Machine Learning Applications in Industry Safety: Analysis and Prediction of Industrial Accidents

Md Mahbubur Rahman
Industrial and Systems Engineering
Lamar University
Beaumont, Texas, USA
mrahman41@lamar.edu

Amjad Hossain
Industrial and Systems Engineering
Lamar University
Beaumont, Texas, USA
ahossain9@lamar.edu

Md Arafat Sikder
Industrial and Systems Engineering
Lamar University
Beaumont, Texas, USA
msikder@lamar.edu

2024 International Conference on Smart Systems for applications in Electrical Sciences (ICSSES) | 979-8-3503-6404-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICSSES62373.2024.10561314

***Abstract*— This study leverages machine learning techniques to analyze industrial accident data from 12 manufacturing plants across 3 countries, with the objective of understanding the underlying patterns and predicting future accidents. By employing classification, clustering, and time-series forecasting methods, we aim to identify significant risk factors, characterize accident severity levels, and forecast accident trends. The classification model predicts accident severity with a focus on less severe incidents, while clustering analysis reveals distinct patterns related to industry sectors and specific risks. Time-series forecasting predicts a slight increase in accident counts, suggesting areas for proactive safety measures. Our findings contribute to the development of targeted safety interventions and policy formulation, emphasizing the need for industry-specific safety protocols. This research provides a foundational framework for predictive safety analytics, offering insights that could significantly enhance workplace safety management.**

Keywords—machine learning, industry, predictive safety, risk factors

I. Introduction

Work related injuries, within the manufacturing industry are a concern for those involved in the sector as they impact employee well-being and result in financial losses for companies. Despite improvements in safety measures and technology, the complex nature of operations continues to present challenges in preventing accidents. Traditional safety management methods often react to incidents based on data analysis. However, the use of machine learning and data analytics provides an approach to enhancing safety.

This article examines a dataset on accidents at manufacturing facilities in countries collected by IHM Stefanini, a company based in Brazil. The dataset includes factors such as accident severity levels, industry sectors, and risk elements, offering insights for analysis. Our study utilizes three machine learning techniques: classification, clustering and time series forecasting.

The classification method focuses on predicting the severity of accidents to identify high-risk situations before they happen. Clustering analysis aims to reveal patterns within the data that highlight risks associated with industry sectors. Lastly, time series forecasting is employed to anticipate accident trends and provide insights into periods of heightened risk.

This research makes three contributions to the field.

Our study introduces a model to anticipate accident severity to help identify potential severe accidents. Additionally, our analysis of accident patterns, through clustering, highlights the need for customized safety measures tailored to industry sectors. Furthermore, our forecasting model predicts accident numbers over time, aiding safety planning and resource distribution.

By utilizing machine learning techniques on industrial accident data, this research contributes to advancing workplace safety by offering perspectives and approaches to minimize accidents and strengthen safety protocols. The outcomes of our study have implications for safety managers, policymakers, and industry stakeholders by providing a data-driven foundation for making informed decisions and implementing strategic safety measures...

II. Literature Review

Recent studies have emphasized the importance of combining machine learning (ML) and deep learning (DL) technologies to enhance safety practices in settings, on construction sites. One researcher in a study [1] highlights the value of modeling in preventing accidents a viewpoint supported by another researcher [2] who showcases the effectiveness of convolutional neural networks (CNNs) for real time hazard detection. The role of technology in monitoring worker wellbeing is further investigated in another study [3] offering a proactive safety management approach. Exploring safety narratives through natural language processing as discussed in a study [4] reveals risk factors that can improve traditional safety protocols. Additionally an author in a paper [5] discusses the potential of transfer learning for industry safety applications indicating broader uses for DL models. Ethical concerns surrounding data privacy and model transparency are points raised in another paper [6] advocating for a balanced implementation of technology. In a study [7] integrating ML models into existing safety management systems as proposed by another author [8] represents progress, towards a comprehensive and data driven safety strategy.In the field of safety the construction industry encounters challenges as evidenced by the troubling fact that 1008 workers lost their lives in 2018 according to the US Occupational Safety and Health Administration (OSHA). This was mainly due, to accidents on site such as falls. Being hit by falling objects. Traditional safety monitoring methods that rely on patrols and surveillance are inadequate in addressing the dynamics of construction sites and accurately assessing worker fatigue. The rise of machine learning (ML) and deep learning (DL) driven by advancements in graphics processing units (GPUs) has introduced an era in safety monitoring and risk assessment in construction. Technologies like networks (CNNs) have proven to be highly effective in tasks like image recognition opening up new opportunities to improve job site safety management through automated hazard detection and worker safety monitoring. Furthermore combining DL with natural language processing (NLP) for analyzing reports, on construction safety has shown progress enabling identification of potential risks and the development of preventive measures.

