*International Journal of Research in Engineering and Science (IJRES)*
*ISSN (Online): 2320-9364, ISSN (Print): 2320-9356*
*www.ijres.org Volume 14 Issue 7 ǁ July 2026 ǁ PP. 155-161*

# Research on Prediction of ESG Performance and Corporate Financial Performance Based on Stacking Ensemble Learning

Zhongchen Wang[a], Z h u Y uanyuan[b], Fengling Zhao[c], Xvsheng Wu[a]
[a]*Shenzhen Campus, Jinan University, Shenzhen, Guangdong, China. 518053*
[b]*ShenzhenTourism College, Jinan University, Guangzhou, Guangdong, China. 518053*
[c]*School of Management, Jinan University, Guangzhou , Guangdong, China. 518053*

***ABSTRACT***
*With the increasing global focus on sustainable development, Environmental, Social, and Governance (ESG) metrics have become a crucial dimension for assessing the long-term value of enterprises. However, a complex nonlinear relationship exists between ESG factors and corporate financial performance, which traditional linear regression models struggle to accurately capture. This paper proposes a financial performance prediction frame-work based on Stacking ensemble learning, utilizing a dataset covering 1,000 companies across 9 industries and 7 regions globally from 2015 to 2025. First, the intrinsic connections between ESG metrics and financial indicators are analyzed via correlation heatmaps. Second, base learners including CatBoost, Random Forest, LightGBM, and XGBoost are constructed, and Ridge Regression is employed as the meta-learner for Stacking integration. Finally, the interpretability of the model is analyzed using the SHAP (SHapley Additive exPlanations) method. Experimental results demonstrate that the Stacking model outperforms individual models in terms of $R^2$ score, MAE, and RMSE, exhibiting stronger generalization ability. SHAP analysis further reveals that carbon emis-sions and the ESG composite score are critical factors influencing financial prediction. This study provides an effective tool for investors and corporate managers to leverage machine learning for ESG evaluation.*

***Keywords:*** *ESG, Financial Performance Prediction, Ensemble Learning, Stacking, SHAP Interpretability*

---

Date of Submission: 14-07-2026 Date of Acceptance: 27-07-2026

---

## I. INTRODUCTION

In recent years, the philosophy of ESG investment has evolved from a niche concept to a mainstream strategy, serving as a significant orientation for global capital markets. Numerous studies have attempted to investigate the relationship between ESG performance and corporate financial performance, yet the conclusions remain controversial. Some scholars argue that ESG investments increase operational costs and may impair financial performance in the short term. Conversely, another school of thought posits that robust ESG practices can mitigate risks and enhance corporate reputation, thereby generating long-term excess returns.[1, 2]

When addressing such issues, traditional econometric methods often assume linear relationships among vari-ables and struggle to effectively handle data characteristics involving high dimensionality, multiple sources, and complex interactions. With the advancement of artificial intelligence technology, machine learning, leveraging its powerful non-linear fitting capabilities, has provided a novel perspective for financial forecasting.[3] In particular, algorithms based on Gradient Boosting Decision Trees (GBDT), such as XGBoost, LightGBM, and CatBoost, have demonstrated superior performance in processing structured data.

To further improve prediction accuracy and mitigate the risk of overfitting associated with single models, this paper proposes an ensemble learning model based on the Stacking strategy. The main contributions of this paper are as follows:

1. A prediction model for ESG and financial performance is constructed utilizing a global multi-industry dataset covering the period from 2015 to 2025.

Further author information: (Send correspondence to Zhu
Yuanyuan)Z. Wang: E-mail: 481769790@qq.com

2. The performance of CatBoost, Random Forest, LightGBM, XGBoost, and the Stacking ensemble model is systematically compared.
3. SHAP values are introduced to perform attribution analysis on the black-box models, quantifying the marginal contributions of specific ESG factors to financial performance, thereby enhancing the interpretabil-ity and credibility of the model.

## II. DATA AND PRERROCESSING

### 2.1 Data Source and Description

The dataset provides operational data for 1,000 companies globally spanning the period from 2015 to 2025. It covers 9 major industries, including technology, healthcare, and finance, across 7 geographical regions such as North America, Europe, and Asia-Pacific. The dataset includes the following categories of key features:

- **Financial Metrics**: Revenue, assets, profit margins, market capitalization, etc.
- **ESG Metrics**: Carbon emissions, energy consumption, workforce diversity, governance scores, and com-posite ESG scores.
- **Other Attributes**: Year, country, and industry classification.

