[1] Thuy Trang Ta
[2] Thanh Loc Vo
[3] Ut Em Nguyen

Assessing Machine Learning Algorithms for Predicting Compressive Strength of Normal and High-Early Strength Concrete: A case Study in Binh Thuan, Viet Nam

Abstract: - This study investigates the effectiveness of machine learning models in predicting the compressive strength of both normal and high-early strength concrete. The research, conducted as a case study in Binh Thuan, Vietnam, aims to address the challenges faced by engineers in optimizing concrete mix designs. The Support Vector Machine (SVM) model, in conjunction with regression analysis employing multilinear approaches, yields predictions that are comparatively less accurate than those obtained using deep learning methods such as Artificial Neural Networks (ANN) and Light Gradient Boosting Machine (LightGBM). Particularly, the ANN model exhibits superior predictive performance, boasting an impressive R-squared value of 0.988 and the lowest model error, measured by a Root Mean Square Error of 1.493. Moreover, these deep learning techniques prove adept at capturing the intricate relationship between the water-cementitious material ratio and concrete strength, thereby enhancing the effectiveness of quality control measures at the batching plant. Consequently, engineers are empowered to make precise adjustments to concrete mix proportions during the design phase, leading to a substantial improvement in prediction accuracy and ultimately ensuring the desired performance characteristics of the concrete.

Keywords: Artificial Neural Networks, Compressive strength, Light Gradient Boosting Machine, Support Vector Machine

I. INTRODUCTION

Concrete is an essential material in the construction industry and plays a critical role in ensuring the structural integrity and longevity of building projects, primarily through its compressive strength. However, predicting concrete's compressive strength presents a formidable challenge due to its heterogeneous composition and the variability of constituent materials. Traditionally, researchers have relied on numerous experimental equations for this purpose. Nevertheless, these equations often face limitations due to input conditions, requiring the determination of experimental constants and the formulation of multiple equations to capture the complex relationship between mixture proportions and compressive strength.

Consequently, there is a notable demand for advancing prediction methodologies. In recent years, an increasing number of researchers have focused on leveraging machine learning techniques to predict concrete compressive strength. Exemplary studies include the work of Johnson and Brown (2020) [1], who utilized Artificial Neural Networks (ANN) models to achieve significant predictive accuracy in estimating concrete strength. Additionally, research by Chen and Wang (2018) [2] has provided valuable insights by successfully applying Light Gradient Boosting Machine (LightGBM) to predict compressive strength, thereby opening up new avenues for exploration in this field. In a more recent study, Hai-Van Thi Mai, May Huu Nguyen, and Hai-Bang Ly (2023) [3] delved into specific applications of LightGBM in predicting the compressive strength of fiber-reinforced self-compacting concrete.

The Support Vector Machine (SVM) is known for its good handling of non-linear data, while linear regression is a simple and easily understandable method [4],[5]. On the other hand, the ANN model, constructed based on the structure of the biological nervous system, has the ability to learn from data and simulate nonlinear relationships. The flexibility of ANN makes it a useful tool in predicting concrete compressive strength. LightGBM, an extremely efficient decision tree algorithm, stands out for its speed and ability to handle large data [6]. LightGBM provides model optimization capabilities, making it an attractive choice for predicting concrete compressive strength. ANN and LightGBM, with high computational power, are often used to model complex relationships between independent and dependent variables. Each prediction method possesses its own set of strengths and limitations. SVM, known for its robust performance in handling noisy data, may lack the flexibility inherent in ANN for

[1] Saigon Technology University (STU), Ho Chi Minh City, Vietnam. trang.tathuy@stu.edu.vn

[2] Saigon Technology University (STU), Ho Chi Minh City, Vietnam. dh82006542@student.stu.edu.vn

[3] Saigon Technology University (STU), Ho Chi Minh City, Vietnam. dh82004406@student.stu.edu.vn

capturing intricate relationships within the data. ANN and LightGBM often require large amounts of data, and LightGBM stands out for its fast computational speed. Previous studies in this field have focused on flexible combinations of concrete compressive strength prediction methods, with the goal of improving accuracy and reliability.