The advancement of machine learning (ML) and deep learning (DL) applications marks a change, towards adopting data driven strategies to address the complexities of industrial

safety management. These research findings shed light on how advanced computational methods can be used to reduce risks and protect employees in the demanding field of safety.

III. Methodoloy

The methodology part of this study aims to explore how machine learning (ML) methods can improve safety focusing on predictive analytics and analyzing time series data. This research follows an approach using a mix of analyzing industrial accident data creating algorithms and evaluating performance to understand how ML can forecast and prevent accidents. Below is the architecture we propose in Figure 1.

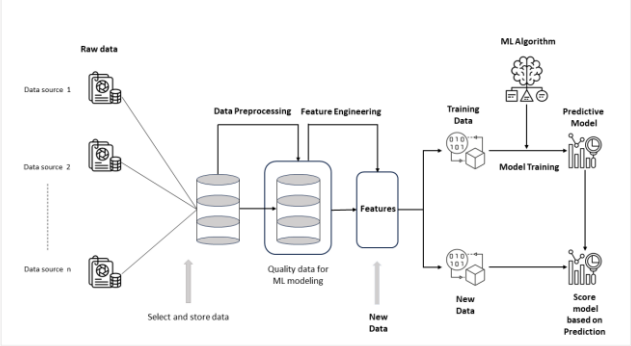

Figure 1: Proposed Architecture

A. Dataset Collection and Describtion :

We have collected the dataset from Kaggle [23] which is a public repository for storing machine leering competition dataset. Our dataset comes from one of Brazil's and the world's biggest industries, shared to address the scarcity of real-world manufacturing plant databases. The urgent need to understand and prevent workplace accidents, injuries, and fatalities drives the sharing of this data with the research community, aiming for new insights and safety improvements. The size of this data is (432, 10) which means it has 432 observations and 10 features.

The dataset encompasses accident records from 12 different manufacturing plants across three countries, representing a wide range of operational contexts. Each record is a unique incident, providing a detailed account of industrial accidents. Key aspects of the dataset include:

- **Date**: Timestamp information for each accident, allowing analysis of accident patterns over time.
- **Countries**: The accidents' countries are anonymized, offering geographical trends without compromising privacy.
- **Local**: The city or plant location is also anonymized, adding context while maintaining anonymity.
- **Industry Sector**: Identifies the sector of each plant, enabling sector-specific safety analysis.
- **Accident Level**: Severity of accidents is categorized from I (least severe) to VI (most severe).

- **Potential Accident Level**: Estimates potential severity, considering various factors, to gauge near-miss severity.
- **Genre**: Gender of the individuals involved, for demographic analysis.
- **Employee or Third Party**: Distinguishes between employees and external individuals affected by accidents.
- **Critical Risk**: Descriptions of risks involved, providing data for risk assessment and mitigation strategies.

B. Data Preprocessing:

Within the preprocessing phase of our dataset analysis, a critical step undertaken was the feature engineering of the 'date' variable, given the dataset's completeness with no missing values necessitating imputation. Initially, the 'date' variable, encapsulating the full date information, was decomposed into discrete components representing the day, month, and year. This decomposition allows for a more granular analysis of temporal patterns in accident occurrences.

