Evidence-Based Machine Learning Algorithm Selection for Construction Data Analytics: A Systematic Review

Stuti Garg[1], Vivek Sharma[1], Dhaval Gajjar[1]

[1] Nieri Department of Construction and Real Estate Development, Clemson University, Clemson, SC 29634, USA
viveks@clemson.edu

Abstract
Machine learning (ML) applications within the construction domain lack a structured decision-making framework for algorithm selection based on dataset characteristics and analysis objectives. The study aims to investigate whether evidence-based guidance is being utilized to formulate an ML algorithm selection that accounts for the nature and characteristics of the dataset and the objectives of the analysis. This study establishes empirical groundwork by analyzing the current use of ML methods in the construction industry and exploring potential correlations between algorithm selection, model objectives, and dataset characteristics. A comprehensive literature review following PRISMA guidelines was conducted, searching the Web of Science, ScienceDirect, IEEE, ICONDA, and ASCE databases. Thirty articles meeting the inclusion criteria were analyzed for the frequency of use of ML algorithms, the objective of the model, and the selection reasoning for the ML applications. The findings indicated "*boosting-gradient-based methods*" (25%) as the most frequently adopted algorithms for developing prediction/classification models for datasets pertinent to the construction industry. "*Regression*" based models were the most widely implemented for prediction models, while "*K-nearest neighbor*" was most frequently adopted for classification. Primary selection criteria most frequently cited for adopting ML methods included "*past use for similar data*" (23%). In comparison, 16% of methods were adopted with "*no reason*" provided, exhibiting a strong reliance on precedent rather than a systematic rationale. The study's results confirmed the absence of a systematic selection approach that directly informs the next phase of the current research. The study's findings include tabulation of concerns, the analysis of machine learning (ML) methods employed in construction, and their approach to adoption, while underscoring the performance of the algorithms corresponding to the dataset's features and the objective of analysis. This study will benefit future researchers and industry analysts by laying down critical groundwork to formulate an evidence-based ML technique selection mechanism that would map the ML algorithm type to the characteristics of the construction-specific dataset and objectives for decision-making efficiency in the construction domain.

Keywords
Machine learning, construction industry, data analysis, algorithm selection, prediction model, classification model.

1. Introduction
The construction sector is characterized by projects with unique features such as low profit margins, limited productivity, fragmented and dynamic nature, and substantial capital requirements for initiation and execution, rendering it inherently high-risk and challenging (Tariq & Gardezi, 2023). This necessitates robust forecasting mechanisms and rigorous evaluation of predictive modeling approaches to support informed decision-making (Hashemi et al., 2020). Early-phase forecasting of cost and schedule overruns, safety performance, and risk exposure is vital to the successful delivery of construction projects (Assaad et al., 2020; Castro Miranda et al., 2022; Egwim et al., 2021; Hashemi et al., 2020; Jaber et al., 2020; Tariq & Gardezi, 2023). The construction sector is generating fast-paced, large volumes of data from diverse sources, in different formats that cannot be utilized efficiently using traditional predictive data analysis tools and techniques (Egwim et al., 2021; Tan & Narayan, 2019; Yu et al., 2020; Zabala-Vargas et al., 2023). Machine Learning (ML) involves using algorithms and statistical models to enable computer systems to learn from data, identify patterns, and make decisions or predictions without being explicitly programmed with fixed rules. It relies on experience and example-based learning rather than traditional rule-based programming (Cali et al., 2021; Egwim et al., 2021). ML techniques present an opportunity to utilize the massive amounts of construction sector data for valuable insights, yielding higher accuracy in predictions (Egwim et al., 2021; Gondia et al., 2019; Poh et al., 2018; Yu et al., 2020).

ML applications hold greater promise for early-phase forecasting for construction projects, with the absence of constraints of statistical assumptions and the ability to manage missing data and imbalanced datasets while capturing linear and non-linear relationships. (Garcia et al., 2022; Gong et al., 2020; Hashemi et al., 2020; Zabala-Vargas et al., 2023). Researchers have applied ML algorithms in the construction industry for cost estimation and prediction. (Chao & Chien, 2009; Jaber et al., 2020; Razi & Ansari, 2024), schedule delays (Alsulamy, 2025; Muntasir Nishat et al., 2024, safety issues (Poh et al., 2018; Sathvik et al., 2024; Yoon et al., 2024), risks (Fan, 2020; Gong et al., 2020; Sanni-Anibire et al., 2022), and building energy consumption and carbon emissions (H. Chen et al., 2025; Hou & Liu, 2024; Razi & Ansari, 2024). However, studies exhibited the absence of any basis for selecting various ML techniques, especially regarding the objectives of analysis and characteristics of the dataset used to address the problem (Garg, 2024). Different algorithms have been developed over time for unique purposes, with their advantages and disadvantages, considering the various studies. Reviewing the literature, it was observed that an empirical approach was adopted where multiple ML algorithms, tools and combinations were deployed to address the varying nature of datasets across the spectrum of the construction sector domains to select the best-performing model (Jain et al., 2024; Turkyilmaz & Polat, 2024; Yoon et al., 2024; Zong et al., 2025). It was documented that selecting the ML algorithm based on specific parameters of the dataset could minimize trial-and-error by providing guidance (Li et al., 2018).

