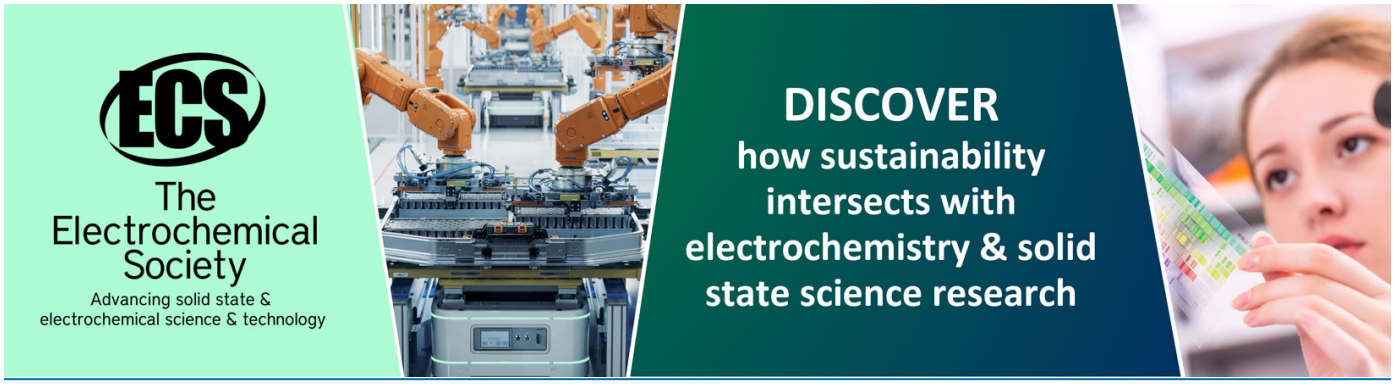
Coarse Aggregate Shape Classification Method Based on Per-Optuna-LightGBM Model



Coarse Aggregate Shape Classification Method Based on Per-Optuna-LightGBM Model

Xin Feng, Zhaoyun Sun*, Zhenzhen Xing, Yulong Wu and Chenyi Lian

School of Information Engineering, Chang'an University, Xi'an, China

chysun@chd.edu.cn

Abstract. To improve the detection level of aggregate shape for automated road use, Per-Optuna-LightGBM model for aggregate shape classification is proposed. Collect aggregate images using industrial camera and extract 48 morphological feature parameters. A feature importance analysis method based on Spearman Correlation and Permutation Importance is proposed to remove redundant factors and select the feature parameters of aggregate morphology. Based on cross-validation, an optimized Optuna-LightGBM model is trained based on the constructed dataset. Compared with GS-XGBoost algorithm, the Optuna-LightGBM model can classify the shape of aggregates more accurately and efficiently. The accuracy value of the proposed model is 82.5%, which increased by 4% compared to before optimization. The proposed model can efficiently classify the shape of aggregates which meet the design requirements, also provide a certain foundation for automated classification of aggregate shapes.

1. Introduction

In recent years, digital image processing technology has been widely applied in the field of aggregate detection, providing a fast and intuitive description of aggregate morphology features. In the research of road-use aggregate morphology features, Hao et al. [1] quantitatively evaluated the angularity of aggregate particles based on 3D point cloud data, and analyzed the angularity of aggregate particles with different particle sizes, rock types, and shapes. Pei et al. [2] and colleagues constructed a neural network model based on multiple feature factors for calculating aggregate particle size, achieving accurate calculation of aggregate particle size. Yang et al. [3] constructed a collection system based on line structure light using 3D structured light point cloud method, collected aggregate point cloud data, and conducted in-depth research on aggregate grading detection.

In terms of particle shape classification, Zhang et al. [4] summarized that the shape of coarse aggregates has a significant impact on the performance of asphalt mixtures, and increasing the angularity of coarse aggregates can enhance the anti-rutting performance of asphalt mixtures. Peng et al. [5] proposed to use shape factor as the evaluation index for single aggregate shape based on the definition of needle-like coarse aggregates, and divided the shape of coarse aggregates into three levels of square, elongated, and needle-like based on the recognition results of shape factors, and studied the influence of different aggregate shape qualities on asphalt mixtures through different experiments. Pei et al. [6] and colleagues used a camera to collect aggregate image data, manually classified aggregate shape into six categories, processed the images with various morphological operations, extracted feature parameters and transformed them into data, and optimized the XGBoost classification model through grid search to achieve six-classification of aggregate shape.

