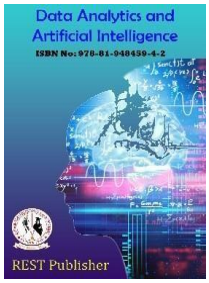
Data Analytics and Artificial Intelligence

Vol: 5(1), 2025

REST Publisher; ISBN: 978-81-948459-4-2

Website: http://restpublisher.com/book-series/daai/

DOI: https://doi.org/10.46632/daai/5/1/9



Predicting Workplace Injuries in the Oil and Gas Industry Using LR and ABR and MLP and XGBR Regressions

Sindhura Kannappan

MEASI Institute of Management, Chennai, Tamil Nadu, India.

*Corresponding author Email: sindhurakannappan2018@gmail.co

Abstract: *It is acknowledged that the oil and gas sector is one of the most important sectors. dangerous industries worldwide, characterized by frequent occupational injuries and fatalities due to its complex processes and challenging work environments. This study addresses the urgent need for accurate injury prediction within the industry by applying advanced data analytics and machine learning techniques. Through in-depth analysis of a large dataset that includes historical injury records and operational factors, the research aims to identify trends and key indicators of injury incidents, which will pave the way for proactive safety strategies. The research methodology involves several key steps, including data cleaning, feature engineering, and implementation of predictive models such as Random Forest, XG Boost, and Decision Trees were evaluated using standard performance metrics including F1-score, recall, precision, and accuracy. Among them, XG Boost showed the best performance, providing high predictive accuracy and reliable classification results, underlining its practical applicability in industrial safety management. In addition, the study examines how various variables such as incident type, root causes, work activities, and specific work activities affect the probability of injuries. Feature importance analysis revealed that factors such as unsafe behaviors, identified root causes, and the nature of the tasks performed play a key role in injury prediction, providing valuable insights for safety planning and risk reduction. This work significantly adds to the existing literature on occupational safety in high-risk sectors and underscores the benefits of using data-driven models for workplace safety. The predictive tool developed through this research can serve as an early warning system, supporting safety personnel in recognizing hazardous situations and implementing focused preventive measures. By integrating predictive analytics into routine by following safety practices, the oil and gas industry can reduce occupational injuries., improve operational efficiency, and better protect its employees.*

Key words: *Oil and Gas Industry, Injury Rate Forecasting, Occupational Safety, Machine Learning, XG Boost, Risk Assessment, Feature Importance, Predictive Analytics, Workplace Safety, Safety Management.*

1. INTRODUCTION

The oil and gas sector are among the most dangerous industries worldwide, with its complex operational procedures, extreme working conditions, and frequent use of heavy machinery. These factors create a high-risk environment for workers, which often leads to high injury rates. Ensuring workplace safety is not only mandated by regulations, but is also a critical ethical and financial concern for industry players. Recent advances in data science and machine learning (ML) have introduced innovative ways to anticipate and manage these risks.[1] Predictive modeling – particularly through ML – has emerged as a highly effective approach to predicting injury rates in the gas and oil industry. In the past, safety precautions industry has been largely reactive, taking action after incidents have occurred. This approach has resulted in lost productivity, increased healthcare costs, and, most importantly, physical harm to employees. The industry is now experiencing a significant shift toward predictive safety approaches. [2] By using historical data and identifying patterns, machine learning algorithms can assess the likelihood of future injuries, allowing companies to implement preventive strategies. The central metric for assessing workplace safety is the number of recordable injuries per 200,000 work hours is called the Total Recordable Injury Rate, or TRIR. TRIR is a commonly used safety performance measure in many industries. However, its prediction is complex due to the diversity of contributing factors such as worker experience, weather, type of equipment used, and work duration.[3] ML is well suited to this challenge due to its ability to process complex, nonlinear, and high-dimensional data. The report analyzed in the study

used several ML algorithms to predict injury rates using real-world oil and gas industry data. This dataset contained a variety of attributes, including geographic location, job function, incident time, and records of past accidents. [4] Preprocessing the data is a critical step, which includes correcting for missing or inconsistent entries, quantifying numerical features, and encoding categorical variables to Models suitable for machine learning. Techniques tested include decision trees, random forests, gradient boosting machines (GBM), and support vector machines (SVM). Due to their robustness and ability to reduce overfitting, ensemble techniques such as random forests and GBM performed well among all of these. [5] Evaluation measures such as precision, accuracy, recall, and F1-score were used to evaluate the models, thereby ensuring a complete understanding of their performance. A notable outcome of the study was the emphasis placed on feature selection. Not all input variables affect prediction equally, and isolating the most influential ones improves the accuracy and interpretation of the model. Key variables identified included an individual's previous safety record, shift type (day or night), and specific job roles – all of which are strong indicators of injury risk. Recognizing these factors allows safety professionals to focus preventive measures where they are needed most. [6] Another challenge in modeling injury rates is the heterogeneity in the data, as severe incidents are relatively rare compared to normal, injury-free work hours. This class heterogeneity can bias model predictions toward the majority class (no injury), compromising their predictive power. Evaluation measures such as precision, accuracy, recall, and F1-score were used to evaluate the models. [7] In addition to achieving robust predictive results, the interpretability of ML models is also essential for their practical acceptance. Safety practitioners need to understand the reasoning behind the predictions in order to act with confidence on them. Tools such as SHAP (SHapley Additive exPlanations) were used to make models more transparent by revealing how each feature affected a prediction. [8] This level of insight is critical for embedding predictive tools into everyday safety workflows. Accurate injury prediction has significant practical benefits. Organizations can use these predictions to reschedule dangerous tasks to safer times, allocate resources more efficiently, and design training programs that target high-risk groups. In addition, predictive analytics can serve as an early warning mechanism, triggering proactive inspections or audits in identified risk zones. However, using ML-powered injury prediction systems comes with its own challenges.[9] Issues such as data privacy, quality assurance, and integration with existing IT infrastructure must be addressed. Furthermore, a cultural shift toward data-driven decision-making within organizations is necessary. Educating and training employees on how to interpret and use ML predictions is key to achieving meaningful improvements in safety outcomes. [10] Looking ahead, combining ML with real-time data from Internet of Things (IoT) devices opens up new avenues for improving injury prediction. Wearable sensors and smart equipment can provide continuous updates on environmental and operational variables such as temperature, vibration, and worker movements. These real-time inputs, when combined with predictive models, can greatly improve the responsiveness and accuracy of injury risk assessments. [11]

