IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 9 September 2024, accepted 27 September 2024, date of publication 7 October 2024, date of current version 24 October 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3475380

RESEARCH ARTICLE

Explainable Image Recognition With Graph-Based Feature Extraction

BASIM AZAM[1,2], DEEPTHI P. KUTTICHIRA[1], BRIJESH VERMA[1], ASHFAQUR RAHMAN[3], (Senior Member, IEEE), AND LIPO WANG[4], (Senior Member, IEEE)

[1]School of Information and Communication Technology, Griffith University, Brisbane, QLD 4111, Australia
[2]School of Computing and Information Systems, The University of Melbourne, Melbourne, VIC 3010, Australia
[3]CSIRO, Hobart, TAS 7005, Australia
[4]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798

Corresponding authors: Basim Azam (basimazam0@gmail.com) and Lipo Wang (elpwang@ntu.edu.sg)

This work was supported by Australian Research Council's Discovery Projects Funding Scheme under Project DP210100640.

ABSTRACT Deep learning models have proven remarkably adept at extracting salient features from raw data, driving state-of-the-art performance across many domains. However, these models suffer from a lack of interpretability; they function as black boxes, obscuring the feature-level support of their predictions. Addressing this problem, we introduce a novel framework that combines the strengths of convolutional layers in extracting features with the adaptability of Graph Neural Networks (GNNs) to effectively represent the interconnections among neuron activations. Our framework operates in two phases: first, it identifies class-oriented neuron activations by analyzing image features, then these activations are encapsulated within a graph structure. The GNN in our system utilizes the connections between neuron activations to yield an interpretable final classification. This approach allows for the backtracking of predictions to identify key contributing neurons, enhancing the model's explainability. The proposed model not only matches, but at times exceeds, the accuracy of current leading models, all the while providing transparency via class-specific feature importance. This novel integration of convolutional and graph neural networks offers a significant step towards interpretable and accountable deep learning models.

INDEX TERMS Graph neural networks, convolutional neural networks, deep learning.

I. INTRODUCTION

Deep Learning (DL) approaches have made major improvements in the field of image classification in recent years. The ability to extract reliable and discriminative features from unprocessed image data, the cornerstone of any efficient image classification model, is a crucial component of these techniques. The model incorporates a variety of features, ranging from simple aspects like color and texture to more complex elements such as patterns and objects. These varied features equip the model with the necessary information to formulate accurate predictions.

In the areas of machine learning and computer vision, feature extraction and image categorization have received a great deal of attention. The Histogram of Oriented Gradients (HOG), Scale-Invariant Feature Transform (SIFT) [1] and Speeded Up Robust Features (SURF) [2] are conventional techniques for feature extraction. However, these techniques demand manual engineering and are frequently ineffective for difficult tasks like image recognition.

Convolutional Neural Networks (CNNs) have become an effective image classification technique in recent years. Due to CNNs' ability to automatically build feature representations with hierarchical structure from raw pixel data, image classification tasks have significantly improved [3]. The core component of CNNs is a feature extractor and classifier built on convolutional layers. These layers take the raw input data and extract its features. The CNN architecture's classifier component converts feature values into output class labels. CNNs are successful, yet despite this, they are frequently

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Wang.

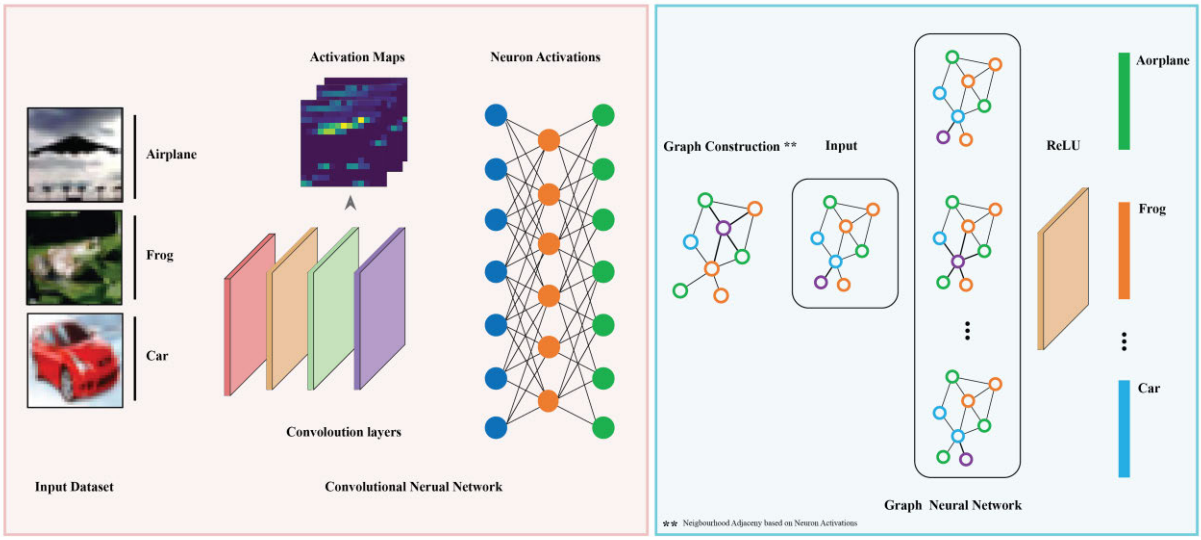


FIGURE 1. An overview of the proposed architecture providing explanations at each step using graph neural networks.

