Proceedings of the 10th World Congress on Civil, Structural, and Environmental Engineering (CSEE 2025)
Barcelona, Spain - April, 2025
Paper No. ICSECT 170
DOI: 10.11159/icsect25.170

Data-Driven Strength Prediction of Recycled Aggregate Concrete: Insights from Boosting-Based Machine Learning Models

Mahan Samiadel[1], Farahnaz Soleimani[1]
[1]Oregon State University
Kearney Hall, 1491 SW Campus Way, Corvallis, Oregon, USA
samiadem@oregonstate.edu; farahnaz.soleimani@oregonstate.edu

Abstract **-** Accurate prediction of the compressive strength (CS) of recycled aggregate concrete (RAC) is crucial for optimizing mix design and ensuring structural integrity. This study compares the predictive performance of six tree-based and ensemble learning models—Decision Tree, Random Forest, Adaptive Boosting, Gradient Boosting, Light Gradient Boosting Machine, and Extreme Gradient Boosting—using a dataset comprising RAC compositions and testing age. The models are evaluated based on predicted versus actual CS values, residual distributions, and statistical performance metrics, including the coefficient of determination (R^2) and root mean squared error (RMSE). The results indicate that boosting-based models, particularly Extreme Gradient Boosting and Light Gradient Boosting Machine, achieve the highest predictive accuracy, with R^2 values of 0.94 and the lowest RMSE scores, demonstrating their effectiveness to capture complex nonlinear relationships. In contrast, Decision Tree and Adaptive Boosting exhibit greater variance and lower reliability, primarily due to their sensitivity to data partitioning and noise. These findings underscore the effectiveness of ensemble learning techniques in predicting RAC properties and highlight the potential for further improvements through hybrid modeling approaches and hyperparameter optimization. This study contributes to advancing sustainable construction practices by enhancing the accuracy and reliability of machine learning-based predictive models for recycled concrete applications.

Keywords**:** Recycled Aggregate Concrete, Machine Learning, Ensemble Learning, Compressive Strength Prediction, Gradient Boosting, Extreme Gradient Boosting, Sustainable Construction, Data-Driven Modeling

1 Introduction

RAC is a sustainable material that replaces natural aggregates with recycled ones from construction and demolition waste, reducing landfill use and conserving resources [1-2]. Due to adhered mortar, recycled coarse aggregates exhibit higher water absorption, lower density, and increased porosity, affecting RAC's mechanical properties, including compressive and tensile strength, carbonation, and chloride penetration. However, optimized mix design, aggregate pre-treatment, and admixtures can enhance its performance to match conventional concrete [1], [3]. Studies have shown that seismically detailed RAC columns exhibit comparable strength and ductility to those made with natural aggregates, confirming its suitability for structural applications [4]. RAC is effectively used in pavements and structural elements like beams, with up to 25% recycled aggregates showing minimal impact on shear strength [2]. Its adoption supports sustainability by reducing carbon emissions and environmental impact [1], [3].

The application of machine learning (ML) techniques for predicting the mechanical properties of RAC has gained significant attention due to their ability to model complex relationships. Behnood et al. utilized the M5' model tree algorithm to predict the elastic modulus of RAC, achieving an accuracy of over 80%, outperforming traditional regression models that often fail to account for the unique properties of recycled aggregates [5], [6]. Similarly, Duan et al. demonstrated the effectiveness of artificial neural networks in predicting the elastic modulus by training models with comprehensive datasets, which included diverse sources of recycled aggregates, ensuring broad applicability [7].

In the context of resilient modulus prediction for pavement applications, Kaloop et al. [8] compared ANN models with regression techniques, concluding that ANNs provide more accurate estimations for blends of recycled concrete and clay masonry, highlighting their adaptability to different mix designs and material combinations. For predicting nominal shear strength, Ababneh et al. [9] employed ANN models to estimate the contribution of RAC to the shear capacity of beams without transverse reinforcement. The study confirmed that ANN models could closely replicate experimental results, with variations as low as 8%, making them a reliable tool for structural applications.

The integration of ML models like ANNs and M5' trees with parametric sensitivity analysis, as demonstrated by Xu et al. [10], further enhances the understanding of how different factors—such as aggregate properties and mix proportions—influence RAC's mechanical behavior. These advanced techniques not only provide accurate predictions but also facilitate optimization in the design of RAC mixtures.

