Coupled Systems Mechanics, *Vol. 15, No. 1 (2026) 61-82*
https://doi.org/10.12989/csm.2026.15.1.061

Techno-Press

A machine learning framework for smart labor forecasting in PEB: integrating classification and profitability analysis

Ringle Raja*[1], Hemalatha[1], Elizabeth Amudhini Stephen[2], Athish[3], Charles Climent Fliex[4a]

[1]*Department of Civil Engineering, Karunya Institute of Technology and Sciences, Coimbatore, India*
[2]*Department of Mathematics, Karunya Institute of Technology and Sciences, Coimbatore, India*
[3]*Department of AI and ML Karunya Institute of Technology and Sciences, Coimbatore, India*
[4]*Elshaddai Engineering Private Limited, Chennai, India*

(Received June 9, 2025, Revised January 15, 2026, Accepted March 25, 2026)

Abstract. Efficient labor forecasting remains a critical yet underexplored challenge in industrialized construction, particularly in Pre-Engineered Building (PEB) fabrication environments characterized by repetitive workflows, trade-specific labor dynamics, and cost-sensitive schedules. This study proposes a classification-enhanced machine learning framework that integrates Random Forest regression with real-time biometric labor data to predict workforce requirements. Projects are systematically classified using fixed thresholds for labor intensity and variability, yielding behaviorally distinct groups that improve model specialization and reduce forecast variance. Dedicated Random Forest models are trained for each classification group, leveraging structured biometric attendance logs to ensure input data fidelity. Model performance is assessed using RMSE and MSE metrics, while a profitability-based evaluation quantifies financial outcomes associated with prediction deviations. Experimental results show that over 88% of forecasts fall within an acceptable ±5% tolerance range, and over 91% of cases result in profit-positive deployment decisions. The framework demonstrates superior accuracy for stable labor types and high adaptability across varying project complexities. By unifying behavioral classification, biometric integration, and financial assessment, this study delivers a robust, scalable forecasting approach for data-driven workforce planning in modular construction systems.

Keywords: biometric data; classification; labor forecasting; machine learning in construction; Pre-Engineered Building (PEB); profitability analysis; random forest

1. Introduction

Labor forecasting in industrialized construction, particularly in Pre-Engineered Building (PEB) fabrication, plays a pivotal role in ensuring optimal resource utilization, cost efficiency, and on-time project delivery. Unlike conventional site-based construction, PEB environments follow modular, factory-driven workflows where labor intensity fluctuates with component complexity, fabrication sequencing, and project-specific production plans. Traditional labor estimation approaches—such as historical trend analysis or rule-of-thumb calculations—lack the ability to

*Corresponding author, Ph.D., E-mail: ringleraja@karunya.edu.in
[a] Director

ISSN: 2234-2184 (Print), 2234-2192 (Online)

accommodate the dynamic, non-linear patterns of modern industrial operations, leading to inefficiencies in both operational planning and financial forecasting.

With the increased digitization of construction workflows, Machine Learning (ML) has emerged as a promising approach for modeling labor demand using real-time and historical data. However, forecasting labor in industrialized environments remains complex due to variations in fabrication sequences, trade-level deployment, and behavioral heterogeneity across projects. In particular, generic prediction models often struggle to account for workforce behavior variations stemming from project intensity and execution patterns.

While prior studies have successfully applied ML techniques to predict construction resource usage, most focus on on-site scenarios or generalized frameworks. There remains a significant research gap in adapting ML-driven forecasting specifically to PEB environments—where tasks are semi-repetitive, structured, and backed by real-time labor attendance data. Furthermore, very few studies consider the economic impact of prediction errors, despite their direct implications on project profitability.

To address these gaps, this study proposes a classification-driven labor forecasting framework for PEB fabrication, built upon real-time biometric attendance data. Projects are segmented based on labor intensity and variability, forming three behaviorally distinct classes. For each class, a Random Forest Regressor (RFR) is trained to predict trade-level labor demands across ten predefined job categories. The model's performance is evaluated using RMSE and MSE, with additional analysis on the profitability implications of forecast deviations.

This study makes the following key contributions:

- Develops a classification-aware forecasting model tailored to the behavior of labor deployment in PEB projects.
- Integrates real-time biometric attendance data to enhance input fidelity and forecasting responsiveness.
- Introduces a profitability-based evaluation of forecast deviations, linking predictive accuracy to economic outcomes.
- Demonstrates the framework's effectiveness across multiple project types with varying operational characteristics.

This research offers a scalable, data-driven tool for workforce prediction in modular construction, combining predictive precision with operational and financial relevance.

2. Literature review

2.1 Machine learning for labor and resource forecasting

In the evolution of construction automation, machine learning (ML) has gained substantial prominence for forecasting labor requirements, resource allocation, cost, and project duration. Ensemble learning models such as Random Forest (RF), Gradient Boosting, and XGBoost have emerged as leading algorithms due to their ability to handle non-linear data distributions and extract complex patterns from noisy datasets [12, 13, 20, 32]. In labor forecasting specifically, Random Forest models have shown high predictive performance in structured fabrication environments [9], while XGBoost has delivered reduced generalization error in repetitive workflows [4, 31].

Artificial Neural Networks (ANNs), Support Vector Regression (SVR), and hybrid ML models

have also been deployed. Raju et al. [22] implemented an ANN for predicting site workforce in steel structures, while Xia et al. [33] combined SVR with fuzzy logic for dynamic workforce scheduling. Sun et al. [28] provided a comparative assessment of tree-based and neural models, finding that ensemble models maintain stability even in high-dimensional and unbalanced labor datasets.

Several studies have explored the application of Artificial Neural Networks (ANN) for estimating construction productivity. For instance, Haddad [8] developed an ANN-based model for predicting labor productivity in foundation works by incorporating multiple influencing factors such as labor experience, material availability, and site conditions. The study demonstrated that ANN models can effectively capture nonlinear relationships between influencing variables and productivity outcomes, achieving high prediction accuracy. However, such models are typically limited to single-stage activity-level productivity estimation and do not address multi-stage fabrication processes or project-level labor forecasting complexities.

2.2 Classification-based forecasting and grouping strategies

ML model generalization is often hindered by heterogeneous datasets, which has prompted interest in classification-based project grouping. Segmenting data based on labor intensity, component type, or fabrication complexity can significantly enhance forecasting accuracy by reducing intra-group variance. Zermane et al. [37] introduced a behavioral classification system to separate fabrication stages for steel structures. You and Chen [34] used labor variability patterns to create behaviorally homogeneous project clusters. Alemão et al. [2], Yusoff [35] reinforced that project classification before model training enables more context-aware and interpretable learning.

Recent developments in multi-tier grouping methods further validate the value of classification. Salihu et al. [25] presented a two-stage classification pipeline combining project scale and activity sequencing. Tao et al. [30] suggested that grouped model training results in superior convergence and lower forecasting variance. Cheng et al. [7] introduced component clustering based on geometry and processing time, which improved early-stage workforce estimation. Despite these advances, most models still treat classification as a preprocessing step rather than embedding it into the predictive framework.

Construction projects are inherently complex due to fragmented supply chains involving multiple stakeholders, processes, and material flows. Cheng et al. [16] highlighted that construction supply chains consist of numerous interconnected entities, where inefficiencies in coordination, procurement, and delivery significantly affect project performance. The use of structured frameworks such as the Supply Chain Operations Reference (SCOR) model enables better understanding of process interdependencies, bottlenecks, and workflow dynamics. However, existing studies primarily focus on supply chain modeling and monitoring rather than predictive labor or fabrication duration forecasting.

2.3 Real-time, biometric, and sensor data integration

The advent of real-time sensing, biometric monitoring, and IoT technologies has opened new possibilities for accurate labor tracking and forecasting. Hanane et al. [36] utilized biometric attendance data in PEB environments to measure labor utilization rates. Reuter and Brambring [23] combined RFID signals and camera-based tracking to build adaptive crew deployment forecasts. Leo [5] implemented facial recognition in modular construction, achieving real-time

workforce mapping at high spatial and temporal resolution.

Other studies extended these insights. Héctor et al. [6] proposed integrating sensor data directly into regression pipelines. Shaheen et al. [26], Tang et al. [29] introduced wearable-based datasets for identifying worker-task associations in production lines. These innovations demonstrate the potential of biometric and real-time inputs not just for monitoring, but also for proactive labor forecasting—an area underutilized in most factory-oriented construction systems.

2.4 Economic and profitability-based evaluation of forecasts

Accurate labor forecasting holds direct implications for project financial outcomes. While conventional evaluation relies on metrics like RMSE and MAE, a growing number of studies call for economic evaluation frameworks. Gopal and Murali [21] illustrated how labor misprediction led to significant project overruns, both in cost and timeline. Shinde and Shah [27] highlighted the financial impact of idle labor arising from forecast overestimation.

Profitability-driven modeling has been proposed as an alternative. Li et al. [18] developed a zone-based classification of forecast errors (profitable, tolerable, risky), while Guo et al. [11] embedded cost-awareness into the loss function. Fahle et al. [10] introduced a margin-aware optimization layer, which evaluated the sensitivity of labor forecasts to unit wage costs and time delays. These studies advocate for prediction pipelines that not only maximize accuracy but also minimize financial deviation.