criticized for being difficult to understand [4]. These models' internal operations are frequently referred to as "black boxes," making it challenging to comprehend how they reach at their predictions [5].

Due to their ability to autonomously learn hierarchical feature representations from raw pixel data [6], CNNs are extensively utilized in image classification [7]. However, their complexity often leads to a lack of interpretability, a significant issue in critical applications like medical imaging and autonomous driving [8].

To address this, Explainable AI (XAI) has emerged, focusing on making AI decision processes transparent. Techniques such as saliency maps, layer-wise relevance propagation, deep Taylor decomposition [9], and Graph Neural Networks (GNNs) [10] are being explored to enhance model explainability. The traditional methods in for visualizing neuron activations such heatmaps [11] and activation maps [12], offer limited insights into the complex and dynamic interactions between features within neural networks. These static representations often fail to capture the nuanced relationships and intricate dependencies that exist across different layers of the network. To that end, we intend to investigate graph constructs ranging from neuron activations through classification output in order to give interpretability of features. The proposed approach transforms neuron activations into graph structures in a dynamic way.

While current approaches, particularly convolutional neural networks, have amazing prediction powers, they have a critical limitation: they are black boxes [13]. Despite their ability to extract features from raw data automatically, CNNs frequently provide little to no information into which characteristics or neurons contributed to a specific prediction [14]. This lack of openness and interpretability is problematic, particularly in situations when understanding the reasons behind a forecast is as crucial as the prediction itself. Handling high-resolution images poses a challenge as they can be represented by large graphs with numerous nodes and edges, complicating the analysis [18], [19], [20]. Additionally, dealing with dynamic graphs, which change structure over time, presents another layer of complexity in this domain [21], [22], [23].

Our research contributes significantly to computer vision community as:

- Developing an innovative framework that merges convolutional layer feature extraction with GNNs for modelling neuron activation relationships, enhancing model interpretability.
- Providing an architecture that enables explainability by tracing back predictions to contributing neurons, unlike traditional deep learning methods.
- Transforming neuron activations into graph structures, effectively capturing, and illustrating feature relationships.
- Presenting a detailed comparison with leading models, showing our model's superior accuracy, explainability, and applicability in decision-making processes.

In subsequent sections, we will explore in depth our proposed architecture, detailing the experimental setup and the achieved results. This will underscore the effectiveness and unique benefits of our approach.

II. PROPOSED METHODOLOGY

Our proposed method utilizes graph neural networks for feature extraction and image classification. GNNs excel in capturing complex node relationships within a graph, an attribute ideal for structured data tasks [14], [15], [16]. In image classification, each image pixel is treated as a graph node, with pixel relationships represented as graph

edges. This graph-based image representation allows GNNs to effectively extract features that represent spatial pixel relationships, thereby enhancing image classification capabilities [15], [17].

Algorithm 1 The Algorithmic Flow of Estimating Graph Embeddings and Explainable Classification

Input: *ImageI , Target Label L*
Output: *Classification Output*

```
for each Image I
    compute convolutional layers
    for each convolutional layer
        extract feature maps
        apply pooling operations to reduce spatial dimensions
        flatten feature maps
        obtain neuron activations using fully connected layers
        for each set of neuron activations
            compute similarity measure between feature vectors
        end
        for each set of feature vectors
            estimate the neuron activations associated with each label
        end
        for each graph G
            apply graph convolutions
            apply pooling operations
            pass graph-level representations through fully connected layer with SoftMax activations to obtain classification output
        end
    end
end
```

The proposed architecture aims to leverage the strengths of both convolutional operations for local image structure extraction and graph-based operations for global relational reasoning. The overall model involves a feature extraction component, a graph construction component, a graph convolution component, and a classification component. These components are not isolated but interconnected, forming an end-to-end trainable system.