### 2.2 Data Preprocessing

To ensure the quality of model training, the following preprocessing steps were performed on the raw data:

1. **Missing Value Treatment**: Missing values in the dataset were detected. Numerical variables were imputed using the mean, while categorical variables were imputed using the mode.
2. **Feature Encoding**: Given that the dataset contains categorical variables such as industry and region, One-Hot Encoding and Label Encoding were applied to facilitate model recognition.
3. **Data Standardization**: To eliminate the influence of different feature scales, StandardScaler was em-ployed to standardize numerical features, resulting in a mean of 0 and a variance of 1.
4. **Feature Correlation Analysis**: To preliminarily understand the linear relationships between features and assist in feature selection, a feature correlation heatmap was plotted, as shown in Figure 1.

The intensity of the colors in the figure indicates the strength of the correlations. Darker red signifies a more significant positive correlation, while blue represents a negative correlation. By analyzing this figure, the linear relationships between features and potential data characteristics can be clearly identified.

First, the target variable Revenue exhibits an extremely strong positive correlation with MarketCap, with a correlation coefficient reaching 0.84. This aligns with economic intuition, suggesting that the scale of enterprise market capitalization is often a robust indicator of revenue. Furthermore, Revenue presents a moderate positive correlation with both CarbonEmissions and WaterUsage, implying that the expansion of corporate production scale is typically accompanied by higher resource consumption.

It is worth noting that extremely high collinearity exists among certain environmental metrics. For instance, the correlation coefficient between EnergyConsumption and CarbonEmissions approaches 1.00, and the coefficient between WaterUsage and EnergyConsumption is as high as 0.97. This significant multicollinearity indicates that feature selection or regularization techniques may be required in the subsequent modeling process to avoid instability in model parameter estimation and to enhance computational efficiency. Concurrently, the correlation between Size ESG Interaction and MarketCap reaches 0.98, indicating that this interaction feature is primarily dominated by the factor of enterprise size.