Traditional methods that were developed earliest in the timeline of ML algorithms, such as Artificial Neural Networks, K-nearest neighbor, Decision Tree, Regression, and Support Vector Machine (Debero & Sinesilassiea, 2024; Jain et al., 2024; Poh et al., 2018; Zong et al., 2025), have been deployed for methodological strength and the specific nature of the analyses. Moreover, advanced methods like the Ensemble approach, Random Forest, Deep Neural Networks, and Gradient Boosting Model (Hou & Liu, 2024; Lee & Son, 2021; Muntasir Nishat et al., 2024; Sanni-Anibire et al., 2021, 2022) have exhibited better performance in scenarios of imbalanced, high cardinality datasets, reducing bias, solving overfitting and improving the generalization of the models. Construction data analysis can achieve greater efficiency and precision by leveraging the distinct advantages of each technological approach (Garg, 2024). Lack of an even preliminary basis for selection of algorithms urged the study's direction towards mapping algorithms in traditional, regular, and advanced categories based on the development timeline and complexity level. Formulation of a quadrant map assignment with the x-axis calibrating the traditional nature of the algorithm and the y-axis recording the regular to advanced characteristics became part of the larger study. *The current study's objective is to confirm the lack of evidence-based selection of ML applications for studies researching construction domains and to validate the correlation between the characteristics of the dataset and goal of analysis with the adopted algorithm.* Henceforth, the scope of the current study is to lay the foundation for a decision-making framework for future researchers by evaluating the frequency of adoption of ML methods and investigating the relationship, if any, with the goal of the model and characteristics of the dataset.

2. Methods

Data collection was conducted through a systematic literature review as illustrated in Fig. 1. Literature review sourced data from five (5) databases: American Society of Civil Engineers (ASCE), International CONstruction Database (ICONDA), IEEE (Institute of Electrical and Electronics Engineers) Xplore digital library, Web of Science, and ScienceDirect. The search string adopted was: "machine learning" AND "construction industry" with articles sorted by relevance and including the first 100 results. Four hundred (400) studies were collected from ASCE, IEEE Xplore, Web of Science, ScienceDirect, while the ICONDA database showed 26 studies. A citation search for the study was conducted while reviewing literature, which added 24 studies to the collection, resulting in 450 articles. The next step involved the preliminary screening of articles with the exclusion of duplicate records, book chapters, review-type articles, and articles in which the keyword "construction" was absent in the abstract, resulting in 70 articles. Further articles were screened for including articles with "ML algorithm" adopted in the study, as well as the relevance to the construction industry, considering the direct application of findings of the articles, resulting in 30 articles. Full-text screening was conducted for 30 studies, wherein title, year of publication, algorithm adopted, categorization of algorithm adopted in the study, objective of analysis, dependent variable and basis for selection of the method/algorithm were documented for each article.

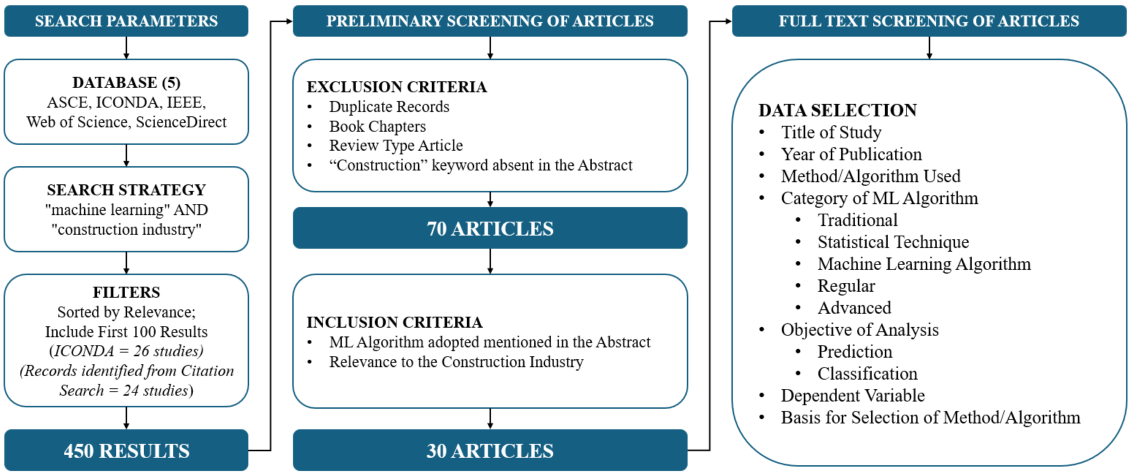

Fig. 1 Schematic Flowchart of Methodology

3. Results

3.1 Frequency of Use of ML Algorithms in Construction

115 algorithms/methods/techniques adopted for developing prediction/classification models were documented and categorized from the 30 articles. With the criterion of a minimum frequency of 3 occurrences, eight (8) categories of models were most frequently used. Each category groups the methods according to the nature of the algorithm used for training and testing the datasets discussed in Table 1, along with a reference to the reviewed articles.

Table 1. Categories of Machine Learning (ML) Algorithms

Category	ML Algorithms	References
Artificial Neural Networks (ANN)	ANN	(Razi & Ansari, 2024; Sammour et al., 2024; Turkyilmaz & Polat, 2024)
	Multi-Layer Perceptron neural network (MLP)	(Gu, 2023; Sanni-Anibire et al., 2021; Zong et al., 2025)
	ANN with Levenberg-Marquardt (LM) Algorithm	(Sanni-Anibire et al., 2021)
	ANN with Bayesian Regularization	(Geyer & Singaravel, 2018)
	Long short-term memory (LSTM)-based recurrent neural networks	(Chen et al., 2025)(Chen et al., 2025)
Boosting/Gradient (Boo/Grad)	AdaBoost	(Muntasir Nishat et al., 2024; Razi & Ansari, 2024)
	Bayesian Optimization Extreme Gradient Boosting (BO-XGBoost)	(Sathvik et al., 2024)
	CatBoost	(Razi & Ansari, 2024; Yoon et al., 2024; Zong et al., 2025)
	Gradient Boosting Model (GBM)	(Lee & Son, 2021; Muntasir Nishat et al., 2024; Sammour et al., 2024; Xia et al., 2024; Zong et al., 2025)
	Light Gradient Boosting Model (LightGBM)	(Alsulamy, 2025; Lee & Son, 2021; Sammour et al., 2024; Yoon et al., 2024; Zong et al., 2025)
	Natural Gradient Boosting (NGBoost)	(Alsulamy, 2025; Xia et al., 2024)
	Stochastic Gradient Tree Boosting (SGTB)	(Razi & Ansari, 2024; Tixier et al., 2016)
	Extreme Gradient Boosting (XGBoost)	(Alsulamy, 2025; Debero & Sinesilassiea, 2024; Hou & Liu, 2024; Sammour et al., 2024; Xia et al., 2024; Yoon et al., 2024; Zong et al., 2025)
	XGBoost optimized by Grid Search (GS-XGBoost)	(Hou & Liu, 2024)
Decision Tree (DT)	Decision Tree	(Debero & Sinesilassiea, 2024; Gu, 2023; Jain et al., 2024; Kifokeris & Xenidis, 2019; Kim et al., 2008; Lee & Son, 2021; Poh et al., 2018; Razi & Ansari,

