Journal of Information Systems and Informatics
Vol. 7, No. 1, March 2025 e-ISSN: 2656-4882 p-ISSN: 2656-5935
DOI: 10.51519/journalisi.v7i1.979 Published By DRPM-UBD

Enhancing Hazard Detection and Risk Severity Assessment in Construction through Multinomial Naive Bayes and Regression

Akaninyene Michael Akwaisua[1], Anietie Ekong[2], Godwin Ansa[3]

[1,2,3]Department of Computer Science, Akwa Ibom State University, Mkpat Enin, Nigeria
Email: [1]akaninyeneakwaisua@gmail.com, [2]anietieekong@aksu.edu.ng, [3]godwinansa@aksu.edu.ng

Abstract

This research delves into the crucial area of hazard detection and risk severity assessment within the construction industry, using machine learning techniques. The dataset utilized is from the Chinese Construction Company (CCECC), Uyo, Nigeria. Comprising over 100,000 instances, it captures various hazard categories prevalent in construction sites, providing a comprehensive foundation for predictive analysis. In the first phase of the study, the system is designed to detect hazards present in construction sites. Leveraging these data, the machine learning models are trained to predict potential hazards based on the information provided. Through TF-IDF vectorization, a feature extraction technique, the textual data is transformed into numerical representations. Multinomial Naive Bayes is employed for hazard classification due to its efficacy in handling text data, and with it, an accuracy of 0.99 was obtained. Subsequently, the trained model was evaluated to assess its performance and the severity of identified hazards are evaluated. The system quantifies the potential risk posed by each hazard using the risk severity attribute. Using the Linear Regression algorithm, the model predicts the severity of risks based on textual descriptions of a hazard. In practical application, the research stresses the significance of risk management strategies in the construction industry to mitigate potential harm to personnel and infrastructure. This research contributes to advancing safety protocols within the construction sector, advocating for a culture of vigilance and precaution to address risks effectively.

Keywords: Hazard, Risk, Construction, Regression Analysis, Naive Bayes, Machine Learning.

1. INTRODUCTION

At construction sites, ensuring the safety of workers, bystanders, and the surrounding environment is crucial. One of the key processes in this effort is hazard identification—basically, spotting potential dangers like heavy machinery, hazardous materials, and risks associated with working at heights or in adverse weather. Risks are of different dimensions and each dimension has its effect. Sometimes, it is preferred to use multi-criteria for such decision-making [1]. With the improvement in computers' computation capacities and the advent of Artificial

Intelligence, computers have come in handier in improving the ways that tasks are performed and this has led to an increase in the accuracy of computations and predictions and a considerable reduction in job delivering time and the number of deaths that were hitherto associated with traditional approaches [2]. When construction managers can systematically assess these hazards, they can implement effective measures to minimize risks and prevent accidents. Once hazards are identified, the next step is to gauge how serious these risks might be.

The Multinomial Naive Bayes Algorithm is particularly effective for handling text-based data, especially in natural language processing tasks. Moreover, creating a culture of safety awareness among everyone involved—project managers, subcontractors, and workers alike—reinforces the importance of these practices. Construction sites are dynamic and complex, filled with potential hazards that require robust management strategies using machine learning approaches and others. The industry often relies on manual inspections and standardized protocols for hazard identification and risk management, which can lead to oversights and slow responses to emerging risks. With an urgent need to enhance safety measures, there's increasing interest in integrating machine learning (ML) techniques, particularly the Multinomial Naive Bayes (MNB) approach, to strengthen hazard identification and risk management. The construction sector generates a lot of unstructured textual data, including incident reports and safety documentation. Traditional methods of sorting through this information can be tedious and often miss important patterns that might indicate potential hazards. Machine Learning Algorithms are especially delivering good outcomes using intricate datasets where Classifiers were trained to detect known classes and to adapt to new ones [3].

Here is where natural language processing—an aspect of artificial intelligence—comes into play. It allows machines to understand and interpret human language, making it easier to extract valuable insights from data. By applying machine learning approaches such as MNB algorithms and logistics regression to the textual data from construction sites, researchers hope to foster a more proactive approach to hazard identification and risk assessment. The existing literature on construction safety highlights a pressing need for innovative solutions to overcome the limitations of traditional hazard identification methods. Accidents in construction often arise from a complex interplay of factors, which calls for a more adaptive, data-driven approach. Tailored machine learning models can analyze historical data, spot recurring patterns, and predict potential hazards before they escalate into serious incidents. MNB techniques enhance these models' ability to process textual data efficiently, providing a clearer understanding of contextual risks.

This research aims to evolve construction site safety by developing and validating machine learning models that leverage MNB for better hazard identification and risk management. Research has shown that there is a huge gap in severity analysis

of hazards. The lack of updated approach that use advance machine learning to analyze hazard in construction industries is one of the motivations for this research. Current systems only detect hazards and does examine severity of those hazards to ascertain damage capability. Some systems that proposed to analyze hazard severity ended up employing manual or unautomated methods, making this work more pertinent. By applying these advanced technologies, we hope to transform safety practices in construction, providing a framework that can more effectively address the industry's unique challenges. We expect that our findings will offer construction professionals practical tools to proactively identify and mitigate potential hazards, ultimately leading to a safer work environment. Given the increasing complexity of industrial processes and the ever-growing volume of data, there's a clear need for advanced techniques in hazard identification and construction risk management. Traditional methods often struggle to keep up with the fast-paced nature of modern work environments. This research focuses on construction sites, where our goal is to identify and manage risks effectively. Every construction company prioritizes employee safety, making it a top concern on every job site, before damage assessment. By using the Multinomial Naive Bayes approach, we aim to pinpoint risk factors and hazardous conditions within the construction industry, ultimately supporting better risk management and worker safety. Our dataset will be compiled from site reports and documentation, blending textual and numerical data to fully leverage the capabilities of the MNB algorithm.

The building and construction industries are inherently dynamic, involving a myriad of processes, materials, and personnel, making hazard identification and risk management crucial aspects of ensuring safety and project success, and classification problem solution design. Over the years, various approaches based on machine learning (ML) techniques have been used for the minimization of these difficulties [4]. In this sector, hazards can range from physical dangers such as working at heights and handling heavy machinery to environmental factors like inclement weather and site conditions. Effective hazard identification begins with a comprehensive assessment of the construction site, considering both the inherent risks associated with the construction processes and potential external factors that could impact the project. Risk management in the building and construction industries encompasses a range of strategies aimed at mitigating identified hazards considering the importance of construction to the global economy and its fatality records [5]. This includes the development of thorough safety protocols, the provision of proper safety training for personnel, and the implementation of engineering controls to reduce risks. Additionally, construction projects often require meticulous planning and monitoring to identify potential risks throughout the various phases of a project, from initial design to project completion. Effective communication among project stakeholders is crucial to ensuring that all parties are aware of potential hazards and are equipped with the knowledge and tools to manage and minimize risks in the industry effectively.

