[1]Rabinder Kr. Prasad

[2]Moirangthem Tiken Singh

[3]Chandan Kalita

[4]Sikdar Md S. Askari

[5]Bikramjit Choudhury

Exploring Neural Network Decision Making: Extended Relevance Propagation and Beyond

Abstract: - This paper introduces Extended Relevance Propagation (ERP), a novel approach designed to enhance the explainability of neural networks. The effectiveness of ERP is evaluated using fidelity-based metrics, benchmarked against established interpretability methods such as Layer-wise Relevance Propagation (LRP) and SHapley Additive exPlanations (SHAP). The ERP model demonstrates superior stability and robustness, producing consistent and trustworthy explanations even under input perturbations. While SHAP delivers highly detailed explanations, it is more sensitive to input changes, whereas LRP offers a balance between interpretative depth and stability. All three methods generate heatmaps that visually emphasize key features influencing model decisions, thereby enhancing transparency and fostering trust. Robustness analyses further validate ERP's high fidelity, underscoring its suitability for applications that demand reliable and interpretable models.

Keywords: Explainability, Extended Relevance Propagation, Robustness, Interpretability.

I. INTRODUCTION

Neural networks, which are modeled after the structure of the human brain, have a remarkable ability to recognize intricate patterns hidden within data. These networks are composed of nodes that are connected to one another and organized into distinct layers. Raw data is initially processed by the input layer through weighted connections, this data is then passed to subsequent hidden layers where activation functions are used to further transform the data, culminating in the output layer's predictions or classifications. Because of their ability to model intricate, non-linear relationships within data, neural networks have achieved significant success in a wide range of domains, including image recognition, natural language processing, and autonomous systems [1].

The impressive capabilities of neural networks are often overshadowed by their transparency issues, which pose significant challenges to their trustworthiness and adoption, especially in critical applications like healthcare, finance, and autonomous driving [2]. Due to its significant relevance in image processing, the black box problem, a prominent issue, needs to be thoroughly explored and understood. This problem underscores the difficulty in deciphering the internal workings of deep neural networks, making it a challenge to ensure that their decisions adhere to principles of fairness, unbiasedness, and reliability, a concern highlighted in the research [3].

The existence of noise in the data exacerbates the inherent complexities of interpreting and relying on these models, making it more challenging to draw accurate conclusions and trust the results. The presence of noisy data poses a significant challenge for neural networks, as it hinders their ability to extract meaningful patterns and leads to predictions and explanations that are less reliable. Methods like Layer-wise Relevance Propagation (LRP) and SHapley Additive exPlanations (SHAP) for understanding the reasons behind model predictions often encounter difficulties in achieving a balance between accuracy and interpretability in the presence of these conditions, as observed in [4]

In an effort to address these challenges, this work presents the proposal of a new interpretable model, referred to as Extended Relevance Propagation (ERP). ERP aims to increase the transparency and trustworthiness of neural

[1,]2Department of Computer Science and Engineering, DUIET, Dibrugarh , India

[3]Department of Information Technology, Gauhati University, India

[4]Sikdar Md S. Askari, Department of Computer Science, Rajiv Gandhi University, India

[5]Department of CSE . Central Institute of Technology, Kokrajhar, India

[2]Corresponding author: Moirangthem Tiken Singh, *Email: tiken.m@dibru.ac.in

networks in crucial decision-making processes by focusing on enhancing the robustness and accuracy of explanations, especially when dealing with scenarios involving noisy data. To clarify the focus of this research, it seeks to provide answers to the following specific questions.

RQ1: In what ways does Extended Relevance Propagation (ERP) improve upon Layer-wise Relevance Propagation (LRP) in terms of handling dead nodes and noise while maintaining interpretability and accuracy? Can the proposed ERP method provide more consistent and reliable relevance scores compared to SHapley Additive exPlanations (SHAP) and LRP under varying conditions of input noise?

RQ2: How does the introduction of noise affect the performance of SHAP in interpreting the contributing pixels for the final output in comparison to LRP and ERP?

The goal of this work is to promote trust in neural networks by examining research questions that focus on improving the robustness and interpretability of their explanations, specifically in environments characterized by noise. Ensuring fairness, accountability, and transparency is paramount in critical applications, making this aspect a crucial factor in their adoption.

