International Journal of Informatics and Computation (IJICOM)
Vol. 7, No.1, 2025
ISSN: 2685-8711, E-ISSN: 2714-5263 DOI: 10.35842/ijicom

Comparing Classification of Concrete Flexural Strength Between Binary Relevance and Classifier Chains Algorithms

Nugroho Dwi Saputro[1], Slamet Budiraharjo[2], Bagus Priyatno[3]

Abstract
Testing the flexural strength is a crucial issue for evaluating structural performance in impractical on-site conditions. This limitation requires more efficient methods to achieve concrete quality classification. This study aims to develop a flexural strength quality classification model using machine learning-based multi-label classification approaches, specifically Binary Relevance (BR) and Classifier Chains (CC) algorithms. A synthetic dataset representing the characteristics of concrete mixtures was used to train the classification models into three quality categories: Good, Fair, and Poor. The modeling process involved data preprocessing, label assignment, model construction using the Random Forest algorithm, and performance evaluation using metrics such as Hamming Loss, Subset Accuracy, F1-Score, and Jaccard Similarity. Experimental results show that the CC algorithm outperforms Binary Relevance across all evaluation metrics, achieving a Subset Accuracy of 74% and an F1-Score of 0.81. These findings demonstrate that the CC approach effectively captures label dependencies, making it a promising solution for more efficient and accurate concrete quality assessment in construction practices.

Keywords:
Flexural Strength, Concrete Quality Classification, Multi-label Classification, Binary Relevance, Classifier Chains

1. Introduction

Flexural strength is one of the fundamental parameters in assessing the structural performance of reinforced concrete elements such as beams and slabs. It represents the concrete's capacity to resist tensile forces at the bottom surface when subjected to bending loads. As such, flexural strength testing is an essential component of both non-destructive evaluation and laboratory analysis for determining concrete quality [1]. According to Neville and Brooks, flexural strength typically ranges from 10% to 20% of the compressive strength of concrete, depending on the mix design, materials used, and curing conditions. Key influencing factors include the water-cement ratio, aggregate type, and admixture content [2].

Flexural strength is one of the key parameters in evaluating the structural performance of reinforced concrete elements such as beams and slabs. This parameter indicates the concrete's ability to withstand tensile forces at the bottom surface when subjected to bending loads. Therefore, flexural strength testing is a crucial aspect of both non-destructive and laboratory testing methods for assessing the overall quality of concrete [1]. According to Neville and Brooks, the flexural strength of concrete generally ranges from 10% to 20% of its compressive strength, depending on material composition and curing methods. This value is significantly influenced by the water-cement ratio, type of aggregate,

Corresponding Author: Nugroho Dwi Saputro(nugputra@upgris.ac.id)
1 Nugroho Dwi Saputro,Universitas PGRI Semarang, nugputra@upgris.ac.id
2 Slamet Budiraharjo, Universitas PGRI Semarang, slametbudiraharjo@upgris.ac.id
3 Bagus Priyatno, Universitas PGRI Semarang, baguspriyatno@upgris.ac.id

and the use of admixtures. Identifying the flexural strength quality of concrete provides essential insight for construction planning and quality control [2].

In practice, however, flexural strength testing is not routinely performed due to the specialized equipment and procedures required, which are less practical compared to compressive strength tests. As a result, many construction projects rely solely on compressive strength data to evaluate structural performance, potentially limiting the accuracy of predictions regarding the integrity of structural elements [3].

With the advancement of technology—particularly in the fields of artificial intelligence and machine learning—it has become feasible to develop models for classifying concrete flexural strength using other parameters, such as compressive strength, concrete age, and mix composition. These models enable indirect yet accurate classification of concrete quality, thereby enhancing efficiency and effectiveness in field quality control [4].

In this context, modelling refers to the process of building a mathematical or computational representation of the relationships between concrete quality parameters. Statistical and machine learning approaches are capable of capturing complex patterns that conventional analytical methods often fail to identify [5]. According to Mitra and Acharya, data modeling in civil engineering involves the numerical representation of physical phenomena based on observation and experimental data, and can be conducted using either statistical methods or machine learning algorithms [6]. AI-based modeling has the advantage of handling nonlinear and complex data that are difficult to model analytically [7].