This study aims to evaluate the effectiveness of advanced machine learning models, including ANN, LightGBM, and SVM combined with linear regression. The objective is to enhance the accuracy and flexibility of predicting concrete compressive strength. The research focuses on a concrete batching plant in Binh Thuan, Vietnam, where two primary concrete types, namely normal concrete (referred to as R28) and high-early strength concrete (referred to as R7), are predominantly used. These concrete types exhibit compressive strengths ranging from 10 MPa to 60 MPa at 28 days. The selection between R28 and R7 concrete is typically based on the specific requirements of construction projects. R28, with its standard curing period and gradual strength progression, is consistently preferred for various structural applications. On the other hand, R7 concrete can achieve more than 70% of its 28-day strength within just 3 days, offering advantages in scenarios requiring rapid construction, early formwork removal, or enhanced structural performance [7]. This study presents a practical case to assess the performance of machine learning models in real production environments. Its objective is to advance the prediction accuracy of concrete compressive strength for both normal and high-early strength concrete, with the potential to provide practical insights into fresh concrete manufacturing technology.

II. DATA COLLECTION

The primary data for this study were collected from the concrete batching plant in Binh Thuan province, Viet Nam. Critical variables were identified, including mix proportions, slumps and curing ages. Data on the compressive strength of both normal and high-early strength concrete were systematically collected over a specified period. Blended Portland cement, compliant with ASTM C1157 specifications (Type GU), was utilized. Alongside cement, fly ash Type C, conforming to ASTM C618, was incorporated into the cementitious materials. A premium mid-range water-reducing additive, with a weight ranging approximately from 1.06 kg to 1.12 kg per liter, was also employed. This admixture is chloride-free and manufactured in accordance with the chemical admixture standards applicable to concrete, specifically ASTM C494 Type F and G. Two types of aggregate used in concrete mixtures with Dmax value as 10mm and 25 mm. Sand from local supplier in Bac Binh, Binh Thuan, Viet Nam, with a fineness modulus greater than 2.3, was utilized. The compressive strength of cubic samples with dimensions of 15cm x 15cm x 15cm conforms to the characteristic compressive strength requirements specified in Eurocode 2, denoted as $f_{ck,cu}$ in units of MPa.

The dataset comprises a data table with dimensions of 1067 rows and 15 columns, which is utilized for implementing the machine learning model. Key variables considered for analysis included:

- Concrete mix proportions (cement,flyash, fine and coarse aggregates, water, and admixtures).
- Main ratio of fresh concrete mixture: water-cement ratio, water-cementitious material ratio (cement and flyash), sand-to-aggregate proportion.
- Concrete type (Normal concrete or high-early strength concrete).
- Different slumps (10 cm, 12 cm, 14cm, 16cm, 18cm).
- Compressive strength at various curing ages (3 days, 7 days, 14 days, 28 days).

Table I presents statistical information for all attributes of the dataset. The relationship between these components and the compressive strength of concrete is highly nonlinear. Consequently, deriving the compressive strength of concrete from these experimental datasets poses a significant challenge.

Table I. Statical information for the dataset

Parameter	Unit	Mean	Min	Max
Water	kg/m^3	186.71	181.31	194.40
Cement	kg/m^3	345.63	176.85	479.35
Flyash	kg/m^3	73.73	0.00	106.52
Sand	kg/m^3	756.40	689.16	823.95
Aggregate 1	kg/m^3	316.11	285.48	364.73
Aggregate 2	kg/m^3	737.25	660.76	853.06

Parameter	Unit	Mean	Min	Max
Admixture	liter/m^3	4.41	2.14	6.28
Sand-Aggregate ratio	%	42.45	41.24	43.40
Water-cementious materials ratio		0.47	0.32	0.80
Water-cement ratio		0.57	0.38	1.08
Flyash-cementious material ratio		0.18	0	0.25
Age of testing	days	14.10	3	28
Slump	mm	138.29	100	180
Type (denoted 1 for R28, 0 for R7)			0	1
Concrete compressive strength	MPa	32.68	3.82	64.89

III. MODEL DEVELOPMENT

A. ANN model

ANN simulates the function of the biological neuron by imitating the working principles of the human brain. ANN is based on a set of connected units called artificial neurons. Each neuron transmits a signal to another neuron by a connection or synapse. Each connection is assigned a weight, which can modify the strength of the signal sent downstream.

In this study, the neural network training employed the Keras sequential model with the Adam optimizer [8]. The Keras sequential model with the Adam optimizer is widely used in deep learning research and applications due to its simplicity and effectiveness [9]. The Keras sequential model is a straightforward approach to building neural networks, allowing layers to be added sequentially, making it easy to understand and implement various architectures. The Adam optimizer, introduced by Kingma and Ba in 2014 [10], is a popular optimization algorithm in deep learning. It adapts the learning rate for each parameter individually. This adaptiveness helps the learning process converge faster and more reliably.