Traditional methods for hazard identification and risk management in construction have long been the backbone of ensuring safety and success on building sites. One widely used approach involves on-site inspections and audits carried out by safety professionals and regulatory bodies. These inspections thoroughly assess the site, equipment, and processes, pinpointing potential dangers ranging from unsafe working conditions to inadequate safety measures. The exceptional abilities of machine learning technique such as Convolutional Neural Network (CNN) can be leveraged for monitoring and detection for enhanced safety [25]. Regular inspections are essential not only for maintaining compliance with safety standards but also for proactively addressing any emerging risks in an intelligent approach, in recent times there has been integration of smart management methods into project management [6]. Another staple of the construction industry is the analysis of incidents and accidents. By digging deep into past incidents, project teams can uncover the root causes of accidents and near misses. This kind of retrospective analysis reveals patterns and systemic issues, which can inform risk management strategies moving forward. Lessons learned from previous incidents help refine safety protocols and training programs, ensuring that similar risks are mitigated in the future, that is why it is necessary to develop some models such as object recognition models that check whether there are construction workers at the site [7] among others. Job Safety Analysis (JSA) and hazard analysis are also crucial tools used to identify and evaluate risks tied to specific tasks. These methods break down job activities into steps, allowing teams to identify hazards at each stage and determine appropriate control measures. Those directly engaged in construction work often have invaluable insights into the specific challenges and hazards they face daily. Toolbox talks and safety meetings are common practices that encourage workers to discuss potential hazards and share their experiences. This bottom-up approach fosters a safety culture where everyone—from project managers to laborers—actively participates in identifying risks and contributing to overall safety. Also, DCNN algorithms could be used to model hazards being that they are good when working on image data [8]

In recent years, machine learning (ML) has emerged as a game-changing technology in the realm of hazard identification and risk management within the construction industry. These algorithms—part of the broader field of artificial intelligence—enable construction professionals to sift through massive amounts of data, identify patterns, and make informed predictions. This shift toward data-driven approaches is revolutionizing traditional safety practices. One of the standout applications of ML in construction is predictive analytics for hazard identification. By analyzing historical data from construction sites—like incidents and near misses—ML algorithms can spot patterns that may suggest potential hazards. Workers at construction sites face numerous risks [9]. This capability allows teams to anticipate similar risks in future projects, enabling proactive measures that enhance safety. Moreover, ML excels at automating the analysis of

complex construction data, including Building Information Modeling (BIM) data, sensor information, and project documentation. This automation speeds up the hazard identification process and minimizes human error, resulting in more accurate risk assessments. Natural Language Processing (NLP), a branch of ML, helps extract valuable insights from textual data, such as project reports and incident logs. By analyzing unstructured text, NLP algorithms can identify trends and recurring themes related to hazards, providing deeper contextual understanding that informs risk assessments. Another significant benefit of ML is its ability to adapt to real-time conditions. It can continuously monitor data from construction sites, including environmental factors and equipment status. This dynamic monitoring allows for immediate hazard identification and timely interventions, shifting the focus from reactive to proactive risk management. However, challenges remain. Effective ML applications require high-quality, labeled data; without large and diverse datasets, these algorithms may struggle to learn effectively. Additionally, many ML models operate as "black boxes," making it tough for construction professionals to understand how they reach their predictions. Building trust in these systems is crucial for their successful adoption in the industry. The potential of machine learning in hazard identification and risk management is immense. It not only enhances predictive analytics and automates data analysis but also facilitates real-time monitoring, leading to more effective and proactive safety practices [10].

Addressing challenges like data quality and model interpretability will be vital for the successful integration of ML in the construction sector. Beyond just identifying hazards, ML can innovate risk management strategies. These algorithms can develop adaptive risk models that evolve alongside project complexities, continuously learning from new data. This flexibility helps construction teams respond more effectively to emerging risks and unexpected challenges during projects. Decision support systems powered by ML offer valuable insights for risk mitigation, analyzing historical data and project-specific parameters to recommend tailored strategies. Construction professionals can implement targeted interventions and optimize resource allocation, leading to a more effective approach to risk management. Furthermore, ML can enhance incident analysis by uncovering deeper insights into root causes and contributing factors of accidents. By examining historical data, these algorithms can reveal correlations and trends that might be missed by traditional methods, allowing teams to address underlying issues more effectively. Incorporating ML into safety training programs is another promising avenue. Algorithms can analyze worker behavior to identify patterns that suggest lapses in safety protocols. This information can inform targeted training interventions, ultimately improving the overall safety culture on construction sites. Despite the potential benefits, there are challenges, including the need for skilled personnel to develop and maintain ML models, concerns about data privacy, and the possibility of bias in algorithms. Additionally, the

construction industry may face resistance to adopting new technologies due to established workflows. Overcoming these hurdles will require collaboration between technology developers, industry stakeholders, and regulatory bodies to establish best practices for responsibly integrating machine learning into hazard identification and risk management in construction.

2. RELATED WORKS

Wei et al. [11] conducted a bibliometric analysis on machine learning in industrial risk assessment. They identified three developmental stages and key research areas, including machine learning algorithms, Industry 4.0, and autonomous driving applications. The study highlighted active research around terms like "Random Forest" and "Internet of Things," indicating ongoing interest and potential advancements in the field. Mamdouh et al. [12] developed a machine learning methodology to predict the impact of construction activities on air traffic during airport expansion projects. The study addressed challenges related to construction disruption and aimed to help planners identify phasing plans that minimize flight delays. The methodology comprises four stages: data collection, preprocessing, model training, and evaluation, illustrated through a case study involving five machine learning models. These models quantify the impact of airport closures on flight ground movement time and compare their performance for accuracy. Notably, the approach allowed for efficient assessment of alternative construction without extensive simulations, facilitating quicker decision-making for planners and enhancing overall project efficiency.

George et al [13] explored the use of ensemble machine learning for construction safety risk assessment, addressing the need to evaluate safety at construction sites. Utilizing a dataset of 4,847 event reports from the Occupational Safety and Health Administration (2015-2017), the study identified key risk factors categorized into four groups. Predictive models were developed using five classifiers, with ensemble methods outperforming other approaches. The resulting predictive model aided safety management teams by providing insights and enabling proactive measures to reduce the risk of accidents, highlighting the value of machine learning in enhancing construction safety practices. Yin et al. [14] developed a machine learning-based security assessment model for marine information systems, crucial for maritime security. The model combined the analytic hierarchy process and fuzzy comprehensive evaluation to create a comprehensive dataset. Three algorithms, decision tree, KNN, and SVM were used to construct the assessment model. The study reported high classification accuracy, with the decision tree achieving 100% on the training set and strong results on the test set (98.20% for decision tree, 97.00% for KNN, and 97.40% for SVM). This research highlights the model's effectiveness as a valuable tool for security professionals, enhancing

the assessment of marine information system security levels through innovative machine learning techniques.

Almahameed et al. [15] investigated the use of machine learning and Particle Swarm Optimization (PSO) for predictive modeling and costs optimization in construction project management. The research focuses on systematically reducing expenses while maximizing value within budget constraints, employing various machine learning algorithms like Linear Regression, Decision Trees, SVM, and more. Key objectives included enhancing cost estimation accuracy and identifying critical cost factors. The study found out that voting regression outperformed individual models and highlighted the importance of feature selection for effective resource allocation. PSO is noted for its effectiveness in reducing construction waste and improving cost estimation. Overall, the research emphasized the potential of these advanced techniques to aid industry professionals in decision-making, resource allocation, and enhancing project profitability.

Elhishi et al. [16] focused on accurately predicting concrete strength, particularly for high-performance concrete used in durable structures. The study evaluated eight machine learning models, including Linear Regression and XGBoost, using a dataset of 1,030 concrete samples. Results showed that XGBoost significantly outperformed other models, achieving an R-Squared (R^2) of 0.91 and a Root Mean Squared Error (RMSE) of 4.37. The research utilized the SHAP technique to provide civil engineers with insights into concrete mix design, enhancing decision-making. Overall, the findings underscore the effectiveness of machine learning, especially ensemble techniques, in optimizing concrete performance and improving construction practices. Jenisha et al. [4] reviewed machine learning approaches for classifying rock masses in the context of tunnels used in hydropower systems. The study emphasizes the challenges of accurately assessing rock mass parameters during the pre-operation, real-time, and post-operation phases, despite existing numerical and empirical methods. It explored the potential of ML techniques to address these challenges. The paper assessed current ML applications in rock mass classification and recommended the development of industry-specific models while highlighting the strengths and weaknesses of various techniques. Overall, they provided an enhanced methodology for rock mass classification in hydropower tunnel projects and guided the integration of ML technologies in this area.