2.5 Application in modular and factory-based construction

Factory-oriented construction workflows such as PEB, modular housing, and precast systems have received growing attention in ML literature. Sadatnya et al. [24] showed that project sequencing in PEB could be used to model expected workforce phases with high regularity. Kang et al. [17] applied a predictive model to panelized housing and captured task wise workforce patterns in standardized production lines. Shinde and Shah [27] developed regression trees for factory-controlled labor analysis, showing stable learning in fixed-cycle fabrication.

Artificial Neural Networks have been widely adopted for modeling complex, nonlinear relationships in engineering and manufacturing systems. Waqar [1], Hobiny et al. [19] demonstrated the effectiveness of ANN in forecasting raw material inventory by capturing dynamic interactions among demand, lead time, supplier reliability, and cost parameters. The study emphasized the ability of ANN models to generalize unseen data and handle fluctuating input variables without requiring explicit mathematical formulations. This capability is particularly relevant to fabrication environments, where multiple interdependent factors influence labor requirements and production duration.

Hwang et al. [14], Imam et al. [15] further explored smart production scheduling using ML in industrial settings. These works emphasized that modular environments, with predictable task durations and structured data collection—are ideal candidates for classification-enhanced forecasting models.

Beyond construction-focused studies, several works in applied mathematics and continuum mechanics also demonstrate how advanced computational modeling and numerical simulation support complex engineering decision-making in thermoelastic and thermofluidic systems, including hyperbolic two-temperature formulations in semiconductor media and EMHD hybrid nanofluid analyses for geothermal applications [3, 19].

2.6 Summary and research gaps

From this comprehensive review, it is evident that the construction ML literature has made strong progress in ensemble modeling, classification-based segmentation, and real-time data integration. However, key research gaps remain:

- First, most studies treat classification as a preprocessing step and do not embed it structurally in the ML pipeline [13, 35].
- Second, while biometric data is increasingly collected, few works use it in multi-output forecasting frameworks [5, 29, 36].
- Third, profit-aware forecasting models remain rare, despite growing recognition of cost-sensitivity in labor planning [10, 11, 18].

This study addresses these gaps by proposing a classification-driven labor forecasting model, integrating Random Forest regression with real-time biometric labor data. Projects are segmented by labor intensity and variability into behaviorally coherent groups. Trade-wise labor forecasts are then generated, with evaluation based on both accuracy metrics and profitability zones. The model is validated in a PEB fabrication environment, offering a unified, scalable solution for economically informed workforce forecasting.

3. Data collection and preprocessing

3.1 Real-time labor attendance data acquisition

Data for this study were collected from a large-scale PEB fabrication unit over a defined operational period. The primary source of information was the factory's face recognition-based biometric attendance system, which records the daily in-time and out-time for each worker. These data were automatically transmitted to a custom-built Google Sheets database, referred to as the EveryDay Worker Attendance (EDWA) sheet, designed for streamlined real-time data storage and retrieval.

The EDWA sheet captures the following fields:

1. Date
2. Labor ID
3. Start Time / End Time
4. Job Code Assignment
5. Lunch Break Hours (entered manually)
6. Overtime Hours (OT) (auto-calculated)
7. Designation ID (auto-extracted based on predefined roles)

The job role of each laborer is predefined at the time of hiring. As such, the system maps each Labor ID to a Designation ID (L1 to L10) corresponding to job types including CO_2 Welder (L1), Fabricator (L2), Fitter (L3), Gas Cutter (L4), Grinder (L5), Helper (L6), Hydra Operator (L7), Rigger (L8), Semi Fitter (L9), and Welder (L10).

3.2 Job-wise and project wise dataset structuring

From the EDWA sheet, the dataset was further refined into two key reporting formats:

1. Job-Wise Daily Worker (JWDW) Report

Table 1. Every Day Worker Attendance Sheet (EDWA)-25,487 entries

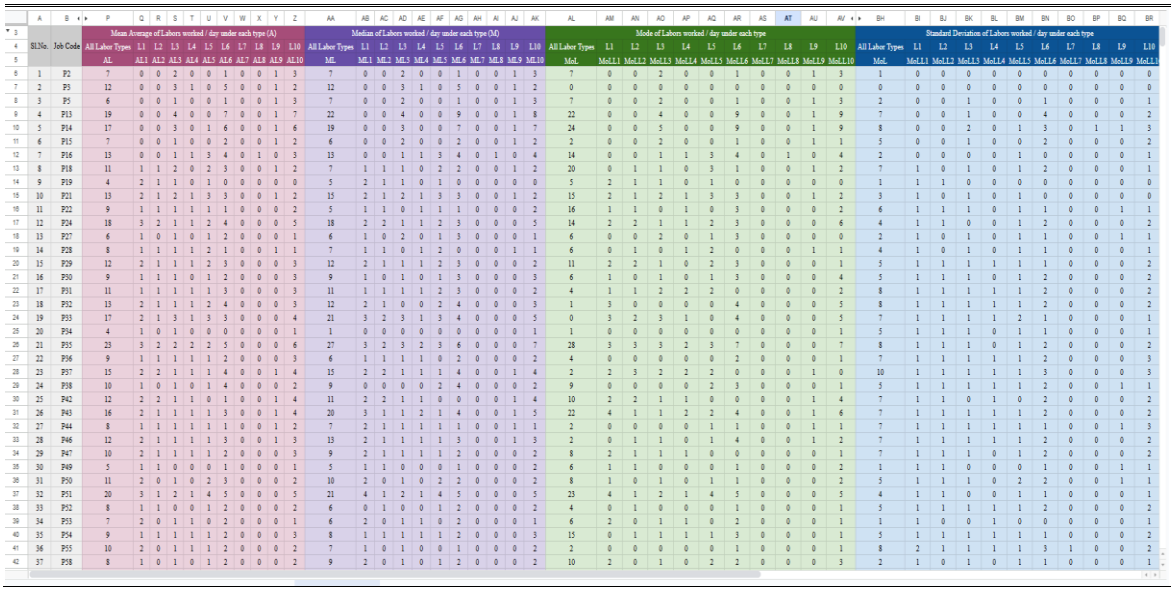
								OT Permitted	Permitted Hours	OT Permitted	Total OT
								2:00	8:00	2:00	10:00
Date	Job No.	Name	Design ID	Designation	Contractor	Start Time	Lunch Hours	End Time	Total hours	Labor ID	OT Hours
18/07/2021	P5	Shailesh	L10	Welder	Ramesh	8:00	0:00	13:00	5:00	L31	
18/07/2021	P5	Pradeep Kumar	L6	Helper	Ramesh	8:00	0:00	13:00	5:00	L28	
19/07/2021	P5	Balindra Manjhi	L10	Welder	Ramesh	8:00	1:00	21:00	12:00	L22	4:00
19/07/2021	P5	Sri Kant Kumar	L6	Helper	Ramesh	8:00	1:00	21:00	12:00	L32	4:00
19/07/2021	P5	Manejar Mahto	L9	Semi Fitter	Ramesh	8:00	1:00	21:00	12:00	L26	4:00
19/07/2021	P5	Basant P	L3	Fitter	Ramesh	8:00	1:00	21:00	12:00	L23	4:00
19/07/2021	P5	Ravindra Kumar	L3	Fitter	Ramesh	8:00	1:00	21:00	12:00	L30	4:00
19/07/2021	P5	Chhotelal Yadav	L10	Welder	Ramesh	8:00	1:00	21:00	12:00	L24	4:00
19/07/2021	P5	Shailesh	L10	Welder	Ramesh	8:00	1:00	21:00	12:00	L31	4:00
19/07/2021	P5	Pradeep Kumar	L6	Helper	Ramesh	8:00	1:00	21:00	12:00	L28	4:00
20/07/2021	P3	Sri Kant Kumar	L6	Helper	Ramesh	8:00	1:00	20:00	11:00	L32	3:00
20/07/2021	P3	Manejar Mahto	L9	Semi Fitter	Ramesh	8:00	1:00	20:00	11:00	L26	3:00
20/07/2021	P3	Basant P	L3	Fitter	Ramesh	8:00	1:00	20:00	11:00	L23	3:00
20/07/2021	P3	Ravindra Kumar	L3	Fitter	Ramesh	8:00	1:00	20:00	11:00	L30	3:00
20/07/2021	P3	Chhotelal Yadav	L10	Welder	Ramesh	8:00	1:00	20:00	11:00	L24	3:00
20/07/2021	P3	Pradeep Kumar	L6	Helper	Ramesh	8:00	1:00	20:00	11:00	L28	3:00
20/07/2021	P3	Jitendar	L3	Fitter	Hanumanth	8:00	1:00	20:00	11:00	L10	3:00
20/07/2021	P3	Arun	L4	Gas Cutter	Hanumanth	8:00	1:00	20:00	11:00	L4	3:00
20/07/2021	P3	Paranjith	L6	Helper	Hanumanth	8:00	1:00	20:00	11:00	L12	3:00
20/07/2021	P3	Sujith	L6	Helper	Hanumanth	8:00	1:00	20:00	11:00	L20	3:00
20/07/2021	P3	Abhishek	L10	Welder	Hanumanth	8:00	1:00	20:00	11:00	L2	3:00
20/07/2021	P3	Ram	L6	Helper	Hanumanth	8:00	1:00	20:00	11:00	L15	3:00
20/07/2021	P5	Balindra Manjhi	L10	Welder	Ramesh	8:00	1:00	20:00	11:00	L22	3:00

Table 1 illustrates a sample EDWA template where the data entry is color-coded for intuitive use. Blue highlights indicate biometric inputs, red indicates manual entries (e.g., lunch breaks), and yellow columns show automatically computed outputs like total working hours.