II. Literature Review

Addressing the opacity of neural networks has sparked the creation of numerous interpretability methods designed to provide insights into their decision-making processes. These methods are designed to demystify the opaque nature of neural networks by offering a deeper understanding of their decision-making processes, thereby providing valuable insights into their internal workings.

Among the wide range of interpretability techniques, Local Interpretable Model-agnostic Explanations (LIME) has emerged as one of the most notable approaches for gaining insights into the workings of diverse machine learning models. In essence, LIME operates by constructing a simplified, interpretable model (like a linear model or decision tree) that closely resembles the black-box model's behavior in the immediate vicinity of a given prediction. This localized approximation allows for understanding the model's decision-making process for that specific prediction. When LIME trains a local surrogate model on perturbed samples of the original data, it is able to capture the local decision boundary of the black-box model. The coefficients and structure of this surrogate model allow practitioners to understand which factors have the greatest influence on specific predictions. Despite its benefits, LIME's reliance on local approximations can result in unstable and inconsistent explanations, particularly when dealing with intricate datasets [5].

SHAP, or SHapley Additive exPlanations, is another popular technique. SHAP combines cooperative game theory to give a single way to evaluate the importance of features in individual predictions. It leverages Shapley values to calculate the average impact of each feature, taking into account all possible feature combinations, leading to impartial and consistent importance measurements. Because SHAP adheres to properties like efficiency, symmetry, and additivity, it is a powerful tool for feature attribution. The computational complexity of this method can be a limitation, especially for large-scale models and high-dimensional data, according to [6].

Neural network predictions can be explained using a technique known as Layer-wise Relevance Propagation (LRP). LRP assesses feature importance by backpropagating the prediction score, assigning scores to neurons and ultimately determining the relevance of each input feature. By highlighting the contributions of individual input features, these relevance scores provide a clear and interpretable visual representation of the decision-making process. Although LRP effectively assigns relevance, its applicability is hindered in scenarios involving noisy or adversarial data, as the relevance scores become less informative [7].

By highlighting key areas, Saliency Maps illustrate how an input image influences the model's decision. By calculating the gradient of the output against the input, these maps are generated. Although saliency maps provide an intuitive and interpretable view of feature importance, their imprecision and potential for noise make it difficult to determine the exact impact of specific features. In addition, their ability to understand the complicated, non-linear patterns found in deep learning models is limited [8].

Counterfactual Explanations offer insights by demonstrating how slight variations in input data can lead to different predictions from the model. When it comes to comprehending decision boundaries and crucial features, this method

is particularly effective. Creating counterfactuals that are both realistic and actionable is a challenge, particularly when dealing with high-dimensional data [9].

Despite progress in neural network interpretability, existing methods are often inadequate when dealing with noisy data. When input is noisy, it can hide the true patterns, leading to explanations that are less trustworthy and less consistent. The accuracy of relevance scores produced by methods like LRP and SHAP is susceptible to noise variations, thus making it more difficult to interpret model behavior [4].

In response to these challenges, researchers have investigated noise-tolerant interpretability methods. One instance is SmoothGrad [4], which enhances saliency map robustness by averaging gradients across multiple slightly altered versions of the input. Despite their advantages, these approaches increase computational cost and could be less effective in situations with significant noise.

The need for more robust interpretability techniques is evident due to the limitations of existing methods, particularly when handling noisy data. This research suggests a new method called Extended Relevance Propagation (ERP) to address these deficiencies. By building on LRP, ERP aims to deliver more dependable and consistent relevance scores, even amidst noise. It seeks to overcome limitations of current methods by addressing problems like dead nodes and the spread of noise through network layers.

The progress in interpretability, enabled by techniques like LIME, SHAP, LRP, and saliency maps, has also unveiled limitations that necessitate continuous innovation. This research focuses on enhancing neural network transparency, robustness, and trustworthiness through ERP development, especially in critical applications requiring interpretability.

III. METHOD

LRP has significantly contributed to the interpretability of neural networks. However, it faces notable limitations, particularly when dealing with negative or zero weights, which can lead to misinterpretations by incorrectly identifying neurons as inactive or "dead." Additionally, LRP's reliance on averaging pixel contributions can compromise granularity, thereby reducing the interpretability of its explanations. To address these challenges, this paper proposes the Extended Relevance Propagation method, designed to enhance robustness and accuracy, particularly in noisy data conditions.

