Integrating Machine Learning and Deep Learning for Compressive Strength Prediction

Siddique Ibrahim S P
School of Computer Science and Engineering
VIT-AP University
Amaravati, India
siddique.ibrahim@vitap.ac.in

Megha Sai Lalith V
School of Computer Science and Engineering
VIT-AP University
Amaravati, India
lalith.22bce9682@vitapstudent.ac.in

Rakesh Y
School of Computer Science and Engineering
VIT-AP University
Amaravati, India
rakesh.22bce9679@vitapstudent.ac.in

Karthik E
School of Computer Science and Engineering
VIT-AP University
Amaravati, India
karthik.22bce7997@vitapstudent.ac.in

Karthik S
School of Computer Science and Engineering
VIT-AP University
Amaravati, India
karthik.22bce9380@vitapstudent.ac.in

Dheeraj K
School of Computer Science and Engineering
VIT-AP University
Amaravati, India
dheeraj.21bce8825@vitapstudent.ac.in

***Abstract*—Prediction of compressive strength in concrete is very important to ensure the safety and durability of any structure. In this paper, we introduce a machine-learning hybrid approach whereby we combine the strengths of two different models: CatBoost, working extremely well with structured data, and a neural network to grasp the underlying complex patterns within the data. Next, these models are combined, and their prediction is fine-tuned by another layer of linear regression to increase overall accuracy. The hybrid model achieved an RMSE of 3.86 and an R^2 score of 0.94, demonstrating its superior accuracy compared to the individual models. Herein, the proposed method has been applied to the standard concrete strength data set, and the results of hybrid performance give more reliable predictions than either model alone. It is a promising technique that can find its practical applications in construction, providing insight into concrete strength and enabling engineers to make better decisions.**

***Index Terms*—Machine learning, artificial neural networks, Compressive Strength of concrete, Hybrid Model, Catboost regressor.**

I. Introduction

Concrete is the backbone of construction. It's a mix of cement, water, gravel, sand, and sometimes a few extras like fly ash or super plasticizers. These ingredients are used to make concrete both strong and durable. Compressive strength is what keeps a building or a bridge from collapsing under weight. Predicting this strength is a big deal. Testing concrete's strength the traditional way? It takes time and money. That's where machine learning steps in, as it has the ability to detect certain patterns in data that would take a human a lot of time to sift through. Though the previous studies have attempted the use of isolated ML models such as the ANNs and the tree– based algorithms for the prediction of concrete strength, these methods often have their limitations. The generalized ANNs is good at detecting non-linear patterns in data but may be less useful in dealing with structured or categorical data, where gradient boosting models like CatBoost thrive. To date, however, no single model has been able to do this accurately and in a robust way that it can be used in practice where data complexity and noise typically reduces predictive capability. In this work, we propose a new hybrid algorithm that integrates CatBoost and ANN model with a linear regression meta-layer for improved prediction accuracy. This hybrid approach takes advantage of both of the models because CatBoost is able to deal with categorical and complex features while the ANN is able to deal with complex, non-linear dependencies. The result is a model that not only improves prediction accuracy but also enhances generalization across diverse concrete mix designs. We've got data from 1,030 concrete samples. It includes cement content, water ratio, gravel, sand, and more. With fine-tuning, this combined approach achieves an RMSE of 3.86 and an R^2 score of 0.94. These metrics show the hybrid model's potential as a more effective tool for engineers and construction

II. Related Work

The prediction of the compressive strength of concrete has been within the focus of multiple authors and several machine learning techniques have been applied such as artificial neural networks (ANN), decision tree based models and ensemble methods. In this section, we review selected works that incorporate neural networks and other machine learning algorithms to predict the strength of concrete, emphasizing their success level in various situations.

A. Neural Networks for Concrete Strength Prediction

Artificial Neural Networks (ANN) have been widely applied in estimating the compressive strength of concrete. For instance, Singh and Khaskil [5] applied ANN to estimate the compressive strength of green concrete with fly-ash and super plasticizer admixtures. For model building, the study utilized a total of nine input features, including among others cement, water content and aggregate proportions. It was possible to make accurate predictions using ANN because the model was capable to bang on non-linear relationships of concrete's strength pattern.

