Journal of Science and Transport Technology
Journal homepage: https://jstt.vn/index.php/en

Experimental Investigation and Machine Learning-Based Prediction of Marshall Stability of Stone Mastic Asphalt Using Steel Slag

Trong-Anh-Minh Nguyen, Hoang-Long Nguyen, Ngoc-Hung Tran, Thanh-Hai Le*
University of Transport Technology, 54 Trieu Khuc, Thanh Liet, Hanoi, Vietnam

Article info
Type of article:
Original research paper

DOI:
https://doi.org/10.58845/jstt.utt.2026.en.6.3.41-57

***Corresponding author:**
Email address:
hailt@utt.edu.vn

Received: 21/03/2026
Received in Revised Form: 15/06/2026
Accepted: 25/06/2026

Abstract: This study applies several machine learning models, including Gradient Boosting Regressor (GBR), CatBoost (CB), Support Vector Machine (SVM), Random Forest (RF), and AdaBoost (AB), to predict the Marshall stability of Stone Mastic Asphalt (SMA) mixtures incorporating steel slag as a partial replacement for conventional coarse aggregates. The dataset consists of 144 Marshall test samples, with input variables including specific gravity, penetration, flash point temperature, softening point temperature, bitumen content, cellulose fiber content, and steel slag content, while Marshall stability is considered as the output variable. The results indicate that all models achieved high predictive accuracy, with CatBoost providing the best performance, achieving R^2 = 0.936, RMSE = 0.227, and MAE = 0.323 on the validation dataset. Additional analyses, including residual analysis and SHAP-based interpretation, suggest the robustness and interpretability of the model within the investigated dataset of the model. These findings highlight the strong potential of CatBoost for accurately predicting the mechanical performance of SMA mixtures and support decision-making in mixture design in pavement engineering.

Keywords: Stone Mastic Asphalt (SMA); Steel slag; Marshall stability; Machine learning; CatBoost.

1. Introduction

In recent years, the use of industrial by-products and recycled materials in pavement construction has received increasing attention in order to enhance sustainability and reduce environmental impacts. Among these materials, steel slag, a by-product of the steel manufacturing industry, has attracted considerable interest as a potential alternative aggregate in asphalt mixtures. Steel slag possesses several advantageous characteristics, including high hardness, rough surface texture, and good mechanical strength, which enhance the interlocking effect between aggregates and improve the load-bearing capacity of asphalt mixtures [1]. In particular, the incorporation of steel slag into Stone Mastic Asphalt (SMA) mixtures has been considered a promising approach to improve the mechanical performance and durability of asphalt pavements. However, the mechanical behavior of SMA mixtures is influenced by multiple factors, including binder content, aggregate structure, and material properties, which interact through complex nonlinear mechanisms [2]. Therefore, accurately evaluating the mechanical performance of these mixtures remains a significant challenge.

In asphalt mixture design, Marshall stability is an important parameter for evaluating the load-bearing capacity and deformation resistance of asphalt mixtures under traffic loading conditions [3]. This parameter is determined through the Marshall test, in which the maximum load that a compacted asphalt specimen can withstand before failure is measured according to standard procedures such as ASTM D6927 or AASHTO T245 [4]. Therefore, accurate determination of Marshall stability plays an important role in the design and optimization of asphalt mixtures.

Traditionally, the mechanical properties of asphalt mixtures are evaluated through laboratory experiments, particularly the Marshall test [5]. Although these experimental methods provide reliable results, they require complex procedures for specimen preparation, compaction, and mechanical testing, making the process time-consuming, labor-intensive, and costly, especially when a large number of mixture design alternatives need to be investigated [6]. Consequently, many studies have attempted to develop empirical and statistical models to estimate Marshall stability based on mixture design parameters. However, conventional regression-based models often rely on simplified assumptions and are limited in their ability to capture the complex nonlinear relationships between mixture components and the mechanical performance of asphalt materials [6].

In recent years, machine learning (ML) techniques have emerged as powerful tools for addressing complex engineering problems due to their ability to automatically learn nonlinear relationships directly from data without requiring explicit mathematical formulation [7]. ML methods have been widely applied in pavement engineering, including the prediction of asphalt mixture performance, pavement distress, and material properties [8]. Several studies have demonstrated that models such as Random Forest (RF) [9], Gradient Boosting (GB) [10], and Support Vector Machine (SVM) [11] can effectively predict asphalt mixture properties with high accuracy based on mixture composition and material characteristics [12]. These models are capable of identifying complex relationships among input variables, thereby improving prediction accuracy compared with traditional statistical approaches. Despite the growing application of machine learning techniques in pavement engineering, studies focusing on the prediction of Marshall stability for SMA mixtures incorporating steel slag remain relatively limited [13]. In particular, the comparative performance of different machine learning algorithms for predicting Marshall stability using experimentally derived datasets has not been sufficiently explored. Moreover, advanced boosting algorithms such as CatBoost, which have demonstrated strong predictive performance in many engineering applications, have rarely been investigated in the context of asphalt mixture performance prediction [14].

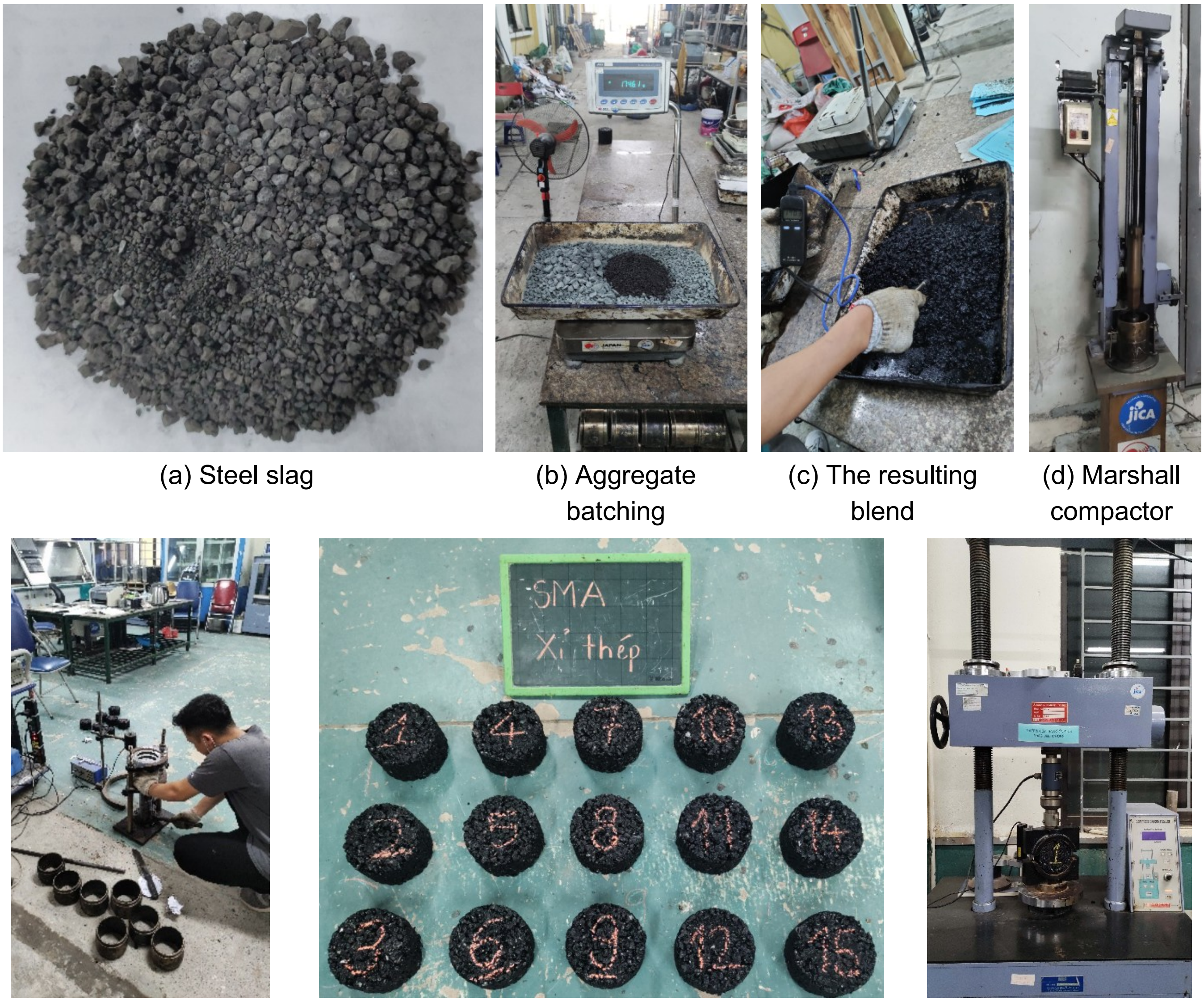
Motivated by these research gaps, this study investigates the application of machine learning techniques for predicting the Marshall stability of Stone Mastic Asphalt (SMA) mixtures incorporating steel slag. The proposed approach aims to explore the capability of data-driven models to capture the complex relationships between mixture design parameters and the mechanical performance of asphalt mixtures [15]. The main objectives of this study are: (i) to develop predictive models for estimating Marshall stability based on mixture design variables; (ii) to evaluate and compare the predictive performance of different machine learning approaches; and (iii) to provide insights into the influence of mixture design parameters on the mechanical behavior of SMA mixtures. The findings of this study are expected to contribute to improving the efficiency of asphalt mixture design and supporting the sustainable use of industrial by-products in pavement engineering.