Pham et al. [17] explored damage assessment in beam-like structures using machine learning, addressing the challenge of evaluating damage in actively operating constructions without relying on excitation source information. The study introduced a supervised machine learning method that utilized spectral signal correlations as input features for artificial neural networks (ANN) and decision tree algorithms. This innovative approach enabled real-world applications by

identifying damage characteristics, such as new cuts and their positions, without needing detailed excitation data. To validate the method, an experimental setup with a supported beam model was used, demonstrating the effectiveness of the proposed spectral correlation features extracted from vibration data for assessing structural integrity. Olena et al [18] examined the integration of artificial intelligence (AI) in the construction industry, highlighting its transformative potential across five key stages: Planning and Design, Risk Assessment, Resource Management, Automation and Monitoring, and Quality Assessment. In the Planning and Design phase, AI automates design processes by analyzing relevant data. During Risk Assessment, forecasting algorithms identify potential risks and inform management strategies. AI optimized logistics and labor scheduling in Resource Management. In Automation and Monitoring, AI utilizes drones and real-time systems for efficient task execution and progress tracking. Finally, in Quality Assessment, AI analyzed completed work to ensure compliance with standards. The paper emphasized AI's role in resource planning, showcasing its benefits and future prospects in enhancing efficiency and effectiveness in construction processes.

Mostofi et al. [19] explored the use of unsupervised machine learning (ML) for safety risk management in construction, motivated by high accident rates. The study investigates how machines can autonomously analyze safety datasets to identify relationships between risk factors and accident outcomes without human intervention. It highlighted clustering and dimensionality reduction techniques, specifically Principal Component Analysis (PCA) and K-means clustering, to uncover patterns and similarities in occupational safety data. This research emphasized the potential of unsupervised ML to enhance decision-making processes in construction safety risk assessments. Le et al. [20] developed a machine learning approach for risk assessment of expressway bridges in Vietnam, addressing the complexities of bridge construction and operation. The research identifies key risk factors and utilizes algorithms like artificial neural networks and the Random Forest Algorithm to enhance accuracy in risk classification. By focusing on impactful factors, the model improved decision-making for bridge management, enabling proactive risk management and contributing to sustainable expressway development. The findings highlight the effectiveness of machine learning in providing insights for operational safety and quality assurance.

Sobhan et al. [21] investigated machine learning techniques for predicting injury severity in occupational accidents using both reactive and proactive data. The study combined investigation and inspection reports to address challenges with unstructured texts and class imbalance. By applying topic modeling and the k-means SMOTE (KMSMOTE) algorithm, the research achieved higher prediction accuracy. Results showed that incorporating proactive data significantly improved injury severity predictions. Additionally, the study generated 19 safety decision

rules that inform factors influencing injury outcomes, providing practical guidance for safety management and interventions to enhance workplace safety. Ahmed et al. [5] studied the prediction of road accident severity, emphasizing the importance of human factors like alcohol, drugs, age, and gender. Utilizing single and ensemble machine learning methods, they framed the problem as a classification task. The Random Forest (RF) algorithm outperformed others, achieving 86.64% accuracy in binary classification and 67.67% in multiclass classification. While other methods like logistic regression and XGBoost showed promise, RF led in accuracy. The research stressed the need to consider diverse factors in accident severity prediction, contributing valuable insights for improving road safety strategies.

2.1. Research Problem

In the construction industry, ensuring safety at job sites is paramount to protect workers and mitigate risks of accidents or injuries. Despite the implementation of safety protocols and regulations, hazards remain prevalent, and accurately predicting the severity of risks associated with these hazards poses a significant challenge. Traditional methods of hazard identification and risk assessment often rely on subjective evaluations, which may not fully capture the dynamic nature of construction site hazards and their potential impacts. Therefore, there is a need for a more objective and data-driven approach to hazard identification and risk severity prediction in construction, leveraging machine learning techniques to analyze large of data from reliable sources and improve the accuracy of risk assessments.

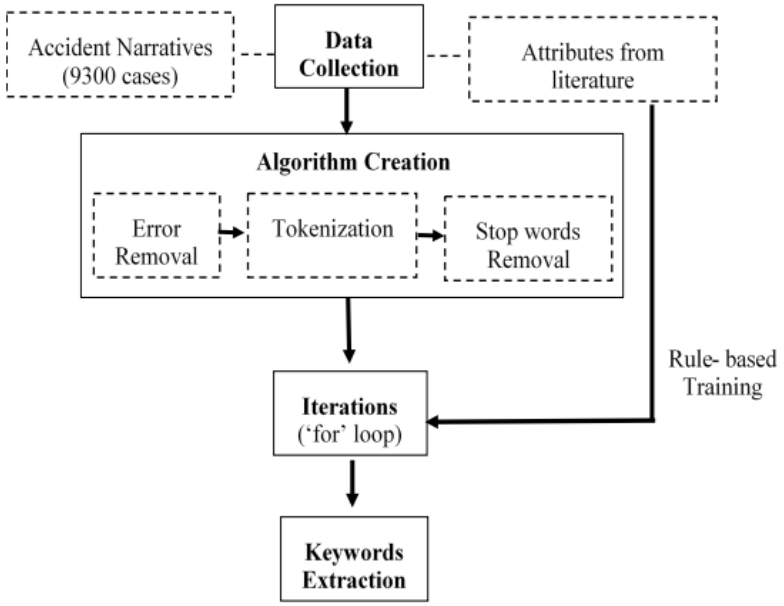

Figure 1. Architecture of the Existing System [22]

In a work modeled by [22], to improve construction site safety by identifying potential hazards through the analysis of accident narratives, they used NLP techniques to extract meaningful information from accident reports. The research methodology involves gathering accident narratives, preprocessing the data, and applying a rule-based iteration approach to extract hazards using NLTK. In the work, the authors used accident narratives in construction sites, meaning that the work was on only observed accidents or hazard type, but not predicting hazard based on the nature of the work, equipment, site arrangement figure 1 is the architecture or the system by [22]. In the proposed system we will bring these factors into consideration to have a better and a more robust system for today's construction sites which is dynamic in nature and new hazards are happening.

3. METHOD

The proposed system begins with data collection, where information on various hazards observed at construction sites, CCECC is gathered. This data serves as the foundation for training and testing the Multinomial Naive Bayes classification model. Next, preprocessing techniques are applied to the collected data to ensure its quality and relevance for analysis. This includes tasks such as data cleaning, feature extraction, and normalization to prepare the dataset for training. Data shuffling helps to reduce variance and in ensuring that the models remain general and not affected by overfitting [24]. Once the data is preprocessed, the Multinomial Naive Bayes algorithm is trained on the labeled dataset to learn patterns and relationships between different hazard classes. During the training phase, the algorithm calculates the probabilities of occurrence for each feature (i.e., hazard characteristic) given a particular hazard class. These probabilities are then used to classify new instances of hazards based on their features. Additionally, risk severity assessment is integrated into the classification process, where the severity of risks associated with each classified hazard is evaluated. This methodology enables the proposed system to provide comprehensive hazard identification and risk assessment capabilities, empowering construction professionals to proactively address safety concerns and enhance overall safety practices on construction sites. He is a complete methodology for the research:

3.1 Data Collection

As part of this research, data was collected from a construction company known as China Civil Engineering Construction Corporation (CCECC) in Uyo, Nigeria. The dataset included detailed site analysis reports and associated information about possible hazards and risk severity encountered in various construction projects. This data served as the foundation for building predictive models aimed at enhancing safety and risk management in the construction industry.

