Evaluation and Comparison of the Performance of Artificial Intelligence Algorithms in Predicting Construction Safety Incidents

F. Alsakka[a], Y. Mohamed[a], and M. Al-Hussein[a]

[a]Department of Civil and Environmental Engineering, University of Alberta, Canada
E-mail: falsakka@ualberta.ca, Yasser.Mohamed@ualberta.ca, malhussein@ualberta.ca

Abstract

Predicting the outcomes of safety incidents on construction projects is of a great value to various project stakeholders. Accurate estimates allow construction managers to take appropriate preventive measures based on the severity of the outcomes. Such estimates can be predicted using machine learning algorithms, although the quality of these estimates is dictated by factors including the types of algorithms employed and the dataset used to train them. Moreover, the metrics used to evaluate the algorithms can be misleading, indicating satisfactory performance when this may not be the case. In light of these considerations, this study trains a set of machine learning algorithms to predict the severity of safety incidents, highlighting the importance of confirming the credibility of performance evaluation results, and compares the performance of the algorithms. The results show that the support vector machine and k-nearest neighbors prediction models exhibit the best overall performance, with support vector machine achieving a mean absolute percentage error value of 18.78% and k-nearest neighbors an accuracy of 64.84%. On the other hand, the results also reveal that the models performed poorly in predicting some classes as a result of a high degree of imbalance identified in the dataset used for training and testing the models. The study's main contribution is to highlight the possibility of making biased performance evaluations of machine learning algorithms, depending on the performance measures used for the evaluation.

Keywords –

Construction, Safety, Machine Learning Algorithms, Prediction, Performance, Data Imbalance

1 Introduction

Given their hazardous nature, construction projects involve numerous high-risk activities during which safety incidents of high frequency and impact may occur. In Canada, more than 450 deaths and 63,000 injuries were recorded over a period spanning the period 2006 to 2017 [1]. High-risk activities can result in injuries, fatalities, schedule delays, and financial losses. In fact, healthcare expenditures in response to construction safety incidents in Canada amount to about 19.8 billion dollars each year [1]. The growing costs of incidents and the increased related pressure imposed by owners have increased contractors' awareness of the significance of safety risks [2]. Accordingly, there have been major efforts made and strict regulations introduced to control and minimize the risks associated with safety incidents.

Knowing the possible outcomes of safety incidents could help in mitigating the risk of their occurrence by assisting decision makers in taking the necessary preventive measures. With the large volumes of data being collected thanks to modern technology, artificial intelligence algorithms can be used for predicting the outcomes of incidents. In fact, multiple studies have developed models to predict and classify different aspects of construction incidents [3-12]. However, different algorithms can vary in performance depending on the dataset used for the training, and, hence, the algorithm to be used must be carefully selected. Moreover, evaluating the performance of the algorithms is a critical step that must be handled with special care. The metrics used to evaluate the algorithms could lead to an errant indication of satisfactory performance. The risk of this occurring is especially high when dealing with poor quality, insufficient, or complex data. This issue is manifest in a recent study by Ayhan & Tokdemir [3] aimed at predicting the severity of safety incidents on construction projects. The dataset used in their study is characterized by a high level of

heterogeneity and could thus be considered highly complex. The authors attempted to address the issue of heterogeneity by clustering the data prior to training the algorithms. Nevertheless, their results still showed high values of error, with the mean absolute percentage error reaching 62.3% in the case of the algorithm that exhibited the best performance. Despite this high error value, their study relied on an overall error of 18%—which was biased due to the high degree of imbalance found in the dataset—to judge the performance of their algorithms.

In this context, the present study aims to train a set of machine learning algorithms using the same dataset used by Ayhan & Tokdemir [3] in order to predict the severity of safety incidents on construction projects while targeting the following objectives: (1) highlight the importance of properly evaluating the performance of algorithms; (2) compare the performance of various algorithms when trained using the same dataset; and (3) find the best performing algorithm for predicting the severity of safety incidents.

2 Relevant Applications

Many studies have endeavoured to train machine learning algorithms for applications related to safety incidents on construction projects. These algorithms have been deployed for a wide variety of applications related to enforcing safety measures, predicting safety risks, identifying causal factors, and detecting hazards, to name a few. A summary of some of these applications is presented in Table 1.

