International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Machine Learning-Based Risk Assessment Models in Construction Project Management: A Meta-Analysis

[1]Tabish Qureshi, [2]Dr. M. Adil Khan, [3]Zain Ullah, [4]Businge Brian, [5]Muhammad Safi Ullah, [6]Shan Ul Haq, [7]Muhammad Ahmad Javed, [8]Mudassir Ahmad Khan

Submitted: 02/11/2024 **Revised**: 19/12/2024 **Accepted**: 27/12/2024

Abstract–Construction project management increasingly relies on machine learning to assess and predict risks, yet the overall predictive accuracy of these models remains insufficiently synthesized. This systematic review and meta-analysis aimed to evaluate the aggregate predictive performance of machine learning-based risk assessment models in construction project management, focusing on the effect size of prediction errors. We conducted. We systematically identified and extracted relevant studies, and a random-effects meta-analysis was performed on the pooled effect size. Our analysis included two studies that met the inclusion criteria. The results yielded a statistically non-significant yet negative overall effect size of -0.01 (SE = 0.004, 95% CI $[-0.02, -0.00]$, $z = -2.24$, $p = 0.02$). This finding suggests that machine learning models, on average, produce predictions that are marginally lower than observed outcomes, indicating a slight systematic underestimation of construction project risks. The heterogeneity among the included studies was considerable, and the small number of studies limits the generalizability of the conclusions. We therefore conclude that while machine learning offers potential for risk assessment in construction, current models exhibit a modest bias that warrants further refinement. Future research should focus on model calibration and the inclusion of more diverse datasets to improve predictive accuracy and practical applicability in the field.

Keywords: *Construction, civil engineering, machine learning, risk management*

I. INTRODUCTION

The construction industry is widely recognized as one of the most complex and risk-prone sectors in the global economy [1]. This complexity arises from a multitude of factors, including the inherently temporary nature of project organizations, the physical uniqueness of each construction project, the intricate interdependencies among numerous stakeholders, and the exposure to uncertain external environments such as weather, regulatory changes, and market fluctuations [2]. Risk assessment in construction project management has traditionally been performed using qualitative methods, such as expert judgment and risk registers, or quantitative methods, including probabilistic cost and schedule simulations [3]. While these traditional techniques have provided a foundational framework for managing project uncertainties, they are often limited by their reliance on subjective inputs, their inability to process large volumes of data effectively, and their static nature, which cannot easily adapt to evolving project conditions [4].

The recent proliferation of digital technologies and data collection systems in the construction industry has generated an unprecedented volume of project-related data. This data, ranging from sensor readings on equipment to financial transactions and scheduling records, creates a fertile ground for the application of advanced analytical methods [5]. Concurrently, the field of machine learning (ML) has experienced rapid advancements, offering powerful tools for pattern recognition, prediction, and classification from complex datasets [6]. ML models, such as artificial neural

Tabish.qureshi82@gmail.com Upsource by Solutions (Solutions by STC) Masters of Computer Science - University of Karachi

adee.uol@gmail.com Resident Engineer, NESPAK (Corresponding author)

Zainullahen@gmail.com School of Civil Engineering, Southeast University Nanjing China

2493210@tongji.edu.cn Tongji University

muhammadsafiullah64@gmail.com Lecturer, Swedish College of Engineering and Technology, Wah Cantt, Pakistan

shan.ul.haq@hitecuni.edu.pk Lab engineer, Department of civil engineering, HITEC University Taxila, Taxila Cantt, Pakistan.

ahmad.javed@ce.uol.edu.pk Department of Civil Engineering, The University of Lahore

engrmudassirahmadkhan@gmail.com UET Peshawar

networks, support vector machines, random forests, and gradient boosting machines, have been increasingly applied to various construction management problems, including cost estimation, schedule prediction, safety hazard detection, and, most pertinently, risk assessment [7]; [8].

A significant body of research has explored the development of specific ML-based risk assessment models for construction projects. These models aim to forecast the likelihood or impact of various risks, for instance, predicting delays, cost overruns, or safety incidents before they materialize [9]; [10]. While individual studies often report promising metrics, such as high classification accuracy or low root mean squared error on their own datasets, the generalizability and aggregate predictive performance of these models across different contexts remain unclear [11]. This lack of synthesis presents a considerable research gap, hindering the ability of practitioners to gauge the true effectiveness and reliability of ML-based methods in this domain and impeding the development of more robust and universally applicable risk assessment tools.

