

Multi-Metric Weighted Ensemble Learning for Train Delay Prediction

1st Masoomeh Golchehreh

School of Computing Sciences University of East Anglia
Norwich, United Kingdom
m.golchehreh@uea.ac.uk

2nd Wenjia Wang

School of Computing Sciences University of East Anglia
Norwich, United Kingdom
Wenjia.Wang@uea.ac.uk

Abstract—Train delays are a common problem in railway operations. Accurate delay prediction using machine learning can help operators adjust decisions to prevent or reduce delays. This paper proposes a Multi-Metric Weighted Ensemble Method (MWEM), a heterogeneous ensemble that combines diverse learning models through adaptive, performance-driven weighting for station-to-station delay prediction. The key idea is a metric-informed cross-validation scheme that assigns ensemble weights based on multiple evaluation criteria, enabling complementary model strengths to be integrated.

The proposed MWEM is evaluated on two real-world UK railway routes and compared with a range of baseline and ensemble models, including GB and XGB. Results show that MWEM consistently outperforms baseline and ensemble models across multiple evaluation metrics. These findings suggest that multi-metric, data-driven weighting provides an effective and lightweight approach for improving predictive reliability in railway delay forecasting.

Index Terms—Train delay prediction, Ensemble learning, Machine learning, Railway operations

I. INTRODUCTION

Accurate prediction in real-world machine learning remains challenging due to uncertainty, dynamic operational conditions, and heterogeneous data distributions. Model performance can vary across time, locations, and operational environments, making it difficult for a single model or evaluation metric to consistently perform well. From a computational intelligence perspective, this highlights the need for adaptive model combination strategies that can balance multiple aspects of predictive performance and improve prediction stability instead of relying on a single objective, particularly in operational environments where prediction accuracy and stability are both important.

In railway networks, train services operate under tightly constrained timetables and are highly sensitive to disruption [1], [2]. Even small delays can propagate across connected stations, leading to widespread performance degradation and reduced passenger satisfaction. Statistics from the UK Office of Rail and Road indicate that between 2020 and 2025, over 30% of the train services were delayed (arrived after 1 minute of the planned arrival time) [3]. These delays have caused serious disruptions to passengers' normal travel and high financial costs to the train operating companies (TOC). Therefore, accurate delay prediction may support train controllers in

adjusting operational decisions aimed at mitigating the spread of delays [4].

A wide range of techniques has been developed for train delay prediction [5]–[7]. However, despite the progress of existing machine learning and ensemble approaches, most methods rely on a single evaluation metric or fixed weighting schemes. Such strategies may overlook complementary aspects of model performance and limit the ability of ensembles to adapt to varying operational conditions [8]. Consequently, there is increasing interest in ensemble strategies that can balance multiple performance criteria and improve predictive stability across heterogeneous railway environments.

More advanced architectures have also been explored to improve predictive accuracy by incorporating hybrid learning schemes, including ensemble methods [9], [10]. Ensemble learning combines multiple models [11], and homogeneous and heterogeneous ensembles have both improved performance over individual predictors. However, existing ensemble approaches rely on fixed or heuristic weighting schemes or optimise performance using a single evaluation metric, which can overlook complementary strengths revealed under alternative performance criteria and may reduce robustness across different operational contexts.

To address these challenges and limitations, this paper introduces the *Multi-Metric Weighted Ensemble Method* (MWEM), which integrates diverse learning algorithms through an adaptive weighting mechanism. Rather than relying on fixed or heuristic combinations, the MWEM assigns weights to base models based on their validation performance across multiple evaluation criteria. This design aims to capture complementary strengths of heterogeneous models and improve prediction stability across station-to-station segments, thereby supporting more reliable train delay forecasting under varying operational conditions through a lightweight and data-driven ensemble weighting strategy rather than introducing a new learning paradigm.

The main contributions of this work are summarised as follows:

- A novel heterogeneous ensemble framework, namely MWEM, that utilises multiple evaluation metrics to adaptively determine the weights of member models, rather than relying on fixed ensemble configurations. This metric-informed weighting mechanism is designed

to capture complementary model strengths and enhance predictive robustness across varying operational conditions.

- A comprehensive empirical evaluation conducted using two real-world UK railway datasets and a diverse set of baseline models, including linear, tree-based, neural network, and ensemble methods. The results demonstrate that MWEM consistently improves predictive performance across multiple evaluation metrics and exhibits stable behaviour across different routes, network sections, and delay magnitudes, providing evidence of the effectiveness and practical applicability of the proposed novel multi-metric ensemble framework.