Subsequently, to capture seasonal variations that could influence accident rates, the month component was further transformed into a 'season' variable. The division, into spring, summer, autumn and winter corresponds to the weather seasons giving us a glimpse into how changes in seasons could impact the frequency and types of accidents in environments. This shift is based on the idea that each season may present safety challenges allowing for an examination of accident trends. This precise feature engineering work is aimed at improving the dataset for machine learning models to recognize and understand patterns related to time and seasons, in accident occurrences. By including these calculated variables, we expect a analysis that can provide deeper insights into the factors influencing workplace safety over various time periods.

- **Spring**: September to November
- **Summer**: December to February
- **Autumn**: March to May
- **Winter**: June to August

C. Exploratory Data Analysis:

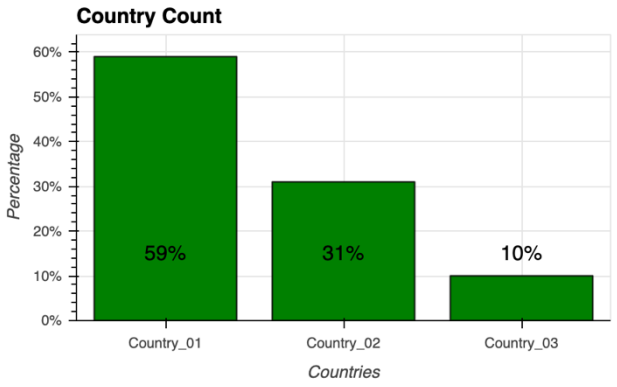

Figure 2: The percentage of accidents among anonymous countries.

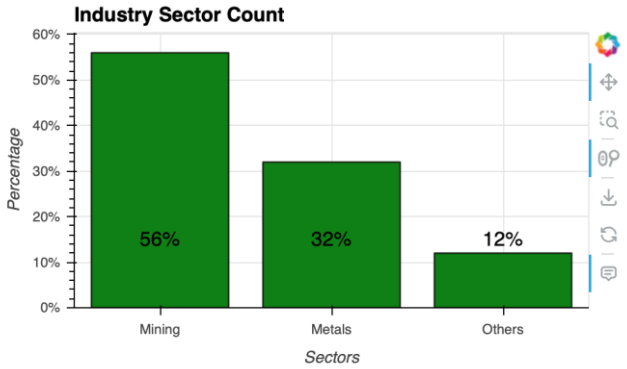

Figure 3: The percentage of accidents among anonymous three industry sector.

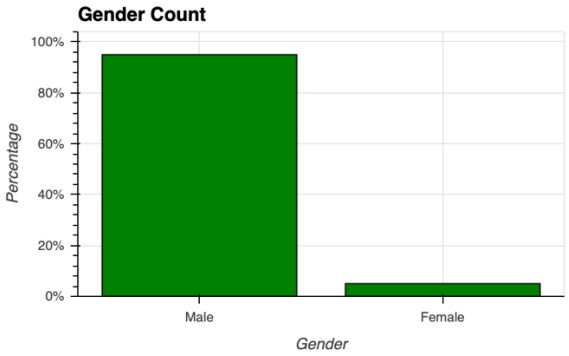

Figure 4: The percentage of accidents among two genders (Male and Female)

According to the figure 2, we can see that the maximum number of accidents happened in countries 1 (59%), figure 3 describe that the number of accidents is more in Mining industry (57%), figure 4 shown that the male having more accident rather than female.

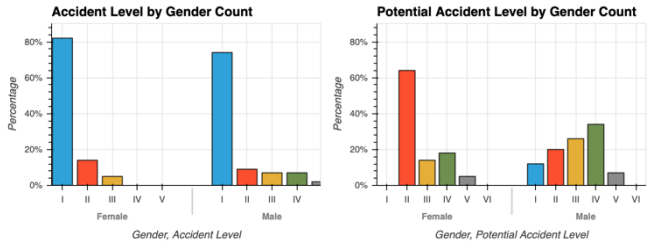

Figure 5: Accident Level by Gender

Figure 5, describe the accident level and potential accident level by gender (female and male)

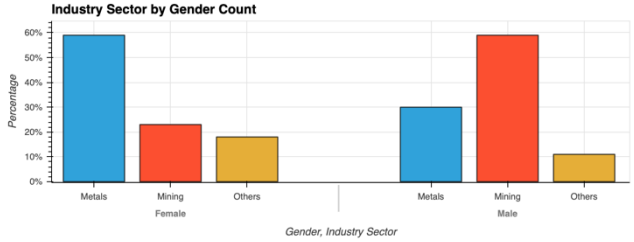

Figure 6: Industry Sector by Gender

Figure 6, describe the accident level and potential accident level by gender (female and male) working in different industries.