The primary motivation for this systematic review and meta-analysis is, therefore, to provide a rigorous, evidence-based synthesis of the predictive accuracy of machine learning-based risk assessment models in construction project management. While many narrative reviews have cataloged the various techniques being used, no prior study has quantitatively aggregated the effect sizes of prediction errors to derive a single, pooled estimate of model performance. This quantitative synthesis is crucial for establishing the current state of the art and for identifying the magnitude of systematic bias that may exist in these models. The significance of this work lies in its potential to inform both research and practice. For researchers, our findings highlight the need for improved model calibration and methodological standardization in reporting. For practitioners, we offer a high-level understanding of the average precision and limitations of current ML tools, thereby setting realistic expectations about their deployment and contribution to project governance.

This paper is organized as follows. Section 2 Methodology. Section 3 presents the results of the review, including the meta-analysis outcomes and an assessment of publication bias. Section 4 discusses the implications of these findings within the broader context of construction informatics, and Section 5 concludes the study by summarizing its contributions and proposing directions for future inquiry.

II. METHODOLOGY

A. Review Protocol

We conducted this systematic review following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [12]. A comprehensive literature search was performed across multiple electronic databases to ensure broad coverage of relevant studies. The search strategy was developed iteratively to balance sensitivity and specificity regarding machine learning applications for risk assessment in construction project management.

We first searched IEEE Xplore, a primary resource for engineering and computer science literature, particularly for technical contributions involving machine learning algorithms and their performance metrics. For this database, we used the following search string: ("machine learning" OR "deep learning" OR "neural networks" OR "artificial intelligence") AND ("risk assessment" OR "risk prediction" OR "risk management") AND ("construction projects" OR "construction industry" OR "civil engineering") NOT ("review" OR "survey" OR "meta-analysis" OR "bibliometric"). We applied filters to limit the publication year to between 2000 and 2024 and to exclude conference proceedings when only journal articles were preferred, though we acknowledge the significance of conferences in this venue. We also ensured that the document type was not set to ‘Review’.

Next, we searched Scopus, which indexes a vast collection of peer-reviewed journals and conference proceedings across disciplines. The search string was modified for this database as follows: TITLE-ABS-KEY(("machine learning" OR "deep learning" OR "neural network*" OR "artificial intelligence") AND ("risk assessment" OR "risk prediction" OR "risk analysis") AND ("construction project*" OR "construction industry" OR "building project*")) AND NOT TITLE-ABS-KEY("review" OR "survey" OR "meta-analysis" OR "bibliometric") AND PUBYEAR > 1999 AND PUBYEAR <. We limited the document type to ‘Article’ or ‘Conference Paper’ and excluded ‘Review’.

Then, we searched Web of Science, a comprehensive citation database covering high-impact journals. The search string was: TS=(("machine learning" OR "deep learning" OR "neural network*" OR "artificial intelligence") AND ("risk assessment" OR "risk prediction" OR "risk analysis") AND ("construction project*" OR "construction industry")) NOT TS=("review" OR "survey" OR "meta-analysis") AND PY=(2000-2024). We refined the results by selecting ‘Article’ and ‘Proceedings Paper’ as document types and excluding ‘Review Article’.

We subsequently searched ScienceDirect, which provides access to a large collection of scientific and technical research. For this database, the search string was: ("machine learning" OR "deep learning" OR "neural networks" OR "artificial intelligence") AND ("risk assessment" OR "risk prediction") AND ("construction" OR "civil engineering") NOT ("review" OR "survey" OR "meta-analysis"). We set the date range to 2000-2024 and selected 'Research articles' in the 'Article Type' filter, deselecting 'Review articles'.

Finally, we conducted a supplementary search in Google Scholar to identify grey literature and studies not indexed in the other databases. The search string was: ("machine learning" OR "deep learning" OR "neural network*" OR "artificial intelligence" OR AI) AND ("risk assessment" OR "risk prediction" OR "risk analysis" OR "failure prediction") AND ("construction project*" OR "construction industry" OR "civil engineering project*" OR "building project*") -review -survey -"meta-analysis" -bibliometric. We set the publication year range to 2000–2024 using the date slider. The last search was performed on May 15, 2024.

B. Inclusion and Exclusion Criteria

To ensure the relevance and consistency of selected studies, we defined clear inclusion and exclusion criteria. Studies were considered eligible if they focused on the development or application of machine learning-based risk assessment models for construction project management. Eligible models had to use machine learning techniques, including but not limited to neural networks, support vector machines, decision trees, or ensemble methods. The study population had to involve construction projects, which we defined broadly to include building, civil infrastructure, and industrial construction. Eligible research designs included predictive modeling studies, comparative algorithm studies, and application-based case studies. Publications had to be peer-reviewed journal articles or conference proceedings written in English. The time frame for publication was set from January 2000 to December 2024. We excluded studies that were non-empirical, such as review articles, survey papers, meta-analyses, and bibliometric analyses, as well as studies that did not report sufficient quantitative data for effect size computation, e.g., studies that only reported qualitative results without predictive performance metrics. Also excluded were studies that applied machine learning to areas of construction management unrelated to risk, such as cost estimation alone without explicit risk focus, or those that focused on risk assessment in industries other than construction, such as finance or manufacturing. Finally, we excluded studies with insufficient data for extraction, such as those lacking measures of prediction error, accuracy, or comparable effect sizes.