II. RELATED WORK

Research on train delay prediction has evolved alongside increasing network complexity and the availability of operational railway data. Early studies relied mainly on statistical and probabilistic formulations, while more recent work has adopted machine learning techniques to capture nonlinear delay patterns. Although these approaches have improved accuracy, empirical evidence shows that predictive performance often varies across network segments and operational contexts. This has motivated growing interest in ensemble-based approaches that combine multiple models to exploit complementary strengths. Accordingly, this section reviews statistical and machine learning methods for train delay prediction, with an emphasis on ensemble and hybrid learning frameworks.

A. Statistical and Probabilistic Delay Models

Initial work on train delay prediction focused on statistical formulations such as autoregressive models and Markov-based transition processes [4]. These approaches offered interpretable representations of temporal delay dynamics but were limited in modelling nonlinear interactions and cascading effects across interconnected stations. Probabilistic extensions, including Bayesian networks and stochastic delay propagation models, were later introduced to capture dependencies between trains and shared infrastructure [6], [12]. Despite improved uncertainty handling, these models often required strong domain assumptions and scaled poorly to large or heterogeneous networks.

B. Machine Learning for Delay Prediction

With the growing availability of historical and real-time railway data, machine learning (ML) approaches have become a dominant paradigm for delay prediction. Classical models such as decision trees, support vector machines, random forests, and shallow neural networks have been widely applied to learn nonlinear relationships between operational features and delay outcomes [1], [13], [14]. ML models' performance could be enhanced by incorporating timetable information and contextual variables such as weather conditions [15], [16].

More recently, deep learning architectures have been explored for capturing complex temporal and spatial dependencies. Recurrent neural networks, particularly LSTM-based

models, have been used to model delay evolution over time [17], while graph neural networks represent network structure and inter-station dependencies [18], [19]. Hybrid deep learning models combining convolutional and temporal components have further improved predictive accuracy [7]. Nevertheless, they usually require more computing resources than are available in the rail industry. Comparative studies [1], [2], [20] indicate that different ML paradigms perform differently under different conditions and no single model consistently dominates.

C. Ensemble and Hybrid Learning Frameworks

To address variability in individual model performance, ensemble learning has been increasingly adopted in railway delay prediction. Homogeneous ensembles, such as gradient boosting and random forests, have demonstrated strong predictive capability by aggregating multiple learners of the same type [5], [21]. Heterogeneous ensembles extend this idea by combining models from different algorithmic families, allowing complementary behaviours to be exploited [10], [11].

Hybrid learning frameworks further integrate predictive models with optimisation or decision-making components, including reinforcement learning for timetable adjustment [9] and optimisation-based approaches for delay mitigation and predictive maintenance [22]. However, many ensemble designs still rely on fixed or manually specified weighting schemes [23], [24], limiting their ability to adapt to varying data characteristics.

D. Motivation for Metric-Weighted Ensemble

Given the heterogeneous nature of railway operations and the absence of a universally dominant predictive model, there is strong motivation for ensemble strategies that adaptively combine diverse learners. Assigning ensemble weights based on validation performance across multiple evaluation criteria provides a principled, data-driven mechanism for emphasising stronger contributors while reducing weaker ones. Motivated by these considerations, this paper proposes a metric-weighted heterogeneous ensemble framework that leverages multi-criteria validation to guide adaptive weight assignment.

III. MULTI-METRIC WEIGHTED ENSEMBLE METHOD (MWEM)

This section describes the proposed Multi-Metric Weighted Ensemble Method (MWEM) for predicting train delays between consecutive stations. The method models delay evolution along a route using segment-level operational data and a set of standard regression models. These base models are then combined using the proposed MWEM, which assigns weights based on validation performance across multiple evaluation metrics. The key components of the modelling pipeline are described in the following subsections.

A. Problem Formulation

Train delay prediction is treated as a supervised regression task between consecutive stations. Delays are predicted step

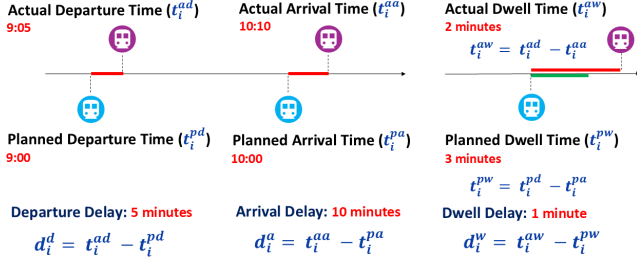


Fig. 1. Illustration of arrival, departure, and dwell delays based on planned and actual train timings at a station.

by step along a route, which allows local effects to be captured while delays naturally propagate from one station to the next. At each station, planned and actual arrival and departure times are available. From these, arrival, departure, and dwell delays are defined to describe timing deviations, as illustrated in Fig. 1. The prediction task is to estimate the arrival delay at the next station using information available up to the current station.