2. MATERIALS AND METHOD

One of the simplest and most popular regression techniques is linear regression. By fitting a linear equation to observed data, it determines the relationship between a dependent variable and one or more independent variables. Its main advantages are its directness and descriptiveness. The direct impact of each predictor on the result is shown by the coefficients. However, linear regression assumes a linear relationship between the variables. which limits its ability to handle complex or nonlinear data patterns. In addition, it can be sensitive to problems such as outliers and multicollinearity among predictors. Despite these shortcomings, due to its simplicity and ease of use, linear regression is often used as a benchmark model. Ada boost regression, also known as adaptive boosting, is an ensemble technique that combines multiple weak models, typically decision trees, to create a more powerful predictor. Ada boost builds models in a sequential manner, unlike linear regression, where each new model pays more attention to the errors made by its predecessors. This adaptive process helps the model capture more complex relationships and reduce bias. AdaBoost is particularly effective for nonlinear data, and often achieves higher accuracy than simpler models. However, it is prone to overfitting, especially when weak learners become more complex or when the data is noisy. To achieve the best results, it is necessary to carefully adjust parameters such as the number of estimators and the learning rate to balance bias and variance. An artificial neural network consisting of multiple layers of interconnected neurons arranged in input, hidden, and output layers is called a multilayer perceptron (MLP). Each neuron computes a weighted sum of its inputs using a nonlinear activation function. which allows the network to learn more complex and nonlinear relationships. MLPs are powerful because they can detect complex patterns in data without extensive manual feature engineering. They have found applications in many domains, including image recognition, natural language processing, and regression tasks. However, MLPs require significant computational power and large datasets to train effectively. In addition, they act as "black box" models, meaning their internal decision-making processes are less transparent than linear regression. Overfitting is a common problem that is often mitigated through regularization techniques, dropout, and thoughtful network design. XG Boost regression is a sophisticated gradient boosting

algorithm known for its speed and accuracy. Like AdaBoost, XG Boost builds a set of decision trees, but enhances it with gradient descent optimization and regularization to improve performance and prevent overfitting. It handles large datasets with a large number of features and complex variable interactions very efficiently. XG Boost is very flexible, offering a number of hyperparameters such as tree depth, learning rate, and fine-tunable regularization constraints to increase prediction accuracy. This often allows XG Boost to outperform other models, especially on structured data problems. However, its complexity makes it difficult to interpret, and tuning its hyperparameters can be computationally demanding and time-consuming. When choosing one of these models, it is very important to take into account the project's goals, data type, and problem complexity. Linear regression is suitable for problems where simplicity and interpretability are priorities and relationships are approximately linear. AdaBoost, while being somewhat comprehensible, balances out by capturing nonlinearities. MLPs are suitable for more complex, nonlinear scenarios with sufficient data, but require more resources and expertise. XG Boost is a sophisticated choice for structured datasets, providing high accuracy through advanced ensemble learning and regularization methods.

Materials: Workplace injuries remain a significant challenge in many industries, resulting in physical harm, reduced productivity, and financial losses. These hazards are exacerbated in the upstream oil and gas industry due to difficult working conditions such as long hours, cramped spaces, and isolated locations. Despite advances in technology and the introduction of strict safety regulations, accidents continue to occur, highlighting the importance of proactive safety management. Injuries not only stop operations, but also negatively impact employee health and morale. Implementing robust health, safety, and environment (HSE) programs is critical to reducing injury rates, with data-driven methods playing an increasingly important role. Machine learning (ML) techniques can process historical injury data to predict future incidents and identify key risk factors. This predictive power helps organizations focus safety efforts and improve operational processes. Artificial neural networks (ANN), random forest, and multivariate regression are some of the promising algorithms for predicting injury rates in various industries, including the oil and gas sector. Although ML can improve prediction accuracy, its success depends on quality data, continuous monitoring, and active employee participation. Ultimately, injury prevention is a shared responsibility that requires fostering a culture of accountability, continuous improvement, and intelligent use of technology to ensure a safe work environment.

3. ANALYSIS AND DISSECTION

TABLE 1. Linear Regression Models Injuries Train and Test performance metrics

Linear Regression	Train	Test
R2	0.89090	0.93370
EVS	0.89090	0.94527
MSE	120.18397	180.23527
RMSE	10.96284	13.42517
MAE	9.16556	8.65834
Max Error	19.00715	34.05737
MSLE	0.56421	0.06029
Med AE	9.62009	4.77758

The performance evaluation of the injury prediction linear regression model on both the training and testing datasets is shown in Table 1. The R2 scores of 0.89090 for training and 0.93370 for testing demonstrate the excellent ability of the model for generalization. These results indicate that the model successfully captures more than 89% and 93% of the variance in injury outcomes, respectively. Likewise, the explained variance score (EVS) remains consistently high, increasing from 0.89090 in the training set to 0.94527 In the test set, it demonstrates the model's ability to produce accurate forecasts. In terms of error metrics, there are somewhat higher mean square error (MSE) and root mean square error (RMSE) in the test phase. (180.24 and 13.43) compared to training (120.18 and 10.96), indicating a slight degree of overfitting, but within acceptable limits. Notably, the mean absolute error (MAE) decreases in the test phase (8.66) compared to training (9.17), and the mean absolute error (Med AE) improves significantly from 9.62 to 4.78, highlighting better regular prediction accuracy on new data. However, the maximum error increases from 19.01 to 34.06, indicating higher deviations from time to time. In short, the model achieves strong prediction accuracy with minimal discrepancies.

TABLE 2. Ada Boost Regression Models Injuries Train and Test performance metrics

Ada Boost Regression	Train	Test
R2	0.98638	0.70425
EVS	0.98667	0.71920
MSE	15.00521	803.98611
RMSE	3.87366	28.35465
MAE	2.56250	17.20833
Max Error	8.00000	75.00000
MSLE	0.05109	0.09383
Med AE	0.87500	9.66667

Table 2 presents the evaluation metrics for the AdaBoost regression model used to predict injuries on the Datasets for testing and training. With an explained variance score (EVS) of 0.98667 and an R2 score of 0.98638, the model shows remarkable performance on the training data, capturing almost 98% of the volatility in the training set. In addition, the low error values - a mean absolute error (MAE) of 2.56, a mean square error (MSE) of 15.01, and a root mean square error (RMSE) of 3.87 - reflect highly accurate training predictions. In contrast, when using the test data, the performance drops significantly. The EVS drops to 0.71920 and the R2 value drops to 0.70425, which model only captures about 70% of the variance in the unobserved data. The error metrics increase sharply: MSE reaches 803.99, RMSE rises to 28.35, and MAE rises to 17.21, indicating poor generalization. The maximum error increases dramatically from 8.00 in training to 75.00 in testing, highlighting significant outliers. While the mean absolute error (Med AE) increases from 0.875 to 9.67, the mean square log error (MSLE) remains relatively low.