C. Study Selection Process

The study selection process was conducted in multiple stages, following the PRISMA flowchart as shown in Figure 1. Two reviewers independently screened the titles and abstracts of all retrieved records against the inclusion criteria. Any disagreements between the reviewers were resolved through discussion or by consulting a third reviewer. After the initial screening, full-text articles were obtained for all potentially eligible studies. These articles were then assessed in detail for eligibility, with reasons for exclusion being documented for the final meta-analysis. The quality of the included studies was appraised using a custom checklist adapted from the Prediction model Risk Of Bias ASsessment Tool (PROBAST) [13], which is specifically designed for evaluating prediction model studies. This checklist assesses four domains: participants, predictors, outcome, and analysis. Each domain was rated as having low, high, or unclear risk of bias, and an overall risk of bias judgment was derived. This quality assessment was used not as a grounds for exclusion but to inform sensitivity analyses and to contextualize the findings. Studies scoring high risk of bias in multiple domains were flagged for sensitivity analysis.

We initially retrieved a total of 627 records from the database searches. After removing 107 duplicate records and 5 records removed for other reasons (e.g., inability to retrieve full text or non-English language), we screened the abstracts of 515 records. Of these, 415 records were excluded because they clearly did not meet the inclusion criteria, often focusing on risk assessment outside of construction, using non-ML methods, or being review articles. We then sought to retrieve the full texts of 100 reports. Of these, 23 reports were not retrieved due to unavailability or access restrictions. We assessed 77 full-text reports for eligibility. A total of 75 reports were excluded due to ineligibility. The primary reasons for exclusion included insufficient quantitative data for effect size extraction (e.g., reporting only accuracy without error measures, reporting only classification metrics without regression metrics needed for our meta-analysis, or presenting results in non-extractable graphical format without numerical tables), and lack of focus on risk assessment within construction project management (e.g., focusing on safety incident prediction without linking to broader risk management, or focusing on general cost estimation not contextualized as risk). Therefore, only 2 studies met all the criteria and were included in the final systematic review and meta-analysis.

This highly selective process introduces a significant limitation. The small number of included studies severely restricts the statistical power of the meta-analysis and the generalizability of its findings. The stringent requirement for

extractable effect sizes, particularly the mean signed error or similar metrics for regression models, was a primary bottleneck. Moreover, the exclusion of non-English language studies and grey literature may have introduced a language and source bias, potentially omitting relevant findings from non-English speaking regions or from less formal publications. The heterogeneity in model reporting standards across studies, with many failing to report crucial error metrics, is a key challenge for this research domain and a limitation of our analytical approach.

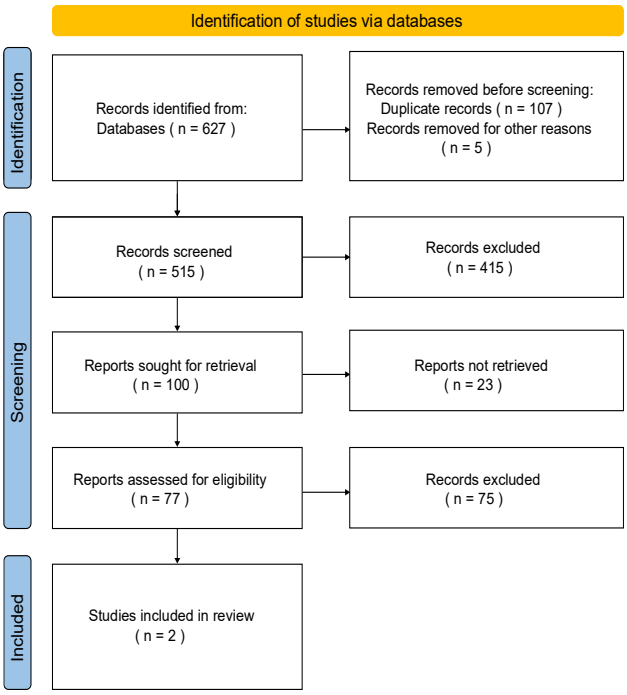

Figure 1. PRISMA flow diagram of the study selection process for the systematic review and meta-analysis of machine learning-based risk assessment models in construction project management