A. FEATURE EXTRACTION

Given an input image $x_i \in \mathbb{R}^{H\times W\times C}$, where H, W, and C are height, width, and number of channels of the image, respectively. We utilize the deep neural network, which we denote as $F(\cdot)$, to transform the input image into a set of features defined by the mapping function F_i, where each feature corresponds to the output of a specific filter. This transformation can be represented as:

Where $F_i \in \mathbb{R}^{H'\times W'\times C'}$ are the height, width, and number of channels of the feature maps, θ_F denotes the parameters for the feature extraction.

$$F_i = F(x_i; \theta_F)$$

TABLE 1. Time and space complexity analysis of the algorithm.

	Time Complexity	Space Complexity
CNN	$O(q * m * n * k^2 * p * m * n * r * s)$	$O(m^2 * n^2 * p * (r + s)$
Similarity Computation	$O(t^2 * u)$	$O(t * u)$
GNN	$O(w * v^2 * r * s)$	$O(v^2(r + s))$
Overall	Dominated by GNN (as pre-trained CNN can be used to compute activations)	

where $m \times n$ is the input image size, k is the kernel size, p is the number of kernels, q is the number of convolutional layers, r, s are the number of neurons in the input and output layers, t is the number of feature vectors, u is the dimensionality of the feature vectors, v is the number of nodes in the input graph, w is the number of graph convolution layers.

B. GRAPH CONSTRUCTION

The next step is to construct a graph $G_i = V_i, A_i$ from feature maps F_i. Each node $v_{i,j} \in V_i$ in the graph corresponds to a region of the image and is assigned a feature vector $f_{i,j}$ extracted from F_i. The edges of the graph represent the relationships between different regions of the image. The adjacency matric $A_i \in \mathbb{R}^{n\times n}$, where n is the number of nodes, is defined based on the relationship between the feature vectors of the nodes as f the nodes as

$$A_{i,j,k} = \frac{f_{i,j} \cdot f_{i,k}}{\|f_{i,j}\|_2 \, \|f_{i.k}\|_2}$$

where $A_{i,j,k}$ id the entry at the *jth* row and *kth* column of A_i, and $\cdot$ denotes the dot product.

C. GRAPH CONVOLUTIONS

We then perform graph convolution operation, denoted as $GC(.)$, to propagate information through the graph. The operation updates the node features based on their own features and the features of their neighbours capturing the relational information between different regions of the image. The node embedding N_i^l, after l graph convolution layers can be represented as:

$$N_i^l = \sigma(GC(N_i^{(l-1)} A_i; \theta_{GC}^l))$$

where $N_i^l \in \mathbb{R}^{n\times d_l}$ denotes the node embeddings after l layers, $N_i^o = V_i$, θ_{GC}^l denotes the parameters of the lth graph convolution layer, d_l is the dimension of the node embeddings after l layers, and $\sigma(\cdot)$ and is a non-linear activation function, such as the ReLU function.

D. CLASSIFICATION

The node embeddings N_i^L after L layers of graph convolution are then aggregated to generate a graph embedding $G_i \in \mathbb{R}^{d_L}$.

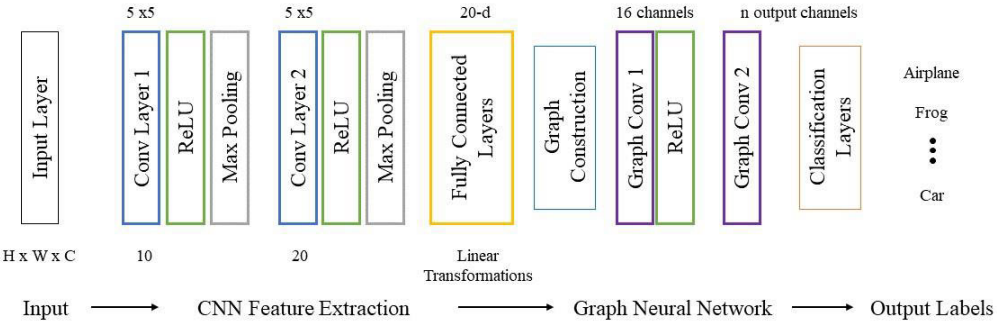


FIGURE 2. Schematic of deep learning model architecture combining convolutional neural networks and graph neural networks for image classification.

This is done using the mean pooling:

$$G'_i = \frac{1}{n}\sum_{j=1}^{n} N^L_{i,j}$$

where $N^L_{i,j}$ denotes the node embedding at Layer L and n is the total number of nodes.

The graph embedding G'_i is then passed through a classifier to predict the output label $y_{\hat{i}}$:

$$y_{\hat{i}} = argmax(C'(G'_i; \theta_C))$$

$$\mathcal{L}(\theta) = -\frac{1}{N}\sum_{i=1}^{N} \log(softmax\left(C'\left(G'_i; \theta_C\right)\right)[y_i])$$

where N is the number of samples, C′ denotes the classifier function, θ_c its parameters. The algorithm 1 details the process of proposed architecture, while table 1 presents the time and space complexity analysis of the algorithm.