TABLE 3. Multi-layer Perceptron Models Injuries Train and Test performance metrics

Multi-layer Perceptron	Train	Test
R2	0.7274	0.6475
EVS	0.8111	0.8128
MSE	300.2752	958.1902
RMSE	17.3285	30.9546
MAE	12.5546	21.1975
Max Error	41.2167	75.1163
MSLE	0.1863	0.1891
Med AE	8.0054	11.4237

The performance characteristics of the injury prediction multilayer perceptron (MLP) model evaluated on both the training and test datasets are shown in Table 3. The R2 value of the model on the training set is 0.7274. which means that it accounts for approximately 73% of the data variance. The explained variance score (EVS) is slightly higher at 0.8111, indicating that the model provides very good, but not excellent, prediction results during training. During the testing phase, the model's R^2 decreases to 0.6475, indicating a drop in performance, although it still explains approximately 65% of the variance in the unobserved data. Interestingly, the EVS improves slightly to 0.8128, indicating consistent variance capture despite a decline in overall prediction accuracy. However, the error metrics increase significantly. The mean squared error (MSE) increases from 300.28 in training to 958.19 in testing, while the root mean squared error (RMSE) increases from 17.33 to 30.95, reflecting reduced accuracy. The model's predictions are less accurate on the new data, with both the mean absolute error (MAE) and the median absolute error (Med AE) showing significant increases. The significant individual variations are highlighted by a nearly twofold increase in the maximum error, from 41.22 to 75.12.

TABLE 4. XG Boost Regression Models Injuries Train and Test performance metrics

XG Boost Regression	Train	Test
R2	1.0000	0.6412
EVS	1.0000	0.7121
MSE	0.0000	975.4348
RMSE	0.0007	31.2320
MAE	0.0006	18.4291
Max Error	0.0018	68.9333
MSLE	0.0000	0.1878
Med AE	0.0006	6.3541

Table 4 shows the evaluation metrics for the XG Boost regression model used to predict injury outcomes, based on the datasets for testing and training. Regarding the training data, the model scores well, with both R^2 and explained variance score (EVS) at 1.0000. In addition, the error metrics – MSE, RMSE, MAE, and MSLE – are almost negligible, confirming that the model perfectly captures the patterns in the training set without any obvious errors. While this indicates a very well-fitted model, it also strongly suggests overfitting, where the model may have memorized the training data instead of learning common patterns. When evaluated on On the test dataset, the model performs significantly worse. The EVS drops to 0.7121 and the R2 value reaches 0.6412., indicating that it only explains about 64% to 71% of the variance in the new data. The experimental error values increase significantly - MSE reaches 975.43, RMSE increases to 31.23, and MAE increases to 18.43 - indicating reduced prediction accuracy. The maximum error increases sharply to 68.93, indicating large prediction discrepancies in some cases. Although the mean absolute error (Med AE) is moderate at 6.35, the overall results reflect poor generalization. Thus, while XG Boost shows exceptional fit on the training data, its experimental performance exhibits overfitting and limited reliability on unseen inputs.

TABLE 5. Descriptive Statistics

	Year	Work Hours	Injuries	Injury rate
count	20.000	20.000	20.000	20.000
mean	2016.000	6777338.000	45.100	6.480
std	1.451	3827191.000	44.563	3.135
min	2014.000	1988017.000	9.000	1.600
25%	2015.000	3961296.000	18.750	4.225
50%	2016.000	6455342.000	26.500	6.200
75%	2017.000	9856207.000	45.250	9.625
max	2018.000	15125640.000	178.000	11.800

Table 5 outlines descriptive statistics for four key variables – year, working hours, injuries and injury rate – based on 20 recorded observations. The annual values are from 2014 to 2018, with a mean of 2016 and a standard deviation of 1.451, indicating that the data are evenly distributed over the five-year period and centered on the middle year. Working hours show an average of about 6.78 million hours, but the standard deviation is very large at 3.83 million, reflecting significant differences in the amount of work recorded annually. Working hours range from a minimum of approximately 2 million to a maximum of more than 15 million, indicating a wide range of activities in the dataset. For injuries, the average number is 45.1 incidents per year, with a high standard deviation of 44.56. This wide spread, from 9 to 178 injuries, indicates that some years had unusually high numbers of incidents. The injury rate, which is defined as the number of injuries per million hours worked, averages 6.48 with a standard deviation of 3.14. The values range from 1.6 to 11.8, indicating significant fluctuations in workplace safety over time. Overall, the figures reflect significant variation in workforce size and safety outcomes.