While ML models like ANNs and M5' trees have shown high accuracy in predicting RAC properties, their reliance on diverse and high-quality datasets limits their practicality. Additionally, challenges like overfitting and adapting to field conditions remain underexplored, requiring further investigation to enhance model robustness and real-world applicability.

ML techniques have proven to be highly effective in predicting the CS of RAC, leveraging their ability to handle the complex, nonlinear relationships between various influencing factors. Duan et al. [11] utilized artificial neural networks with 14 input parameters, including water-cement ratio, aggregate types, and replacement ratios, achieving high accuracy in predicting the 28-day CS of RAC using datasets from multiple sources. Similarly, Khademi et al. [12] compared the performance of ANN, adaptive neuro-fuzzy inference systems, and multiple linear regression, concluding that ANN outperformed others with a determination coefficient of 0.92.

Advanced approaches such as deep learning and ensemble models have also been applied. Deng et al. developed a deep learning model incorporating convolutional neural networks, demonstrating superior prediction precision and efficiency compared to traditional ANNs [13]. Furthermore, Hoang [14] proposed an ant colony-optimized extreme gradient boosting machine model, which achieved outstanding accuracy with an RMSE of 4.98 and demonstrated significant improvements over baseline models.

To further improve predictions, studies like those by Hosseinzadeh et al. [15] combined recycled aggregates with supplementary materials like fly ash, using ML algorithms such as random forests and extreme gradient boosting to achieve prediction accuracies exceeding 95%, thus demonstrating the adaptability of ML to varying RAC compositions and properties. These advancements highlight the critical role of ML in optimizing RAC mix designs and reducing the need for exhaustive experimental testing.

ML improves RAC CS prediction but faces challenges like data inconsistency and computational demands. This study compares tree-based and ensemble models, including decision tree, random forest, adaptive boosting, gradient boosting, light gradient boosting machine, and extreme gradient boosting, using a dataset with RAC components and testing age. The research evaluates predictive accuracy, efficiency, and robustness, aiming to enhance mix design optimization and support sustainable construction.

2 Dataset description

In this study, a dataset of 1,100 samples sourced from the work of Hoang [14] was used. This dataset was collected from 46 different sources. This dataset includes input and output variables relevant to RAC. The input variables consist of the contents of cement (C), silica fume (SF), fly ash (FA), water (W), natural fine aggregate (NFA), natural coarse aggregate (NCA), recycled fine aggregate (RFA), recycled coarse aggregate (RCA), and the age of testing (A), measured in days. The output variable is the CS of the concrete, recorded in MPa. These variables, detailed in Table 1, provide a comprehensive representation of the components and testing conditions for RAC mixes, enabling robust modeling and prediction of its mechanical properties.

Table 2 presents descriptive statistics for the variables, highlighting their variability. For instance, the cement content (C) ranges from 140 to 600 kg/m³, with a mean of 356 kg/m³ and a standard deviation of 72.7 kg/m³. The recycled coarse aggregate (RCA) shows a broad range, from 0 to 1632 kg/m³, with an average of 518.5 kg/m³. The testing age (A) spans from 1 to 180 days, with a median of 28 days, reflecting the diversity in curing periods. Variables such as silica fume (SF) and recycled fine aggregate (RFA) exhibit skewness, indicating asymmetry in their distributions. This variability across inputs ensures the dataset's suitability for evaluating complex, nonlinear relationships in RAC properties.

Table 1: Variables, notations, and units

	Variable	Notation	Unit
Input Variables	Cement Content	**C**	Kg/m^3
	Silica Fume Content	**SF**	Kg/m^3
	Fly Ash Content	**FA**	Kg/m^3
	Water Content	**W**	Kg/m^3
	Natural Fine Aggregate	**NFA**	Kg/m^3
	Natural Coarse Aggregate	**NCA**	Kg/m^3
	Recycled Fine Aggregate	**RFA**	Kg/m^3
	Recycled Coarse Aggregate	**RCA**	Kg/m^3
	Age of Testing	**A**	Days
Output Variable	Compressive Strength	**CS**	MPa