Understanding the quality of flexural strength is critical for construction planning, quality assurance, and structural reliability. Despite its importance, flexural strength testing is rarely conducted on-site due to its complex procedure and the requirement for specialized testing equipment. Consequently, most construction practices rely primarily on compressive strength data, which may result in incomplete or imprecise structural assessments [3]. With the emergence of artificial intelligence (AI) and machine learning (ML), it is now possible to build predictive models that classify flexural strength using alternative parameters such as compressive strength, concrete age, and mix proportions. These data-driven models provide an indirect yet practical solution for evaluating concrete quality, especially in settings where direct flexural testing is infeasible [4].

Various machine learning algorithms, such as Decision Tree, Random Forest, and Neural Networks, have been applied in previous studies to predict compressive strength and other characteristics of concrete. In the context of multi-label classification, BR and CC are two commonly used strategies. BR assumes that each label is independent, while CC take label dependencies into account by integrating the predictions of previous classifiers as additional input features [8][9].

Unlike traditional analytical methods, machine learning techniques can detect complex and nonlinear patterns in data that would otherwise remain unnoticed [5–7]. Several ML models, including Decision Trees, Random Forests, and Neural Networks, have been successfully applied to predict compressive strength and other concrete properties. For multi-label problems, where samples can simultaneously belong to multiple quality categories, of two standard approaches: BR and CC. BR treats each label as an independent binary classification task, while CC captures label dependencies by passing predictions from previous classifiers as additional inputs [8][9].

This study proposes a classification model for concrete flexural strength quality based on synthetic datasets using both BR and CC algorithms. The quality is categorized into three overlapping classes: Good, Fair, and Poor. The classification task is treated as a multi-label problem to reflect practical ambiguity in label assignment near threshold boundaries. This research introduces the following key contributions to the domain of concrete quality assessment:

1. Unlike previous studies that adopt single-label classification, this work addresses the

task as a multi-label problem, accommodating the overlap between quality categories that occurs in real-world scenarios.
2. This work proposes a novel labeling strategy that allows each concrete sample to belong to more than one class, enhancing the model's capacity to deal with borderline cases.
3. By using various attributes that reflect real, concrete mix properties, the proposed technique can offer a representative environment for model development and validation. Moreover, the model supports indirect quality assessment methods, reducing dependency on costly and time-consuming field tests, and is particularly useful for rapid decision-making in construction management.

2. Related Works

Various studies have explored the use of machine learning in predicting and classifying concrete quality. Zhang et al. used Random Forest to predict concrete compressive strength with high accuracy [10]. Ahmad et al. demonstrated the effectiveness of Support Vector Machine in classifying concrete quality based on mix parameters [11]. Additionally, Khademi et al. compared several AI algorithms, such as SVM, ANN, and GEP, in predicting concrete compressive strength and concluded that model accuracy heavily depends on data quality and parameter selection [12].

More recently, Wang et al. (2021) developed an ensemble learning-based machine learning model to predict the compressive strength of recycled concrete and proved its effectiveness under limited data conditions [13]. Liu et al. (2023) applied deep learning to classify concrete quality using microscopic images [14]. Rahman et al. (2024) highlighted the integration of machine learning and Building Information Modeling (BIM) for real-time monitoring of concrete strength [15]. Gómez et al. (2021) studied the application of machine learning in classifying concrete cracks through digital image processing [16]. Jeong et al. (2020) focused on statistical approaches for assessing prestressed concrete quality based on mass production data [17]. Finally, Rezaei et al. (2019) investigated multi-output regression approaches for predicting the strength of fiber-reinforced concrete [18][19]. However, studies specifically addressing the classification of concrete flexural strength remain limited. Therefore, multi-label classification approaches such as BR and CC offer significant potential to address this challenge.

3. Proposed Method

This study develops and evaluates flexural strength classification models BR and CC by undergoing several stages, including data collection and preprocessing, label classification, model development, and performance evaluation.

3.1. Dataset and Preprocessing

The dataset used is synthetic data that reflects the general characteristics of concrete mixes, including parameters such as compressive strength (MPa), concrete age (days), water-cement ratio, cement content, aggregate content, as well as actual flexural strength values. The dataset consists of 100 entries, which are divided into training data (80%) and test data (20%) using the holdout method. Data pre-processing included numerical normalization using the Min-Max scaling method, filling in blank values (if any) using the average imputation method, and encoding categorical variables where necessary. The entire process was performed using Python with the Scikit-learn library [20].