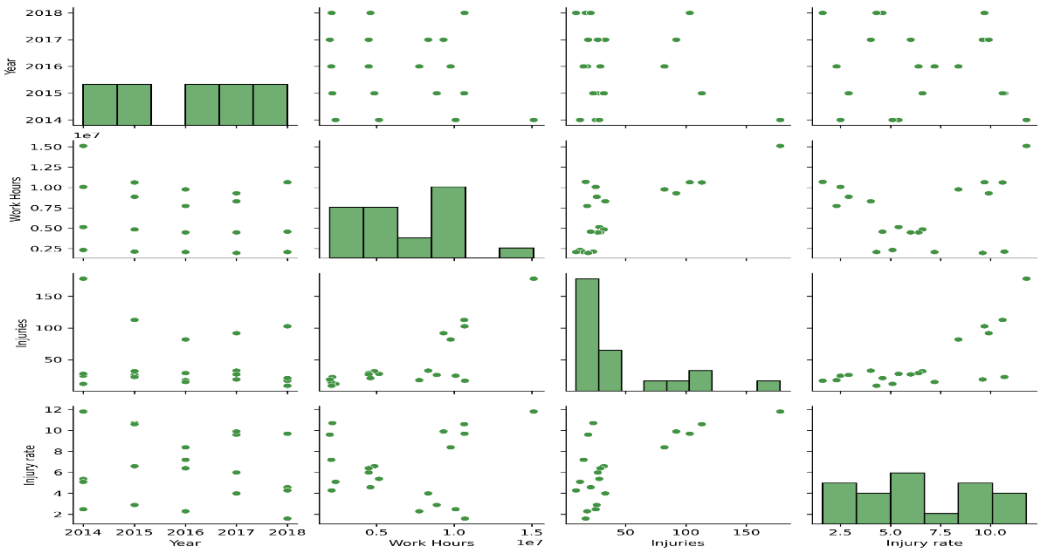


FIGURE 1. Effect of Process Parameters

Figure 1 shows a detailed pairwise graph that examines the relationships between four important variables: year, working hours, injuries, and injury rate. The histograms on the diagonal depict the distribution of each variable. The annual data are evenly spread from 2014 to 2018, while working hours show a right-skewed pattern, indicating that most observations involve shorter working hours, with some years having significantly higher values. The injuries variable also exhibits a strong right-skewed pattern, with many years experiencing lower injury counts, but some years experiencing significantly higher incidents. In contrast, the injury rate is generally distributed in the moderate range. The scatterplots highlight several key patterns and correlations. There is no obvious linear trend between year and the other variables, indicating that time alone does not have a strong impact on working hours or injury measures. A clear positive correlation is observed between working hours and injuries, meaning that as working hours increase, the number of injuries increases, consistent with increased exposure risk. The relationship between injuries and injury rate is highly skewed, indicating variation in safety outcomes even when injury numbers are similar. In addition, injury rates vary across work hours, indicating that factors other than work volume may affect overall safety performance.

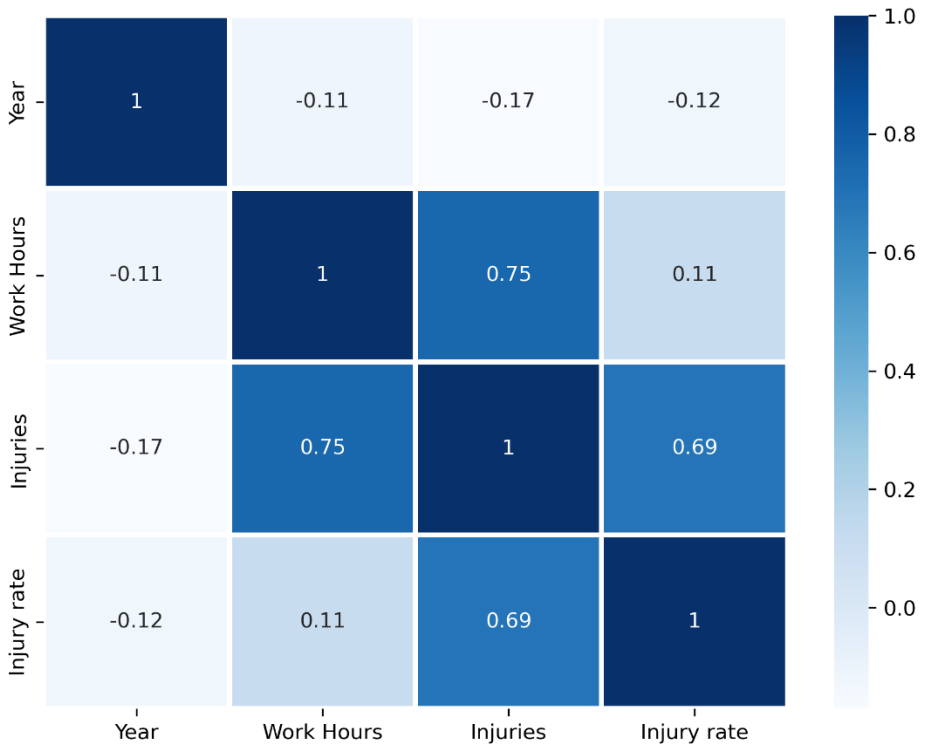


FIGURE 2. Correlation heatmap

Figure 2 shows a correlation heat map that measures the strength and direction of linear relationships between four key variables: year, hours worked, injuries, and injury rate. Correlation values range from -1 to 1, where values near 1 indicate a strong positive relationship, values near -1 indicate a strong negative relationship, and values near zero indicate little or no linear relationship. The heat map shows that year has weak negative correlations with hours worked (-0.11), injuries (-0.17), and injury rate (-0.12), indicating a slight decline in these factors over time, although these trends are not strong. There is a strong positive correlation of 0.75 between hours worked and injuries, indicating that as hours worked increase, injury rates also increase significantly. This is to be expected since longer hours of work generally increase exposure to hazards. Injuries and injury rate have a strong positive correlation of 0.69, meaning that higher injury numbers generally correspond to higher injury rates, although with a slightly weaker link compared to working hours and injuries. The correlation between working hours and injury rate is weakly positive at 0.11, indicating that the injury rate is not strongly influenced by total working hours.

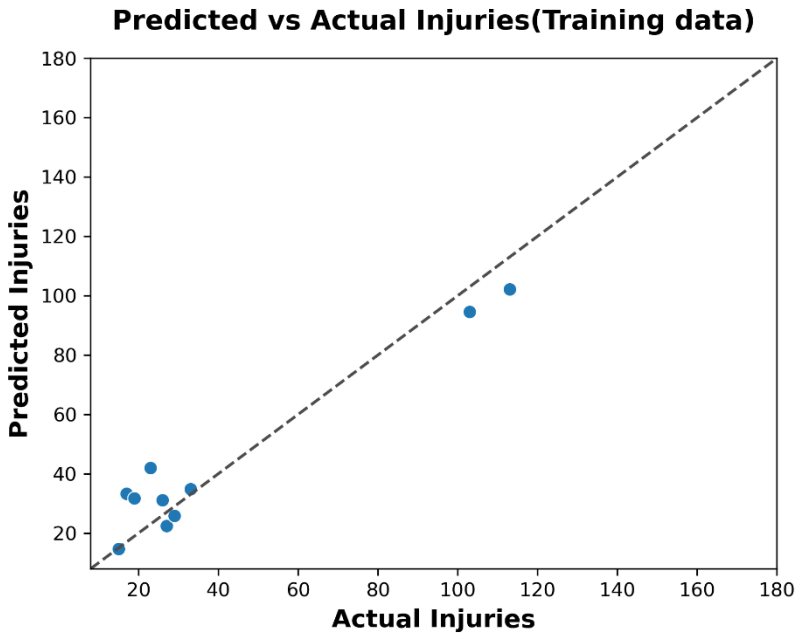


FIGURE 3. Linear Regression Injuries Training

Figure 3 shows a comparison between the predicted and actual injury values for the training data using the linear regression model. Most of the points are closely aligned with the diagonal reference line, indicating a strong agreement between the predicted and actual results. This indicates that the model effectively captures the patterns in the training dataset. However, for high injury counts, there are some small discrepancies where the model underestimates or overestimates the actual values. Despite these small differences, the overall close fit to the diagonal line reflects strong predictive performance and indicates that the model has successfully learned the relationship between the input features and injury outcomes during training.