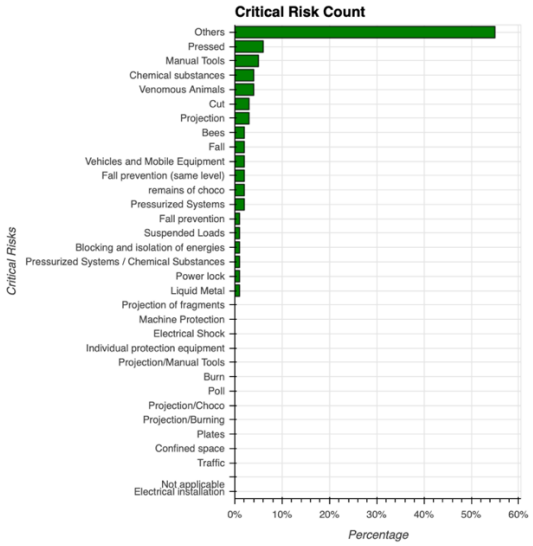

Figure 7: Critical Risk Types

In figure 7, Because most part of the Critical Risks are classified as 'Others', it is thought that there are too many risks to classify precisely. And it is also thought that it takes so much time to analyze risks and reasons why the accidents occur.

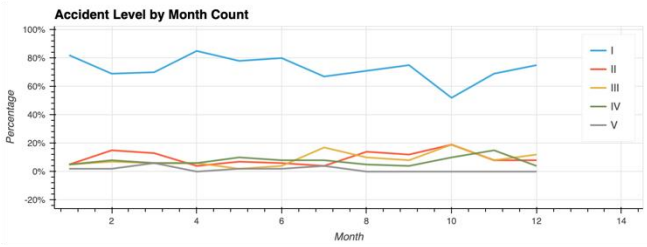

Figure 8: Accident Level by Month Count

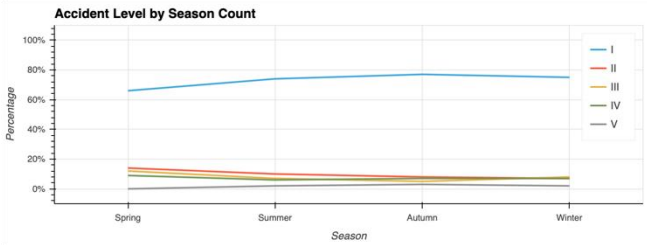

Figure 9: Accident Level by Season Count

Figure 8 and 9 represents the accident level occurs in moth and season respectively.

D. Classfiicaiton Methods:

In our project we've used classification techniques to forecast accident severity from a dataset each, with its advantages and suitable for different aspects of the task. The K Nearest Neighbors (KNN) method relies on the idea that similar cases tend to be near each other in feature space making it highly effective for data with decision boundaries. Logistic Regression offers a framework for estimating the likelihoods of accident severity levels providing clear interpretability and suitability for situations with linear relationships between features. The Decision Tree Classifier is adept at capturing linear connections without requiring feature scaling offering

simple decision rules. The Random Forest technique boosts prediction accuracy and resilience by combining decision trees effectively handling overfitting and diverse data characteristics. Lastly the Gradient Boosting Classifier improves prediction accuracy by correcting errors from models focusing on challenging cases to enhance overall predictions steadily. Together these methods create a strategy by leveraging their strengths to address the intricacies involved in forecasting accident severity levels accurately and comprehensively.

E. Clustering:

During our investigation, into accident data grouping we started by changing accident severity levels from words to numbers and picked out factors for grouping. By using the K means technique we identified five groupings based on the bend in the data method. Our examination unveiled trends among these groups showing differences in industry distribution, accident seriousness and common risk elements. "Others" emerged as the category, in all clusters. The outcomes were summarized in a organized table and detailed report emphasizing the significance of recognizing safety trends to industries and individual groups. This underscores how clustering can guide targeted safety actions and interventions effectively.