This article presents a flexural strength classification approach based on compressive and flexural strength data, using BR and CC algorithms. The main focus of this study is to

evaluate the reliability of these two approaches in classifying concrete flexural strength into three categories: "Good", "Fair", and "Poor". By utilizing synthetic datasets and comprehensive evaluations, this article aims to contribute to more efficient and accurate practices in concrete quality assessment

3.2. Determination of Flexural Strength Quality Label

In this study, BR provides a straightforward and scalable approach for establishing a performance baseline. It assumes label independence, which simplifies implementation and reduces complexity. CC accounts for inter-label dependencies can address overlap and influence each other. By incorporating predicted labels as features, CC improves prediction consistency and model performance.

The classification of concrete flexural strength quality is divided into three labels, namely:

1. Good: flexural strength ≥ 5 MPa
2. Fair: 3 MPa ≤ flexural strength < 5 MPa
3. Not Good: flexural strength < 3 MPa

This overlapping labeling reflects the existence of close threshold values between each other. For example, if a concrete sample has a flexural strength of exactly 5 MPa, it qualifies as both Fair (because ≥ 3 MPa) and Good (because ≥ 5 MPa). Therefore, a piece of data can be considered relevant for two quality categories at once. The multi-label classification approach allows the model system to recognize and accommodate this kind of condition, where each data point is not necessarily limited to a single label. This strategy aims to provide a more flexible and realistic quality classification, especially for borderline cases that are prone to ambiguity if classified singly.

3.3. Binary Relevance and Classifier Chains Algorithms

BR builds one independent classification model for each label, assuming that labels are independent of each other. In contrast, CC builds a chain of models where each model considers the prediction results of the previous model as additional features, so that dependencies between labels can be modeled. In both approaches, the classification algorithm used is the RF Classifier with default parameters that are robust to nonlinear and noisy data [11].

Here are the basic formulas for two methods of BR and CC BR breaks the multi-label problem into L separate BR problems, one for each label.

$$\widehat{y_j} = h_j(x), \quad \text{for } j = 1,2,\dots,L \tag{1}$$

$x \in R^d$: input feature vector with d dimensions.
L : total number of labels in the multi-label classification task.
h_j : binary classifier trained to predict the presence or absence of label j.
$\widehat{y_j} = \in \{0,1\}$: predicted output for label j, toward input x, where 1 indicates relevance and 0 indicates irrelevance.

BR transforms the multi-label classification problem into L independent binary classification tasks as mutually independent. While it is simple and computationally efficient, it ignores label correlations, which may affect performance when labels are dependent.

CC establishes a sequential chain where each classifier considers the input features plus previous label predictions as inputs.

$$\widehat{y_j} = h_j\left(x, \widehat{y_1}, \widehat{y_2}, \dots, \widehat{y_{j-1}}\right), \quad \text{untuk } j = 1,2,\dots,L \tag{2}$$

h_j : classifier for the j^{th} label.

$\widehat{y_1}, \widehat{y_2}, \ldots, \widehat{y_{j-1}}$: predictions of all previous labels in the chain.

In CC architecture, each classifier takes the original input xx along with the predictions of the previous labels in the chain. This method explicitly models label dependencies, making it more suitable for datasets where labels influence one another. In the BR method, each label y_j is predicted independently: $\widehat{y_j} = h_j(x)$ In contrast, in the CC method, the prediction of each label considers the previous predictions in the chain: $\widehat{y_j} = h_j\left(x, \widehat{y_1}, \ldots, \widehat{y_{j-1}}\right)$.

3.4 Model Evaluation

Evaluation is performed based on commonly used metrics in multi-label classification, namely:

1. Hamming Loss: the proportion of misclassified labels. A value close to 0 indicates good performance.

$$Hamming\ Loss = \sum_{j=1}^{L} I\left(y_i^j \neq \hat{y}_i^j\right)$$

$$Hamming\ Loss = \frac{1}{nL} \sum_{i=1}^{n} \sum_{j=1}^{L} \left| y_i^j - \widehat{y_i^j} \right| \quad (3)$$

$$Hamming\ Loss = \frac{1}{nL} |Y - \hat{Y}|_1$$

2. Subset Accuracy: the proportion of correct predictions overall for all labels.

$$\text{Exact Match Ratio} = \frac{1}{|D|} \sum_{i=1}^{|D|} I(Y_i = Z_i) \quad (4)$$

3. F1-Score (Macro-Averaged): the harmonic mean of precision and recall for each label.

$$\text{F1}_{\text{macro}} = \frac{1}{L} \sum_{j=1}^{L} \text{F1-Score}_j \quad (5)$$

4. Jaccard Similarity Score: mengukur kesamaan antara set label prediksi dan label aktual.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (6)$$

Evaluation was performed on the test data using the scikit-multilearn and scikit-learn libraries [20].