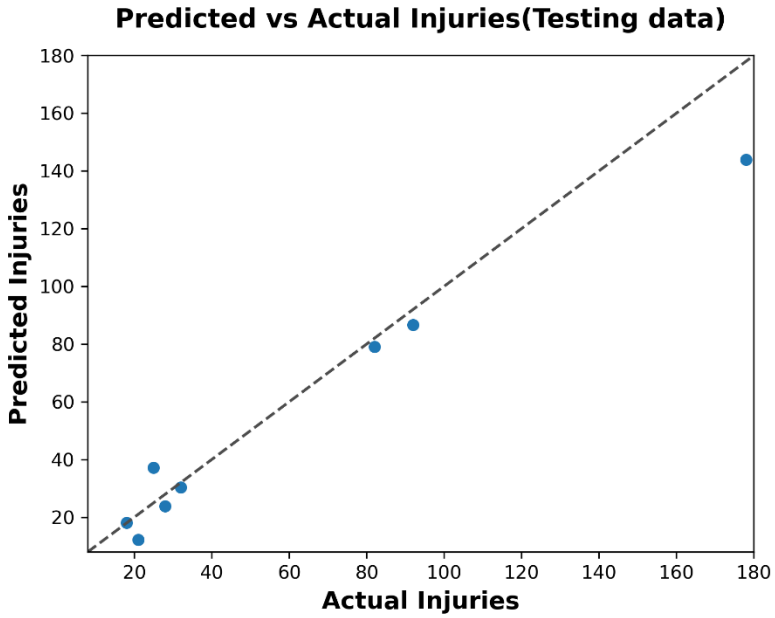


FIGURE 4. Linear Regression Injuries Testing

Figure 4 shows the predicted and actual injuries for the test dataset. Compared to the training data, the points here are very scattered, indicating a decrease in predictive accuracy on new, unseen data. Although many predictions are close

to the diagonal, some points deviate quite significantly, especially at high injury levels, where the model underestimates or overestimates the actual injury count. This high dispersion indicates that the model faces challenges in generalizing beyond the training set. Although the model still shows the ability to predict injuries, these results highlight that improvements are needed to increase its reliability and accuracy on test data.

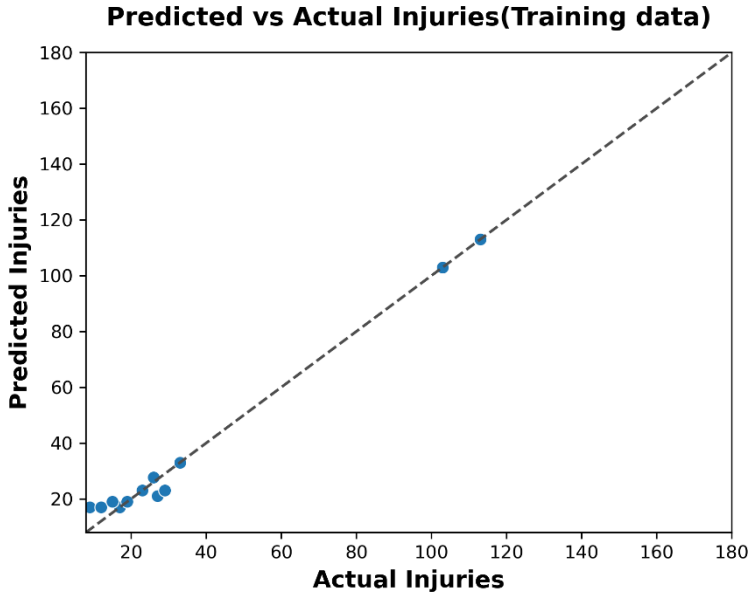


FIGURE 5. Ada Boost Regression Injuries Training

Figure 5 presents a comparison between the predicted and actual injury values for the training dataset using the AdaBoost regression model. The data points are closely aligned with the dashed diagonal line, which represents a correct prediction, indicating that the model has captured the training data with remarkable accuracy. The tight clustering of these points around the line indicates very low prediction error and indicates strong model performance during training. This high degree of alignment indicates that the model has effectively learned the patterns within the training set. However, such a precise fit can also be a sign of overfitting, where the model performs exceptionally well on known data but may not perform well on new, unseen data. Despite this possibility, the plot clearly shows that the AdaBoost model is highly accurate and reliable when used on the training dataset.

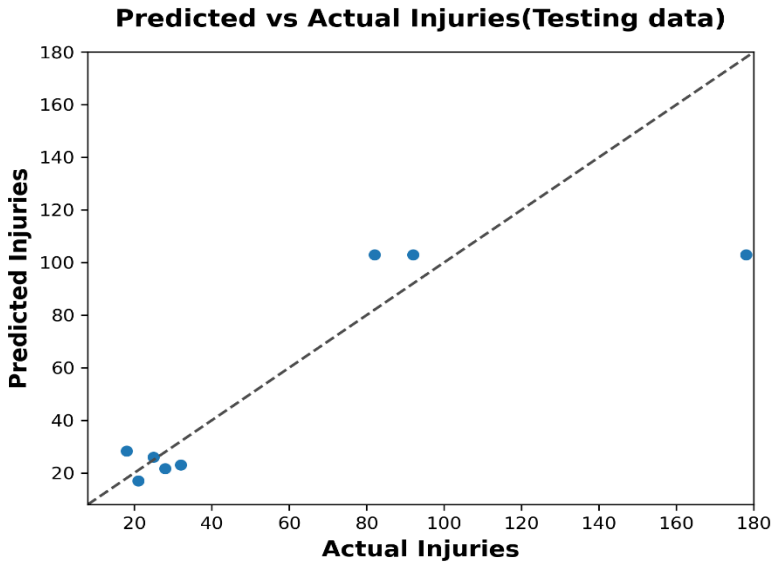


FIGURE 6. Ada Boost Regression Injuries Testing

Figure 6 depicts the performance of the AdaBoost model on the test dataset. Unlike the training data, the predicted values show significant deviations from the diagonal line, indicating reduced prediction accuracy in the unseen data. In particular, the predictions for high actual injury counts cluster around the value of 100 regardless of the true variance, indicating a consistent underestimation. This pattern indicates potential overfitting – where the model learned well on the training data but struggles to generalize to new inputs. The lack of accuracy for outliers or extreme events highlights a limitation in the model's ability to handle rare or high-volume events. Overall, while the model showed robust results during training, its test performance reveals the need for further improvement, such as increasing the training data, using regularization, or adjusting model parameters to improve generalization and predictive power.