Table 1. Literature summary

Study	Goal	Algorithms
[3]	Predict the severity level of safety incidents	Artificial Neural Network
		Case-Based Reasoning
[4]	Predict safety climate (i.e., employees' perceptions of existing safety practices) on construction projects	Artificial Neural Network
[5]	Assess construction workers' risk perceptions in terms of the probability and severity of the consequences of safety hazards	Gaussian Support Vector Machine
		K-Nearest Neighbor
		Decision Tree
		Bagging Tree
[6,7]	Detect safety helmet wearing on construction sites	Deep learning
		Convolutional Neural Network
[8]	Analyse site fall accidents in order to identify related causal factors, classify the factors, and identify the correlation between the type of accident and the causal factor(s)	Text mining
[9]	Predict safety risk factors on construction projects	Back Propagation Neural Network
[10]	Detect safety hazard issues based on the project's schedule and the sounds generated by work activities and equipment operating on construction sites	K-Nearest Neighbor
[11]	Detect fires on construction sites in real-time	Convolutional Neural Network
[12]	Identify construction safety hazards	Case-Based Reasoning

3 A Brief Overview of the Previous Study

The study conducted by Ayhan & Tokdemir [3] aimed at predicting the severity of safety incidents that occur on construction projects. The prediction outcomes subsume six levels of severity including, from low to high: (1) Level 1: At risk behavior/near miss; (2) Level 2: Accident with material damage; (3) Level 3: Accident with first aid; (4) Level 4: Partial failure/accident with medical intervention; (5) Level 5: Lost workday cases; and (6) Level 6: Fatalities. The dataset used by Ayhan & Tokdemir [3] covers 5,224 cases of actual incidents that occurred on megaprojects in the Euro-Asia region. The data provides information on the severity level of each incident and a corresponding list of 60 attributes classified into nine categories, subsuming the time of the day, age, experience, occupation, activities, hazardous cases, risky behaviours, human factors, and workplace factors. The time of the day is binned into eight 3-hour intervals starting from 6:00 a.m., age is binned into four categories (18–25 years, 25–35 years, 35–45 years, and 45–65 years), experience is binned into six categories (1 month, 1–3 months, 3–6 months, 6–12 months, 12–24 months, and 24 months), and occupation could be any of seven specified positions including administrative affairs, construction equipment operator, repairman, rough work crew, finishing work crew, mechanical assembly crew, engineer or, otherwise,

labelled as others. Meanwhile, the attributes belonging to the remaining categories are modelled as binary variables that take the value of "1" if the factor is found in the incident case and "0" if not.

To address the issue of the high level of heterogeneity found in the collected dataset, the authors clustered the data using Latent Class Clustering Analysis (LCCA). The authors used Artificial Neural Networks (ANN) and Case-Based Reasoning (CBR) to predict the severity of the outcome of an incident.

4 Methodology

RapidMiner software was employed in this study to train and test the machine learning models. Figure 1 depicts a sample of the machine learning process built using RapidMiner software, and the steps followed to design the process are described in the following subsections.