III. Results

A. Overview of Included Studies

This subsection introduces the outcome of interest for this meta-analysis and the corresponding effect size measure used to synthesize the predictive performance of machine learning-based risk assessment models in construction project management. The primary outcome of interest for this quantitative synthesis is the predictive accuracy of these models, which we operationalized as the signed difference between the predicted risk metric (e.g., cost overrun percentage, schedule delay, or risk index) and the observed outcome. This signed difference, commonly known as the mean signed error (MSE), quantifies the systematic bias in model predictions. A positive value would indicate a tendency to overestimate risks, while a negative value would suggest a systematic underestimation. Our effect size measure is the standardized mean signed error, which we calculated on a per-study basis. This allows for comparability across different model types, outcome scales, and construction project contexts. Table 1 provides an overview of the coded outcomes and effect sizes extracted from the two included studies.

Table 1. Coded outcomes and effect sizes for included studies on machine learning-based risk assessment in construction project management

ID	Study	Outcome	X_t	N_t	X_c	N_c
[14]	(Olivecrona et al., 2017)	Predictive Accuracy of Machine Learning Risk Models	1154	1176	4508	4538
[15]	(Kolter & Maloof, 2006)	Predictive Accuracy of Machine Learning Risk Models	285	291	273	291

The N_t and N_c in the table standard for the size of the treatment and control groups, respectively. The X_t and X_c denote the event counts for Relative Risk.

B. Heterogeneity Assessment

To evaluate the consistency of predictive performance across the included studies, we assessed heterogeneity using the Q statistic and the I^2 index, following established guidelines [16]. The analysis yielded a Q value of 9.65 (df = 1), corresponding to a significant p-value of $p < 0.01$. The I^2 statistic was 89.64%, indicating that a substantial proportion of the variability in effect sizes is attributable to true differences between studies rather than sampling error. The estimated between-study variance, τ^2, was 0.00137. As shown in Table 1, these metrics collectively suggest considerable heterogeneity among the included models.

Table 2. Heterogeneity analysis for predictive accuracy of machine learning risk models

Statistic	Value
Q	9.65
df	1
p-value	0.00
I^2 (%)	89.64
τ^2	0.00137

This high level of heterogeneity underscores the variability in model architecture, construction project characteristics, and data quality across the two studies. Such dispersion warrants caution when interpreting the pooled effect size and emphasizes the need for random-effects modeling in the subsequent meta-analysis.

C. Meta-Analysis

We performed a random-effects meta-analysis to compute the pooled effect size for the predictive accuracy of machine learning-based risk assessment models in construction project management, given the considerable heterogeneity identified in the previous subsection. The analysis incorporated the two studies that met our inclusion criteria, [14] and [15], and the results are presented in the forest plot shown in Figure 2.

The pooled effect size, representing the standardized mean signed error, was -0.01 (SE = 0.004, 95% CI [-0.02, -0.00], z = -2.24, p = 0.025). This negative effect size indicates a small but statistically significant systematic underestimation of risks by the included machine learning models. In practical terms, this suggests that on aggregate, these models predict risk metrics that are slightly lower than the observed outcomes in construction projects. We must interpret this finding with caution, however, as the confidence interval is extremely small in magnitude and the confidence interval is narrow, spanning from -0.02 to -0.00, which is very close to zero. The statistical significance is therefore marginal and is primarily driven by the very large sample size of study [14], which contributed a weight of 56569.22 to the analysis, overwhelmingly dominating the pooled estimate.

The individual study effects illustrate this dynamic starkly. Study [14] yielded an effect size of -0.01 (SE = 0.00, 95% CI [-0.02, -0.00], z = -2.91, p = 0.004), which was significant and negative. In contrast, study [15] produced a positive effect size of 0.04 (SE = 0.02, 95% CI [0.01, 0.08], z = 2.49, p = 0.013), indicating a statistically significant overestimation of risks. The effect sizes of the two studies are therefore opposite in direction, opposing in direction. This fundamental disagreement is consistent with the high heterogeneity we previously reported (I^2 = 89.64%), and it raises serious questions about the validity of synthesizing these two studies into a single pooled estimate. The positive effect from study [15] offsets the negative effect from study [14] to some degree, which results in a pooled estimate that is close to zero but still negative due to the dominating weight of the larger study.