2. Project-Wise Labor Dataset (PWLD)

Raw biometric attendance data spanning 25,487 entries was first cleaned and structured across

Table 2. Statistical analysis of labor working data-2,465 rows



four stages of preprocessing to yield project-level and trade-wise labor summaries. Each successive table builds on the previous, with derived statistics computed via automated spreadsheet formulas and visualized through Python-based ML outputs.

The JWDW Report consolidates the count of workers per labor type assigned to different jobs on each day. It automatically tallies total workers and provides a daily count for each of the 10 labor types. This transformation was enabled through advanced built-in functions in Google Sheets. This table forms the intermediary stage from daily biometric entries to aggregated project-level insights.

The PWLD was developed by grouping JWDW records under project IDs. For each project, the total number of days and cumulative laborer count under each designation were calculated. This table served as the foundational dataset for further statistical and machine learning modeling.

3.3 Statistical characterization of labor data

To understand the distribution of labor efforts across projects and roles, a comprehensive statistical analysis was performed. The metrics calculated included:

1. Mean laborers per project
2. Median laborers per project
3. Mode laborers per project
4. Standard deviation of laborers per project

These metrics were computed both at the aggregate level (irrespective of labor type) and for each of the 10 individual labor categories. Table 2 summarizes these statistics, with columns “All Labor Types” representing aggregate statistics, and other columns L1-L10 representing the same across individual labor types.

This statistical baseline was essential for the classification of labor intensity and variability in later stages, as well as for model training and evaluation.

3.4 Data quality and preprocessing measures

Data preprocessing steps included:

- Handling Missing Data: Ensuring completeness in biometric entries and manually entered lunch breaks.
- Outlier Detection: Projects with abnormally high or low labor values were cross-validated with project logs.
- Label Encoding: Designation IDs were treated as categorical labels for model input.
- Normalization: Daily working hours were normalized to a standard scale to minimize distortion across roles with varying work durations.

These measures ensured that the dataset maintained high integrity and usability for the subsequent machine learning phase.

In summary, Section 3 presents a comprehensive pipeline from raw biometric entries to structured, statistically enriched datasets suitable for predictive analytics. This robust data foundation enabled reliable forecasting of labor needs across PEB projects in Section 4.

3.5 Data quality and preprocessing measures

The step-by-step pipeline described in this section in detail is summarized below,

Raw Data → EDWA → JWDW → PWLD → Statistical Table → ML Input

4. Labor behavior classification

4.1 Overview of classification strategy

Understanding how labor is distributed across various PEB projects is essential for forecasting accuracy. Before proceeding with machine learning modeling, each project was carefully examined for its labor behavior in terms of intensity (overall labor usage per day) and variability (consistency in daily labor deployment). These two dimensions play a crucial role in characterizing the predictability of labor trends and subsequently influence model performance.

To validate the suitability of fixed thresholds, this sensitivity analysis was conducted comparing fixed thresholds with percentile-based thresholds. Results showed that fixed thresholds provided more consistent grouping across projects and aligned better with practical labor planning norms in PEB environments.

4.2 Definitions and measurement

Labor intensity for each project was defined as the mean number of laborers employed per day over the project duration. Labor variability was quantified by taking the absolute difference between the mean and median laborers per day. A higher variability indicated fluctuating workforce patterns, while a lower difference reflected stable deployment.

4.3 Fixed threshold approach for classification

To categorize projects consistently, we adopted a fixed threshold methodology based on

Table 3. Mean intensity and variability analysis

Job Code	Mean Labor	Labor Intensity Category	Abs Diff All Labor	Labor Variability Category	Combined Labor Behavior
EEIPL/05/102	7	Low Intensity	0	Low Variability	Low Intensity – Low Variability
EEIPL/05/103	12	Low Intensity	0	Low Variability	Low Intensity – Low Variability
EEIPL/05/104.1	6	Low Intensity	1	Low Variability	Low Intensity – Low Variability
EEIPL/05/109	19	Medium Intensity	3	Medium Variability	Medium Intensity – Medium Variability
EEIPL/05/110	17	Medium Intensity	3	Medium Variability	Medium Intensity – Medium Variability
EEIPL/05/112	13	Low Intensity	0	Low Variability	Low Intensity – Low Variability
EEIPL/05/114	11	Low Intensity	4	Medium Variability	Low Intensity – Medium Variability
EEIPL/05/115	4	Low Intensity	1	Low Variability	Low Intensity – Low Variability
EEIPL/05/117.1	13	Low Intensity	2	Low Variability	Low Intensity – Low Variability
EEIPL/05/117.2	9	Low Intensity	4	Medium Variability	Low Intensity – Medium Variability
EEIPL/05/118	18	Medium Intensity	0	Low Variability	Medium Intensity – Low Variability
EEIPL/05/119	6	Low Intensity	0	Low Variability	Low Intensity – Low Variability
EEIPL/08/121	12	Low Intensity	0	Low Variability	Low Intensity – Low Variability
EEIPL/08/122	9	Low Intensity	0	Low Variability	Low Intensity – Low Variability
EEIPL/08/123	11	Low Intensity	0	Low Variability	Low Intensity – Low Variability
EEIPL/08/124	13	Low Intensity	1	Low Variability	Low Intensity – Low Variability
EEIPL/08/125	17	Medium Intensity	4	Medium Variability	Medium Intensity – Medium Variability
EEIPL/08/126	4	Low Intensity	3	Medium Variability	Low Intensity – Medium Variability
EEIPL/08/127	23	Medium Intensity	3	Medium Variability	Medium Intensity – Medium Variability
EEIPL/08/128	9	Low Intensity	3	Medium Variability	Low Intensity – Medium Variability
EEIPL/08/129	15	Low Intensity	0	Low Variability	Low Intensity – Low Variability
EEIPL/08/130	10	Low Intensity	1	Low Variability	Low Intensity – Low Variability
EEIPL/08/134	12	Low Intensity	2	Low Variability	Low Intensity – Low Variability
EEIPL/08/135	16	Medium Intensity	4	Medium Variability	Medium Intensity – Medium Variability

commonly observed patterns in the PEB industry. As shown in the classification criteria below:

Labor Intensity

- Low Intensity: 0-15 workers/day
- Medium Intensity: 16-30 workers/day
- High Intensity: >30 workers/day

Labor Variability

- Low Variability: 0-2 workers/day
- Medium Variability: >2-5 workers/day
- High Variability: >5 workers/day

These classifications reflect realistic groupings found across similar industrial environments and help maintain consistency with industry benchmarks.

4.4 Consideration of dynamic thresholds

Although a dynamic thresholding method using 33rd and 66th percentiles was explored, it was ultimately set aside. The dynamic cutoffs (e.g., 8 workers/day for low intensity) conflicted with established norms and introduced inconsistencies when comparing across projects or organizations. Therefore, the fixed method was retained due to its practicality, transparency, and alignment with operational planning.

4.5 Cross-classification analysis and modeling implications

Each project was assigned to one of the nine possible behavioral groups formed by the

Figure 1. Cross classification of labor intensity vs variability

combination of labor intensity and variability categories. The resulting distribution is summarized in Table 3 and a visual representation of this distribution is presented in Fig. 1, which displays a matrix of the number of projects falling under each class.

Each cross-classified group revealed different labor behavioral traits:

- High Intensity-High Variability: Typically, chaotic projects; harder to predict.
- High Intensity-Low Variability: Large-scale but well-organized.
- Medium Intensity-Medium Variability: Balanced scale with moderate complexity.
- Low Intensity-Low Variability: Predictable and stable projects.

This distribution clearly shows that a large portion of the dataset (approximately 72%) belongs to the Low Intensity-Low Variability class. Such projects are typically characterized by simple fabrication scopes and stable workforce needs, making them ideal for baseline modeling.

However, the dataset exhibits limited representation in the Medium Intensity categories and an absence of High Intensity or High Variability projects. This skew introduces modeling constraints in terms of generalizability, especially for applications involving more complex or dynamic labor scenarios.

Despite this, the clearly segmented structure of the dataset enables the construction of classification-specific models tailored to stable, low-to-medium intensity projects, with the potential to expand the model in future phases as additional high-variability projects are incorporated.

The classification of labor behavior not only added context but also improved the interpretability and grouping logic for ML model training. As we will discuss in Section 5, each classification group was treated as a separate data subset (e.g., P1C1 vs P2C1), ensuring that labor forecasting models could adapt to specific project characteristics.

In summary, the labor behavior classification framework laid a strong analytical foundation, allowing for more targeted modeling strategies and offering clearer insights into forecasting accuracy under different labor deployment scenarios.