F. Time-series Forecasting:

In our project we are exploring time series analysis and forecasting which's crucial, for understanding and predicting trends based on historical data. The core of this method is the ARIMA model, for AutoRegressive Integrated Moving Average. This model is well known for its effectiveness in forecasting time series data by identifying patterns like trends and seasonality adjusting them to ensure the data remains consistent. By incorporating elements that consider autoregression (AR) differencing (I) for achieving consistency and moving averages (MA) ARIMA provides a framework, for making predictions. By determining the model parameters (p, d q) through autocorrelation and partial autocorrelation analyses followed by validating the model we can forecast values with a certain level of precision. This systematic approach enables us not to understand the dynamics of data but also to estimate future accident frequencies demonstrating how time series forecasting can be applied practically in real world situations.

IV. Results and Discussion

In this section, we talk about results got from machine learning models [24], closeting effect, and time-series forecasting. First, we have shown the output for machine learning algorithms, second, we have shown output for k-means clustering and thirdly and finally, we have shown the output for ARIIMA model [25].

Table 1: Accuracy, Precision, Recall and F1-score for Five Machine Learning Models

Model	Accuracy	Precision	Recall	F1-score
KNN	88%	81%	82%	84%
LR	94%	91%	93%	92%
RF	92%	85%	88%	90%
DT	78%	76%	78%	77%
GB	92%	88%	86%	89%

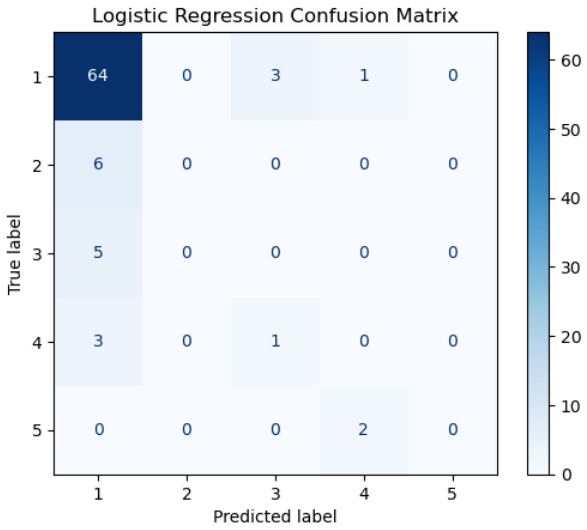

Figure 10: Confusion Matrix for Best Machine Learning Model (Logistic Regression)

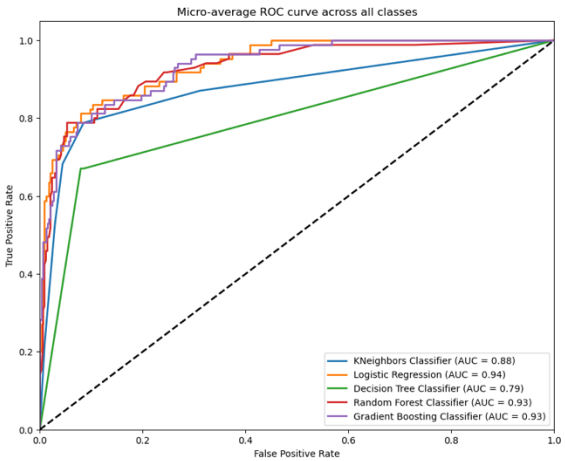

Figure 11: AUC-ROC curve for all ML models.

From Table 1, figure 10 and 11, It is clearly identified that the Logistic Regression model surpass all the other model for accuracy, precision, recall, f1-scroe and roc value.