3.2 Feature Selection/Engineering

Feature engineering involves transforming raw data into meaningful features that better represent the underlying problem for the predictive models. In this work, feature engineering included text cleaning, normalization, stemming, and the creation of new features. Text cleaning involves removing unnecessary characters, punctuation, and stop words from the site analysis text to ensure the data is clean and uniform. Normalization converted all text to lowercase to maintain consistency, while lemmatization/stemming reduced words to their base or root form to standardize variations of words (e.g., "running" to "run"). Additionally, generating new features that might capture relevant information, such as the length of the site analysis text or the frequency of specific terms related to hazards, further enriched the dataset.

3.3 Feature Extraction

Feature extraction refers to the process of transforming raw text data into numerical representations that machine learning models can process. In this study, the TF-IDF (Term Frequency-Inverse Document Frequency) vectorizer was used for feature extraction. TF-IDF captures the importance of words in a document relative to a collection of documents, assigning higher weights to words that are frequent in a document but rare across all documents, thus highlighting significant terms. This process involved tokenization (splitting the site analysis text into individual words or tokens), calculating term frequencies (TF), determining inverse document frequencies (IDF), and computing TF-IDF scores.

3.4 Categorical Data Encoding and Tokenization

Tokenization is the process of breaking down text into smaller units, such as words or phrases, called tokens. For this research, word tokenization was employed, splitting the site analysis texts into individual words. This step is crucial as it converts text into a format that can be analyzed and processed by machine learning algorithms. One-hot encoding is a method of converting categorical variables into binary vectors. Each category is represented as a vector with all elements set to 0 except for one element set to 1, indicating the presence of that category. In this study, one-hot encoding was applied to categorical features such as possible hazards. For example, if the possible hazards were categorized as "fall," "electrical," and "chemical," each hazard would be represented as a binary vector [1, 0, 0], [0, 1, 0], and [0, 0, 1] respectively. Model formation involved training machine learning models to predict hazards and risk severity based on the processed data. Multinomial Naive Bayes, a probabilistic classifier suitable for discrete data [23], particularly for text classification tasks, was used for hazard prediction. It calculates the probability of each possible hazard given the site analysis text and selects the

hazard with the highest probability. The site analysis texts were transformed using the TF-IDF vectorizer, and the resulting numerical features were used to train the Naive Bayes model on the labeled hazard data. For risk severity prediction, Linear Regression was employed. This regression analysis method models the relationship between a dependent variable (risk severity) and one or more independent variables (features extracted from site analysis). The transformed site analysis texts were used to predict the continuous values of risk severity. The model learned the weights of the features that best fit the training data, minimizing the error between predicted and actual risk severity values. The study developed robust models to predict hazards and assess risk severity, aiding construction companies like CCECC in making informed safety decisions and improving their risk management processes. Figure 2 is the analysis of the proposed system. Table1 is a Snippet of Dataset, Figure 2 is the framework of the proposed hazard detection and severity analysis system.

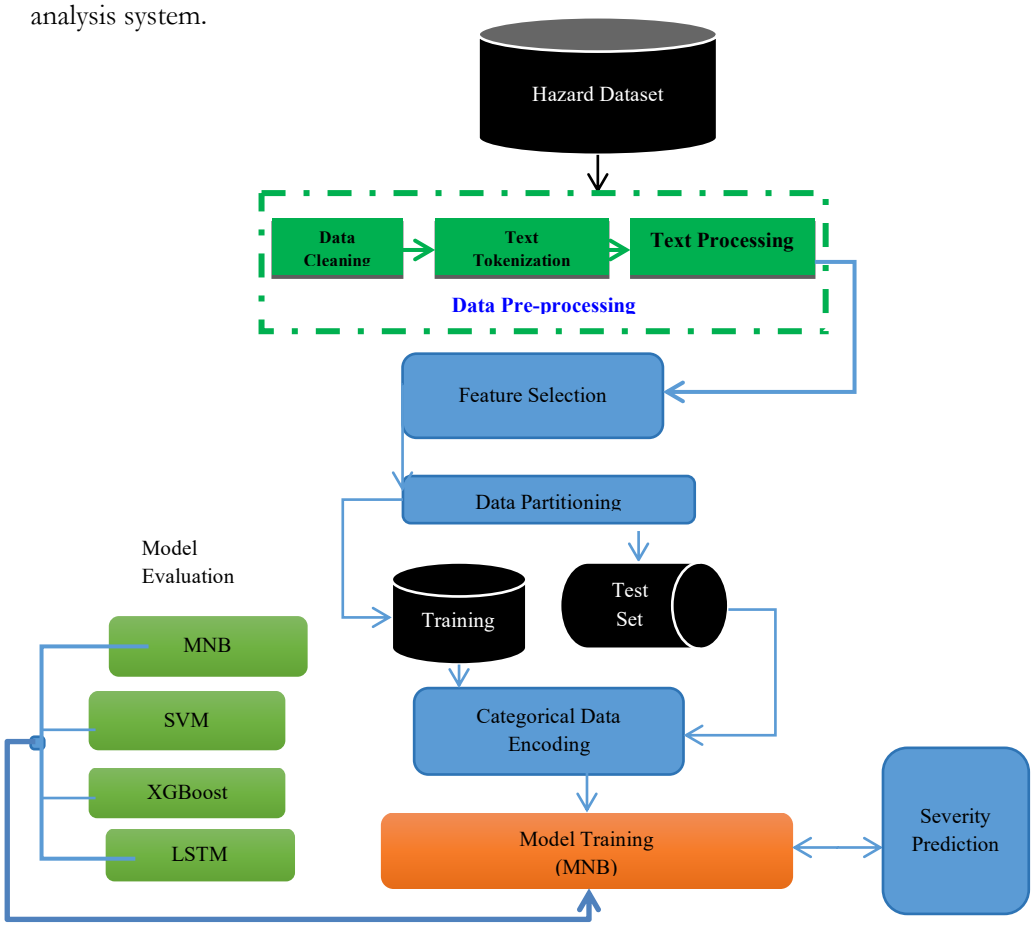

Figure 2. Architecture of the Proposed System

In Figure 2, the proposed frame for hazard detection and severity assessment begins with the dataset of hazard as its first module. The preprocessing activities happen immediate the system receives data. Feature Selection and data partitioning is done to choose the needed feature and the size of dataset for both test and train sets. Since the algorithm used is a semantic based algorithm, the system receives text (site analysis text) as input, this condition aided the feature selection process, making site analysis an important feature in the dataset. The Multinomial Naïve Bayes algorithm performs one-hot encoding of categorical data, then model is built. The model built performs hazard detection and examines its severity. Table 1 is shown with all possible features used.