The model's architecture was derived from insights gained from a real concrete mix proportioning database, where the input layer consisted of 14 neurons representing dataset features, while a single neuron in the output layer predicted the compressive strength of concrete. The neural network comprised two hidden layers, each containing 64 neurons, constructed using the “keras.Sequential” method. This method facilitated the sequential addition of layers without cross-connections. Known as "Dense" layers, each ensured that every neuron was connected to all neurons in both the previous and subsequent layers. Additionally, the activation function for each neuron was the Rectified Linear Unit (ReLU), enabling non-linearity within the network [11].

To prepare the model for training, the model compilation step was employed. During compilation, specifications such as the Adam optimizer and the mean squared error loss function were set to guide the training process, crucial for optimizing the model's performance. Subsequently, the model underwent training using the model fitting method, adjusting its weights iteratively over 200 epochs. Additionally, a batch size of 32 was defined, determining the number of samples used in each training iteration.

Throughout the process, various Python libraries were instrumental for data manipulation and analysis [12]. These included "pandas" and "numpy" for data processing, "train_test_split" from "sklearn.model_selection" for data splitting, and "StandardScaler" from "sklearn.preprocessing" for standardizing numerical features. Finally, the model's performance was assessed using metrics such as root mean squared error (RMSE) and R-squared, obtained from functions within the "sklearn.metrics" library [13].

The performance of the ANN model was evaluated using a test-size of 0.4 (60% training; 40% testing). The results revealed a high level of predictive accuracy, with an R-squared value of 0.988 and a RMSE of 1.493. These metrics indicate that the model effectively captured the underlying patterns in the data and provided precise predictions of concrete compressive strength.

To ensure the generalization performance of the model, cross-validation was conducted using TensorFlow and scikit-learn. The mean RMSE across multiple folds was found to be 1.4742, indicating consistent performance across different subsets of the data. Additionally, the R-squared scores obtained from cross-validation were [0.9899, 0.9882, 0.9859, 0.9886, 0.9874], with a mean R-squared value of 0.9880. These results further validate the

robustness and reliability of the ANN model, demonstrating its ability to generalize well to unseen data and maintain high predictive accuracy across different data distributions.

Fig.1 shows the the learning process of the ANN model visualized through two key metrics: RMSE and R-squared value. The learning curve of RMSE depicts the evolution of prediction errors as the ANN model iteratively learns from the training data. A decreasing trend in RMSE suggests that the model is progressively improving its predictive accuracy over epochs. Moreover, the learning curve of R-squared provides valuable insights into the overall goodness of fit of the model. A rising trend in R-square indicates that the model is capturing more variance in the target variable as training progresses, demonstrating its capability to explain the observed variability in the data.

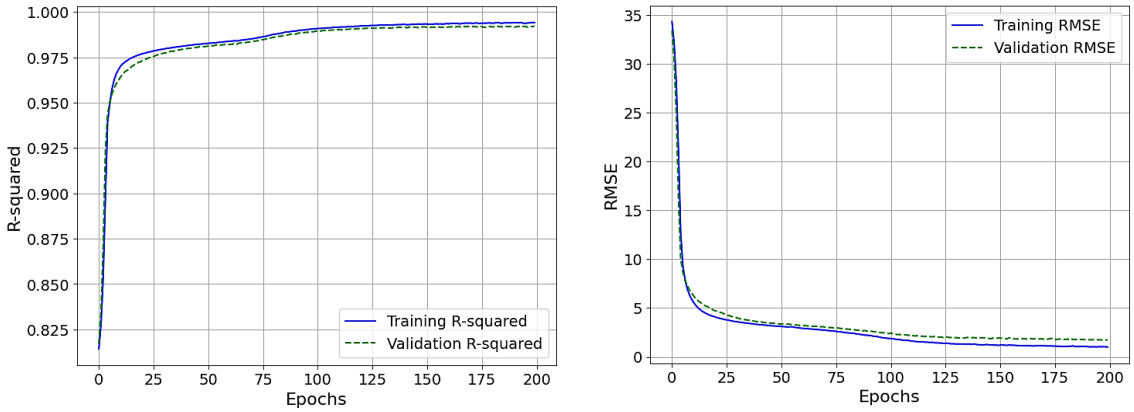


Fig. 1- Learning curve of ANN model

B. LightGBM Model

The LightGBM method was employed in this study for regression tasks. Utilizing libraries such as pandas, numpy, sklearn's train_test_split, LightGBM, StandardScaler, and matplotlib.pyplot, the data was preprocessed and prepared for model training and evaluation. StandardScaler was applied for numerical feature scaling to ensure standardized data distribution, facilitating optimal performance of the LightGBM algorithm. LightGBM datasets were constructed for training and testing, tailored for efficient model training and evaluation within the LightGBM framework.