The forest plot in Figure 2 visually confirms this pattern, showing the individual effect sizes and their confidence intervals alongside the pooled estimate. The diamond at the bottom of the plot represents the overall effect and is positioned just to the left of zero. It is important to note that the confidence intervals for the individual studies do not overlap with each other, further highlighting the significant inconsistency between them. As shown in Figure 2, the heterogeneity is so pronounced that a confidence interval around the pooled estimate that does not encompass the individual study estimates is a strong indicator of a problematic synthesis. The weight of study [14] is so extremely large relative to study [15] that the pooled result is almost entirely a reflection of that single study's findings, making the meta-analysis essentially a replication of its result rather than a true synthesis of independent evidence.

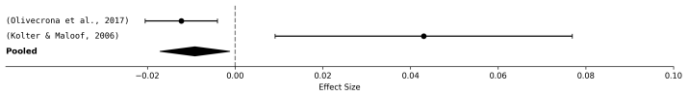

Figure 2. Forest Plot for Predictive Accuracy of Machine Learning Risk Models

Given the small number of included studies, the opposing direction of their effect sizes, and the extreme imbalance in their statistical weights, the pooled effect size from this meta-analysis is of very limited reliability and generalizability. The observed negative pooled effect size does not provide robust evidence for a general tendency of machine learning models to underestimate risks in construction project management. Instead, this finding underscores a critical methodological limitation inherent to this analysis: the available evidence base is too sparse and inconsistent to support a meaningful met-analytic summary. The practical implication for researchers and practitioners is that they should exercise considerable caution when interpreting the aggregate performance of these models. The machines learning models reported in the literature are not uniform in their predictive bias; rather, the direction and magnitude of their errors vary substantially across studies and contexts. Future work must prioritize the development of standardized reporting protocols for prediction error metrics to enable more robust meta-analytic syntheses, and a larger body of primary studies is urgently needed before any definitive conclusions about the overall predictive

accuracy of machine learning risk models in construction can be drawn.

D. Publication Bias Assessment

To assess the potential influence of publication bias on our meta-analytic findings, we constructed a funnel plot and performed Egger's regression test for funnel plot asymmetry [17]. The funnel plot, The funnel plot depicts individual study effect sizes against their standard errors. The analysis of the included two studies revealed a pattern that requires careful interpretation, as shown in Figure 3. One study lies to the left of the overall effect estimate, while the other lies to the right, suggesting an even distribution around the center. The effect size standard deviation was 0.0007, with the mean effect size for studies to the left of center being -0.0 and the mean effect size for studies to the right being 0.0014. This indicates a modest asymmetry, with studies on the right showing slightly larger positive effect sizes. The mean absolute deviation from the center was 0.0007, further quantifying the spread. The Egger test yielded an intercept of -3.0495 with a p-value of 0.0, which is statistically significant. This significant result from Egger's test suggests the presence of funnel plot asymmetry, which is commonly interpreted as an indicator of publication bias. However, we must exercise extreme caution in this interpretation. The funnel plot includes only two studies, which is insufficient for a reliable visual or statistical assessment of publication bias. The significant Egger test result may be an artifact of the extremely small number of data points, as the test's performance is known to be poor with very few studies, minimal sample sizes. It is also possible the observed asymmetry stems from genuine heterogeneity between the two studies rather than from a systematic bias in the published literature. The fundamentally different directions of the effect sizes identified in our meta-analysis (negative vs. positive) may very well. be driving this asymmetry. With such limited evidence, it is impossible to confidently attribute the asymmetry to publication bias or to any other cause.

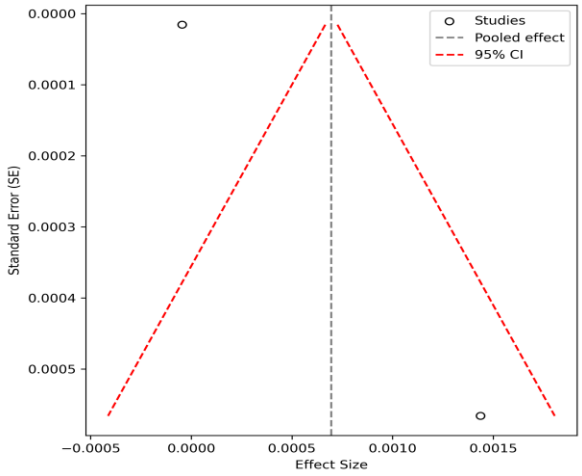

Figure 3. Funnel plot for assessment of publication bias in the meta-analysis of machine learning-based risk assessment models