By summarizing the research results of scholars from domestic and abroad on aggregates, it is concluded that the external features of aggregates are closely related to their road performance. Therefore, this paper conducts research on road-use aggregate shape based on machine learning, with the main innovations being:

(1) Using Spearman correlation analysis and Permutation Importance methods to reduce redundant features, select features with high importance for the results, which can improve model efficiency.

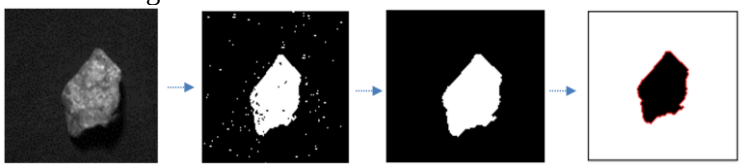
(2) Proposing an Optuna-optimized LightGBM model, adaptively selecting hyperparameter optimization strategies based on the characteristics of the model parameters, quickly finding the optimal parameters within the given range, while maintaining high accuracy and improving model efficiency.

2. Material

The study discuss the shape characteristics of the aggregates. There are 1624 aggregates in this data sample. According to the classification method Al-Batah et al. [7], the aggregates were classified into four categories: angular, cubical, flake elongated and irregular.

2.1. Image set processing

The aggregate image acquisition system consists of a light source, a camera, a lens and a background plate. Firstly, the aggregates are manually classified into four categories of shapes and labeled with the categories. Secondly, based on the requirements of aggregate image acquisition, the design of the aggregate image acquisition system is carried out, including the selection of supplementary light sources, cameras and lenses, and background plates. The built aggregate image acquisition system is used to capture images of the aggregates and build the aggregate image dataset. Finally, the aggregate images in the dataset are processed, including the processes of smoothing, sharpening, denoising, binarization, and morphology, as shown in Figure 1.



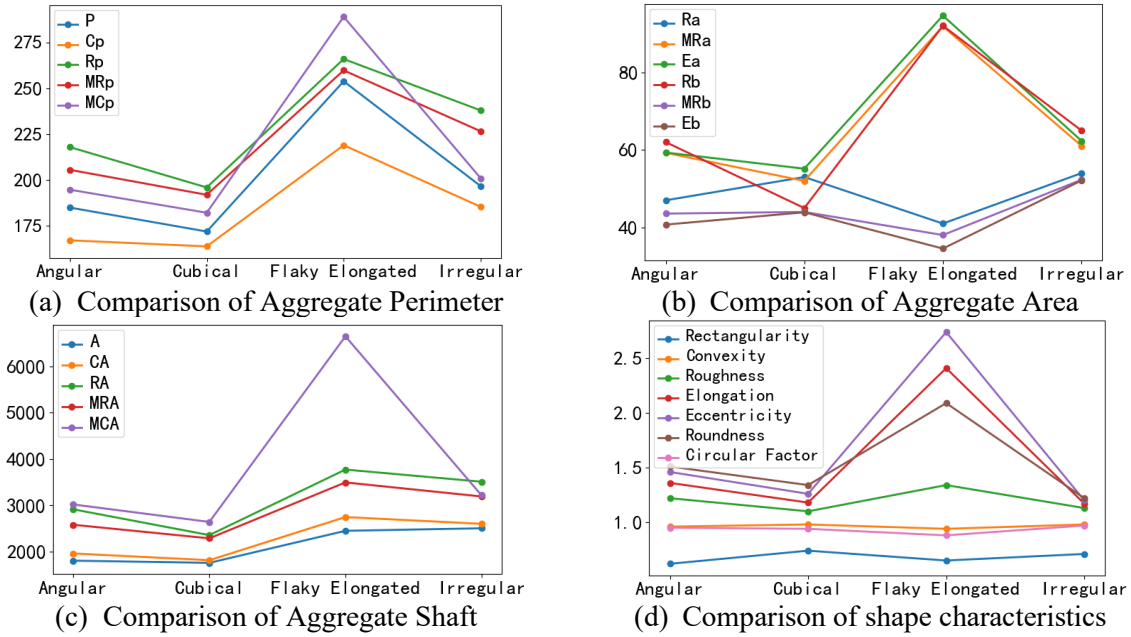
(a) Original Image (b) Binarization (c) Denoising (d) Edge Detection.