Table 2: Descriptive statistics for the variables

	C	SF	FA	W	NFA	NCA	RFA	RCA	A
Max	600.0	50.0	227.5	271.0	1065.0	1366.0	1000.0	1632.0	180.0
Min	140.0	0.0	0.0	120.0	0.0	0.0	0.0	0.0	1.0
Average	356.0	1.1	27.3	193.3	650.1	529.9	33.0	518.5	32.7
Mode	350.0	0.0	0.0	205.0	642.0	0.0	0.0	0.0	28.0
Median	361.0	0.0	0.0	193.1	685.0	543.2	0.0	496.5	28.0
StD	72.7	6.1	58.1	26.0	222.2	447.2	134.7	427.7	36.4
Skewness	-0.9	6.1	2.0	-0.3	-1.4	0.0	4.5	0.3	2.1

The pairplots, correlation matrix, and variables distributions are all shown in Fig. 1 to explore relationships and patterns among the variables. In the pairplots, CS shows a moderate positive relationship with cement content and age of testing, indicating their significant role in enhancing RAC strength. However, variables like silica fume, fly ash, and recycled fine aggregate exhibit a more scattered relationship with CS, suggesting their weaker influence or potential non-linear effects. Notably, recycled coarse aggregate shows a slight negative trend with CS, potentially reflecting the adverse impact of adhered mortar or lower quality of recycled materials.

