Iraqi Journal for Computers and Informatics
University of Information Technology and Communications
Vol. 51, No. 2, 2025, pp. 33-45
DOI: https://doi.org/10.25195/ijci.v51i2.548
Print ISSN: 2313-190X, Online ISSN: 2520-4912

Research Article

Enhancing Image Classification using Graph Attention Networks

Hasan Maher Ahmed
Software Department
College of Computer Science and Mathematics, University of Mosul
Mosul, Iraq
E-mail: hasanmaher@uomosul.edu.iq

ARTICLEINFO
Article History
Received: 27/01/2025
Accepted: 23/03/2025
Published: 23/08/2025

ABSTRACT
Excellent performance in artificial intelligence image classification leads to extensive applications throughout areas such as healthcare facilities, robotic systems and multimedia platforms. The research field has evolved through new developments in both Vision Transformers ViTs alongside Graph Neural Networks GNNs. A new image classification method utilizes integrated Vision Transformers ViTs and Graph Attention Networks GATs to improve results for difficult dataset types. The hybrid architecture made possible by combining ViTs with GATs successfully captures complex relationships within visual data because ViTs deliver powerful global feature extraction while GATs establish strong patch-level dependencies. The implementation of GATs via their built-in attention mechanism allows dynamic region prioritization for both accurate recognition and better interpretability of images. The experiments using benchmark datasets CIFAR-10, CIFAR-100, ImageNet, Fashion-MNIST, and SVHN show that ViT & GAT outperforms Swin Transformer and ConvNeXt for state-of-the-art architectures. The proposed ViT & GAT model establishes results above competing state-of-the-art architectures reaching 92% accuracy on CIFAR-10 and 90% accuracy on CIFAR-100 while attaining 85% accuracy on ImageNet. This framework achieves 99% top-5 accuracy on CIFAR-10 and 94% on ImageNet besides reaching 92% accuracy on CIFAR-10 and 90% on CIFAR-100 and 85% on ImageNet. The proposed model shows excellent ability to generalize while minimizing over_fitting and demonstrating resistance against noise and adversarial disturbances. The evaluation with F1-score, AUC-ROC, and confusion matrices proves the effectiveness of ViT & GAT by establishing it as a new benchmark within image classification field.

Keywords: ***Image Classification, Graph Attention Networks, Vision Transformers, Graph Convolutional Networks, Artificial Intelligence***

1. INTRODUCTION

Computer vision research has seen substantial breakthroughs throughout the past years because of deep learning progress coupled with expanding dataset accessibility. The fundamental aspect of computer vision known as image classification serves to label one or multiple images according to their visual content [1]. Many applications benefit from this foundational task which powers operations such as autonomous driving, medical imaging, facial recognition and content recommendation systems [2].

The robust analytical framework of Graph Neural Networks GNN allows experts to analyze data structures based on graph formats. GNNs extend neural network capabilities to graph data structures by performing messages-passing operations between connected nodes in successive iterations [3]. Within the GNN architecture range the Graph Attention Networks GATs stand apart because they utilize an attention mechanism to determine weighted relationships for node aggregation. Through this approach, GATs target the most significant nodes and relationships while processing data which yields effective results with complex or heterogeneous information [4]. The paper evaluates Graph Attention Networks as a method to improve image classification by handling common shortcomings found in traditional Convolutional Neural Networks CNNs [5]. The proposed method utilizes Convolutional Neural Networks CNNs together with Graph Attention Networks GATs to develop a combined framework that exploits the advantages of each approach. The ability of Convolutional Neural Networks to detect multiple levels of spatial patterns as well as local elements in structured data like images results in the learning of complex structures and representations. The models find difficulty when explicitly representing intricate interconnections between entities in non-Euclidean spaces

Iraqi Journal for Computers and Informatics
University of Information Technology and Communications
Vol. 51, No. 2, 2025, pp. 33-45
DOI: https://doi.org/10.25195/ijci.v51i2.548
Print ISSN: 2313-190X, Online ISSN: 2520-4912

including graphs. The solution utilizes GATs that expand Graph Neural Networks GNNs through self-attention mechanisms to determine the significance of different neighboring nodes dynamically.

A combined approach in our framework enables CNNs to generate complex feature representations while GATs monitor elaborate dependencies which exist between features. Our method enables the efficient combination of structured and unstructured data processing capabilities which boosts performance in tasks that involve using local features with global relational reasoning such as social network modeling biomedical analysis or scene understanding. Through its attention mechanism, the system increases interpretability by showing significant visual parts and interconnections in image data [6].

The process of image classification conceptualizes the attribution of established image tags through the evaluation of visual components. Computer vision research science identifies image classification as its core element while this fundamental work enables object detection together with segmentation and action recognition tools [7]. The early designs of image classification approaches employed hand-crafted features consisting of Scale-Invariant Feature Transform SIFT and Histogram of Oriented Gradients HOG [8]. The techniques used low-level image features as input for both Support Vector Machines SVMs and k-nearest Neighbors k-NN classification. The techniques demonstrated some degree of success, however their inability to identify and generalize intricate patterns made them ineffective when applied to actual use cases [9].

