International Journal of Occupational Safety and Health

ISSN: 2091-0878 (Online) ISSN: 2738-9707 (Print)

Original Article

Application of machine learning modeling for the upstream oil and gas industry injury rate prediction

Desalegn Y[1], Daniel K[1], Mesfin B[2]

[1] School of Mechanical and Industrial Engineering, Addis Ababa Institute of Technology, Addis Ababa University, Addis Ababa, Ethiopia.

[2] University of Stavanger, Department of Energy and Petroleum Engineering, Stavanger, Norway.

Corresponding authors:

Desalegn Yeshitila, &
Professor Daniel Kitaw (PhD)
School of Mechanical and Industrial Engineering,
Addis Ababa Institute of Technology, Addis Ababa University, Addis Ababa, Ethiopia
Tel.: +251932768383
E-mail: desuselam@yahoo.co.uk
E-mail: danielkitaw@yahoo.com
ORCID ID: https://orcid.org/0000-0002-2696-3548

Date of submission: 21.02.2023
Date of acceptance: 11.09.2023
Date of publication: 01.04.2024

Conflicts of interest: None
Supporting agencies: None
DOI:https://doi.org/10.3126/ijosh.v14i2.52668

Copyright: This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

ABSTRACT

Introduction: Yearly, the International Labor Organization report indicates many workplace accident occurrences. The degree of the happenings depends on the workplace environment setting and the incident regulatory measures implemented. By the nature of its work environment, the oil and gas upstream sector is susceptible to high incident rates. In the current fierce business competition and practices, improving productivity, quality, and other processes, such as Safety, is vital. Implementing well-designed safety procedures is the key to managing and reducing the risk level of workplace incidents.

Methods: Recently, the application of Machine learning (ML) modeling for accident/injury prediction has been reported in the construction, mining, transport, and health sectors. Likewise, the objective of this paper was to implement three machine-learning-based models to predict injury rates in a drilling operation. The Petroleum Safety Authority of Norway provided the datasets. First, the dataset was pre-processed, and then the selected features and target dataset were used for the modeling. Finally, the model prediction and performance accuracy analysis were performed.

Results: Results showed that multivariable regression (MVR), Random Forest (RF), and Artificial Neural Network (ANN) machine learning algorithms-based models predict the test data with R^2 values of 0.9576, 0.793, and 0.97036, respectively.

Conclusion: As the common saying goes, 'prevention is better than cure.' For this, implementing methods such as improved work processes and Health, Safety, and Environment (HSE) mitigation procedures, workplace injuries, and accidents allow for reducing the risk level of workplace injuries. The application of integrated machine learning tools, along with carefully built-in workplace accident database implementation, will provide early detection and possible remedial precautions that can be taken to prevent workplace injuries/accidents/fatalities. However, extensive research and development are required to deploy the method in real life. Combining Machine Learning modeling and carefully designed safety measures is vital for successful and robust predictive tools.

Keywords: ANN, HSE, Multivariate Regression, Occupational injury, Random Forest, Safety Management

Introduction

Occupational accidents and occupational injuries can happen anytime and in any field. Occupational injury includes personal injury or fatality from work accidents. The consequences can affect the employees' performance and personal life outside of work. A report from the International Labor Organization estimated that every year, a considerable number of work-related accidents and diseases cause death, fatal accidents, and fatal work-related diseases.[1] Hämäläinen et al. stated that the degree of injuries and their causes are associated with the working environments and the safety procedure implemented.[2] Safety management at workplaces is becoming the paramount consideration, and it is being implemented in several industries as a routine management activity.[3] This endeavor's main objective is to reduce workplace accident risk levels. The upstream oil and gas sector is susceptible to high incident rates in the petroleum industry due to its work environment, remote location, and confined spaces. For instance, Morken et al. analyzed the 12 years (1992–2003) of offshore work-related incidents, which is 6725 cases obtained from the Petroleum Safety Authority of Norway.[4] The dataset includes information such as worker's diagnosis, age, occupational category and occupation, and types of exposure. Four main categories (Maintenance, Catering, Drilling, and Administration) were considered. The analysis showed that the dominant occupational categories were maintenance work (40%) and catering (21%). The authors have also indicated the higher occupational incident rate associated with maintenance work and catering in Denmark and the U.K. offshore sector. It is also noted that the rate of injuries varies from region to region. The workplace injuries problem can be avoided, by implementing properly designed HSE guidelines, and the risk level can be minimized.[5] Dyreborg et al. presented case studies of the impact of safety interventions on injury prevention.[6] The result, among others, has shown strong evidence that the safety intervention approach has shown more effectiveness in preventing injuries. In line with this, it is of great importance to develop an accident-identifying tool that allows us to take appropriate preventive measures to minimize or avoid the risk of incidents.

Recently the application of machine learning (ML) for modeling and predicting future event HSE-associated risks has been tested in engineering, management, healthcare, and medicine. Examples of research papers that used machine learning-based modeling for workplace injury analysis, among others, are in the construction industry, the shipping industry, transportation, high-risk flight environments, the tourism sector, the public sector, and the petroleum industry. The following highlights some of the reviewed papers.

Ciarapica et al. assessed the risk of occupational injury, considering the probability and consequences of injuries. The authors have used five years (2002–2006) of occupational injury data in an Italian region. They developed ML modeling and reported that Neuro-fuzzy networks are found to be a powerful tool for prediction.[3]

In industrial, mining, construction, and services sectors, Matías et al. have presented ML methods for analyzing workplace accidents; specifically, floor-level falls. In this paper, they implemented different ML algorithms, such as Bayesian networks, classification trees, and support vector machines.[7] The dataset (2003 -2006) was based on accidents recorded in the industrial, mining, construction, and services sectors in Vigo, Spain. The analysis results show the prediction of the Bayesian network is relatively better and allows for the provision of recommendations for an accident prevention policy. Sánchez et al. have successfully applied the ML model, that is, the Support vector machines (SVMs) learning method, to forecast occupational accidents. They performed a nationwide Survey on work conditions to access HSE and risk prevention, among others. The Authors have performed SVM modeling based on the interview result dataset as input. According to the authors, the results indicated that the SVM performance was good in terms of prediction and the possible overfitting of the data.[8]