5. Machine learning model development and evaluation

5.1 Model selection rationale

In selecting an appropriate machine learning algorithm for labor forecasting in PEB fabrication

projects, several criteria were considered: the nature of the dataset, the interpretability of the model, robustness to non-linearity, and adaptability to diverse project types. After comparative reviews and preliminary experiments, the Random Forest Regressor (RFR) was selected as the core predictive engine for this study.

The labor dataset used in this research was derived from real-time factory attendance systems, with features that included discrete labor categories, continuous working hour records, and project-level classification labels. Such data often exhibits non-linear relationships and heterogeneous distributions, making linear models inadequate in capturing underlying dynamics.

Random Forest, an ensemble learning algorithm based on decision trees, excels in such scenarios due to the following advantages:

- Non-linear modeling capability: RFR can handle complex feature interactions without the need for manual transformations or assumptions about data distribution.
- Robustness to noise and overfitting: By aggregating predictions over multiple randomized trees, the model reduces the likelihood of overfitting, especially in datasets with moderate sample sizes and variable quality.
- Feature importance evaluation: RFR provides intrinsic measures of variable importance, enabling an understanding of which project or labor parameters most influence prediction outcomes.
- Stability across diverse groups: Since the dataset was partitioned into project-based classification groups (as defined in Section 4), the model needed to generalize well across varying project scales and labor deployment behaviors. RFR's structure allows for independent modeling of such groupings without loss of consistency.

Preliminary testing with alternative models such as linear regression and decision trees showed inferior generalization, particularly in handling labor types with high variability. In contrast, the Random Forest Regressor yielded more stable and interpretable outputs, which justified its selection as the primary model for forecasting labor demand in this work.

The steps involved in this algorithm is briefed below:

- Step 1: Input EDWA dataset
- Step 2: Aggregate to project-level dataset
- Step 3: Compute mean and variability
- Step 4: Classify projects (Low/Medium/High)
- Step 5: Split dataset into classification groups
- Step 6: Train Random Forest model per group
- Step 7: Predict labor categories (TL1-TL10)
- Step 8: Compute RMSE, MSE
- Step 9: Evaluate profitability zones

5.2 Training methodology

The training process for the Random Forest Regressor (RFR) model was conducted on a well-structured and filtered dataset, derived from an initial pool of 71 real-time PEB projects. A systematic preprocessing pipeline was implemented to ensure data quality, starting with the elimination of incomplete or non-fabricated projects. Projects were then segregated based on the presence of key structural components—namely columns and rafters—using logical groupings (AND/OR) to define two core types: P1 (projects containing both) and P2 (projects containing either).

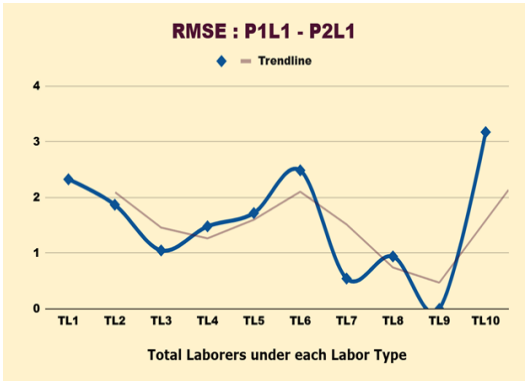


Figure 2(a). RMSE results for classification P1C1-P2C1

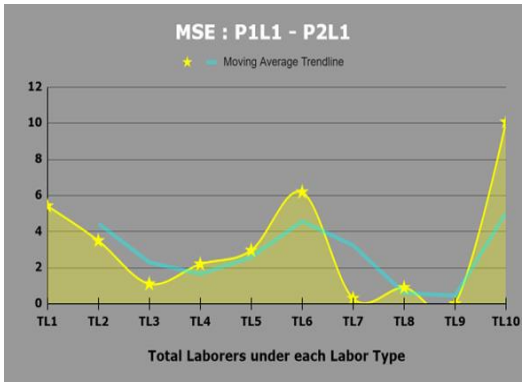


Figure 3(a). MSE results for classification P1C1-P2C1

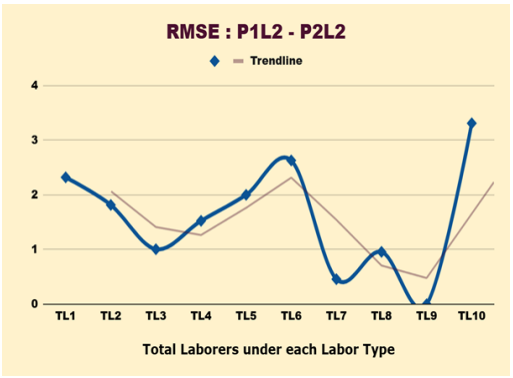


Figure 2(b). RMSE results for classification P1C2-P2C2

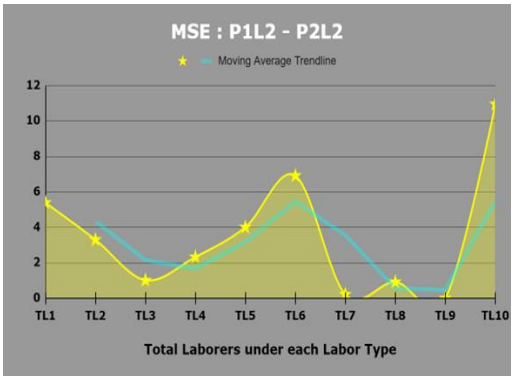


Figure 3(b). MSE results for classification P1C2-P2C2

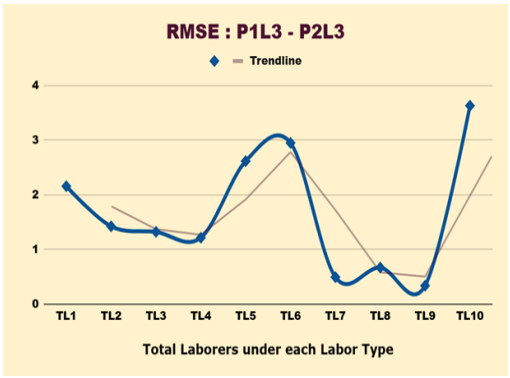


Figure 2(c). RMSE results for classification P1C3-P2C3

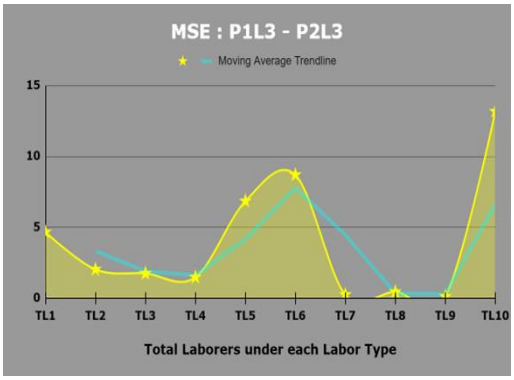


Figure 3(c). MSE results for classification P1C3-P2C3

Further grouping was done based on fabrication quantity thresholds (in metric tons), narrowing the dataset to medium and large-scale projects to ensure consistency in labor demand. Projects were then categorized into three complexity levels—C1, C2, and C3—reflecting the presence of

primary, secondary, and load-bearing structural components, respectively.

Combining these two dimensions (P1/P2 and C1/C2/C3), the data was classified into three core training groups: P1C1 vs P2C1, P1C2 vs P2C2, and P1C3 vs P2C3. This classification ensured homogeneity in structure, scale, and complexity across each dataset subset.

Each classification group was independently used to train a multi-output regression model to predict daily labor demand across ten labor categories (TL1-TL10). The training pipeline included standardized feature scaling and 5-fold cross-validation to validate model stability and generalization. This classification-aware training methodology enabled the RFR to deliver context-sensitive forecasts tailored to the specific characteristics of each PEB project group.

5.3 Model evaluation metrics

To evaluate the predictive performance of the RFR models, two standard regression error metrics were used:

- Root Mean Squared Error (RMSE): Provides a direct interpretation of the average magnitude of error, maintaining the same unit as the original data.
- Mean Squared Error (MSE): Emphasizes larger deviations more heavily by squaring error terms, thus highlighting labor categories with irregular behaviors.

These metrics were computed for each labor type across all classification groups.

5.4 RMSE evaluation

The Root Mean Squared Error (RMSE) was used to evaluate the prediction accuracy of the Random Forest Regressor (RFR) model across different labor categories. RMSE quantifies the average magnitude of prediction error, with lower values indicating better alignment between actual and predicted labor quantities.

The RMSE results are visually presented in Figs. 2 (a)-(c), corresponding to the three classification pairs:

Across all three groups, the model consistently achieved lower RMSE values for TL7 (Hydra Operator), TL8 (Rigger), and TL9 (Semi Fitter). These roles are typically involved in repetitive, mechanical, or equipment-based tasks such as rigging structural components, handling cranes, or semi-automated fitting—all of which follow a relatively uniform schedule and quantity requirement across projects. Their demand does not fluctuate significantly, which makes them inherently easier to predict.

In contrast, higher RMSE values were observed for TL6 (Helper) and TL10 (Welder). Helpers are typically multi-functional and are deployed based on short-term needs, such as material handling, cleaning, or assisting skilled labor. Their numbers can change rapidly depending on day-to-day site activity, making them harder to model. Welders, though a skilled category, often have demand spikes tied to project-specific welding lengths, complexity, or rework needs — leading to greater variance that is difficult to predict using static project features alone.