ERP refines the computation of relevance scores using the following equation:

$$R_j = \sum_k \left(\frac{a_j \times \rho(W_{jk})}{\epsilon + \sum_j a_j \times \rho(W_{jk})} \right) \times R_k, \qquad (1)$$

where R_j represents the relevance score assigned to neuron j, indicating its contribution to the final prediction. The term a_j denotes the activation of neuron j, calculated after applying the activation function, and W_{jk} is the weight connecting neuron j to neuron k in the subsequent layer. The function $\rho(W_{jk})$ ensures that only positive contributions from the weights are considered, thereby focusing on meaningful activations and excluding negative weights. A small constant ϵ is included to maintain numerical stability and prevent division by zero.

The ERP method improves interpretability by focusing on positive contributions, normalizing weighted activations, and smoothing the effects of noise. This approach ensures stable, granular relevance scores, making ERP particularly effective in noisy environments. The backward propagation of relevance scores through each layer begins with the output layer, where initial relevance scores are assigned. Weighted activations are computed for each connection between neurons, normalized using the total activation, and relevance scores are updated iteratively for all layers. The process continues until relevance scores for the input layer are determined.

The neural network considered in this study comprises an input layer, multiple hidden layers, and an output layer. The input layer processes the feature vector X with dimensions $n \times 1$. Each hidden layer i contains m_i neurons and performs computations involving a linear transformation followed by a non-linear activation function. Specifically, for the i^{th} hidden layer, the pre-activation output is calculated as $Z_i = W_i \times A_{i-1} + b_i$, where

W_i is the weight matrix, b_i is the bias vector, and A_{i-1} is the activated output from the previous layer. The activation function applied is the Rectified Linear Unit (ReLU), defined as $\mathrm{ReLU}(x) = \max(0, x)$. The output layer, which consists of h neurons, computes the final predictions using the softmax function, transforming the linear outputs into probabilities that sum to one.

The ERP algorithm (Algorithm 1) outlines the steps to compute relevance scores by backpropagating through the network layers. The ERP algorithm iteratively computes relevance scores for each layer by propagating information backward from the output layer to the input layer. Its time complexity is $O\left(\sum_{i=1}^{L} m_i \times m_{i+1}\right)$, where L is the number of layers, and m_i and m_{i+1} are the number of neurons in the current and next layers, respectively. By addressing the limitations of LRP, ERP enhances the interpretability and robustness of neural networks, particularly under noisy conditions.

Algorithm 1 Extended Relevance Propagation (ERP)

Require: Forward propagation outputs (*outputs*), weights (W), output layer relevance scores (R_{final}), small constant for numerical stability (ϵ), and a positive function for weights (ρ)

Ensure: Relevance scores for each layer (R)

Initialize list R with R_{final} for the final layer
for each layer i from last to first **do**
  Initialize R_i as a zero vector of the same length as *outputs*[i]
  for each neuron j in layer i **do**
    for each neuron k in layer $i+1$ **do**
      Compute $weighted\_activation_{jk} \leftarrow outputs[i][j] \times \rho(W[i][j][k])$
      Normalize contribution using total activation sum and ϵ
      Update $R_i[j] \leftarrow R_i[j] + \frac{weighted\_activation_{jk}}{\epsilon + \sum_j weighted\_activation_{jk}} \times R[i+1][k]$
    end for
  end for
  Prepend R_i to list R
end for
return R

IV. Result and Discussion

To evaluate the performance of the proposed model, the paper introduces a workflow depicted in Figure 1. This workflow outlines the evaluation process for comparing interpretability methods—Layer-wise Relevance Propagation (LRP), Extended Relevance Propagation (ERP), and SHapley Additive exPlanations (SHAP)—using a fidelity-based performance measure. The process begins with a dataset, which is divided into training and testing subsets.