In the shipbuilding industry, Fragiadakis, et al. presented the machine learning predictive and occupational risk assessment model developed with an adaptive neuro-fuzzy inference system.[9] Tsoukalas & Fragiadakis also presented multivariable linear regression and genetic algorithm analysis. Results comparing the predicted values with the recorded data, the work has shown that the proposed model indicates the risk of occupational injury.[10]

In the mining industry, van den Hon et al. have used Artificial Neural Networks (ANN) to model, validate and predict the continuous risk of accidents. Moreover, the authors identified patterns between the input attributes. Results based on the case study data showed that ANN produced a correlation between the predicted continuous risk and actual accidents.[11]

In the transportation sector, Mahdi et al. have presented Machine Learning (Artificial Neural Network (ANN)) and Adaptive-Neuro Fuzzy Inference System (ANFIS) modeling to classify the severity of road accidents.[12] They investigated the application of the combined clustering classification system for categorizing severity in road accidents. Bedard et al. have also used several transportation-related accident datasets that affect the fatality risk of drivers in crashes. They used multivariate logistic regression techniques that reveal the fatal injury associated with factors such as age, sex, and speed. They also indicated that the risk of fatal injuries is associated with not using a seatbelt and practicing over-speed.[13]

In the petroleum industry, Zaranezhad et al. have applied ML algorithms to the workplace accidents dataset related to repair and maintenance at oil refineries. The ML models used are artificial neural networks, fuzzy systems, genetic algorithm (G.A.), and ant-colony optimization algorithm. Based on the considered features, results showed that the perceptron neural network was found to have the highest prediction accuracy of 90.9%.[14] They also evaluated the prediction of hybrid models, and the neural-GA network obtained the highest prediction accuracy of 95.9%. The authors proposed the neural-GA hybrid model to predict early accident predictions caused by repair and maintenance.

In the public sector, Sukumar et al. have implemented different ML algorithms such as random forest, k-nearest neighbor, and decision trees for predicting workplace injury/ workplace incidents. They used the dataset (2015 to 2017) obtained from the Occupational Safety and Health Administration (OSHA) database, which comprises about 61% fatal and 39% non-fatal injuries. The target feature of the research was to predict the nature of the injury, i.e., fatal or non-fatal. The authors' results showed that the statistical performance of the decision tree model was higher than the other two algorithms employed in the case.[15] However, referring to the work of Wang et al, Sukumar et al. recommended a Random forest algorithm for high dimensional data.[15,16] The recommendation is in line with the work of Capitaine et al.[17]

In the construction industry, Tetik et al. have modeled occupational injuries and fatalities. They used datasets (2010 and 2012) that are the main factors for the occurrence of construction accidents. In this study, they employed a decision tree algorithm for the modeling. The results show the relationship between the injury status of workers and the attributes, and the accuracy rate of the model was 70.26%. They also proposed applying the model to the prevention and mitigation strategies for construction accidents.[18] Zhu et al. used machine learning techniques to predict the consequences of construction accidents based on 16 incident factors. The authors have implemented eight algorithms: Logistic regression, Decision tree, Support vector machine, Naive Bayes, K-nearest neighbor, Random forest, Multi-Layer Perceptron, and AutoML. According to the authors, results show that Naive Bayes and Logistics regression achieve the best F1-Score of 78.3 % on a raw data set. They also reported that the "Type of accident" and "Accident reporting and handling" are the most critical factors, and "Emergency management" and "Safety training" are critical subsystems that have a significant impact on the severity of the accident.[19]

In the flight sector, Maynard et al. have used neural networks and Machine learning modeling to predict high-risk flight environments from accident and incident data. Results indicated the potential application of the ANN model to identify the most significant flight risks.[20]

In the tourism sector, Chadyiwa et al. have investigated the application of Machine Learning Applications in the Prediction of Occupational Injuries in South African National Park. The authors compare the performance of the SVM, k-nearest neighbors (KNN), X.G. boost classifier, and deep learning neural networks (DNN) machine learning models concerning the prediction of occupational injuries. Based on the considered datasets, the author's results show that the SVMs had the best performance in prediction.[21]

In a recent public sector study, Khairuddin et al. presented a Machine Learning Approach and Feature Optimization for Smart Workplace Surveillance. For the analysis, the authors have used 66,405 data from the public occupational injury records from OSHA. The idea is to develop a possible occupational Injury Risk Mitigation method. They employed five machine learning algorithms: Support Vector Machine, K-Nearest Neighbors, Naïve Bayes, Decision Tree, and Random Forest. The comparison result reveals that the Random Forest outperformed other models with higher accuracy and F1-score. The authors have also proposed a feature optimization technique, and from the study, they highlight the promising potential for smart workplace surveillance for future injury corrective and preventive strategies.[22]

In a recent injury analysis in the transportation sector, Augustine et al. applied a machine learning modeling approach to predict a road accident. For this, they compared the accident prediction of machine learning models such as Logistic Regression, Random Forest, Decision Tree, K-Nearest Neighbor, XGBoost, and Support Vector. They used the government record accident datasets in a district in India. The comparison result reveals that the Random Forest algorithm gave the highest accuracy of 80.78%.[23] Pandaa et al. also analyzed the statewide accident dataset in India for the period 2008–2019 with four different machine learning methods, including support vector machine (SVM), random forest (RF), Gradient Boosting Machine (GBM), and extreme gradient boosting (XGB). The authors have considered features such as commercial vehicles, excess speed, national highways, and pedestrian faults which are the factors for accidental road killings. The authors' findings suggest that the Machine learning model predicts the accident severity. Among the considered ML models, the gradient boosting machine achieved the best test accuracy.[24]

Research motivation and objectives: From the reviewed research works, we can observe that the application of ML for predicting workplace injuries/accidents in various sectors has shown promising results. However, up to the authors' knowledge, the application of ML for occupational injuries in the petroleum industry is limited. Therefore, this paper aims to present the application of machine learning modeling to predict the possible occupational injury rate based on the available relevant dataset obtained from drilling activities.