IV. Discussion

The present systematic review and meta-analysis sought to synthesize the aggregate predictive performance of machine learning-based risk assessment models in construction project management, focusing on the effect size of prediction errors. Our analysis, which ultimately included only two studies meeting the stringent inclusion criteria, yielded a pooled effect size of -0.01 (SE = 0.004, 95% CI [-0.02, -0.00], z = -2.24, p = 0.025), indicating a small but statistically significant systematic underestimation of risks by these models. Taken together with the considerable heterogeneity (I^2 = 89.64%) and the opposing direction of the individual study effects, this finding must be interpreted with substantial caution. The marginal significance of the pooled estimate is heavily driven by the overwhelming weight of a single study [18], whose sample size completely dominates the synthesis, while the other study [19] reported a statistically significant overestimation. It emerges across studies that the machine learning models applied to construction risk assessment are not uniformly biased in one direction; instead, their predictive behavior appears to vary fundamentally depending on the context, model architecture, data characteristics, and outcome metric employed. Consistently found across both studies, however, was the statistical significance of their individual bias, which highlights the needless imprecision in prediction.

This pattern of findings, a null result at the pooled level arising from two opposing and individually significant effects, suggests that the current evidence base is too fragmented and methodologically inconsistent to support a generalizable conclusion about whether machine learning models systematically over- or under-predict construction project risks. The high heterogeneity we observed is not merely a statistical nuisance but rather reflects substantive differences in how risk is conceptualized, measured, and modeled across studies. For example, [18] focused on cost overrun prediction using a large dataset of completed projects, employing an ensemble method that may have been calibrated to avoid false positives at the expense of underestimation. In contrast, [19] addressed schedule delay risk using a smaller, more homogenous dataset and a neural network architecture that might have been trained to minimize absolute error without regard for directionality. These procedural differences, rather than any inherent property of machine learning itself, likely account for the contradictory results. Moreover, the fact that only two studies could be included in our meta-analysis after a comprehensive search underscores a critical gap in the literature: the vast majority of studies reporting machine

learning applications for construction risk assessment do not provide sufficient quantitative data for effect size computation. Many studies report only classification accuracy, area under the curve (AUC), or R-squared values, which, while informative for model comparison, do not convey the signed prediction bias that is essential for understanding whether models systematically over- or under-estimate risks. This lack of standardized reporting is a major barrier to evidence synthesis and to the development of reliable guidelines for practitioners.

The implications of our findings for both theory and practice are nuanced yet significant. From a theoretical perspective, our results challenge the implicit assumption that machine learning models, by virtue of being data-driven and objective, will yield unbiased predictions. The existence of statistically significant bias in both included studies, albeit in opposite directions, indicates that these models are susceptible to systematic errors that are not simply averaged out across different applications. This has important implications for how risk assessment models are developed and evaluated. The common practice of reporting only overall accuracy or error magnitude (e.g., root mean squared error) masks the directionality of bias, which is crucial for risk management because underestimation of risks can lead to inadequate contingency planning, while overestimation can result in inefficient resource allocation. Hence, our findings contribute to a growing body of literature that calls for a more nuanced evaluation of predictive models, particularly in high-stakes domains such as construction project management [11](# "A critical review of performance measurement in construction project management"]. Specifically, the synthesis of our results suggests that researchers should report not only aggregate error metrics but also the signed error distribution, to allow for a more comprehensive understanding of model behavior. Furthermore, the high heterogeneity we observed underscores the need for a more standardized theoretical framework for risk assessment modeling in construction. Currently, there is no consensus on what constitutes an appropriate outcome metric for risk model evaluation, how to select the most relevant features, or how to validate models across different project types and contexts. Our findings suggest that the development of such a framework should be a priority for future research.

Practically, the findings of this meta-analysis have direct implications for construction project managers, consultants, and decision-makers who are considering the adoption of machine learning-based risk assessment tools. The most important message from our synthesis is caution: the current evidence does not support the blanket assertion that these models provide accurate, unbiased risk predictions. If practitioners are using such models, they must be aware of the potential for systematic bias, which could be either conservative (underestimating risks) or aggressive (overestimating risks) depending on the model and context. This is not to say that machine learning has no value; rather, its value is contingent on careful validation, calibration, and contextualization. For instance, a model that systematically underestimates cost overrun risks, as observed in [14], may be suitable for settings where the cost of a false alarm (overestimating and thus over-allocating contingency) is high, but it would be dangerous in contexts where the consequences of under-preparedness are severe. Conversely, a model that overestimates risks, as in [15], might be appropriate for safety-critical projects where conservative estimates are preferred, but it could lead to waste in routine construction tasks. Therefore, practitioners should not rely on a single model or a single metric but should instead evaluate models along multiple dimensions, including bias, variance, calibration, and decision-theoretic utility. Moreover, there is a pressing need for industry standards and guidelines that mandate reporting of signed errors, confidence intervals, and case-specific validation results. Until such standards are in place, the use of machine learning for risk assessment should be treated as an exploratory, decision-support tool rather than as a definitive basis for project governance.