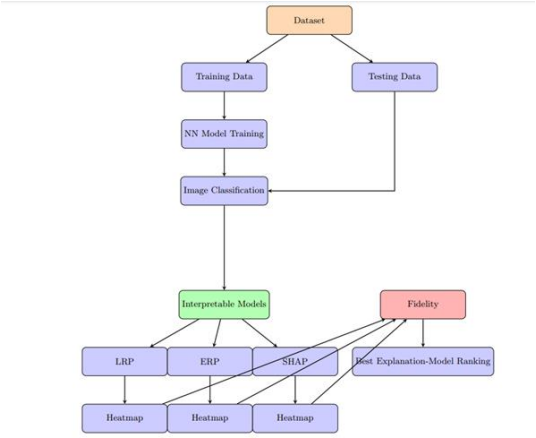

Figure 1: Workflow of The Model. This workflow systematically evaluates the effectiveness of different interpretability methods (LRP, ERP, SHAP).

The neural network (NN) model is trained on the training data and subsequently evaluated on the testing data. Post-training, the interpretability methods (LRP, ERP, and SHAP) are employed to generate relevance scores for image classification predictions. These scores are visualized as heatmaps, highlighting the most influential image regions in the model's decision-making process. A fidelity metric assesses the accuracy of these relevance scores in reflecting the model's reasoning, ranking the methods to identify the most effective one.

Analyses to explore the interpretability of the trained neural network are conducted on the MNIST dataset. Predictions are generated through forward propagation, followed by gradient computation via backward propagation to identify the pixels most impactful on the predictions. This process provides insights into the model's reasoning, enhancing transparency. Figure 3 illustrate the resulting heatmaps, which visually represent pixel contributions and highlight regions critical to the model's decisions. These visualizations foster an intuitive understanding of the model's behavior.

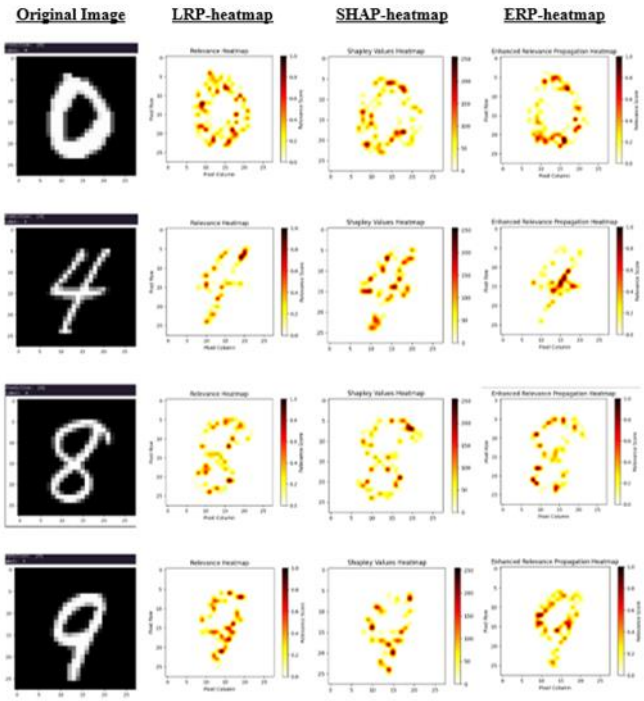

Figure 2: Comparison of interpretability techniques applied to MNIST digits using Layer- wise Relevance Propagation (LRP), SHapley Additive exPlanations (SHAP), and Extended Relevance Propagation (ERP) heatmaps. Heatmaps show which pixels are most crucial for the model's predictions, with brighter areas signifying greater influence on the final decision.

Figure 2 compares the interpretability techniques applied to MNIST digit predictions. Each heatmap method visualizes the pixel contributions differently:

- LRP highlights relevant pixels by redistributing the model's predictions backward through its layers.
- SHAP employs game-theoretic values to explain feature importance, approximating Shapley values for each pixel.
- ERP refines pixel relevance, providing sharper, more precise indications of critical regions.

To assess robustness, the mean difference in heatmaps between original and perturbed inputs was computed across all test samples. A low mean difference indicates stable, consistent explanations, whereas a higher value suggests sensitivity to input variations, reflecting lower reliability. Table 1 compares the mean values of the three methods across random samples, revealing their unique properties. LRP demonstrates moderate sensitivity, as seen in its relatively consistent values. SHAP exhibits greater variability, indicating higher sensitivity to input changes. ERP,

with near-zero values, showcases the highest stability and minimal vulnerability to perturbations, making it the most reliable method.