The model parameters were meticulously specified to customize the behavior of the LightGBM model, ensuring optimal performance. Parameters such as 'num_leaves', 'learning_rate', and 'feature_fraction' were fine-tuned to control both the model's complexity and its generalization ability. Setting 'num_leaves' to 64 balanced the capture of intricate data patterns while mitigating overfitting risks. A learning rate of 0.05 regulated the model's learning step size, facilitating efficient convergence. Additionally, 'feature_fraction' set at 0.6 facilitated effective feature selection during each iteration, meaning the model utilizes 60% of the available features in each iteration, thus striking a balance between model complexity and generalization ability.

Following model training with 200 boosting rounds, predictions were generated on the testing dataset. The evaluation involved metrics like mean squared error and R-squared to assess accuracy and model fit. The observed RMSE of 1.733 and R-squared of 0.983 indicate the model's robust performance. Further assessment of cross-validated R-squared scores, ranging from 0.924 to 0.949 for testing data, with an average of 0.939, underscores the model's consistent performance across different data subsets. Similarly, the average testing RMSE of 1.407 corroborates the model's ability to generalize well to unseen data.

Fig 2 illustrates the learning curve, showcasing the robust performance of LightGBM across various training dataset sizes. It demonstrates swift learning with smaller datasets, achieving high accuracy early in the training process. With an increase in dataset size, LightGBM consistently enhances its performance, highlighting its capability to effectively manage large-scale datasets and discern complex patterns. Eventually, the curve stabilizes, implying that further increments in training data might yield diminishing returns in performance improvements.

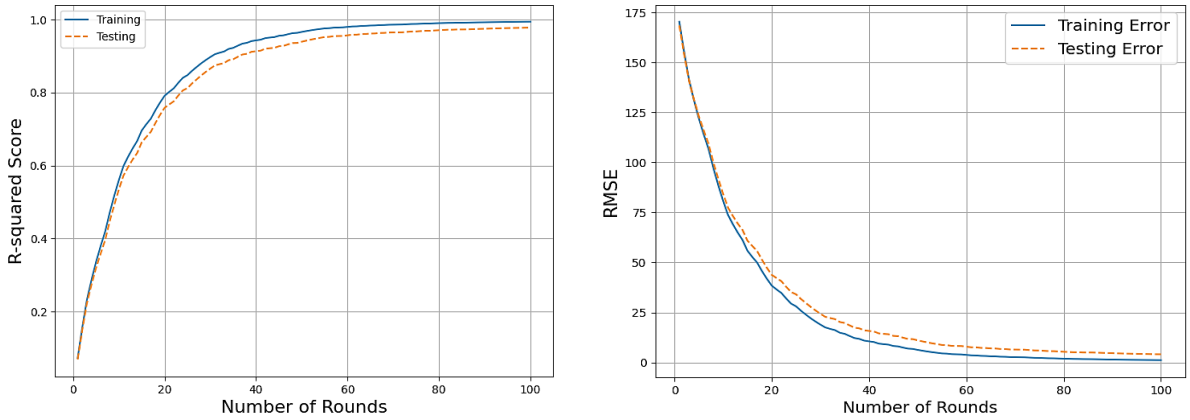


Fig. 2- Learning curve of LightGBM

C. SVM combined with linear regression

When employing a simple machine learning algorithm such as linear regression to construct a numerical model using the entire dataset as described above, the predictive results exhibit low accuracy (R-squared of 0.65 and RMSE of 8.37). This is due to the inability of linear regression to effectively learn from the entirety of the data, unlike deep learning methods such as ANN or LightGBM. Therefore, with the dataset primarily consisting of two concrete groups: high-early strength concrete (R7) and normal strength concrete (R28), this study adopts a method of preliminary classification by using SVM, followed by linear regression for each classified group when employing simple machine learning techniques.

SVM methodology was utilized to address the classification task of R7 and R28 based on their compressive strength. The dataset was preprocessed and split into training and testing sets with a test size of 0.4. After scaling the features using StandardScaler to ensure uniformity in data distribution, an SVM classifier with a radial basis function (RBF) kernel, regularization parameter C set to 2, and a gamma value of 0.1 was constructed. The model exhibited outstanding accuracy on the testing set, achieving a perfect accuracy score of 1. To assess the model's generalization performance, a learning curve was plotted using a five-fold cross-validation strategy. The cross-validation results demonstrated consistent and robust performance across different subsets of the data, with minimum and maximum accuracy scores ranging from 0.984375 to 1.0 across the folds. The consistently high accuracy scores indicate the model's ability to effectively generalize to unseen data and maintain high predictive accuracy across various data distributions.