These findings highlight that while the model performs reliably for labor roles with consistent task structures, more variable roles may require either additional feature inputs (e.g., task duration, rework rates) or finer-grained time-series modeling to improve prediction accuracy.

5.5 MSE evaluation

The corresponding Mean Squared Error (MSE) was also calculated to assess model performance and are illustrated in Figs. 3 (a)-(c) for the same classification categories. MSE represents the average of squared differences between predicted and actual values, and it penalizes larger errors more heavily than RMSE. While RMSE is more interpretable in real units, MSE is valuable for identifying the influence of outliers and measuring model sensitivity to large deviations.

The patterns observed in the MSE plots were largely consistent with those seen in the RMSE evaluation. Labor categories TL7 (Hydra Operator), TL8 (Rigger), and TL9 (Semi Fitter) continued to exhibit lower error values across all project classifications. This reaffirms the stable nature of their deployment and the model's capability to predict them accurately.

Conversely, TL6 (Helper) and TL10 (Welder) again showed elevated MSE scores, reinforcing their higher forecasting uncertainty due to dynamic or project-specific demand patterns. The MSE metric's sensitivity to large deviations also highlights the presence of occasional outlier behavior in these categories—potentially driven by sudden labor surges, delays, or rework requirements.

Together, the RMSE and MSE results validate the robustness of the model in forecasting consistent labor roles, while also indicating areas where further feature enrichment or alternative modeling strategies may be beneficial.

5.6 Comparative summary and observations

A comparative analysis across the three classification sets—P1C1 vs P2C1, P1C2 vs P2C2, and P1C3 vs P2C3—revealed important performance differences in terms of forecasting accuracy. Among these, the P1C3 vs P2C3 group consistently outperformed the others in both RMSE and MSE metrics across most labor categories. This can be attributed to the lower within-group variability and higher structural consistency of the projects, which allowed the model to learn more coherent patterns and generalize effectively.

In contrast, classification sets involving higher labor variability—particularly P1C1 vs P2C1—exhibited greater prediction errors. This confirms earlier observations in Section 4.5, where high-variability groups were identified as more complex and less predictable due to inconsistent labor deployment.

These findings validate the strategic benefit of classification-based grouping in ML forecasting for PEB projects. By aligning model training with labor behavior patterns, the forecasting system demonstrates greater adaptability and interpretability. Furthermore, the dual-metric evaluation (RMSE and MSE) provided a comprehensive understanding of the model's strengths and highlighted specific labor types, such as TL6 and TL10, where enhancements in data representation or modeling depth may be required.

This performance analysis sets the stage for the next section, which compares forecasted outputs with real project data, and examines the financial and operational implications of prediction deviations—an essential dimension for assessing the practical value of the proposed forecasting framework.

6. Results and profitability analysis

6.1 Comparison of model predictions with real-world data

Table 4(a). ML Predictive model performance on P1C1 Vs P2C1

Original Data			P1L1 - P2L1										
Sl.No.	Code	Total MT	TL	TL1	TL2	TL3	TL4	TL5	TL6	TL7	TL8	TL9	TL10
1	P2	33.61	10	0	0	3	0	0	2	0	0	1	4
2	P29	45.20	32	4	2	3	2	5	8	0	0	1	7
3	P36	42.78	35	5	3	3	2	4	8	0	0	1	9
4	P43	99.08	47	6	5	4	4	4	8	1	2	1	12
5	P46	90.13	31	5	3	2	2	4	6	1	0	1	7
6	P51	42.97	30	5	2	2	1	5	6	0	0	1	8

ML Predictions													
Sl.No.	Code	Total MT	TL	TL1	TL2	TL3	TL4	TL5	TL6	TL7	TL8	TL9	TL10
1	P2	33.61	28	4	2	3	1	4	6	0	0	1	7
2	P29	45.20	31	4	2	3	2	5	7	0	0	1	7
3	P36	42.78	31	4	2	3	2	4	7	0	0	1	8
4	P43	99.08	35	5	3	3	2	4	7	0	1	1	9
5	P46	90.13	31	5	3	2	2	5	6	0	0	1	8
6	P51	42.97	30	5	2	2	1	5	6	0	0	1	8

Difference %													
Sl.No.	Code	Total MT	TL	TL1	TL2	TL3	TL4	TL5	TL6	TL7	TL8	TL9	TL10
1	P2	33.61	54.36%	11.72%	6.04%	-1.19%	4.14%	11.19%	11.72%	0.51%	0.36%	0.00%	9.88%
2	P29	45.20	-2.65%	0.84%	0.27%	-0.95%	-0.84%	-0.73%	-2.30%	0.24%	0.00%	0.00%	0.82%
3	P36	42.78	-9.51%	-1.19%	-1.71%	-0.96%	-1.05%	0.77%	-2.81%	0.19%	0.05%	0.00%	-2.81%
4	P43	99.08	-11.94%	-0.97%	-1.97%	-1.21%	-1.84%	0.33%	-1.31%	-0.55%	-1.35%	0.00%	-3.08%
5	P46	90.13	0.49%	-0.08%	-0.51%	0.24%	-0.39%	0.60%	0.40%	-0.63%	0.09%	0.00%	0.77%
6	P51	42.97	0.70%	-0.72%	0.37%	0.65%	0.79%	-0.88%	0.61%	0.33%	0.14%	0.00%	-0.58%

Excess Prediction (above 5%)	min	54.36%	11.72%	6.04%	0.00%	0.00%	11.19%	11.72%	0.00%	0.00%	0.00%	9.88%	Cumulative Average to see overall trends
	max	54.36%	11.72%	6.04%	0.00%	0.00%	11.19%	11.72%	0.00%	0.00%	0.00%	9.88%	
	Total %	16.67%	16.67%	16.67%	0.00%	0.00%	16.67%	16.67%	0.00%	0.00%	0.00%	16.67%	
Shortfall Prediction (below -5%)	min	-11.94%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	
	max	-9.51%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	
	Total %	33.33%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	
Profitable (between 0 to 5%)		33.33%	16.67%	33.33%	33.33%	33.33%	50.00%	33.33%	66.67%	66.67%	0.00%	33.33%	
More Profitable (between 0 to -5%)		16.67%	66.67%	50.00%	66.67%	66.67%	33.33%	50.00%	33.33%	33.33%	100.00%	50.00%	
Acceptable Profit (between -5% to 5%)		50.00%	83.33%	83.33%	100.00%	100.00%	83.33%	83.33%	100.00%	100.00%	100.00%	83.33%	

Cumulative Average to see overall trends
9.54%
9.54%
9.09%
1.09%
0.86%
3.03%
36.36%
51.52%
87.88%

Following the training and evaluation of the Random Forest Regressor (RFR) models across the three classification groups, the next step involved comparing the predicted labor quantities with actual labor deployment data showcased in Table 4 (a)-(c). This comparison was conducted for each test project across all ten labor types (TL1-TL10).

Each table consists of three sections:

1. Original Data: Ground-truth labor quantities extracted from project logs.
2. ML Predictions: Output values generated by the trained RFR model.
3. Difference %: The percentage deviation of predicted values from actual data.

Predictions with exact matches (0% difference) are highlighted in green, predictions within an acceptable ±5% range are marked in yellow, and underpredictions or overpredictions beyond this threshold are flagged in red.

6.2 Visual analysis of forecast deviations

The distribution of these differences is visually represented in Figs. 4 (a)-(c). These figures highlight:

Table 4(b). ML predictive model performance on P1C2 Vs P2C2

Original Data			P1L2 - P2L2									
SLNo.	Code	Total MT	TL	TL1	TL2	TL3	TL4	TL5	TL6	TL7	TL8	TL9	TL10
1	P2	33.61	10	0	0	3	0	0	2	0	0	1	4
2	P29	45.20	32	4	2	3	2	5	8	0	0	1	7
3	P31	73.82	45	6	3	3	3	7	10	0	0	1	12
4	P36	42.78	35	5	3	3	2	4	8	0	0	1	9
5	P43	99.08	47	6	5	4	4	4	8	1	2	1	12
6	P46	90.13	31	5	3	2	2	4	6	1	0	1	7
7	P51	42.97	30	5	2	2	1	5	6	0	0	1	8

ML Predictions												
SLNo.	Code	Total MT	TL	TL1	TL2	TL3	TL4	TL5	TL6	TL7	TL8	TL9	TL10
1	P2	33.61	32	4	3	3	2	4	6	0	0	1	8
2	P29	45.20	33	5	2	3	2	5	7	0	0	1	8
3	P31	73.82	38	5	3	3	2	5	8	0	0	1	10
4	P36	42.78	34	5	3	3	2	5	7	0	0	1	9
5	P43	99.08	39	5	4	3	3	4	7	1	1	1	10
6	P46	90.13	34	5	3	2	2	4	7	1	0	1	8
7	P51	42.97	31	5	2	2	2	4	6	0	0	1	8

