The rapid adoption of AI models in precision-critical sectors, such as finance, precise manufacturing, and cybersecurity, has called upon the development of eXplainable AI (XAI), as a crucial tool for enhancing transparency and interpretability in AI systems [1]. XAI addresses the imperative need for comprehending the decision-making processes of artificial intelligence, fostering trust and accountability [2]. Currently, XAI finds extensive application in Natural Language Processing (NLP) and image recognition. In the realm of image recognition, the underlying neural network algorithms, encompassing recurrent neural networks (RNNs), residual networks (ResNets), and the widely utilized convolutional neural networks (CNNs), introduce intricate complexities due to strong non-linearities within and between multi-layer perceptrons (MLPs). Although previous visualization attempts could graphically present the each layer’s reliance on the receptive fields and conclude what layer is employing what kind of information (edges/ strips/ color), the analysis process is not well-established in a mathematical foundation, and only presents a preliminary comparison that hinders more profound investigations towards the model decision process. This underscores the urgency of shifting focus from merely constructing these cryptic models to explicitly revealing the intricate interconnections among the various layers that constitute image recognition systems.

Traditional XAI methods rely on decompositional approaches which involves analyzing how individual input features or components independently influence a machine learning model's output or performance. It aims to break down complex model behavior into understandable parts, revealing the impact of each feature on predictions.[3]. In the context of two-dimensional image prediction, most studies resort to feature importance tests like permutation importance or SHAP (Shapley Additive Explanations) methods applied to the entire pixel image [4]. However, these methods exhibit statistical limitationsbecause they fail to account for the nuanced interactions between different pixel blocks, where the prediction of one point is subtly influenced by its surrounding contextual information. Establishing one-to-one correspondence in such interactions remains inconclusive if the correlation between neighboring pixel grids is not properly addressed . Moreover, extrapolating from one opaque "black box" model to another lacks a solid mathematical foundation and introduces randomness into the decision-making process [5].

To address these limitations, pedagogical approaches have been proposed. Instead of dissecting input features, they create surrogate models or alternative architectures that capture the essential aspects of the original neural network's behavior. These simplified models aim to provide insights into how the original network works without directly inspecting individual features but by replicating the overall dynamics and relationships within the network. Previously, Oclay introduced a method known as Decision Tree Extractor (DecText). DecText processes labeled data, unlabeled data, and random data through a neural network to obtain prediction labels, subsequently representing the decision process in a transparent model through a tree structure diagram [6].

While it demonstrates potential for illuminating the logic within black box models, this established pedagogical method, as the author claims in their work, suffers from effective node filtering. This is because the inherent decision tree structure is prone to overfitting when the hyperparameters of the tree components is not adequately controlled. And there are no explicit assessment approaches to determine the deepness of the tree structure because it is highly dependent on the sparsity of the dataset. As a result, they can capture noise in the data, leading to the inclusion of spurious decision nodes.

(放Figure 1）

（引出方法前 缺一个更高维的概括 如下）

Based on the above analysis, decomposition methods show effectiveness in feature extraction via neural networks, but they often fall short in providing rigorous mathematical and statistical diagnostic tools to evaluate their results. Instead, existing teaching methods provide a level of mathematical logic that is interpretable. However, these methods are prone to overfitting or potential information loss during the feature extraction phase. This reminds us of the need for a balanced approach that bridges the gap between efficient feature extraction and sound mathematical logic, addressing the limitations of current approaches.

In response to these challenges, we present an innovative approach that leverages Principal Component Analysis (PCA)-extracted features in conjunction with Multiple Linear Regression (MLR). The PCA extracted features are mutually orthogonal, which serves to eliminate the correlation between features that hinders the interpretability of the regression analysis. The advantages of MLR also inspired us to apply it to our study. First, MLR brings key aspects of rigorous hypothesis testing to the forefront, allowing for meticulous evaluation of the relationship between independent and dependent variables. In addition, the feature diagnosis methods provided by MLR play a key role in determining the importance and relevance of features in the dataset. This, in turn, enables more informed decisions about which features to include or exclude in the model. A significant advantage of MLR over decision trees is its ability to automate the feature selection process by means of forward, backward, or step-by-step selection. This automation not only saves time, but also reduces the subjectivity of manually determining nodes in the decision tree, effectively simplifies the process of model construction, and improves the reliability and interpretability of model construction. Consequently, our proposed method establishes diagnostics on robust statistical foundations, reducing feature correlations, and enhancing the model's interpretable components.

This study focuses specifically on neural network decision analysis, utilizing the MNIST dataset—a well-recognized benchmark in computer vision [7]. (Add three highlights of the model afterwards)The structure of this paper is organized as follows: Section 2 outlines the data collection process, Section 3 explicates the methodologies employed in our study, Section 4 delves into the analysis of regression results, and finally, Section 5 discusses three application scenarios where our proposed method outperforms traditional approaches.