involved in the feature Optimization process.

This research used a k-value of 10 and conducted 200 iterations to optimize the suggested model. Using k = 10 is prevalent in applied ML models since its practicality lowers test error rates associated with elevated bias or variation. In theory, the disparity in size between the learning set and the re-sampling sections will diminish as k grows. Simultaneously, the prejudice of the methodologies diminishes as this disparity decreases. Subsequently, the cross-validation accuracy metrics are calculated for all hyperparameter configurations. The baseline model using feature inputs will examine the mean cross-validation accuracy rating. The design with the highest accuracy rating is selected as the final model.

Results and Discussion

An analysis is conducted on a publicly accessible collection of 65,518 workplace injury reports from OSHA, using four distinct ML models: Support Vector Machine (SVM), Extreme Gradient Boosting (XGB), Decision Tree (DT) and Random Forest (RF) and proposed EML. Experiments have been conducted on a high-performance system with eight cores, 64 GB of RAM and a 100 GB drive. TensorFlow and Keras frameworks have been used in the study.

Fig. 4 displays a performance comparison between existing ML models and the proposed EML method for IAA-PM-WHP. Among classic ML models, the RF method had the greatest accuracy (0.89), indicating robust overall prediction power. The EML method outperformed all models, attaining the greatest accuracy (0.92), precision (0.99), recall (0.899), F1-score (0.94) and AUC (0.92). This shows that the EML

model is the best at identifying risks in the workplace because it is the most stable and well-rounded. Also, the XGB and SVM models had high accuracy scores of 0.985 and 0.984 respectively, showing they were good at finding dangerous work situations with few false positives.