Category	ML Algorithms	References
		2024; Sammour et al., 2024; Sathvik et al., 2024; Turkyilmaz & Polat, 2024; Zong et al., 2025)
Ensemble	AdaBoost +SVM, AdaBoost +KNN	(Gong et al., 2020)
	ANN Combiner: ANN(MLP) + KNN(Ibk) + SVM(SMOReg)	(Cheng et al., 2010)
	Association Rule Mining + Bayesian Network	(Fan, 2020)
	Bagging Regression	(Razi & Ansari, 2024)
	Binary Particle Swarm Optimization (BPSO)+AdaBoost+SVM	(Kim et al., 2008)
	Evolutionary Support Vector Machine Inference Model (ESIM) = hybrid of SVM and fast messy genetic algorithm (fmGA) and K-means clustering.	(Fan, 2020)
	Extremely Randomized Trees	(Razi & Ansari, 2024; Sammour et al., 2024)
	KNN Ensemble Method: MLRA (Linear Regression) + SVM (SMROeg)	(Sanni-Anibire et al., 2021)
	Rule Fit	(Sammour et al., 2024)
	Voting - Combine ANN, KNN and SVM by majority voting Stacking - Combine ANN and SVM with ANN as combining classifier Stacking - Combine ANN and SVM with SVM as combining classifier	(Sanni-Anibire et al., 2022)
K-Nearest Neighbors (KNN)	KNN	(S. Chen et al., 2025; Debero & Sinesilassiea, 2024; Gong et al., 2020; Poh et al., 2018; Sanni-Anibire et al., 2022; Turkyilmaz & Polat, 2024)
Random Forest (RF)	Random Forest	(Debero & Sinesilassiea, 2024; Geyer & Singaravel, 2018; Gong et al., 2020; Gu, 2023; Hou & Liu, 2024; Jain et al., 2024; Lee & Son, 2021; Muntasir Nishat et al., 2024; Poh et al., 2018; Razi & Ansari, 2024; Sammour et al., 2024; Sathvik et al., 2024; Turkyilmaz & Polat, 2024; Zong et al., 2025)
Regression (Reg)	Elastic-Net Regression	(Hou & Liu, 2024)
	Eureqa	(Sammour et al., 2024)
	Linear Regression	(Alshannaq & Awawdeh, 2024; Debero & Sinesilassiea, 2024; Razi & Ansari, 2024; Tixier et al., 2016)
	Logistic Regression	(Poh et al., 2018; Zong et al., 2025)(
	Machine Learning Regression Techniques (MLRT)	(Jaber et al., 2020)
	Multiple Linear Regression	(Chao & Chien, 2009; Jain et al., 2024)
	Polynomial Regression	(Debero & Sinesilassiea, 2024; Razi & Ansari, 2024)
Support Vector Machine (SVM)	SVM	(Jain et al., 2024; Poh et al., 2018; Razi & Ansari, 2024; Sammour et al., 2024; Sathvik et al., 2024; Turkyilmaz & Polat, 2024; Zong et al., 2025)
	SVM - SMO (Sequential Minimal Optimization)	(Sanni-Anibire et al., 2022)
	SVM – Soft Margin	(Sanni-Anibire et al., 2022)

The frequency of adoption of each of the categories has been summarized in Fig. 2, indicating 25% of the studies adopted "*boosting/gradient-based ML models*" while 13% adopted "*random forest*", followed by 12% using "*regression*" techniques and 11% "*ensemble*" methods.

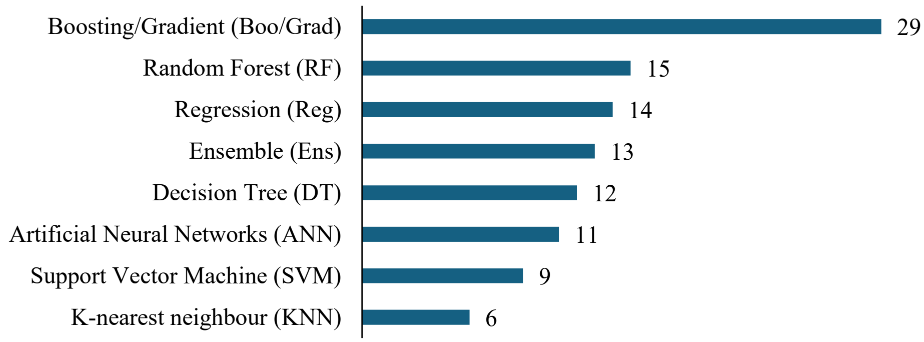

Fig. 2 Frequency of Adoption of ML Algorithms in the Construction Industry

3.2 Objective of ML Algorithm

The studies were also investigated for the objective of ML algorithms adopted – prediction or classification. The articles in which the model's outcome/dependent variable/result was continuous were categorized as "prediction"; however, if the result was discrete, the method was categorized for the objective of analysis as "classification". Fig. 3 below illustrates the percentage of ML algorithms used for prediction vis-à-vis classification. With the overall data, 55.7% of methods were adopted for prediction, while 44.3% were utilized for classification tasks for construction industry datasets.