2. Materials

In this study, the coarse and fine aggregates were crushed basalt sourced from the Phu Man quarry (Quoc Oai, Ha Noi), while the mineral filler was obtained from the Kien Khe quarry (Ninh

Binh). Polymer Modified Bitumen (PMB III), supplied by Petrolimex Asphalt Co., Ltd. (Viet Nam), was utilized as the binder. To ensure mixture stability, an organic cellulose fiber (ARBOCEL® ZZ 8/1, manufactured by JRS, Germany) was incorporated as a stabilizing additive at a dosage of 0.3% by total weight of the mixture. Additionally, Electric Arc Furnace (EAF) steel slag was provided in a crushed form by the Thanh Dai Phu My Joint Stock Company (Nam Cau Kien Industrial Park, Hai Phong city).

The experimental tests were conducted in accordance with the specifications of AASHTO M325-08 and TCCS 36:2021/TCĐBVN to evaluate the mechanical characteristics of the SMA mixtures [16]. All specimens were prepared using an automatic Marshall compactor manufactured by Matest (Italy). Cylindrical specimens with a height of 63.5 mm and a diameter of 101.6 mm were compacted with 50 blows on each face to ensure consistent compaction energy and specimen uniformity throughout the experimental process. The Marshall specimen preparation and testing procedure is illustrated in Fig. 1.



(a) Steel slag (b) Aggregate batching (c) The resulting blend (d) Marshall compactor

(e) Specimen extrusion (f) Samples at room temperature (g) Marshall Compression

Fig. 1. Marshall test procedure for SMA mixtures

The dataset used in this study consists of 144 specimens, developed to evaluate the influence of bitumen content, cellulose fiber content, and steel slag replacement ratio on the Marshall stability of asphalt mixtures. A total of 75 Stone Mastic Asphalt (SMA) specimens were prepared by partially replacing conventional coarse aggregates with steel slag. Two primary mixture design parameters were considered, namely bitumen content and steel slag replacement ratio, in order to evaluate their influence on the mechanical properties of SMA mixtures. The bitumen content was varied from 5.5% to 7.5%, while the steel slag replacement ratio ranged from 0% to 20%. In all mixtures, the cellulose fiber content was maintained at 0.3%, and the nominal maximum aggregate size of the SMA mixture was fixed at 12.5 mm according to standard SMA mixture design requirements. The remaining 69 specimens were sourced from the previous research conducted by Le [17], where the bitumen content ranged from 5.5% to 7.5%. In that research, cellulose fiber dosages were investigated at proportions of 0%, 0.2%, 0.3%, and 0.5% by weight of the total mixture. Since both datasets were based on SMA mixtures and followed comparable Marshall stability testing procedures, they were considered suitable for integration into a unified dataset for machine learning model development.

3. Datasets and methods

3.1. Datasets

The descriptive statistics of the experimental dataset are presented in Table 1, reflecting the variability of mixture design parameters and mechanical properties within the investigated range. In particular, the bitumen content was varied from 5.5% to 7.5%, in combination with the simultaneous adjustment of cellulose fiber content (0–0.5%) and steel slag replacement ratio (0–20%), resulting in a diverse range of SMA mixture compositions. The physical properties of bitumen exhibit low variability, indicating consistent material quality throughout the experimental program. Variations in mixture composition led to noticeable changes in Marshall stability, with a mean value of 10.504 kN and a range from 8.65 kN to 13.521 kN, highlighting significant differences in the load-bearing capacity of the mixtures. Notably, most values are concentrated within the interquartile range of 9.635–11.245 kN, indicating a generally moderate to high level of stability, while higher values suggest the potential for substantial improvement in mechanical performance under optimal mixture conditions.

Table 1. Descriptive statistics of the variables used in the SMA experimental dataset

Parameter	Mean	Std	Min	25%	50%	75%	Max
Specific gravity	1.036	0.002	1.034	1.034	1.037	1.037	1.037
Penetration	49.438	1.504	48	48	48	51	51
Flash point temperature	251.646	3.509	248	248	255	255	255
Softening point temperature	87.438	1.504	86	86	86	89	89
Bitumen content	6.517	0.693	5.5	6	6.5	7	7.5
Cellulose fiber content	0.281	0.124	0	0.3	0.3	0.3	0.5
Steel slag content	0.052	0.072	0	0	0	0.1	0.2
Marshall stability (kN)	10.504	1.156	8.65	9.635	10.23	11.245	13.521

The statistical distributions of the dataset variables are shown in Fig. 2, where most variables exhibit relatively narrow ranges due to the controlled laboratory experimental conditions. However, Marshall stability presents a wider distribution, reflecting the substantial variation in the mechanical performance of SMA mixtures with different steel slag contents. This variability confirms that the dataset captures meaningful changes in mixture performance and provides a suitable basis for subsequent statistical analysis and machine learning model development.

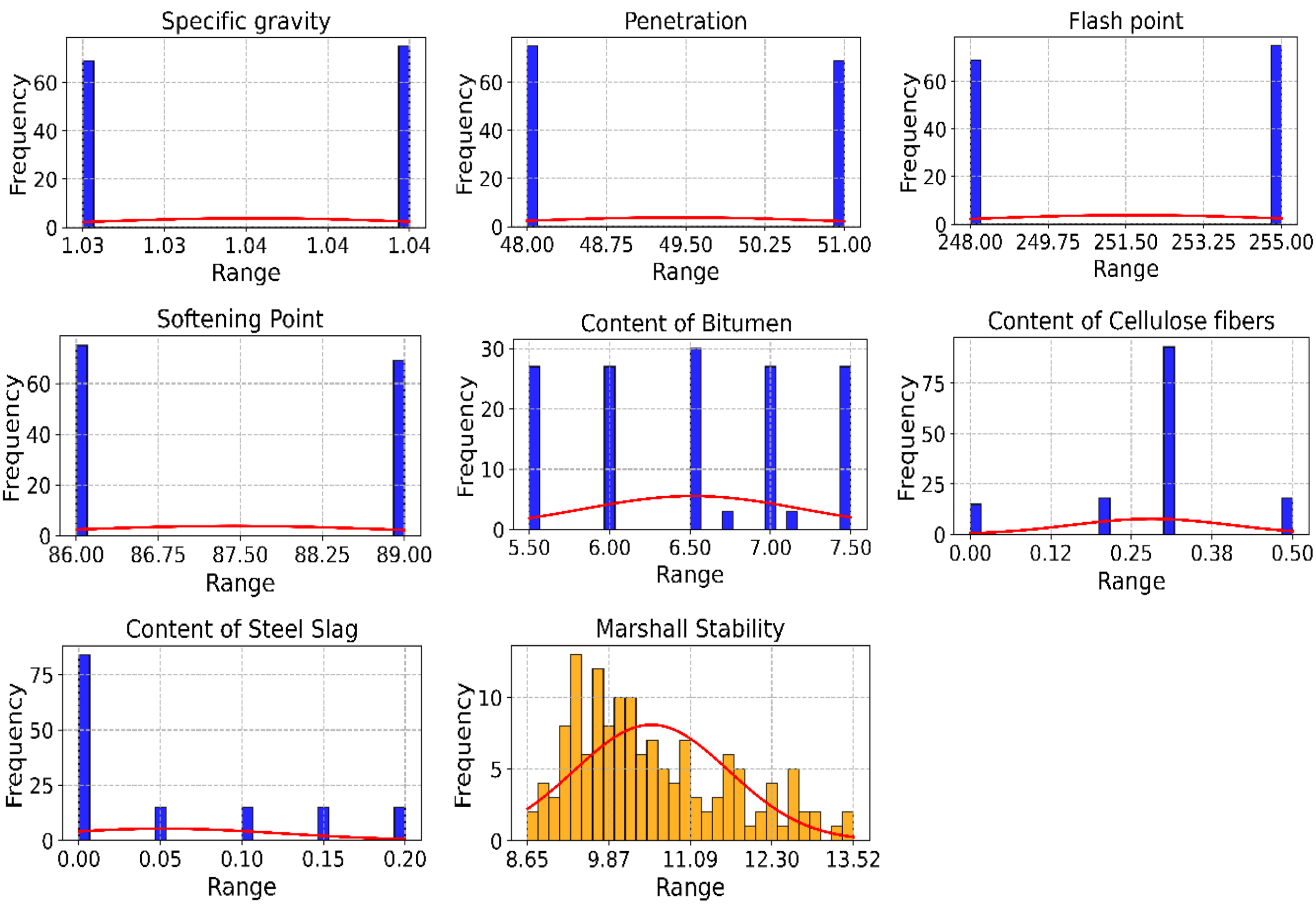


Fig. 2. Statistical distributions of the input variables and Marshall stability in the dataset

The correlation matrix in Fig. 3 illustrates the relationships among the investigated variables and Marshall Stability. Steel slag content showed a strong positive correlation with Marshall Stability (r = 0.82), indicating that the incorporation of steel slag plays an important role in improving the load-bearing capacity of SMA mixtures. Although this correlation is relatively high, it represents the relationship between an input variable and the target variable; therefore, steel slag content was retained as an important predictor in the model. Meanwhile, severe inter-correlation was observed among the binder-related variables, including Specific gravity, Penetration, Flash point, and Softening Point, with correlation coefficients approaching ±1.00. This indicates that these variables contain highly overlapping information when used simultaneously as model inputs. To reduce redundancy and improve the reliability of model interpretation, Specific gravity, Flash point, and Softening Point were removed from the input dataset, while Penetration was retained as a representative binder property. Penetration was selected because it directly reflects the consistency and hardness of bitumen, which are closely related to the stiffness and load-resisting behavior of asphalt mixtures. In addition, bitumen content and cellulose fiber content showed weak correlations with Marshall Stability within the investigated range; however, they were retained because they are essential mixture design parameters. Consequently, the final input variables used for model development included Penetration, Content of Bitumen, Content of Cellulose fibers, and Content of Steel Slag, while Marshall Stability was selected as the output variable.