B. Feature Engineering and Preprocessing

A unified set of features was constructed to represent operational behaviour and temporal conditions affecting train delays, including arrival, departure, and dwell delays, preceding checkpoint delays, hour of day, weekday/weekend, month, public holidays, peak and off-peak periods, and historical mean delays. After removing inconsistencies and duplicates, missing values were addressed using forward propagation and statistical imputation. Outliers were filtered, and features were normalised where required.

C. Baseline Models

A diverse set of regression models was used as benchmarks, including linear methods (Linear Regression (LR), Bayesian Ridge Regression (BR), Elastic Net Regression (ENet)), tree-based learners (Decision Tree (DT)), kernel-based models (Support Vector Regression (SVR), Kernel Ridge Regression (KR)), instance-based k-Nearest Neighbours (KNN), Multi-Layer Perceptron (MLP), and several ensemble-based methods: Random Forest (RF), Gradient Boosting (GB), Extra Trees (ET), AdaBoost (AB), and Extreme Gradient Boosting (XGB). XGB is widely used as a strong tree-based boosting method and has demonstrated competitive performance across a range of structured prediction problems. Recent studies [2], [20], [25] emphasise the value of maintaining diverse baselines, as different routes and operational settings may favour different model families, providing a reliable basis for evaluating the proposed ensemble.

D. Heterogeneous Ensemble Framework

Performance of individual models varies significantly across routes, peak/off-peak periods, and data availability. Linear models may fail to capture nonlinear relationships, tree-based models may overfit sparse data, and neural networks require

extensive tuning. To mitigate these limitations, heterogeneous ensembles combine multiple model families to exploit complementary strengths. A simple Average Ensemble (AE) serves as a baseline, while MWEM introduces adaptive weighting based on multi-metric performance.

E. Average Ensemble (AE)

AE combines predictions from all base models by assigning equal weight to each and taking their mean. While simple and easy to implement, this approach assumes all models are equally reliable, regardless of their individual performance.

F. Proposed MWEM Framework

The ensemble strategy proposed in this study applies a metric-weighted approach that employs cross-validation to assign weights to base models (Algorithm 1). Unlike traditional ensemble methods that apply equal weights or rely on a single evaluation criterion, MWEM evaluates models across multiple performance metrics and integrates these results into the weighting scheme. The justification is that different metrics highlight different aspects of model performance, so this design enhances the ensemble's ability to adapt across varied conditions and improves both reliability and predictive accuracy.

Given a training dataset $D_{\text{train}} = [X_{\text{train}}, Y_{\text{train}}]$ and a set of candidate base models $l_1, l_2, \dots, l_m$, the method performs f -fold cross-validation. Unlike conventional cross-validation, where a single evaluation metric is applied across all folds, MWEM deliberately assigns a different evaluation metric E_i to each fold. For example, one fold may use Mean Absolute Error (MAE), another Root Mean Squared Error (RMSE), and another the coefficient of determination (R^2). This design ensures that complementary aspects of predictive performance are incorporated into the weighting process, rather than biasing weight estimation toward a single error criterion.

Different evaluation metrics capture complementary aspects of predictive performance. For example, MAE reflects average prediction error, RMSE penalises larger deviations, and R^2 measures the proportion of variance explained by the model. Assigning different metrics across cross-validation folds, therefore, enables the ensemble weighting process to consider diverse performance characteristics. The final weights are obtained by aggregating fold-specific weights, which helps balance the influence of individual models and improves the robustness and stability of predictions across varying operational conditions.

For building the ensemble, the dataset D_{train} is partitioned into f folds. $(f - 1)$ folds are used as the training subset T_i^D , and the remaining fold is used as the validation subset V_i^D . This procedure is repeated f times in a round-robin manner. In each iteration, the base models are trained on T_i^D and evaluated on V_i^D . Models are then ranked from highest to lowest performance. A candidate model subset is constructed iteratively by sequentially adding the best-ranked models, with ensemble predictions recalculated after each inclusion. Models

are retained only if their addition leads to measurable improvement according to the fold-specific metric. Stronger models, therefore, receive higher weights, while models with negligible contribution are assigned zero weight. This guarantees that non-contributing models do not influence the final ensemble.

The fold-specific weight vectors are aggregated into an overall weight vector $\mathcal{W} = [\omega_1^*, \dots, \omega_m^*]$ by averaging across folds.