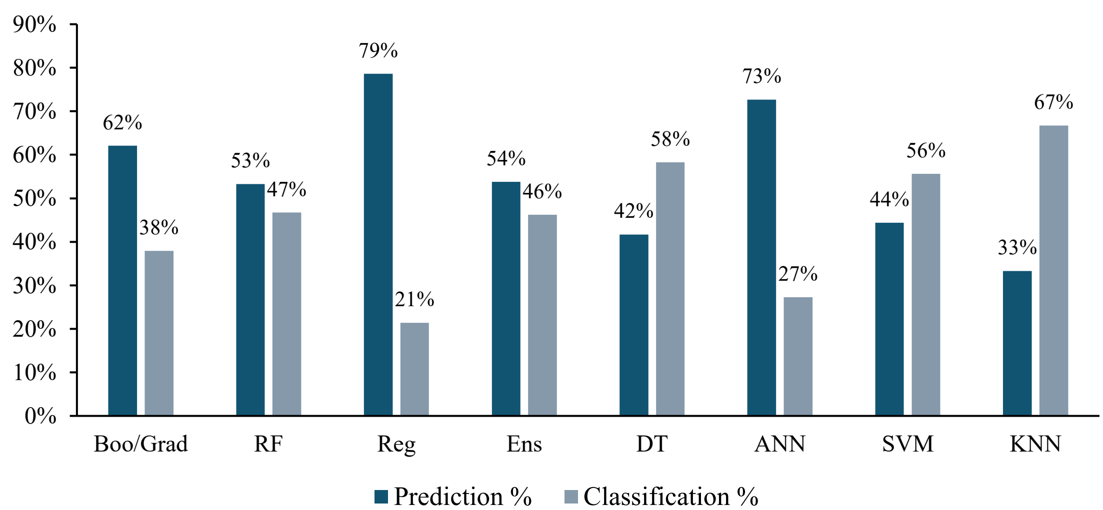

Fig 3. Percentage of ML Algorithm's Objective of Analysis – Prediction vis-à-vis Classification

3.3 Basis of Selection of ML Algorithm

On reviewing the articles in the final stage of full-text screening, the "*basis for selection*" of the algorithm/technique/method was documented for each of the 115 methods. Fig. 4 below summarizes the results of the "basis for selection/reasoning" frequency for adopting ML algorithms in the studies. The key insights were that the top-ranked reasons cited in studies for adopting a particular algorithm were "*past use for similar data*" at 23%, followed by "*better accuracy*" at 17% and "*past use for similar objective*" at 16%.

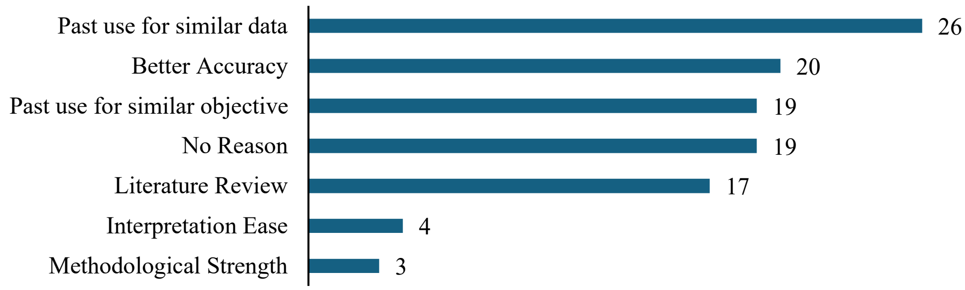

Fig 4. Frequency of "Basis of Selection" of ML Algorithms

4. Discussion

4.1 ML Algorithms in Construction

Categorization of the different ML algorithms, as evident in Table 1, enabled investigation of the advantages and disadvantages along with the unique characteristics of each type of algorithm listed below.

1) *Artificial Neural Networks (ANN):* ANNs are conceptualized as interconnected systems for parallel distributed information processing that map input vectors into output vectors. Structurally, they comprise multiple processing elements, referred to as neurons or nodes, typically organized into three primary layers: input, hidden, and output. These networks acquire knowledge by learning through structured layers of neurons, utilizing transition functions to mediate information flow and adaptation - repeatedly adjusting numerical weights associated with the interconnections between neurons.

   Advantages: Ability to handle complex nonlinear relationships and handle variety of data types with high accuracy and predictive power.

   Disadvantages: Poor interpretability and limited transparency, often described as a "black box" nature, making it difficult to understand how the predictions are generated; prone to overfitting; complicated regulation of hyperparameters; significant computational requirements for large datasets.

2) *Boosting/Gradient:* These algorithms aggregate multiple weak learners to form a strong predictive model with improved accuracy. The methods follow a sequential learning strategy, progressively refining predictions by addressing and correcting the mistakes of prior iterations. The training process involves iteratively computing gradients and constructing trees, where each subsequent tree targets the residual errors left by preceding trees. A crucial feature of such approaches is the inclusion of regularization techniques, which manage model complexity and mitigate the risk of overfitting.

   Advantages: High accuracy and generalization; reduced overfitting; robust results addressing biases; handles complex nonlinear relationships; some variants, such as XGBoost, are optimized for parallelization and speed.

   Disadvantages: Poor interpretability; high computing resource demands; high model complexity.

