Interpretability Analysis of Battery Capacity Model

Quanxi Guo[a], Tong Sun [b], Xin Li [b]
[a]*Department of Energy, Politecnico di Milano, 20156, Milan*
[b]*Department of ABC, Politecnico di Milano, 20156, Milan*
Email:(guo.quanxi, tong.sun, xin1.li)@polimi.it

Abstract—**Accurate prediction of the remaining capacity of the battery is the focus of current battery life research. Common data-driven methods usually use features related to capacity as input to build a capacity prediction model. However, it is difficult to reveal the intrinsic relationship between capacity and input features based on prediction accuracy alone. Therefore, it is necessary to use interpretable modeling methods or use post-interpretability techniques to perform feature contribution analysis on black box models. This study first uses ridge regression to perform interpretable modeling on the capacity of the CACLE battery dataset and then uses SHAP to analyze the relationship between LSTM input features and capacity. Experimental results show that ridge regression can quickly output the fitting formula between input features and target variables, but in the presence of strong collinearity, its regression coefficient may be disturbed. SHAP can avoid the impact of this collinearity and independently analyze the contribution value of each input feature to the target variable.**

Keywords—Ridge Regression, SHAP, Explainability, LSTM

I. Introduction

With the widespread application of lithium-ion batteries in electric vehicles, energy storage systems and other fields, accurate prediction of the remaining capacity of the battery is of great significance for ensuring the safe operation of the equipment and extending the battery life. Traditional physical model-based methods are often difficult to achieve high-precision and real-time capacity prediction due to the limitations of complex electrochemical mechanisms and environmental changes. In recent years, machine learning technology has become an important tool in the field of battery performance prediction due to its powerful data-driven capabilities, significantly improving the accuracy of predictions. However, complex machine learning models such as deep neural networks are often regarded as "black boxes" and their decision-making processes are difficult to explain, which limits the trust and application of models in engineering practice. Therefore, interpretable machine learning models and post-hoc Interpretability methods have become a solution to this problem.

In the study of modeling battery remaining capacity using interpretable machine learning models, Kumarappa and Manjunatha [1] evaluated the prediction effects of multiple algorithms such as regression, support vector machine, deep neural network, and identified the key factors affecting battery degradation through feature selection techniques, thereby enhancing the explanatory power of the model, In response to the limitations of traditional models in terms of training speed and robustness. Liu et al. [2] proposed an improved Extreme Learning Machine (ELM) model that combines grey correlation analysis and Bayesian optimization to achieve high-precision capacity prediction with an error rate of less than 0.2%, while also improving the model's interpretability and training efficiency. Wang et al. [3] constructed a battery capacity prediction model based on extreme random trees (ERT) and introduced Shapley additive explanations (SHAP) to analyze the contribution of each feature to the model output, thereby significantly enhancing the transparency and reliability of the model. In addition to the battery field, interpretable machine learning models are also widely used in other fields, such as Douglas et al [4] . used interpretable machine learning methods such as gradient boosting regressor (GBR) and extreme gradient boosting (XGBoost) to accurately simulate the pyrolysis of waste biomass. In fact, GBR and XGBoost are a type of integrated model based on decision trees. Luo et a [5] . proposed a new carbon emission characterization and prediction model based on interpretable machine learning and land use. It does not rely on socioeconomic indicators and is therefore able to predict carbon emissions after the decoupling effect. It can also reflect the spatial distribution characteristics of carbon emissions and exhibit high accuracy and interpretability. Fan et al [6]. used two typical ensemble learning models, random forest and XGBoost, to simulate the toxicity of ILs to Vibrio fischeri. The hyperparameters of the model were fine-tuned using Bayesian optimization, and its robustness was enhanced by 5-fold cross validation. Model comparison results showed that the XGBoost model exhibited good generalization ability.