Methods

A total of three ML algorithms were used to compare which one best suits and ensure the data used can be used for modeling and predicting workplace injury rates. Figure 1 shows the methodology implemented in this paperwork, which comprises three main parts. These are data pre-processing, machine learning modeling, and model performance accuracy analysis. The data pre-processing was performed to evaluate the data correlation among the features as well as with the target injury rate.

Once the features were identified, the second phase was splitting the dataset into training and testing to be used for the machine learning modeling and model predictions. In this paper, three learning algorithms were selected, namely, Multivariable regression, Random forest regression, and Artificial Neural network. The

model prediction was analyzed with the test - and training dataset. The statistical model's performance accuracy evaluation methods used in this paper were coefficient of determination (R^2), Mean Square Error, and Root Mean Square Error. The details of how they work are presented in the following sections.

The authors of this paper used multivariable regression for multiple independent variables/features (x_1, x_2, x_3… x_n) to predict the target variable, y [25]

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \varepsilon \quad (1)$$

Where

y = the predicted value, which is the dependent variable or target variable

β_0 = the y-intercept (i.e., the value of y when all other independent variables are set to 0)

β_1 = the regression coefficient of the first independent variable x_1

β_n = the regression coefficient of the last independent variable x_n

ε = model error (how much variation there is in the estimate of y)

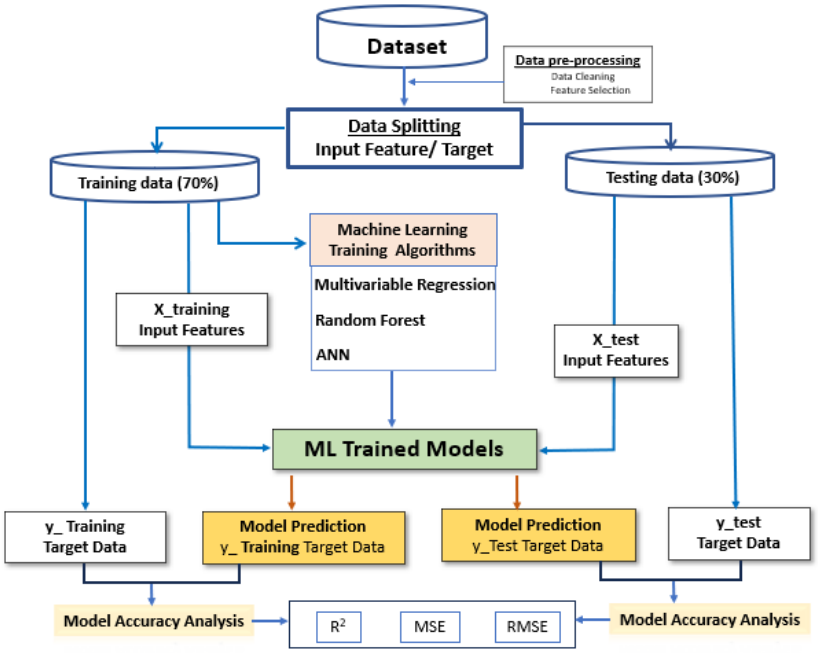

Figure 1: The workflow implemented in this study.

Tsoukalas et al. used a multivariable linear regression method for workplace injury analysis. In this paper, for the multivariable regression modeling, input features used were working hours and injuries within the year range, and the target output is injury rate by splitting the data into 70% for training and 30% for testing.[10]

Random forest is a Supervised Machine Learning Algorithm. It is used in Classification and Regression problems. The concept behind the Random forest algorithms is that they build decision trees based on different inputs and take their majority vote for classifying and averages in case of regression.[26] A random forest algorithm is constructed with a collection of decision trees. The random forest algorithm is a two-step process. First, building n decision trees regressor. Each decision tree regression predicts an output for a given input. The final output is then obtained from the Random forest regression by taking the average of those predictions. Random forest reduces overfitting since it averages over the independent trees.[27] Random forest machine learning algorithm has been employed in several injury studies.[7,15 19,22, 23, 28-30]

In this paper, Random Forest regression is performed, splitting the dataset by 70% for training, and the rest of the dataset (30%) is used for testing the model prediction. The splitting was done to yield a statistically meaningful result.

An artificial neural network (ANN), also known as a neuron network is the mathematical model of

a system that simulates similarly as biological neural networks operate in the human brain capable of learning, prediction, and recognition.[31] Several authors have used ANN learning algorithms for the analysis of workplace injuries.[11,12,14,19,20,32.] ANN uses nodes, similar to neurons building the same sorts of complex interconnections between them (synapses). The neural network comprises three parts, namely the input layer, the hidden layer, and the output layer. The artificial neurons have weighted inputs, transfer functions, and target output. The activation of the neuron uses the weighted sum of the inputs. The single output of the neuron is generated after passing the activation signal through the transfer function. The ANN model is built by using a feed-forward backpropagation network. The training algorithm used in this study was the Levenberg–Marquardt algorithm (TRAINLM). The network training function updates weight and bias values. In addition, the LEARNGDM adaptation learning function is used to calculate the changing weight and update returns the weight change and a new learning state. The network was built with three layers, an input layer, a hidden layer, and an output layer. The input layer consists of three neurons the hidden layers are five neurons and a tangent sigmoid transfer function (TANSIG) transfer function and the output layer has one neuron and the sigmoid TANSIG transfer function. ANN model was developed using inputs and divided into ratios of 70 % for training and 15% for testing, and 15% for validation.

Once the model is built and tested for prediction, the final stage is to evaluate the model's performance accuracy. For this, we used the commonly used statistical parameters such as mean square error (MSE), root mean square error (RMSE), and regression coefficient (R^2), Montgomery.[33]

Mean Square Error (MSE):

MSE provides a measure of how close a regression model is to a measured data point. The closer the MSE value to 0, the more accurate the regression model is.

$$\mathrm{MSE} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\left(\mathrm{y}_{\mathrm{i}}^{\mathrm{predicted}} - \mathrm{y}_{\mathrm{i}}^{\mathrm{Actual}}\right)^2 \qquad (2)$$

Root Mean Square Error (RMSE):

RMSE is also another regression model performance indicator. It is the measure of the mean difference between the actual and the values predicted by a model. It also estimates the accuracy of the model to predict the true, target value.