3.2. Methods

3.2.1. Gradient Boosting Regression (GBR)

The GBR algorithm is a machine learning method belonging to the ensemble learning family, in which multiple weak learners are constructed sequentially to progressively improve the predictive capability of the overall model [18]. Instead of building a single complex model, GBR forms the final prediction by combining a series of simple regression models, where each newly added model is trained to approximate the residual errors produced by the previous models. The training process typically begins with a simple initial model, such as the mean value of the target variable. At each iteration, the discrepancy between the predicted and observed values is used to train a weak regression model, commonly a decision tree [19]. The newly trained model is then incorporated into the overall model through a learning rate parameter that controls the magnitude of each update. By repeatedly learning from the residuals, GBR gradually reduces prediction errors and improves the overall predictive accuracy of the model.

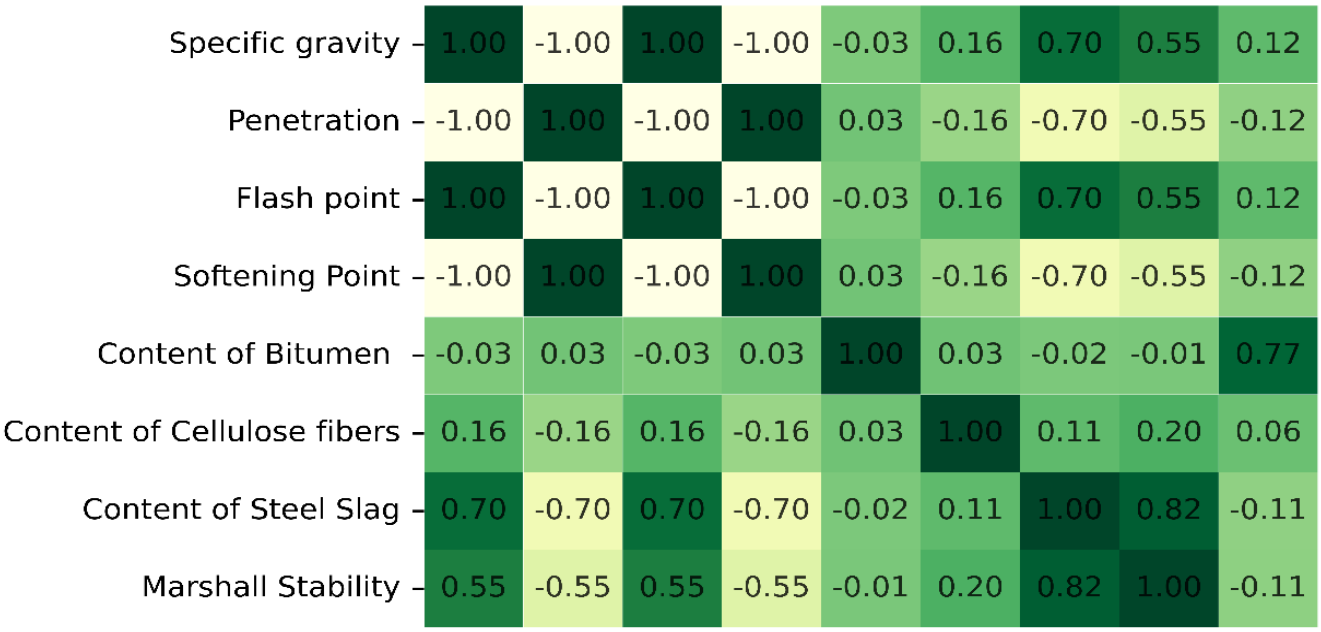


Fig. 3. Correlation matrix among the variables of the SMA experimental dataset

3.2.2. CatBoost (CB)

CatBoost is a machine learning algorithm belonging to the gradient boosting family based on decision trees, developed to address limitations commonly observed in traditional boosting methods such as Gradient Boosting and XGBoost, particularly prediction bias and overfitting when the dataset size is relatively small [20]. The core idea of CatBoost lies in its ordered boosting strategy, where training samples are processed in a randomly permuted order and the gradient at each iteration is computed only from samples that appear earlier in that order. This mechanism effectively reduces prediction bias during the training process. In addition, CatBoost employs symmetric decision trees, in which the same splitting condition is applied to all nodes at the same depth. This structure simplifies the tree architecture, improves computational efficiency, and enhances model stability and generalization capability. Owing to these characteristics, CatBoost can achieve high predictive performance even when trained on relatively limited datasets and has been widely applied in various regression tasks [21].

3.2.3. Support Vector Machine (SVM)

SVM is a machine learning algorithm developed based on the principle of optimizing a separating boundary in the feature space [22]. For regression tasks, SVM is extended to Support Vector Regression (SVR), which aims to find an optimal function that approximates the relationship between input variables and continuous output values. Instead of maximizing the margin between classes, SVR seeks to determine a function that fits the data within a predefined error tolerance, while simultaneously maintaining model simplicity. In this framework, only a subset of training data points, known as support vectors, contributes to defining the regression function. When the relationship between variables is nonlinear, kernel functions can be employed to map the input data into a higher-dimensional feature space, where a linear regression function can be constructed [23]. Common kernel functions include linear, polynomial, and radial basis function (RBF). Due to its strong generalization ability and robustness against overfitting, SVR has been widely used in various regression problems.

3.2.4. Random Forest (RF)

RF is an algorithm constructed from an ensemble of decision trees. Each decision tree in the forest is trained using a randomly generated subset of the original dataset, where both samples and features are selected through a resampling process at different stages of tree construction. During the splitting process of each tree, only a randomly selected subset of features is considered to determine the optimal split. This strategy reduces the correlation among trees, increases model diversity, and helps mitigate the risk of overfitting [24]. When predicting a new sample, the algorithm aggregates the outputs from all trained decision trees to produce the final prediction.

3.2.5. Adaptive Boosting (AB)

AB is a machine learning algorithm belonging to the ensemble learning family, in which multiple simple learners are constructed sequentially to progressively enhance the predictive capability of the overall model [25]. Rather than training models independently, AB introduces a mechanism that dynamically adjusts the weights of training samples after each iteration. Initially, all samples are assigned equal importance; however, during the training process, the weights of incorrectly predicted samples are increased, while those of correctly predicted samples are reduced. As a result, subsequent weak learners focus more on the difficult instances that are harder to predict. By iteratively combining these weak learners, AB gradually improves the predictive performance of the ensemble model while enhancing its robustness and generalization ability, particularly when dealing with datasets characterized by uneven data distributions [26].

3.2.6. Validation indicators

The predictive performance of the model was evaluated using three statistical indicators: the coefficient of determination (R^2), mean absolute error (MAE), and root mean square error (RMSE) [27]. These metrics are widely used to assess the predictive effectiveness of a model. The R^2 value ranges from 0 to 1, where values closer to 1 indicate a stronger agreement between predicted and observed values, reflecting higher model accuracy and better predictive performance. Meanwhile, MAE and RMSE are used to quantify the magnitude of prediction errors between observed and predicted values [27]. MAE represents the average absolute deviation, whereas RMSE emphasizes larger deviations because the errors are squared before averaging. Therefore, smaller MAE and RMSE values indicate better predictive capability of the model [28]. The mathematical formulations of these evaluation metrics are presented as follows:

$$R^2=1-\frac{\sum_{i=1}^{n}\left(c_i-p_i\right)^2}{\sum_{i=1}^{n}\left(c_i-\bar{c}\right)^2} \quad (1)$$

$$MAE=\frac{\sum_{i=1}^{n}\left|c_i-p_i\right|}{n} \quad (2)$$

$$RMSE=\sqrt{\frac{\sum_{i=1}^{n}\left(c_i-p_i\right)^2}{n}} \quad (3)$$

where c_i and p_i denote the observed and predicted values for the i-th sample, respectively, c ̄ is the mean of the observed values, and n represents the total number of samples.