3) *Decision Tree (DT)*: DT Regression observes the features of an object and trains a model in a tree's structure to predict future data to produce meaningful continuous output. The DT Classifier operates based on the majority decision principle, focusing on how independent factors divide data records into class labels by analyzing training data. It generates a hierarchical structure with nodes, branches, and leaves, where the root node represents the entire dataset, sub-nodes are formed by data division based on specific characteristics, and leaves reflect the final determined class labels.

   Advantages: Easy to understand and visualize, making them suitable for deriving rules and comprehending complicated problems in an understandable hierarchical structure. Ability to handle different data types (categorical and numerical) and apply them to both linear and nonlinear relationships. Require no data preprocessing for specific tasks and are adaptive to nonlinear data.

   Disadvantages: Prone to overfitting, especially on noisy datasets; high computing resource demands for large datasets; sensitive to noise in the data.

4) *Ensemble*: These methods represent a class of ML techniques designed to enhance the effectiveness of a single classifier by aggregating the outputs of multiple base classifiers. Also known as multiple classifier systems or integrated learning approaches, these methods involve constructing several models whose predictions are combined, frequently resulting in superior overall performance—even when individual models are relatively weak. Popular ensemble strategies include combining boosting algorithms, such as AdaBoost and Stochastic Gradient Boosting, and bagging algorithms like Random Forest and Bagging Regression.

Advantages: High accuracy and generalization; reduced overfitting and addresses biases found in singular methods; robust to noise and outliers; proficient at modeling intricate, nonlinear patterns within data; ability to handle large volumes of data, high-dimensional data, imbalanced datasets, and missing values; adaptable to various data types.
Disadvantages: Increased computational demands; sensitivity to base learner weaknesses, e.g., if a base classifier struggles with specific challenges like minority group classification, even with an ensemble framework, the performance for those groups might still not meet application needs.

5) *K-Nearest Neighbors (KNN)*: KNN operates based on a proximity mechanism, which involves identifying the "k" closest instances or samples from the training data to a new, unseen data point. KNN applies a simple majority decision rule for classification: the new instance is assigned to the class most represented among its "k" nearest neighbors. The algorithm typically outputs the mean value of its "k" nearest neighbors for regression tasks.
Advantages: Simple and easy to use; reduced computing complexity and improved categorization speed; applicable to linear and non-linear relationships; capturing local patterns in datasets.
Disadvantages: Sensitivity to parameters; performance degradation with high-dimensional data; unsuitable for imbalanced datasets and sensitive to noise.
6) *Random Forest (RF)*: Random Forest algorithms build numerous decision trees during training and merge their individual outputs to create a single, more reliable prediction. This process uses bootstrap aggregating (bagging), where each tree is developed using a randomly selected portion of training data and randomly chosen input variables. The final output is usually determined by taking the mean of all tree results in regression tasks, or through majority consensus in classification problems. This collective methodology helps minimize overfitting risks and improves the model's overall performance. Random Forest commonly employs the CART (Classification and Regression Trees) algorithm as the splitting criterion for building individual trees to reduce node impurity.
Advantages: High accuracy and strong generalization; reduced overfitting; robust to noise and outliers; versatility with data types and dimensions; simple to determine each variable's significance.
Disadvantages: Scalability issues with large datasets; high model complexity; limited interpretability.
7) *Regression*: Linear regression estimates the relationship between one independent and dependent variable using a straight line, adjusting this linear relationship to predict new data. Multilinear regression extends this to numerous independent variables predicting a single dependent variable. Logistic regression is primarily used for classification tasks where the outcome is categorical. It calculates probabilities using the sigmoid function to classify new observations into distinct categories.
Advantages: Simplicity and interpretability; effective for specific applications and predictive power for calculating forecasts; identification of significant predictors.
Disadvantages: Assumption of linear relationship; statistical assumptions of normal distribution constrain linear regression; poor predictive performance with complex data; logistic regression is less effective for multi-class problems.
8) *Support Vector Machine (SVM)*: SVM is a supervised learning method used for classification and regression, functioning by combining the aspects of linear and instance-based modeling. This method aims to find an "optimal hyperplane" – a decision boundary for classification or a "best fit" line for regression that maximizes the space between the varying classes of data or includes as many data points as possible. Key data points, "support vectors," are identified to achieve the objective. At the same time, SVM uses the "kernel trick" for more complex and non-linear datasets to map the data into a higher dimension for distinct linear separation.
Advantages: High accuracy and strong performance; effective with small and high-dimensional data; strong for non-linear modeling; robust to outliers.
Disadvantages: High computing resource demands for large datasets; loss of precision with boundary overlaps during distinction; increased complexity for multi-class problems since it is designed for binary classification.

4.2 Selection of ML Algorithms in Construction

An analysis of descriptive statistics for the data collected through literature review showed that "*boosting and gradient-based*" methods and "*random forest*" algorithms were the most adopted ML algorithms for developing prediction and classification models in the construction domain. Studies have associated these methods with being highly effective in handling complex, non-linear relationships between variables, a common feature in datasets within the construction industry. Moreover, with imbalanced distribution and skewed allocation of data points across input variables, these methods solve overfitting and generalize reasonably well, reducing potential biases. Considering the adoption of techniques for specific objectives of analysis, it can be deduced from Fig. 3 that "*regression*" and "*ANN*" are the top-ranked algorithms for prediction models. The top two selected techniques for developing classification models were "*KNN*" and "*decision tree*". Algorithms like "*random forest*" and "*ensemble*" methods showed versatility in prediction and classification tasks. In Fig. 4, the "basis for selection" as stated in the articles for adopting a particular technique/method/algorithm was mapped across the studies. The most prevalent reason, "*past use for similar data*," to adopt the ML algorithm underscores the critical correlation between the nature of the dataset and the technique selected for developing a prediction/classification model. Studies depended on the precedent of performance of ML applications with similarities in specific data characteristics, including but not limited to size, dimensionality, type (continuous or categorical), imbalanced, and missing values. This was followed by "*better accuracy*," which is reasoned by researchers as an important criterion for selecting a specific ML method to ensure robust performance of the algorithm, as well as better generalization over unseen data, exhibiting strength in the methodology of the algorithm. It is important to note that at 16%, the reason for "*past use for similar objective*" was also a vital basis for selecting algorithms, considering the past documented performance of the method and established track record in achieving a similar goal of prediction/classification for the specific context of construction domain (e.g. cost, schedule, safety, risk, energy consumption, etc.). It is critical to note that 16% of the methods were adopted for "*no reason*," cited following the empirical principle of testing diverse algorithms to compare the performance, concluding the algorithm best according to the performance evaluation metrics. To investigate this further, in the current study, the algorithm and the reason for selection were mapped to create combinations as highlighted in Table 2.