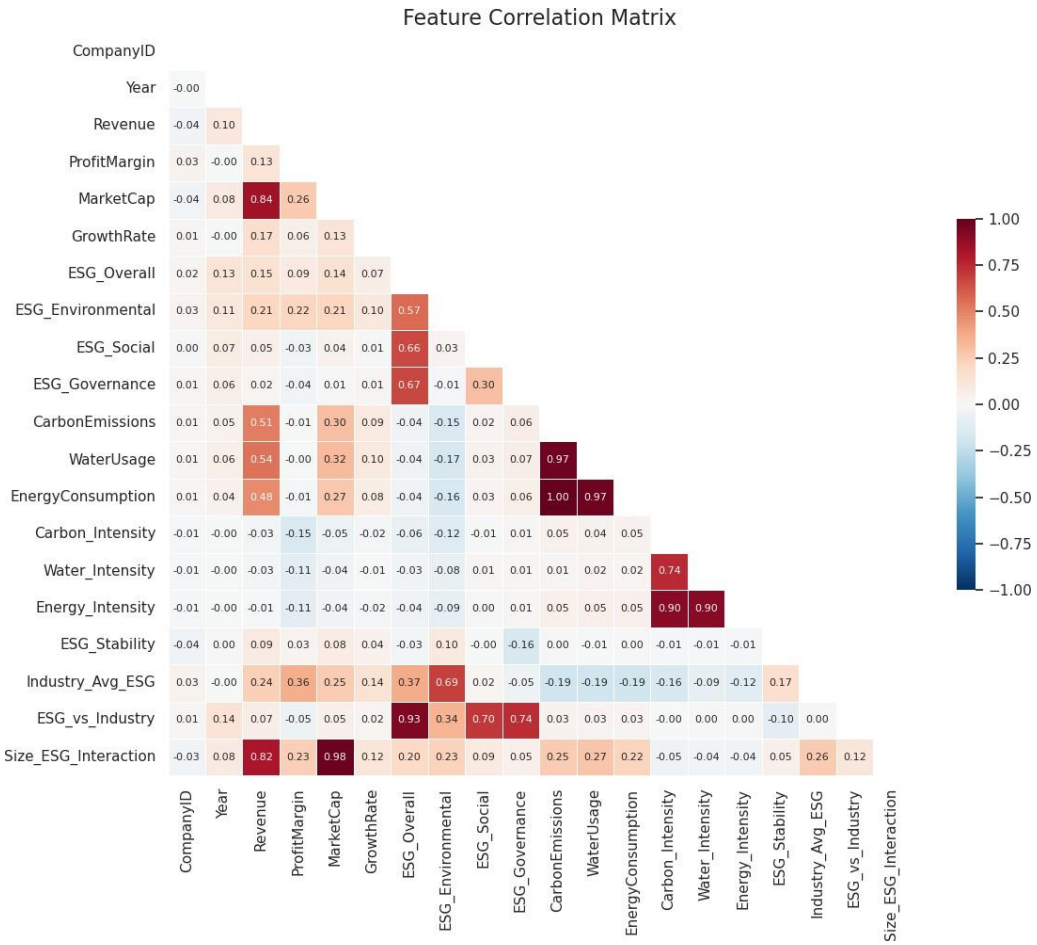


Figure 1. Feature Correlation Heatmap

## III. MODEL CONSTRUCTION

### 3.1 Base Model Selection

This study selects four machine learning algorithms widely used in industry as base models:

- **Random Forest (RF)**: Based on the Bagging methodology, RF reduces variance by constructing multiple decision trees and averaging their outputs, exhibiting good robustness to noise.[4]
- **XGBoost**: An efficient implementation of Gradient Boosting Decision Trees (GBDT), which introduces regularization terms to control overfitting and supports parallel computing.[5]
- **LightGBM**: Based on histogram algorithms and Gradient-based One-Side Sampling (GOSS), LightGBM demonstrates extremely fast training speeds when processing large-scale data.[6]
- **CatBoost**: Specifically optimized for categorical features, CatBoost employs the Ordered Boosting method to effectively address the prediction shift problem.[7]

### 3.2 Stacking Ensemble Strategy

To combine the strengths of different models, this paper adopts the Stacking ensemble strategy.[8] The overall architecture comprises base models and a meta-model, structured as follows:

- **Layer 1**: The training set is divided into 5 folds for cross-validation to train the four aforementioned base models respectively. The prediction results of each base model on the validation sets are concatenated to serve as input features for the second-layer model.
- **Layer 2**: Ridge Regression is employed as the meta-model. By introducing an L2 regularization term, Ridge Regression effectively handles potential multicollinearity among the prediction values of the base models, thereby outputting the final prediction results.

## IV. EXPERIMENTAL RESULTS AND ANALYSIS

### 4.1 Experimental Settings and Evaluation Metrics

The dataset was randomly split into a training set and a testing set at a ratio of 8:2. To comprehensively evaluate the performance of the regression models, this paper selects the following three metrics:

### 4.2 Evaluation Metrics Definitions

Three metrics are used to evaluate the regression model performance, with their mathematical definitions as follows:

#### 4.2.1 Coefficient of Determination ($R^2$ Score)

Measures the extent to which the model explains the variance in the data.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{1}$$

where $y_i$ represents the actual value of the $i$-th sample, $\hat{y}_i$ represents the predicted value of the $i$-th sample, $\bar{y}$ represents the mean of all actual values, and $n$ is the total number of samples.

#### 4.2.2 Mean Absolute Error (MAE)

The average of the absolute differences between the predicted values and the actual values.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \tag{2}$$

where $y_i$ represents the actual value of the $i$-th sample, $\hat{y}_i$ represents the predicted value of the $i$-th sample, and $n$ is the total number of samples.

#### 4.2.3 Root Mean Square Error (RMSE)

The square root of the mean of squared errors between predicted and actual values, which is sensitive to outliers.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{3}$$

where $y_i$ represents the actual value of the $i$-th sample, $\hat{y}_i$ represents the predicted value of the $i$-th sample, and $n$ is the total number of samples.

Table 1. Performance Comparison of Different Models

| Model | $R^2$ score | MAE | RMSE |
|---|---|---|---|
| CatBoost | 0.9237 | 0.729 | 3.254 |
| Random Forest | 0.9028 | 0.828 | 3.672 |
| LightGBM | 0.8988 | 0.818 | 3.747 |
| XGBoost | 0.9041 | 0.710 | 3.648 |
| Stacking | 0.9381 | 0.504 | 2.930 |

### 4.3 Performance Comparison of Different Models

Table 1 presents the performance of the five models in the prediction task. We performed a comparative analysis based on three key evaluation metrics: $R^2$, MAE, and RMSE.