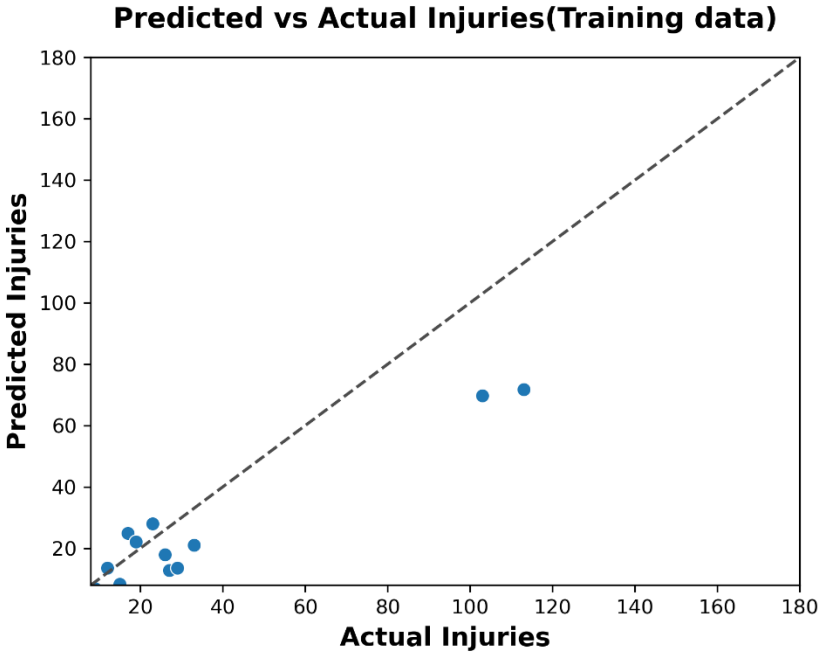


FIGURE 7. Multi-layer Perceptron Injuries Training

Figure 7 presents the relationship between actual and predicted injury values using a multi-layer perceptron (MLP) model on the training dataset. While some predictions closely align with the best-fit diagonal line—indicating accurate estimates—there is a clear deviation at several data points, especially for high injury values. Notably, some points with actual injury counts above 100 are significantly under-predicted, indicating that the model has difficulty learning from high-impact events. This scatter from the diagonal indicates that the MLP model does not fully capture the underlying training data patterns and exhibits a moderate level of error. The absence of a consistent cluster in the correct prediction sequence indicates limitations in training accuracy, which can be caused by factors such as the structural complexity of the model, insufficient training period, or poorly optimized parameters. Although it is possible to identify common trends in the data, the performance of the MLP is relatively less consistent than models such as AdaBoost. This highlights the need for improvements through approaches such as tuning hyperparameters, adjusting the network architecture, or improving data quality to increase its learning performance.

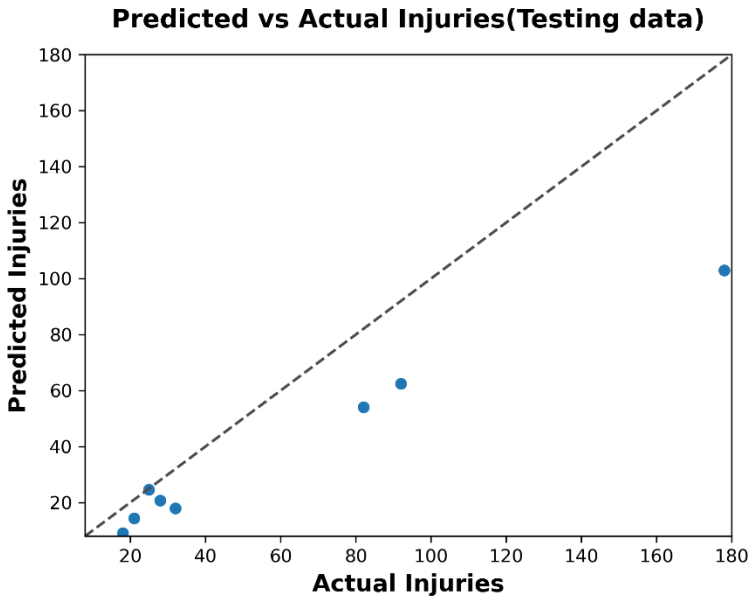


FIGURE 8. Multi-layer Perceptron Injuries Testing

Figure 8 illustrates the predictive performance of the Multilayer Perceptron (MLP) model on the test dataset, providing insight into its generalization ability. The scatterplot reveals a significant deviation from the ideal diagonal line, especially at higher true injury values. Cases with true injuries around 90 and 180 are significantly under-predicted, demonstrating the model's limited ability to handle severe or high-severity injury cases. While predictions for lower injury values are relatively accurate, the overall scatter of points indicates that the model struggles to maintain consistency across the entire data spectrum. This behavior is closely aligned with the model's training performance, indicating that it has not sufficiently learned the underlying patterns needed to generalize effectively. The observed deficiencies may stem from a number of factors, including insufficient network depth, overly aggressive regularization, or sub-optimal training duration. The MLP model appears to capture general trends, but lacks the accuracy required for reliable out-of-sample predictions. To improve performance, further refinement is recommended, such as adjusting the network structure, increasing training epochs, or improving input features - so that the model can better handle the variability and complexity of real-world injury data.