Table 2. Algorithm-Reason for Selection Combination Frequency

Algorithm	Reason for Selection	Frequency	Percentage
Boosting/Gradient	Past use for similar data	9	7.8%
Boosting/Gradient	Better Accuracy	8	7.0%
Boosting/Gradient	No Reason	5	4.3%
Random Forest	Better Accuracy	5	4.3%
Regression	Literature Review	5	4.3%
ANN	No Reason	4	3.5%
Boosting/Gradient	Past use for similar objective	4	3.5%
Ensemble	No Reason	4	3.5%
Ensemble	Past use for similar objective	4	3.5%
Random Forest	Past use for similar objective	4	3.5%
Decision Tree	Literature Review	3	2.6%
Decision Tree	Past use for similar data	3	2.6%
Ensemble	Better Accuracy	3	2.6%
Regression	Past use for similar data	3	2.6%
SVM	Literature Review	3	2.6%

Considering the criterion of documenting only occurrences with a frequency of 3 or higher, it can be noted that the reason for adopting methods for previous use for construction datasets includes "*boosting/gradient-based models*", "*decision tree*" and "*regression*". Table 2 further highlights that methods like "*boosting/gradient-based models*", "*random forest*", and "*ensemble*" have been adopted for better accuracy, indicating records of strong performance of these algorithms for the construction industry. However, it is integral to understand that methods like "*boosting/gradient-based models",* "*ANN*", and "*Ensemble*" have been adopted with no reason stated in the studies, which further emphasizes the lack of a decision-making framework present for selecting an ML algorithm for developing prediction/classification models for construction datasets.

5. Conclusions

The current study conducted a systematic literature review for articles utilizing ML in the construction industry, investigating 115 methods from 30 articles. The most adopted ML techniques were "*boosting/gradient-based models*" (25%), "*random forest*" (13%), "*regression*" techniques (12%) and "*ensemble*" methods (11%). The preferred methods for developing prediction models were "*regression*", "*ANN*" and "*boosting/gradient*" and for classification, "*KNN*", "*Decision tree*" and "*SVM*". The most frequently cited rationale for algorithm selection was "*past use for similar data*" (23%), suggesting that dataset characteristics could influence the method adopted. However, this finding cannot determine whether such precedent-based selections represent optimal data-algorithm mapping or are simply a reflection of researchers' habits and methodological convenience. This ambiguity, coupled with the finding that 16% of methods were adopted with "*no reason*", highlights a significant methodological gap where algorithm selection is often based on prior practice rather than a systematic analysis of data characteristics and problem requirements.

Considering the documented prevalence of precedent-based selection (39% - on combining '*past use for similar data*' and '*past use for similar objective'*) indicates that researchers intuitively recognize the dependence of the algorithms' performance on the dataset's size and cardinality, the relationship between variables, missing values, imbalanced distributions, etc. however lack a systematic framework to guide the decisions. The next phase of this research aims to empirically validate the algorithm-dataset-objective combinations that yield optimal performance – transforming the current trial-and-error approach into a systematic evidence-based methodology. ML applications vary based on the specific problems being tackled, with the choice of methodology directly tied to the nature of available data. Different ML approaches are suited to distinct data formats, each with strengths and limitations. Additionally, combining multiple ML techniques is often necessary to achieve optimal results when dealing with intricate challenges and large datasets. Future directions of research that have been identified during the study are listed below:

- The current study becomes the basis of a larger study to develop norms for the landscape of the selection matrix of ML techniques for the construction industry. It represents a critical step towards establishing evidence-based ML practices in the construction industry. By documenting current algorithm adoption patterns and revealing gaps in selection methodologies, the study provides the foundation for developing comprehensive decision-making frameworks that can significantly enhance the effectiveness and efficiency of ML applications in construction.
- Another future scope of the study would be the assignment of ML techniques in a conceptual quadrant to aid time-saving selection of the ML algorithm best suited for the nature of data, analysis and desired prediction for datasets in the construction industry. Further, the conceptual quadrant would categorize the algorithms according to the application's traditional, regular or advanced nature.
- Similar quadrant assignments for algorithms can be conducted for other industries (namely, retail, manufacturing, finance, etc.)and comparative analysis can be conducted between different industry sectors to explore algorithmic versatility, assess cross-domain applicability of ML solutions, and establish a comprehensive taxonomy of algorithm-industry compatibility.

The ultimate goal is to transform construction data analytics from a trial-and-error process to a systematic, evidence-based discipline that can fully harness the potential of ML to address the industry's complex challenges, benefiting future academicians and industry researchers.