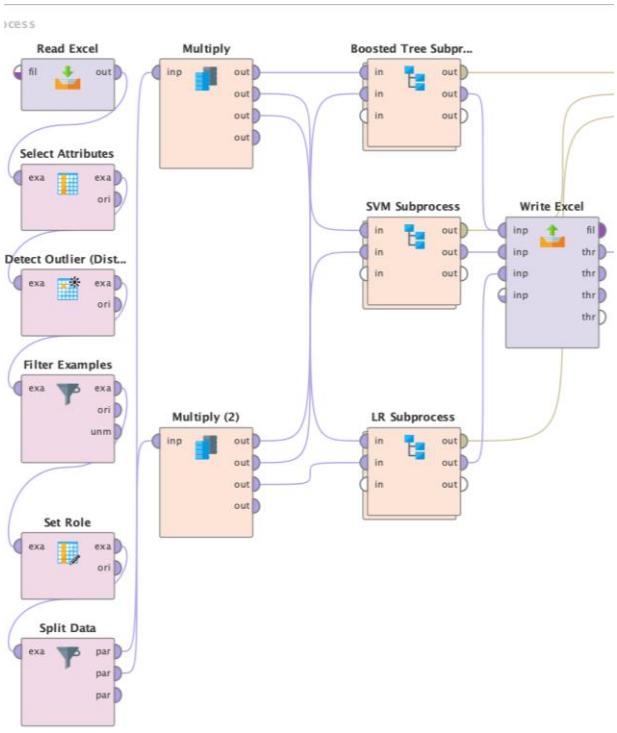

Figure 1. Machine learning process

4.1 Data Pre-processing & Exploration

4.1.1 Type of Prediction Problem

The type of class to be predicted (i.e., incident severity) is nominal, as it only assumes the value of one of the six severity levels in the provided dataset. However, the numbers corresponding to these levels (i.e., 1 to 6) are ordinal, since the severity of an incident is higher for higher levels. In reality, data on safety incidents could be considered continuous since the data could fall between the specified categories. For instance, an incident that results in a minor injury that does not necessitate first aid (e.g., a scratch) is less severe than Level 2 incidents but more severe than incidents that only involve material damage. As such, a decimal number falling between any two levels is helpful in terms of accurately representing reality. Hence, one set of machine learning algorithms were trained to estimate a numerical value of the severity level, and another set was trained to classify the outcome of the incidents. However, it should be noted that, in the interest of simplicity, the difference between the severity levels is assumed to be linear. In the study by Ayhan & Tokdemir [3], the class was treated as a numerical value where the severity level was estimated as a decimal number.

4.1.2 Data Cleansing

The dataset considered in this study was tidy and did not require significant cleansing. The data was evaluated to identify any missing attribute values. Only the experience attribute had missing values, and in only four of the 5,224 cases. Given the small proportion of missing values, these instances were simply excluded from the dataset. No additional errors or duplicates were identified.

4.1.3 Data Preparation

Typically, data is aggregated based on specific attributes in order to absorb variability and increase accuracy. As the age, experience, and time of day attributes have been already divided into bins, no further processing was deemed necessary at this stage.

It is integral to detect outliers in the data and remove them in order to minimize noise and, consequently, improve the accuracy of the prediction models. The process of removing outliers was undertaken in an iterative manner to avoid any detrimental effect on the performance of the prediction models as a result of removing core points. Density-based outlier detection was undertaken using Euclidean distances and ten neighbours. In this method, it should be noted, the neighbourhood of each datapoint is checked for the existence of ten neighbour datapoints. Accordingly, a point is considered an outlier if its neighbourhood does not contain enough datapoints. The number of outliers to be identified was initially set to ten points only. Then, this number was iteratively modified to maximize the accuracy of the prediction models. Two hundred outliers were ultimately identified and removed from the dataset.

It is critical to note that the data was highly imbalanced, meaning that some classes had a significantly higher frequency than others. As shown in

Figure 2, Level 3 severity had 3,070 observations, as compared to only four observations in the case of Level 6 severity. Such an imbalance in the dataset could result in building inaccurate models that exhibit satisfactory performance even when tested on a portion of the data that was not part of the training data. This is due to the fact that, if the model predicts that all the instances in the testing dataset belong to Level 3 class, its accuracy would still be high since the majority of the instances are of Level 3. The problem of imbalance is typically addressed by under-sampling or over-sampling the dataset [13]. Oversampling refers to the practice of duplicating instances from the minority classes to increase their cardinality, while under-sampling consists of taking subsets of the majority classes in order to reduce their frequency relative to that of the minority classes [13]. These techniques translate into replicating instances in Level 1, Level 2, Level 5, and Level 6 classes or taking subsets of Level 3 and Level 4 classes. Both of these strategies were tested when building the models.