It can be clearly observed from Table 1 that the Stacking ensemble model demonstrates significant advantages across all evaluation metrics. Specifically, the Stacking model achieved the highest $R^2$ score of 0.9381, indicating the strongest ability to explain data variance and the best goodness of fit. Meanwhile, regarding error metrics, the Stacking model achieved the lowest MAE of 0.504 and RMSE of 2.930, significantly outperforming other individual models. In comparison, the CatBoost model ranks second, with an $R^2$ of 0.9237 and MAE/RMSE of 0.729 and 3.254 respectively, demonstrating good robustness. Random Forest, LightGBM, and XGBoost performed relatively similarly, slightly trailing the former two. Overall, the experimental results robustly prove that the Stacking strategy effectively improves prediction accuracy and generalization ability by integrating the strengths of multiple base models.

### 4.4 Visualization of Prediction Results of the Stacking Model

To intuitively demonstrate the prediction capability of the Stacking model, we selected a subset of samples from the test set and plotted a comparison graph of predicted versus actual values, as shown in Fig. 2.

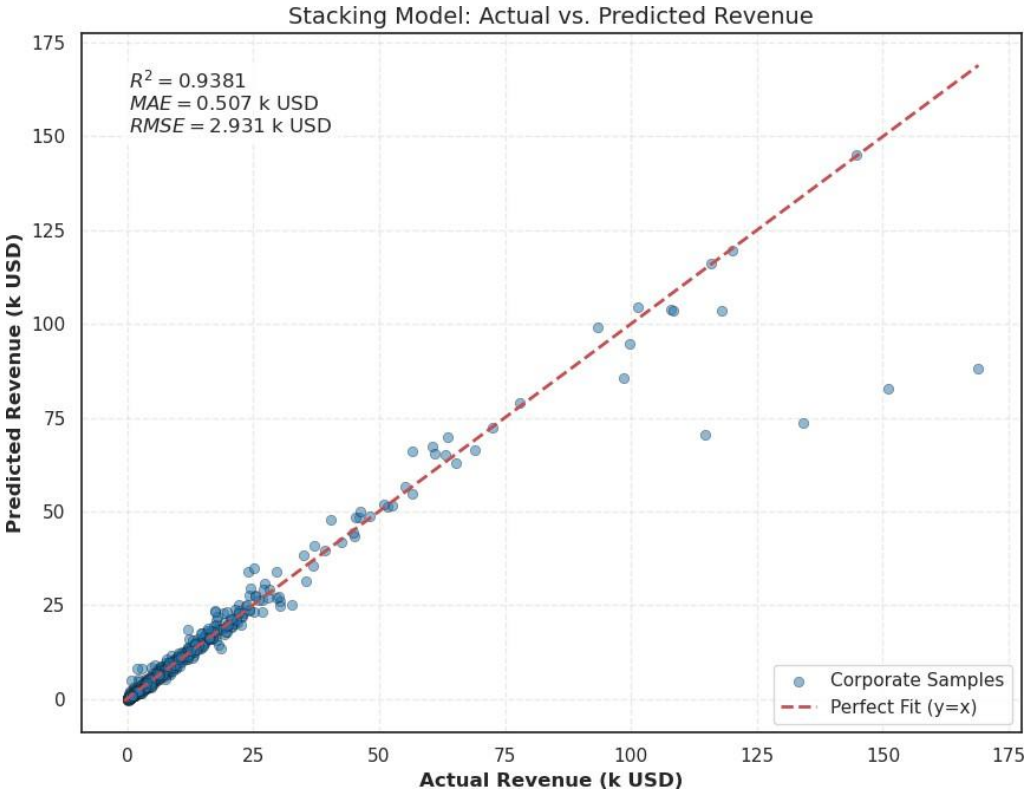


Figure 2. Comparison of Predicted and Actual Values of the Stacking Model

In the figure, the x-axis represents the actual corporate revenue, and the y-axis represents the predicted revenue output by the Stacking model. The red dashed line represents the perfect prediction line in an ideal state, where the predicted value exactly equals the actual value. It can be observed that the vast majority of sample points are tightly distributed around the diagonal red line. This high degree of clustering indicates that the model possesses extremely high prediction accuracy and fitting capability. Specifically, the model achieved an excellent $R^2$ of 0.9381, while controlling the MAE at 0.507 k USD and the RMSE at only 2.931 k USD. Although there are a few outliers in the higher revenue range, these generally belong to extreme cases, and the overall trend maintains good consistency. This further verifies the effectiveness of the Stacking ensemble strategy in capturing data characteristics and performing accurate regression predictions.