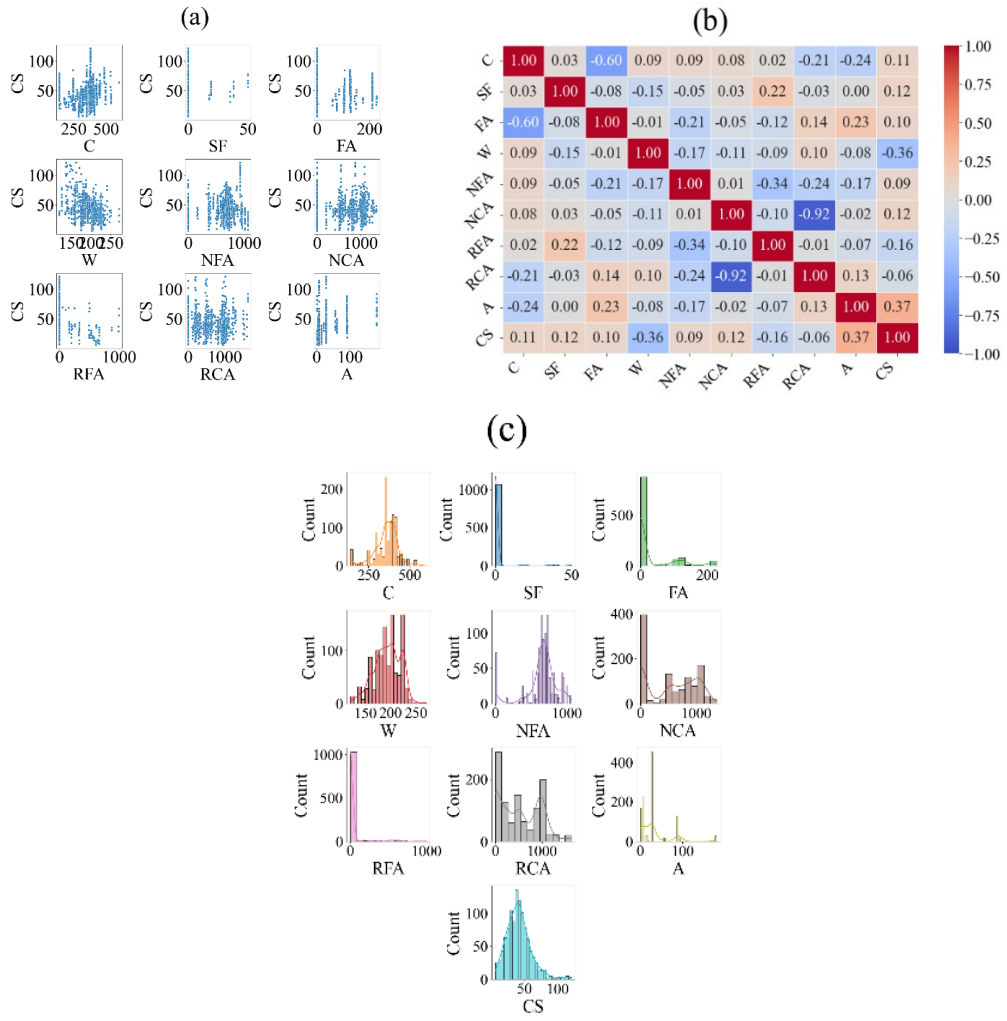

Fig. 1: a) Pairplots, b) Correlation matrix, and c) Distributions of variables

The correlation heatmap further quantifies these observations. Cement content and age exhibit moderate positive correlations with CS (0.37 and 0.36, respectively). Conversely, RCA and RFA show weak negative correlations (-0.24 and -0.21), indicating their detrimental effect on CS. Other variables, such as natural coarse aggregate and water content, exhibit negligible correlations with CS, highlighting their limited direct impact in the current dataset.

The variable distributions reveal significant variability among the features. Cement content and water content exhibit relatively normal distributions, while RCA and age display skewed distributions, indicating the predominance of certain mix designs or testing durations. The CS distribution shows a slightly right-skewed pattern, reflecting higher frequencies of lower-strength RAC samples. These insights underline the diverse and complex nature of the dataset, which is well-suited for exploring the predictive capabilities of ML models.

3 Methodology

The prediction capability of six tree-based and ensemble learning models using the described RAC dataset is comprehensively investigated in this study. Specifically, Decision Tree, Random Forest, Adaptive Boosting, Gradient Boosting, Light Gradient Boosting Machine, and Extreme Gradient Boosting are employed to analyze the complex relationships between the components of RAC, testing age, and the resulting CS. Models are systematically optimized and evaluated to ensure robust and reliable predictions. Each algorithm is briefly explained below to provide a foundational understanding of its working principles and relevance to the analysis.

To evaluate the models' performance, two key metrics are utilized: RMSE, which measures the average magnitude of prediction errors, and the R^2 score, which assesses how well the model explains the variability of the target variable. A 10-fold cross-validation approach is adopted to optimize hyperparameter values for each model and ensure generalization by minimizing overfitting and bias. This approach involves splitting the dataset into ten subsets, iteratively training the model on nine subsets, and validating it on the remaining subset, thereby providing a robust framework for model evaluation. By leveraging these methods, this study aims to identify the most effective model for predicting the CS of RAC.

A regression decision tree (DT) predicts continuous values by recursively splitting data to minimize variance. It creates simple, interpretable rules but can overfit without pruning. Despite its efficiency, it often lacks generalization, making ensemble methods like Random Forest and Gradient Boosting preferable for improved accuracy [16].

Random Forest (RF) is an ensemble method that builds multiple DTs and averages their predictions to improve accuracy and reduce overfitting. It has been widely used in structural engineering for analyzing complex responses and identifying critical factors influencing performance, making it well-suited for predicting the mechanical properties of recycled aggregate concrete [17]. It handles complex, nonlinear data well but requires hyperparameter tuning for optimal performance [18].

Adaptive Boosting (AdaBoost) enhances weak learners by adjusting sample weights, focusing on harder-to-predict data. It improves accuracy but is sensitive to noise and requires careful tuning for stability [19].

Gradient Boosting (GB) is an ensemble learning technique that builds trees sequentially, with each tree correcting the residual errors of the previous one. It effectively captures complex nonlinear relationships and delivers high predictive accuracy. However, it requires careful tuning of hyperparameters like learning rate and tree depth to balance performance and prevent overfitting [20].

Light Gradient Boosting Machine (LightGBM) is an optimized gradient boosting method that grows trees leaf-wise for better accuracy and efficiency. It excels in handling large datasets and complex patterns but requires careful tuning to prevent overfitting [21].

Extreme Gradient Boosting (XGBoost) is a high-performance GB algorithm that improves accuracy with regularization and parallel processing. It efficiently handles complex data but requires hyperparameter tuning for optimal performance [22].

4 Results and Discussion

The predictive performance of six tree-based and ensemble learning models—DT, RF, AdaBoost, GB, LightGBM, and XGBoost—was evaluated using the RAC dataset. The comparison was based on predicted versus actual CS values, residual distributions, and statistical performance metrics, including the coefficient of determination (R^2) and RMSE. The results are shown in Fig. 2.