Table 1. Snippet of Dataset

Site analysis	Possible hazard	Risk severity
Working on roof without proper fall protection. Objects not secured on scaffolding.	Falls, Falling Objects	0.99
Materials falling from upper floors. Faulty electrical equipment.	Falling Objects, Electrical Hazards	0.95
Elevated work without proper fall protection. Inadequate safety measures in trench.	Falls, Excavation Hazards	0.9
Trench collapse during excavation work. Struck by falling equipment during lifting.	Excavation Hazards, Struck-By Hazards	0.85
Unsafe signaling around moving equipment. Struck by overhead crane load.	Struck-By Hazards, Caught-In or Between Hazards	0.8
Inadequate lighting leading to slips and falls. Tripping on construction materials.	Slips, Trips, and Falls, Collapse of Structures	0.75
Icy surfaces leading to slips and falls. Unsafe scaffolding during construction.	Slips, Trips, and Falls, Hazardous Materials	0.7
Concrete formwork collapse during construction. Tools falling during equipment operation.	Collapse of Structures, Mechanical Hazards	0.65
Vibrating tool usage without proper precautions. Materials falling from elevated work areas.	Vibration Hazards, Falling Objects	0.6
Objects falling from scaffold above. Vehicle collisions on construction site.	Falling Objects, Vehicle and Traffic Hazards	0.55

Site analysis	Possible hazard	Risk severity
Working on roof without proper fall protection. Objects not secured on scaffolding.	Falls, Falling Objects	0.99
Materials falling from upper floors. Faulty electrical equipment.	Falling Objects, Electrical Hazards	0.95
Elevated work without proper fall protection. Inadequate safety measures in trench.	Falls, Excavation Hazards	0.9
Trench collapse during excavation work. Struck by falling equipment during lifting.	Excavation Hazards, Struck-By Hazards	0.85
Unsafe signaling around moving equipment. Struck by overhead crane load.	Struck-By Hazards, Caught-In or Between Hazards	0.8
Inadequate lighting leading to slips and falls. Tripping on construction materials.	Slips, Trips, and Falls, Collapse of Structures	0.75
Icy surfaces leading to slips and falls. Unsafe scaffolding during construction.	Slips, Trips, and Falls, Hazardous Materials	0.7
Concrete formwork collapse during construction. Tools falling during equipment operation.	Collapse of Structures, Mechanical Hazards	0.65
Vibrating tool usage without proper precautions. Materials falling from elevated work areas.	Vibration Hazards, Falling Objects	0.6
Objects falling from scaffold above. Vehicle collisions on construction site.	Falling Objects, Vehicle and Traffic Hazards	0.55
Working on roof without proper fall protection. Objects not secured on scaffolding.	Falls, Falling Objects	0.99
Materials falling from upper floors. Faulty electrical equipment.	Falling Objects, Electrical Hazards	0.95
Elevated work without proper fall protection. Inadequate safety measures in trench.	Falls, Excavation Hazards	0.9
Trench collapse during excavation work. Struck by falling equipment during lifting.	Excavation Hazards, Struck-By Hazards	0.85
Unsafe signaling around moving equipment. Struck by overhead crane load.	Struck-By Hazards, Caught-In or Between Hazards	0.8

Site analysis	Possible hazard	Risk severity
Inadequate lighting leading to slips and falls. Tripping on construction materials.	Slips, Trips, and Falls, Collapse of Structures	0.75
Icy surfaces leading to slips and falls. Unsafe scaffolding during construction.	Slips, Trips, and Falls, Hazardous Materials	0.7
Concrete formwork collapse during construction. Tools falling during equipment operation.	Collapse of Structures, Mechanical Hazards	0.65

4. RESULTS AND DISCUSSION

In this section of this research, we present a comprehensive analysis of the findings derived from our dataset on construction hazards. This section aims to interpret the results obtained from various analytical methods, including Multinomial Naive Bayes and regression analysis, highlighting their implications for risk assessment in the construction industry. We will explore the significance of the identified hazards, their prevalence, and how they contribute to overall safety risks on construction sites.

4.1. Experimental Performance

The dataset used in this research comprises a comprehensive collection of information related to various hazards encountered in the construction industry. It includes critical variables such as site analysis, hazard types, and severity ratings, allowing for a thorough examination of risk factors. The dataset captures a diverse range of hazards, including falls, falling objects, excavation risks, and more, providing a rich foundation for analysis. By encompassing multiple aspects of construction safety, this dataset facilitates a deeper understanding of the prevalent risks and their potential impacts on workers

In addition to its breadth, the dataset is characterized by its quantitative measurements, which support statistical analysis and model training. Each entry is carefully curated to ensure accuracy, with risk severity ratings indicating the potential danger associated with each hazard type. This structured approach enables the application of advanced analytical techniques, such as Multinomial Naive Bayes and regression analysis, to derive meaningful insights. Ultimately, the dataset serves as a vital tool in assessing risk severity and informing effective safety protocols in the construction sector, contributing significantly to the research outcomes. Figure 3 is the MNB Classification performance.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

	precision	recall	f1-score	support
Collapse of Structures, Mechanical Hazards	0.99	1.00	0.99	2028
Excavation Hazards, Struck-By Hazards	0.99	1.00	0.99	1950
Falling Objects, Electrical Hazards	0.99	1.00	0.99	2030
Falling Objects, Vehicle and Traffic Hazards	0.99	1.00	1.00	1997
Falls, Excavation Hazards	0.99	1.00	1.00	2018
Falls, Falling Objects	0.99	1.00	1.00	1990
Slips, Trips, and Falls, Collapse of Structures	0.99	1.00	0.99	1923
Slips, Trips, and Falls, Hazardous Materials	0.99	1.00	1.00	2065
Struck-By Hazards, Caught-In or Between Hazards	0.99	1.00	0.99	1988
Vibration Hazards, Falling Objects	0.99	1.00	0.99	2003
a flow flow	0.00	0.00	0.00	209
accuracy			0.99	20201
macro avg	0.90	0.91	0.90	20201
weighted avg	0.98	0.99	0.98	20201

Figure 3. MNB Classification Report

The classification report for the Multinomial Naive Bayes model reveals impressive performance metrics, showcasing an accuracy score of 99 in predicting risk categories. This high level of accuracy indicates that the model effectively identifies and classifies the different hazard types within the dataset. Furthermore, the precision score of 0.99 signifies how reliable the model is, as it correctly identifies nearly all actual positive cases with minimal false positives. The F1-score of 0.99 further underscores the model's robustness, balancing both precision and recall to ensure that it is not only accurate but also sensitive to the various hazards present.

Categorical Plot of Hazard in Figure 4(a) highlights the contributions of different hazard types to the model's accuracy, providing insight into which risks are most influential in the classification process. This information is invaluable, as it allows for targeted interventions in risk management, emphasizing the hazards that significantly affect safety outcomes. The outcomes of our Naive Bayes model in risk classification demonstrate its efficacy in providing a reliable and comprehensive analysis of hazards in the construction industry, paving the way for improved safety protocols and informed decision-making.

Journal of Information Systems and Informatics
Vol. 7, No. 1, March 2025
p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