For higher-is-better metrics, the weight assigned to model l_j in fold i is defined as:

$$\Omega_{i,j}^* = \frac{E_i(l_j)}{\sum_{k=1}^{|\Phi|} E_i(l_k)} \quad (1)$$

where negative values of $E_i(l_j)$ (e.g., negative R^2 scores) are clipped to zero before normalisation. This ensures non-negative ensemble weights and improves numerical stability.

and for lower-is-better metrics, the inverse form is applied:

$$\Omega_{i,j}^* = \frac{1/(E_i(l_j) + \epsilon)}{\sum_{k=1}^{|\Phi|} 1/(E_i(l_k) + \epsilon)} \quad (2)$$

where $\epsilon > 0$ is a small smoothing constant introduced to improve numerical stability and prevent excessively large weights when the evaluation error approaches zero.

The fold-specific weights are then aggregated into:

$$\omega_k^* = \frac{1}{f} \sum_{i=1}^f \Omega_{i,k}^* \quad (3)$$

At inference time, predictions from the selected base models are combined using the learned weights to produce the final ensemble output.

G. Evaluation Metrics

Predictive performance was assessed using a set of widely adopted regression metrics commonly used in railway delay forecasting. These include Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2), which together capture different aspects of accuracy and error behaviour [27]. To improve robustness assessment under noisy conditions, Huber loss was also considered [28]. In addition, a tolerance-based accuracy measure, denoted as $ACC@k$, was used to quantify the proportion of predictions falling within a predefined error margin (e.g., 1, 3, or 5 minutes), which is particularly relevant in railway operational contexts [23]. Lower values of MAE, RMSE, and Huber loss indicate better performance, while higher values of R^2 and $ACC@k$ reflect improved predictive accuracy.

It is important to distinguish between metrics used for ensemble weight learning and those used for final performance reporting. In this study, five evaluation metrics (MAE, RMSE, R^2 , Huber loss, and $ACC@k$) are used during the MWEM weight estimation process to capture complementary performance characteristics. For final model comparison, a subset of widely adopted metrics (MAE, RMSE, R^2 , and $ACC@1$) is reported to maintain clarity and comparability with existing railway delay prediction literature.

Algorithm 1: Multi-Metric Weighted Ensemble Method

Input: D_{train} (training data); $L = \{l_1, \dots, l_m\}$ (base models); $E = \{E_1, \dots, E_f\}$ (evaluation metrics)
Output: Ensemble $\mathcal{E} = \{L, \mathcal{W}\}$ with final weight vector $\mathcal{W}$
for $i \leftarrow 1$ **to** f **do** // f -fold cross-validation
 Split D_{train} into training set T_i^D and validation set V_i^D ;
 foreach $l_j \in L$ **do**
 Train l_j on T_i^D ;
 Evaluate l_j on V_i^D using metric E_i ;
 Rank all models from best to worst according to E_i ;
 Initialise candidate ensemble $\Phi \leftarrow \emptyset$;
 Initialise best ensemble score θ_{best} ;
 for models l_j in ranked order **do**
 Add l_j to Φ ;
 Compute ensemble predictions on V_i^D ;
 Evaluate ensemble performance under E_i ;
 if ensemble performance improves **then**
 update θ_{best} ;
 else
 remove l_j from Φ ;
 Compute fold-specific weights for models in Φ using Eqs. (1) or (2);
 Assign zero weight to all models $l_j \notin \Phi$;
Aggregate fold-specific weights across all folds using Eq. (3);

IV. DATA AND EXPERIMENTAL SETUP

Two real-world datasets from UK railway operations were used, each representing a service route operated by a different Train Operating Company (TOC) [29]. The NL dataset covers a short commuter route with eight stations over six months and is characterised by generally low delays, while the WW dataset spans a longer, busier corridor with 20 stations over two years and exhibits higher delay variability.

Both datasets were preprocessed by removing duplicate and inconsistent records, imputing missing values, and excluding implausible delay observations. Delay measurements were standardised to minutes, and continuous features were normalised where required.

All models were trained and evaluated separately per route using five-fold cross-validation. Given the temporal nature of railway delay data, this evaluation provides a general comparison of model performance across the available dataset.

V. RESULTS AND DISCUSSION

This section reports the experimental results of the proposed MWEM. For comparison, results from the AE and the baseline regression models introduced in Section III-C are also presented. Performance is evaluated using R^2 , MAE, RMSE, and $ACC@1$. Unless stated otherwise, reported values correspond to averages across both routes and all station-to-station sections.

A. Overall Model Performance

Table I summarises the mean predictive performance of all models. Among the baseline regressors, decision tree-based

ensemble methods such as GB, RF and XGB achieved comparatively strong results, while linear and instance-based models such as ENet and KNN performed less well. The simple AE outperformed most individual models, confirming the benefit of aggregating predictions from diverse learners. The proposed MWEM consistently achieved the best overall performance across all evaluation metrics. Compared with the strongest single baseline (GB), MWEM reduced RMSE by approximately 8% while achieving higher R^2 and $ACC@1$. These results demonstrate that metric-weighted model aggregation provides a systematic advantage over both individual predictors and uniform averaging. Reported values are averages across all route-wise folds and sections. Similar ranking tendencies were also observed for MAE and R^2 , although RMSE was selected as the primary metric for formal rank-based testing.