## V. MODEL INTERPRETABILITY

While ensemble learning enhances prediction accuracy, it often compromises model interpretability. To address this issue, this study employs the SHAP (SHapley Additive exPlanations) method.[9] Rooted in game theory, SHAP values quantify the marginal contribution of each feature to the model's final prediction. As shown in Figure 3.

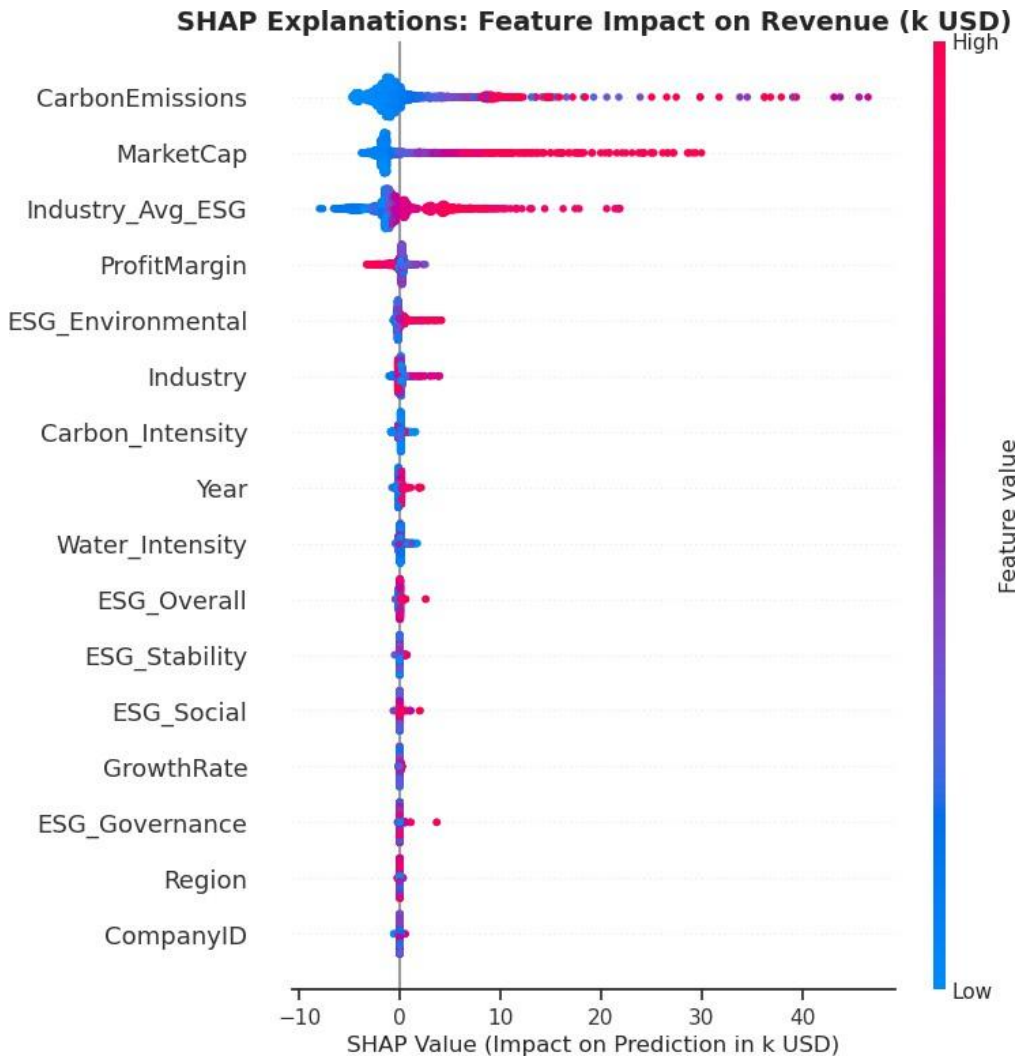


Figure 3. SHAP Value Analysis of the Stacking Model

As observed from Figure 3, *CarbonEmissions* emerges as the most critical feature influencing corporate revenue prediction. Its SHAP value distribution exhibits the widest range (indicating a large span on the horizontal axis), signifying a decisive impact on the prediction results. The distribution of red and blue colors reveals specific directional impacts: red dots are predominantly distributed in the positive SHAP value region, while blue dots (representing low feature values) are concentrated in the negative region. This implies a positive correlation between higher *CarbonEmissions* and higher predicted revenue, which may be attributed to the high-output characteristics of specific sectors, such as energy or heavy industry.

Following closely are *MarketCap* and *IndustryAverageESGScore*. *MarketCap* similarly demonstrates a distinct positive influence; that is, a higher market capitalization corresponds to higher predicted revenue, aligning with economic intuition. Interestingly, *IndustryAverageESGScore* presents a complex non-linear relationship. Although its overall impact is significant, the distribution of high and low values is relatively intertwined, suggesting heterogeneity in the impact of the industry environment on individual corporate revenue.

In contrast, features such as *ProfitMargin* and *EnvironmentalScore*, while contributing to some extent, have SHAP values concentrated near the zero point, indicating relatively minor marginal contributions to the model prediction. Furthermore, features like *CompanyID* and *Region* rank at the bottom, suggesting that within the decision logic of this model, mere identity markers or geographical locations are not the core drivers determining revenue. Through SHAP analysis, we not only verified that the model's decision basis conforms to business logic but also uncovered the potential significance of environmental metrics in revenue forecasting.