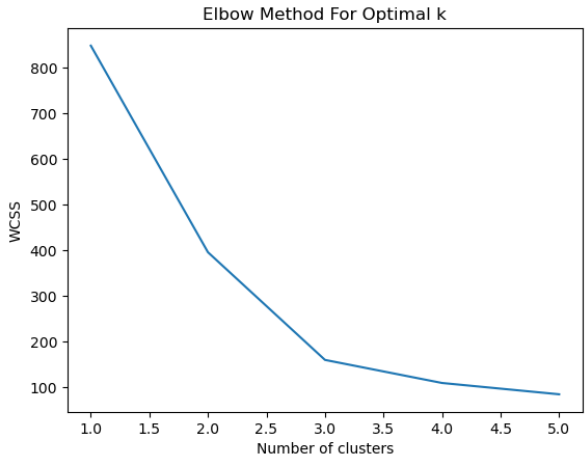

Figure 12: Finding Optimal K

We chose optimal K value 4 shown in Figure 12.

Table 2: Clustering Analysis Results

Cluster	Industry - Metals	Industry -Mining	Industry - Others	Average Accident Level	Average Potential Accident Level
0	43	60	0	1.14	3.00
1	12	34	4	2.62	4.04
2	55	49	40	1.05	1.66
3	16	70	4	1.00	4.11
4	8	27	2	4.19	4.43

The Table 2 describe the cluster analysis results.

1. Distribution of Industry Sectors within each Cluster

- **Cluster 0:** Predominantly Mining (60) with a significant presence of Metals (43).
- **Cluster 1:** A mix of Mining (34) and Metals (12), with a few in Others (4).
- **Cluster 2:** A balanced distribution across Mining (49), Metals (55), and Others (40).
- **Cluster 3:** Dominated by Mining (70) with Metals (16) and Others (4).
- **Cluster 4:** Mainly Mining (27) and Metals (8), with a few in Others (2).

2. Average Severity of Accidents across Clusters

- **Cluster 0:** Average Accident Level of 1.14 and Potential Accident Level of 3.00.
- **Cluster 1:** Higher severity with an Average Accident Level of 2.62 and Potential Accident Level of 4.04.
- **Cluster 2:** Lower severity with an Average Accident Level of 1.05 and Potential Accident Level of 1.66.
- **Cluster 3:** Lowest current severity with an Average Accident Level of 1.00 and a high Potential Accident Level of 4.11, indicating potentially severe accidents.
- **Cluster 4:** Highest severity with an Average Accident Level of 4.19 and Potential Accident Level of 4.43.

3. Most Common Risk Factors in each Cluster

- **All Clusters:** The most common risk factor identified across all clusters is "Others."

These results provide a nuanced view of the dataset, revealing how accidents are distributed across different industry sectors, the severity of accidents within each cluster, and the most common risk factors associated with each cluster. This information can be instrumental in targeting safety measures and interventions more effectively. If there's anything more specific, you'd like to explore or any other assistance you need, feel free to ask!

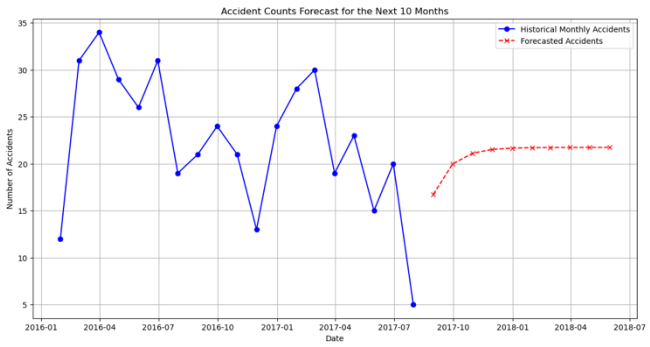

Figure 13: Accidents counts Forecast for the Next 10 months.