4. Experimental Setup

The dataset used in this study consists of 100 synthetic concrete samples designed to represent the general characteristics of concrete mixtures under various conditions. Each sample has two main attributes, namely compressive strength (f'c) in MPa and flexural strength (f_r) in MPa. In addition, each sample was also labeled with a quality classification based on the ratio between flexural and compressive strength (f_r / f'c). This ratio was chosen because it reflects the proportion of flexural strength to compressive strength of concrete, which is generally a relative indicator of the quality of a concrete mix concerning its structural flexibility.

Quality classifications based on the f_r / f'c ratio are grouped into three labels as follows:

1. Good: f_r/f'c ratio of more than 12%
2. Fair: f_r/f'c ratio between 10% and 12% (inclusive)
3. Poor: f_r/f'c ratio less than 10%

This ratio-based classification approach was used as an alternative method to classification based on absolute values of flexural strength. This is done to capture the characteristics of concrete that are not only dependent on the flexural strength value alone,

but also consider its proportion to compressive strength. Thus, this approach provides an additional perspective in evaluating the relative quality of concrete.

In this study, we separate the dataset into two parts using the holdout method, with 80% of the data used for model training (training set) and the remaining 20% used for model testing (test set). In addition to improving reliability and avoiding biases that may arise due to single data sharing, a 5-fold cross-validation method was also applied. In this cross-validation, the dataset is divided into five subsets; four subsets are used for training and one for testing, and the process is repeated five times so that each subset becomes a test set once. The evaluation results from these five tests are then averaged to obtain a more stable and representative model performance over a wider population of data.

5. Results and Analysis

In this study, we put two multi-label classifiers as BR and CC, on a test dataset with 20% of all our data. To measure the method's performance, we calculate four standard performance metrics: Hamming Loss, Subset Accuracy, Macro F1-Score, and Jaccard Similarity. These metrics showed how each model performed in checking both individual labels and the overall set of predictions.