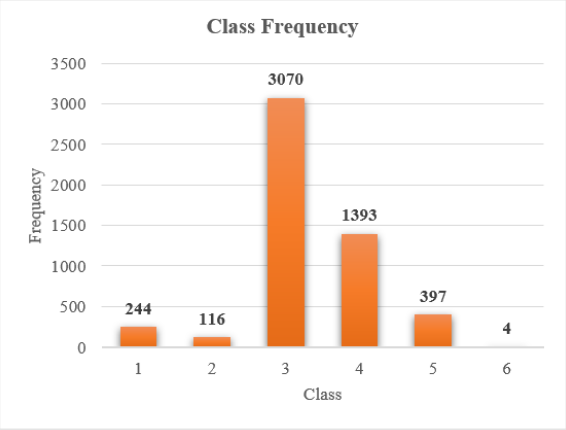

Figure 2. Class frequency

4.2 Machine Learning Models

4.2.1 Model Selection and Description

A trial-and-error approach was adopted to select the machine learning models exhibiting the best performance in predicting the severity level of incidents. The models selected for the numerical estimation of the severity level were Support Vector Machine (SVM), Linear Regression (LR), and Gradient Boosted Trees (GBT). For predicting the severity level as a nominal class, K-Nearest Neighbour (KNN), GBT, Random Forest (RF), and Generalized Linear Model (GLM) were selected.

A brief description of the selected models is summarized in Table 2 below.

Table 2. Description of models [14]

Model	Description
SVM	SVM is a non-probabilistic binary linear classifier that takes input data and forecasts which of two possible classes contains the input.
LR	LR models the relationship between a scalar variable and one or more explanatory variables by fitting a linear equation to the labelled training data.
GBT	GBT is an ensemble of classification tree models or regression models. It predicts classes through estimations that are gradually improved.
KNN	KNN compares a new example with k examples from the training dataset that are the nearest neighbours to the new example.
RF	RF is an ensemble of random trees. When given new examples, each random tree predicts the label of the input by following the corresponding branches. Class predictions are based on the majority of the trees' predictions, while estimations are the average of the trees' predictions.
GLM	GLM is a generalization of the traditional regression model that allows for the use of variables that are not normally distributed.

4.2.2 Models' Development and Validation

It is important to set aside a portion of the data to be used for testing purposes once the models have been trained. This helps mitigate the risk of overfitting, which is more likely to occur if the machine learning model is trained and tested using the same dataset. In specific, overfitting occurs when the model captures the noise in the training dataset and, consequently, fails to accurately predict new data [15]. Hence, the models were trained using 80% of the dataset, and their performance was evaluated using the performance measures explained in Section 4.2.3. The parameters of each model were continuously tuned to optimize their performance. Finally, the models were tested and validated using the remaining 20% of the dataset.

4.2.3 Models' Performance Evaluation

Ten-fold cross validation was used to train and evaluate the selected models. For the numerical estimation models, the mean absolute percentage error (MAPE) (1) and the root mean squared error (RMSE) (2) were used to evaluate performance.

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \times 100 \qquad (1)$$

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(F_t - A_t)^2}{n}} \quad (2)$$

where A_t is the actual instance, F_t is the predicted instance, and n is the total number of instances. As for the class prediction models, classification error (3) and accuracy (4) were used to evaluate their performance.

$$E = \frac{f}{n} \times 100 \quad (3)$$

$$\%\ Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \quad (4)$$

where f is the number of incorrectly classified classes, TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.

Recall (5) and precision (6) metrics were also computed in order to provide more context for understanding the accuracy evaluation metric. Computing recall values, it should be noted, addresses the following question: among the safety incidents that will actually occur, how many are we able to predict using our models? The recall measure is significant for the purpose of this study, as failing to anticipate incidents could result in financial and/or human losses. As for the precision metric, it answers the following question: among the incidents that are predicted to occur, how many will actually occur? When incidents are expected to occur, mitigation measures are undertaken accordingly to minimize the risk of occurrence. This translates into additional project costs associated with the safety measures. Therefore, the recall metric is deemed more critical than the precision metric in the context of this study, as it could be related to human losses.

$$\%\ Recall = \frac{TP}{TP + FN} \times 100 \quad (5)$$

$$\%\ Precision = \frac{TP}{TP + FP} \times 100 \quad (6)$$

5 Results and Discussion

5.1 Imbalanced Data Problem

The results of both techniques employed for addressing the problem of data imbalance (i.e., under-sampling and over-sampling) were found to be unsatisfactory. This was mainly due to the significant gap between the frequencies of different classes. Under-sampling resulted in poor performance of the models, largely attributable to the fact that the size of the dataset had to be significantly reduced in order to achieve balance, and thus it was not of a sufficient size to both train and test the models. On the contrary, over-sampling served to increase considerably the accuracy of the models. However, this accuracy is misleading, as it was largely the result of the testing data having contained instances of data points that were used in both testing and training of the models as a result of making duplicates. Over-sampling in the case of this dataset introduced the problem of overfitting. In light of this, the dataset was left as is, and the models were evaluated against each other to assess their overall performance and select the most optimal ones.