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\left(\mathrm{y}_{\mathrm{i}}^{\mathrm{predicted}} - \mathrm{y}_{\mathrm{i}}^{\mathrm{Actual}}\right)^2} \qquad (3)$$

Regression Coefficient (R^2):

R-square(R^2) measures the goodness of the best-fit regression line. It defines the degree of variance in the target value that can be explained by the input Features. The R^2 value varies from 0 to 1. A score of 1 is ideal where 100% variation can be explained by the input feature variable.

$$\mathrm{R}^2 = 1 - \frac{\sum_{\mathrm{i}=1}^{\mathrm{N}}\left(\mathrm{y}_{\mathrm{i}}^{\mathrm{predicted}} - \mathrm{y}_{\mathrm{i}}^{\mathrm{Actual}}\right)^2}{\sum_{\mathrm{i}=1}^{\mathrm{N}}\left(\mathrm{y}_{\mathrm{Actual}}^{\mathrm{Mean}} - \mathrm{y}_{\mathrm{i}}^{\mathrm{Actual}}\right)^2} \qquad (4)$$

The secondary workplace injury dataset obtained from Norway's Petroleum Safety Authority[34] was used for ML modeling and analysis. Except for the dataset, the details of the causes, categories, kinds of activities, operators, and other relevant information associated with the injury dataset were not reported in the database. Table 1 shows the dataset, which was recorded from 2009 to 2018, and consists of four variables (Year, Work-hour, Number of injuries, and injury rate). The number of injuries and the associated working hours are not consistently occurring. The injury rate is calculated from the dataset as:

([Number of injuries in the reporting period] x 1,000,000) / (Total hours worked)

The data pre-processing was performed with Pandas/Python library to clean and select features to be used as input for the machine learning algorithms.

Table 1: Workplace injury datasets used in this study.[34]

Year	Work-Hours	Injuries	Injury-rate	Year	Work-Hours	Injuries	Injury-rate
2009	8920468	39	4.4	2014	10084881	25	2.5
2009	6363025	48	7.5	2014	5166295	28	5.4
2009	2221184	28	12.6	2014	2347674	12	5.1
2009	11079666	133	12	2014	15125636	178	11.8
2010	8975538	28	3.1	2015	8869938	26	2.9
2010	5893739	47	8	2015	4856239	32	6.6
2010	2321410	23	9.9	2015	2154055	23	10.7
2010	11834044	122	10.3	2015	10636021	113	10.6
2011	8715265	22	2.5	2016	7744388	18	2.3
2011	5594466	43	7.7	2016	4499170	29	6.4
2011	2402714	24	10	2016	2090811	15	7.2
2011	14951055	154	10.3	2016	9779982	82	8.4
2012	8997539	40	4.4	2017	8329241	33	4
2012	5149376	40	7.8	2017	4503183	27	6
2012	2466948	14	5.7	2017	1988017	19	9.6
2012	15408376	157	10.2	2017	9309383	92	9.9
2013	9386604	38	4	2018	10699902	17	1.6
2013	5553985	41	7.4	2018	4598378	21	4.6
2013	2426849	26	10.7	2018	2101929	9	4.3
2013	15721547	137	8.7	2018	10661638	103	9.7

Results

This section presents the training and test results obtained from the three Machine Learning modeling methods. Here, the training and testing datasets were compared with the respective model predictions. Moreover, the degree of the model accuracy evaluations will be discussed.

Using the whole dataset, the multivariable regression model for the injury rate is obtained as:

$$\text{Injury rate} = \beta_0 + \beta_1 * \text{Year} + \beta_3 * \text{Workhours} + \beta_3 * \text{Injury} \quad (5)$$

Where the coefficients are:

$\beta_o = 304.119737$, $\beta_1 = -0.14682652$, $\beta_2 = -9.6909E\text{-}07$, and $\beta_3 = 0.10734475$

The multivariable regression model was built using the scikit-learn/Python library. Figures 2-3 show the comparison between the training and the testing datasets with model-predicted values, respectively. From the model performance accuracy analysis results presented in Table 2, it is shown that the training dataset and the test datasets correlated with the model predictions with R^2 values of 92.6% and 95.7%, respectively.

Table 2: Multivariable model performance accuracy analysis summary.

Performance	RMSE	MSE	R^2
Training	1.19875	1.4370	0.9264
Testing	0.94129	0.8860	0.9576

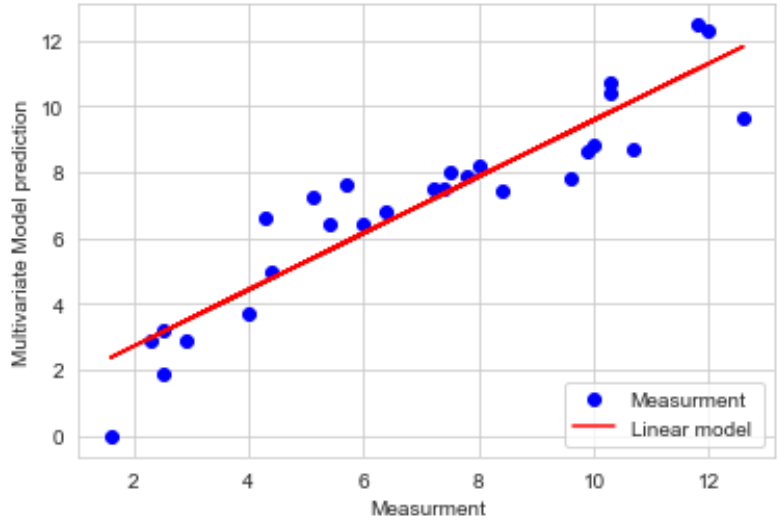

Figure 2: Scatter plot of 70% training injury rate data vs. multivariable regression model prediction.