3.2.7. SHapley Additive exPlanations (SHAP)

The SHAP (SHapley Additive exPlanations) method was employed to analyze and interpret the predictive mechanism of the machine learning model by quantifying the contribution of each input feature to the predicted output [29]. This approach is grounded in the Shapley value concept from cooperative game theory, in which each feature is treated as a player contributing to the formation of the model prediction. The influence of a particular feature is determined by examining the change in the model output when that feature is introduced into different subsets of features. By computing the average marginal contribution of a feature across all possible combinations of the remaining variables, SHAP ensures that the attribution of feature importance is conducted in a consistent manner with a clear theoretical foundation [30]. Through this mechanism, SHAP enables the identification of the relative influence of individual input variables on the model predictions while clarifying the relationships between features and the predicted outcomes. In addition to providing insights into feature importance at the global level, the method also allows for detailed interpretation of the contribution of each feature to individual predictions.

4. Results and discussion

In this study, the dataset was normalized using the Min–Max scaling method, where the variables were transformed into a range between 0 and 1 to ensure consistency in the scale of the input variables [31]. After normalization, the dataset was divided into two subsets for model training and validation. Specifically, 70% of the data were used as the training set, while the remaining 30% were used as the validation set to evaluate the predictive performance and generalization capability of the machine learning models [32].

In addition, to optimize model performance, the Grid SearchCV method was employed to identify the optimal set of hyperparameters within a predefined parameter space [33]. This approach evaluates multiple parameter combinations through a cross-validation mechanism, allowing the configuration that yields the best model performance to be selected [34]. Furthermore, an early stopping mechanism was incorporated during the training process to mitigate overfitting and reduce computational time [35]. The training procedure automatically terminates when the model performance no longer improves after a certain number of iterations. The main hyperparameters used in the training and optimization processes are presented in Table 2.

Table 2. Optimized hyperparameters of the machine learning models used in this study

Hyperparameters	GBR	CB	SVM	RF	AB
C	-	-	100	-	-
depth	-	4	-	-	-
epsilon	-	-	0.1	-	-
gamma	-	-	scale	-	-
iterations	-	600	-	-	-
kernel	-	-	rbf	-	-
l2_leaf_reg	-	5	-	-	-
learning_rate	0.1	0.03	-	-	0.1
loss	-	-	-	-	square
loss_function	-	RMSE	-	-	-
max_depth	2	-	-	20	-
max_features	-	-	-	log2	-
min_samples_leaf	2	-	-	1	-
min_samples_split	-	-	-	2	-
n_estimators	250	-	-	300	250
subsample	0.8	-	-	-	-

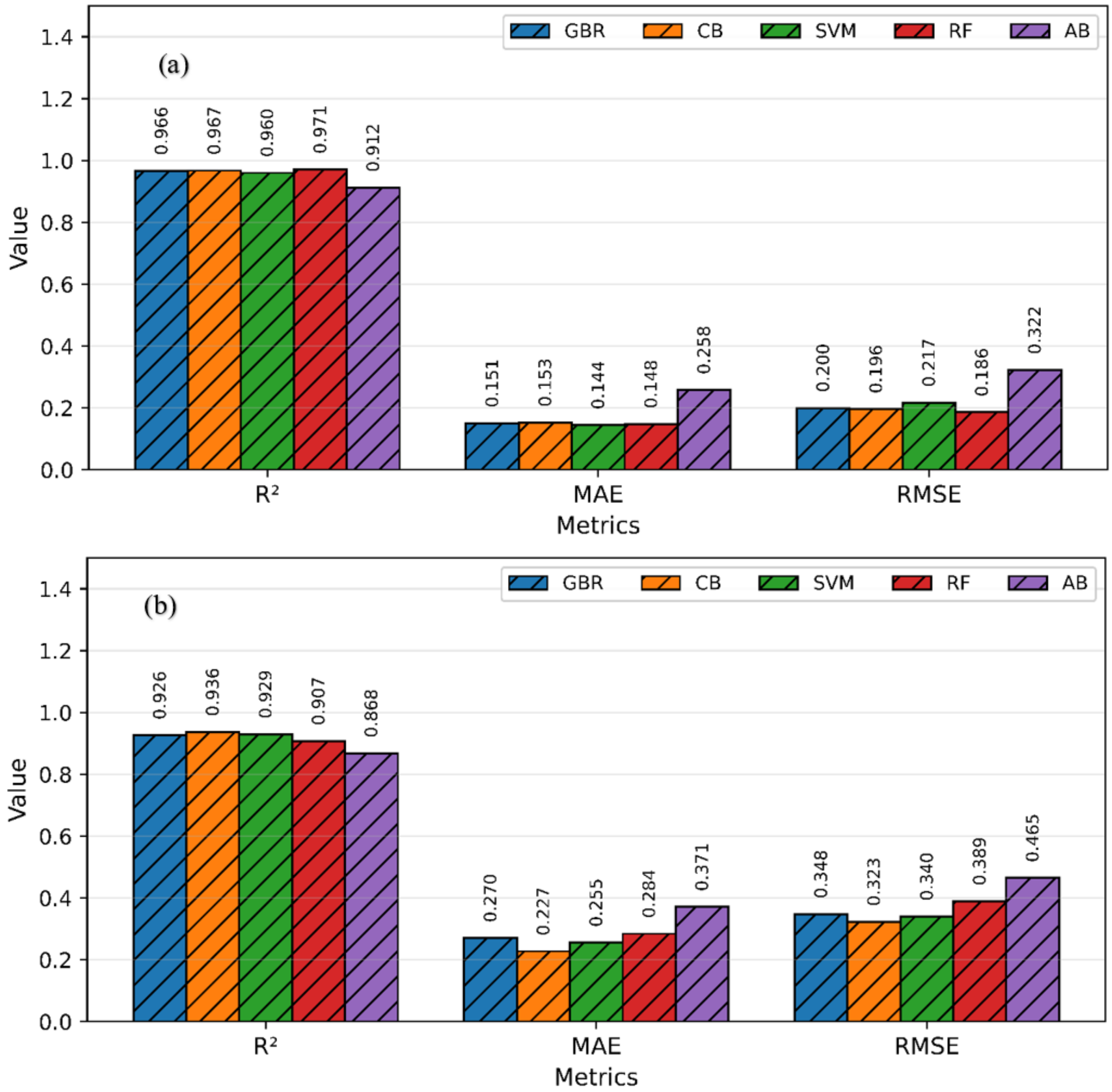


Fig. 4. Performance of machine learning models in predicting Marshall stability on (a) the training dataset and (b) the validation dataset

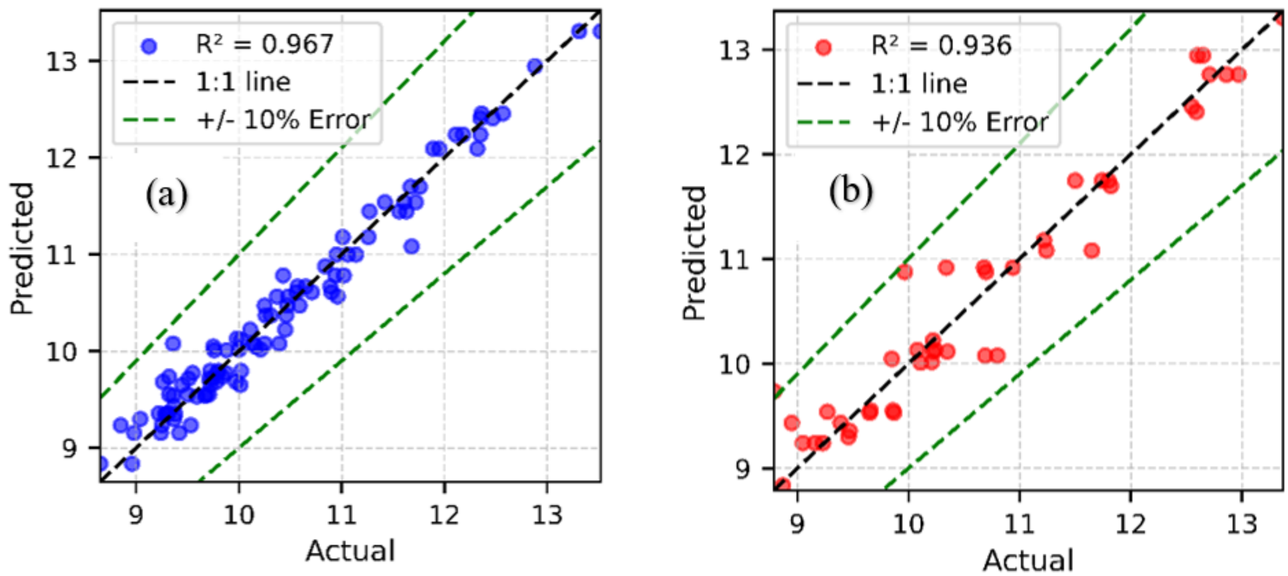


Fig. 5. Coefficient of determination (R^2) of Marshall stability on (a) the training dataset and (b) the validation dataset