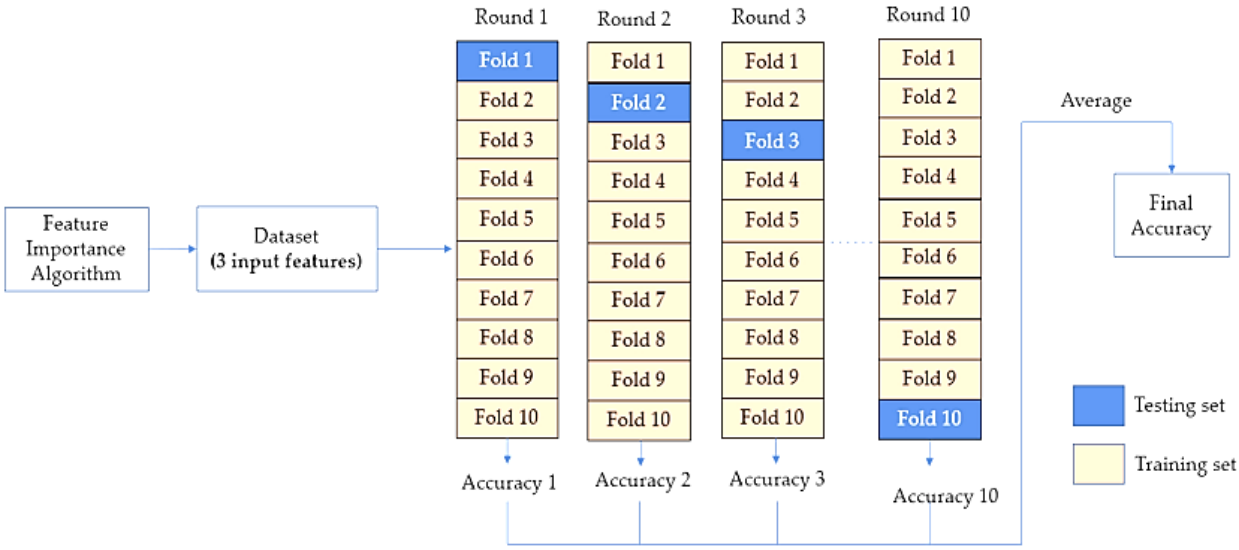

Fig. 3: Steps involved in Feature Optimization

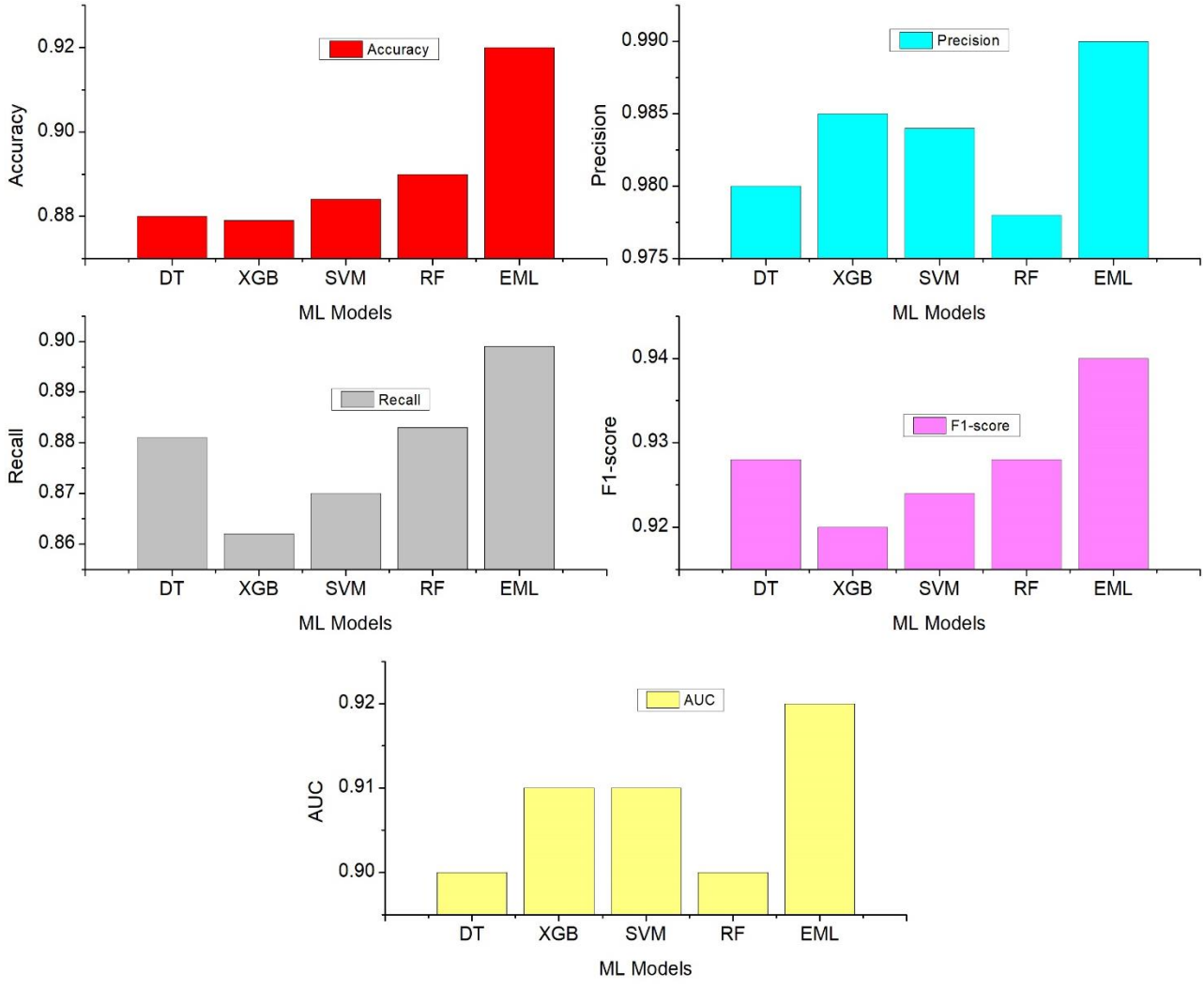

Fig. 4: Performance analysis of traditional ML and proposed EML algorithms for IAA-PM-WHP

The EML algorithm performed better than all previous models, making it the best way to investigate workplace accidents and lower the risk of workplace harm. The high memory value of EML (0.899) shows that it is good at finding the most real risk cases, which lowers the chance of missing important workplace issues. The F1-score (0.94) and AUC (0.92) also show that it is very good at predicting the future and is reliable. Even though the RF and DT models did well, they had slightly worse memory and the AUC values show that EML has a better mix of accuracy and recall. This study emphasizes the advantages of advanced ML methods for improving workplace safety. It shows how better ML can help to avoid accidents and to lower risks in the workplace.