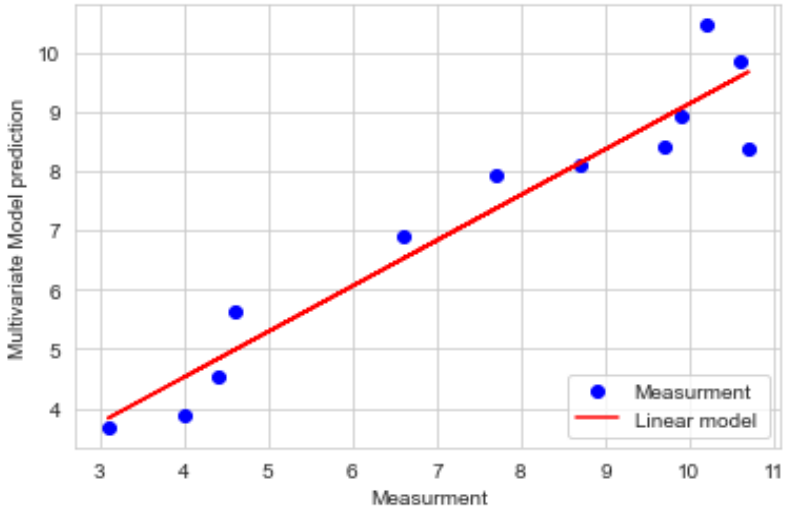

Figure 3: Scatter plot of 30% testing injury rate data vs. multivariable regression model prediction.

The random forest regression was implemented in Python. The basic concept of random forest regression modeling is presented in the methods section, above. Figure 4 displays the comparison of the random forest model prediction, which is based on the 70% dataset and the true training dataset. As provided in Table 3, the model performance accuracy analysis result shows that the random forest model predicts the training dataset with an R^2 value of 0.9875. Further, to evaluate the model prediction performance, 30% of the test datasets were used. Figure 5 shows that the random forest model predicts the test dataset with an R^2 of 79.3%.

Table 3: Random Forest model performance accuracy analysis summary.

Performance	RMSE	MSE	R^2
Training	0.56487	0.3190	0.9875
Testing	1.6946	2.8719	0.793

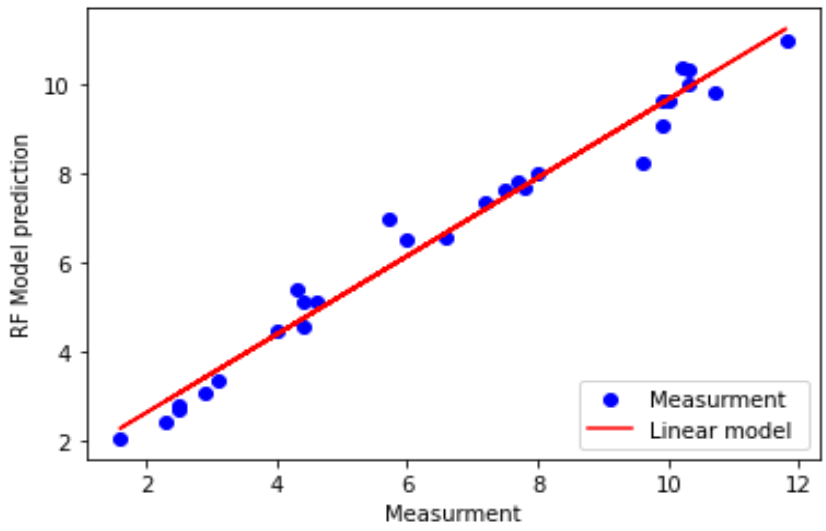

Figure 4: Scatter plot of 70% training injury rate data vs. Random forest regression model prediction.

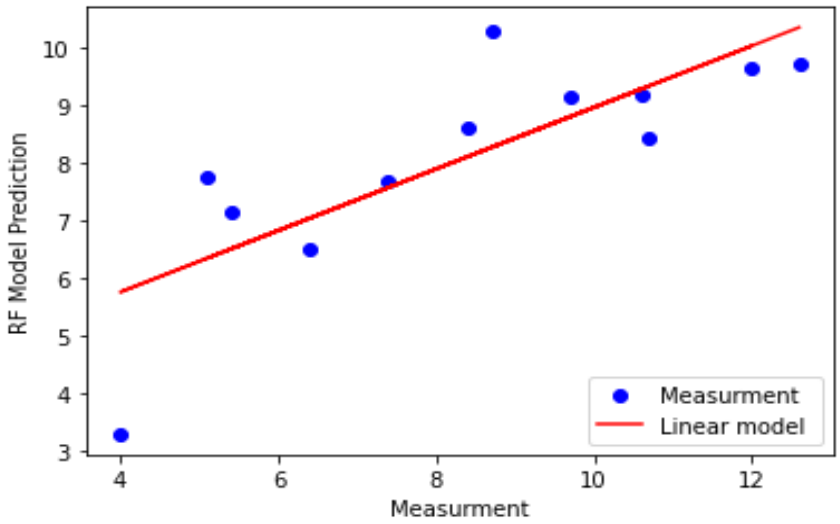

Figure 5: Scatter plot of 30% testing injury rate data vs. Random forest regression model prediction.

A three-layer ANN was also built. To avoid the possible overfitting issue, there are different rule-of-thumb methods for the selection of the appropriate number of neurons for the hidden layers.

The number of hidden layers should be:

- Between the number of the input- and the output layer.
- 2/3 of the number of the input layer plus the size of the output layer.
- Less than twice the size of the input layer.

To satisfy these three conditions, we selected the number of the hidden layer to be five. ANN model is built- in MATLAB/nftool library.[35] Figure 6 displays the results obtained from the measurement (dataset) and ANN model prediction. Table 4 shows the summary of the model performance accuracy analysis obtained from the ANN training, validation, and testing datasets. As provided in the table, the ANN-based model's R^2 values of the training dataset, validation, testing, and all datasets showed a strong correlation with 0.99975, 0.9979, 0.97036, and 0.99283, respectively.