Figure 1. Aggregate image processing

2.2. Aggregate morphology feature parameter extraction

Relevant feature parameters are extracted from the acquired 2D images to represent various shapes of aggregates, and the shape of the aggregates is described by the feature parameters. 48 features are extracted in total, including parameters such as perimeter, internal and external circles, internal and external rectangles, and fitted circles. The unit measures of different aggregate features are different, and they are divided into four groups in order to better observe the distribution of aggregate feature parameters. The perimeter characteristics are compared as a group, the area characteristics are compared as a group, and the length and width of the long and short axes as well as the diameter are compared as a group. The following four line graphs are drawn according to the data in the table, which can be more intuitive to observe the different aggregates under the same characteristic latitude, as shown in Figure 2.

The Figure 2 show the comparison of perimeter, area, shaft (long and short axes) and shape characteristics, and it can be seen that the indicators of flaky elongated are significantly higher than the other three types, fully illustrating their long shape characteristics. The angular aggregates and cubical are smaller, and the long and short axes vary slightly. Overall, most of the indicators vary significantly with the shape of the aggregate, but the area of the aggregate A, the area of the convex pack C_A, the area of the outer rectangle $R_A$and the area of the minimum outer rectangle MR_Avary less with the shape. Most of the shape features show more significant differences depending on the shape of the aggregate, but the $Rectangularity$, $Convexity$, and $Circular\ Factor$ are not significant. In view of this, correlation and importance analysis is required for the features.



(a) Comparison of Aggregate Perimeter (b) Comparison of Aggregate Area

(c) Comparison of Aggregate Shaft (d) Comparison of shape characteristics

Figure 2. Comparison of the shape characteristics of different aggregates

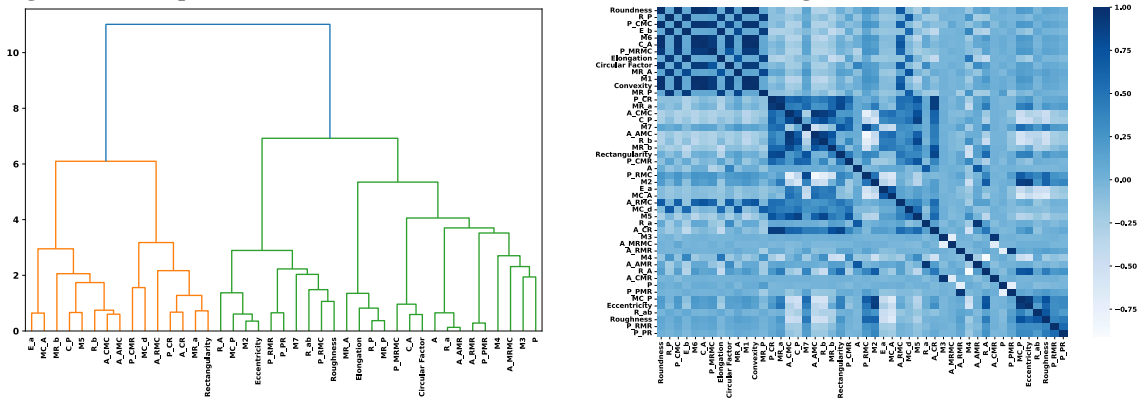
3. Methodology

3.1. Data pre-processing

The data characterization work starts with feature coding [8]. The aggregate shapes were classified into four categories, namely cubical, angular, Flaky Elongated, and irregular, which were binary named as 00, 01, 10, and 11 to be fed into the model. After that, the data are normalized by scaling the feature parameters so that they fall into the same dimensional space to ensure that all the feature data are in the same unit magnitude and are not affected by the difference in units.

3.2. Feature importance analysis

In the feature importance analysis process, for the co-linear features, we investigated the correlation between the features using the Spearman correlation coefficient method [9]. Hierarchical clustering was performed using Spearman rank correlation with a threshold of 0.95. One feature was retained from each cluster. We selected the feature parameters highly correlated with Roundness, P_CMC, Convexity, M6, M1, E_b, and A_CMR, and performed an importance analysis on the remaining 41 feature parameters using the feature permutation method. The results are shown in the Figure 3.