On the other hand, post-hoc Interpretability methods are also widely used in the battery field. Most of these methods are combined with neural networks to explain the direct relationship between the input and output of the neural network. Huang et al. [7] used SHAP to analyze feature contributions in a temporal convolutional network (TCN) and combined it with the Equilibrium Optimizer to optimize model parameters, ultimately significantly improving prediction accuracy and feature selection rationality. Li et al. [8] proposed an interpretable online prediction method that combines support vector regression and SHAP to reflect the battery capacity regeneration process and quantitatively explain the impact of different health indicators. The transfer learning method proposed by Lin et al [9]. combining empirical knowledge with limited information in the target dataset to achieve excellent prediction performance for battery life. Not only in the battery field, but also in other fields, post hoc explanation is still widely used. Yang et al [10]. proposed a dual-branch convolutional neural network (DBCNN) and its post-hoc interpretability, using the SHAP framework to explain the prediction results by exploring feature contributions. Comparative experiments showed that compared with traditional models, the proposed model can enhance feature representation and fusion capabilities, thereby improving the performance of mineral exploration mapping. Ma et al [11]. proposed an intrinsic and post-hoc interpretable method combining Kolmogorov-Arnold network and genetic algorithm, using SHAP method to provide global and local insights into the overall and sample-specific effects on breakpoint tensile strength. Validation of optimization results confirmed that the proposed method outperforms other state-of-the-art methods.

This study will use the ridge regression model and Long Short-Term Memory Network (LSTM) combined with the SHAP method to perform interpretable modeling of the remaining capacity of lithium batteries. The goal is to enhance the comprehensibility and physical rationality of the model decision-making process and provide reliable and credible support for battery capacity evaluation.

II. DATASET

The data used in this study comes from the lithium-ion battery aging data set provided by the CALCE Laboratory of the University of Maryland. The selected battery number is CS2_35. The charging strategy of this battery is a standard constant current-constant voltage (CC-CV) mode: first, constant current charging is performed at a constant current of 0.5C until the battery voltage reaches 4.2V, and then it enters the constant voltage stage, maintaining the voltage at 4.2V until the charging current decays to below 0.05A. The discharge process adopts a constant current discharge mode, discharging at a constant current of 1C, and the termination voltage is set to 2.7V.

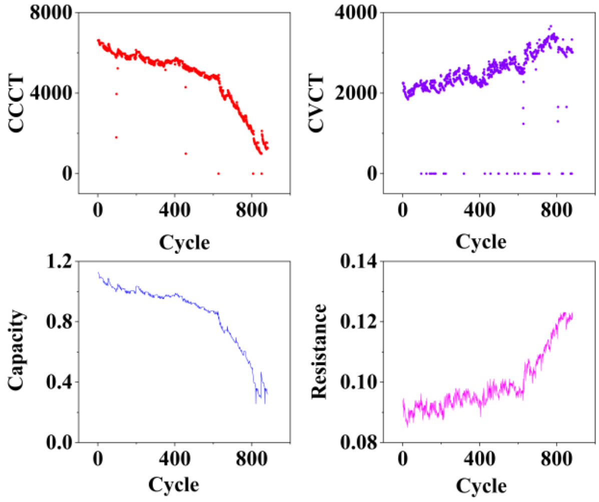

Fig. 1. Dataset Visualization

The data collection interval is 30 seconds, and the entire data set contains about 900 complete charge and discharge cycles. In the data preprocessing stage, key parameters such as battery capacity (Capacity), constant current charging time (Constant Current Charging Time, CCCT), constant voltage charging time (Constant Voltage Charging Time, CVCT) and internal resistance (R) are extracted from each cycle as features and target variables for modeling in subsequent studies. The figure below shows the trend of these parameters as a function of cycle number; the specific parameters change with the cycle as shown in Fig. 1.

III. METHODOLOGYT

This study plans to use an interpretable model (ridge regression) and a post-hoc interpretation method (SHAP) to perform interpretability analysis on the same set of data. This section will introduce the methods used.

A. Ridge Regression

Ridge regression is an improved version of linear regression, which alleviates multicollinearity and overfitting problems by introducing L2 regularization terms [12][13]. The core idea of ridge regression is to penalize the size of model parameters and limit the size of parameters based on the least squares method, thereby improving the generalization ability of the model on the test set.

For a regression problem, it can be simplified in the form of Eq.1.

$$y = X\beta + \varepsilon \tag{1}$$

Where y is the target variable, X is the feature matrix, β is the regression coefficient vector, and ε is the error term. In this study, the target variable is capacity, and the feature matrix is CCCT, CVCT and R.

The solution formula for ridge regression is as follows:

$$\min_{\beta}\left\{\|y - X\beta\|^2 + \lambda\|\beta\|^2\right\} \tag{2}$$

Where λ is the regularization strength, which is a hyperparameter, $\|\beta\|^2$ is the L2 norm of the parameter, these two together are the regularization term used to prevent the model from overfitting and underfitting.