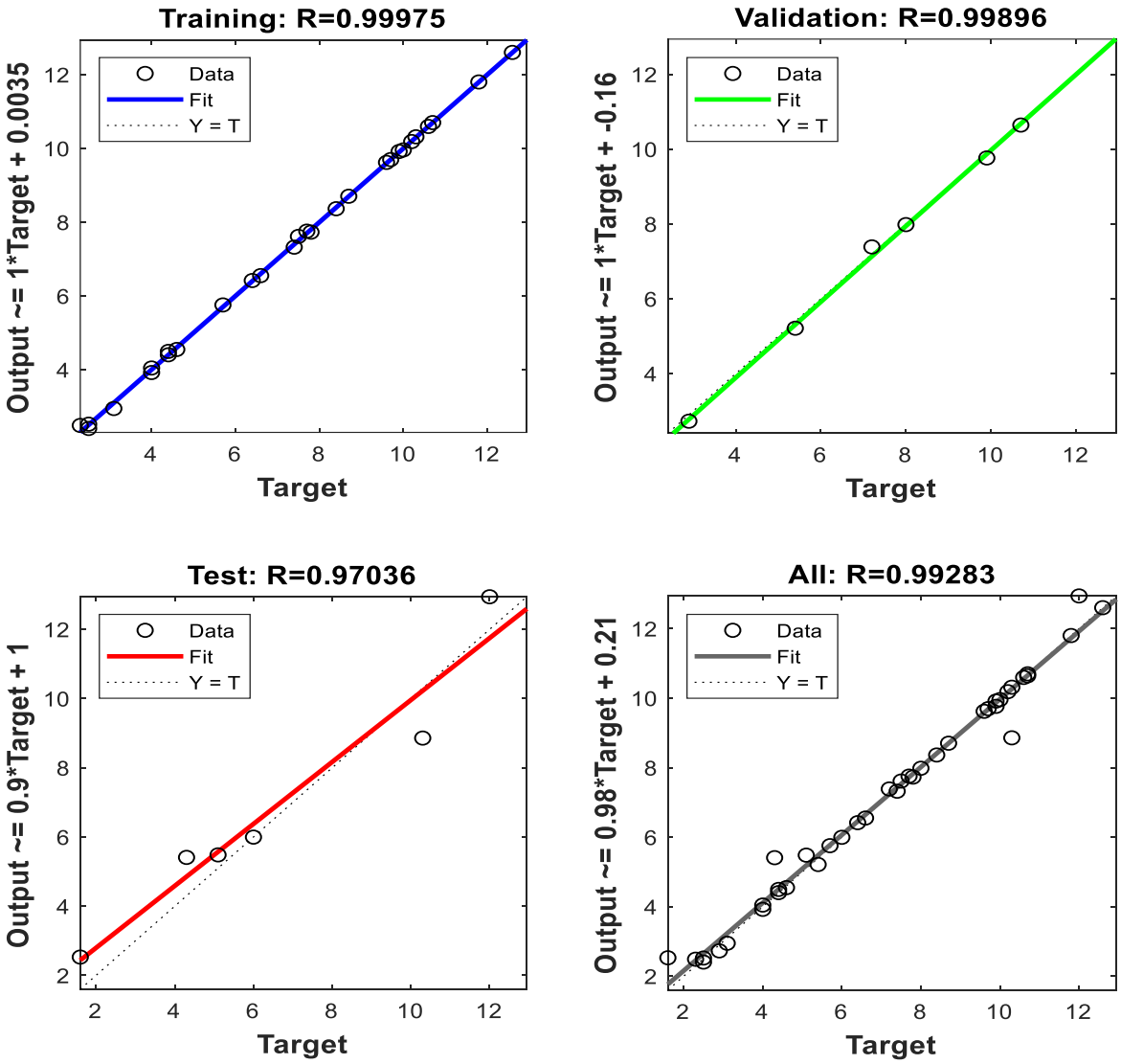

Figure 6: Comparison of ANN model prediction and target dataset.

Table 4: ANN model performance accuracy analysis summary

Performance	Samples	MSE	RMSE	R^2
Training	28	0.00447	0.06686	0.99995
Validation	6	0.02012	0.14184	0.99896
Testing	6	0.86798	0.93165	0.97036

Discussion

Workplace injuries are ordinary happenings in every work environment setting. The severity and the number of occurrences of the injuries may vary due to several factors and sometimes could be deadly. By the nature of its work environment, remote location, confined spaces, and long working hours the oil and gas upstream sector is susceptible to high incident rates. This is in line with the findings of Palathoti S. et al.[36] In the current business competition and practices, it is customary to make every effort to continuously improve; among others, productivity, quality, and Safety are no different. The current trend in safety practices is Safety is left alone for the safety officer and safety department. However, Safety should also be every employee's responsibility.

Despite technological development, improved work processes, and Health, Safety, and Environment (HSE) mitigation procedures,

workplace injuries, and accidents are continuously reported worldwide.[1] The degree of workplace injuries and accidents varies depending on the environmental conditions and safety measures.

The use of contemporary tools plays a vital role in analyzing and accident rate forecasting. As reviewed in the introduction, many studies have applied Artificial intelligence techniques (both regression and classification models) in predicting injury outcomes in various fields, including medical, mining, tourism, transport, and construction sectors. Model analysis results showed that ML is a promising tool for forecasting and injury analysis. The application of the models for injury prediction allows for learning from previous injuries and positive developments related to risk controls and mitigation measures. In addition to modeling, continuous improvement approaches are vital in updating safety measures and precautions to minimize the risk and improve the workplace with the involvement of every employee.

ML application in Safety would help top management in making knowledgeable decision making. It can help them in making general rules from substantial amounts of cases belonging to highly dimensional spaces and is, therefore, a way to ground safety-related decisions under uncertainty on empirical knowledge. ML application could lead to improved decision-making and reduce the accident rate.

The application of artificial intelligence for injury assessment study is limited in the petroleum up - and downstream. Since the ML model performance analysis has shown promising results in various sectors, this paper also aimed to seek the potential application of Machine Learning modeling and prediction of injuries based on the considered dataset in the petroleum industry. For the evaluation, multivariable, Random forest, and ANN machine-learning regression models were selected. Their modeling and optimization procedures are different.

The multivariable regression model is a linear combination of weighted input features related to the target variable. The best regression model is obtained first by writing the error square function, which is the sum of the square of the difference between the multivariable model and the measured, target variable. Applying partial differentiation on the error function concerning the curve fitting coefficients results in optimized coefficients and hence the best-fit model is obtained.

On the other hand, the ANN modeling applies feed-forward-backward propagation training algorithms to achieve the best weight and bias parameters resulting in minimized error.