III. EXPERIMENTAL SETUP & RESULTS

In this section, the description of dataset, evaluation metrics, details of hardware and software setup, and the experiments carried out are discussed.

A. EXPERIMENTAL SETUP

The study proposes a novel architecture that integrates CNNs and GNNs to provide explainable insights into image classification tasks. The CNN component extracts spatial features from the input image through a series of convolutional layers. These activations are then used to construct graphs which are processed by GNN component. Figure 2 presents the layer-wise details of how the CNN and GNN components are combined in the architecture. The architecture begins with the input image being processed through several convolutional layers of the CNN, which extract spatial features. These features are then used to construct a graph where each node represents a neuron activation, and edges represent the relationships between these activations. The GNN component processes this graph to capture the complex dependencies and interactions between the features. The final output layer of the GNN provides the classification labels, making the model both efficient and interpretable by allowing backtracking of predictions to identify key contributing neurons.

Three benchmark datasets were used to evaluate the proposed methodology MNIST [26], CIFAR-10 [27], and Kaggle Cats and Dogs [28] dataset. Table 2 provides detailed description of the these datasets. The following evaluation metrics were employed to ass the model's performance: accuracy, precision, recall, and F1-score. The mathematical formulations for calculating these metrics are given in Table 3.

TABLE 2. Detaiied description of dataset used in this study.

Dataset	Description
Kaggle Cats and Dogs	The Kaggle Cats and Dogs dataset is a collection of images from a Kaggle competition. It contains 25,000 images of dogs and cats. The classes in this dataset are cat and dog. The images sizes range from 100 x 100 pixels to 2000 x 1000 pixels. 
MNIST	The Modified National Institute of Standards and Technology (MNIST) dataset is a large database of handwritten digits. It contains 60,000 training images and 10,000 testing images. The classes in this dataset are the digits from 0 to 9. Each image is 28 x 28 (784) pixels and each pixel is a grayscale integer between 0 and 255. 
CIFAR-10	The Canadian Institute for Advanced Research (CIFAR-10) dataset consists of 60,000 32x32 color images in 10 classes, with 6,000 images per class. The classes in this dataset are airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. This is RGB dataset. 

TABLE 3. Detailed descripiion of evaluation metrics used.

Metric	Mathematical Formulation
Accuracy	$\frac{\text{TP+TN}}{TP+TN+FP+FN}$
Precision	$\frac{\text{TP}}{TP+FP}$
Recall	$\frac{\text{TP}}{TP+FN}$
F1-Score	$2 * \frac{Precision * Recall}{Precision + Recall}$

B. RESULTS

1) MNIST DATASET

Table 4 outlines the performance of the proposed model on the MNIST handwritten digit dataset. The model achieved remarkable accuracy of 99.26%, correctly classifying nearly all test images. Both precision and recall metrics were also impressive at 99.26%, indicating the model's strong capability in accurately identifying correct digits while minimizing false identifications. The F1 score, which provides a balanced measure by taking the harmonic mean of precision and recall, was 99.2% - underscoring the model's overall excellent performance.

Table 5 further breaks down the model's performance on MNIST by providing class-wise precision, recall, and F1 scores. Across all digit classes from 0 to 9, the model demonstrated consistently high scores, with F1 scores ranging from 0.95 for digits 2, 5, and 6, to 0.99 for digits 1, 4, and 7. This granular analysis highlight the model's robust performance in recognizing the handwritten digits of all types.

2) CIFA-10 DATASET

The model is further evaluated on more complex CIFAR-10 dataset, which contains colour images across 10 different classes such as airplanes, automobiles, birds, and others. As shown in Table 5, the model achieved an accuracy of 87.37% on this dataset. The precision and recall were similarly high at 87.35% and 87.37% respectively, demonstrating the model's effectiveness in accurately classifying objects across multiple diverse categories while minimizing incorrect predictions. The F1 score of 87.31% indicates a balanced trade-off between precision and recall.

Table 8 presents a class-wise breakdown of the model's performance on CIFAR-10. Across object categories like airplanes, automobiles, birds, cats, dogs, and others, the F1 scores ranged from 0.88 for birds, dogs, and ships to 0.92 for automobiles, deer, and horses. This analysis provides insights into the model's nuanced capabilities in recognizing different types of objects.

3) CATS AND DOGS DATASET

On the Kaggle Cats and Dogs binary classification dataset, the model exhibited an accuracy of 91.44% as represented in Table 8. Precision and recall are well-balanced at 91.45% and 91.44% respectively, with an F1 score of 91.43%, presented in table 9. This demonstrates the model's proficiency in distinguishing between images of cats and dogs.

TABLE 4. Evaluation on MNIST dataset.

Metric	Value
Accuracy	99.266
Precision	99.268
Recall	99.266
F1 Score	99.266

TABLE 5. Class-wise performance evaluation on MNIST dataset.