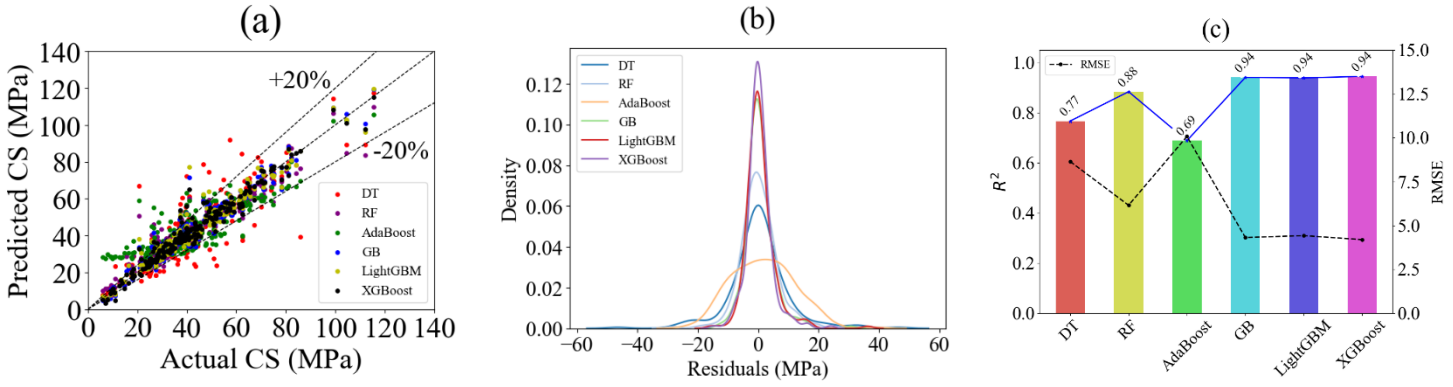

Fig. 2: a) Predicted vs. actual CS values, b) residuals distributions, and c) R^2 and RMSE plots

The scatter plot comparing actual and predicted CS values demonstrates the accuracy and distribution of predictions for each model. The XGBoost model exhibits the closest alignment with the diagonal reference line, indicating superior predictive accuracy. Most predictions for GB and LightGBM also fall within the ±20% error margins, signifying their reliability. Conversely, DT and RF models exhibit greater dispersion, particularly for higher CS values, highlighting their limitations in capturing complex relationships.

The residuals distributions plot further illustrates model performance by analyzing the distribution of prediction errors. The XGBoost and LightGBM models show the highest peak and the narrowest spread, indicating minimal prediction errors and consistent performance across data points. In contrast, the DT model exhibits a broader distribution, with a higher occurrence of large residuals, confirming its susceptibility to overfitting and limited generalization capability. RF also presents wider residual variability, suggesting its reduced effectiveness in capturing nonlinear patterns.

The bar and line chart compares the R^2 and RMSE values for each model. XGBoost, GB, and LightGBM achieved the highest R^2 values (0.94), indicating a strong correlation between predicted and actual values. These models also recorded the lowest RMSE values, demonstrating their high predictive accuracy. RF followed closely, with an R^2 of 0.88, though its RMSE was slightly higher. DT and RF performed the weakest, with R^2 values of 0.77 and 0.69, respectively, and significantly higher RMSE values, reinforcing their lower reliability in predicting CS accurately.

The results highlight the advantages of ensemble learning methods, particularly boosting-based algorithms, in predicting the CS of RAC. XGBoost consistently outperformed other models, owing to its ability to minimize overfitting while capturing complex nonlinear interactions, which aligns with previous findings on the effectiveness of ensemble methods in structural modeling contexts [23]. LightGBM and GB also demonstrated robust performance, benefiting from optimized tree structures and efficient handling of feature importance. RF, while providing reliable predictions, showed slightly higher variance compared to boosting methods. Its strength lies in evaluating feature importance and identifying key predictors, which has been effectively demonstrated in various sensitivity analyses and structural modeling contexts [24].

In contrast, DT exhibited high variability in predictions and larger residuals, reflecting its sensitivity to data partitioning and lack of generalization. RF, despite being an ensemble method, struggled with larger errors, likely due to its reliance on weak learners and sensitivity to noise in the dataset.