Unlike the Multivariable and ANN modeling, the Random forest regression is based on building n decision trees regressor (estimator). Then, the final output is then obtained from the average of those predictions. Random forest reduces overfitting since it averages over the independent trees.

Implementing the above three ML training algorithms on available injury data obtained from the North Sea offshore drilling sector, the model's assessment results have shown that the model's predictions are pretty good. However, since the results obtained were from limited datasets, it is difficult to make conclusions for model deployment unless more research is conducted. Regardless of the predictions, the work presented in this paper was to demonstrate the application of machine learning models to predict injuries in the petroleum industry, as also shown in the review of several other public and industrial sectors.

In addition to the selected ML regression models, in the future,

- develop a classification-based model that could predict accident or injury occurrences.
- implement the regression algorithms that were not used in this paper
- include more features that affect workplace accidents or injuries.

Morken et al. presented injuries during drilling operations under five categories, each having

different sub-factors. The database to be used would include Administration, Catering, Drilling and operation, Construction, maintenance, and other injury-related operations. Moreover, the details of the accident, fatal, injury on which part of the body, work department, gender, age, etc. The detailed information on the input features allows for accurate and reliable prediction.[4]

Conclusion

In recent years, the application of ML modeling for injury and safety studies has been increasingly used. Several types of ML algorithms are utilized to model accident/injury data obtained in different sectors, such as mining, transportation, and construction. The performance of the model predictions is also reported to be a potential tool for injury detection and forecasting, which would be necessary for safety practitioners and policymakers.

Due to the limited ML-based injury studies in the oil and gas industry, this paper presents a preliminary ML modeling on drilling-related injury datasets. Model performance accuracy analysis results obtained from the three ML models show that:

- The model accuracy of Multivariable regression, Random Forest, and Artificial Network machine learning algorithms-based models predict the test data with R^2 values of 0.9576, 0.793, and 0.97036, respectively.
- The ANN model has shown a better R^2 value, which is due to the backpropagation-feed forward iteration computations allowing for reducing the error. However, it is important to note that all ML modeling algorithms have pros and cons.

To sum up, the authors believe that implementing integrated machine learning tools along with a carefully built-in workplace accident database will provide early detection and possible remedial precautions that can be taken to prevent workplace injuries /accidents /fatalities. However, extensive research and development are required for deploying the machine learning methods to be utilized in real life. Combining Machine Learning with carefully designed safety measures is the key to successful and robust predictive tools. Moreover, along with new and improved technologies, the application of artificial intelligence on big data could contribute to innovating Safety management systems.

Acknowledgments

The authors acknowledge the Addis Ababa School of Engineering's administrative and technical support.

References

1. International Labour Organization (ILO). Promoting safe and healthy jobs: The ILO Global Programme on Safety, Health and the Environment (Safework): [Internet] 2006. Available from: https://www.ilo.org/global/publications/world-of-work-magazine/articles/WCMS_099050/lang--en/index.htm
2. Hämäläinen P, Saarela KL, Takala J. Global trend according to estimated number of occupational accidents and fatal work-related diseases at region and country level. Journal of Safety Research 2009;40(2):125-39 Available from: https://doi.org/10.1016/j.jsr.2008.12.010
3. Ciarapica FE, Giacchetta G. Classification and prediction of occupational injury risk using soft computing techniques: An Italian study. Safety Science. 2009 Jan; 47(1):36-49. Available from: https://doi.org/10.1016/j.ssci.2008.01.006
4. Morken T, Mehlum IS, and Moen BE. Work-related musculoskeletal disorders in Norway's offshore petroleum industry. Occupational Medicine. 2007; 57(2):112–7 Available from: https://doi.org/10.2118/111622-MS
5. Haugan T, Størseth F, Lootz E, Weggeberg H, Zachariassen S. Work-Related Disorders and Personal Injuries in the Norwegian Petroleum Industry: Achieving a Broader Picture by Combining Data Sources. SPE 111622. Health, Safety, and Environment in Oil and Gas Exploration and Production held in Nice, France, 15–17 April 2008. Available from https://doi.org/10.2118/111622-MS
6. Dyreborg J, Lipscomb HJ, Nielsen K, Törner M, Rasmussen K, Frydendall KB, et. Al.. Safety interventions for the prevention of accidents at work: A systematic review. Campbell Systematic

Reviews: 2022 Jun;18(2). Available from: https://doi.org/10.1002/cl2.1234

7. Matías JM, Rivas T, Martín JE, Taboada J. A machine learning methodology for the analysis of workplace accidents. International Journal of Computer Mathematics. 2008; 85(3):559-78. Available from: https://doi.org/10.1080/00207160701297346

8. Sánchez S, Fernández PR, Lasheras FS, de Cos Juez FJ, García Nieto PJ. Prediction of work-related accidents according to working conditions using support vector machines. Applied Mathematics and Computation. 2011 Dec; 218(7):3539-52. Available from: https://doi.org/10.1016/j.amc.2011.08.100

9. Fragiadakis NG, Tsoukalas VD, Papazoglou VJ. An adaptive neuro-fuzzy inference system (anfis) model for assessing occupational risk in the shipbuilding industry. Safety Science. 2014; 63: 226-35. Available from: https://doi.org/10.1016/j.ssci.2013.11.013

10. Tsoukalas VD, Fragiadakis NG. Prediction of occupational risk in the shipbuilding industry using multivariable linear regression and genetic algorithm analysis. Safety Science. 2016 Mar;83:12-22. Available from: https://doi.org/10.1016/j.ssci.2015.11.010

11. Van den Honert AF, Vlok PJ. Estimating The Continuous Risk Of Accidents Occurring In The Mining Industry In South Africa. S. Afr. J. Ind. Eng. 2015; 26(3):71–85. Available from: http://dx.doi.org/10.7166/26-3-1121

12. Alikhania M, Nedaiea A, Ahmadvand A. Presentation of clustering-classification heuristic method for improvement accuracy in classification of severity of road accidents in Iran.Safety Science. 2013 Dec; 60:. 142-50. Available from: https://doi.org/10.1016/j.ssci.2013.06.008