## VI.CONCLUSION

Based on a global corporate dataset spanning from 2015 to 2025, this paper explores the application of machine learning techniques to predict the impact of ESG factors on corporate financial performance. Experimental results demonstrate that the Stacking ensemble model, which integrates CatBoost, Random Forest, LightGBM, and XGBoost, significantly outperforms individual models in terms of prediction accuracy. Through SHAP analysis, the pivotal role of ESG metrics in financial prediction is confirmed, revealing the substantive impact of non-financial indicators on corporate value.
Future work could focus on incorporating more granular real-market data and combining time-series models, such as LSTM[10] or Transformer,[11] to further capture the lag effects and dynamic trends of ESG impacts.

## ACKNOWLEDGMENTS

This research was supported by [Ministry of Education Industry-Academy Cooperation and Collaborative Edu-cation Program] (No. 241200363244457) and [Guangdong Province Undergraduate Teaching Quality Engineering Construction project in 2023 - Teaching and Research Office on Business Digitalization and Business Intelligence Curriculum Group] and National Innovation and Entrepreneurship Training Program for Undergraduate (No. 202510559147X).

## REFERENCES

[1] Friede, G., Busch, T., and Bassen, A., "Esg and financial performance: aggregated evidence from more than 2000 empirical studies," *Journal of Sustainable Finance & Investment* **5**(4), 210–233 (2015).
[2] Eccles, R. G., Ioannou, I., and Serafeim, G., "The impact of corporate sustainability on organizational processes and performance," *Management Science* **60**(11), 2835–2857 (2014).
[3] Henrique, B. M., Sobreiro, V. A., and Kimura, H., "Literature review: Machine learning techniques applied to financial market prediction," *Expert Systems with Applications* **124**, 226–251 (2019).
[4] Breiman, L., "Random forests," *Machine Learning* **45**(1), 5–32 (2001).
[5] Chen, T. and Guestrin, C., "Xgboost: A scalable tree boosting system," in [*Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*], 785–794 (2016).
[6] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., and Liu, T. Y., "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in Neural Information Processing Systems* **30**, 3146–3154 (2017).
[7] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A., "Catboost: unbiased boosting with categorical features," *Advances in Neural Information Processing Systems* **31** (2018).
[8] Wolpert, D. H., "Stacked generalization," *Neural Networks* **5**(2), 241–259 (1992).
[9] Lundberg, S. M. and Lee, S. I., "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems* **30** (2017).
[10] Hochreiter, S. and Schmidhuber, J., "Long short-term memory," *Neural Computation* **9**(8), 1735–1780 (1997).
[11] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I., "Attention is all you need," in [*Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*], 5998–6008 (2017).