Difference %												
SLNo.	Code	Total MT	TL	TL1	TL2	TL3	TL4	TL5	TL6	TL7	TL8	TL9	TL10
1	P2	33.61	65.31%	12.68%	7.47%	-0.36%	5.71%	11.66%	13.21%	0.92%	1.25%	0.00%	12.77%
2	P29	45.20	1.75%	1.26%	0.91%	-0.66%	-0.18%	-0.62%	-1.81%	0.49%	0.35%	0.00%	2.01%
3	P31	73.82	-9.04%	-0.84%	-0.09%	-0.33%	-0.76%	-2.11%	-2.55%	0.38%	0.24%	0.00%	-2.98%
4	P36	42.78	-1.31%	-0.09%	-0.47%	-0.70%	0.14%	1.38%	-1.92%	0.75%	0.65%	0.00%	-1.05%
5	P43	99.08	-8.51%	-0.72%	-1.42%	-0.88%	-1.28%	0.26%	-0.87%	-0.37%	-1.03%	0.00%	-2.20%
6	P46	90.13	3.15%	0.12%	-0.02%	0.50%	0.08%	0.43%	0.67%	-0.44%	0.36%	0.00%	1.46%
7	P51	42.97	2.65%	-0.88%	0.70%	1.07%	1.28%	-1.21%	0.88%	0.47%	0.42%	0.00%	-0.07%

Excess Prediction (above 5%)	min	65.31%	12.68%	7.47%	0.00%	5.71%	11.66%	13.21%	0.00%	0.00%	0.00%	12.77%	Cumulative Average to see overall trends
	max	65.31%	12.68%	7.47%	0.00%	5.71%	11.66%	13.21%	0.00%	0.00%	0.00%	12.77%	11.71%
	Total %	14.29%	14.29%	14.29%	0.00%	14.29%	14.29%	14.29%	0.00%	0.00%	0.00%	14.29%	9.09%
Shortfall Prediction (below -5%)	min	-9.04%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.82%
	max	-8.51%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.77%
	Total %	28.57%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	2.60%
Profitable (between 0 to 5%)		42.86%	28.57%	28.57%	28.57%	42.86%	42.86%	28.57%	71.43%	85.71%	0.00%	28.57%	38.96%
More Profitable (between 0 to -5%)		14.29%	57.14%	57.14%	71.43%	42.86%	42.86%	57.14%	28.57%	14.29%	100.00%	57.14%	49.35%
Acceptable Profit (between -5% to 5%)		57.14%	85.71%	85.71%	100.00%	85.71%	85.71%	85.71%	100.00%	100.00%	100.00%	85.71%	88.31%

Table 4(c). ML predictive model performance on P1C3 Vs P2C3

Original Data			P1L3 - P2L3									
SLNo.	Code	Total MT	TL	TL1	TL2	TL3	TL4	TL5	TL6	TL7	TL8	TL9	TL10
1	P2	33.61	10	0	0	3	0	0	2	0	0	1	4
2	P24	62.58	54	8	4	4	3	4	13	0	0	2	16
3	P31	73.82	45	6	3	3	3	7	10	0	0	1	12
4	P32	100.99	52	6	4	5	3	9	11	0	0	1	13
5	P36	42.78	35	5	3	3	2	4	8	0	0	1	9
6	P43	99.08	47	6	5	4	4	4	8	1	2	1	12
7	P46	90.13	31	5	3	2	2	4	6	1	0	1	7
8	P47	33.88	34	6	4	2	1	4	7	0	0	1	9
9	P51	42.97	30	5	2	2	1	5	6	0	0	1	8

ML Predictions												
SLNo.	Code	Total MT	TL	TL1	TL2	TL3	TL4	TL5	TL6	TL7	TL8	TL9	TL10
1	P2	33.61	31	4	3	3	1	4	6	0	0	1	8
2	P24	62.58	43	6	4	3	2	4	10	0	0	1	12
3	P31	73.82	44	6	3	3	3	6	9	0	0	1	12
4	P32	100.99	45	6	4	4	3	6	10	0	0	1	12
5	P36	42.78	36	5	3	3	2	5	8	0	0	1	10
6	P43	99.08	40	6	4	3	3	4	7	1	1	1	10
7	P46	90.13	36	5	3	3	2	5	7	0	0	1	9
8	P47	33.88	35	6	3	2	2	5	7	0	0	1	9
9	P51	42.97	34	5	3	2	2	5	7	0	0	1	9

Difference %												
SLNo.	Code	Total MT	TL	TL1	TL2	TL3	TL4	TL5	TL6	TL7	TL8	TL9	TL10
1	P2	33.61	62.79%	13.24%	7.68%	-0.92%	4.34%	11.63%	12.85%	0.45%	0.54%	0.03%	12.94%
2	P24	62.58	-16.94%	-2.57%	-0.74%	-1.15%	-1.09%	0.72%	-5.42%	0.19%	0.26%	-0.94%	-6.20%
3	P31	73.82	-1.35%	0.05%	0.60%	0.31%	-0.34%	-1.38%	-0.80%	0.20%	0.33%	0.12%	-0.45%
4	P32	100.99	-6.56%	0.16%	-0.22%	-1.28%	-0.41%	-2.93%	-1.44%	0.15%	0.28%	0.16%	-1.03%
5	P36	42.78	3.44%	1.05%	0.40%	-0.51%	-0.35%	1.71%	-0.93%	0.33%	0.42%	0.09%	1.24%
6	P43	99.08	-7.23%	-0.48%	-1.25%	-0.80%	-1.27%	0.29%	-0.58%	-0.42%	-1.03%	0.03%	-1.72%
7	P46	90.13	5.56%	0.39%	0.14%	0.71%	0.13%	0.70%	1.22%	-0.60%	0.40%	0.03%	2.43%
8	P47	33.88	2.57%	-1.48%	-2.45%	1.39%	1.48%	1.59%	0.68%	0.30%	0.24%	0.15%	0.68%
9	P51	42.97	9.12%	0.61%	1.54%	1.05%	1.26%	-0.14%	1.89%	0.30%	0.42%	0.05%	2.16%

Excess Prediction (above 5%)	min	5.56%	13.24%	7.68%	0.00%	0.00%	11.63%	12.85%	0.00%	0.00%	0.00%	12.94%	Cumulative Average to see overall trends
	max	62.79%	13.24%	7.68%	0.00%	0.00%	11.63%	12.85%	0.00%	0.00%	0.00%	12.94%	11.01%
	Total %	33.33%	11.11%	11.11%	0.00%	0.00%	11.11%	11.11%	0.00%	0.00%	0.00%	11.11%	8.08%
Shortfall Prediction (below -5%)	min	-16.94%	0.00%	0.00%	0.00%	0.00%	0.00%	-5.42%	0.00%	0.00%	0.00%	-6.20%	2.60%
	max	-6.56%	0.00%	0.00%	0.00%	0.00%	0.00%	-5.42%	0.00%	0.00%	0.00%	-6.20%	1.65%
	Total %	33.33%	0.00%	0.00%	0.00%	0.00%	0.00%	11.11%	0.00%	0.00%	0.00%	11.11%	5.05%
Profitable (between 0 to 5%)		22.22%	55.56%	44.44%	44.44%	44.44%	55.56%	33.33%	77.78%	88.89%	88.89%	44.44%	54.55%
More Profitable (between 0 to -5%)		11.11%	33.33%	44.44%	55.56%	55.56%	33.33%	44.44%	22.22%	11.11%	11.11%	33.33%	32.32%
Acceptable Profit (between -5% to 5%)		33.33%	88.89%	88.89%	100.00%	100.00%	88.89%	77.78%	100.00%	100.00%	100.00%	77.78%	86.87%

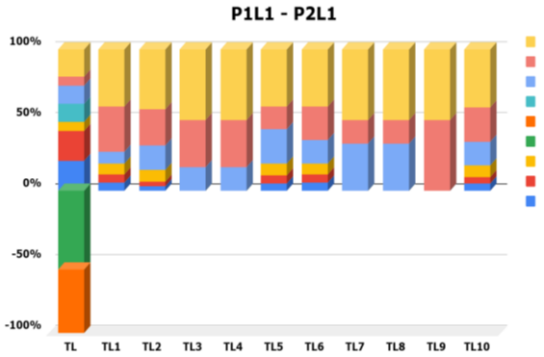


Figure 4(a). Statistical analysis and logical functions of P1C1 Vs P2C1

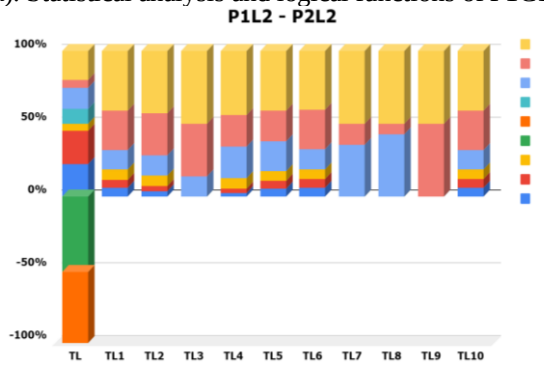


Figure 4(b). Statistical analysis and logical functions of P1C2 Vs P2C2

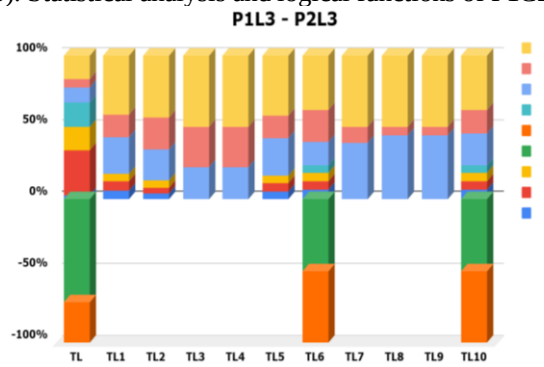


Figure 4(c). Statistical analysis and logical functions of P1C3 Vs P2C3

• The percentage of test projects showing overestimated labor values (excess predictions)
• Those with underestimated values (shortfall predictions)
• Predictions that fall within an allowable tolerance range (±5%)

Figs. 4 (a)-(c) collectively depict the distribution of labor forecast deviations across the three classification groups—P1C1 vs P2C1, P1C2 vs P2C2, and P1C3 vs P2C3—offering a comprehensive perspective on the model's predictive reliability.