13. Bedard M, Guyatt GH, Stones MJ, Hirdes JP. The independent contribution of driver, crash, and vehicle characteristics to driver fatalities. Accident Analysis and Prevention. 2002 Nov; 34(6): 717–27 Available from: https://doi.org/10.1016/S0001-4575(01)00072-0

14. Zaranezhad A, Mahabadi HA, Dehghani MR. Development of prediction models for repair and maintenance-related accidents at oil refineries using artificial neural network, fuzzy system, genetic algorithm, and ant colony optimization algorithm. Process Safety and Environmental Protection. 2019 Nov;131: 331-48. Available from: https://doi.org/10.1016/j.psep.2019.08.031

15. Sukumar D, Zhang J, Tao X, Wang X, Zhang W. Predicting Workplace Injuries Using Machine Learning Algorithms. In Proceedings of the 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), Sydney, Australia, 2020. 763-64 Available from: http://dx.doi.org/10.1109/DSAA49011.2020.00104

16. Wang Q, Nguyen TT, Huang JZ, Nguyen TT. An efficient random forests algorithm for high dimensional data classification. Advances in Data Analysis and Classification. 2018;12: 953-72. Available from:http://doi.org/10.1007/s11634-018-0318-1

17. Capitaine L, Genuer R, Thiébaut R. Random forests for high-dimensional longitudinal data. Statistical methods in medical research. 2020; 30(1):166-84. Available from: https://doi.org/10.1177/0962280220946080

18. Tetik YO, Kale OA, Bayram I., Baradan S. Applying decision tree algorithm to explore occupational injuries in the Turkish construction industry. Journal of Engg. Research. 2022;10(3B). Available from: https://doi.org/10.36909/jer.12209

19. Zhu R, Hu X, Hou J, Li X. Application of machine learning techniques for predicting the consequences of construction accidents in China. Process Safety and Environment. Protection. 2021; 145: 293–302. Available from:https://doi.org/10.1016/j.psep.2020.08.006

20. Maynard E, Harris D. Using neural networks to predict high-risk flight environments from accident and incident data. International Journal of Occupational Safety and Ergonomics. 2021;28(2):1204-12. Available from: https://doi.org/10.1080/10803548.2021.1877455

21. Chadyiwa M, Kagura J, Stewart A. Investigating Machine Learning Applications in the Prediction of Occupational Injuries in South African National Parks. Mach. Learn. Knowl. Extr. 2022; 4(3): 768–78. Available from: https://doi.org/10.3390/make4030037

22. Khairuddin MZF, Hui PL, Hasikin K, Abd Razak NA, Lai KW, Mohd Saudi AS, et. al. Occupational Injury Risk Mitigation: Machine Learning Approach and Feature Optimization for Smart Workplace Surveillance. Int. J. Environ. Res. Public Health. 2022; 19(21):. Available from: https://doi.org/10.3390/ijerph192113962

23. Augustine T, Shukla S. Road Accident Prediction using Machine Learning Approaches. 2022 2nd

International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE) Greater Noida, India, 2022. 808-11. Available from: https://doi.org/10.1109/ICACITE53722.2022.9823499

24. Pandaa C, Mishrab AK, Dashc AK, Nawabd H. Predicting and explaining severity of road accident using artificial intelligence techniques, SHAP, and feature analysis. International Journal of Crashworthiness. 2022; 28(2):186-201. Available from: https://doi.org/10.1080/13588265.2022.2074643

25. Anderson TW. An Introduction to Multivariate Statistical Analysis. 3rd Edition. Wiley; 2003 752.Available from: https://www.wiley.com/en-us/An+Introduction+to+Multivariate+Statistical+Analysis%2C+3rd+Edition-p-9780471360919

26. Breiman L. Random forests. Machine learning. 2001;45:5-32. Available from: https://link.springer.com/article/10.1023/A:1010933404324

27. Han J, Kamber M, Jian PJ. Data Mining: Concepts and Techniques. 3rd Edition. USA: Elsevier; 2011. 694p. Available from: https://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf

28. Chang L, Wang H. Analysis of traffic injury severity: An application of non-parametric classification tree techniques. Accident Analysis & Prevention. 2006; 38(5):1019–27. Available from: https://doi.org/10.1016/j.aap.2006.04.009

29. Cheng C, Leu S, Cheng Y, Wu T, Lin C. Applying data mining techniques to explore factors contributing to occupational injuries in Taiwan's construction industry. Accident Analysis & Prevention. 2012; 48:214–22. Available from: https://doi.org/10.1016/j.aap.2011.04.014

30. Choi J, Gu B, Chin S, Lee JS. Machine learning predictive model based on national data for fatal accidents of construction workers. Automation in Construction. 2020;110. Available from: https://doi.org/10.1016/j.autcon.2019.102974

31. Agatonovic-Kustrin S, Beresford R. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. Journal of pharmaceutical and biomedical analysis. 2000; 22(5): 717-27. Available from: https://doi.org/10.1016/S0731-7085(99)00272-1

32. Ivaz J, Nikolić RR, Petrović D, Djoković JM, Hadzima B. Prediction Of The Work-Related Injuries Based On Neural Networks. System Safety: Human - Technical Facility - Environment. 2021;3(1):19-37. Available from: https://doi.org/10.2478/czoto-2021-0003

33. Douglas CM. Introduction to Statistical Quality Control. 8th Edition; Wiley; 2019. 768. Available from: https://www.wiley.com/en-us/Introduction+to+Statistical+Quality+Control%2C+8th+Edition-p-9781119399308

34. Petroleum Safety Authority Data. Available from: https://www.ptil.no/contentassets/525fcd91f1f24aa49147857927989c23/eng-faste-innretn-2018.htm

35. Abiodun OI, Jantan A, Omolara AE, Dada KV, Mohamed NA, Arshad H. State-of-the-art in artificial neural network applications: A survey. Heliyon. 2018; 4(11). Available from: https://doi.org/10.1016/j.heliyon.2018.e00938

36. Palathoti S, Al Aghbari AHM, Otitolaiye VO. Effect of Long Extended Working Hours on the Occupational Health and Safety of Oil and Gas Workers in the Sultanate of Oman. International Journal of Occupational Safety and Health. 2023; 13(4): 419–28. Available from: https://doi.org/10.3126/ijosh.v13i4.48968