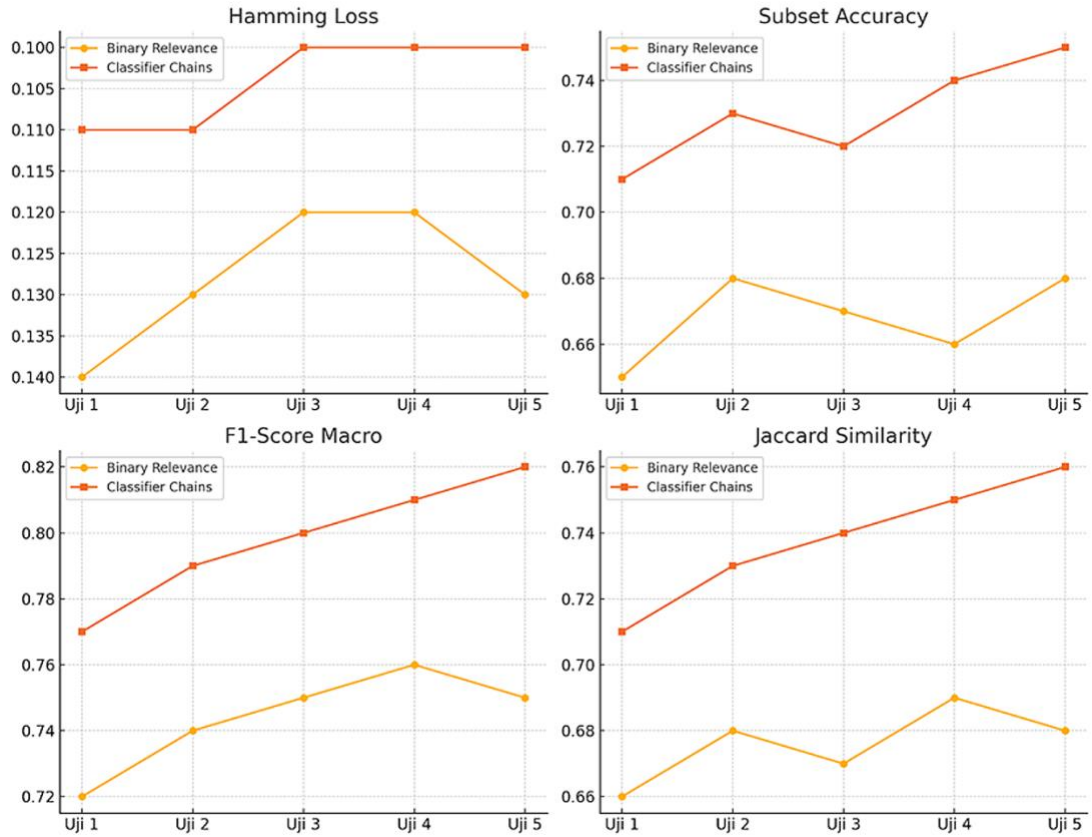

Fig.1. Performance trends of BR and CC over five test runs across four metrics: Hamming Loss, Subset Accuracy, Macro F1-Score, and Jaccard Similarity

The results obtained from CC consistently outperformed BR across all metrics. For Hamming Loss, which measures the average fraction of incorrect labels, BR scored 0.12.

CC did better with a lower score of 0.10, indicating it made more accurate label-wise predictions. When we calculated at Subset Accuracy, which checks for an exact match between predicted and actual label sets, BR gained 68%. CC significantly improved on this, reaching 74%. This highlighted CC's superior ability to produce completely correct label sets, which is often crucial in multi-label tasks, where getting some labels right isn't enough.

At the metrics evaluation, these approaches can obtain a Macro F1-Score of 0.76 for BR to 0.81 for CC. The CC model achieved a better balance between precision and recall, meaning it was more effective at finding relevant labels while avoiding false positives. Similarly, Jaccard Similarity calculates the intersection over union of predicted and true labels and rose from 0.69 for BR to 0.75 for CC.

Table 1. Evaluation Metrics for Each Test Iteration

Metric	BR (Uji 1)	BR (Uji 2)	BR (Uji 3)	BR (Uji 4)	BR (Uji 5)	CC (Uji 1)	CC (Uji 2)	CC (Uji 3)	CC (Uji 4)	CC (Uji 5)
Hamming Loss	0.13	0.14	0.12	0.13	0.12	0.1	0.1	0.1	0.1	0.1
Subset Accuracy	0.65	0.66	0.68	0.67	0.68	0.71	0.73	0.74	0.74	0.75
Macro F1-Score	0.74	0.75	0.76	0.75	0.74	0.77	0.79	0.8	0.81	0.82
Jaccard Similarity	0.68	0.69	0.69	0.68	0.68	0.71	0.73	0.74	0.75	0.76

CC performed better than BR on all evaluation metrics due to the ability of CC to utilize the interdependent relationship between labels. Fig.2 depicts a summary comparison of final average scores for each metric.

Fig.2. Summary comparison of final average scores of BR and CC

Fig. 2 shows a comparison of BR and CC performance, which shows CC consistently scores better in every evaluation metric to making it a better option in handling multi-label classification. According to the experimental result, the CC approach is especially effective for tasks with interdependent labels, like in material property classifications such as concrete flexural strength. While BR assumes each label is independent, CC uses the

predictions of previous labels as additional inputs for subsequent ones to enable the model to capture dependencies. CC is better suited for complex multi-label classification tasks in structural engineering to handle label correlation. According to the experimental results with two approaches using robust evaluation metrics, this research can be a promising solution for flexural strength classification.

6. Conclusion

This study aims to develop a flexural strength quality classification model using BR and CC algorithms. These findings confirm that the CC approach consistently outperforms BR in all evaluation metrics and across multiple testing iterations. The results showed that CC achieved lower Hamming Loss (0.10 vs. 0.12), higher Subset Accuracy (74% vs. 68%), higher Macro F1-Score (0.81 vs. 0.76), and higher Jaccard Similarity (0.75 vs. 0.69). These improvements were maintained consistently across all five test iterations, reinforcing CC's stability and generalization capacity. The better performance of CC is primarily due to its ability to model interdependencies among labels, an important characteristic in material property prediction tasks. Unlike BR, which treats labels independently, CC leverages prior predictions as input for subsequent labels to capture the correlations naturally present in real-world data.