5.2 Performance Evaluation Results

The values of evaluation metrics computed for the different models are summarized in Table 3.

Table 3. Evaluation results

Numerical Estimation		
Model	MAPE	RMSE
SVM	18.29% +/− 2.68%	0.850 +/− 0.051
LR	20.83% +/− 2.31%	0.738 +/− 0.044
GBT	20.01% +/− 1.04%	0.732 +/− 0.029

Class Prediction		
Model	Classification Error	Accuracy
GBT	37.16% +/− 2.26%	62.84% +/− 2.26%
KNN	37.48 % +/− 1.65%	62.52% +/− 1.65%
RF	39.35% +/− 0.43%	60.65% +/− 0.43%
GLM	37.61% +/− 1.58%	62.39% +/− 1.58%

For the numerical estimation models, SVM was found to have the lowest MAPE and GBT the lowest RMSE, while LR exhibited poorer performance. As for the class prediction models, the classification error and the accuracy values for GBT, KNN, and GLM were found to be relatively close, while RF had higher error and lower accuracy. Moreover, the value of recall and precision metrics were found to be acceptable for all the models, the one exception being the RF model, in which the value of class' recall was 99.75% for Level 3 and 0.28% for Level 4, as the model predicted that most of the instances belong to class Level 3. This means that the RF model failed to predict Level 4 incidents, which include those necessitating medical intervention. Such values of the recall metric confirm the criticality of verifying the credibility of some performance evaluation measures (i.e., the accuracy metric in this case) to avoid misleading results. If the accuracy value was solely used to judge the performance of the algorithms, the RF model's performance would not have been considered significantly inferior to that of others.

Therefore, the LR and RF models were excluded at this stage, and the final selection among the remaining

models was based on the validation results as explained in the following section.

5.3 Validation Results

In conducting the validation, the trained models did not yield promising results, as the error reached 37.15%. However, similar results were recorded for all the models, mainly attributable to the low quality of the training dataset.

The validation results were used to select the best-performing models. As shown in Table 4, for the class prediction models, KNN was found to perform better than GBT and GLM, with a classification error of 35.16% and an accuracy of 64.84%. As for the numerical estimation models, the notion of an optimal model is contingent on the choice of the performance evaluation measure. In other words, SVM is the better option if MAPE is used for identifying the best model, while GBT is considered a better choice if RMSE is used. For the purpose of comparison with the previous study, MAPE was chosen as the decision-making criterion, and SVM was correspondingly selected.

Table 4. Validation results

Model	MAPE	RMSE
SVM	18.78% +/- 37.44%	0.844 +/− 0.000
GBT	20.28% +/- 35.89%	0.710 +/− 0.000
Model	Classification Error	Accuracy
GBT	37.15%	62.85%
KNN	35.16%	64.84%
GLM	36.35%	63.65%

5.4 Additional Assessment of the Models

A satisfactory value of MAPE computed for the whole dataset does not guarantee good performance of the models. This is especially critical given the high degree of imbalance found in the dataset. Hence, MAPE was also computed for each class separately to ensure that the large number of Level 3 instances is not skewing the results. This was done for both SVM and KNN; the results are plotted in Figure 3.