Class/Metric	Precision	Recall	F1-Score
0	0.97	0.99	0.98
1	0.99	0.99	0.99
2	0.96	0.94	0.95
3	0.97	0.99	0.98
4	0.99	0.99	0.99
5	0.96	0.94	0.95
6	0.97	0.99	0.98
7	0.99	0.99	0.99
8	0.96	0.94	0.95
9	0.97	0.99	0.98

C. MODEL EXPLAINABILITY

The key advantage of the proposed architecture is its ability to provide explainable predictions. Figure 3 visually demonstrates how the model achieves this by highlighting the important features used for classification at step. The figure represents the images being processed in the first column. The second column shows visualized extracted convolutional features, with brighter yellow areas indicating higher importance for the final classification. This helps identify which regions of the images were most influential in the model's decision making process. The third column depicts on of the fully connected layers, providing a high level view of the initial computations and abstract activations. These activations are then transformed into a graph structure shown in the fourth columns, where reach node represents a neuron and edges illustrate connection between them. The percentage on the nodes indicate the activation level of each neuron, enabling interpretability into how the network arrives at its final output classification displayed in the last column. The step-by-step visualization of model's internal structures allows users to retrace the reasoning behind each prediction, providing insights that are often lacking in tradition black-box models.

Figure 3 in the document demonstrates the image classification process. The diagram shows how input images are processed through convolution layers to extract key features, depicted in the second column. These features are visually distinguished by their importance, with highly important features highlighted in bright yellow, while areas of least

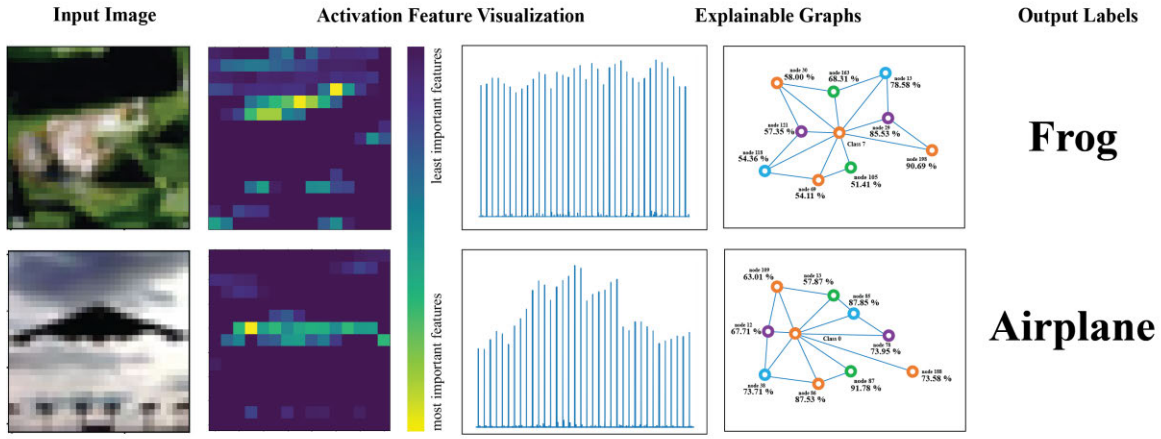


FIGURE 3. A step-by-step illustration of the explainable image classification process using our novel graph-based framework.

TABLE 6. Evaluation on CIFAR-10 dataset.

Metric	Value
Accuracy	87.37
Precision	87.35
Recall	87.37
F1 Score	87.31

TABLE 7. Class-wise performance evaluation on CIFAR 10 dataset.

Class/Metric	Precision	Recall	F1-Score
Airplane	0.91	0.89	0.90
Automobile	0.93	0.91	0.92
Bird	0.87	0.89	0.88
Cat	0.91	0.89	0.9
Deer	0.93	0.91	0.92
Dog	0.87	0.89	0.88
Frog	0.91	0.89	0.90
Horse	0.93	0.91	0.92
Ship	0.87	0.89	0.88
Truck	0.91	0.89	0.90

TABLE 8. Evaluation on Kaggle cats and dogs dataset.

Metric	Value
Accuracy	91.44
Precision	91.45
Recall	91.44
F1 Score	91.43

importance are shown in darker colours. This representation underscores the model's focus on certain features deemed essential for accurate classification.

In figure 3, the saliency maps highlight areas within images that are pivotal for the neural network's classification

TABLE 9. Class-wise performance evaluation on kaggle dataset.