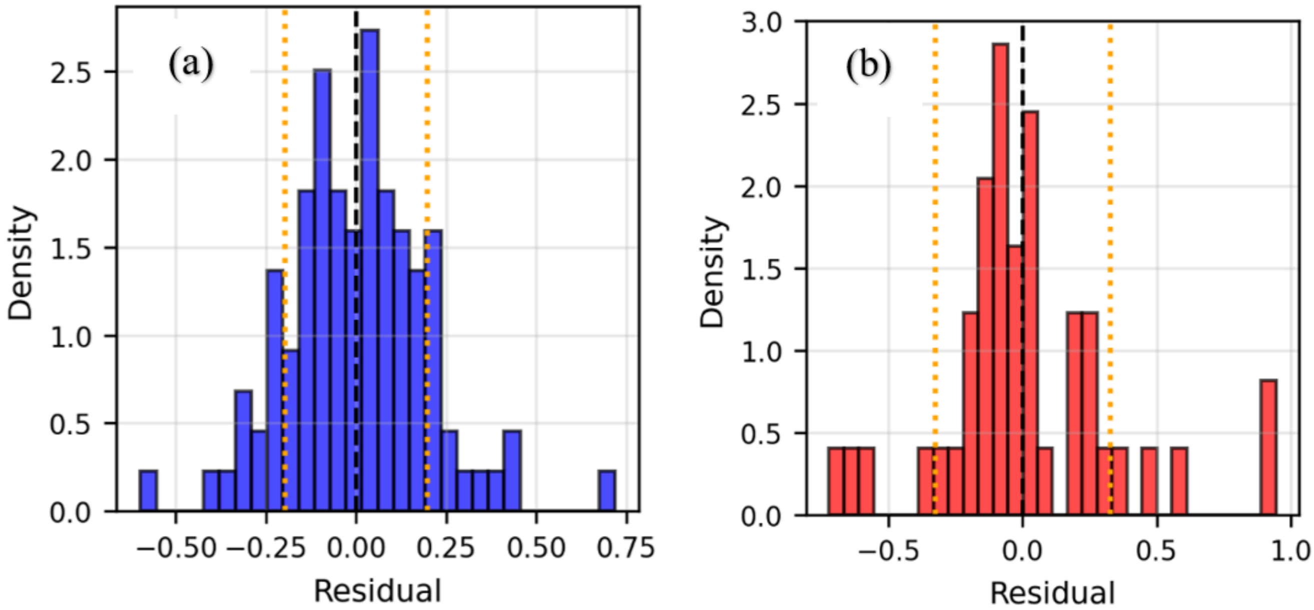


Fig. 6. Residual distribution of the CatBoost model for predicting Marshall stability on (a) the training dataset and (b) the validation dataset

The performance evaluation results of the five machine learning models, including GBR, CB, SVM, RF, and AB, based on the three evaluation metrics R^2, MAE, and RMSE on the training dataset (Fig. 4a) and validation dataset (Fig. 4b), reveal noticeable differences in their learning capability and prediction performance. On the training dataset, all models achieved high R^2 values (>0.90), with RF and CB demonstrating the best performance, achieving R^2 values of 0.971 and 0.967, respectively. These were followed by GBR (0.966) and SVM (0.960), while AB obtained a lower value (R^2 = 0.912). In addition, the MAE and RMSE values of the models were relatively low, ranging from 0.144–0.258 and 0.186–0.322, respectively, indicating that the models effectively learned the patterns in the training data. On the validation dataset, the prediction accuracy of all models decreased compared with the training dataset. Among them, CB achieved the best performance with R^2 = 0.936, RMSE = 0.323 and MAE = 0.227, indicating better predictive capability compared with the other models. The GBR and SVM models also produced relatively stable results with R^2 values of 0.926 and 0.929, respectively, while RF achieved R^2 = 0.907. In contrast, AB showed the lowest performance, with R^2 = 0.868, MAE = 0.371, and RMSE = 0.465, reflecting its weaker predictive capability for the current dataset.

The plots in Fig. 5 present the relationship between the actual and predicted values of Marshall stability obtained by the CatBoost (CB) model on the training dataset (a) and the validation dataset (b), expressed through the coefficient of determination (R^2). The predicted points are closely distributed around the 1:1 line, indicating a strong agreement between the observed and predicted values. On the training dataset, the model achieved R^2=0.967, demonstrating its ability to effectively capture the relationship between the input variables and Marshall stability. On the validation dataset, the model maintained a high predictive accuracy with R^2=0.936, and most prediction points fall within the ±10% error bounds, confirming the good generalization capability and reliability of the model for predicting Marshall stability.

The residual distribution results in Fig. 6 illustrate the differences between the actual and predicted values of the CatBoost model on the training dataset (a) and the validation dataset (b). Most residuals are concentrated around zero, indicating a good agreement between the predicted and observed values. On the training dataset, the residuals are relatively narrowly distributed and mainly concentrated within ±0.25, although a few larger residuals are still observed. This result reflects the strong learning capability of the model on the training data. On the validation dataset, the residual distribution becomes wider, with some residuals exceeding ±0.5, indicating a slight increase in prediction error when the model is applied to unseen data. However, most residuals remain close to zero, suggesting that the model still maintains acceptable predictive performance in estimating Marshall Stability.

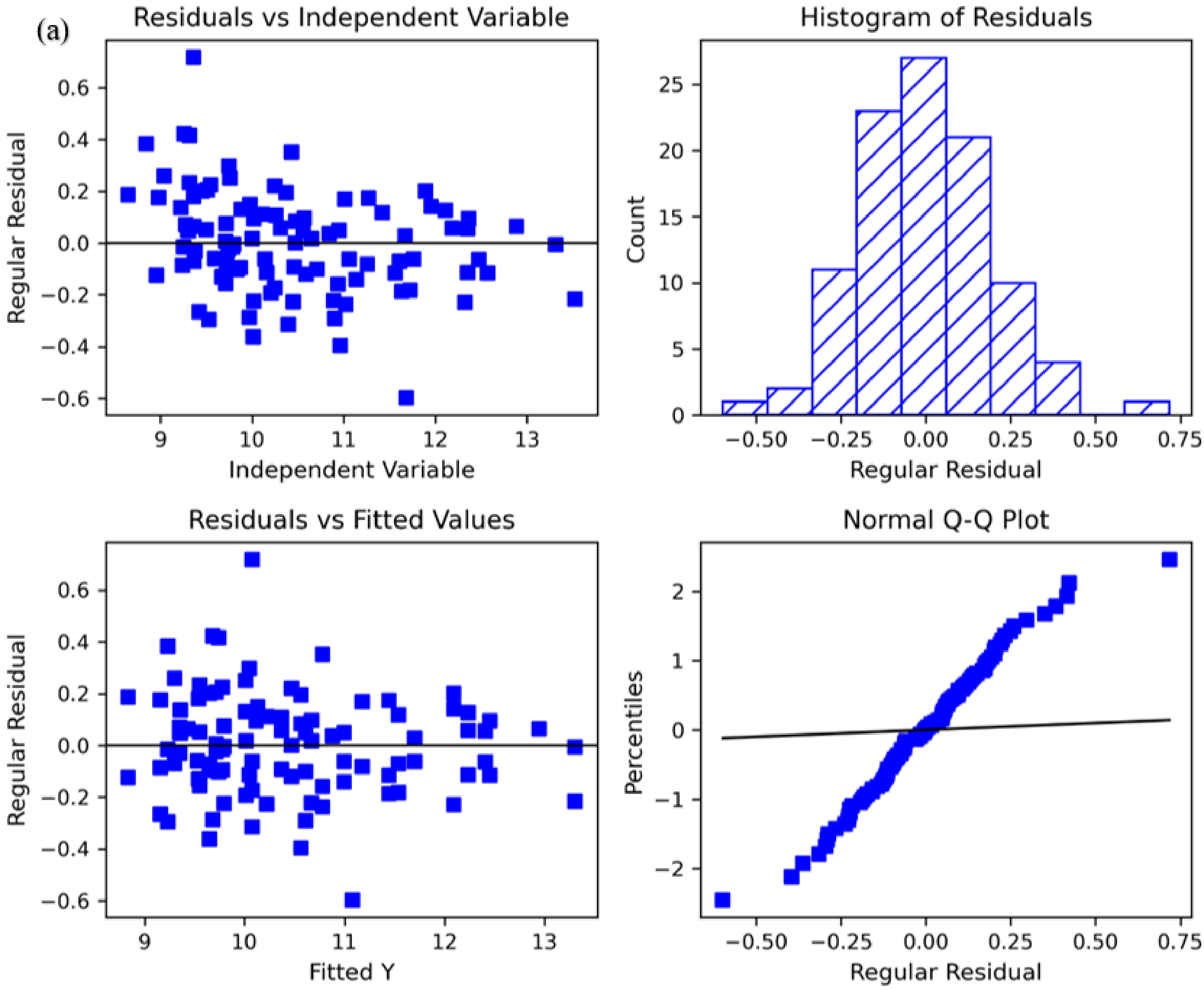


Fig. 7. Residual distribution and Q–Q plot of the optimal CatBoost (CB) model for the training dataset (a) and validation dataset (b)

Residual analysis of the optimal CatBoost (CB) model in Fig. 7 for the training dataset (a) and validation dataset (b) indicates that the model generally exhibits stable prediction behavior. For the training dataset, the residuals are distributed around the zero line when plotted against the independent variable and fitted values, with no clear systematic pattern observed. The histogram of residuals is mainly centered near zero, suggesting that most prediction errors are relatively small. However, the Q–Q plot shows some deviations from the reference line, particularly at the tails, indicating that the residuals do not perfectly follow a normal distribution. For the validation dataset, the residuals also fluctuate around zero, but with a wider dispersion and several larger errors compared with the training dataset. This result is consistent with the lower prediction accuracy observed on the validation dataset. Overall, the residual analysis suggests that the CB model provides acceptable predictive performance, although minor deviations and outliers remain in the residual distribution.