(a) Hierarchical clustering diagram (b) Correlation Heat Map

Figure 3. Spearman correlation analysis

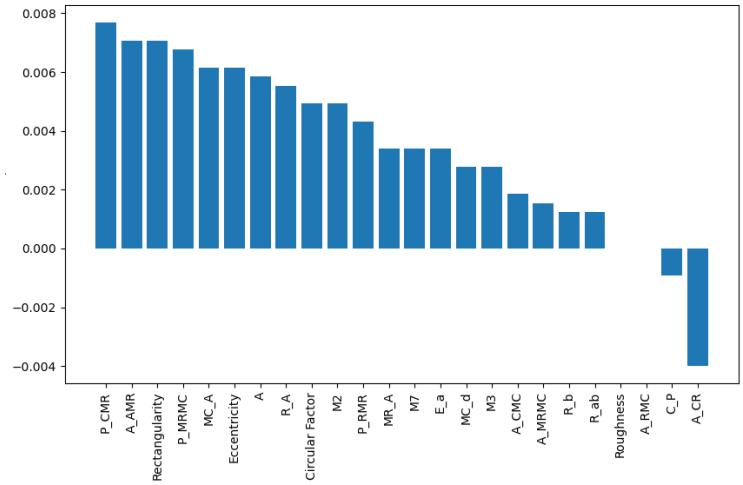


Figure 4. Permutation-based aggregate shape characterization

Based on permutation to select features [10], which measures the contribution of a feature to the model by observing changes in model performance when a feature value is randomly permuted. By evaluating and ranking the importance of the features, we obtained the importance rankings of some features, where a value greater than zero indicates that the inclusion of the feature can improve model performance, while a value less than or equal to zero indicates that the inclusion of the feature has little or even negative impact on accuracy. Figure 4 shows the importance rankings of some of the features. To consider the impact of redundant features on model training, we selected 20 features highly correlated with shape, such as P_CMR, A_AMR, E_a, and Eccentricity, from the features that had a significant impact on aggregate shape as shown in the figure.

3.3. Light Gradient Boosting Machine

LightGBM is a gradient boosting framework based on decision trees [11], which can efficiently handle large-scale datasets and high-dimensional features, and has strong stability and generalization ability. Its main principle is to improve the predictive accuracy of the model by integrating multiple decision trees. The algorithm constructs multiple regression trees to form a tree ensemble and makes the predicted values of the tree ensemble close to the true values to improve the predictive accuracy of the model. When constructing the t[th] tree of the model, its predicted value can be represented by formula (1).

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} (f_k(x_i)) = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{1}$$

The symbol f_k represents the k[th] tree, $f_k(x_i)$ is the output score of the k[th] tree for input x_i, and $\hat{y}_i^{(t)}$ is the prediction result of the t-tree ensemble model for sample x_i.

One of the advantages of LightGBM is the histogram-based sorting method. Figure 5 shows the process of histogram sorting. This method divides continuous values into discrete intervals, namely data binning.

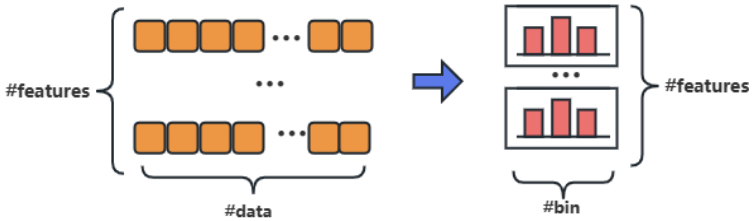


Figure 5. Histogram optimization

3.4. Optuna-LightGBM model construction

Optuna [12] is an automatic hyperparameter tuning framework based on various search strategy optimization algorithms, including random search, grid search [13], and Bayesian optimization [14]. It can help us find the best hyperparameter combination in a shorter time, thereby improving the performance and stability of the model. Compared to other hyperparameter tuning methods, Optuna has higher efficiency and scalability and can automatically handle both discrete and continuous hyperparameters.

A total of 400 sets of data were selected from each category, with about 75% of the data were selected as the training set and 25% as the test set. In the training set, Optuna was used to optimize the LightGBM model parameters. Five-fold cross-validation was used, with 80% of the data set for training and 20% for validation, using accuracy as the fitness function for 100 rounds of training to find the optimal solution, and finally save the model. The model construction process is shown in Figure 6.