Class/Metric	Precision	Recall	F1-Score
Cat	0.91	0.89	0.90
Dog	0.93	0.91	0.92

decisions. These visual representations are key to understanding which specific attributes of an image are most influential. For example, the saliency map of the frog image shows a prominent yellow region, indicating that certain features are critical for recognizing the frog. These characteristics likely differ significantly from other regions and images, aiding the model in distinguishing the frog from its surroundings. Similarly, in the airplane image, the focused region on the wing suggests that unique structural details are essential for the network to identify the aircraft. These saliency maps provide insight into the model's reasoning process, revealing how it discriminates between different objects and assigns relevance to particular image features, thus enhancing our understanding of how neural networks interpret and classify visual data.

Figure 4 represents the t-SNE visualizations of graph embeddings generated for each dataset, the top image represents for cat and dog dataset, while the middle image represents the graph embeddings generated for the MNIST dataset and final the bottom image describes the embeddings for CIFAR-10 dataset. The t-SNE represents the efficacy of graphs generated and provide insight into the separability using the proposed architecture. The t-SNE can be visually inspected to validate the fact that proposed architecture has learned to separate different classes in the lower-dimensional space. As the model has learned discriminative features, the clusters corresponding to different classes will be well-separated, thus making it easier to distinguish them.

Table 10 presents the results of ablation study of the proposed architecture across three datasets: Cats and Dogs, MNIST, and CIFAR-10. The study compares the performance

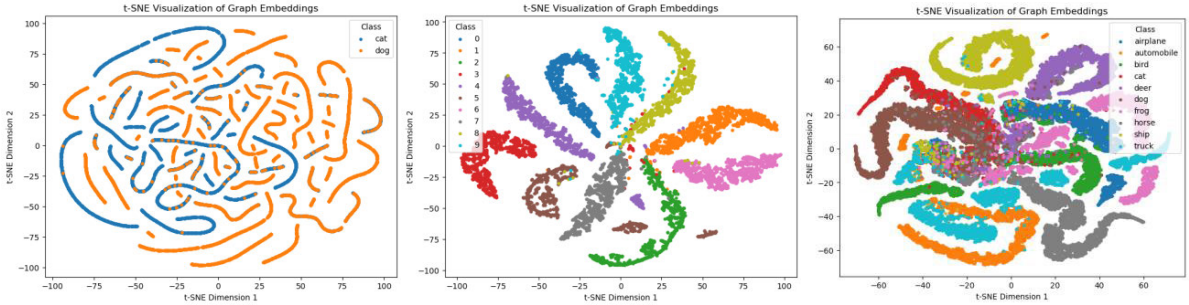


FIGURE 4. T-SNE visualizations of graph embeddings for (left) Kaggle dataset, (middle) MNIST dataset, and (Right) CIFAR-10 dataset.

TABLE 10. Ablation study of the proposed architecture.

Dataset	Model Variation	Accuracy	Precision	Recall	F1- Score
Cats and Dogs Dataset	Full Model	91.44	91.45	91.44	91.43
	Without GNN	90.87	90.84	90.86	90.85
MNIST Dataset	Full Model	99.27	99.27	99.27	99.27
	Without GNN	98.35	98.36	98.35	98.35
CIFAR-10 Dataset	Full Model	87.37	87.35	87.37	87.31
	Without GNN	85.65	85.64	85.63	85.64

of the full model, which integrates both CNNs and GNNs, against a variation of the model without the GNN component. The results demonstrate that the full model consistently outperforms the model without GNNs, while also highlighting the significant contribution of neuron activations in terms of overall accuracy, precision, recall, and F1-score. The consistent enhancement in scores across all datasets reinforces the ability of the GNN to capture and elucidate meaningful features, thereby facilitating more precise and interpretable classifications. This indicates that the integration of GNN into the model framework not only boosts overall accuracy but also enhances the consistency and reliability of the predictions, making the model more robust to variations in data.

The model integrates several components: convolutional neural networks for initial feature extraction, Neuron Activations for identifying significant neural responses, Graph Construction for representing these activations in a structured format, and graph neural networks for processing this graph data. We assessed the model's effectiveness across various classes using different combinations of these elements. The assessment was based on Accuracy, Precision, Recall, and F1 Score, thereby demonstrating the model's capability and its ability to provide explainable outcomes. The demonstrated robustness across multiple standard datasets, including MNIST and CIFAR-10, underscores our framework's potential for high performance in real-world scenarios.

IV. CONCLUSION

This paper proposed a novel graph neural network-based method for accurate classification, emphasizing explainability. It significantly improved the network performance by elucidating features and constructing graph connections derived from neuron activations tailored to each class, enhancing the interpretability of the classification process. The proposed graph neural network component models the complicated interaction between neuron activations. The proposed method achieved overall accuracy of 99.26%, 91.44%, and 87.37%, on MNIST, Cat and Dog, and CIFAR-10 datasets. Compared to existing methods, the proposed method not only delivers superior performance but also sheds light on the workings of its features and neuron activations, including their interconnections.