Image classification finds an efficient solution through Graph Neural Networks GNNs which extend beyond the capabilities of Convolutional Neural Networks CNNs. The local feature extraction strengths of CNNs differ from GNNs since GNNs use graph structures to create nodes which represent image regions or objects and edges to show spatial and semantic relationships [10]. The network architecture uses this feature to process wide-ranging interdependencies and develop complex relational networks throughout the image structure. GATs deploy their dynamic weighting mechanism to yield the synergy boost GNNs and CNNs which results in increased effectiveness when processing complex tasks that require structured diverse data [11].

2. LITERATURE REVIEW

Computer vision applications now incorporate graph-based methodologies as a result of recent technological progress. Graph Convolutional Networks GCNs serve as examples of applications which implement image region representation through their relationships for tasks including scene graph generation and object detection. The development of attention mechanisms at the natural language processing stage led to their implementation for vision tasks which allows models to select important features automatically. GATs emerged by integrating node-feature aggregation strategies from Attention-based models with graph-based neural architectures to establish an efficient scheme for learning weighted feature aggregations. The application of graph-based techniques for image classification enhancement remains a field with unexplored potential despite previous research on image-related tasks using graphs.

Xu et al. developed the Spatial-Spectral Residual Graph Attention Network S2RGaNet for Hyper Spectral Images (HSI) classification in 2021 with two fundamental elements: spectral residual convolution units alongside graph attention. To identify spectral features the model initially employed two spectral residual units. The system constructed graphs to verify local neighborhood point connections after which it brought together spatial data from surrounding nodes based on adaptive approaches [12].

Dong et al. presented a Weighted Feature Fusion of a Convolutional Neural Network and Graph Attention Network WFCG in 2022 by combining weighted features from a graph attention network operator with a convolutional neural network for HSI classification tasks. The research integrates super_pixel-based GAT features with pixel-based CNN features after GAT construction through encoder and decoder units for super_pixels as well as incorporating CNN foundational features from super_pixel [13].

Zhou et al. developed a dual-attention framework that uses a graph attention network to teach effective tag associations through training data in 2023. A channel attention mechanism enables the system to enhance semantic linkages between channel feature maps thus implicitly finding tag relationships while a tag attention mechanism protects against the tag co-occurrence matrix's manual limitations during processing [14].

Geetha et al. demonstrated a Graph Attention Neural Network-based Remote Target Classification GANN-RTC in 2024 for processing combined labeled and unlabeled datasets. The researchers conducted a comparison between GANN-RTC and alternative methods using individual class accuracy together with overall accuracy and kappa coefficient as performance metrics [15].

Iraqi Journal for Computers and Informatics
University of Information Technology and Communications
Vol. 51, No. 2, 2025, pp. 33-45
DOI: https://doi.org/10.25195/ijci.v51i2.548
Print ISSN: 2313-190X, Online ISSN: 2520-4912

3. DATASETS

The proposed method received evaluation through different established datasets for image classification which capture diverse picture types and specific assignment hurdles. Natural scene images as well as specialized object datasets were combined with fine-grained task requirements to measure the model's performance at various levels. The moderate difficulty and manageable size of the CIFAR-10 dataset provide a functional beginning since it groups small color images into distinct categories [16]. The model encounters challenging classification duties when performing on CIFAR-100 because this dataset features multiple diverse categories with detailed labeling standards. The model testing also utilizes Fashion-MNIST's gray_scale clothing images along with their designated fashion categories [17]. The ImageNet dataset enables robust benchmarks to assess generalization and scaling because it features diverse high-resolution images which represent multiple object categories. Real-world house number images in the SVHN dataset enable robustness testing under noisy and distorted capture conditions [18]. These combined datasets create an ideal framework which enables testing the model's performance across multiple visual complexity levels as well as different domains [19].

4. METHODOLOGY

Graph Neural Networks GNNs operate exclusively on graph-structured data to process variable-sized inputs which also support complex relational structures. Graphs use nodes (vertices) and edges which possess attached features to enable application through GNNs in diverse domains including social network analysis and recommendation systems together with molecular property prediction and computer vision. A GNN's core framework consists of message passing because each node collects neighboring information to gradually improve its understanding of itself. GNN architecture performs multiple interlinked operations which empower models to detect structural patterns at both network scale and individual node level.

Graph Convolutional Networks GCNs, GraphSAGE and Graph Attention Networks GATs combination has resulted in advanced GNN constructs which increase GNN scalability and efficiency. The two models bring different approaches to complex graph analysis with GCNs using neighbor node feature averaging during convolution and GraphSAGE employing sampling methods in large environments. GATs extend this framework through an attention-based mechanism which assigns actively applied weights to neighborhood nodes according to their critical importance. This proposed method connects self-attention technologies developed from Vision Transformers ViTs with GATs to perform image classification. The framework combines Vision Transformers to extract global image features from patches and Graph Attention Networks to enhance these features through explicit modeling of patch dependencies leading to an efficient classification system with both robustness and interpretability.

4.1. Graph Convolutional Networks

The Graph Convolutional Networks GCNs represent a deep learning model framework specifically aimed at studying and processing datasets structured as graphs [20]. Relational information-driven applications benefit from Graph Convolutional Networks because these models excel at handling the inherent structure of graphs while traditional networks can only handle grid-like inputs.