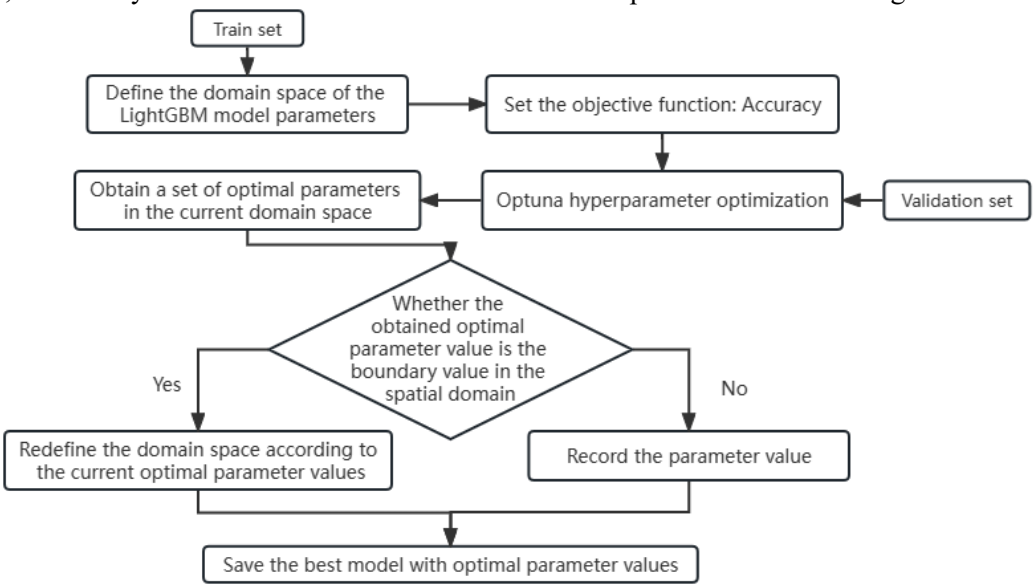
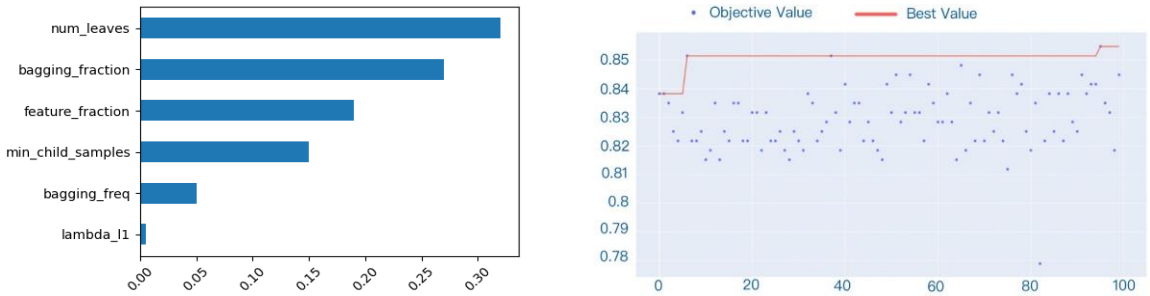


Figure 6. Optuna-LightGBM process

The data was learned and visualized using the Optuna-LightGBM model. The Figure 7 shows the importance of each parameter, with important parameters such as num_leaves and bagging_fraction given higher weights during optimization. After 100 iterations of model training, the average accuracy on the validation set was about 0.83, with a maximum of 0.85. After 100 rounds of training, the best model and its parameters have been saved and can be used for prediction on a new test set.



(a) Parameter importance ranking (b) Model parameter optimization process

Figure 7. Optuna-LightGBM training process

4. Experiments and results

4.1. Model Evaluation

To evaluate and analyze the model, precision, accuracy, recall, and F1 [15] score are introduced as evaluation metrics, using a test set consisting of 400 samples, with 100 samples for each category. The confusion matrix is also introduced as a metric, with each row representing the number of samples predicted by the model to have a certain shape type, and each column representing the actual number of samples with that shape type. In addition, macro avg and weighted avg are introduced as comprehensive metrics that provide more information about the model's performance across different categories. These metrics are particularly useful for imbalanced data.

4.2. Results of Per-Optuna-LightGBM classification mode

In this study, the Per-Optuna-LightGBM model was used for predictive classification of aggregate shapes. The results presented confusion matrices with colour shades reflecting the number of aggregates as Figure 8 is shown.