TABLE I
MEAN TEST PERFORMANCE PER MODEL AND TWO DATASETS. OUR PROPOSED MWEM PRODUCED THE BEST RESULTS (IN BOLD) ON ALL THE MEASURES.

Model	R^2	MAE	RMSE	$ACC@1$ (%)
LR	0.895	0.696	1.163	90.991
ENet	0.735	1.164	2.137	81.224
BR	0.895	0.695	1.163	90.999
SVR	0.826	0.761	1.728	89.697
KNN	0.665	1.361	2.438	74.541
MLP	0.896	0.691	1.158	90.722
DT	0.789	0.843	1.689	86.208
RF	0.885	0.664	1.258	91.445
AB	0.860	0.864	1.444	87.645
ET	0.865	0.719	1.342	89.150
GB	0.890	0.650	1.213	91.747
XGB	0.888	0.656	1.230	91.746
AE	0.900	0.658	1.160	91.719
MWEM	0.903	0.637	1.119	92.094

B. Section-Level Performance

Figure 2 presents section-level R^2 and RMSE values for both routes, where MWEM consistently outperformed or matched the best baseline models across most sections and exhibited more stable behaviour than individual regressors, particularly in operationally complex segments (e.g., NL6, WW9, WW14, WW18) characterised by higher traffic density, shared junctions, and accumulated upstream delays that introduce irregular and nonlinear delay patterns.

C. Performance Across Delay Magnitudes

Model performance was further examined across increasing delay magnitudes by analysing cumulative RMSE over positive delay bins, where zero indicates on-time operation. Delays were grouped into successive ranges from 1 to 30 minutes, with an additional bin for delays exceeding 30 minutes. As shown in Fig. 3, RMSE increases steadily with delay severity for all models, reflecting the growing difficulty of predicting larger delays. This increase is most pronounced for weaker baselines such as KNN and ENet, while tree-based models (GB, RF, XGB) display more gradual degradation. Across all delay ranges, MWEM consistently achieves the lowest RMSE