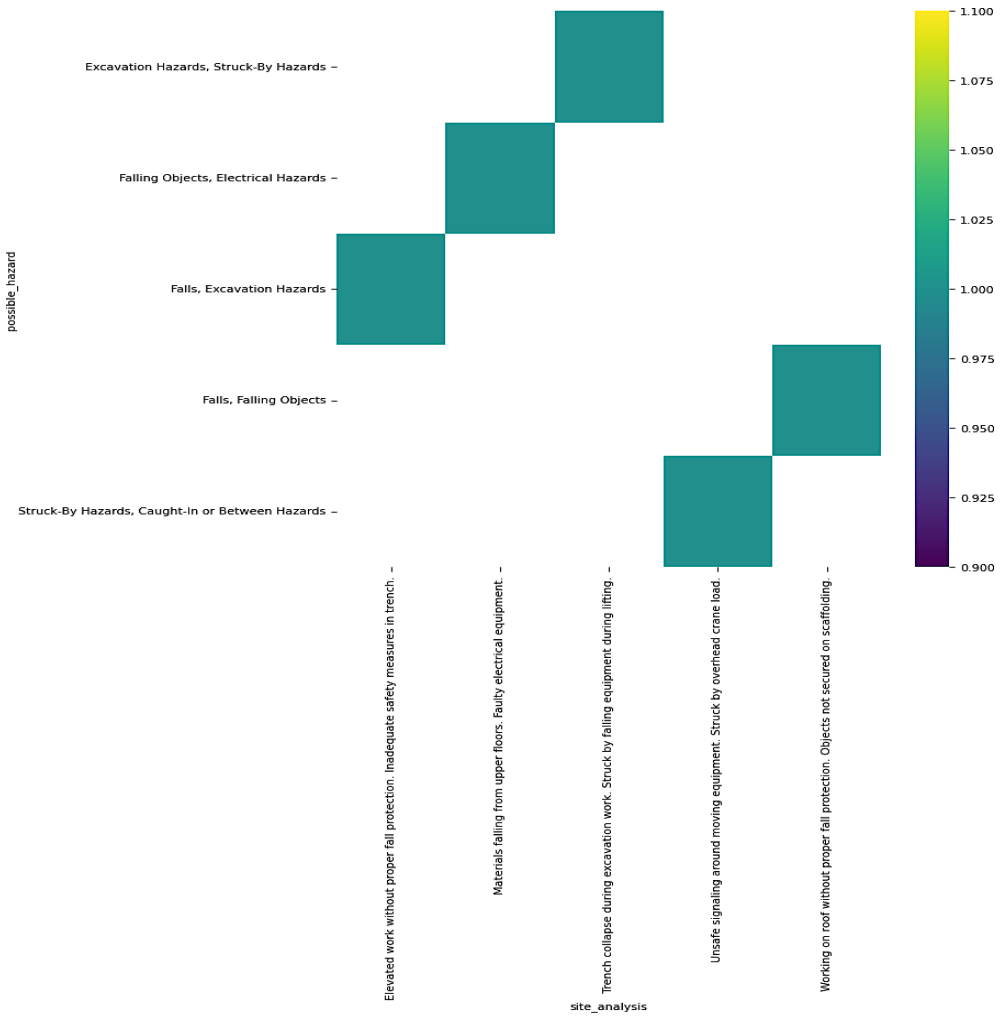

Figure 4(a). Categorical Plot of Hazard

This plot is the categorical feature plot that shows all the hazard type and how concentrated they are in the dataset. It also measures closeness and dependability of one feature or data point to another, allowing a more indepth analysis. The categorical feature plot provides a comprehensive overview of the various hazard types present in the dataset, illustrating their concentration and distribution. By visualizing how these hazards are categorized in this research, the plot enables us to assess the proximity and interdependence of different features or data points within the dataset. This analysis allows for a deeper understanding of the relationships between various hazards, highlighting which types may be more prevalent and potentially interconnected. Ultimately, this insight is crucial for

developing targeted strategies for risk management and safety improvements in the construction industry. Fig. 4(b) is a Pie Chart showing the Distribution of each Hazard

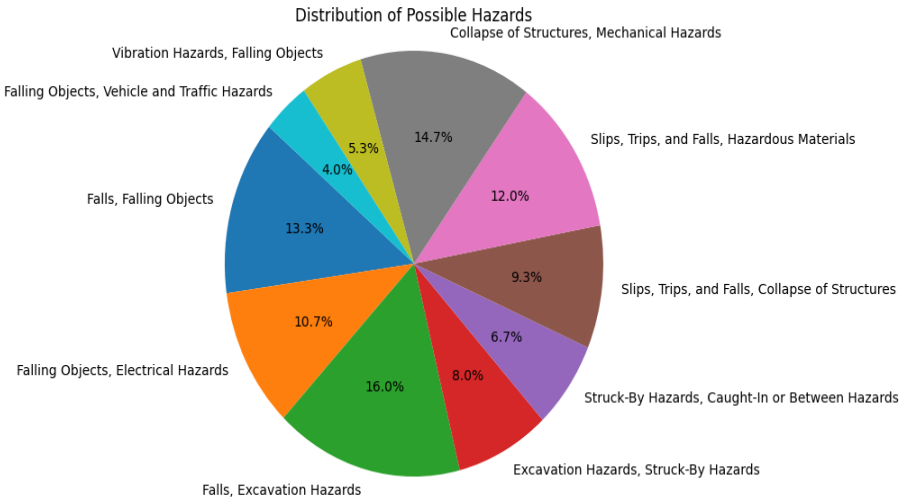

Figure 4(b). Distribution of each Hazard

The chart illustrates the distribution and contribution of each hazard type within the dataset, presented through a pie chart that clearly delineates the percentage composition of various hazards. The data reveals that falls, falling objects, and excavation hazards account for the highest percentages, indicating their prevalence on construction sites. This finding indicates the critical importance of addressing these specific hazards in safety protocols and risk management strategies. By highlighting the most significant risks, our research emphasizes the need for targeted interventions to mitigate these dangers, ultimately promoting a safer working environment in the construction industry.

The confusion matrix in Figure 5 provides a clear insight into how effectively the model understands and detects hazard patterns within the dataset. By comparing predicted classifications against actual outcomes, the confusion matrix enables the system to evaluate its performance and calculate the accuracy score. In this model, we achieved an impressive accuracy score of 0.99, indicating that the model is highly proficient in correctly identifying hazards. This high level of accuracy not only reflects the model's capability in hazard detection but also reinforces its reliability in informing safety measures within the construction industry. The Graphical user interface of the develop application is a s presented in Figure 6.

Journal of Information Systems and Informatics
Vol. 7, No. 1, March 2025
p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