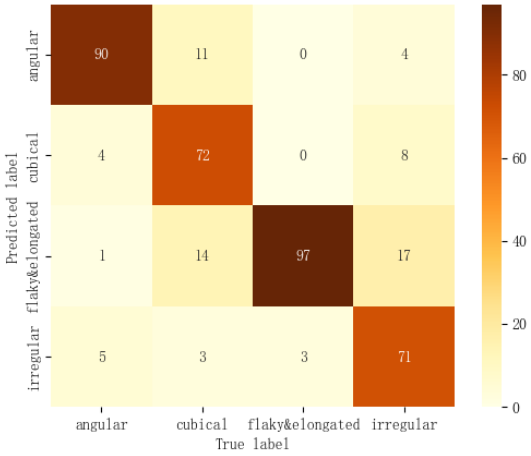


Figure 8. Confusion matrix of Per-Optuna-LightGBM results.

As is shown in Figure 8, the diagonal part of the presents the darkest colour, indicating a good effect of the four categories. Out of the 100 aggregates with actual angular shapes, 90 were correctly predicted. Out of the 100 aggregates with actual cubic shapes, 72 were correctly predicted, which is relatively average and may be caused by less significant changes in parameters, such as the number of cube edges and outer circles. Among the 100 aggregates with actual shape of flaky elongated, 97 were correctly predicted, which was attributed to its shape with obvious difference of large parameters. Among the 100 aggregates with actual irregular shape, 71 were correctly predicted, and some were misclassified into angular and flaky elongated, which indicates the difficulty of irregular shape.

According to the confusion matrix of the Per-Optuna-LightGBM model, the precision, recall, and F1-score of the four aggregates are listed in Table 1.

Table 1. Per-Optuna-LightGBM model aggregate shape classification

Typical shapes	Precision	Recall	F1-score	Support
Cubical	0.86	0.90	0.88	100
Angular	0.86	0.72	0.78	100
Flaky Elongated	0.75	0.97	0.85	100
Irregular	0.87	0.71	0.78	100
macro avg	0.83	0.82	0.82	400
weighted avg	0.83	0.82	0.82	400
accuracy			0.825	

From the Table 1, it can be observed that: the best classification results were obtained for flake elongated. The classification results were better for angular and cubical, and worse for irregular shapes, with each index lower than the former. And the Per-Optuna-LightGBM-based aggregate shape classification model reached above 0.8, and the overall accuracy of the model reached 0.825, which would represent that the predicted category of the model for aggregate shape is basically consistent with the real category.

4.3. Comparison with other models

To verify the validity of the model proposed in this paper, Per-Optuna-LightGBM Model was compared with LightGBM, Optuna-LightGBM and GS-XGBoost. Accuracy and model runtime were used as model evaluation metrics, and the results of the four model classifications are shown in Figure 9.

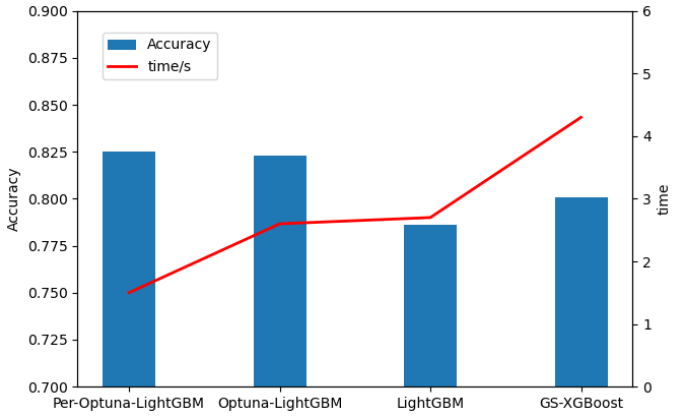


Figure 9. Performance comparison of different models

It can be seen from Figure 9 that the Per-Optuna-LightGBM model outperforms the other models in terms of Accuracy and running time, and the index reaches more than 0.825. The Optuna-LightGBM model is comparable to the former, but the running time doubles without the effect of Permutation feature screening, which indicates that the work of feature importance analysis can improve the efficiency of aggregate shape detection while maintaining accuracy. The Optuna-based powerful hyperparameter search capability makes the LightGBM model 4% more accurate and more efficient than the grid-search XGBoost model, thanks to the histogram optimization approach of LightGBM. XGBoost uses a level-wise hierarchical splitting strategy, which splits all nodes at each level of the tree This increases the computation time of the XBGoost model without considering the splitting gain of the nodes. In summary, it can be concluded that the final optimized Per-Optuna-LightGBM model can efficiently classify and identify the aggregate shapes.