References

Alshannaq, A. A., & Awawdeh, A. R. (2024). Implementation of Machine Learning in Predicting Pin-Bearing Strength of Aged and Nonaged Pultruded GFRP Composites. *Journal of Composites for Construction*, *28*(3). https://doi.org/10.1061/JCCOF2.CCENG-4483

Alsulamy, S. (2025). Predicting construction delay risks in Saudi Arabian projects: A comparative analysis of CatBoost, XGBoost, and LGBM. *Expert Systems with Applications*, *268*. https://doi.org/10.1016/j.eswa.2024.126268

Assaad, R., El-Adaway, I. H., & Abotaleb, I. S. (2020). Predicting Project Performance in the Construction Industry. *Journal of Construction Engineering and Management*, *146*(5). https://doi.org/10.1061/(asce)co.1943-7862.0001797

Cali, U., Kuzlu, M., Pipattanasomporn, M., Kempf, J., & Bai, L. (2021). Foundations of Big Data, Machine Learning, and Artificial Intelligence and Explainable Artificial Intelligence. In *Digitalization of Power Markets and Systems Using Energy Informatics*. Springer, Cham. https://doi.org/https://doi.org/10.1007/978-3-030-83301-5_6

Castro Miranda, S. L., Del Rey Castillo, E., Gonzalez, V., & Adafin, J. (2022). Predictive Analytics for Early-Stage Construction Costs Estimation. *Buildings*, *12*(7). https://doi.org/10.3390/buildings12071043

Chao, L.-C., & Chien, C.-F. (2009). Estimating Project S-Curves Using Polynomial Function and Neural Networks. *Journal of Construction Engineering and Management*, *135*(3), 169–177. https://doi.org/10.1061/ASCE0733-93642009135:3169

Chen, H., Wang, J., Shen, Q. G., Chen, B., Dong, J., Feng, Z., & Liu, Y. (2025). Application of hybrid machine learning algorithms for life cycle carbon prediction and optimization of buildings: A case study in China. *Sustainable Cities and Society*, *122*. https://doi.org/10.1016/j.scs.2025.106248

Chen, S., Bangaru, S. S., Yigit, T., Trkov, M., Wang, C., & Yi, J. (2025). Machine Learning-Based Real-Time Walking Activity and Posture Estimation in Construction With a Single Wearable Inertial Measurement Unit. *IEEE Transactions on Automation Science and Engineering*, *22*, 753–758. https://doi.org/https://doi.org/10.1109/TASE.2025.3573966

Cheng, M. Y., Wu, Y. W., & Wu, C. F. (2010). Project success prediction using an evolutionary support vector machine inference model. *Automation in Construction*, *19*(3), 302–307. https://doi.org/10.1016/j.autcon.2009.12.003

Debero, D. W., & Sinesilassiea, E. G. (2024). Machine Learning Model for Construction Time Prediction: A Case of Selected Public Building Projects in Hosanna, Ethiopia. *Journal of Engineering (United Kingdom)*, *2024*. https://doi.org/10.1155/2024/5653690

Egwim, C. N., Alaka, H., Toriola-Coker, L. O., Balogun, H., & Sunmola, F. (2021). Applied artificial intelligence for predicting construction projects delay. *Machine Learning with Applications*, *6*, 100166. https://doi.org/10.1016/j.mlwa.2021.100166

Fan, C.-L. (2020). Defect Risk Assessment Using a Hybrid Machine Learning Method. *Journal of Construction Engineering and Management*, *146*(9). https://doi.org/10.1061/(ASCE)CO.1943

Garcia, J., Villavicencio, G., Altimiras, F., Crawford, B., Soto, R., Minatogawa, V., Franco, M., Martínez-Muñoz, D., & Yepes, V. (2022). Machine learning techniques applied to construction: A hybrid bibliometric analysis of advances and future directions. *Automation in Construction*, *142*. https://doi.org/10.1016/j.autcon.2022.104532

Garg, S. (2024). *Assessing the Efficacy: Statistical Models vs. Machine Learning (ML) Approaches for Prediction Modeling In the Construction Industry* [Clemson University]. https://open.clemson.edu/all_theses/4362

Geyer, P., & Singaravel, S. (2018). Component-based machine learning for performance prediction in building design. *Applied Energy*, *228*, 1439–1453. https://doi.org/10.1016/j.apenergy.2018.07.011

Gondia, A., Siam, A., El-Dakhakhni, W., & Nassar, A. H. (2019). *Machine Learning Algorithms for Construction Projects Delay Risk Prediction*. https://doi.org/10.1061/(ASCE)

Gong, P., Guo, H., Huang, Y., & Guo, S. (2020). Safety risk evaluations of deep foundation construction schemes based on imbalanced data sets. *Journal of Civil Engineering and Management*, *26*(4), 380–395. https://doi.org/10.3846/jcem.2020.12321

Gu, S. (2023). Construction Cost Index Prediction Based on Machine Learning. *International Conference on Applied Intelligence and Sustainable Computing, ICAISC 2023*. https://doi.org/10.1109/ICAISC58445.2023.10199679

Hashemi, S. T., Ebadati, O. M., & Kaur, H. (2020). Cost estimation and prediction in construction projects: a systematic review on machine learning techniques. *SN Applied Sciences*, *2*(10). https://doi.org/10.1007/s42452-020-03497-1

Hou, Y., & Liu, S. (2024). Predictive Modeling and Validation of Carbon Emissions from China's Coastal Construction Industry: A BO-XGBoost Ensemble Approach. *Sustainability (Switzerland)* , *16*(10). https://doi.org/10.3390/su16104215

Jaber, F. K., Jasim, N. A., & Al-Zwainy, F. M. S. (2020). Forecasting techniques in construction industry: Earned value indicators and performance models. *Scientific Review Engineering and Environmental Sciences*, *29*(2), 234–243. https://doi.org/10.22630/PNIKS.2020.29.2.20