B. Traditional Machine Learning Models

Researchers have also explored traditional machine learning models, with ensemble methods such as Gradient Boosting Machine (GBM), Random Forest regressor(RF) proving especially effective. Zhu et al. [2] used both RF and GBM to achieve R^2 scores of 0.9167 and 0.942, respectively. This high performance underscores these methods' ability to manage complex relationships within the data, making them reliable for predicting strength, even with noisy datasets.

Advanced algorithms like CatBoost, particularly with hyperparameter tuning, have also shown promise. Shi et al. [3] used a CatBoost model optimized with Random Search (RS) to predict strength, achieving better accuracy than RF and LightGBM, especially in reducing RMSE and MAPE. CatBoost's robustness in handling categorical data and tuning parameters effectively highlights its potential in concrete prediction tasks.

C. Hybrid and Feature Selection Models

Some studies combine models to enhance predictive performance further. For instance, Abuodeh et al. [1] used Sequential Feature Selection (SFS) with ANN, creating a robust hybrid model that leverages each model's strengths to improve accuracy—especially in datasets where noise and irrelevant features could otherwise skew results.

It has been observed in previous research papers that compression strength of concrete can be estimated fairly accurately using machine aided systems such as Artificial Neural Networks, Random Forest, Gradient Boosting Machine or even CatBoost. But all these techniques have certain constraints. For example; ANN models, as used by Singh & Khaskil in 2020, are capable of learning from complex and intricate non-linear relationships among the variables, but these models tend to be overfitted and ignorant of the noise in data. Random Forest and Gradient Boosting Machines, as described by Zhu et al. in 2023, perform reasonably well for tabular data but require a large amount of training and have troubles with categorical variables. CatBoost approaches, as described by Shi et al. in 2024, can handle tabular and categorical data quite effectively and output good results. But by their nature, they are quite poor at deciphering continuous target variables due to their inability to learn non-linear relationships sufficiently. In the same vein, Abuodeh et al. in the year 2019, implemented SFS to improve the results of ANN network, but even so this type of model cannot fully grasp other subtleties in structured data. Our hybrid approach addresses these gaps by combining CatBoost with ANN, leveraging CatBoost's structured data handling and ANN's ability to capture non-linear patterns.

This is further enhanced by a linear regression meta-layer, which optimally balances the strengths of both models to achieve higher predictive accuracy and robustness. Unlike single-model approaches, our hybrid model provides superior generalization across diverse data types, as demonstrated by an RMSE of 3.86 and an R^2 of 0.94, which surpasses the performance of the cited standalone models.

III. METHODOLOGY

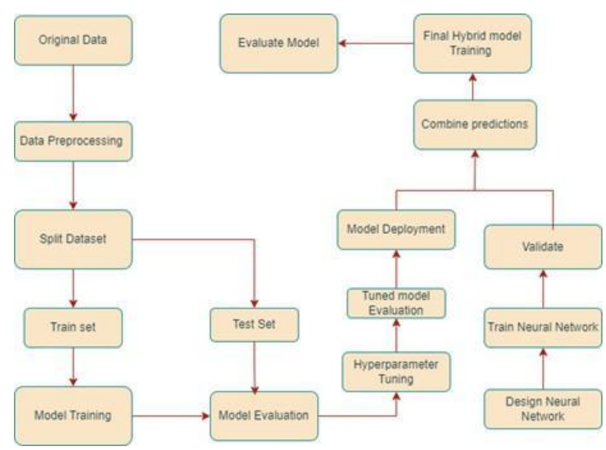

Fig. 1. Flow Diagram

A. Dataset

The dataset contains 1030 records of concrete mixtures with various components such as

- Cement (kg/m^3): A fundamental and binding component which sets after contact with water.
- Blast Furnace Slag (kg/m^3): A waste from steel production which enhances strength and lowers water absorption.
- Fly Ash (kg/m^3): A residue produced from coal burning that is added to concrete to improve its workability and durability.
- Water (kg/m^3): Used for mixing and wetting, activating cement to set and harden.
- Super plasticizer (kg/m^3): A component which increases the workability of concrete while not proportionately increasing water.
- Coarse Aggregate (kg/m^3): Bigger stones or gravel which support the mass and impart strength to the concrete.