Fig. 3 illustrates the resulting PCA biplot, visually representing the data points in a two-dimensional space defined by the principal components. Additionally, the decision boundary of the SVM model is superimposed on the biplot, showcasing how the model delineates the classes in the reduced feature space. The biplot clearly demonstrates a discernible separation between the “normal concrete” and “high-early strength concrete” classes, with a well-defined decision boundary. This indicates that the SVM model effectively distinguishes between the two types of concrete, showcasing its robustness in classifying concrete samples.

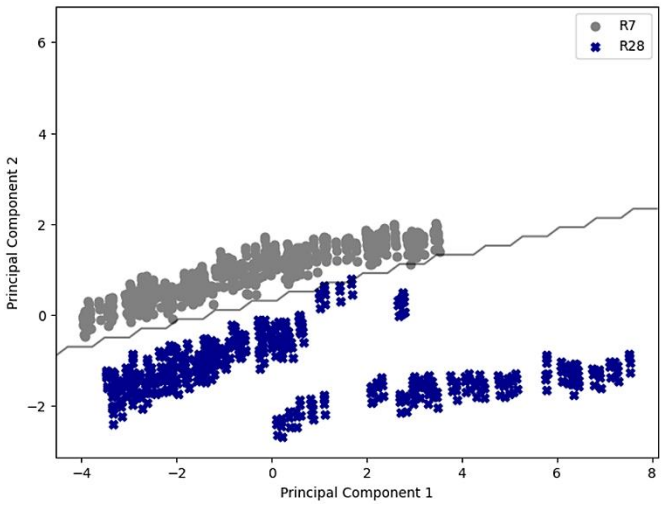


Fig 3. PCA Biplot with Decision Boundary

After classifying the concrete samples into the two groups, R7 and R28, linear regression was employed individually for each classified dataset. For the R7 dataset, the RMSE was found to be 3.3398, with an R-squared score of 0.9213. Meanwhile, for the R28 dataset, the RMSE was calculated as 4.1070, accompanied by an R-squared score of 0.9064. Upon comparing the predicted compressive strength values for both R7 and R28 with the actual values, the overall R-squared score was determined to be 0.938, with an RMSE of 3.37. These results suggest that the combined SVM and linear regression approach yields relatively good predictions of concrete compressive strength; however, it falls short compared to methods such as ANN or LightGBM in terms of predictive accuracy and capturing the intricate patterns within the data.

D. Emperical Equation

In practice, at the batching plant, relying on a large number of tests conducted over a considerable period with relatively stable material sources, the relationship between the concrete strength at the age of 28 days (f_{ck}) with the actual strength of cement (f_{ce}) and the water-cementitious ratio ($\frac{W}{C+F}$) is in line with the following formula:

$$f_{ck} = \alpha_a . f_{ce} (\frac{C+F}{W} - \alpha_b) \quad (1)$$

Where α_a, α_b are experimental cooefficients. α_a =0.48; α_b = 0.33.