The SHAP summary plot results in Fig. 8 show that Content of Steel Slag is the most influential variable in the predictions of the CatBoost model. Low steel slag content is associated with negative SHAP values, mainly ranging from approximately −0.7 to −0.4, whereas higher steel slag content produces positive SHAP values of about 0.6–1.7, indicating a strong positive contribution to Marshall Stability. Content of Bitumen PMB III also shows a noticeable influence, with SHAP values varying from negative to positive ranges, suggesting that its effect on the predicted output depends on the bitumen content level. Content of Cellulose fibers exhibits a moderate effect, with most SHAP values distributed between approximately −0.8 and 0.4. In contrast, Penetration has a limited contribution, as most of its SHAP values are concentrated close to zero. These findings indicate that the CB model predictions are primarily driven by steel slag content, followed by bitumen PMB III and cellulose fibers, while penetration has the least influence among the selected input variables.

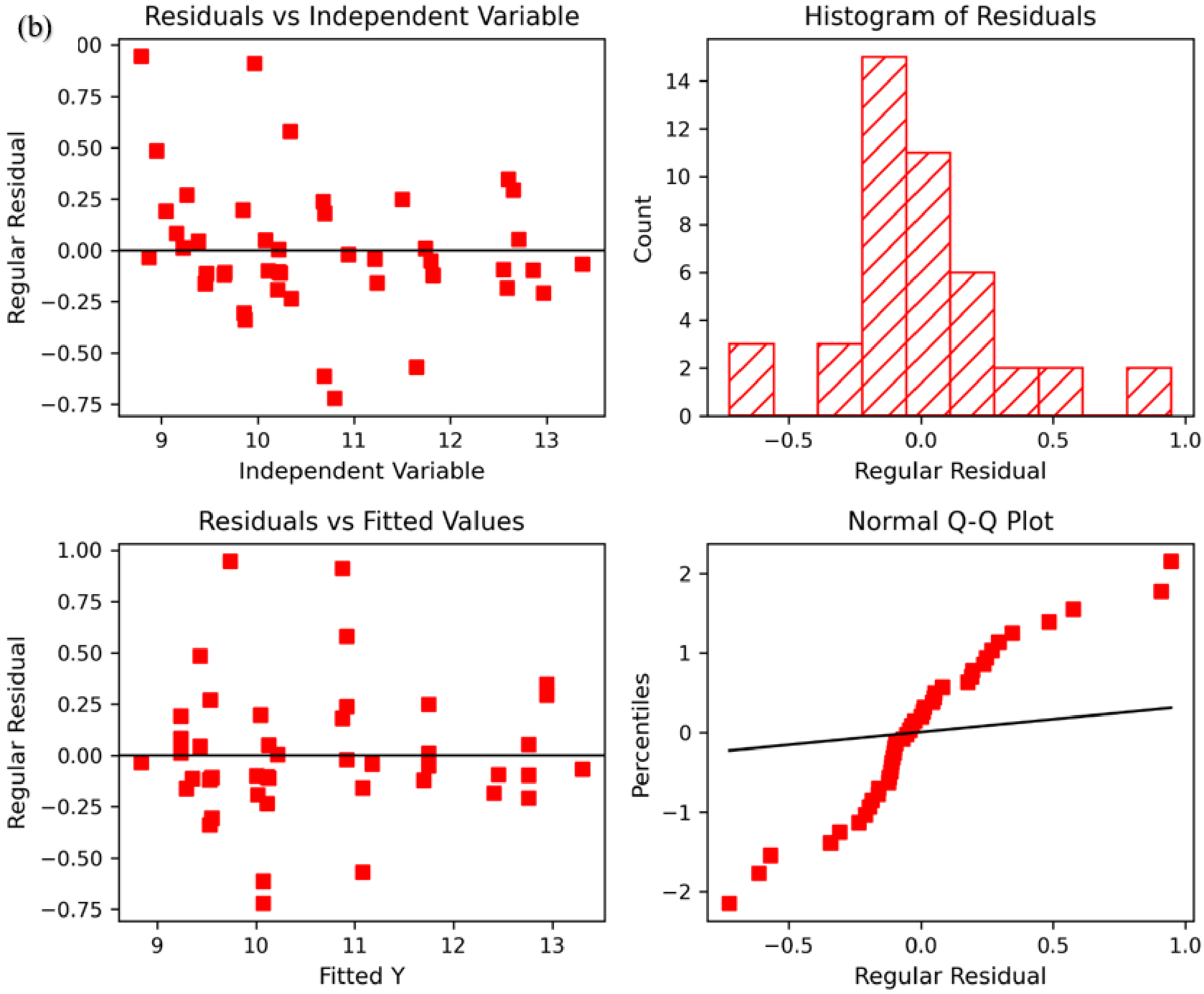


Fig. 7. (continued)

SHAP waterfall analysis illustrates the contribution of each input variable to the CatBoost model prediction for a specific sample (Fig. 9). The prediction starts from the baseline value E[f(X)] = 10.432 and, after incorporating the contributions of individual variables, reaches the final prediction f(x) = 9.716. Among the variables, Content of Steel Slag produces the largest negative contribution (−0.53), followed by Content of Cellulose fibers (−0.30), while Content of Bitumen PMB III provides a positive contribution (+0.12). Penetration has almost no effect on the predicted value. These results indicate that the reduction in the predicted value is mainly driven by steel slag and cellulose fibers, and also highlight that the influence of variables in the SHAP waterfall plot, as a local explanation, may differ from the overall trend observed in the SHAP summary plot, which represents the global explanation.

The results indicate that Content of Steel Slag shows an approximately monotonic increasing relationship with the model predictions (Fig. 10). When the steel slag content is very low or equal to zero, the SHAP values are mainly negative, ranging from approximately −0.7 to −0.4, indicating that this variable reduces the model output. At a steel slag content of around 0.25 on the normalized scale, the SHAP values remain negative, suggesting that its positive influence on the prediction is still limited at low content levels. However, when the steel slag content increases to 0.50 and above, the SHAP values become positive and increase substantially as the content reaches 0.75–1.00, with values ranging approximately from 0.6 to 1.7. This demonstrates a strong positive contribution of steel slag content to the predicted Marshall Stability. This pattern suggests the existence of a threshold effect, where the influence of steel slag becomes more significant at medium to high content levels, highlighting its important role in enhancing the predictions of the CatBoost model.