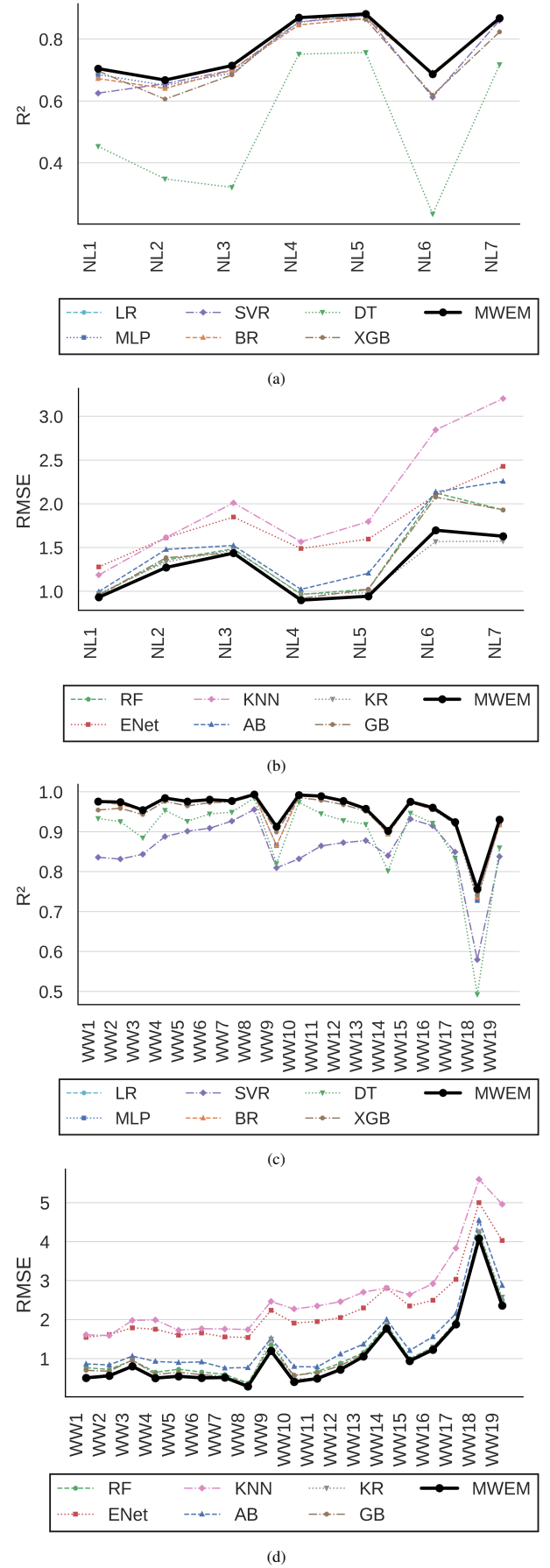


Fig. 2. Section-level performance comparison across the two UK rail routes: (a) NL — R^2 , (b) NL — RMSE, (c) WW — R^2 , (d) WW — RMSE.

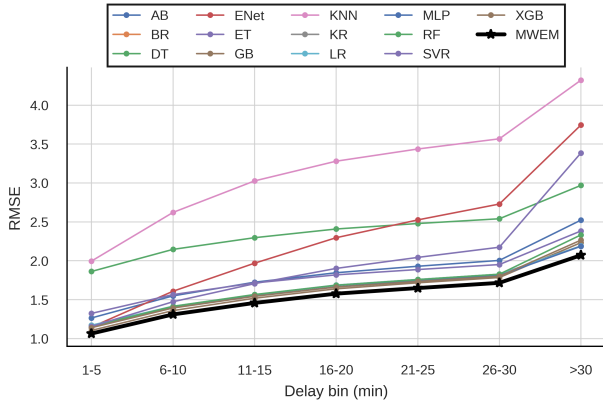


Fig. 3. Cumulative RMSE across increasing delay bins.

and exhibits the slowest error growth, indicating that metric-weighted aggregation effectively balances model contributions and limits error escalation for more severe delays.

D. Taylor Diagram Analysis

To provide a compact and integrated comparison of model performance, a Taylor diagram was used to jointly assess agreement between predicted and observed delays in terms of standard deviation, root mean square deviation (RMSD), and correlation coefficient. Results are aggregated across all routes, sections, and cross-validation folds, providing a holistic view of predictive behaviour. Figure 4 shows the Taylor diagram for all baseline models and MWEM. The reference point represents the observed delays, characterised by unit correlation and zero RMSD. Models located closer to this point exhibit better overall agreement, combining high correlation, low error, and appropriate variance. MWEM is positioned closest to the reference, indicating the most favourable balance between error reduction and variance preservation while maintaining the highest correlation with observations. Strong decision tree-based ensemble models such as XGB, RF, and GB also perform competitively, but remain farther from the reference due to higher RMSD or variance mismatch.

E. Statistical Significance Analysis

Statistical significance of performance differences was assessed using the Friedman test followed by Nemenyi's post-hoc analysis at $\alpha = 0.05$. Models were ranked per section using RMSE, and average ranks were computed. The Critical Difference diagram (Figure 5) shows that MWEM achieved a significantly better average rank than most baseline models, especially better than the ensemble models - XGB and GB, which formed a non-significant cluster with linear regressors, while weaker learners such as DT, ENet, and KNN were clearly separated and statistically inferior. These results confirm that the observed improvements of the MWEM are statistically consistent and significant, not due to randomness.

F. Discussion

The results indicate that the MWEM consistently produces competitive or slightly improved performance compared

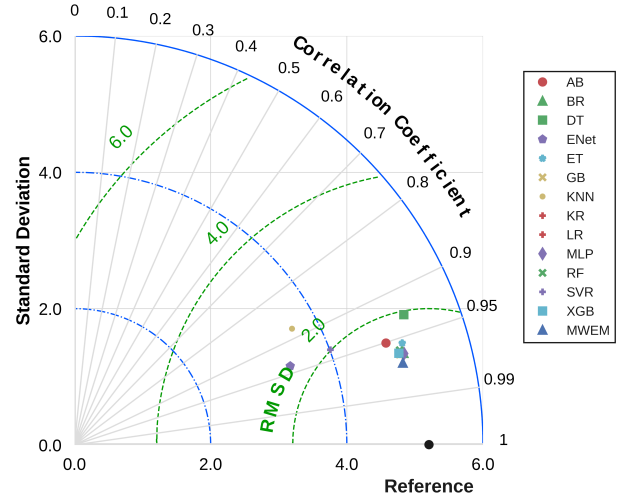


Fig. 4. Taylor diagram comparing baseline models and MWEM.