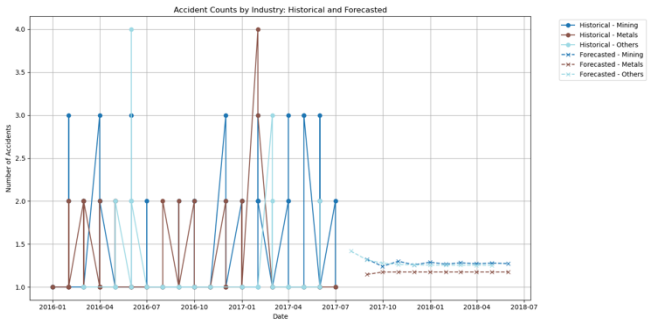

Figure 14: Accident Counts by Industry: Historical and Forecasted

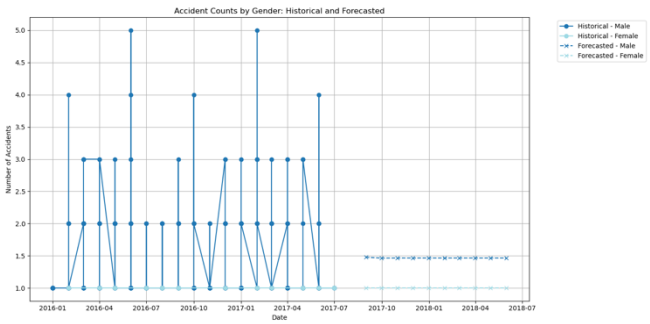

Figure 15: Accident Counts by Industry: Historical and Forecasted

Figure 13 shown the forecast accident number for next 10 moths which is from 2017-08 to 2018-08, figure 14 and figure 15 shown the accident number-based industry type and gender type respectively. More specifically, it tells us that the male is in Mining Industry will have more accident in the next 10 months which can be reduce by taking proper steps in the beforehand.

This study faces limitations due to its narrow dataset scope, confined to the years 2017 and 2018, without explicit country, location names, or detailed risk categories. The use of "Others" as a catch-all category for critical risks introduces ambiguity, hindering the identification of specific hazard

patterns and the development of targeted mitigation strategies. This limitation restricts the depth of analysis and specificity of the findings, making it challenging to generalize the results to broader contexts or to accurately pinpoint areas for improvement in industrial safety measures.

V. CONCLUSION

Our investigation into the predictive power of machine learning for assessing accident severity in industrial settings, while limited to data from 2017 and 2018 and lacking geographical and detailed risk categorization, underscores the potential for data-driven safety enhancements. Despite these constraints, the study highlights the viability of machine learning models, particularly Logistic Regression, in forecasting accident occurrences and severity. The absence of detailed risk information points to a critical area for future research: refining risk classification systems to enhance predictive accuracy and safety interventions. As such, future efforts should focus on expanding data collection to include more diverse variables, extending the temporal and geographical scope of the dataset, and improving risk categorization to better inform safety practices and policies.

REFERENCES

[1] Oh, J., Washington, S.P. and Nam, D., 2006. Accident prediction model for railway-highway interfaces. *Accident Analysis & Prevention*, *38*(2), pp.346-356.

[2] Chan, K.Y., Abu-Salih, B., Qaddoura, R., Ala'M, A.Z., Palade, V., Pham, D.S., Del Ser, J. and Muhammad, K., 2023. Deep Neural Networks in the Cloud: Review, Applications, Challenges and Research Directions. *Neurocomputing*, p.126327.

[3] Mejia, C., Ciarlante, K. and Chheda, K., 2021. A wearable technology solution and research agenda for housekeeper safety and health. *International Journal of Contemporary Hospitality Management*, *33*(10), pp.3223-3255.

[4] Demner-Fushman, D., Elhadad, N. and Friedman, C., 2021. Natural language processing for health-related texts. In *Biomedical Informatics: Computer Applications in Health Care and Biomedicine* (pp. 241-272). Cham: Springer International Publishing.

[5] Newrzella, S.R., Franklin, D.W. and Haider, S., 2021. 5-dimension cross-industry digital twin applications model and analysis of digital twin classification terms and models. *IEEE Access*, *9*, pp.131306-131321.

[6] Whittlestone, J., Nyrup, R., Alexandrova, A., Dihal, K. and Cave, S., 2019. Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research. *London: Nuffield Foundation*.

[7] Sattari, F., Lefsrud, L., Kurian, D. and Macciotta, R., 2022. A theoretical framework for data-driven artificial intelligence decision making for enhancing the asset integrity management system in the oil & gas sector. *Journal of Loss Prevention in the Process Industries*, *74*, p.104648.