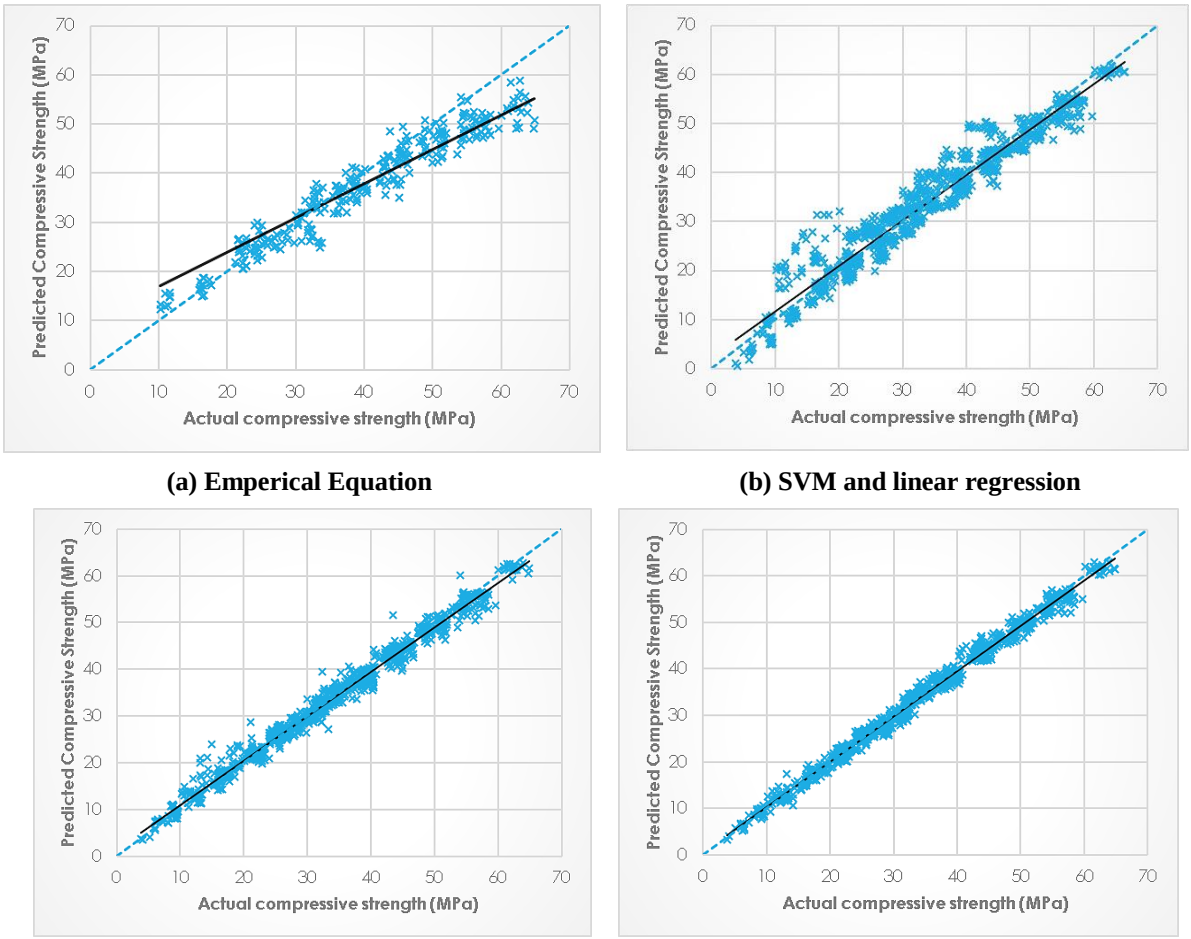
IV. PERFORMANCE EVALUATION

Performance assessment plays a pivotal role in evaluating the efficacy of diverse predictive modeling methodologies. In this investigation, we undertake a comparative analysis of four distinct techniques: Empirical equation, SVM with linear regression, LightGBM, and ANN. The findings, delineated in **Table II**, elucidate the efficacy of each approach based on their R-square and RMSE metrics. Particularly noteworthy is the superior performance exhibited by the ANN model, as evidenced by its highest R-square value and lowest RMSE score. This underscores the ANN model's superiority in predictive accuracy and overall model fit compared to the alternative methodologies.

Table II. Performance evaluation

Parameter	Empirical equation	SVM and linear regression	LigthGBM	ANN
R-square	0.938	0.938	0.983	0.988
RMSE	6.65	3.37	1.733	1.493

Additionally, **Fig.4** illustrates the comparison between actual and predicted results for each method. These visualizations provide further insights into the predictive capabilities of the models, allowing for a more comprehensive assessment of their performance.



(a) Emperical Equation **(b) SVM and linear regression**

(c) LightGBM **(d) ANN**

Fig. 4- Actual and predicted compressive strength

In addition to the aforementioned methods, the performance of predicting concrete strength can be further evaluated by exploring the relationship between the water-cement ratio (W/C) and concrete strength. The water-cement ratio is a widely used metric in concrete mix design due to its strong correlation with both the strength and durability of Portland cement concrete (PCC). Generally, lower water-cement ratios result in higher strength and increased durability of PCC. When natural pozzolans, such as fly ash, are incorporated into the mix, the ratio transforms into a water-cementitious material ratio (W/(C+F)), where the cementitious material includes both Portland cement and pozzolanic material [14].

The empirical formula commonly used in practice at the batching plant for concrete mix design, aimed at predicting concrete strength, may exhibit a significant deviation from actual results. **Fig. 5(a)** illustrates that these two curves intersect at approximately 30 MPa of concrete strength. This disparity could be attributed to the phenomenon wherein, for concrete strengths below 30 MPa, the achieved strength tends to be lower when using predictive formulas. This outcome arises from insufficient cement content, which fails to provide a cohesive environment. Consequently, fly ash remains suspended in the concrete structure, contributing to an unstable component that lacks proper bonding due to inadequate cement content. Consequently, fly ash does not contribute to strength development and may even diminish the overall strength.