Through the above solution process, we can finally get the regression coefficient matrix β, which clearly describes the mapping relationship between the feature matrix X and the target variable y, thus completing the modeling of the target variable. It is worth emphasizing that the model constructed by ridge regression has good interpretability, which means that the model structure is transparent and can intuitively reveal the specific impact of the input features on the output variables.

For the implementation of ridge regression, it is proposed to directly call Ridge in sklearn.linear_model in Python, and the regularization strength L is set to 0.1

B. SHAP

SHAP is a widely used model interpretation method with a solid theoretical foundation. It is based on the Shapley value in game theory and aims to fairly distribute the contribution of each feature to the model prediction results [14]. In this study, SHAP will be placed at the back end of LSTM to perform post hoc interpretable analysis of the input and output of LSTM.

The specific formula is as follows:

$$f(x) = \phi_0 + \sum_{i}^{M} \phi_i \tag{3}$$

Where $f(x)$ is the model's predicted output for input x and ϕ_0 is the baseline prediction value of the model, which is the expected prediction value of the model, that is, what is the average output of the model when there is no feature information. ϕ_i The marginal contribution of feature i to the prediction result (Shapley value). Corresponding to this study, the input CCCT, CVCT and R of LSTM will be calculated respectively for their contribution to the output Capacity.

The SHAP value calculation for feature i is as follows:

$$\phi_i = \sum_{S \subseteq F/\{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!}\left(f_{s\cup\{i\}}\left(x_{s\cup\{i\}}\right) - f_s\left(x_s\right)\right) \tag{4}$$

Where F is the set of all features, $S \subseteq F/\{i\}$ is any subset that does not include the feature i and $\frac{|S|!(|F|-|S|-1)!}{|F|!}$ is the weight item, $f_{s\cup\{i\}}\left(x_{s\cup\{i\}}\right)$ represents the model prediction output after adding feature i, and $f_s\left(x_s\right)$ represents the model output without feature i.

Eq. 4 can be used to calculate the contribution of each feature to the final output, thus making the direct relationship between input and output explainable.

In this study, three input variables (CCCT, CVCT, and R) will be used to calculate their SHAP values for the output variable (Capacity) to quantify the contribution of each input feature to the model prediction results. The specific calculation process will be implemented based on the SHAP package in the Python environment to achieve the interpretation and analysis of the model prediction.

The LSTM model used in this study consists of three layers: input layer, one hidden layer, and output layer. The number of neurons in the hidden layer is set to 8. During the training process, data will be input in batches, with a single batch size of 8, a learning rate of 0.001, and an epoch of 100.

C. *Dataset Partitioning*

The CALCE dataset used in this experiment contains data from about 900 battery cycles. To ensure the objectivity of the model evaluation, the dataset will be divided into a training set and a test set in a ratio of 7:3 to avoid the model from knowing the test set information in advance during the training process, thereby ensuring the validity and reliability of the evaluation results.

D. *Evaluation*

For model evaluation, Relative Error (RE), Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) will be used to evaluate the difference between the predicted results and the actual data. The specific formulas for the three evaluation methods are as follows:

$$RE = \frac{|y_i - \hat{y}_i|}{|y_i|} \tag{5}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{6}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{7}$$

Where n is the total number of samples, y_i is the true value, and $\hat{y}_i$ is the predicted value.

IV. Results and Discussion

A. *Ridge Regression*

This study uses the ridge regression method to model the input characteristics and output variables of the battery. The model is trained based on the training set divided in the previous chapter, and the model performance is evaluated through the test set. The relevant results are shown below.

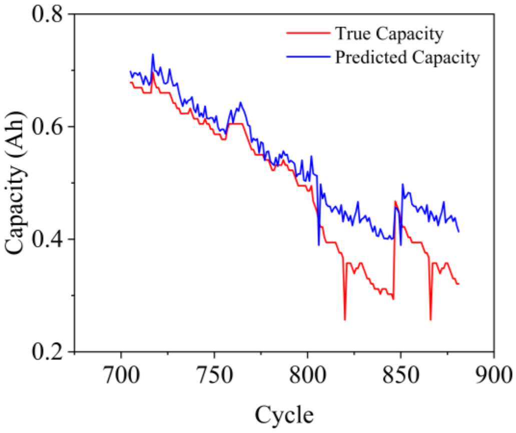

Fig. 2. Comparison of ridge regression predictions and true values

As can be seen from Fig.2, the ridge regression prediction effect on battery capacity is acceptable. However, since the capacity of the battery data changes suddenly at the end of the cycle, the model cannot accurately predict the capacity at the end of the cycle. This experiment aims to verify the effectiveness of the constructed model, that is, to evaluate whether it can accurately predict the output variables based on the mapping relationship learned from historical data when faced with unseen input data.