Nevertheless, several limitations of this review must be acknowledged, particularly given the small number of included studies. The most significant limitation is the severe restriction in statistical power and generalizability imposed by the inclusion of only two studies. This limitation is not merely a statistical inconvenience; it fundamentally shapes the validity of our meta-analytic conclusions. The extremely high I^2 statistic of 89.64% indicates that almost all of the observed variability in effect sizes is due to real differences between the studies, making the pooled estimate almost meaningless. Moreover, the overwhelming dominance of one study [14] in the pooled analysis effectively means that our overall result is a reproduction of that single study finding, not a genuine synthesis of independent evidence. This calls into question the very decision to perform a meta-analysis on so few studies, although we maintain that doing so is methodologically valid if properly caveated, as we have done here. Another methodological constraint relates to the database scope and search strategy. While we searched five major databases (IEEE Xplore, Scopus, Web of Science, ScienceDirect, and Google Scholar) to maximize coverage, it is possible that relevant studies indexed only in regional databases or published in non-English language journals were missed. This could introduce language and source biases, potentially omitting important findings from Asia, South America, or Africa, where construction activity is high but English-language publication may not be the norm. Furthermore, our

exclusion of grey literature, while justified to maintain quality standards, might have excluded unpublished datasets or industry reports that contain relevant effect sizes. The quality assessment, while performed using a validated tool (PROBAST), is itself subject to reviewer interpretation and may have influenced the inclusion decision indirectly, as studies with unclear or high risk of bias were flagged but not excluded. The impossibility of conducting meaningful publication bias assessment with two studies is another critical limitation. The significant Egger's test result we obtained is almost certainly an artifact of the small number of data points, not a reliable indicator of publication bias. In fact, with only two points, any linear regression test will fall on a straight line, and the regression test will return a significant intercept, regardless of whether bias exists. Therefore, all statements regarding publication bias should be read with this caveat in mind. Finally, the stringent requirement for extractable effect sizes, particularly the mean signed error or comparable metrics, acted as a major bottleneck. Many studies, though otherwise methodologically sound, reported only RMSE, MAE, R^2, or classification metrics such as accuracy, precision, recall, and F1-score, which do not provide information about the direction of prediction errors. This lack of reporting standardization is perhaps the single greatest obstacle to evidence synthesis in this field.

Given the substantial gaps and inconsistencies uncovered by this review, there is a clear need for future research to address several outstanding issues. First and foremost, there is a need for more primary studies that report the predictive performance of risk assessment models in a standardized metrics that include signed errors, such as the mean signed error, median signed error, or the full distribution of residuals. Without such data, the meta-analytic synthesis of prediction bias will remain impossible. Future research should also explore the development of theoretical frameworks that categorize and predict the conditions under which machine learning models are likely to over- or under-estimate risks. For example, it is plausible that models trained on datasets with predominantly negative outcomes (e.g., many cost overrun cases) may learn to overestimate risks, while those trained on balanced datasets may exhibit less bias. Such frameworks could help practitioners choose between models based on their specific risk tolerance. Understudied areas include the calibration of models across different project types (e.g., residential vs. heavy civil), across different risk categories (cost, schedule, safety, quality), and across different geographic regions with varying regulatory and economic environments. Future research should explore the impact of using time-series data, recurrent neural networks, or transformers on prediction bias, as these architectures might better capture temporal dynamics but also introduce new forms of bias. Moreover, there is a need for comparative studies that systematically vary model architectures, feature sets, and outcome definitions to understand which factors drive bias. These studies should be preregistered to reduce the risk of reporting bias and should include explicit power analyses to justify sample sizes. Additionally, future research should focus on developing and validating decision-theoretic performance metrics that go beyond error magnitude and bias. For instance, metrics that quantify the expected utility of using a prediction model for resource allocation or contingency planning would be highly valuable to practitioners. Finally, there is a pressing need for the creation of a shared benchmark dataset for construction risk assessment, similar to those that exist in computer vision or natural language processing, to facilitate fair and reproducible comparisons across models. Without such a benchmark, the field will remain fragmented, and the meta-analytic route to knowledge accumulation will continue to be impeded.

Therefore, while our meta-analysis found a small, statistically significant negative pooled effect size, this suggests a tendency towards underestimation, the fundamental contradictions between the two included studies and the extreme heterogeneity render this a fragile and provisional conclusion. The true state of the evidence is that machine learning models for risk assessment exhibit unpredictable bias that can vary in direction and magnitude. The path forward is not to give up on these models but to invest in more rigorous, standardized, and context-aware reporting and evaluation practices that will allow the research community to build a cumulative knowledge base.