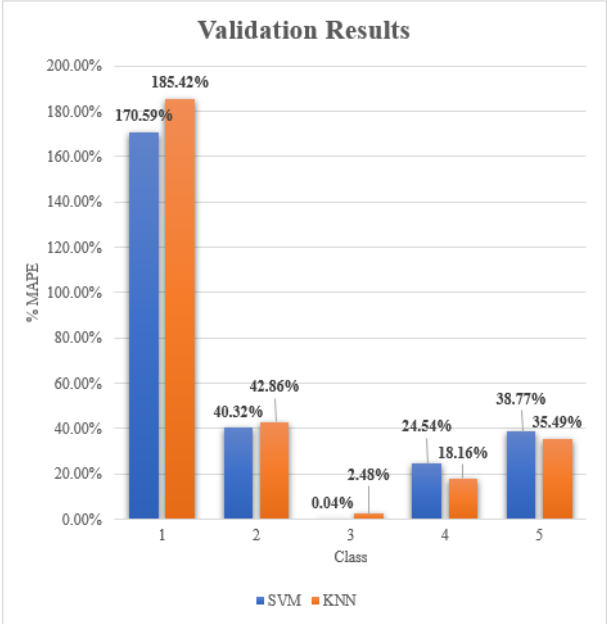

Figure 3. Validation results

The results show comparable performance for the two models. The error, however, was found to be significant for Level 1 predictions (170.59% for SVM and 185.42% for KNN), while that of the Level 3 predictions was found to be very low (0.04% for SVM and 2.48% for KNN). Meanwhile, the error was found to be reasonably acceptable for the other classes. It should be noted that the testing dataset did not include any Level 6 incidents resulting in null values for the Level 6 class prediction performance measures.

As anticipated, the reasonably acceptable performance of the models is a result of the misleadingly low error value obtained in predicting the Level 3 class instances. Given the high degree of imbalance in the dataset, it stands to reason that predicting that any new instance belongs to Level 3 class would give relatively acceptable results as compared to those obtained using the selected algorithms. In fact, the performance of the selected models was only slightly better than that of the Zero Rule classifier, as shown in Figure 4. Hence, the results obtained were deemed to be unsatisfactory. In other words, although the overall error of these models was found to be acceptable, this was due to imbalance skewing the Level 3 class predictions towards zero. As such, the developed models are not generally recommended for use, as they are not capable of reliably generalizing new datasets, given the low quality of the dataset used for training.

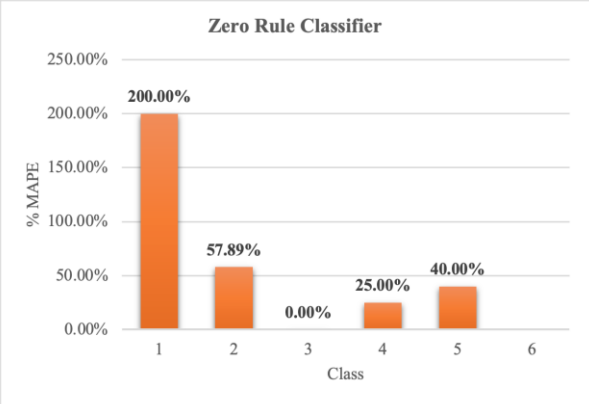

Figure 4. Zero Rule classifier performance results

6 Comparison

The error pattern found in the data predicted using SVM and KNN models matched that corresponding to the ANN and CBR models developed by Ayhan & Tokdemir [3], as shown in Table 5. The same issue identified in the SVM and KNN predictions was observed in the results generated by their models: the MAPE values were very high in the case of the Level 1 prediction, very low for Level 3 prediction, and relatively acceptable for the other classes. However, the variation between the MAPE values for different classes was found to be smaller in the case of CBR and ANN, and the highest error in the case of CBR was significantly lower (62.3%). This is presumably a result of the clustering analysis performed by Ayhan & Tokdemir [3].

Table 5. Summary of MAPE values

	Previous Study	
	ANN	CBR
Lowest MAPE (found for Level 3 predictions)	7.95%	8.66%
Highest MAPE (found for Level 1 predictions)	192.3%	62.3%

	Current Study	
	SVM	KNN
Lowest MAPE (found for Level 3 predictions)	0.04%	2.48%
Highest MAPE (found for Level 1 predictions)	170.59%	185.42%

Based on the results summarized in Table 5, it can be concluded that the clustering analysis performed by Ayhan & Tokdemir [3] did not play a significant role in improving the quality of the dataset in the case of ANN although their results were more favorable in the case of CBR. Despite the significant effort on the part of Ayhan & Tokdemir [3] to improve the dataset, the error patterns (i.e., low error for Level 3 predictions and high error for Level 1 predictions) were close to those obtained in the present study in which minor data preparation was performed prior to the models' development.