More importantly, ridge regression not only has good predictive ability, but also can clearly reveal the relationship between input features and output variables, thereby constructing an interpretable regression model.

$$Capacity = 0.404CCCT - 0.013CVCT - 0.242R \tag{8}$$

Eq. 8 shows the relationship between input and output in this study. It can be seen that CCCT is positively correlated with capacity, while CVCT and R are negatively correlated with capacity. This means that as the battery capacity decreases, CCCT also decreases, while CVCT and R increase. In addition, the ridge regression also retains the weight coefficient of each feature, which can help determine which features are the main influencing factors.

B. *SHAP*

This section uses LSTM to predict battery capacity. The model input and data set division method are consistent with the ridge regression model.

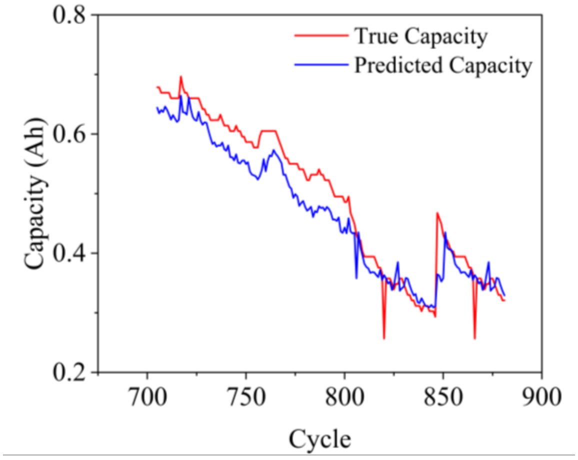

Fig. 3. Comparison of LSTM predictions and true values

Fig. 3 shows the prediction results of battery capacity based on the LSTM model. As a recursive neural network structure that specializes in processing time series data, LSTM shows better prediction accuracy than ridge regression in capacity prediction tasks. Although at the end of the battery cycle, due to the mutation of data points, LSTM also shows a increase in prediction error, a trend similar to the ridge regression model, but overall, LSTM more effectively captures the dynamic characteristics of capacity evolution over time.

Fig. 4 shows a feature importance graph based on SHAP analysis, which is used to explain the marginal contribution of different input variables when LSTM predicts battery capacity. The vertical axis in the figure is the feature name (CCCT, resistance, CVCT), and the horizontal axis is the corresponding SHAP value, which represents the degree and direction of influence of each feature on the model output. Each point represents a sample, and the color represents the value of the feature in the sample. As can be seen from the figure, the SHAP values of CCCT and resistance are mainly distributed in the negative area, indicating that in most samples, these two features have a negative impact on the battery capacity; that is, the increase of these features often leads to a decrease in the predicted capacity value. In addition, most of the high-value (red) samples of CCCT correspond to smaller SHAP values, further indicating that its larger values usually have a negative contribution to the model output. Similarly, high values of resistance also correspond significantly to negative SHAP values.

In contrast, the SHAP values of CVCT are concentrated near zero and the distribution is relatively symmetrical, indicating that the marginal contribution of this feature in the current model is small and its impact on the prediction of battery capacity is not significant.

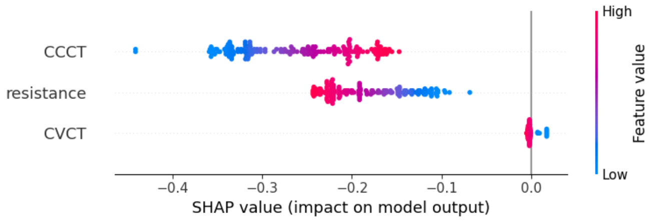

Fig. 4. SHAP value of different inputs

However, in the previous section, the regression coefficient of CCCT in the ridge regression model is positive (0.404), which theoretically indicates that CCCT has a positive effect on capacity, but the SHAP analysis results show that the SHAP value of CCCT is generally negative, indicating that CCCT has a negative effect on capacity in most samples. To further explore the reasons for the difference between the two, CCCT and R were collinearly analyzed, and the results are as Fig. 5.