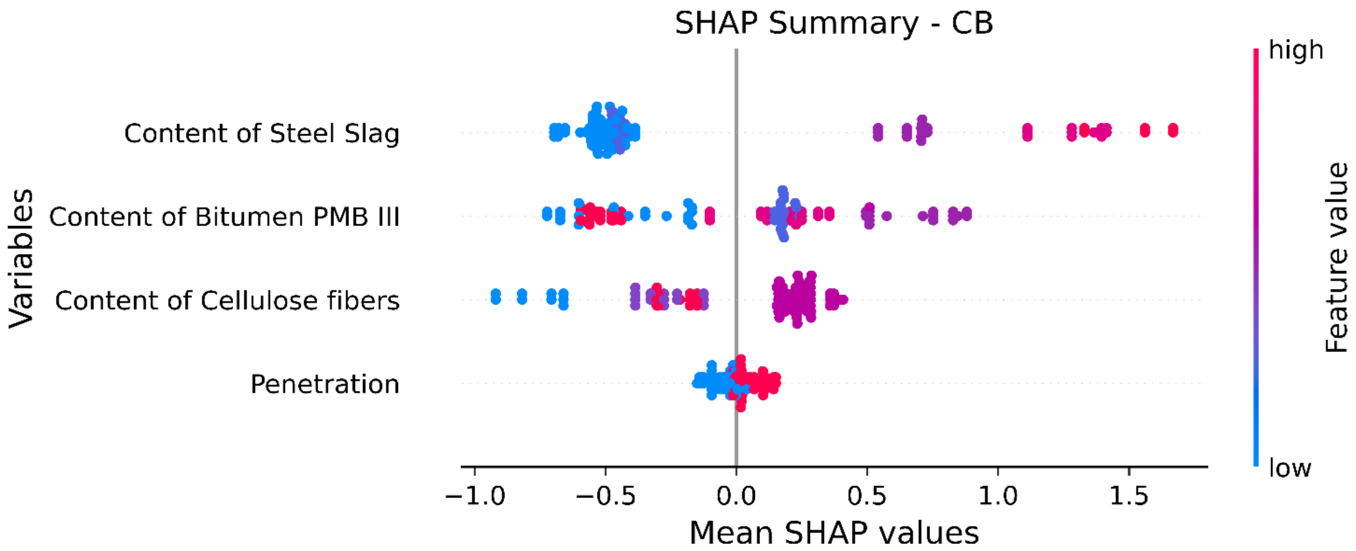


Fig. 8. SHAP summary plot for the CatBoost (CB) model

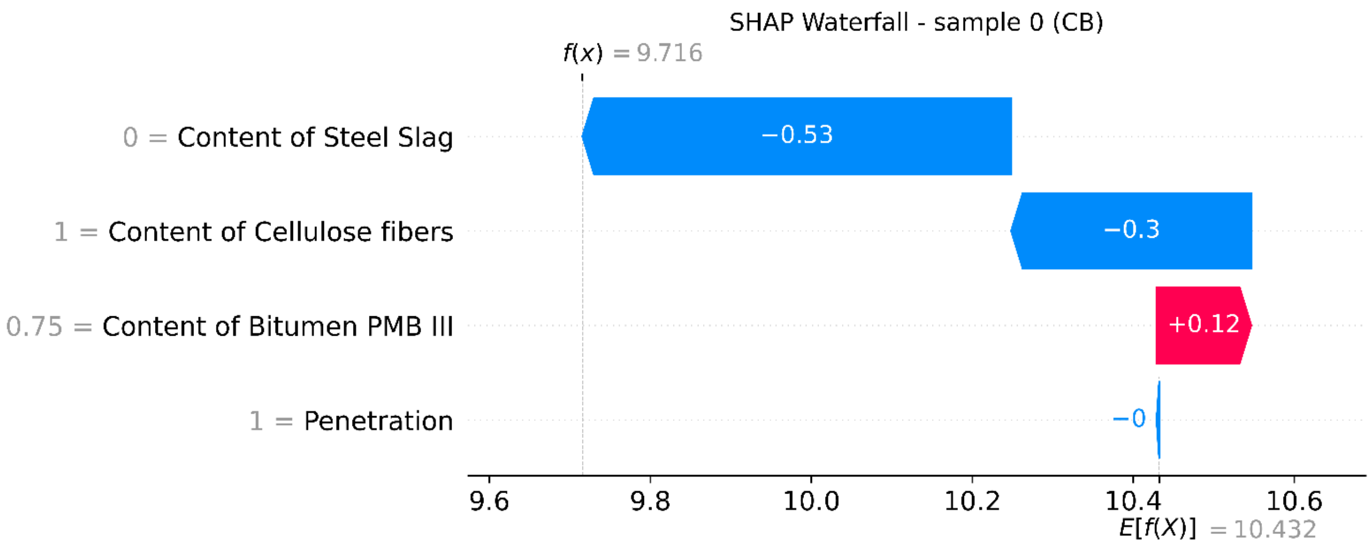


Fig. 9. SHAP waterfall plot for a representative sample predicted by the CatBoost (CB) model

Overall, the comparative evaluation results indicate that the CatBoost (CB) model demonstrates the best predictive performance among the considered machine learning models for estimating Marshall Stability. This finding is consistent with previous studies showing that machine learning models can effectively predict Marshall Stability and Marshall Flow of asphalt mixtures. Gul et al applied supervised machine learning algorithms to predict Marshall Stability and Marshall Flow, while Mistry and Roy used ANFIS to estimate Marshall properties of bituminous mixtures containing waste fillers [36, 37]. In addition, recent studies have shown that tree-based ensemble models and boosting algorithms are highly suitable for asphalt mixture prediction problems. Asi et al reported that CatBoost, LightGBM, XGBoost, and Extra Trees can be used to predict Marshall Stability and Flow, with CatBoost showing strong predictive performance [38]. Similarly, Zahoor et al found that Random Forest and other machine learning models achieved reliable prediction results for Marshall Stability and Flow [39]. Therefore, the high prediction accuracy obtained by the CB model in this study is reasonable and consistent with the findings of previous studies. From a materials engineering perspective, Marshall Stability is governed by the combined and nonlinear effects of binder characteristics, bitumen content, cellulose fiber, aggregate skeleton, and steel slag content. The SHAP results further support this interpretation by identifying steel slag content as the dominant factor contributing to Marshall Stability prediction.

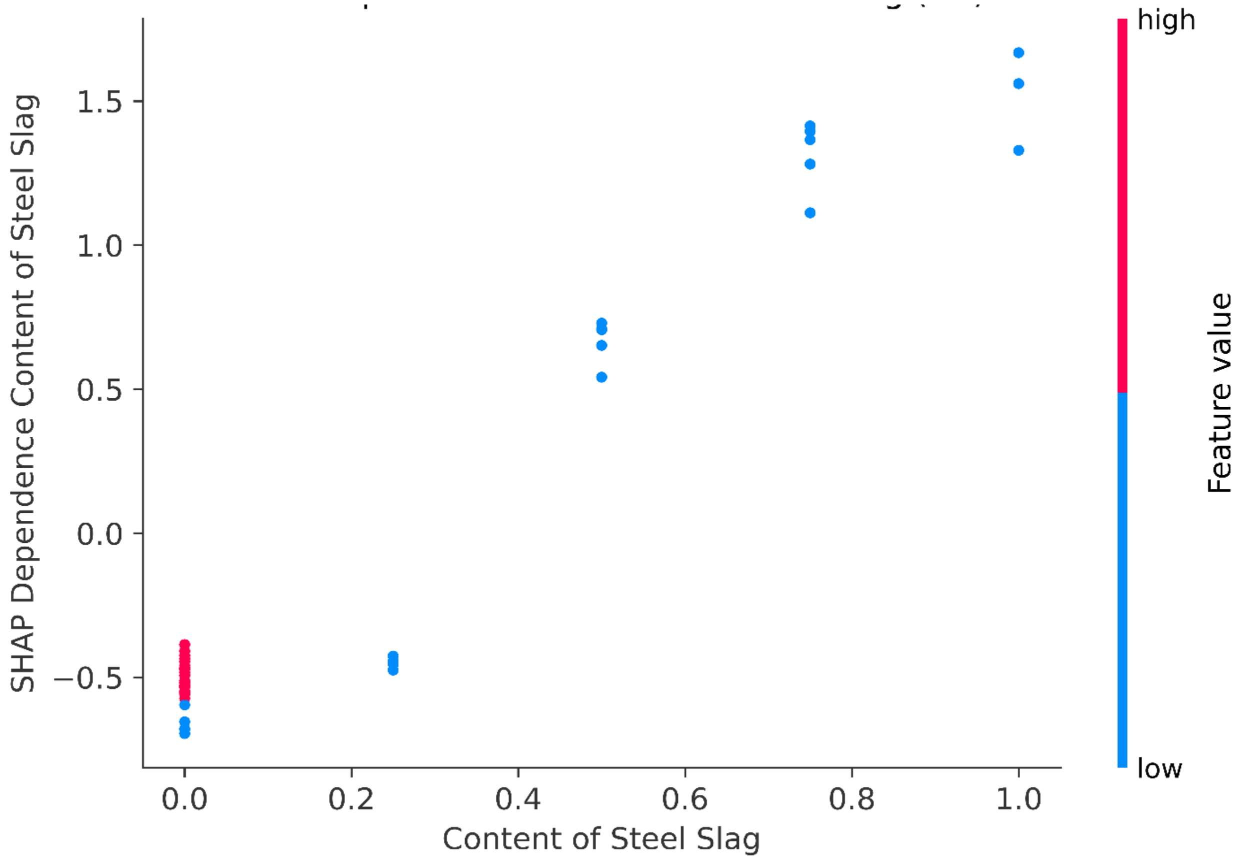


Fig. 10. SHAP dependence plot of steel slag content for the CatBoost (CB) model

5. Conclusions

The results of this study indicate that the application of machine learning models is an effective approach for predicting the Marshall stability of Stone Mastic Asphalt (SMA) mixtures. The final input variables used for model development included penetration, bitumen content, cellulose fiber content, and steel slag content, while Marshall stability was selected as the output variable. Among the evaluated models, CatBoost achieved the best predictive performance, with R^2 = 0.936, RMSE = 0.227, and MAE = 0.323, while also showing good agreement between predicted and observed values, with a concentrated residual distribution and most predictions falling within the ±10% error range. In addition, the model demonstrates the ability to capture complex nonlinear relationships between mixture design parameters and the mechanical response of SMA mixtures, which are influenced by the combined effects of binder content, aggregate structure, and the incorporation of steel slag through nonlinear interaction mechanisms that are difficult to represent using conventional empirical approaches. However, the dataset used in this study remains relatively limited and may not fully reflect the variability of mixture compositions and material properties encountered in practice. Therefore, future research should focus on expanding the dataset to include a wider range of mixture designs and material conditions, thereby improving the reliability and practical applicability of the proposed prediction framework.

References

[1] Y. Yan, W. Li, C. Liu, B. Pan. (2025). Comprehensive review of thermally induced self-healing behavior in asphalt mixtures and the role of steel slag. *Coatings*, 15(6), 668. https://doi.org/10.3390/coatings15060668.

[2]. T. Zhong, Y. Zheng, Z. Chen, L. Yao, W. Zhang, Y. Zhu, L. Fu. (2024). Utilization of steel slag as coarse aggregate and filler in stone mastic asphalt (SMA) mixture: Engineering performance, environmental impact and economic benefits analysis. *Journal of Cleaner Production*, 450, 141891. https://doi.org/10.1016/j.jclepro.2024.141891.

[3]. S. Wang, W. Yu, Y. Miao, L. Wang. (2023). Review on load transfer mechanisms of asphalt mixture meso-structure. *Materials*, 16(3), 1280. https://doi.org/10.3390/ma16031280.