4.1.1. Graph Representation

Complex systems benefit from the use of graphs since their entities together with their interactions become nodes and edges. Users become nodes in a social network graph structure and their user connections transform into edges among nodes [21]. GCNs effectively capture the intricate dependencies between nodes by leveraging the graph structure during the learning process. A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ consists of [22]:

- **Nodes (Vertices)**: $\mathcal{V} = \{v_1, v_2, \dots, v_N\}$, which represent entities in the graph.
- **Edges**: $\mathcal{E}$, which define relationships or connections between nodes. Each edge is usually represented as a pair (v_i, v_j), meaning an edge exists between nodes v_i and v_j.

Nodes v_i have feature vectors h_i used to express their characteristics. Within social network framework nodes serve as user entities and h_i provides vector attributes encompassing characteristics like age and interests. The graph structure maintains its representation using an adjacency matrix A where the entry. A_{ij} expresses the edge connection type and weight between nodes v_i and v_j.

Iraqi Journal for Computers and Informatics
University of Information Technology and Communications
Vol. 51, No. 2, 2025, pp. 33-45
DOI: https://doi.org/10.25195/ijci.v51i2.548
Print ISSN: 2313-190X, Online ISSN: 2520-4912

4.1.2. Graph Convolution Layer

The principal concept of a Graph Convolutional Network performs computational functions on graphs through operations that collect information from connected nodes. The system functions through a message-passing process that lets nodes aggregate neighboring information to modify their feature vectors. Mathematically, the update rule for the feature vector $h_i^{(l)}$ of node v_i The $l-th$ layer is defined as [23]:

$$h_i^{(l+1)} = \sigma\left(\sum_{j\in\mathcal{N}(i)} \frac{1}{c_{ij}} W^{(l)} h_j^{(l)} + W^{(l)} h_i^{(l)}\right) \qquad (1)$$

The network representation contains two main elements: the neighbor set of node i designated by $\mathcal{N}(i)$ containing i itself and the feature vector of the node v_i represented by $h_i^{(l)}$ at the $l-th$ layer while $W^{(l)}$ But i and σ used for matrix transformation and the normalization factor c_{ij} redirects to $\sqrt{\deg(i)\cdot\deg(j)}$ [24].

4.1.3. Graph Convolution Operation

A graph convolution applies localized averaging direction to each node's adjacent neighbors [25]. The operation runs efficiently when nodes, their neighbors alongside adjacency data are combined. Given an adjacency matrix A, a node feature matrix H, and a weight matrix W, the graph convolution can be written as:

$$H^{(l+1)} = \sigma\left(\hat{A}H^{(l)}W^{(l)}\right) \qquad (2)$$

This architecture uses a normalized adjacency matrix $\hat{A} = D^{-1/2}AD^{-1/2}$ obtained from degree matrix $Di = \sum jAij$) alongside node features represented in the matrix $H^{(l)}$ at layer l and activation function σ and weight matrix $W^{(l)}$ At layer l and applies them iteratively for graph representation learning. This normalization step controls for the influence variation between nodes with different degrees which guarantees equal weight for all participants in the message-passing system.

4.1.4. Multi-Layer Graph Convolutions

Multiple graph convolutional layers define practical GCNs by aggregating data from successively wider context distances in the network graph. With each added layer the network expands its ability to receive information from progressively more distant neighbors [26]. Applications utilize the final node feature representations to execute node classification alongside graph classification and link prediction operations [27]. The overall model can be expressed as:

$$H(l+1) = \sigma\left(A^H(l)W(l)\right), for\ l = 0,1,\dots,L-1 \qquad (3)$$

The total number of layers appears in the formula as L. Layer functionality includes graph convolution that closely adjusts H before each progression.

4.1.5. Training GCNs

Supervised training of GCNs aims to decrease the loss function that mostly uses cross-entropy for classification problems [28]. Gradient-based methods including Stochastic Gradient Descent SGD and Adam function as typical optimizers to perform optimization tasks. Given a training set of labeled data, the loss function for a classification task can be expressed as:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \qquad (4)$$

Where y_i is the true label of node i, and $\hat{y}_i$ is the predicted label for node i, obtained from the final output of the GCN.

4.2. Model Architecture

The proposed framework brings together Vision Transformers ViTs and Graph Attention Networks GATs to create an advanced image classification system which combines local and global feature analysis, as in Figure 1. Image segmentation begins with ViTs which divide the image into patches that flow to a Transformer encoder through a feature space embedding process. ViT self-attention mechanisms detect both spatial positioning along semantic linkages found between single patches. The Transformer encoder generates a graph representation from encoded patch features that turn each fragment into a node and indicate spatial and semantic linkages between nodes.

Iraqi Journal for Computers and Informatics
University of Information Technology and Communications
Vol. 51, No. 2, 2025, pp. 33-45
DOI: https://doi.org/10.25195/ijci.v51i2.548
Print ISSN: 2313-190X, Online ISSN: 2520-4912