5 Conclusion

Overall, this study underscores the importance of selecting appropriate ML models for predicting the mechanical properties of RAC. Boosting-based models, particularly XGBoost, demonstrated superior accuracy and reliability in capturing the complex dependencies within the dataset. However, despite their advantages, these models still require careful hyperparameter tuning and validation to ensure their generalizability across different datasets. Further investigations into hybrid modeling approaches and deep learning techniques could enhance prediction accuracy and broaden applicability in real-world construction scenarios. Additionally, future studies should also consider integrating domain-specific knowledge,

such as material microstructure analysis, to improve model interpretability and decision-making. The findings from this research contribute to advancing sustainable construction practices by enabling more accurate strength predictions and optimizing the use of recycled materials in concrete production.

References

[1] B. Wang, L. Yan, Q. Fu, and B. Kasal, "A Comprehensive Review on Recycled Aggregate and Recycled Aggregate Concrete," *Resources, Conservation and Recycling*, vol. 171, p. 105565, Aug. 2021, doi: 10.1016/j.resconrec.2021.105565.

[2] M. Etxeberria, A. R. Marí, and E. Vázquez, "Recycled aggregate concrete as structural material," *Mater Struct*, vol. 40, no. 5, pp. 529–541, Feb. 2007, doi: 10.1617/s11527-006-9161-5.

[3] J. Xiao, "Erratum to: Recycled Aggregate Concrete Structures," in *Recycled Aggregate Concrete Structures*, in Springer Tracts in Civil Engineering. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2018, pp. E1–E1. doi: 10.1007/978-3-662-53987-3_16.

[4] F. Soleimani, M. McKay, C. S. W. Yang, K. E. Kurtis, R. DesRoches, and L. F. Kahn, "Cyclic Testing and Assessment of Columns Containing Recycled Concrete Debris," *ACI Structural Journal*, vol. 113, no. 5, Sep. 2016, doi: 10.14359/51689024.

[5] A. Behnood, J. Olek, and M. A. Glinicki, "Predicting modulus elasticity of recycled aggregate concrete using M5′ model tree algorithm," *Construction and Building Materials*, vol. 94, pp. 137–147, Sep. 2015, doi: 10.1016/j.conbuildmat.2015.06.055.

[6] E. M. Golafshani and A. Behnood, "Application of soft computing methods for predicting the elastic modulus of recycled aggregate concrete," *Journal of Cleaner Production*, vol. 176, pp. 1163–1176, Mar. 2018, doi: 10.1016/j.jclepro.2017.11.186.

[7] Z. H. Duan, S. C. Kou, and C. S. Poon, "Using artificial neural networks for predicting the elastic modulus of recycled aggregate concrete," *Construction and Building Materials*, vol. 44, pp. 524–532, Jul. 2013, doi: 10.1016/j.conbuildmat.2013.02.064.

[8] M. R. Kaloop, A. R. Gabr, S. M. El-Badawy, A. Arisha, S. Shwally, and J. W. Hu, "Predicting resilient modulus of recycled concrete and clay masonry blends for pavement applications using soft computing techniques," *Front. Struct. Civ. Eng.*, vol. 13, no. 6, pp. 1379–1392, Dec. 2019, doi: 10.1007/s11709-019-0562-2.

[9] A. Ababneh, M. Alhassan, and M. Abu-Haifa, "Predicting the contribution of recycled aggregate concrete to the shear capacity of beams without transverse reinforcement using artificial neural networks," *Case Studies in Construction Materials*, vol. 13, p. e00414, Dec. 2020, doi: 10.1016/j.cscm.2020.e00414.

[10] J. Xu, X. Zhao, Y. Yu, T. Xie, G. Yang, and J. Xue, "Parametric sensitivity analysis and modelling of mechanical properties of normal- and high-strength recycled aggregate concrete using grey theory, multiple nonlinear regression and artificial neural networks," *Construction and Building Materials*, vol. 211, pp. 479–491, Jun. 2019, doi: 10.1016/j.conbuildmat.2019.03.234.

[11] Z. H. Duan, S. C. Kou, and C. S. Poon, "Prediction of compressive strength of recycled aggregate concrete using artificial neural networks," *Construction and Building Materials*, vol. 40, pp. 1200–1206, Mar. 2013, doi: 10.1016/j.conbuildmat.2012.04.063.