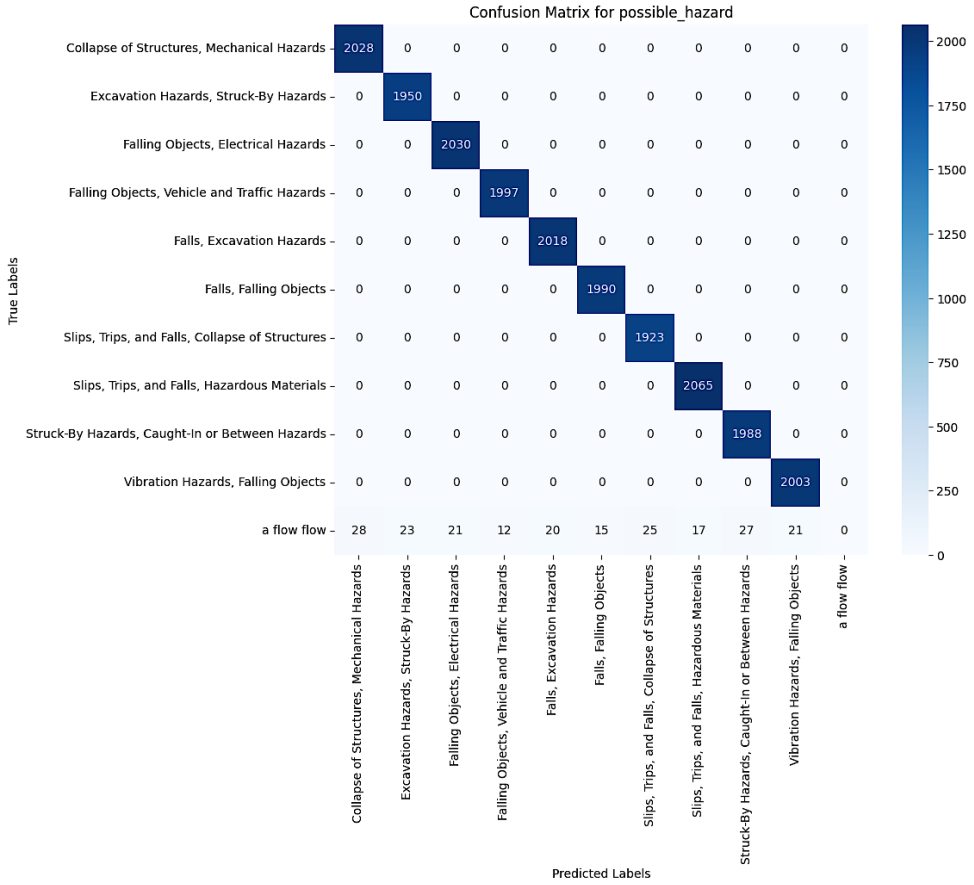

Figure 5. Confusion Matrix for Possible Hazard

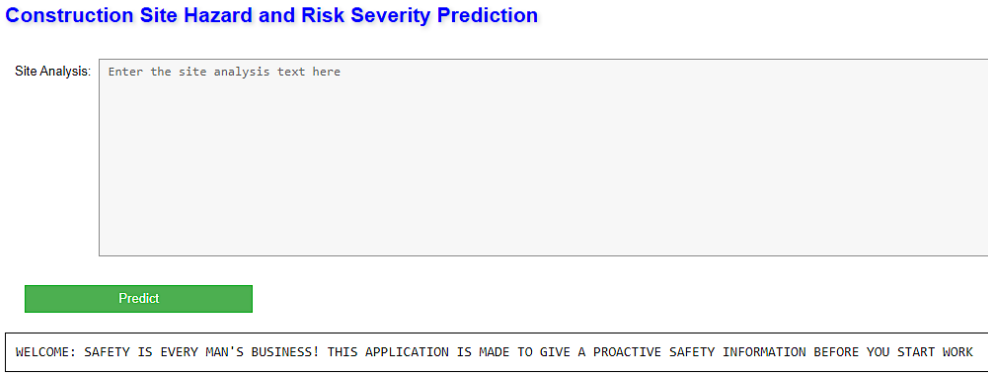

Figure 6. Graphical User Interface (GUI)

The graphical user interface (GUI) created for this research provides users with an intuitive and visually appealing platform to input site analysis text and receive predictive insights simultaneously. Designed with user-friendly elements such as a large, embossed text box with a shadow effect for inputting the site analysis, the interface ensures ease of use and readability. Upon entering the text, users can simply click the "Predict" button to obtain predictions for both possible hazards and risk severity concurrently, displayed clearly in separate output sections. This dual-functionality interface not only enhances user experience but also streamlines the process of obtaining critical safety assessments, making it a valuable tool for construction site managers and safety officers to quickly and accurately evaluate risks and take necessary preventive measures.

4.2. Discussion

The findings of this study highlight the significant potential of machine learning models in enhancing hazard detection and classification within the construction industry. The high performance of the Multinomial Naive Bayes (MNB) model, demonstrated by its 99% accuracy, suggests that predictive algorithms can serve as valuable tools for identifying potential risks with minimal human intervention. This is a crucial advancement, given that traditional safety assessments often rely on manual inspections and subjective judgment, which can be prone to human error. The ability to automate and standardize hazard classification allows for consistent and data-driven decision-making, reducing the likelihood of overlooked risks and improving overall safety management.

Despite the impressive model accuracy, it is important to consider the limitations and potential biases within the dataset. The dataset used in this study is well-structured, containing diverse hazard types such as falls, falling objects, and excavation risks. However, the model's high performance could be partially influenced by class imbalance, where certain hazard types may be overrepresented. If the dataset is dominated by a few major hazard categories, the model might perform exceptionally well on those while struggling with underrepresented hazards. Future studies should explore techniques such as data augmentation, resampling, or weighting strategies to ensure balanced learning across all hazard types.

The categorical feature plot and hazard distribution analysis provide further insight into how different risks contribute to the model's predictions. The concentration of certain hazards, such as falls and excavation risks, aligns with well-documented industry trends where these hazards are among the leading causes of accidents in construction. This suggests that the model is effectively capturing real-world risk patterns. However, the visualization also raises questions about potential correlations between different hazards. For example, excavation hazards may often

co-occur with structural instability risks, and falling objects might be associated with inadequate protective measures. Exploring these relationships through advanced statistical analysis or feature engineering could enhance the model's predictive power by incorporating contextual dependencies between hazards.

The confusion matrix analysis further validates the reliability of the MNB model, showing that it can accurately distinguish between different hazard types. However, while a 99% accuracy rate is impressive, it is essential to critically examine the few misclassifications that do occur. Are there specific hazard types that the model struggles to differentiate? Misclassification in a safety-critical context can have serious consequences, particularly if high-risk hazards are incorrectly labeled as low-risk. Analyzing misclassified cases and introducing misclassification cost functions in future iterations of the model could help mitigate these risks, ensuring that errors in hazard classification do not lead to safety oversights.

One of the most practical contributions of this research is the development of a graphical user interface (GUI) for real-time hazard classification. The ability to input site analysis text and receive immediate predictions provides an intuitive and accessible solution for construction site managers and safety officers. However, the success of this tool in real-world applications depends on user adoption and integration into existing workflows. Future improvements could involve real-time hazard monitoring through IoT sensors, voice recognition for verbal site reports, or integration with wearable safety devices to enhance usability and practicality in dynamic construction environments.

The implementation of machine learning in construction safety represents a shift toward proactive risk management, where hazards can be identified before accidents occur rather than after-the-fact analysis. While this study demonstrates the effectiveness of supervised learning techniques, future research could explore unsupervised or semi-supervised approaches to detect emerging hazards that may not be explicitly labeled in the dataset. Additionally, the inclusion of real-time environmental data (e.g., weather conditions, equipment usage, worker fatigue levels) could refine the model's predictions, making it more adaptive to changing site conditions.

Another key area for improvement is the interpretability and explainability of AI models in safety-critical industries. While Naive Bayes offers transparency in probability-based decision-making, construction site managers may require additional insights beyond simple predictions. How does the model determine risk severity? Which features contribute most to a given hazard classification? Incorporating explainable AI (XAI) techniques, such as SHAP (SHapley Additive Explanations) values, could provide clearer justifications for predictions,

increasing trust and adoption of AI-based safety tools. Finally, policy implications must be considered. The use of AI-driven hazard classification systems could inform regulatory frameworks and safety standards, leading to data-driven safety compliance measures. Organizations could implement automated hazard monitoring as part of mandatory site assessments, ensuring that predictive models complement and enhance human oversight rather than replace it.

This study demonstrates that machine learning models can effectively classify construction hazards with high accuracy, providing valuable insights for safety management. However, for long-term impact, it is crucial to address data limitations, misclassification risks, and real-world integration challenges. Future research should explore multi-modal data sources, real-time hazard detection, and explainable AI techniques to refine and expand the applicability of AI-driven safety solutions. By leveraging these advancements, the construction industry can transition toward a smarter, more proactive approach to workplace safety, reducing accidents and saving lives.

5. CONCLUSION

This research represents a significant stride towards enhancing safety practices within the construction industry by using machine learning techniques for hazard detection and risk severity assessment. Through the comprehensive analysis of a large dataset sourced from real-world construction operations, the study has elucidated the diverse array of hazards prevalent in construction sites, ranging from falls and structural collapses to electrical hazards and equipment-related risks. By harnessing advanced algorithms such as Multinomial Naive Bayes and Linear Regression, the research has demonstrated the efficacy of predictive models in accurately identifying potential hazards and quantifying their severity. Furthermore, the findings of this research reveal the paramount importance of proactive risk management strategies within the construction sector. By preemptively identifying and assessing hazards, construction stakeholders from CCECC can implement targeted safety measures to mitigate risks and safeguard the well-being of personnel and infrastructure by making use of the software tool created in this research. The predictive capabilities of the developed models empower construction site managers and workers to make informed decisions regarding safety protocols, resource allocation, and project planning, thereby fostering a culture of safety and accountability. Moreover, the research highlights the imperative of continuous monitoring and adaptation in response to evolving hazards and site conditions. As construction projects progress and environments change, the dynamic nature of risk necessitates ongoing vigilance and adjustment of safety protocols. By integrating machine learning-based hazard detection and risk severity assessment into routine safety practices, construction organizations can cultivate a proactive approach to risk mitigation, thereby enhancing overall

safety performance and reducing the incidence of workplace accidents and injuries. Ultimately, this research reveals the overarching principle that no risk within the construction industry is too insignificant to be overlooked. By acknowledging and addressing even seemingly minor hazards, construction stakeholders can preemptively avert potentially catastrophic consequences, thereby fostering a culture of safety, resilience, and excellence within the construction sector. As the industry continues to evolve, the insights gleaned from this research can serve as a foundation for advancing safety protocols, driving continuous improvement, and ultimately ensuring the well-being of all individuals engaged in construction activities.