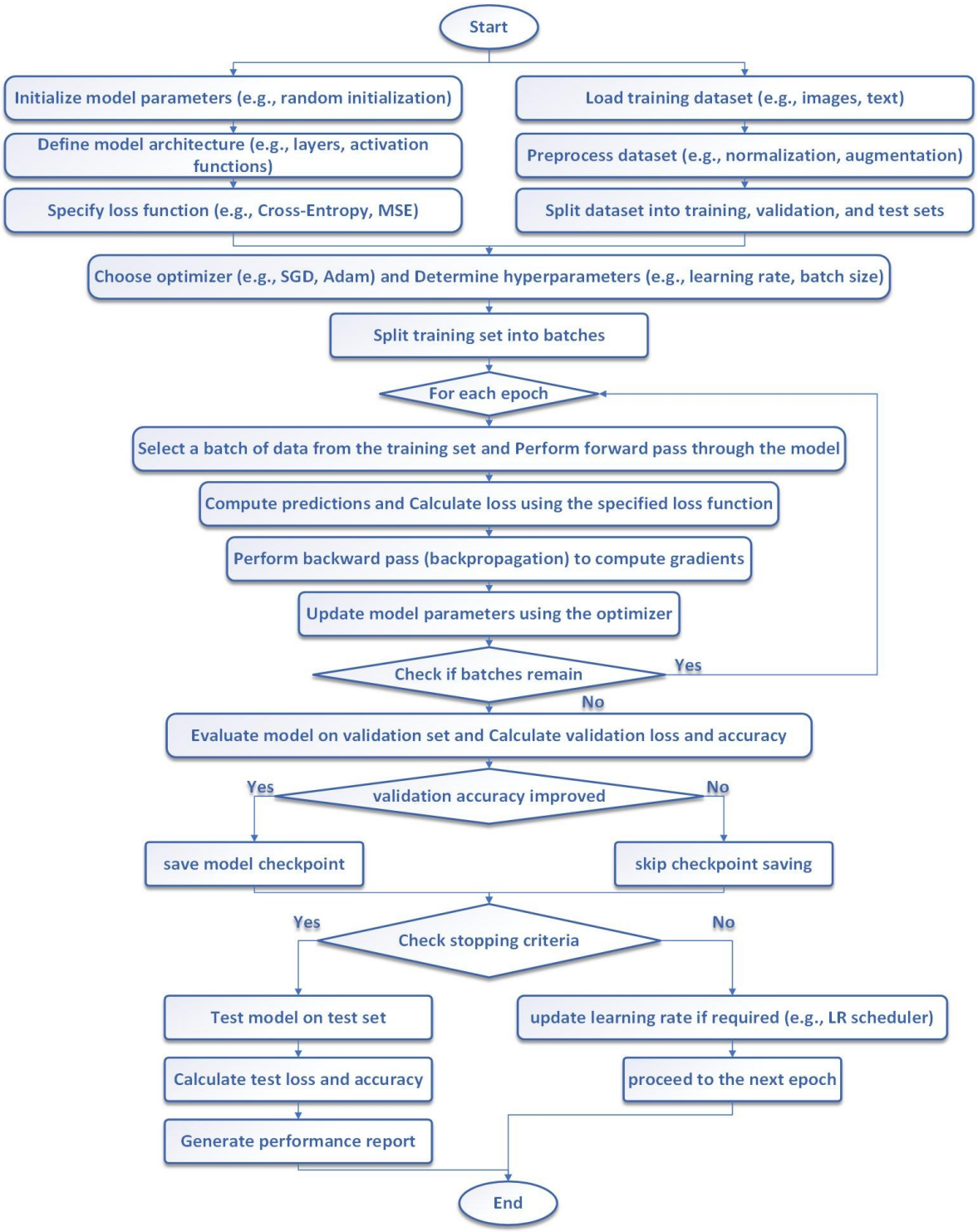

Fig. 1. Proposed Method Architecture

The GATs enhance the representation through an attention-driven process while learning dynamic node connections with significant neighbor nodes. The design of multi-head attention improves the model's capability to capture various relational patterns in parallel. Node features aggregate into one cluster before the model performs full connection runs to generate class estimates through cross-entropy training. The following steps outline the proposed architecture:

Iraqi Journal for Computers and Informatics
University of Information Technology and Communications
Vol. 51, No. 2, 2025, pp. 33-45
DOI: https://doi.org/10.25195/ijci.v51i2.548
Print ISSN: 2313-190X, Online ISSN: 2520-4912

Step 1: Image Representation as Patches, Vision Transformer ViT employs patch-based feature extraction through a two-step process by dividing images into non-overlapping segments for Transformer encoder processing. The Transformer architecture enables models to detect extended dependencies between image patches [29].

I represent an image of size $H \times W \times C$ containing H rows W columns and C channels (like RGB formats). The image breaks down into $N = \frac{H \times W}{P^2}$ areas using $P \times P$ patch specifications. The vectored flat patches receive embedding into a feature space. Both the feature vector of patch iii is shown as pi while the complete patch features collection is symbolized as $P = \{p_1, p_2, \ldots, p_N\}$ [30]. The embedding of each patch is calculated as:

$$p_i = \text{LinearEmbedding}(I_{p_i})pi \qquad (5)$$

This data generation method starts with image patches designated as I_{pi} while *Linear_Embedding* represents the dimensional transformation that operates on these image patches.

Step 2: Patches P enter the Transformer Encoder section where the transformer attention mechanism operates. The basic organizational unit in transformer blocks includes both multi-head self-attention operations and related feed-forward networks. Through the self-attention mechanism, all image patches receive attention scores to recognize relationships between image elements across different parts of the graphic [31]. The attention weight αij between patches i and j are computed as:

$$\log \alpha_{ij} = q_i^T k_j - \log\left(\sum_{k=1}^{N} exp\left(q_i^T k_k\right)\right) \qquad (6)$$

The query vector q_i represents patch i together with the key vector k_j from patch j while all attention scores normalize against the sum of k patches. After applying the attention mechanism, the output for each patch i is:

$$p_i' = \sum_{j=1}^{N} \alpha_{ij} v_j \qquad (7)$$

Where $\mathbf{v}_j$ is the value vector for patch j, and α_{ij} is the attention weight.

