**1. Introduction**

The adoption of AI models in accuracy-sensitive industries such as investment, precise manufacturing, and cyber security has been steadily increasing. As a result, Explainable AI (XAI) has gained unprecedented popularity as a means to better understand the decision-making process of artificial intelligence systems [1]. XAI addresses the need for transparency and interpretability, enabling users to comprehend the underlying factors influencing AI predictions and fostering trust and accountability [2]. At present, XAI is mainly applied in Natural Language Processing (NLP) and image recognition. In the field of image recognition, there is inherent complexity in the neural network algorithms underlying XAI systems, such as recurrent neural networks (RNNS), residual networks, and widely recognized convolutional neural networks (CNNS). Nevertheless, the inner workings of neural networks remain cryptic, eluding comprehensive human comprehension. Consequently, there arises an increasingly pressing call to transcend the mere construction of these enigmatic models, redirecting efforts towards the elucidation of the intricate interconnections that reside among the various layers comprising image recognition systems.

Traditionally, XAI applies decompositional approaches which classify the extracted pixels for feature analysis. These approaches encompass model selection, feature extraction, explainability techniques, and evaluation of prediction results [3]. In the case of two-dimensional image prediction, challenges arise from the high dimensionality of features and the high collinearity between local pixel points. Consequently, most studies rely on feature importance tests such as permutation importance or SHAP (Shapley Additive Explanations) methods based on the entire pixel image [4]. However, these methods have limitations from a statistical perspective. There is a clear correlation between different pixel blocks, where the prediction of a particular point is subtly influenced by its surrounding contextual information. In this interaction, one-to-one correspondence cannot be established, and the accuracy of prediction remains to be investigated. Moreover, the practice of extrapolating from one opaque "black box" model to another lacks a sound mathematical basis. It relies on the predictions produced by these black-box models, thus introducing random factors into the model's decision-making

process [5].

In order to solve the drawbacks of such traditional methods, pedagogical approaches have been proposed. These approaches involve simulating the input-output dynamics of a complex neural network using a simplified model. The aim is to explore the complex internal relationships in complex neural networks by studying the mathematical correlations in these simplified models. Previously, Oclay proposed a new method, Decision Tree Extractor (DecText). In this method, the labeled data, unlabeled data and random data were processed by the neural network to obtain prediction labels, so as to represent the decision process in a transparent model through the diagram of the tree structure [6]. It showed the potential to illustrate the logic inside the black box, however, it contains some drawbacks. The most serious point is that due to the overfitting of data, useless decision layers would significantly affect the processing and judgment of the effective diagnostics.

In addressing the issue, we introduced an innovative approach that leverages features extracted through Principal Component Analysis (PCA) in conjunction with Multiple Linear Regression (MLR). Diverging from the complexity inherent in the construction process of decision trees, our method initiated a deconstruction of the data set during the collection and assembly phase, subsequently subjecting it to analysis through the more straightforward and explanatory MLR. The features derived through Principal Component Analysis (PCA) exhibit orthogonality, signifying the absence of correlations among distinct features, which made them amenable to subsequent analysis with MLR. Therefore, our proposed method realized the diagnostics with strong statistical foundations. Specificly, it reduced feature correlations.and enhanced the effective components of the model with explainable significance.

In this study, we specifically focus on neural network decision analysis by utilizing the MNIST dataset, a widely recognized benchmark dataset in the field of computer vision [7]. The structure of the paper is organized as follows. Section 2 demonstrates the process of data establishment. Section 3 explains the methodologies which we refer to set up the study. Section 4 elaborates the analysis of the regression results. Section 5 discusses three application scenarios of our proposed method compared to traditional approaches.

**2. Data establishment**

The adoption of AI models in accuracy-sensitive industries such as investment, precise manufacturing, and cyber s

Reference

[1] P. J. Phillips, C. Hahn, P. Fontana, A. Yates, K. K. Greene, D. Broniatowski, and M. A. Przybocki, “Four principles of explainable artificial intelligence,” 2021.

[2] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why Should I Trust You?: Explaining the Predictions of Any Classifier,” in 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (B. Krishnapuram, M. Shah, A. J. Smola, C. C. Aggarwal, D. Shen, and R. Rastogi, eds.), pp. 1135–1144, ACM, 2016.

[3] C. Molnar, Interpretable machine learning: A guide for making Black Box models explainable. Munich, 2022.

[4] S. M. Lundberg and S. Lee, “A unified approach to interpreting model predictions,” pp. 4765–4774, 2017.

[5] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning: with Applications in R. Springer, 2013.

[6] Oclay B. Extracting Decision Trees From Trained Neural Networks. SIGKDD. 2002

[7] L. Deng, “The mnist database of handwritten digit images for machine learning research,” IEEE Signal Processing Magazine (SPM), vol. 29, no. 6, pp. 141–142, 2012.