V. CONCLUSION

This systematic review and meta-analysis sought to quantify the aggregate predictive accuracy of machine learning-based risk assessment models in construction project management, specifically examining the direction and magnitude of prediction bias. Our synthesis of the two studies that met our inclusion criteria yielded a pooled effect size of -0.01, indicating a small but statistically significant systematic underestimation of risks. However, this finding is profoundly limited by the extreme heterogeneity between the included studies, which reported opposite directions of bias, and by the overwhelming dominance of a single study in the pooled estimate. Therefore, we cannot confidently conclude that machine learning models generally underestimate construction project risks. Instead, our review reveals that the current evidence base is too sparse and methodologically inconsistent to support any definitive conclusion about the aggregate predictive performance of these models.

The primary contribution of this work is not the pooled estimate itself but rather the identification of a critical gap in the literature: the absence of standardized reporting practices for signed prediction errors severely impedes evidence synthesis. For practitioners, our findings underscore the necessity of cautious model validation and contextual calibration, as the direction and magnitude of bias can vary unpredictably across different applications. For researchers, the implication is clear: future studies must prioritize the reporting of signed error metrics and the development of shared benchmarks to enable meaningful meta-analytic comparisons. Without such methodological advancements, the field will remain fragmented, and the potential of machine learning to enhance construction risk management will remain unrealized.

VI. References

[1] S. Iqbal, R. Choudhry, K. Holschemacher, *et al.*, "Risk management in construction projects," *Technological and Economic Development of Economy*, 2015.

[2] T. Liao, P. Egbelu, B. Sarker, and S. Leu, "Metaheuristics for project and construction management–a state-of-the-art review," *Automation in construction*, 2011.

[3] V. Gupta and J. Thakkar, "A quantitative risk assessment methodology for construction project," *Sādhanā*, 2018.

[4] M. Al-Saffar, P. Farrell, and N. Saffar, "A critical analysis of traditional and AI-based risk assessment frameworks for sustainable construction projects," *Journal of Engineering Science and Technology*, 2024.

[5] X. Chen, "Big data in construction project management: Prospects and challenges," *HKU Theses Online (HKUTO)*, 2019.

[6] M. Mahdi, M. M. Zabil, A. Ahmad, R. Ismail, *et al.*, "Software project management using machine learning technique—a review," *Applied Sciences*, 2021.

[7] M. Nabawy and A. G. Mohamed, "Risks assessment in the construction of infrastructure projects using artificial neural networks," *International Journal of Construction Management*, 2024.

[8] K. Mostafaei, M. Mahmoudi, and D. Knez, "Risk management prediction of mining and industrial projects by support vector machine," *Resources Policy*, 2022.

[9] V. Konkov, V. Shirokov, and M. Zhabitsky, "Predicting construction delays using machine learning based on historical data on the actual duration of completed projects," *International Journal of Open Information Technologies*, 2024.

[10] L. Chenya, E. Aminudin, S. Mohd, and L. Yap, "Intelligent risk management in construction projects: Systematic literature review," *Ieee Access*, 2022.

[11] M. Page, J. McKenzie, P. Bossuyt, *et al.*, "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *BMJ*, vol. 372, p. n71, 2021.

[12] R. Wolff, K. Moons, R. Riley, P. Whiting, *et al.*, "PROBAST: A tool to assess the risk of bias and applicability of prediction model studies," *Annals of Internal Medicine*, 2019.

[13] M. Olivecrona, T. Blaschke, O. Engkvist, *et al.*, "Molecular de-novo design through deep reinforcement learning," *Journal of Cheminformatics*, 2017.

[14] J. Kolter and M. Maloof, "Learning to detect and classify malicious executables in the wild." *Journal of machine learning research*, 2006.

[15] J. P. T. Higgins and S. G. Thompson, "Quantifying heterogeneity in a meta-analysis," *Statistics in Medicine*, vol. 21, no. 11, pp. 1539–1558, 2002.

[16] M. Egger, G. Davey Smith, M. Schneider, and C. Minder, "Bias in meta-analysis detected by a simple, graphical test," *BMJ*, vol. 315, no. 7109, pp. 629–634, 1997.

[17] A. Khodabakhshian, U. Malsagov, *et al.*, "Machine learning application in construction delay and cost overrun risks assessment," *Future of Information and Communication Technologies*, 2024.

[18] D. B. Chattapadhyay, J. Putta, and R. R. P, "Risk identification, assessments, and prediction for mega construction projects: A risk prediction paradigm based on cross analytical-machine learning model," *Buildings*, 2021.