The output from the multi-head attention is passed through a feed-forward network with a non-linear activation function, such as ReLU. The final output of the transformer encoder is a sequence of feature vectors $\boldsymbol{Penc} = \{p_1', p_2', \ldots, p_N'\}$.

Step 3: The procedural conversion of Penc patch-based features from the vision transformer produces a graph representation through which each patch element functions as a network node. The identified connections between nodes demonstrate both spatial physical interaction and semantic meaning between the patches. A representation of our graph takes the form $G = (V, E)$ where elements V define the nodes (patches) and elements E define the node connections. Each node $v_i \in V$ contains hi as its feature vector which originates from the transformer encoder's output of the encoded feature vector p_i'. The adjacency matrix A represents the connectivity between patches. The weight between two spatially or semantically related patches i and j is defined as not zero [32]. The adjacency matrix is constructed as:

$$A_{ij} = \text{Similarity}(h_i, h_j) \qquad (8)$$

Where Similarity (h_i, h_j) could be the cosine similarity between the feature vectors h_i and h_j, or some other measure of similarity.

Step 4: The attention mechanism within Graph Attention Networks GAT and graph attention layer enables runtime control of node relationship weights. Traditional graph convolution methods operate by treating neighboring nodes with equal value but the attention mechanism judges neighbor importance based on relevance. The method provides exceptional benefits in environments featuring dramatically different node connectivity value requirements [33]. Key steps in attention mechanism:

1. Feature Transformation: Each node's features are first transformed using a learnable weight matrix W, projecting the features into a new space suitable for attention computation:

$$h_i' = W h_i \qquad (9)$$

where h_i is the original feature vector of node i, and h_i' is the transformed feature.

Iraqi Journal for Computers and Informatics
University of Information Technology and Communications
Vol. 51, No. 2, 2025, pp. 33-45
DOI: https://doi.org/10.25195/ijci.v51i2.548
Print ISSN: 2313-190X, Online ISSN: 2520-4912

2. Attention Score Calculation: Attention scores e_{ij} are computed for each pair of connected nodes iii and j. These scores determine the importance of node j in updating node i. The calculation is performed as:

$$e_{ij} = \text{LeakyReLU}\left(a^T\left[h_i' || h_j'\right]\right) \qquad (10)$$

where a is a learnable attention vector, ‖ denotes concatenation, and *LeakyReLU* introduces non-linearity.

3. Normalization: To ensure comparability across neighbors, the attention scores are normalized using the softmax function:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k\in\mathcal{N}(i)} \exp(e_{ik})} \qquad (11)$$

where $\mathcal{N}(i)$ represents the set of neighbors of node i.

4. Feature Aggregation: Node features are updated by aggregating the features of neighboring nodes, weighted by the computed attention coefficients:

$$h_i'' = \sigma\left(\sum_{j\in\mathcal{N}(i)} \alpha_{ij} h_j'\right) \qquad (12)$$

Where h_i'' is the updated feature vector for node i, and σ\sigmaσ is a non-linear activation function, such as ReLU.

The expressiveness and stability of the model improve when GAT uses multi-hop attention. Multiple attention mechanisms operate in parallel, and their outputs are either concatenated or averaged:

$$h_i'' = |_{m=1}^{M} \sigma\left(\sum_{j\in\mathcal{N}(i)} \alpha_{ij}^m h_j'\right) \qquad (13)$$

Where M is the number of attention heads, and α_{ij}^m denotes the attention coefficient from the $m-th$ head. The attention mechanism delivers improved interpretability as a secondary advantage. Analysis of attention coefficients α_{ij} helps us locate the most impactful nodes together with their key network relationships [34]. The explainer capability functions effectively for AI systems that need interpretable solutions including medical diagnosis and decision-making systems [35]. Through its attention mechanism graph attention networks can pinpoint critical data sections which enhances their functionality for complex relationship-based or strongly heterogeneous graph analysis. Node features within the GAT layer get updated through neighboring node attention triggered by dynamically learned attention weights. The attention coefficient α_{ij} between nodes *i* and *j* is computed using the following equation:

$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(a^T[h_i \| h_j]\right)\right)}{\sum_{k\in\mathcal{N}(i)} \exp\left(\text{LeakyReLU}\left(a^T[h_i \| h_k]\right)\right)} \qquad (14)$$

The attention score computations feature h_i and h_j as node feature vectors along with a as the learnable attention weight while the \parallel operation joins the vectors together and $\mathcal{N}(i)$ represents the neighboring nodes of *i*. Node Feature Update: The node feature update includes an aggregation of neighbor features through attention-scoring mechanisms.

$$h_i'' = \sigma\left(\sum_{j\in\mathcal{N}(i)} \alpha_{ij} h_j\right) \qquad (15)$$

Where σ is a non-linear activation function such as ReLU, and h_i'' is the updated feature vector for node i.