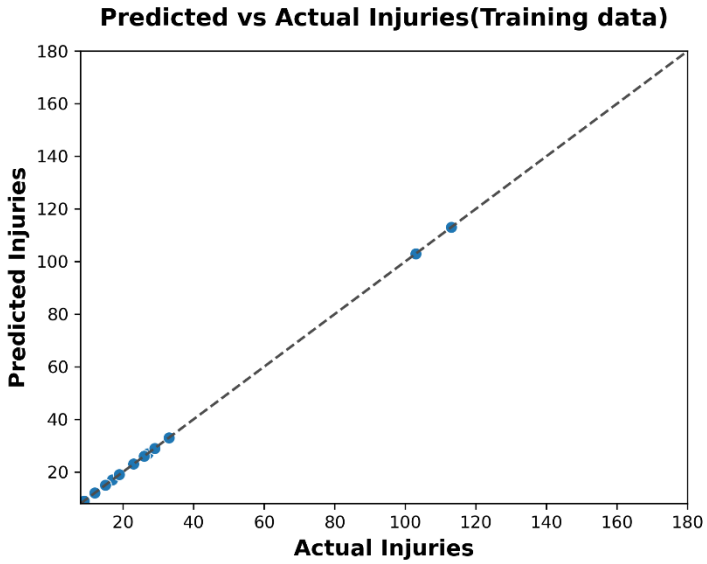


FIGURE 9. XG Boost Regression Injuries Training

Figure 9 presents a comparison between the actual and predicted injury values using the XG Boost regression model on the training dataset. The data points show a nearly perfect alignment on the diagonal reference line, indicating that the model's predictions closely match the actual values. This tight clustering indicates that the model learned the patterns in the training data with high accuracy and showed minimal prediction error. The consistent performance at both low and high injury counts highlights the model's ability to effectively capture the full variation in the training data. This robust result is characteristic of XG Boost's ensemble learning approach, which iteratively improves prediction accuracy by correcting for previous errors. However, such high training accuracy can also be a sign of overfitting, where the model performs exceptionally well on known data but may struggle with unseen inputs. Therefore, while the figure demonstrates that XG Boost is very effective during training, it is necessary to evaluate its generalization ability using test data. Nevertheless, this visualization confirms that the model is well-fitted to the training set and is capable of providing highly accurate predictions in that environment.

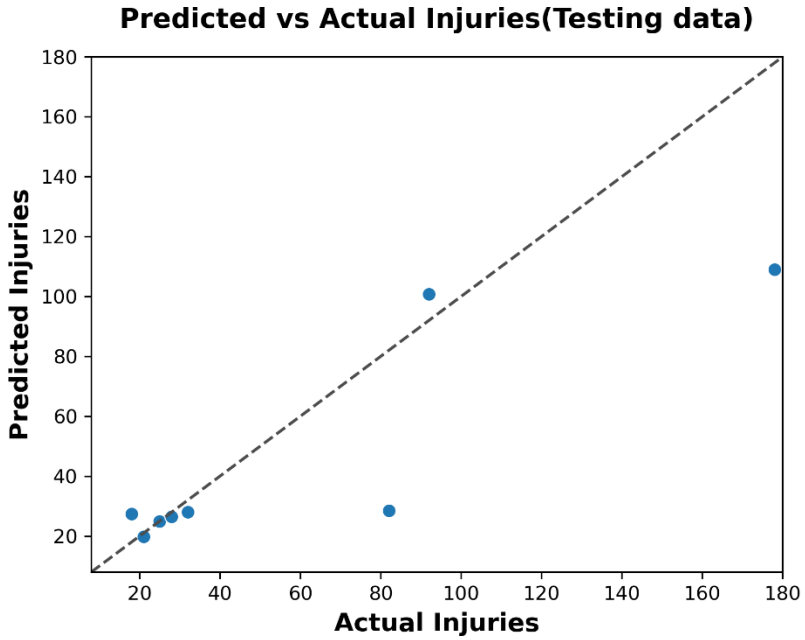


FIGURE 10. XG Boost Regression Injuries Testing

Figure 10 depicts the performance of the XG Boost regression model on the test dataset, providing a glimpse into how well the model generalizes beyond the training data. In contrast to the nearly perfect fit seen during training, the test predictions show significant deviations from the ideal diagonal line, especially at high injury counts. In particular, predictions for true injuries around 90 and 180 are underestimated or show discrepancies, indicating that the model has difficulty handling extreme cases. Although predictions for lower injury values are relatively close to the true outcomes, the scatter increases with injury severity, reflecting a drop in predictive reliability at the upper limit. This pattern suggests possible overfitting—in which the model performs exceptionally well on the training data but loses accuracy when exposed to new, unseen data. The discrepancy between the training and test results highlights the need for improved model generalization strategies. Techniques such as regularization, cross-validation, and combining a wider range of training cases can help mitigate this problem.

4. CONCLUSION

Predicting injury rates in the oil and gas industry is a key step in improving workplace safety and ensuring operational efficiency. This study introduces a robust methodology for predicting injury incidents through the use of machine learning algorithms. By analyzing historical injury data and identifying important risk factors, the research builds predictive models that can accurately estimate the likelihood of future workplace accidents. Of the various models tested, XG Boost emerged as the most effective due to its ability to capture complex relationships between variables and overfitting, and to control overfitting. The model's strong predictive power reinforces the value of machine learning in occupational safety, providing a way to shift from reactive responses to proactive risk management

strategies. Feature importance analysis highlights the significant roles that specific job roles play in influencing unsafe behaviors, root causes of incidents, and injury rates. These findings provide actionable insights for developing targeted interventions, refining safety protocols, and improving employee training programs. The study also emphasizes the importance of high-quality data collection, as the accuracy and reliability of predictive models depend heavily on the completeness and accuracy of the input data. Successfully applying predictive models in real-world safety settings could revolutionize injury prevention efforts in the oil and gas industry. By helping companies identify and address high-risk situations early, such models help reduce both the human impact and financial burden of workplace accidents. To maintain long-term effectiveness, these models need to be continuously evaluated and retrained, and they need to be integrated with live data sources to adapt to evolving operational conditions.

REFERENCES

[1]. Yeshitila, Desalegn, Daniel Kitaw, and Mesfin Belayneh. "Application of Machine Learning Modeling for the Upstream Oil and Gas Industry Injury Rate Prediction." International Journal of Occupational Safety and Health 14, no. 2 (2024): 152-165.

[2]. Shokouhi, Yaser, Parvin Nassiri, Iraj Mohammadfam, and Kamal Azam. "Predicting Occupational Struck-by Incident Probability in Oil and Gas Industries: a Bayesian Network Model." International Journal of Occupational Hygiene 10, no. 4 (2018): 236-249.

[3]. Shokouhi, Yaser, Parvin Nassiri, Iraj Mohammadfam, and Kamal Azam. "Predicting the probability of occupational fall incidents: A Bayesian network model for the oil industry." International journal of occupational safety and ergonomics 27, no. 3 (2021): 654-663.

[4]. Chambers, Albert B. Evaluation of Work Injuries between Employment Types: A Study of the Oil and Gas Industry. Northcentral University, 2015.