Future work aims to explore the adaptation of our framework to extreme variations in dataset quality and conditions, ensuring its effectiveness and reliability in real-world applications. Additionally, future research will focus on enhancing the generalizability of the graph neural network model by developing strategies to dynamically adjust the graph connections and feature elucidation.

REFERENCES

[1] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 886–893.

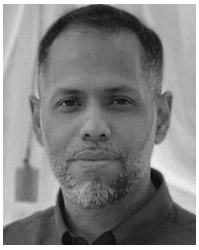
[2] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[4] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Sep. 2014, pp. 818–833.
[5] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
[6] Z. J. Wang, R. Turko, O. Shaikh, H. Park, N. Das, F. Hohman, M. Kahng, and D. H. P. Chau, "CNN explainer: Learning convolutional neural networks with interactive visualization," *IEEE Trans. Vis. Comput. Graphics*, vol. 27, no. 2, pp. 1396–1406, Feb. 2021.
[7] V. Chamola, V. Hassija, A. R. Sulthana, D. Ghosh, D. Dhingra, and B. Sikdar, "A review of trustworthy and explainable artificial intelligence (XAI)," *IEEE Access*, vol. 11, pp. 78994–79015, 2023.
[8] G. Singh and K.-C. Yow, "These do not look like those: An interpretable deep learning model for image recognition," *IEEE Access*, vol. 9, pp. 41482–41493, 2021.
[9] A. Holzinger, B. Malle, A. Saranti, and B. Pfeifer, "Towards multi-modal causability with graph neural networks enabling information fusion for explainable AI," *Inf. Fusion*, vol. 71, pp. 28–37, Jul. 2021.
[10] K. Mahmood, R. Mahmood, E. Rathbun, and M. van Dijk, "Back in black: A comparative evaluation of recent state-of-the-art black-box attacks," *IEEE Access*, vol. 10, pp. 998–1019, 2022.
[11] D. Wang, N. Honnorat, P. T. Fox, K. Ritter, S. B. Eickhoff, S. Seshadri, and M. Habes, "Deep neural network heatmaps capture Alzheimer's disease patterns reported in a large meta-analysis of neuroimaging studies," *IEEE Access*, vol. 12, pp. 1234–1245, 2023.
[12] A. Suwalska, J. Siuda, S. Kocot, W. Zmuda, M. Rudzinska-Bar, and J. Polanska, "Activation maps of convolutional neural networks as a tool for brain degeneration tracking in early diagnosis of dementia in Parkinson's disease based on magnetic resonance imaging," *IEEE Access*, vol. 17, pp. 4115–4121, 2023.
[13] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.
[14] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph neural networks: A review of methods and applications," 2018, *arXiv:1812.08434*.
[15] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 4–24, Jan. 2021.
[16] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lió, and Y. Bengio, "Graph attention networks," 2017, *arXiv:1710.10903*.
[17] Z. Liu, C. Chen, L. Li, J. Zhou, X. Li, L. Song, and Y. Qi, "GeniePath: Graph neural networks with adaptive receptive paths," 2018, *arXiv:1802.00910*.
[18] P. Cui, X. Wang, J. Pei, and W. Zhu, "A survey on network embedding," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 5, pp. 833–852, May 2019.
[19] Z. Liao, "Trainable activation function in image classification," 2020, *arXiv:2004.13271*.
[20] J. Park, J. Lee, and D. Jeon, "A 65 nm 236.5nJ/classification neuromorphic processor with 7.5% energy overhead on-chip learning using direct spike-only feedback," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2019, pp. 140–142.
[21] N. Passalis and A. Tefas, "Training lightweight deep convolutional neural networks using bag-of-features pooling," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 6, pp. 1705–1715, Jun. 2019.
[22] L. Taylor, A. King, and N. Harper, "Robust and accelerated single-spike spiking neural network training with applicability to challenging temporal tasks," 2022, *arXiv:2205.15286*.
[23] R. Wightman, H. Touvron, and H. Jégou, "ResNet strikes back: An improved training procedure in timm," Oct. 2021, *arXiv:2110.00476*.
[24] M. Shafiq and Z. Gu, "Deep residual learning for image recognition: A survey," *Appl. Sci.*, vol. 12, no. 18, p. 8972, Sep. 2022, doi: 10.3390/app12188972.
[25] J. Pranav and A. Sethi, "Vision Xformers: Efficient attention for image classification," Jul. 2021, *arXiv:2107.02239*.
[26] L. Deng, "The MNIST database of handwritten digit images for machine learning research [best of the web]," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 141–142, Nov. 2012.
[27] A. Krizhevsky et al.,, "Learning multiple layers of features from tiny images," Univ. Toronto, Toronto, ON, Canada, 2009.
[28] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar, "Cats and dogs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3498–3505.