As shown in Fig. 5, the vertical axis is the SHAP value of R, the horizontal axis is its numerical value, and the color bar indicates the value of CCCT in the corresponding sample. It can be observed in the figure that with the increase of R, CCCT shows a highly linear downward trend, and the two are almost strictly negatively correlated, indicating that there is a significant collinearity between the two. In the ridge regression model, the regression coefficient of R is negative (-0.242), which means that its increase will significantly reduce the predicted capacity value. Although the corresponding coefficient of CCCT in the linear regression model is positive, due to its strong correlation with R, it often decreases with the increase of R in actual samples, thereby indirectly causing a decrease in capacity. Therefore, the feature importance revealed by SHAP does not contradict the performance of the ridge regression model in the global coefficient but further reflects the influence of the dependency between features on the interpretation of the model.

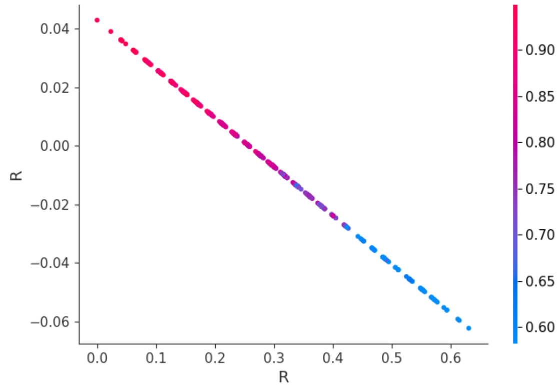

Fig. 5. R vs CCCT

In addition, Pearson correlation analysis was performed between the features and between the features and the target variable. The purpose is to better support the above conclusions.

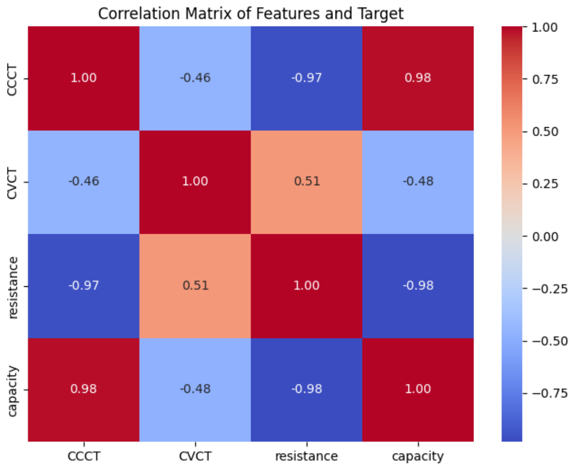

Fig. 6. Pearson correlation analysis

As can be seen from Figure 6, although CCCT is positively correlated with capacity in the original data, its contribution is largely explained or replaced by R. Therefore, in the SHAP analysis, the model assigns more explanatory power to R, making the marginal contribution of CCCT negative. This phenomenon also shows that SHAP analysis is not simply evaluating the statistical correlation between features and targets, but quantifying the conditional contribution of features under a specific model structure, thereby revealing the actual decision logic of the model.

C. Evaluation

This section will evaluate the two models using the evaluation method mentioned in Chapter 3. Although the main goal of this study is to explore the interpretability of the model, we believe that the model's accurate response to input features is the basis for studying the model's interpretability.

TABLE I. ERROR ANALYSIS

	Ridge Regression	LSTM
RMSE	0.0601	0.0392
MAE	0.0476	0.0329
R^2	0.7671	0.9009

Table I shows the performance comparison results of the two models under three evaluation indicators. It can be seen that the LSTM model is superior to the ridge regression model in all evaluation dimensions and shows higher prediction accuracy. This result verifies the rationality of this experimental design and further proves the effectiveness of the model construction and training process from a quantitative perspective.