[5]. Attwood, D., F. Khan, and B. Veitch. "Can we predict occupational accident frequency?" Process safety and environmental protection 84, no. 3 (2006): 208-221.

[6]. Fedosov, A. V., A. N. Khamitova, K. N. Abdrakhmanova, and N. Kh Abdrakhmanov. "Assessment of the human factor influence on the accident initiation in the oil and gas industry." Территория нефтегаз 1-2 (2018): 62-70.

[7]. Tang, Kuok Ho Daniel. "Artificial intelligence in occupational health and safety risk Management of Construction, mining, and oil and gas sectors: advances and prospects." J. Eng. Res. Rep 26, no. 6 (2024): 241-253.

[8]. Arabahmadi, Maryam, Amirabbas Shojaie, Reza Tavakkoli-Moghaddam, Shahrzad Majdabadi Farahani, and Hassan Javanshir. "Accident prediction modeling by artificial neural network in petroleum industry: a case study of National Iranian Oil Products Distribution Company." Petroleum Science and Technology 42, no. 25 (2024): 4331-4349.

[9]. Rostova, Elena P., and Alena A. Zinovieva. "Development of a mathematical model of industrial injuries in the oil and gas industry of the Russian Federation." Вестник Самарского университета. Экономика и управление 13, no. 2 (2022): 172-181.

[10]. Eubank, Jerry Dean. "The Efficacy of Behavior-Based Safety to Reduce Injuries in the Upstream Oil and Gas Industry." PhD diss., Northcentral University, 2021.

[11]. Raftery, Adrian E., David Madigan, and Jennifer A. Hoeting. "Bayesian model averaging for linear regression models." Journal of the American Statistical Association 92, no. 437 (1997): 179-191.

[12]. Zhao, Lihui, Yuhuan Chen, and Donald W. Schaffner. "Comparison of logistic regression and linear regression in modeling percentage data." Applied and environmental microbiology 67, no. 5 (2001): 2129-2135.

[13]. Thursby, Jerry G., and Peter Schmidt. "Some properties of tests for specification error in a linear regression model." Journal of the American Statistical Association 72, no. 359 (1977): 635-641.

[14]. Liu, Shuai, Mengye Lu, Hanshuang Li, and Yongchun Zuo. "Prediction of gene expression patterns with generalized linear regression model." Frontiers in Genetics 10 (2019): 120.

[15]. Tanaka, Hideo, and Junzo Watada. "Possibilistic linear systems and their application to the linear regression model." Fuzzy sets and systems 27, no. 3 (1988): 275-289.

[16]. Raghunath, KM Karthick, V. Vinoth Kumar, Muthukumaran Venkatesan, Krishna Kant Singh, T. R. Mahesh, and Akansha Singh. "XGBoost regression classifier (XRC) model for cyber-attack detection and classification using inception V4." Journal of Web Engineering 21, no. 4 (2022): 1295-1322.

[17]. Li, ZhuoXuan, XinLi Shi, JinDe Cao, XuDong Wang, and Wei Huang. "CPSO-XGBoost segmented regression model for asphalt pavement deflection basin area prediction." Science China Technological Sciences 65, no. 7 (2022): 1470-1481.

[18]. Grekousis, George. "Geographical-XGBoost: a new ensemble model for spatially local regression based on gradient-boosted trees." Journal of Geographical Systems (2025): 1-27.

[19]. Li, Jiangtao, Xingqin An, Qingyong Li, Chao Wang, Haomin Yu, Xinyuan Zhou, and Yangli-ao Geng. "Application of XGBoost algorithm in the optimization of pollutant concentration." Atmospheric Research 276 (2022): 106238.

[20]. Wang, Yuanyuan, Shanfeng Sun, Xiaoqiao Chen, Xiangjun Zeng, Yang Kong, Jun Chen, Yongsheng Guo, and Tingyuan Wang. "Short-term load forecasting of industrial customers based on SVMD and XGBoost." International Journal of Electrical Power & Energy Systems 129 (2021): 106830.

[21]. Xiao, Feiyun, Yong Wang, Liangguo He, Hu Wang, Weihan Li, and Zhengshi Liu. "Motion estimation from surface electromyogram using adaboost regression and average feature values." IEEE Access 7 (2019): 13121-13134.
[22]. Gupta, Krishna Kumar, Kanak Kalita, Ranjan Kumar Ghadai, Manickam Ramachandran, and Xiao-Zhi Gao. "Machine learning-based predictive modelling of biodiesel production—A comparative perspective." Energies 14, no. 4 (2021): 1122.
[23]. Kankanala, Padmavathy, Sanjoy Das, and Anil Pahwa. "AdaBoost $^{+} $: An ensemble learning approach for estimating weather-related outages in distribution systems." IEEE Transactions on Power Systems 29, no. 1 (2013): 359-367.
[24]. Ampomah, Ernest Kwame, Zhiguang Qin, Gabriel Nyame, and Francis Effirm Botchey. "Stock market decision support modeling with tree-based AdaBoost ensemble machine learning models." Informatica 44, no. 4 (2021).
[25]. Kim, Daewon, and Michael Philen. "Damage classification using Adaboost machine learning for structural health monitoring." In Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems 2011, vol. 7981, pp. 659-673. SPIE, 2011.
[26]. Jou, I-Chang, Shih-Shien You, and Long-Wen Chang. "Analysis of hidden nodes for multi-layer perceptron neural networks." Pattern recognition 27, no. 6 (1994): 859-864.
[27]. Ismail, Mohammad Ridwan, Mohd Khalid Awang, M. Nordin A. Rahman, and Mokhairi Makhtar. "A multi-layer perceptron approach for customer churn prediction." International Journal of Multimedia and Ubiquitous Engineering 10, no. 7 (2015): 213-222.
[28]. Sozou, Peter D., Timothy F. Cootes, Christopher J. Taylor, E. C. Di Mauro, and Andreas Lanitis. "Non-linear point distribution modelling using a multi-layer perceptron." Image and Vision Computing 15, no. 6 (1997): 457-463.
[29]. Turkoglu, Bahaeddin, and Ersin Kaya. "Training multi-layer perceptron with artificial algae algorithm." Engineering Science and Technology, an International Journal 23, no. 6 (2020): 1342-1350.
[30]. Abd-elaziem, Ayman H., and Tamer HM Soliman. "A multi-layer perceptron (mlp) neural networks for stellar classification: A review of methods and results." International Journal of Advances in Applied Computational Intelligence 3, no. 10.54216 (2023).