- Fine Aggregate (kg/m³): sand, which is used to fill spaces between coarse aggregates to make the workability and density of concrete better.
- Age (days): The duration since preparations were mixed; Concrete continues to strengthen with age as it undergoes curing.
- Concrete Compressive Strength (MPa)[Target variable]: The strength of the concrete above which the deformation takes place as loads acting on the structure is supported. This indicates the quality and durability of the concrete.

B. Data Preprocessing

- Missing Values: No missing values could be detected, and thus there was no need for imputation or removal of data.
- Scaling: The features need to be rescaled so that all input variables are in the same range, since models such as LR and GB require all variables to be on a similar scale. In this instance, Standard-Scaler was adopted.
- The StandardScaler centered and scaled the features to unit variance so that all the features were on the same scale.

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

μ = Mean

σ = Standard Deviation

C. Exploratory Data Analysis

TABLE I
STATISTICS OF DATA

Statistic	Cement	Blast Furnace Slag	Fly Ash	Water
Count	1030.00	1030.00	1030.00	1030.00
Mean	281.17	73.90	54.19	181.57
Std	104.51	86.28	63.99	21.35
Min	102.00	0.00	0.00	121.80
25%	192.38	0.00	0.00	164.90
50%	272.90	22.00	0.00	185.00
75%	350.00	142.95	118.30	192.00
Max	540.00	359.40	200.10	247.00

TABLE II
STATISTICS OF DATA

Statistic	Super plasticizer	Coarse Aggregate	Fine Aggregate	Age
Count	1030.00	1030.00	1030.00	1030.00
Mean	6.20	972.92	773.58	45.66
Std	5.97	77.75	80.18	63.17
Min	0.00	801.00	594.00	1.00
25%	0.00	932.00	730.95	7.00
50%	6.40	968.00	779.50	28.00
75%	10.20	1029.40	824.00	56.00
Max	32.20	1145.00	992.60	365.00

The Heatmap uses the dataset's correlation matrix for the prediction of concrete's compressive strength with various features. Each cell represents two feature variables, and the correlation coefficient between them varies from -1 to 1.

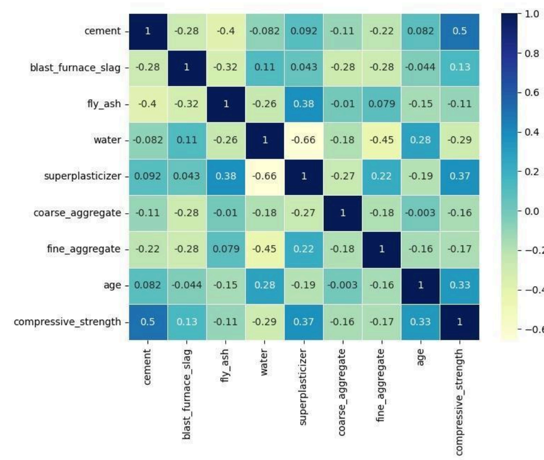

Fig. 2. Heatmap

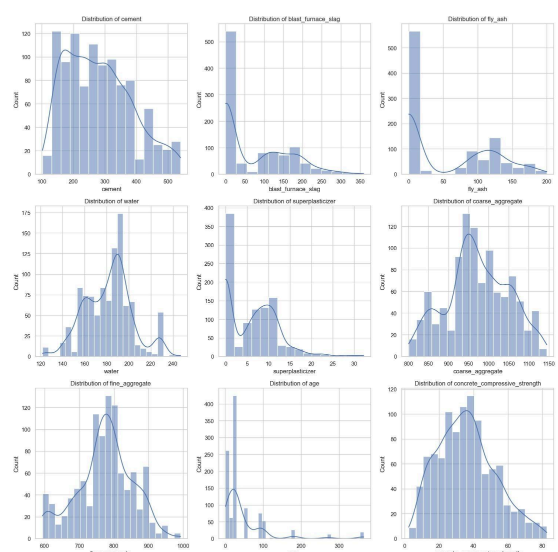

Fig. 3. Data Distribution

D. Train-Test Split

The data was split into 80% for training and 20% for testing. The training constituent provided 824 samples and the testing constituent provided 206 samples.