Fig. 5 illustrates the trends in loss and accuracy throughout several iterations during the training of the EML model for IAA-PM-WHP. At iteration 0, the model starts with a big loss of 0.8 and a very low accuracy of 0.3, which shows that it is just starting to learn. The accuracy improves with each iteration, reaching 0.75 by iteration 30 and a peak of 0.98 by iteration 160.

This shows that the model correctly learns patterns from the training sample over time. At the same time, the loss keeps going down until it reaches 0.58 at iteration 60, which means that the model is getting better at making predictions. Still, the loss does not always go down; it goes up and down in small amounts, which could be because of differences in how the training works, the complexity of the data, or changes in the learning rate.

As training goes on, the model gets much better at accuracy. However, there are changes in loss and accuracy at certain rounds, especially at 50, 100 and 180 where accuracy temporarily drops. This could be because they were either too fit or loose at different training points. Even with these changes, the model becomes stable at many points especially at rounds 90, 110 and 140, where accuracy consistently scores at 0.8. The highest level of accuracy, 0.98, at iteration 160 shows that the model is well-tuned and can make good predictions. The different loss numbers suggest that more tuning of hyperparameters or regularization methods could stabilize the model.

Conclusion

This work introduces Industrial Accident Analysis and Predictive Models for Workplace Hazard Prevention (IAA-PM-WHP) using an Ensemble Machine Learning (EML) methodology. Our study uses four ML models—SVM, DT, RF and XGB to look at a freely available dataset that has 65,518 reports of work-related accidents from the OSHA database. This research showed a new way to improve model development by focusing on three important factors: "type of damage," "kind of occurrence," and "affected organ." The EML model combines predictions from four important machine learning methods using "soft voting".