V. CONCLUSION

Based on CACLE battery data, this study constructed an interpretable model (ridge regression) and a black box model (LSTM), and combined the post-hoc interpretability method (SHAP) to systematically analyze the input feature contribution of the model. The results show that when modeling battery capacity, ridge regression can clearly reveal the relationship between capacity and each input variable in a linear and explicit form, providing a certain explanatory basis for understanding the system. However, ridge regression itself does not have the ability to identify collinearity between input variables, so its regression coefficient may be affected in scenarios where features are highly correlated. In contrast, LSTM combined with SHAP value analysis can make a more detailed assessment of the marginal contribution of each input variable, especially in the presence of severe collinear feature combinations (such as the high negative correlation between CCCT and R). SHAP can effectively avoid redundant information interference between variables and reflect a more realistic feature contribution direction. After a comprehensive comparison of the two methods, it is found that although the interpretable model is more intuitive in expression, in actual scenarios where there is strong collinearity between input features, post-hoc interpretability methods such as SHAP have stronger explanatory robustness and higher credibility.

VI. CODE AND DATA

The code and data used in this study will be publicly available on GitHub at the following link:

QuanxiGuo/Explainable-Course

REFERENCES

[1] Kumarappa, S., M, M.H., 2024. Machine learning-based prediction of lithium-ion battery life cycle for capacity degradation modelling. World Journal of Advanced Research and Reviews 21, 1299–1309

[2] Liu, Z., Huang, Z., Tang, L., Wang, H., 2024. Lithium-Ion Battery Capacity Prediction Method Based on Improved Extreme Learning Machine. Journal of Electrochemical Energy Conversion and Storage 22.

[3] Wang, Y., Kumar, A., Ren, J., You, P., Seth, A., Gopaluni, R.B., Cao, Y., 2024. Interpretable Data-Driven Capacity Estimation of Lithium-ion Batteries. IFAC-PapersOnLine, 12th IFAC Symposium on Advanced Control of Chemical Processes ADCHEM 2024 58, 139–144.

[4] Divine, D.C., Hubert, S., Epelle, E.I., Ojo, A.U., Adeleke, A.A., Ogbaga, C.C., Akande, O., Okoye, P.U., Giwa, A. and Okolie, J.A., 2024. Enhancing biomass Pyrolysis: Predictive insights from process simulation integrated with interpretable Machine learning models. Fuel, 366, p.131346.

[5] Luo, H., Wang, C., Li, C., Meng, X., Yang, X. and Tan, Q., 2024. Multi-scale carbon emission characterization and prediction based on land use and interpretable machine learning model: A case study of the Yangtze River Delta Region, China. Applied Energy, 360, p.122819.

[6] Fan, D., Xue, K., Zhang, R., Zhu, W., Zhang, H., Qi, J., Zhu, Z., Wang, Y. and Cui, P., 2024. Application of interpretable machine learning models to improve the prediction performance of ionic liquids toxicity. Science of The Total Environment, 908, p.168168.

[7] Huang, G., Fu, L., Liu, L., 2025. Lithium-Ion Battery State of Health Estimation Based on Model Interpretable Feature Extraction. J. Electrochem. Soc. 172, 020513.

[8] Li, Z., Shen, S., Ye, Y., Cai, Z., Zhen, A., 2024. An interpretable online prediction method for remaining useful life of lithium-ion batteries. Sci Rep 14, 12541.

[9] Lin, T., Chen, S., Harris, S.J., Zhao, T., Liu, Y. and Wan, J., 2024. Investigating explainable transfer learning for battery lifetime prediction under state transitions. eScience, 4(5), p.100280.

[10] Yang, F., Zuo, R., Xiong, Y., Xu, Y., Nie, J. and Zhang, G., 2024. Dual-branch convolutional neural network and its post hoc interpretability for mapping mineral prospectivity. Mathematical geosciences, 56(7), pp.1487-1515.

[11] Ma, S., Leng, J., Chen, Z., Du, Y., Zhang, X. and Liu, Q., 2025. Intrinsically and Post-Hoc Interpretable Kolmogorov-Arnold Network and Genetic Algorithm for Laser Deep Penetration Welding Parameters Optimization. *IEEE Transactions on Instrumentation and Measurement*.

[12] Hoerl, A.E. and Kennard, R.W., 1970. Ridge regression: applications to nonorthogonal problems. Technometrics, 12(1), pp.69-82.

[13] Marquardt, D.W. and Snee, R.D., 1975. Ridge regression in practice. The American Statistician, 29(1), pp.3-20.

[14] Lundberg, S.M. and Lee, S.I., 2017. A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.