E. Model Training

CatBoost is an advanced gradient boosting algorithm which has been developed to work effectively with both categorical and numerical data and is most effective in the presence of many complicated structures.[3] In the case of CatBoost Regression, the method is aimed at reducing the gap between predicted and true continuous values of the target variable by using an ensemble of decision trees. These trees are constructed in a serial manner. Each of the trees is built with the aim of correcting the errors made by the last constructed trees. The key idea behind CatBoost, like other gradient boosting methods, is to build the model in a stage-wise manner by minimizing a loss function. The optimization of this loss function enhances prediction accuracy throughout successive iterations. Through the process of correcting errors that were produced in earlier decision trees, CatBoost constructs a sequence of decision trees in order to make predictions regarding the concrete's compressive strength. In order to reduce the amount of error that occurs, the model uses the gradient boosting technique to optimize both the predicted and real strength values. A complex relationship that exists between the raw materials of concrete and other variables is well addressed by this approach, which also prevents an overfitting condition. This ensures that the compressive strength estimations are accurate and robust based on the input features.

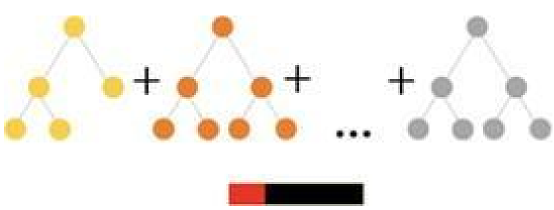

Fig. 4. Catboost regression

F. Hyperparameter Tuning

Randomized-Search CV is used for hyperparameter tuning, as manually selecting parameters for gradient boosting can be so much time-taking process. It helps to find the best parameters for the improvement of model accuracy. Hyperparameter tuning was accomplished via Randomized-Search CV for the objective of optimizing the CatBoost model through the selection of various combinations of hyperparameters such as Learning Rate (controls the size of the step taken in each iteration to minimize the loss function; lower values provide more gradual and stable learning but may require more iterations), Tree Depth (defines the maximum depth of each tree; deeper trees capture more complex patterns), Number of Trees (Iterations) (specifies the total number of boosting rounds or trees added; more trees increase model complexity but can also enhance predictive power), Regularization Factors like L2 Leaf Reg (applies regularization to the leaf values in each tree, helping to reduce overfitting by penalizing large values). These hyperparameters are one of the crucial determinants of the model's performance, and adjusting or 'tuning' them is usually aimed at enhancing the accuracy of the model while achieving a balance between overfitting and underfitting the model.[3] The procedure included 3-fold cross validation, which means the dataset was split into three parts: Out of which 2 folds were used for training, and 1-fold was assigned for validation. This complete cycle was repeated three times for each of the fifty hyperparameters settings tested. Therefore, the total of training and evaluation cycles was three times 50. Which gave one hundred and fifty training cycles in total to measure the best performing combination.

G. Training and Combining Neural network model and ensemble model

Using Keras, a Neural Network (NN) is constructed to predict the compressive strength of target concrete. There is an input layer responsible for scaling of the features, which is followed by two hidden layers of 64 neurons with ReLU activation each.[7] The last layer contains a single neuron for output for regression purposes.

The model uses Adam as its optimizer, and Mean Squared Error (MSE) as the loss function during training, which is executed for a hundred epochs with a batch size of 32 using a training set and verified on the test set. After training, the Neural network model generates predictions on the test data.[5][6]