[4]. C. Hettiarachchi, W. Mampearachchi. (2025). Asphalt Mixture Design. In *Handbook of Asphalt Technology*, Springer, 101-153. https://doi.org/10.1007/978-981-96-4532-9.

[5]. J. Zhu, T. Ma, J. Fan, Z. Fang, T. Chen, Y. Zhou. (2020). Experimental study of high modulus asphalt mixture containing reclaimed asphalt pavement. *Journal of Cleaner Production*, 263, 121447. https://doi.org/10.1016/j.jclepro.2020.121447.

[6]. Y.B. Kebede, M.-D. Yang, H.D. Shiku. (2025). Enhancing Asphalt Mix Design with Transfer Learning and Hybrid Artificial Intelligence. *Henok Desalegn, Enhancing Asphalt Mix Design with Transfer Learning and Hybrid Artificial Intelligence*. http://dx.doi.org/10.2139/ssrn.5235675.

[7]. O.J. Adeyeye, I. Akanbi. (2024). Artificial intelligence for systems engineering complexity: a review on the use of AI and machine learning algorithms. *Computer Science & IT Research Journal*, 5(4), 787-808. https://doi.org/10.51594/csitrj.v5i4.1026.

[8]. A. Yazdi, M.H. Dehnad. (2025). A case study on predicting asphalt pavement distress using advanced machine learning techniques and road surface inspection data. *Case Studies in Construction Materials*, e05291. https://doi.org/10.1016/j.cscm.2025.e05291.

[9]. H.K. Sanij, R. Babagoli, R.M. Elyasi. (2025). Enhancing stone matrix asphalt performance with sugarcane bagasse ash: Mechanical properties and machine learning-based predictions using XGBoost and random forest. *Case Studies in Construction Materials*, e05186. https://doi.org/10.1016/j.cscm.2025.e05186.

[10]. Y. Ali, F. Hussain, M. Irfan, A.S. Buller. (2021). An eXtreme Gradient Boosting model for predicting dynamic modulus of asphalt concrete mixtures. *Construction and Building Materials*, 295, 123642. https://doi.org/10.1016/j.conbuildmat.2021.123642.

[11]. X. Fan, S. Lv, C. Xia, D. Ge, C. Liu, W. Lu. (2024). Strength prediction of asphalt mixture under interactive conditions based on BPNN

and SVM. *Case Studies in Construction Materials*, 21, e03489. https://doi.org/10.1016/j.cscm.2024.e03489.

[12]. X. Yu, Z. Zhang, R. Luo, T. Huang, M. Xiao. (2023). Asphalt content prediction model of asphalt mixtures based on dielectric properties. *Materials and Structures*, 56(1), 10. https://doi.org/10.1617/s11527-022-02095-4.

[13]. H.L. Vo, V.P. Le, V.B. Le. (2024). To review the possibility of using steel slag and coconut fiber additives to manufacture stone mastic asphalt (SMA) (in Vietnamese). *Science and Technology*, 740, 28-31.

[14]. G. Gan, A.A. Zhang, K. Long, J. Shang, J. Zhang, C. Ai. (2025). Continuous targeted prediction of asphalt mixture properties based on machine learning. *International Journal of Pavement Engineering*, 26(1), 2515154. https://doi.org/10.1080/10298436.2025.2515154

[15]. M. Dai, F. Zhang, S. Dai, C. Xing, S. Xiao, H. Lv, Y. Tan. (2024). Optimizing asphalt mix design through predicting volumetric properties using interpretable machine learning. *Powder Technology*, 444, 119954. https://doi.org/10.1016/j.powtec.2024.119954

[16]. A.A.o.S.H.a.T.O. (AASHTO). (2012). Standard Specification for Stone Matrix Asphalt (SMA) (AASHTO M 325-08).

[17]. T.H. Le. (2021). Study on the influence of Stone Mastic Asphalt material on rutting resistance and crack resistance of asphalt concrete pavement under Vietnamese conditions (in Vietnamese). Doctoral Dissertation. *Military Technical Academy*.

[18]. A. Natekin, A. Knoll. (2013). Gradient boosting machines, a tutorial. *Front. Neurorobot*, 7, 21. https://doi.org/10.3389/fnbot.2013.00021

[19]. J.H. Friedman. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 29(5). 1189-1232. https://doi.org/10.1214/aos/1013203451

[20]. J.T. Hancock, T.M. Khoshgoftaar. (2020). CatBoost for big data: an interdisciplinary review. *Journal of big data*, 7(1), 94. https://doi.org/10.1186/s40537-020-00369-8

[21]. L. Prokhorenkova, G. Gusev, A. Vorobev, A.V. Dorogush, A. Gulin. (2018). CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31. https://doi.org/10.1186/s40537-020-00369-8

[22]. C. Cortes, V. Vapnik. (1995). Support-vector networks. *Machine learning*, 20, 273-297. https://doi.org/10.1007/BF00994018

[23]. D.A. Pisner, D.M. Schnyer. (2020). Support vector machine. In *Machine learning*, *Elsevier:* 101-121. https://doi.org/10.1016/B978-0-12-815739-8.00006-7

[24]. P. Probst, M.N. Wright, A.L. Boulesteix. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: data mining and knowledge discovery*, 9(3), e1301. https://doi.org/10.1002/widm.1301

[25]. A. Shahraki, M. Abbasi, Ø. Haugen. (2020). Boosting algorithms for network intrusion detection: A comparative evaluation of Real AdaBoost, Gentle AdaBoost and Modest AdaBoost. *Engineering Applications of Artificial Intelligence*, 94, 103770. https://doi.org/10.1016/j.engappai.2020.103770

[26]. A. Beygelzimer, S. Kale, H. Luo. (2015). In Optimal and adaptive algorithms for online boosting. *International Conference on Machine Learning*, PMLR: 2323-2331. https://doi.org/10.48550/arXiv.1502.02651

[27]. D. Chicco, M.J. Warrens, G. Jurman. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *Peerj computer science*, 7, e623. https://doi.org/10.7717/peerj-cs.623

[28]. C.J. Willmott, K. Matsuura. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in

assessing average model performance. *Climate research*, 30(1), 79-82. https://doi.org/10.3354/cr030079

[29]. Y. Wang. (2024). A comparative analysis of model agnostic techniques for explainable artificial intelligence. *Research Reports on Computer Science*. 3(2). 25–33. https://doi.org/10.37256/rrcs.3220244750

[30]. A.M. Salih, Z. Raisi-Estabragh, I.B. Galazzo, P. Radeva, S.E. Petersen, K. Lekadir, G. Menegaz. (2025). A perspective on explainable artificial intelligence methods: SHAP and LIME. *Advanced Intelligent Systems*, 7(1), 2400304. https://doi.org/10.1002/aisy.202400304

[31]. M. Shantal, Z. Othman, A.A. Bakar. (2023). A novel approach for data feature weighting using correlation coefficients and min–max normalization. *Symmetry*, 15(12), 2185. https://doi.org/10.3390/sym15122185

[32]. I. Muraina. (2022). In Ideal dataset splitting ratios in machine learning algorithms: general concerns for data scientists and data analysts. *7th international Mardin Artuklu scientific research conference*, 496-504.

[33]. G.N. Ahmad, H. Fatima, S. Ullah, A.S. Saidi. (2022). Efficient medical diagnosis of human heart diseases using machine learning techniques with and without GridSearchCV. *IEEE Access*, 10, 80151-80173. https://doi.org/10.1109/ACCESS.2022.3165792

[34]. I. Syarif, A. Prugel-Bennett, G. Wills. (2016). SVM parameter optimization using grid search and genetic algorithm to improve classification performance. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, 14(4), 1502-1509. http://doi.org/10.12928/telkomnika.v14i4.3956

[35]. H. Li, G.K. Rajbahadur, D. Lin, C.-P. Bezemer, Z.M. Jiang. (2024). Keeping deep learning models in check: A history-based approach to mitigate overfitting. *IEEE Access*, 12, 70676-70689. http://doi.org/10.1109/ACCESS.2024.3402543

[36]. M.A. Gul, M.K. Islam, H.H. Awan, M. Sohail, A.F. Al Fuhaid, M. Arifuzzaman, H.J. Qureshi. (2022). Prediction of marshall stability and marshall flow of asphalt pavements using supervised machine learning algorithms. *Symmetry*, 14(11), 2324. https://doi.org/10.3390/sym14112324

[37]. R. Mistry, T.K. Roy. (2020). Predicting Marshall stability and flow of bituminous mix containing waste fillers by the adaptive neuro-fuzzy inference system. *Revista de la Construcción*, 19(2), 209-219.

[38]. I. Asi, Y.I. Alhadidi, T.I. Alhadidi. (2024). Predicting Marshall stability and flow parameters in asphalt pavements using explainable machine-learning models. *Transportation Engineering*, 18, 100282. https://doi.org/10.1016/j.treng.2024.100282

[39]. M.F. Zahoor, A. Hussain, A. Khattak. (2025). Machine Learning-Based Prediction Performance Comparison of Marshall Stability and Flow in Asphalt Mixtures. *Infrastructures*, 10(6), 142. https://doi.org/10.3390/infrastructures10060142