BASIM AZAM received the master's degree in electrical engineering from the Institute of Space Technology, Pakistan, and the Ph.D. degree from Griffith University, his thesis examined "Deep Learning-Based Context Adaptative Architectures for Image Parsing." He is currently a Researcher in the field of computer vision, machine learning, and explainable AI. He is a Postdoctoral Research Fellow with the School of Computing and Information Systems (CIS), The University of Melbourne. He is also a Research Fellow with Griffith University, where he investigates explainability in AI algorithms. He has investigated and developed computer vision techniques for object detection, image segmentation, remote sensing, and medical imaging applications. He is passionate about developing solutions in the field of computer vision and AI. He was awarded Australian Research Council's Discovery Project Scholarship for his Ph.D. from Griffith University and the Merit Scholarship for his master's from the Institute of Space Technology.

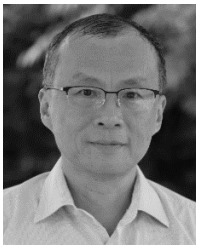
DEEPTHI P. KUTTICHIRA received the Ph.D. degree in machine learning (ML) from Deakin University. She has been a Postdoctoral Research Fellow in computer vision with the Institute for Intelligent and Integrated Systems, Griffith University, Brisbane, Australia. She is currently a Researcher, doing her postdoctoral research in explainable features for neural network models. She is interested in explainable artificial intelligence (XAI), fairness in ML (FairML), and optimization problems, and has a broad interest in computer vision problems.

BRIJESH VERMA is currently a Professor with Griffith University, Brisbane, Australia. He is the Chief Investigator on two ARC Discovery Projects: Deep learning architecture with context adaptive features for image parsing and A novel automatic neural network feature extractor with explanations and non-iterative learning. He has authored/co-authored/co-edited 13 books, including *Roadside Video Data Analysis: Deep Learning*, nine book chapters, and more than 200 articles in areas, such as neural networks, deep learning, evolutionary algorithms, pattern recognition, computer vision, image processing, digital mammography, and web information retrieval. His main research interests include computational intelligence and pattern recognition.

He was a member of Australian Research Council College of Experts. He is an Editorial Board Member of several international journals, including *Applied Soft Computing* and *Neural Computing and Applications*. He has been awarded many discovery and linkage research grants from Australian Research Council (ARC). He was the General Co-Chair of the International Joint Conference on Neural Networks (IJCNN 2023). He was the Chair of the IEEE Computational Intelligence Society's Queensland Chapter and under his leadership, the chapter won the Outstanding Chapter Award from IEEE CIS. He was an Associate Editor of IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS (TNNLS). He is the Editor-in-Chief of *International Journal of Computational Intelligence and Applications* (IJCIA).



ASHFAQUR RAHMAN (Senior Member, IEEE) received the Ph.D. degree from Monash University, Australia, in 2008. He is currently a Principal Research Scientist with Data61/CSIRO, Sandy Bay, Australia. He is the Leader of the Statistical Machine Learning Group, Data61/CSIRO. He leads a group of researchers who are enthusiastic about solving foundational machine-learning problems and applying them to solve real-world problems. He led multiple projects that provided data-driven solutions across numerous industries across Australia. He is involved in the organization of key workshops and conferences. He has published more than 100 peer-reviewed journal articles, book chapters, and conference papers. He supervises several Ph.D. students in collaboration with multiple Australian universities. He served as a reviewer for prestigious journals and conferences. He is an Associate Editor of *Information Processing in Agriculture* (Elsevier).



LIPO WANG (Senior Member, IEEE) received the bachelor's degree from the National University of Defense Technology, China, and the Ph.D. degree from Louisiana State University, USA. He is currently with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. He has more than 400 publications, a U.S. patent in neural networks, and a patent in systems. He has more than 14,000 Google Scholar citations, with an H-index of 52. He was a keynote speaker for more than 40 international conferences. His research interests include artificial intelligence for image and data processing. He was a member of the Board of Governors of the International Neural Network Society, the IEEE Computational Intelligence Society (CIS), and the IEEE Biometrics Council. He received the APNNA Excellent Service Award. He is the Co-Editor-in-Chief of *International Journal of Computational Intelligence and Applications* (IJCIA), a Senior Editor of IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS, and an Associate Editor of IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS. He is/was an Associate Editor/an Editorial Board Member of more than 30 international journals, including two other IEEE Transactions and a guest editor for more than ten journal special issues. He served as the CIS Vice President for Technical Activities, the Chair for the Emergent Technologies Technical Committee, and the Chair for the Education Committee of the IEEE Engineering in Medicine and Biology Society (EMBS). He was the President of the Asia-Pacific Neural Network Assembly (APNNA, recently renamed as APNNS–"Society"). He was the Founding Chair of the EMBS Singapore Chapter and CIS Singapore Chapter.