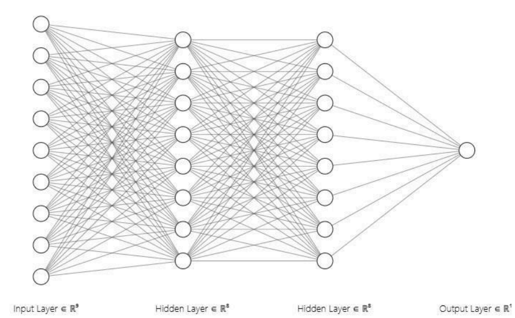

Fig. 5. Neural Network

Combining Catboost and Neural network models: Neural networks helps to find complex , non-linear relations between features. Catboost performs well in regression tasks. CatBoost and neural networks capture different types of patterns. The predictions of both the CatBoost and Neural network models are placed in a new data set, where each row comprises predictions from both the models. This combined data set is further used for training a Linear Regression model. Finally, the combined regression model learns how best to combine the predictions of CatBoost and the Neural network in order to achieve an improved accuracy. The model's results are finally validated by using metrics like RMSE, Mean absolute error(MAE) and the R^2 score.

H. Model Evaluation Metrics

The model is evaluated by comparing with other Models using MSE, MAE and R2 to find accurate model.[8] RMSE measures the square root of the mean of the squared differences between predicted and actual values.

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N} (y_t - \hat{y})^2} \qquad (2)$$

R^2 measures the proportion of the variance in the dependent variable that is predictable from the independent variables.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2} \qquad (3)$$

Where,

$\hat{y}$ = predicted value of y

$\bar{y}$ = mean value of y

MAE measures the average of the absolute differences between actual and predicted values.

$$MAE = \frac{1}{N}\sum_{t=1}^{N} |y_t - \hat{y}| \qquad (4)$$

IV. Results and Discussion

CatBoost works better than other models due to the ability to handle diverse feature structures and prevent overfitting through progressive error correction. On the other hand, ANN finds the non-linear relationship between features, A hybrid approach that combines these predictions with linear regression is followed. The results from evaluating various ensemble techniques, including Random Forest, Linear Regression , Gradient Boosting, CatBoost, Decision Tree showed that CatBoost demonstrated the highest accuracy in prediction of concrete's compressive strength. CatBoost had the lowest RMSE and MAE, and the highest R^2 score, i.e 0.934 , indicating that it captured the relationships between concrete ingredients and compressive strength more accurately than other models. This success can be attributed to CatBoost's gradient boosting method, which systematically reduces prediction errors.

TABLE III
Model evaluation metrics for different algorithms

Model	RMSE	MAE	R_2
Catboost regressor	4.121622	2.662943	0.934073
XGBoost Regressor	4.874561	3.014511	0.907786
Random Forest	5.246098	3.656626	0.893194
Gradient Boosting	5.319677	3.885305	0.890177
Decision Tree	6.852775	4.223471	0.817487
KNN Regressor	8.725250	6.985301	0.704553
Support Vector Regressor	9.460397	7.536116	0.652630
Linear Regression	9.785262	7.756544	0.628405

The Hyperparameter tuning, especially for CatBoost using RandomizedSearchCV, played a key role in improving model performance by finding optimal combinations of parameters. While the Neural Network model was trained on standardized data using a two-hidden-layer architecture. The predictions of both models were utilized to train a Linear Regression metamodel. The hybrid (stacked) model's predictions were tested using RMSE and R^2 scores to provide a comprehensive evaluation of ensemble performance. The hybrid model achieved an R^2 value of 0.9422, explaining 94% of the variance in concrete compressive strength and RMSE of 3.85, indicating accurate predictions.

V. Ethical Considerations and Limitations

Although there are some limitations of our hybrid CatBoost - ANN model, which can be termed as real world challenges, it also shows a lot of promise.

A. *Generalization with Different Mixtures:*

This model was trained and tested on a particular set of concrete data, and hence, may not be applied to other concrete environments, which can be an entirely different setting. To genuinely render this model usable in diverse settings, it will require data from a diverse training environment.

B. *Computation Needs:*

Our hybrid model makes use of CatBoost, an artificial neuron network, and a linear regression layer, and as a result, it consumes significant computational resources. Hence, resource demanding training and fine-tuning of these models can bring about inaccessibility within small organizations or projects with a low level of computation resources.

C. *Difficulties Associated with the Real-Time Application:*

The integration of this model within a construction site in real time would seem another difficult task. In such cases, we would like to look into simplifying the model or some other approaches such as purpose built hardware that would allow for quicker responses without compromising functionality.