For concrete strengths exceeding 30 MPa, a minimum cement content of 290 to 300 kg/m³ is necessary to establish a cohesive environment comprising calcium silicate hydrate (C-S-H) and calcium hydroxide ($Ca(OH)_2$. In this scenario, fly ash can replace approximately 20% of the cement, effectively reacting with $Ca(OH)_2$ to produce additional C-S-H, filling voids, enhancing density, and ultimately bolstering concrete strength [15].

Using the predicted concrete strength from mechine learning method, a relationship between the water-cementitious material ratio ratio and concrete strength can be depicted. While the SVM model combined with linear regression shows relatively good predictive performance on the dataset, it may not accurately capture the relationship between

the water-cementitious material ratio and the compressive strength of concrete at 28 days as shown in **Fig.5(b)**. This limitation could hinder its applicability in adjusting the component proportions in concrete mix designs.

Alternatively, both the ANN and LightGBM models showcase remarkable accuracy and faithfully capture the nearly identical correlation between the water-cementitious material ratio and concrete strength, as depicted in **Fig. 5(c)** and **Fig. 5(d)**. These models hold great promise for accurately forecasting concrete strength based on the water-cementitious material ratio, thus enabling meticulous adjustments in concrete mix designs.

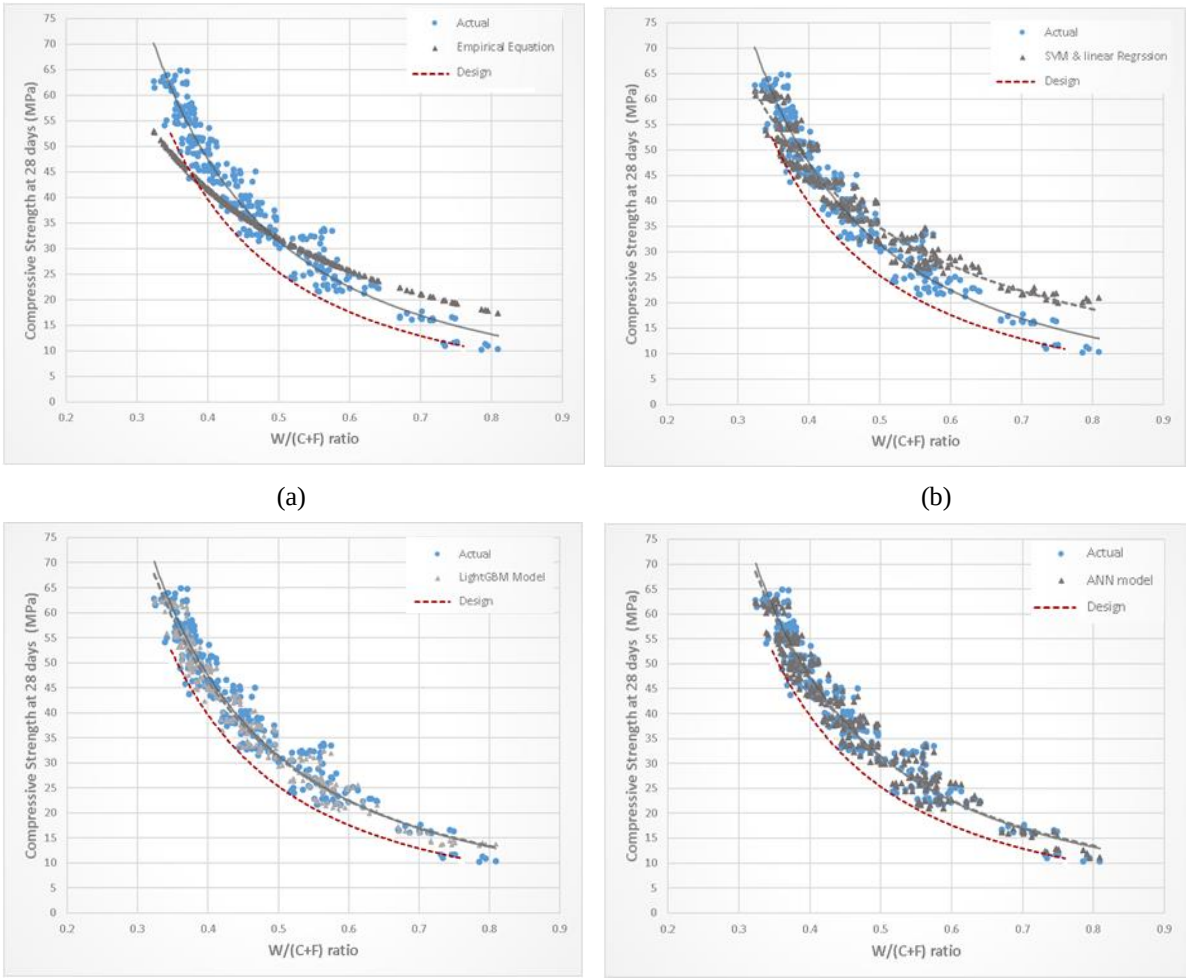


Fig. 5- The correlation between water-cementitious materials ratio and compressive strength