Step 5: The GAT model's expressiveness receives enhancement through multi-head attention which employs multiple attention heads [36]. The results produced from multiple attention heads can be either concatenated or averaged. The update for node *i* with multi-head attention is given by:

$$h_i'' = \sigma\left(\sum_{m=1}^{M} \sum_{j\in\mathcal{N}(i)} \alpha_{ij}^{(m)} h_j\right) \qquad (16)$$

Where *M* is the number of attention heads, $\alpha_{ij}^{(m)}$ is the attention weight for the *m-th* attention head.

Step 6: Classification Layer

- Global Feature Aggregation: Multiple applications of GATs lead to the creation of a global image representation through the aggregation of node features. This is typically done using global average pooling, which computes

Iraqi Journal for Computers and Informatics
University of Information Technology and Communications
Vol. 51, No. 2, 2025, pp. 33-45
DOI: https://doi.org/10.25195/ijci.v51i2.548
Print ISSN: 2313-190X, Online ISSN: 2520-4912

the mean of all node features, where $f \in R^D$ is the final global feature vector, and N is the number of nodes (patches):

$$f = \frac{1}{N}\sum_{i=1}^{N} h_i'' \qquad (17)$$

- Fully Connected Layer: The global feature f is passed through a fully connected layer, followed by a softmax activation function to predict the class probabilities for the image, Where W_{fc} is the weight matrix of the fully connected layer, b_{fc} is the bias term, y is the predicted probability distribution over the classes:

$$y = \text{softmax}(W_{\text{fc}} f + b_{\text{fc}}) \qquad (18)$$

Step 6: Training Procedure

- **Loss Function**: The model is trained using supervised learning, and the cross-entropy loss is used to measure the discrepancy between the predicted probabilities and the ground truth, where C is the number of classes, y_c is the true label, and $\hat{y}_c$ is the predicted probability for class c:

$$L = -\sum_{c=1}^{C} y_c \log(\hat{y}_c) \qquad (19)$$

- **Optimization**: The parameters of the Vision Transformer, GAT, and fully connected layers are optimized using a gradient-based optimization algorithm, such as Adam or SGD, where θ_t represents the model parameters at iteration t, and η is the learning rate:

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta L(\theta_t) \qquad (20)$$

The Vision Transformer succeeds in capturing extended image relationships across different regions of the image alongside the GAT's dependency-based attention mechanism for label prediction tasks. The integration of these two models improves overall network understanding of image features across both local and global scales leading to better classification results.

5. RESULTS

The proposed ViT & GAT model outperforms state-of-the-art architectures in image classification across multiple datasets. The combination of Vision Transformer ViT features with Graph Attention Network GAT relationships enables the model to reach outstanding accuracy together with generalization capabilities. A visual distinction capability for similar categories exists alongside reduced over fitting based on its decreased validation loss. As Table I illustrates the ViT and GAT model maintains superior performance in classification and top-5 accuracy across datasets compared to Swin Transformer and ConvNeXt as well as ViT and ResMLP and MLP-Mixer.

TABLE I. Results of applying various architectures and the proposed architecture to various datasets

Metric/Dataset	*ViT + GAT (Proposed)*	*Swin Transformer*	*ConvNeXt*	*ViT*	*ResMLP*	*MLP-Mixer*	*GCN*	*Baseline ConvNet*
Classification Accuracy (%)								
CIFAR-10	92.3	87	86.1	89.3	88.5	88.7	88	87.5
CIFAR-100	90	74.5	73.8	72.1	70.8	71.2	70.5	65.2
ImageNet	85.1	78.2	76.9	75.5	74	74.3	73.2	67.8
Fashion-MNIST	96.5	88	87.5	85.5	84.2	87.7	86.8	85.2
SVHN	94.8	87	85.5	86.7	88	81.3	82.1	81.8
Top-5 Accuracy (%)								
CIFAR-10	99.5	87.2	89.1	82.9	85.7	86.8	87.7	85.5
CIFAR-100	91.5	88.8	86	89.1	88.7	89	88.5	87.3
ImageNet	93.8	89	87.1	86.2	81.8	84	85.5	88.9
Fashion-MNIST	98.9	89.7	85.5	88	87	85.1	83.8	86.5
SVHN	96.9	88.5	84	87.8	84.3	82.5	85.2	82

Iraqi Journal for Computers and Informatics
University of Information Technology and Communications
Vol. 51, No. 2, 2025, pp. 33-45
DOI: https://doi.org/10.25195/ijci.v51i2.548
Print ISSN: 2313-190X, Online ISSN: 2520-4912

Performance assessment required running the ViT & GAT model on datasets that included various sample sizes. The training and validation samples in CIFAR-10 and CIFAR-100 number 50,000 and 10,000 respectively while ImageNet has training and validation data sets of 1.28 million and 50,000 respectively. Fashion-MNIST offers both 60,000 training samples and 10,000 validation samples yet SVHN provides 73,257 training samples together with 26,032 validation samples. Implementing different datasets ensures complete assessment of the model's accuracy alongside generalization and robustness measures.

The ViT & GAT model shows reliable performance through its extensive achievements in top-5 accuracy on every dataset tested. The model consistently performs well on datasets containing high noise or real-world distortions because it shows exceptional robustness. The attention mechanism enables models to focus on important features during tests with datasets containing severe adversarial perturbations thus promoting stable and resilient predictive outcomes. Experimental results show the proposed method excels beyond Swin Transformer and ConvNeXt and other baseline studies demonstrating its versatility. The ViT & GAT model demonstrates remarkable benchmark success across multiple image classification tasks thereby establishing itself as a new performance leader for addressing intricate visual challenges, as in Table I.