Jain, S., Jauhar, S. K., & Piyush. (2024). A machine-learning-based framework for contractor selection and order allocation in public construction projects considering sustainability, risk, and safety. *Annals of Operations Research*, *338*(1), 225–267. https://doi.org/10.1007/s10479-024-05898-6

Kifokeris, D., & Xenidis, Y. (2019). Risk source-based constructability appraisal using supervised machine learning. *Automation in Construction*, *104*, 341–359. https://doi.org/10.1016/j.autcon.2019.04.012

Kim, H., Soibelman, L., & Grobler, F. (2008). Factor selection for delay analysis using Knowledge Discovery in Databases. *Automation in Construction*, *17*(5), 550–560. https://doi.org/10.1016/j.autcon.2007.10.001

Lee, S. H., & Son, J. (2021). Development of a safety management system tracking the weight of heavy objects carried by construction workers using fsr sensors. *Applied Sciences (Switzerland)*, *11*(4), 1–15. https://doi.org/10.3390/app11041378

Li, Z., Zhong, X., & Cui, Z. (2018). Evaluating forecasting algorithm of realistic datasets based on machine learning. *ACM International Conference Proceeding Series*, *Part F137692*, 72–76. https://doi.org/10.1145/3194206.3194238

Muntasir Nishat, M., Magnussen Neraas, S., Marsov, A., & Olsson, N. O. E. (2024). Prediction of project activity delays caused by variation orders: A machine-learning approach. *IOP Conference Series: Earth and Environmental Science*, *1389*(1). https://doi.org/10.1088/1755-1315/1389/1/012038

Nguyen Van, T., & Nguyen Quoc, T. (2021). Research Trends on Machine Learning in Construction Management: A Scientometric Analysis. *Journal of Applied Science and Technology Trends*, *2*(03), 96–104. https://doi.org/10.38094/jastt203105

Poh, C. Q. X., Ubeynarayana, C. U., & Goh, Y. M. (2018). Safety leading indicators for construction sites: A machine learning approach. *Automation in Construction*, *93*, 375–386. https://doi.org/10.1016/j.autcon.2018.03.022

Razi, N., & Ansari, R. (2024). A prediction-based model to optimize construction programs: Considering time, cost, energy consumption, and CO2 emissions trade-off. *Journal of Cleaner Production*, *445*. https://doi.org/10.1016/j.jclepro.2024.141164

Sammour, F., Alkailani, H., Sweis, G. J., Sweis, R. J., Maaitah, W., & Alashkar, A. (2024). Forecasting demand in the residential construction industry using machine learning algorithms in Jordan. *Construction Innovation*, *24*(5), 1228–1254. https://doi.org/10.1108/CI-10-2022-0279

Sanni-Anibire, M. O., Zin, R. M., & Olatunji, S. O. (2021). Machine learning - Based framework for construction delay mitigation. *Journal of Information Technology in Construction*, *26*, 303–318. https://doi.org/10.36680/j.itcon.2021.017

Sanni-Anibire, M. O., Zin, R. M., & Olatunji, S. O. (2022). Machine learning model for delay risk assessment in tall building projects. *International Journal of Construction Management*, *22*(11), 2134–2143. https://doi.org/10.1080/15623599.2020.1768326

Sathvik, S., Alsharef, A., Singh, A. K., Shah, M. A., & ShivaKumar, G. (2024). Enhancing construction safety: predicting worker sleep deprivation using machine learning algorithms. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-65568-2

Tan, H. C., & Narayan, S. (2019). Adopting Big Data to Forecast Success of Construction Projects: A Review. *Malaysian Construction Research Journal*, *6*(1). https://www.researchgate.net/publication/335570239

Tariq, J., & Gardezi, S. S. S. (2023). Study the delays and conflicts for construction projects and their mutual relationship: A review. *Ain Shams Engineering Journal*, *14*(1). https://doi.org/10.1016/j.asej.2022.101815

Tixier, A. J. P., Hallowell, M. R., Rajagopalan, B., & Bowman, D. (2016). Application of machine learning to construction injury prediction. *Automation in Construction*, *69*, 102–114. https://doi.org/10.1016/j.autcon.2016.05.016

Turkyilmaz, A. H., & Polat, G. (2024). Risk-Based Completion Cost Overrun Ratio Estimation in Construction Projects Using Machine Learning Classification Algorithms: A Case Study. *Buildings*, *14*(11). https://doi.org/10.3390/buildings14113541

Xia, X., Xiang, P., Khanmohammadi, S., Gao, T., & Arashpour, M. (2024). Predicting Safety Accident Costs in Construction Projects Using Ensemble Data-Driven Models. *Journal of Construction Engineering and Management*, *150*(7). https://doi.org/10.1061/JCEMD4.COENG-14397

Yoon, S., Chang, T., & Chi, S. (2024). Developing an Integrated Construction Safety Management System for Accident Prevention. *Journal of Management in Engineering*. https://doi.org/10.1061/JMENEA.MEENG-6074

Yu, T., Liang, X., & Wang, Y. (2020). Factors Affecting the Utilization of Big Data in Construction Projects. *Journal of Construction Engineering and Management*, *146*(5). https://doi.org/10.1061/(asce)co.1943-7862.0001807

Zabala-Vargas, S., Jaimes-Quintanilla, M., & Jimenez-Barrera, M. H. (2023). Big Data, Data Science, and Artificial Intelligence for Project Management in the Architecture, Engineering, and Construction Industry: A Systematic Review. *Buildings*, *13*(12). https://doi.org/10.3390/buildings13122944

Zong, H., Yi, W., Chan, A. P. C., Yang, H., Wu, P., & Xiao, B. (2025). Developing a Fatigue Model for Construction Workers: An Interpretable Machine Learning Approach. *Journal of Management in Engineering*, *41*(4). https://doi.org/10.1061/JMENEA.MEENG-6662