Table 1 : Mean Values for LRP, SHAP, and ERP

Random Sample Numbers	LRP Mean Value	SHAP Mean Value	ERP Mean Value
7	-0.3157	-2.5818	0.0002
9	-0.3144	-4.7100	-0.0008
1	-0.3068	-2.9322	-0.0007
4	-0.3081	-5.0607	-0.0013
6	-0.3081	-1.9229	-0.0009
2	0.0025	-4.9080	-0.0025
3	-3.1331	-2.6461	-0.0011

Table 2: Average Mean value

Models	Mean Values
LRP	-0.4045
ERP	-0.0010
SHAP	-3.5373

Table 2 presents the average mean values for the ERP, LRP, and SHAP models. The model with the mean difference closest to zero indicates the highest fidelity, meaning it provides the most consistent and reliable explanations. In this case, the ERP model, with a mean difference of -0.0010, demonstrated the greatest stability and robustness in its explanations, followed by LRP and SHAP. A lower mean difference value suggests that the model's explanations remain consistent even with minor input perturbations, making it more trustworthy and less prone to overfitting. High fidelity also implies the model's ability to focus on meaningful features, handling noise effectively without being influenced by irrelevant data variations.

V. CONCLUSION

This research evaluates three interpretability methods - LRP, ERP, and SHAP - through fidelity-based performance metrics. Among the methods tested, ERP emerged as the most reliable, with stable and consistent explanations, SHAP provided detailed but sensitive interpretations, and LRP offered a balance of consistency and depth. By enhancing transparency, heatmaps generated through these methods offered insights into the neural network's decision-making, fostering trust in its predictions. Future studies can extend this work by testing interpretability methods on larger datasets and advanced architectures, such as Transformers, applying them in practical domains like healthcare and finance, and developing new metrics for evaluating both human understanding and computational efficiency. The utility and reliability of these methods in practical applications can be further enhanced by exploring hybrid approaches, testing their robustness under adversarial settings, and conducting user studies. These efforts will guarantee the deployment of more transparent and reliable AI systems across different fields.

REFERENCES

[1] *R. Miikkulainen, J. Liang, E. Meyerson, A. Rawal, D. Fink, O. Francon, B. Raju, H. Shahrzad, A. Navruzyan, N. Duffy, et al., "Evolving deep neural networks," in Artificial Intelligence in the Age of Neural Networks and Brain Computing, pp. 269–287, Elsevier, 2024.*

[2] *W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Müller, "Explaining deep neural networks and beyond: A review of methods and applications," Proceedings of the IEEE, vol. 109, no. 3, pp. 247–278, 2021.*

[3] J. Renkhoff, W. Tan, A. Velasquez, W. Y. Wang, Y. Liu, J. Wang, S. Niu, L. B. Fazlic, G. Dartmann, and H. Song, "Exploring adversarial attacks on neural networks: An explainable approach," in *2022 IEEE International Performance, Computing, and Communications Conference (IPCCC)*, pp. 41–42, 2022.

[4] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "Smoothgrad: Removing noise by adding noise," *arXiv preprint arXiv:1706.03825*, 2017.

[5] D. Garreau and U. Luxburg, "Explaining the explainer: A first theoretical analysis of LIME," in *International Conference on Artificial Intelligence and Statistics*, pp. 1287–1296, 2020.

[6] G. Van den Broeck, A. Lykov, M. Schleich, and D. Suciu, "On the tractability of SHAP explanations," *Journal of Artificial Intelligence Research*, vol. 74, pp. 851–886, 2022.

[7] J. Sun, S. Lapuschkin, W. Samek, and A. Binder, "Explain and improve: LRP-inference fine-tuning for image captioning models," *Information Fusion*, vol. 77, pp. 233–246, 2022.

[8] T. Gomez, T. Fréour, and H. Mouchère, "Metrics for saliency map evaluation of deep learning explanation methods," in *International Conference on Pattern Recognition and Artificial Intelligence*, pp. 84–95, 2022.

[9] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, and S. Lee, "Counterfactual visual explanations," in *International Conference on Machine Learning*, pp. 2376–2384, 2019.