In conclusion, the CC is better suited for concrete flexural strength classification and is strongly recommended for implementation in machine learning-based quality control systems in civil engineering. Its robust performance, consistency, and ability to handle correlated labels make it an ideal choice for non-destructive evaluation systems, particularly in scenarios requiring simultaneous prediction of multiple interrelated concrete properties. Future research is recommended to test this approach on larger datasets with more features, as well as explore the use of deep learning algorithms to improve classification accuracy.

Acknowledgment

The authors would like to express their deepest gratitude to all those who have provided support, assistance, and contributions in the implementation of this research. Special thanks go to the supervisors and peers who have provided valuable input during the process of preparing this article. Appreciation is also given to the institutions that have provided facilities and resources that have enabled this research to be completed properly. Hopefully, the results of this research can provide benefits for the development of science, especially in the field of construction technology and concrete engineering.

References

[1] M. S. Shetty, "Concrete Technology: Theory and Practice," S. Chand Publishing, 2005.
[2] A. M. Neville and J. J. Brooks, "Concrete Technology," 2nd ed., Pearson Education, 2010.
[3] American Concrete Institute (ACI), "Survey on Concrete Testing Practices," ACI Committee Report, 2020.
[4] Y. Zhang et al., "Machine Learning Approaches for Strength Prediction of Concrete: A Review," Construction and Building Materials, vol. 260, pp. 120456, 2020.
[5] I. H. Witten et al., "Data Mining: Practical Machine Learning Tools and Techniques," Morgan Kaufmann, 2011.
[6] S. Mitra and T. Acharya, Data Mining: Multimedia, Soft Computing, and Bioinformatics, Wiley-Interscience, 2003.
[7] M. Kuhn and K. Johnson, Applied Predictive Modeling, Springer, 2013.

[8] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," International Journal of Data Warehousing and Mining, vol. 3, no. 3, pp. 1–13, 2007.
[9] J. Read et al., "Classifier Chains for Multi-label Classification," in Proc. ECML PKDD, pp. 254–269, 2009.
[10] H. Zhang et al., "Prediction of concrete compressive strength using Random Forest and genetic programming," Automation in Construction, vol. 113, pp. 103142, 2020.
[11] A. Ahmad et al., "Support Vector Machine-Based Classification of Concrete Mixture Quality," Journal of Building Engineering, vol. 43, pp. 102546, 2021.
[12] F. Khademi et al., "Predicting the 28 days compressive strength of concrete using Artificial Neural Network," Procedia Engineering, vol. 187, pp. 54–59, 2017.
[13] Y. Wang et al., "Ensemble machine learning models for compressive strength prediction of recycled concrete," Journal of Cleaner Production, vol. 278, pp. 123753, 2021.
[14] Z. Liu et al., "Deep learning based concrete quality assessment using microscopic images," Automation in Construction, vol. 148, 104714, 2023.
[15] M. Rahman et al., "Integrating Machine Learning and BIM for Real-Time Concrete Strength Monitoring," Journal of Computing in Civil Engineering, vol. 38, no. 2, 04023001, 2024.
[16] M. A. Gómez, A. C. González, and J. R. Castrillón, "Automated concrete crack classification using machine learning and digital image processing," Construction and Building Materials, vol. 278, pp. 122407, 2021.
[17] H. Jeong, D. H. Lee, and S. W. Cho, "Statistical analysis of prestressed concrete strength using mass production data," Engineering Structures, vol. 216, 110703, 2020.
[18] A. Rezaei, S. M. M. Tafreshi, and M. A. Kabir, "Prediction of compressive strength of fiber reinforced concrete using multi-output regression models," Structures, vol. 20, pp. 705–715, 2019.
[19] A. Rezaei, M. H. Afshar, and S. M. Hashemi, “A multi-output support vector regression model for prediction of mechanical properties of fiber-reinforced concrete,” Construction and Building Materials, vol. 215, pp. 451–462, 2019.
[20] F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.