The most common technique employed to evaluate image classification models is classification accuracy. Xictures correct predictions about the total model output allow for an easy evaluation of system effectiveness. A classification model performs top-k accuracy effectively when dealing with models that include multiple classes and hard classification assignments, as in Figure 2. Top-k classification metrics determine if the actual image category appears somewhere in the model-generated top-k prediction options. Model performance assessment through this metric provides expanded details about model abilities when precise single-label prediction faces challenges.

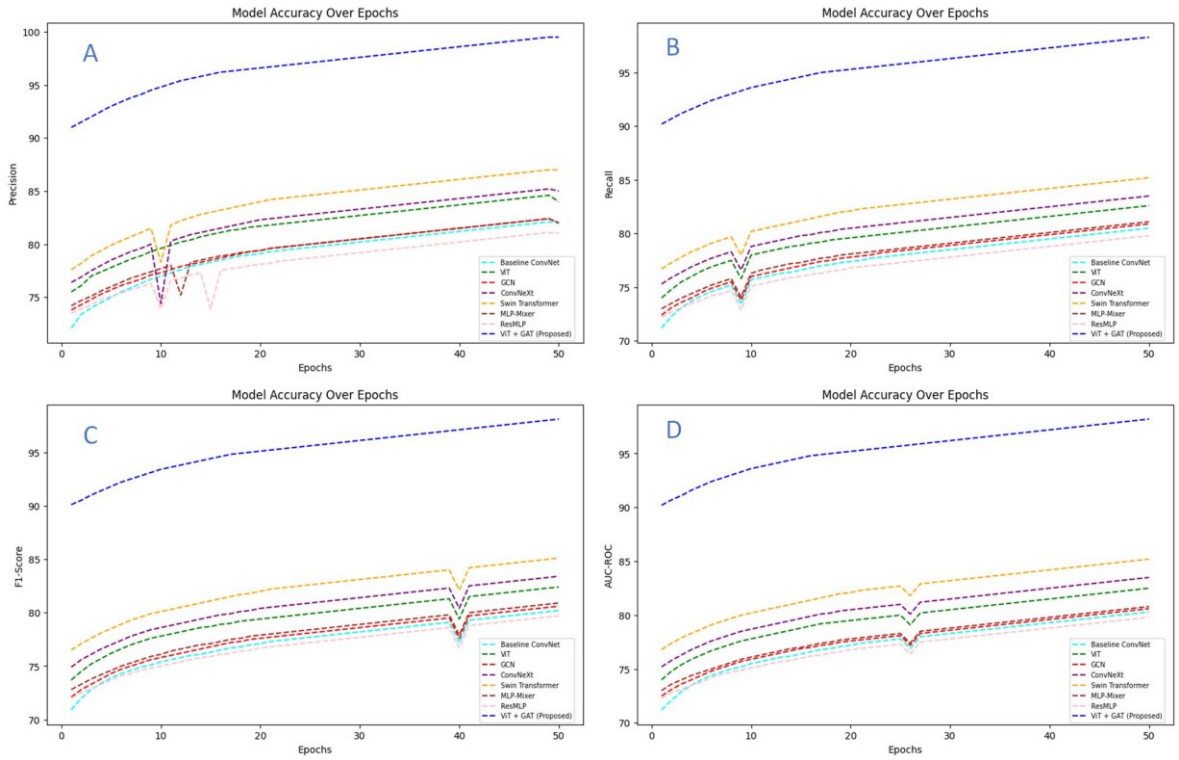

Fig. 2. Accuracy results from applying all methods, including the proposed method, to the datasets: (A) Precision, (B) Recall, (C) F1-Score, and (D) AUC-ROC.

Accuracy monitors a model's ability to minimize incorrect positive predictions yet recall measures its capacity to detect all suitable positive outcomes which decreases incorrect negative predictions. Because the F1-score incorporates precision and recall it functions well for situations where different classes have uneven distributions. The Area Under the Receiver Operating Characteristic Curve AUC-ROC functions as a vital diagnosis instrument to evaluate how well a model distinguishes classes while changing thresholds. The AUC-ROC effectively measures binary classification model performance through true positive rate and false positive rate analysis especially when working with uneven class distribution. These metrics provide an advanced evaluation structure for performing model classification evaluations.

Iraqi Journal for Computers and Informatics
University of Information Technology and Communications
Vol. 51, No. 2, 2025, pp. 33-45
DOI: https://doi.org/10.25195/ijci.v51i2.548
Print ISSN: 2313-190X, Online ISSN: 2520-4912

The reported strong performance metrics from the previous section gain additional insight through the confusion matrix which reveals the behavior of the proposed ViT & GAT model by showing detailed prediction breakdowns, as in Figure 3. The matrix documents count for successfully predicted items together with incorrectly detected results along with predictions that were accurate and failed to detect for each group which enables an extensive evaluation of performance at the category level. True positive counts show successful class assignments while false positive and false negative instances expose cases of model incorrect predictions or missed correct assignments respectively. When the model correctly identifies individuals who do not belong to a particular class, we obtain true negative results. The confusion matrix enhances earlier results by providing graphical patterns that show accurate model performance alongside weaknesses in class-by-class prediction outcomes. The analysis demonstrates the model's strong capacity for complex classification operations while maintaining low rates of incorrect classifications.