5. Conclusion

The study takes road aggregate shape as the research object, which based on feature importance, hyperparameter training, and machine learning model to carry out research, the main conclusions are as follows:

(1) Based on Spearman correlation analysis and Permutation Importance for evaluation, the top 20 important features are selected to effectively improve the learning ability of the model and reduce the running time, and the reference standard for aggregate classification is also given.

(2) The Per-Optuna-LightGBM model is proposed and compared with LightGBM model, Optuna-LightGBM model and GS-XGBoost, respectively, and the comparison shows that the Per-Optuna-LightGBM model outperforms the other models in terms of accuracy and efficiency.

The proposed model effectively achieves the classification of aggregate shape, and can quickly detect aggregate shape, which is important for the development of road construction and transportation.

6. References

[1] Hao Xueli, Sun Chaoyun, Geng Fangyuan, Li Wei, Pei Lili, and Zhang Xin. Quantitative Evaluation of Aggregate Particle Angularity Based on Three-dimensional Point Cloud Data [J]. Journal of South China University of Technology (Natural Science Edition), 2021, 49(01): 142-152.

[2] Lili Pei, Ting Yu, Ruichi Ma, Wei Li and Xueli Hao. Automatic Classification of Coarse Aggregate Particle Size Based on Light Gradient Boost Machine. 7th Annual International Conference on Material Engineering and Application. (ICMEA 2020),2020.

[3] Yang Ming,Ding Jiangang,Li Wei,Tian Aojia,Pei Lili,Hao Xueli. A coarse aggregate gradation detection method based on 3D point cloud[J]. Construction and Building Materials,2023,377.

[4] Zhang Dong, Hou Shuguang, and Bian Jiang. Research Status of the Influence of Coarse Aggregate Morphology on the Performance of Asphalt Mixture. Journal of Nanjing Tech University (Natural Science Edition), 2017, 39(06): 149-154.

[5] Peng Yuming. Rapid Detection Technology of Aggregate Shape and Research on the Influence of Shape on Mixture Performance [D]. Chongqing Jiaotong University, 2022. DOI: 10.27671/d.cnki.gcjtc.2022.000290.

[6] Lili Pei,Zhaoyun Sun,Ting Yu,Wei Li,Xueli Hao,Yuanjiao Hu,Chunmei Yang. Pavement aggregate shape classification based on extreme gradient boosting[J]. Construction and Building Materials,2020,256(C).

[7] Al-Batah M S, Isa N A M, Zamli K Z, et al. A novel aggregate classification technique using moment invariants and cascaded multilayered perceptron network[J]. International Journal of Mineral Processing, 2009, 92(1-2): 92-102

[8] Saraf Tara Othman Qadir,Fuad N.,Taujuddin N. S. A. M.. Feature Encoding and Selection for Iris Recognition Based on Variable Length Black Hole Optimization[J]. Computers,2022,11(9).

[9] Essam F. El Hashash,Raga Hassan Ali Shiekh. A Comparison of the Pearson, Spearman Rank and Kendall Tau Correlation Coefficients Using Quantitative Variables[J]. Asian Journal of Probability and Statistics,2022.

[10] Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. Bioinformatics, 26(10), 1340-1347.

[11] Wu Shanshan,Zheng Haifeng,Chen Chi,Zhang Kun. Research On The Daily Electricity Forecast Model Based On LightGBM[J]. Journal of Physics: Conference Series,2023,2477(1).

[12] Srinivas Polipireddy,Katarya Rahul. hyOPTXg: OPTUNA hyper-parameter optimization framework for predicting cardiovascular disease using XGBoost[J]. Biomedical Signal Processing and Control,2022,73.

[13] Belete Daniel Mesafint,Huchaiah Manjaiah D.. Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results[J]. International Journal of Computers and Applications,2022,44(9).

[14] Chen Yifang,Li Feng,Zhou Siqi,Zhang Xiao,Zhang Song,Zhang Qiang,Su Yijie. Bayesian optimization based random forest and extreme gradient boosting for the pavement density prediction in GPR detection[J]. Construction and Building Materials,2023,387.

[15] Kamal, M. S., & Sangaiah, A. K. (2020). Evaluation Metrics for Machine Learning: A Survey. IEEE Access, 8, 108952-108978. doi: 10.1109/ACCESS.2020.3004649.