In Fig. 4(a) (P1C1 vs P2C1), approximately 50% of labor predictions fall within the ±5% tolerance range, demonstrating strong alignment with actual labor deployment for the majority of projects. Shortfall predictions account for about 33%, with underestimations ranging from

-11.94% to -9.51%, which may pose operational risks if labor supply is not adequately adjusted. Excess predictions comprise roughly 17%, including some significant outliers reaching up to 54.36%. Labor categories such as TL3, TL4, TL7, TL8, and TL9 show exceptional forecasting accuracy, while TL1, TL2, TL5, TL6, and TL10 tend to be slightly overpredicted, reflecting a conservative approach that favors resource availability.

Fig. 4(b) (P1C2 vs P2C2) exhibits a slightly more conservative prediction pattern, with 57.14% of forecasts within the acceptable range. Shortfall predictions decrease to 28.57%, while excess predictions account for 14.29%. The maximum excess prediction here is notably higher at 65.31%, suggesting occasional overestimation in more complex project settings. Labor types TL3, TL7, TL8, and TL9 maintain perfect accuracy, with other categories showing moderate overprediction without significant shortfalls, continuing the model's conservative bias in resource estimation.

For the highest complexity group, Fig. 4(c) (P1C3 vs P2C3) reveals a more balanced distribution, with equal proportions (33.33%) of excess, shortfall, and acceptable predictions. Overestimations range from 5.56% to 62.79%, while underestimations vary from -16.94% to -6.56%. Labor categories TL3, TL4, TL7, TL8, and TL9 consistently achieve 100% accuracy within profitable bounds. However, TL6 and TL10 show a mixture of excess and shortfall predictions, highlighting areas where model improvements may be necessary to better capture dynamic labor demands.

Across all groups, a significant portion of forecasts fall within the acceptable ±5% tolerance range, indicating robust alignment between predicted and actual labor deployments. Specifically, the proportion of predictions within this range varies from approximately 33% in the highest complexity group (P1C3 vs P2C3) to about 50-57% in the lower complexity groups (P1C1 vs P2C1 and P1C2 vs P2C2). This trend suggests that while the model performs reliably across project types, prediction accuracy tends to be higher in less complex, more stable projects.

When examining individual labor categories, certain roles such as TL3 (Semi Fitter), TL4, TL7 (Hydra Operator), TL8 (Rigger), and TL9 consistently exhibit high predictive accuracy, with most forecasts falling within the profitable range across all classification sets. This reflects the model's strength in capturing the stable, routine deployment patterns of these labor types.

Conversely, labor categories like TL6 (Helper) and TL10 (Welder) show higher variability in forecast deviations, with a mix of excess and shortfall predictions. These roles typically involve more dynamic, demand-driven tasks, explaining the model's comparatively lower precision and indicating areas where further feature engineering or adaptive modeling could improve outcomes.

Overall, the integrated analysis across Figs. 4 (a)-(c) underscores the robustness and practical utility of the classification-based Random Forest forecasting framework. It performs strongly for stable labor roles and less complex projects, while highlighting specific categories and conditions where additional refinement is warranted. This holistic view informs targeted improvements to the model and provides valuable insights for operational workforce planning in PEB fabrication.

6.3 Cumulative forecasting performance and profitability analysis

Fig. 5 summarizes the cumulative percentages of test predictions falling under excess, shortfall, and allowable categories for all three classification sets.

Fig. 5 illustrates the cumulative forecasting trend across all three classification groups—P1L1 & P2L1, P1L2 & P2L2, and P1L3 & P2L3. The model demonstrates remarkable stability, with excess labor predictions remaining between 8% and 9%, and shortfall predictions held consistently low at just 3-5%. Importantly, 87% to 88% of all forecasts fall within the acceptable ±5%

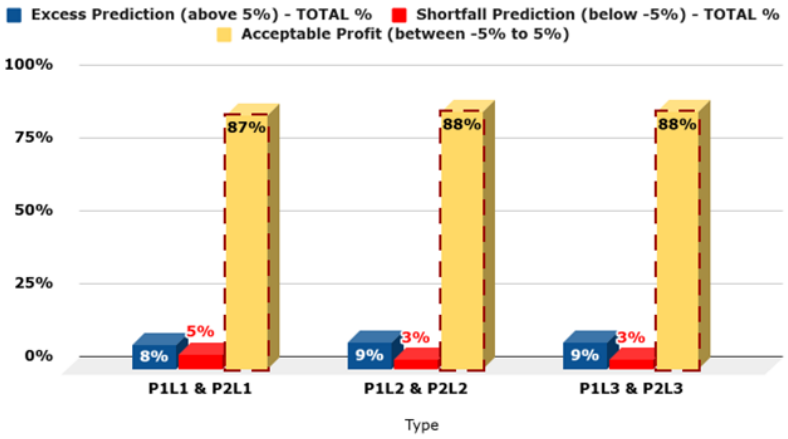


Figure 5. Trend analysis of laborer predictions

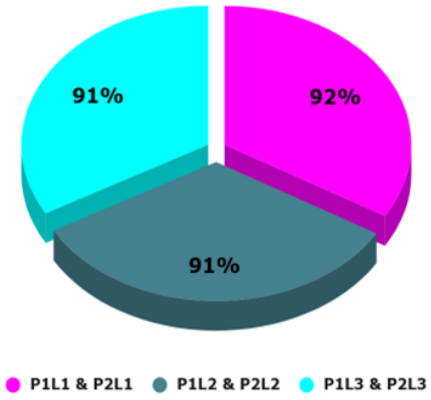


Figure 6. Overall profitability prediction

tolerance range, highlighting the model's precision and its ability to minimize both over- and underestimation risks. This consistent performance across different classification sets reinforces the robustness and generalizability of the proposed forecasting framework in practical labor planning scenarios.

The economic value of the forecasting model is further illustrated in Fig. 6, which presents a consolidated pie chart of profitability distribution.

As shown in Fig. 6, the model's economic viability is further supported by the profitability distribution. Across all classification groups, over 91% of labor forecasts contributed to profitable outcomes, with P1L1 & P2L1 achieving 92%, and P1L2 & P2L2, and P1L3 & P2L3 each recording 91%. These results demonstrate that the Random Forest-based forecasting approach, when combined with classification-driven preprocessing, not only improves predictive accuracy but also delivers meaningful cost-efficiency benefits.

Together, these findings validate the proposed machine learning approach as a reliable and economically sound tool for labor forecasting in PEB fabrication. Its high prediction accuracy and consistent profitability outcomes support its broader application for workforce optimization and strategic resource planning in industrialized construction environments.

7. Conclusions

This study proposed and validated a machine learning-based framework for labor forecasting in Pre-Engineered Building (PEB) fabrication projects using real-time attendance data and classification-driven modeling. The integration of biometric-based labor tracking with structured Google Sheet-based data transformation enabled the creation of a high-quality dataset reflecting actual workforce deployment across 69 projects.

A key innovation of this research was the two-dimensional classification of projects based on labor intensity and labor variability, which enabled tailored training and testing of Random Forest Regressor (RFR) models. This approach ensured that the underlying behavioral patterns of labor deployment were reflected in the predictive model structure.

The RFR model demonstrated high performance in forecasting the quantity of skilled and unskilled labor required across ten distinct labor categories. Dual-metric evaluation using RMSE and MSE indicated stable and accurate performance, particularly for labor types with consistent work routines and minimal variance. Comparative results across classification groups (P1C1-P2C3) revealed that low-variability project groups yielded the most precise forecasts.

From a practical perspective, the model's predictions translated into significant implications for cost control and resource optimization. The profitability analysis revealed that over 88% of forecasts fell within the acceptable or profit-generating threshold, with less than 5% resulting in labor shortages. Furthermore, a consolidated profitability ratio above 91% highlighted the framework's potential for real-world deployment in cost-sensitive fabrication environments.

This framework, by bridging operational labor data with intelligent forecasting, contributes a scalable decision-support tool that can be extended across similar industrial domains. It also offers a methodological template for integrating structured classification and machine learning in other resource prediction problems.