There are some constraints at the moment, and it would be possible to overcome them by performing the future research of this nature by increasing the number of samples and the variability of the environmental conditions measured within the dataset. This will enhance better generalization of the model across the different conditions. Moreover, we could also test other methods, such as transfer learning, that would incorporate changes to the model without most of the retraining. Simplifying the model to reduce computational demands or incorporating real-time data from IoT sensors could also make it more practical for day-to-day use in construction.

VI. Conclusion

This study introduced a hybrid model using CatBoost, Artificial Neural Networks (ANN), and Linear Regression to predict concrete's compressive strength, achieving a solid accuracy of 94%. CatBoost handles categorical features and complex interactions, ANN models non-linear relationships, and Linear Regression adds interpretability. Combining these methods make an accurate model for the prediction task. The model's high accuracy highlights the value of hybrid approaches in handling varied data and complex dependencies, as seen in concrete's strength prediction. This method not only boosts accuracy but also provides flexibility for civil engineers and researchers, especially useful in concrete design and construction.

However, We can improve furthermore. Tuning the model further, applying advanced feature selection, or using more real-world data from different environmental conditions could enhance performance. Future research might even look at using the model in real-time settings, integrating it with IoT sensors to predict strength dynamically. Overall, this hybrid approach represents a step forward in predicting concrete strength, showing how machine learning can enhance engineering efficiency and cut down on the costs of traditional testing.

References

[1] O. Abuodeh, J. A. Abdalla and R. A. Hawileh, "Prediction of Compressive Strength of Ultra-High Performance Concrete using SFS and ANN," 2019 8th International Conference on Modeling Simulation and Applied Optimization (ICMSAO), Manama, Bahrain, 2019, pp. 1-5, doi: 10.1109/ICMSAO.2019.8880452.

[2] J. Zhu, S. Fang, Z. Yang, Y. Qin and H. Chen, " Prediction of Concrete Strength Based on Random Forest and Gradient Boosting Machine," 2023 IEEE 3rd International Conference on Power, Electronics and Computer Applications (ICPECA), Shenyang, China, 2023, pp. 306-312, doi: 10.1109/ICPECA56706.2023.10075839.

[3] C. Shi, B. Xi, L. Shen and C. Liu, "Concrete compressive strength prediction model based on RS-Catboost algorithm," 2024 4th International Symposium on Computer Technology and Information Science (ISCTIS), Xi'an, China, 2024, pp. 428-431, doi: 10.1109/ISCTIS63324.2024.10698947.

[4] SP, Siddique Ibrahim, et al. "Deep Learning and AI-Powered Natural Catastrophes Warning Systems." Utilizing AI and Machine Learning for Natural Disaster Management. IGI Global, 2024. 274-292.

[5] P. Singh and P. Khaskil, "Prediction of Compressive Strength of Green Concrete with Admixtures Using Neural Networks," 2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON), Greater Noida, India, 2020, pp. 714-717, doi: 10.1109/GUCON48875.2020.9231230.

[6] F. H. Chiew, "Prediction of Blast Furnace Slag Concrete Compressive Strength Using Artificial Neural Networks and Multiple Regression Analysis," 2019 International Conference on Computer and Drone Applications (IConDA), Kuching, Malaysia, 2019, pp. 54-58, doi:10.1109/IConDA47345.2019.9034920.

[7] S. P. S. Ibrahim, Nithin, S. M. A. Kareem and G. V. Kailash, "Weed Net: Deep Learning Informed Convolutional Neural Network Based Weed Detection in Soybean Crops", 2023 3rd International Conference on Mobile Networks and Wireless Communications (ICMNWC), Tumkur, India, 2023, pp. 1-8, doi: 10.1109/ICMNWC60182.2023.10435726.

[8] S. P. S. Ibrahim, I. Rakshitha, T. Vasisri, R. H. Aswitha, M. R. Rao and D. V. Krishna,"Revolutionizing Solar Generation Data Mining through Advanced Machine Learning Algorithms: Novel Insights and Results", 2023 7th International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS), Bangalore, India, 2023, pp. 1-8, doi: 10.1109/CSITSS60515.2023.10334112.