ACKNOWLEDGEMENTS

We would like to extend our sincere gratitude to management of CCECC for their valuable support in providing data for this research. Their assistance has been instrumental in enhancing the depth and accuracy of our research, and we are appreciative of their commitment to sharing insights that contribute to a better understanding of risk and hazard.

REFERENCES

[1] V. S. K. Delhi, R. Sankarlal, and A. Thomas, "Detection of personal protective equipment (PPE) compliance on construction site using computer vision-based deep learning techniques," *Front. Built Environ.*, vol. 6, p. 136, 2020.

[2] A. Ekong, H. Odikwa, and O. Ekong, "Minimizing symptom-based diagnostic errors using weighted input variables and fuzzy logic rules in clinical decision support systems," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 10, no. 3, pp. 1567–1575, 2021.

[3] A. Ekong, B. Ekong, and A. Edet, "Supervised machine learning model for effective classification of patients with COVID-19 symptoms based on Bayesian belief network," *Res. J. Sci. Technol.*, vol. 2, no. 2, pp. 27–33, 2022.

[4] J. Dumaru, S. Karki, R. Shrestha, and S. S. Khadka, "A review on machine learning approaches for Himalayan rock mass classification," *J. Phys. Conf. Ser.*, vol. 2629, p. 012016, 2023, doi: 10.1088/1742-6596/2629/1/012016.

[5] S. Ahmed, M. A. Hossain, M. M. I. Bhuiyan, and S. K. Ray, "A comparative study of machine learning algorithms to predict road accident severity," in *Proc. 20th Int. Conf. Ubiquitous Comput. Commun. (IUCC/CIT/DSCI/SmartCNS)*, London, U.K., 2021, pp. 390–397.

[6] C. Liao, E. Aminudin, S. Mohd, and S. S. Yap, "Intelligent risk management in construction projects: Systematic literature review," *IEEE Access*, vol. 10, pp. 72936–72954, 2022.

[7] J. Lee and S. Lee, "Construction site safety management: A computer vision and deep learning approach," *Sensors*, vol. 23, p. 944, 2023, doi: 10.3390/s23020944.
[8] N. Udoetor, G. Ansa, A. Ekong, and A. Edet, "Intelligent system for detection of copyright-protected data for enhanced data security," *Br. J. Comput. Netw. Inf. Technol. (BJCNIT)*, vol. 7, no. 4, pp. 58–80, 2024.
[9] M. Alateeq and M. Ali, "Construction site hazards identification using deep learning and computer vision," *Sustainability*, vol. 15, p. 2358, 2023, doi: 10.3390/su15032358.
[10] A. Edet and G. Ansa, "Machine learning enabled system for intelligent classification of host-based intrusion severity," *Glob. J. Eng. Technol. Adv.*, vol. 16, no. 3, pp. 041–050, 2023.
[11] Z. Wei et al., "Insights into the application of machine learning in industrial risk assessment: A bibliometric mapping analysis," *Sustainability*, vol. 15, p. 6965, 2023, doi: 10.3390/su15086965.
[12] M. Al-Ghzawi and K. El-Rayes, "Machine learning for predicting the impact of construction activities on air traffic operations during airport expansion projects," *Autom. Constr.*, vol. 158, p. 105189, 2024, doi: 10.1016/j.autcon.2023.105189.
[13] M. R. George, M. R. Nalluri, and K. B. Anand, "Application of ensemble machine learning for construction safety risk assessment," *J. Inst. Eng. (India): Ser. A*, vol. 103, pp. 989–1003, 2022, doi: 10.1007/s40030-022-00690-w.
[14] S. Yin, T. Liu, and B. Huang, "Research on security assessment model of marine information system based on machine learning," in *Proc. 2023 Int. Conf. Marine Equip. Technol. Sustainable Dev. (METSD 2023)*, D. Yang, Ed., 2023, pp. 375–386, doi: 10.1007/978-981-99-4291-6_11.
[15] B. A. Almahameed and M. Bisharah, "Applying machine learning and particle swarm optimization for predictive modeling and cost optimization in construction project management," *Asian J. Civil Eng.*, vol. 17, no. 1, pp. 1–15, 2023, doi: 10.1007/s42107-023-00843-7.
[16] S. Elhishi, A. M. Elashry, and S. El-Metwally, "Unboxing machine learning models for concrete strength prediction using XAI," *Sci. Rep.*, vol. 13, p. 19892, 2023, doi: 10.1038/s41598-023-47169-7.
[17] B. T. Pham, C. Vuong, and K. N. Ngo, "Damage assessment in beam-like structures by correlation of spectrum using machine learning," *Frattura Integr. Strutt.*, vol. 17, no. 65, pp. 300–319, 2023, doi: 10.3221/IGF-ESIS.65.20.
[18] O. Lialiuk and R. Osypenko, "Features of the implementation of artificial intelligence in construction," *J. Mod. Technol. Mater. Des. Constr.*, vol. 35, no. 2, pp. 172–176, 2023.

[19] F. Mostofi and V. Toğan, "Explainable safety risk management in construction with unsupervised learning," in *Artificial Intelligence and Machine Learning Techniques for Civil Engineering*, V. Plevris, A. Ahmad, and N. Lagaros, Eds., IGI Global, 2023, pp. 273–305, doi: 10.4018/978-1-6684-5643-9.ch011.

[20] A. Le Duc et al., "A machine learning approach to risk assessment of expressway bridges," *Tạp Chí Khoa Học Giao Thông Vận Tải*, vol. 73, no. 7, pp. 661–673, 2023, doi: 10.47869/tcsj.73.7.1.

[21] S. Sarkar et al., "Predicting and analyzing injury severity: A machine learning-based approach using class-imbalanced proactive and reactive data," *Safety Sci.*, vol. 125, p. 104616, 2020, doi: 10.1016/j.ssci.2020.104616.

[22] S. Ballal, K. A. Patel, and D. A. Patel, "Enhancing construction site safety: Natural language processing for hazard identification and prevention," *J. Eng. Project Prod. Manag.*, vol. 14, no. 2, p. 004, 2024.

[23] B. Ekong et al., "Machine learning approach for classification of sickle cell anemia in teenagers based on Bayesian network," *J. Inf. Syst. Inform.*, vol. 5, no. 4, pp. 1793–1808, 2023.

[24] A. Ekong, "Evaluation of machine learning techniques towards early detection of cardiovascular diseases," *Am. J. Artif. Intell.*, vol. 7, no. 1, pp. 6–16, 2023, doi: 10.11648/j.ajai.20230701.12.

[25] P. Ekong, G. G. James, and I. Ohaeri, "Oil and gas pipeline leakage detection using IoT and deep learning algorithm," *J. Inf. Syst. Inform.*, vol. 6, no. 1, pp. 421–434, 2024, doi: 10.51519/journalisi.v6i1.652.