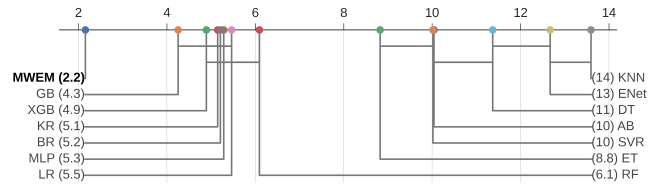


Fig. 5. Nemenyi Critical Difference (CD) diagram based on RMSE. Lower ranks indicate better performance.

with individual baselines and existing ensemble approaches across the evaluated settings [11], [23], [30]. The observed improvements across multiple metrics and network sections suggest that multi-metric weighting can help balance complementary model strengths and improve predictive stability. By assigning weights based on validation performance under multiple evaluation criteria, MWEM captures complementary strengths across learning algorithms and avoids reliance on fixed combination rules.

These findings align with prior evidence that ensemble learning can improve reliability in transport delay prediction. In addition, the proposed framework remains lightweight and straightforward to integrate with existing machine learning pipelines, as the final predictor is obtained through weighted aggregation of base model outputs. However, while the performance differences are consistent, further evaluation is required to assess the operational impact of these improvements and to validate the approach under broader deployment conditions.

VI. CONCLUSION AND FURTHER WORK

This paper presented the Multi-Metric Weighted Ensemble Method (MWEM) for station-to-station train delay prediction. Using two real-world UK railway routes, the results showed that MWEM achieved competitive and consistently improved performance compared with individual baseline models and existing ensemble approaches across multiple evaluation metrics. The findings suggest that assigning ensemble weights through adaptive, metric-informed validation can help balance

complementary model strengths and improve prediction stability relative to individual predictors or uniform weighting schemes.

By integrating diverse learning algorithms, the MWEM provides a lightweight and computationally efficient framework for train delay forecasting. The weighted aggregation design allows the approach to be integrated with existing machine learning pipelines, supporting potential use in operational decision-support and timetable management applications. However, further evaluation is required to assess the broader operational impact and robustness of the proposed framework under different deployment conditions.

Future work will extend the evaluation to additional datasets and longer time horizons to further assess scalability. Incorporating richer contextual features, such as infrastructure characteristics and external disruption indicators, may further improve predictive performance. Another promising direction is the integration of the MWEM with real-time railway monitoring systems and online learning mechanisms, enabling adaptive model updates as new operational data becomes available. Probabilistic extensions of the framework could also be explored to support uncertainty-aware decision-making in railway operations. In addition, future work will investigate time-aware validation strategies, such as rolling or blocked validation, to better reflect real-world forecasting scenarios. These evaluation approaches may provide additional insight into model robustness under realistic operational conditions.