[12] F. Khademi, S. M. Jamal, N. Deshpande, and S. Londhe, "Predicting strength of recycled aggregate concrete using Artificial Neural Network, Adaptive Neuro-Fuzzy Inference System and Multiple Linear Regression," *International Journal of Sustainable Built Environment*, vol. 5, no. 2, pp. 355–369, Dec. 2016, doi: 10.1016/j.ijsbe.2016.09.003.

[13] F. Deng, Y. He, S. Zhou, Y. Yu, H. Cheng, and X. Wu, "Compressive strength prediction of recycled concrete based on deep learning," *Construction and Building Materials*, vol. 175, pp. 562–569, Jun. 2018, doi: 10.1016/j.conbuildmat.2018.04.169.

[14] N.-D. Hoang, "A novel ant colony-optimized extreme gradient boosting machine for estimating compressive strength of recycled aggregate concrete," *Multiscale and Multidiscip. Model. Exp. and Des.*, vol. 7, no. 1, pp. 375–394, Mar. 2024, doi: 10.1007/s41939-023-00220-6.

[15] M. Hosseinzadeh, M. Dehestani, and A. Hosseinzadeh, "Prediction of mechanical properties of recycled aggregate fly ash concrete employing machine learning algorithms," *Journal of Building Engineering*, vol. 76, p. 107006, Oct. 2023, doi: 10.1016/j.jobe.2023.107006.
[16] S. Pathak, I. Mishra, and A. Swetapadma, "An Assessment of Decision Tree based Classification and Regression Algorithms," in *2018 3rd International Conference on Inventive Computation Technologies (ICICT)*, Coimbatore, India: IEEE, Nov. 2018, pp. 92–95. doi: 10.1109/ICICT43934.2018.9034296.
[17] F. Soleimani and D. Hajializadeh, "Analytical seismic resilience identifiers of bridges," in *Bridge Maintenance, Safety, Management, Digitalization and Sustainability*, 1st ed., London: CRC Press, 2024, pp. 1583–1591. doi: 10.1201/9781003483755-185.
[18] J. Zhang, G. Ma, Y. Huang, J. Sun, F. Aslani, and B. Nener, "Modelling uniaxial compressive strength of lightweight self-compacting concrete using random forest regression," *Construction and Building Materials*, vol. 210, pp. 713–719, Jun. 2019, doi: 10.1016/j.conbuildmat.2019.03.189.
[19] D.-C. Feng, Z.-T. Liu, X.-D. Wang, Y. Chen, J.-Q. Chang, D.-F. Wei, and Z.-M, Jiang, "Machine learning-based compressive strength prediction for concrete: An adaptive boosting approach," *Construction and Building Materials*, vol. 230, p. 117000, Jan. 2020, doi: 10.1016/j.conbuildmat.2019.117000.
[20] Z. M. Alhakeem, Y. M. Jebur, S. N. Henedy, H. Imran, L. F. A. Bernardo, and H. M. Hussein, "Prediction of Ecofriendly Concrete Compressive Strength Using Gradient Boosting Regression Tree Combined with GridSearchCV Hyperparameter-Optimization Techniques," *Materials*, vol. 15, no. 21, p. 7432, Oct. 2022, doi: 10.3390/ma15217432.
[21] C. Daniel, "A robust LightGBM model for concrete tensile strength forecast to aid in resilience-based structure strategies," *Heliyon*, vol. 10, no. 20, p. e39679, Oct. 2024, doi: 10.1016/j.heliyon.2024.e39679.
[22] T. Nguyen-Sy, J. Wakim, Q.-D. To, M.-N. Vu, T.-D. Nguyen, and T.-T. Nguyen, "Predicting the compressive strength of concrete from its compositions and age using the extreme gradient boosting method," *Construction and Building Materials*, vol. 260, p. 119757, Nov. 2020, doi: 10.1016/j.conbuildmat.2020.119757.
[23] F. Soleimani and D. Hajializadeh, "Bridge seismic hazard resilience assessment with ensemble machine learning," *Structures*, vol. 38, pp. 719–732, Apr. 2022, doi: 10.1016/j.istruc.2022.02.013.
[24] F. Soleimani, "Analytical seismic performance and sensitivity evaluation of bridges based on random decision forest framework," *Structures*, vol. 32, pp. 329–341, Aug. 2021, doi: 10.1016/j.istruc.2021.02.049.