7 Conclusions

The SVM and KNN prediction models exhibited the highest performance among the various machine learning algorithms for predicting the severity level of safety incidents on construction projects. Nevertheless, although both algorithms yielded acceptable overall values of performance evaluation metrics (an overall MAPE of 18.78% for SVM and an accuracy of 64.84% for KNN), these values were not representative of the actual performance of the models. This was confirmed by computing the MAPE separately for each class, resulting in a value of 185.42% for KNN prediction of Level 1 class as compared to 2.48% for KNN prediction of Level 3 class. The high variation in MAPE values between the different classes is attributable to the high degree of imbalance found in the dataset (i.e., approximately 59% of its instances belong to the Level 3 class).

The results of this study reinforce the following points:

- The perception of the performance of machine learning algorithms could be highly biased depending on the metrics used for performance evaluation. For instance, if the final selection in this study had been solely based on the overall errors computed for validation purposes, the performance of the algorithms would have been considered relatively acceptable, whereas the actual results were unfavourable. A combination of different performance measures and validation techniques should be utilized to ensure that an unbiased decision is made.
- The quality of the training dataset could diminish the value of deploying some advanced machine learning algorithms and make the use of simpler classifiers, such as the Zero Rule classifier, more desirable.
- When the quality of the dataset is questionable, it is critical to perform multiple levels of performance evaluation to confirm the credibility of the evaluation results.

8 References

[1] SPI Health and Safety. Construction Workers: 3 or 4 Times More Accidents. On-line: https://www.spi-s.com/en/blog/ohs-leadership/construction-workers-3-or-4-times-more-accidents. Accessed: 07/07/2021.

[2] Mitropoulos P., Abdelhamid T.S., Howell G.A. Systems model of construction accident causation. *Journal of construction engineering and management*, 131(7):816-25, 2005.

[3] Ayhan B.U. and Tokdemir O.B. Safety assessment in megaprojects using artificial intelligence. *Safety Science*, 118:273-87, 2019.

[4] Patel D.A. and Jha K.N. Neural network approach for safety climate prediction. *Journal of Management in Engineering*, 31(6):05014027, 2015.

[5] Leei G., Choi B., Jebelli H., Lee S. Assessment of construction workers' perceived risk using physiological data from wearable sensors: A machine learning approach. *Journal of Building Engineering*, 7:102824, 2021.

[6] Huang L., Fu Q., He M., Jiang D., Hao Z. Detection algorithm of safety helmet wearing based on deep learning. *Concurrency and Computation: Practice and Experience*. 10:e6234, 2021.

[7] Wei L., Cheng M., Feng M., Lijuan Z. Research on recognition of safety helmet wearing of electric power construction personnel based on artificial intelligence technology. *Journal of Physics: Conference Series*, 1684(1):012013, 2020.

[8] Luo X., Liu Q., Qiu Z. A Correlation Analysis of Construction Site Fall Accidents Based on Text Mining. *Frontiers in Built Environment*, 7, 2021.

[9] Shen T., Nagai Y., Gao C. Design of building construction safety prediction model based on optimized BP neural network algorithm. *Soft Computing*, 24(11):7839-50, 2020.

[10] Lee Y.C., Shariatfar M., Rashidi A., Lee H.W. Evidence-driven sound detection for prenotification and identification of construction safety hazards and accidents. *Automation in Construction*, 113:103127, 2020.

[11] Su Y., Mao C., Jiang R., Liu G., Wang J. Data-Driven Fire Safety Management at Building Construction Sites: Leveraging CNN. *Journal of Management in Engineering*, 37(2):04020108, 2021.

[12] Goh Y.M. and Chua D.K. Case-based reasoning for construction hazard identification: Case representation and retrieval. *Journal of Construction Engineering and Management*. 135(11):1181-9, 2009.

[13] B. Rocca, Handling imbalanced datasets in machine learning. *Towards Data Science*, 2019.

[14] RapidMiner. RapidMiner documentation. On-line: https://docs.rapidminer.com. Accessed: 07/07/2021.

[15] Brownlee J. Overfitting and underfitting with machine learning algorithms. *Machine Learning Mastery*, 21, 2016.