The significance of predicting concrete compressive strength based on the water-cementitious materials ratio lies in its utility for engineers in adjusting concrete mix designs. By accurately predicting compressive strength, engineers can make informed decisions regarding the composition of concrete mixtures. This capability enables them to optimize the mix design process, ensuring that the resulting concrete meets performance requirements. Ultimately, the ability to predict compressive strength based on the water-cement ratio empowers engineers to achieve more efficient and effective concrete designs, leading to improved construction outcomes and enhanced structural integrity.

V. CONCLUSIONS

The use of machine learning methods for predicting concrete strength offers significant benefits for quality control at batching plants. Better accuracy in predicting concrete strength can be achieved by leveraging advanced models such as ANN and LightGBM. This facilitates more precise adjustments in concrete mix designs, ultimately leading to improved overall quality and performance of concrete. The implementation of these predictive models in practice can enhance the control and consistency of concrete production processes, resulting in higher-quality concrete and improved construction outcomes.

REFERENCES

[1] J. Smith, R. Johnson, A. Brown, " Predicting compressive strength of self-repairing concrete using artificial neural networks," Construction and Building Materials, vol. 234, pp.117420, 2020.

[2] Jun-Feng Jia , Xi-Ze Chen, Yu-Lei Bai, Yu-Long Li, Zhi-Hao Wang, "An interpretable ensemble learning method to predict the compressive strength of concrete," Structures, vol.46, pp. 201-213, 2022.

[3] H.V.T.Mai, M.H.Nguyen, H.B.Ly, "Development of machine learning methods to predict the compressive strength of fiber-reinforced self-compacting concrete and sensitivity analysis," Construction and Building Materials, vol.367, pp. 130339, 2023.

[4] C. Cortes, V. Vapnik, "Support-Vector Networks," Machine Learning, vol.20(3), pp.273–297, 1995.

[5] D.C. Montgomery, E.A. Peck, G.G.Vining, "Introduction to Linear Regression Analysis," John Wiley & Sons, Hoboken, vol. 821, pp.17-20, 2012.

[6] Ke.G. Meng, Q. Finley et al., "Lightgbm: A Highly Efficient Gradient Boosting Decision Tree," Advances in Neural Information Processing Systems, 30th Annual Conference on Neural Information Processing Systems, Long Beach,pp. 3146-3154, Dec 2017.

[7] C. Lee, S. Lee, N.Nguyen, "Modeling of Compressive Strength Development of High-Early-Strength-Concrete at Different Curing Temperatures", International Journal of Concrete Structures and Materials, vol.10, pp.205–219, 2016.

[8] A. Johnson, "Optimization Algorithms for Neural Network Training," Proceedings of the International Conference on Artificial Intelligence, vol.27, no. 2, pp.112-125, 2019.

[9] D. Sumathi, K. Alluri, "Deploying Deep Learning Models for Various Real-Time Applications Using Keras," Advanced Deep Learning for Engineers and Scientists, EAI/Springer Innovations in Communication and Computing. Springer, Cham,https://doi.org/10.1007/978-3-030-66519-7_5, pp. 113-143, July 2021.

[10] D. P. Kingma, J. A. Ba, J. Adam, "A method for stochastic optimization," arXiv 2014. arXiv preprint arXiv:1412.6980, vol.1, pp.106, 2020.

[11] I. Goodfellow, Y. Bengio, A. Courville, "Deep Learning," MIT Press, 2016.

[12] A. Eidelman, "Python data science handbook by jake VANDERPLAS", Statistique et Société, vol. 8, no.2, pp. 45-47, 2020.

[13] K. Brown, "Evaluation Metrics for Neural Network Models," IEEE Transactions on Neural Networks, vol. 10, no. 4, pp.75-88, 2021.

[14] M. Sidney, "Section 5 - Calculations Relating to Concrete and Masonry," Construction Calculations Manual, Butterworth-Heinemann: Boston, pp. 211-264, 2012.

[15] S. Goñi , A. Guerrero, M. P. Luxán, A. Macías, "Activation of the fly ash pozzolanic reaction by hydrothermal conditions," Cement and Concrete Research , vol. 33, no. 9, pp. 1399-1405, 2003.