REFERENCES

- [1] T. Spanninger, A. Trivella, B. Büchel, F. Corman, "A review of train delay prediction approaches," *J. Rail Transp. Plann. Manag.*, vol. 22, p. 100312, 2022.
- [2] K. Y. Tiong, Z. Ma, C.-W. Palmqvist, "A review of data-driven approaches to predict train delays," *Transp. Res. C, Emerg. Technol.*, vol. 148, p. 104027, 2023.
- [3] Office of Rail and Road, "How many trains arrive on time?," Accessed: Sep. 13, 2025. [Online]. Available: <https://dataportal.orr.gov.uk/popular-statistics/how-many-trains-arrive-on-time/>
- [4] J. Xu, W. Wang, Z. Gao, H. Luo, Q. Wu, "A novel Markov model for near-term railway delay prediction," *Comput. Ind. Eng.*, vol. 181, p. 109302, 2023.
- [5] H. Laifa, H. H. B. Ghezalaa, et al., "Train delay prediction in Tunisian railway through LightGBM model," *Procedia Comput. Sci.*, vol. 192, pp. 981–990, 2021.
- [6] P. Huang, C. Wen, L. Fu, J. Lessan, C. Jiang, Q. Peng, X. Xu, "Modeling train operation as sequences: A study of delay prediction with operation and weather data," *Transp. Res. E, Logist. Transp. Rev.*, vol. 141, p. 102022, 2020.
- [7] D. Wang, J. Guo, C. Zhang, "A novel hybrid deep learning model for complex systems: A case of train delay prediction," *Adv. Civ. Eng.*, vol. 2024, Art. no. 8163062, 2024.
- [8] S. Alsulamy, "Predicting construction delay risks in Saudi Arabian projects: A comparative analysis of CatBoost, XGBoost, and LGBM," *Expert Syst. Appl.*, vol. 268, p. 126268, 2025.
- [9] Z. Pang, L. Wang, L. Li, "A hybrid machine learning approach for train trajectory reconstruction under interruptions considering passenger demand," *Int. J. Rail Transp.*, vol. 13, no. 2, pp. 352–380, 2025.
- [10] P. Zhou, L. Chen, X. Dai, B. Li, T. Chai, "Intelligent prediction of train delay changes and propagation using RVFLNs with improved transfer learning and ensemble learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 12, pp. 7432–7444, 2020.
- [11] M. Al Ghamdi, G. Parr, W. Wang, "Heterogeneous machine learning ensembles for predicting train delays," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 6, pp. 5138–5153, 2024.
- [12] F. Corman, P. Kecman, "Stochastic prediction of train delays in real-time using Bayesian networks," *Transp. Res. C, Emerg. Technol.*, vol. 95, pp. 599–615, 2018.
- [13] Z. Li, C. Wen, R. Hu, C. Xu, P. Huang, X. Jiang, "Near-term train delay prediction in the Dutch railways network," *Int. J. Rail Transp.*, vol. 9, no. 6, pp. 520–539, 2021.
- [14] P. Lapamompinyo, S. Derrible, F. Corman, "Real-time passenger train delay prediction using machine learning: A case study with Amtrak passenger train routes," *IEEE Open J. Intell. Transp. Syst.*, vol. 3, pp. 539–550, 2022.
- [15] Ö. Ardiç, "Forecasting train arrival delays on the Ankara–Eskişehir high-speed line in Turkey," *J. Rail Transp. Plann. Manag.*, vol. 30, p. 100448, 2024.
- [16] W. Barbour, J. C. M. Mori, S. Kuppa, D. B. Work, "Prediction of arrival times of freight traffic on US railroads using support vector regression," *Transp. Res. C, Emerg. Technol.*, vol. 93, pp. 211–227, 2018.
- [17] C. Wen, W. Mou, P. Huang, Z. Li, "A predictive model of train delays on a railway line," *J. Forecast.*, vol. 39, no. 3, pp. 470–488, 2020.
- [18] Y. D. Zhang, L. Liao, Q. Yu, W. G. Ma, K. H. Li, "Using the gradient boosting decision tree (GBDT) algorithm for a train delay prediction model considering the delay propagation feature," *Adv. Prod. Eng. Manag.*, vol. 16, no. 3, pp. 285–296, 2021.
- [19] J. Li, X. Xu, R. Shi, X. Ding, "Train arrival delay prediction based on spatial-temporal graph convolutional network to sequence model," in *Proc. IEEE Int. Intell. Transp. Syst. Conf. (ITSC)*, 2021, pp. 2399–2404.
- [20] T. K. Yong, Z. Ma, C.-W. Palmqvist, "AP-GRIP evaluation framework for data-driven train delay prediction models: Systematic literature review," *Eur. Transp. Res. Rev.*, vol. 17, no. 1, p. 13, 2025.
- [21] Y. D. Zhang, L. Liao, Q. Yu, W. G. Ma, K. H. Li, "Using the gradient boosting decision tree (GBDT) algorithm for a train delay prediction model considering the delay propagation feature," *Adv. Prod. Eng. Manag.*, vol. 16, no. 3, pp. 285–296, 2021.
- [22] M. Khalilzadeh, D. Pamucar, A. Heidari, "Reducing train delays with machine learning-based predictive maintenance for railways," *Decis. Mak.: Appl. Manag. Eng.*, vol. 8, no. 2, pp. 265–284, 2025.
- [23] M. Al Ghamdi, G. Parr, W. Wang, "Weighted ensemble methods for predicting train delays," in *Proc. Int. Conf. Comput. Sci. Appl.*, 2020, pp. 586–600.
- [24] C. Srinivas, K. Asritha, S. R. Legala, S. Katarpu, et al., "Ensemble learning for accurate train delay prediction," 2025.
- [25] M. Sarhani, S. Voß, "Prediction of rail transit delays with machine learning: How to exploit open data sources," *Multimodal Transp.*, vol. 3, no. 2, p. 100120, 2024.
- [26] I. D. Mienye, Y. Sun, "A survey of ensemble learning: Concepts, algorithms, applications, and prospects," *IEEE Access*, vol. 10, pp. 99129–99149, 2022.
- [27] A. Botchkarev, "Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology," *arXiv*, arXiv:1809.03006, 2018.
- [28] Q. Sun, W.-X. Zhou, J. Fan, "Adaptive Huber regression," *J. Amer. Stat. Assoc.*, vol. 115, no. 529, pp. 254–265, 2020.
- [29] Rail Delivery Group / National Rail, "National Rail Open Data," Accessed: Sep. 13, 2025. [Online]. Available: <https://opendata.nationalrail.co.uk/>
- [30] V. Boateng, B. Yang, "Ensemble stacking with the multi-layer perceptron neural network meta-learner for passenger train delay prediction," in *Proc. IEEE Conf. Artif. Intell. (CAI)*, 2023, pp. 21–22.