[8] Su, Y., Mao, C., Jiang, R., Liu, G. and Wang, J., 2021. Data-driven fire safety management at building construction sites: Leveraging CNN. *Journal of management in engineering*, *37*(2), p.04020108.

[9] Rafindadi, A.D.U., Napiah, M., Othman, I., Mikić, M., Haruna, A., Alarifi, H. and Al-Ashmori, Y.Y., 2022. Analysis of the causes and preventive measures of fatal fall-related accidents in the construction industry. *Ain Shams Engineering Journal*, *13*(4), p.101712.

[10] Ball, J., Vosberg, S.J. and Walsh, T., 2020. A Review of United States Arboricultural Operation Fatal and Nonfatal Incidents (2001-2017): Implications for Safety Training. *Arboriculture & Urban Forestry*, *46*(2).

[11] Tighe, S., 2020. An Experiential Analysis of Job Site Safety: Delineating Between Positive Safety Culture and Excessive Safety (Doctoral dissertation, Indiana State University).

[12] Liu, W., Meng, Q., Li, Z. and Hu, X., 2021. Applications of computer vision in monitoring the unsafe behavior of construction workers: Current status and challenges. *Buildings*, *11*(9), p.409.

[13] Sadeghi, S., Soltanmohammadlou, N. and Rahnamayiezekavat, P., 2021. A systematic review of scholarly works addressing crane safety requirements. *Safety Science*, *133*, p.105002.

[14] Hou, L., Chen, H., Zhang, G. and Wang, X., 2021. Deep learning-based applications for safety management in the AEC industry: A review. *Applied Sciences*, *11*(2), p.821.

[15] Bazaluk, O., Tsopa, V., Cheberiachko, S., Deryugin, O., Radchuk, D., Borovytskyi, O. and Lozynskyi, V., 2023. Ergonomic risk management process for safety and health at work. *Frontiers in Public Health*, *11*.

[16] Javed, M.A., Muram, F.U., Punnekkat, S. and Hansson, H., 2021. Safe and secure platooning of Automated Guided Vehicles in Industry 4.0. *Journal of systems architecture*, *121*, p.102309.

[17] Guo, B.H., Yiu, T.W., González, V.A. and Goh, Y.M., 2017. Using a pressure-state-practice model to develop safety leading indicators for construction projects. *Journal of construction engineering and management*, *143*(2), p.04016092.

[18] Poh, C.Q., Ubeynarayana, C.U. and Goh, Y.M., 2018. Safety leading indicators for construction sites: A machine learning approach. *Automation in construction*, *93*, pp.375-386.

[19] Xu, J., Cheung, C., Manu, P. and Ejohwomu, O., 2021. Safety leading indicators in construction: A systematic review. *Safety science*, *139*, p.105250.

[20] Hinze, J., Thurman, S. and Wehle, A., 2013. Leading indicators of construction safety performance. *Safety science*, *51*(1), pp.23-28.

[21] Masood, R., Mujtaba, B., Khan, M.A., Mubin, S., Shafique, F. and Zahoor, H., 2014. Investigation for safety performance indicators on construction projects. *Science International*, *26*(3), pp.1403-1408.

[22] Versteeg, K., Bigelow, P., Dale, A.M. and Chaurasia, A., 2019. Utilizing construction safety leading and lagging indicators to measure project safety performance: A case study. *Safety Science*, *120*, pp.411-421.

[23] The dataset is download from : https://www.kaggle.com/datasets/ihmstefanini/industrial-safety-and-health-analytics-database

[24] K. Mridha, M. M. Uddin, J. Shin, S. Khadka and M. F. Mridha, "An Interpretable Skin Cancer Classification Using Optimized Convolutional Neural Network for a Smart Healthcare System," in *IEEE Access*, vol. 11, pp. 41003-41018, 2023, doi: 10.1109/ACCESS.2023.3269694

[25] E. Saha, R. Saha and K. Mridha, "Short-Term Electricity Consumption Forecasting: Time-Series Approaches," 2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2022, pp. 1-5, doi: 10.1109/ICRITO56286.2022.9964624.