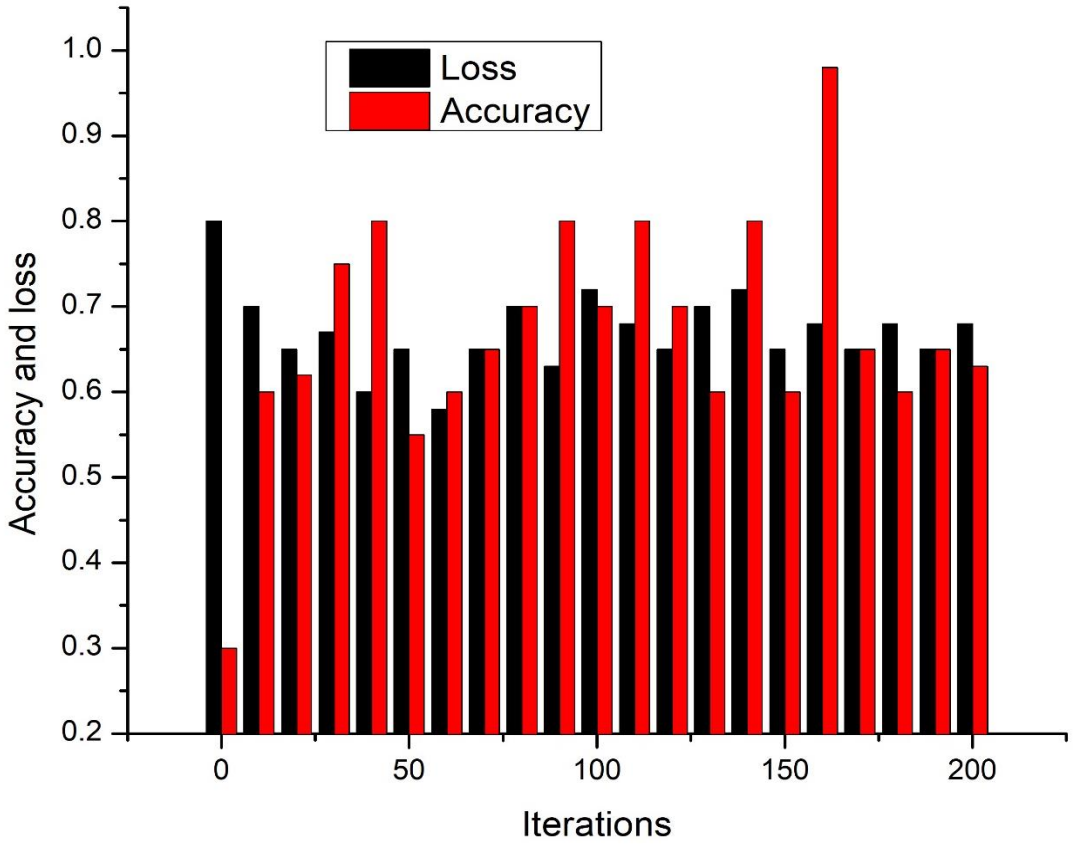

Fig. 5: Loss and accuracy of training database used in the proposed EML for IAA-PM-WHP

The random forest method had the best accuracy (0.89) among standard machine learning models showing that it is very good at making predictions in general. The EML method did better than the others, with the best accuracy (0.92), precision (0.99), recall (0.899), F1-score (0.94) and AUC (0.92). In many places, the suggested EML model becomes stable. This is especially true at iterations 90, 110 and 140 where accuracy always stays at 0.8. The highest level of accuracy (0.98 at iteration 160) shows that the model is well-tuned and is good at making predictions. The changes in loss values show that the model could be more stable with more improvements, like hyperparameter tuning or regularization methods.

References

1. Abbasianjahromi H. and Aghakarimi M., Safety performance prediction and modification strategies for construction projects via machine learning techniques, *Engineering, Construction and Architectural Management*, **30(3)**, 1146-1164 **(2023)**

2. Baker H., Hallowell M.R. and Tixier A.J.P., Automatically learning construction injury precursors from text, *Automation in Construction*, **118**, 103145 **(2020)**

3. Cavalcanti M., Lessa L. and Vasconcelos B.M., Construction accident prevention: a systematic review of machine learning approaches, *Work*, **76(2)**, 507-519 **(2023)**

4. Cheng M.Y., Kusoemo D. and Gosno R.A., Text mining-based construction site accident classification using hybrid supervised machine learning, *Automation in Construction*, **118**, 103265 **(2020)**

5. Choi J., Gu B., Chin S. and Lee J.S., Machine learning predictive model based on national data for fatal accidents of construction workers, *Automation in Construction*, **110**, 102974 **(2020)**

6. Fargnoli M. and Lombardi M., Safety climate and the impact of the COVID-19 pandemic: an investigation on safety perceptions among farmers in Italy, *Safety*, **7(3)**, 52 **(2021)**

7. Gao Y., González V.A., Yiu T.W., Cabrera-Guerrero G., Li N., Baghouz A. and Rahouti A., Immersive virtual reality as an empirical research tool: exploring the capability of a machine learning model for predicting construction workers' safety behaviour, *Virtual Reality*, **26(1)**, 361-383 **(2022)**

8. Goldberg D.M., Characterizing accident narratives with word embeddings: Improving accuracy, richness and generalizability, *Journal of Safety Research*, **80**, 441-455 **(2022)**

9. https://www.osha.gov/severeinjury

10. Kakhki F.D., Freeman S.A. and Mosher G.A., Evaluating machine learning performance in predicting injury severity in agribusiness industries, *Safety Science*, **117**, 257-262 **(2019)**

11. Khairuddin M.Z.F., Lu Hui P., Hasikin K., Abd Razak N.A., Lai K.W., Mohd Saudi A.S. and Ibrahim S.S., Occupational injury risk mitigation: machine learning approach and feature optimization for smart workplace surveillance, *International Journal of Environmental Research and Public Health*, **19(21)**, 13962 **(2022)**

12. Koc K. and Gurgun A.P., Stakeholder-associated life cycle risks in construction supply chain, *Journal of Management in Engineering*, **37(1)**, 04020107 **(2021)**

13. Lee J.Y., Yoon Y.G., Oh T.K., Park S. and Ryu S.I., A study on data pre-processing and accident prediction modelling for occupational accident analysis in the construction industry, *Applied Sciences*, **10(21)**, 7949 **(2020)**

14. Yedla A., Kakhki F.D. and Jannesari A., Predictive modelling for occupational safety outcomes and days away from work analysis in mining operations, *International Journal of Environmental Research and Public Health*, **17(19)**, 7054 **(2020)**.

(Received 12th January 2025, accepted 14th March 2025)