References

1. Waqar, A. (2024). Intelligent decision support systems in construction engineering: An artificial intelligence and machine learning approaches. Expert Systems with Applications, 249, 123503. https://doi.org/10.1016/j.eswa.2024.123503.
2. Alemão, D., Rocha, A.D., Barata, J. (2021). Smart manufacturing scheduling approaches—systematic review and future directions. Applied Sciences, 11(5), 2186. https://doi.org/10.3390/app11052186.
3. Bhatti, M.M., Marin, M., Ellahi, R.F.I.M., Fudulu, I.M. (2023). Insight into the dynamics of EMHD hybrid nanofluid (ZnO/CuO-SA) flow through a pipe for geothermal energy applications. Journal of Thermal Analysis and Calorimetry, 148(24), 14261-14273. https://doi.org/10.1007/s10973-023-12565-8.
4. Bojer, C.S. (2022). Understanding machine learning-based forecasting methods: A decomposition framework and research opportunities. International Journal of Forecasting, 38(4), 1555-1561. https://doi.org/10.1016/j.ijforecast.2021.11.003.
5. Kovacs, E. (2021). The LSST-DESC 3x2pt tomography optimization challenge. The Open Journal of Astrophysics. https://doi.org/10.1023/A:1010933404324.
6. Cañas, H., Mula, J., Campuzano-Bolarín, F., Poler, R. (2022). A conceptual framework for smart production planning and control in Industry 4.0. Computers & Industrial Engineering, 173, 108659. https://doi.org/10.1016/j.cie.2022.108659.
7. Cheng, J., Li, X., Jiang, K., Li, S., Su, A., Zhao, O. (2024). Machine-learning-assisted design of high strength steel I-section columns. Engineering Structures, 308, 118018. https://doi.org/10.1016/j.engstruct.2024.118018.

8. Haddad, E. (2017). Construction productivity estimation model using artificial neural network for foundations works in Gaza strip construction sites. International Journal of Advanced Engineering Research and Science (IJAERS), 4(7), 2456-1908. https://doi.org/10.22161/ijaers.4.7.9.
9. Duymaz, Ş., Güneri, A.F. (2024). The application of machine learning algorithms in the estimation of production lead times: A case study of a steel construction manufacturing company. Journal of Advances in Manufacturing Engineering (JAME), 5(1), 21-28. https://doi.org/10.14744/ytu.jame.2024.00004.
10. Fahle, S., Prinz, C., Kuhlenkötter, B. (2020). Systematic review on machine learning (ML) methods for manufacturing processes–Identifying artificial intelligence (AI) methods for field application. Procedia CIRP, 93, 413-418. https://doi.org/10.1016/j.procir.2020.04.109.
11. Gao, Y., Wang, J., Xu, X. (2024). Machine learning in construction and demolition waste management: Progress, challenges, and future directions. Automation in Construction, 162, 105380. https://doi.org/10.1016/j.autcon.2024.105380.
12. Guo, K., Yang, Z., Yu, C.H., Buehler, M.J. (2021). Artificial intelligence and machine learning in design of mechanical materials. Materials Horizons, 8(4), 1153-1172. https://doi.org/10.1039/D0MH01451F.
13. Guo, S., Yu, J., Liu, X., Wang, C., Jiang, Q. (2019). A predicting model for properties of steel using the industrial big data based on machine learning. Computational Materials Science, 160, 95-104. https://doi.org/10.1016/j.commatsci.2018.12.056.
14. Hwang, S.H., Mangalathu, S., Shin, J., Jeon, J.S. (2022). Estimation of economic seismic loss of steel moment-frame buildings using a machine learning algorithm. Engineering Structures, 254, 113877. https://doi.org/10.1016/j.engstruct.2022.113877.
15. Imam, M.H., Mohiuddin, M., Shuman, N.M., Oyshi, T.I., Debnath, B., Liham, M.I.M.H. (2024). Prediction of seismic performance of steel frame structures: a machine learning approach. Structures, 69, 107547. https://doi.org/10.1016/j.istruc.2024.107547.
16. Cheng, J.C., Law, K.H., Bjornsson, H., Jones, A., Sriram, R.D. (2010). Modeling and monitoring of construction supply chains. Advanced Engineering Informatics, 24(4), 435-455. https://doi.org/10.1016/j.aei.2010.06.009.
17. Kang, Z., Catal, C., Tekinerdogan, B. (2020). Machine learning applications in production lines: A systematic literature review. Computers & Industrial Engineering, 149, 106773. https://doi.org/10.1016/j.cie.2020.106773.
18. Li, Y., Carabelli, S., Fadda, E., Manerba, D., Tadei, R., Terzo, O. (2020). Machine learning and optimization for production rescheduling in Industry 4.0. The International Journal of Advanced Manufacturing Technology, 110(9), 2445-2463. https://doi.org/10.1007/s00170-020-05850-5.
19. Hobiny, A., Abbas, I., Marin, M. (2022). The influences of the hyperbolic two-temperatures theory on waves propagation in a semiconductor material containing spherical cavity. Mathematics, 10(1), 121. https://doi.org/10.3390/math10010121.
20. Oddershede, A.M., Quezada, L.E., Valenzuela, J.E., Palominos, P.I., Lopez-Ospina, H. (2019). Formulation of a manufacturing strategy using the house of quality. Procedia Manufacturing, 39, 843-850. https://doi.org/10.1016/j.promfg.2020.01.417.
21. Gopal, S.R.T.G., Murali, K. (2016). Analysis of factors affecting labour productivity in construction. International Journal of Recent Scientific Research, 7(6), 11744-11747.
22. Raju, S.T.U., Sarker, A., Das, A., Islam, M.M., Al-Rakhami, M.S., Al-Amri, A.M., ... & Albogamy, F.R. (2022). An approach for demand forecasting in steel industries using ensemble learning. Complexity, 2022(1), 9928836. https://doi.org/10.1155/2022/9928836.
23. Reuter, C., Brambring, F. (2016). Improving data consistency in production control. Procedia CIRP, 41, 51-56. https://doi.org/10.1016/j.procir.2015.12.116.
24. Sadatnya, A., Sadeghi, N., Sabzekar, S., Khanjani, M., Tak, A.N., Taghaddos, H. (2023). Machine learning for construction crew productivity prediction using daily work reports. Automation in Construction, 152, 104891. https://doi.org/10.1016/j.autcon.2023.104891.
25. Salihu, A.A., Ibrahim, Y., Muhammad, S. (2022). Impact of labour productivity factors on construction

project cost and time. LAUTECH Journal of Civil and Environmental Studies, 8(1), 1-14. https://doi.org/10.36108/laujoces/2202.80.0110.
26. Shaheen, M.A., Presswood, R., Afshan, S. (2023). Application of machine learning to predict the mechanical properties of high strength steel at elevated temperatures based on the chemical composition. Structures, 52, 17-29. https://doi.org/10.1016/j.istruc.2023.03.085.
27. Shinde, P.P., Shah, S. (2018). A review of machine learning and deep learning applications. 2018 4th international conference on computing communication control and automation (ICCUBEA), 1-6.
28. Sun, W., Wang, Q., Zhou, Y., Wu, J. (2020). Material and energy flows of the iron and steel industry: Status quo, challenges and perspectives. Applied Energy, 268, 114946. https://doi.org/10.1016/j.apenergy.2020.114946.
29. Tang, L., Liu, J., Rong, A., Yang, Z. (2001). A review of planning and scheduling systems and methods for integrated steel production. European Journal of Operational Research, 133(1), 1-20. https://doi.org/10.1016/S0377-2217(00)00240-X.
30. Tao, F., Qi, Q., Liu, A., Kusiak, A. (2018). Data-driven smart manufacturing. Journal of Manufacturing Systems, 48, 157-169. https://doi.org/10.1016/j.jmsy.2018.01.006.
31. Tiwari, S. (2022). Artificial intelligence implications in engineering and production. Engineering Proceedings, 31(1), 16. https://doi.org/10.3390/ASEC2022-13823.
32. Villegas-Ch, W., Navarro, A.M., Sanchez-Viteri, S. (2024). Optimization of inventory management through computer vision and machine learning technologies. Intelligent Systems with Applications, 24, 200438. https://doi.org/10.1016/j.iswa.2024.200438.
33. Xia, Q., Jiang, C., Yang, C., Zheng, X., Pan, X., Shuai, Y., Yuan, S. (2019). A method towards smart manufacturing capabilities and performance measurement. Procedia Manufacturing, 39, 851-858. https://doi.org/10.1016/j.promfg.2020.01.415.
34. You, Z., Wu, C. (2019). A framework for data-driven informatization of the construction company. Advanced Engineering Informatics, 39, 269-277. https://doi.org/10.1016/j.aei.2019.02.002.
35. Yusoff, M.I.M. (2024). Machine learning: an overview. Open Journal of Modelling and Simulation, 12(03), 89-99. https://doi.org/10.4236/ojmsi.2024.123006.
36. Zermane, H., Madjour, H., Bouzghaya, M.A. (2022). Prediction of the amount of raw material in an Algerian cement factory. The Eurasia Proceedings of Science Technology Engineering and Mathematics, 19, 41-46. https://doi.org/10.55549/epstem.1218718.
37. Zermane, H., Madjour, H., Ziar, A., Zermane, A. (2024). Forecasting material quantity using machine learning and times series techniques. Journal of Electrical Engineering, 75(3), 237-248. https://doi.org/10.2478/jee-2024-0029.