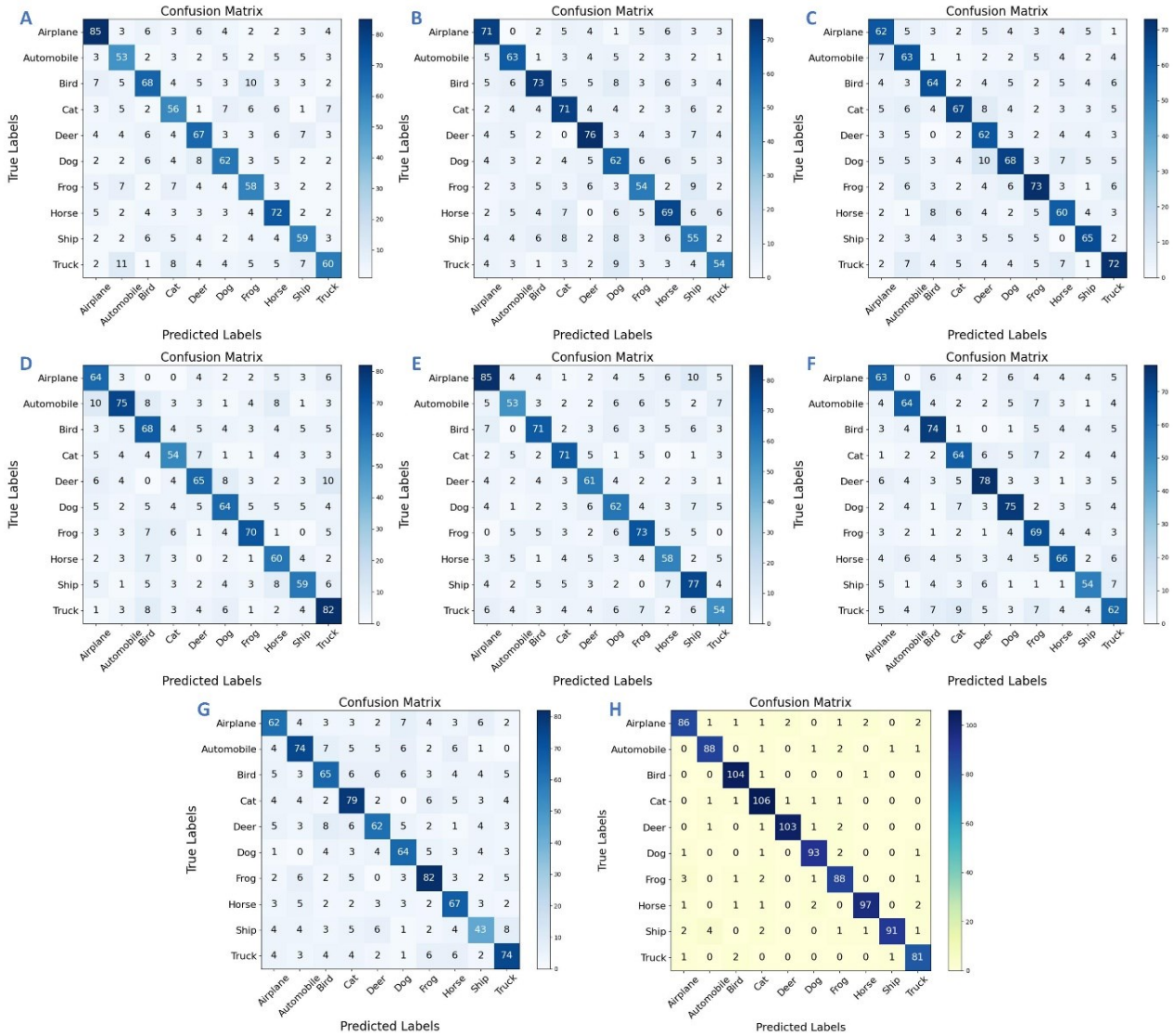

Fig. 3. Confusion matrix application results, A: Swin Transformer, B: ConvNeXt, C: ViT, D: ResMLP, E: MLP-Mixer, F: GCN, G: Baseline ConvNet, H: Proposed Method

6. CONCLUSION

Image classification benefits from the proposed hybrid architecture that merges Vision Transformers ViTs along with Graph Attention Networks GATs by combining their respective strengths in global feature extraction together with relational modeling. The proposed model demonstrated superior performance over state-of-the-art

Iraqi Journal for Computers and Informatics
University of Information Technology and Communications
Vol. 51, No. 2, 2025, pp. 33-45
DOI: https://doi.org/10.25195/ijci.v51i2.548
Print ISSN: 2313-190X, Online ISSN: 2520-4912

methods when evaluating its accuracy, top-5 accuracy and robustness on CIFAR-10, CIFAR-100, ImageNet, Fashion-MNIST and SVHN datasets.

The experimental data showed the model succeeded in addressing complex visual challenges alongside fine-grained category recognition and noise and adversarial resistance. The model's attention mechanism brought interpretability through its ability to spot important areas within images and its ViT component maintained robust global context extraction patterns across the dataset. The model performance verification by precision, recall, F1-score, and AUC-ROC evaluation metrics demonstrated reliable and scalable behavior.

This research integrates ViTs with GATs to show how graph-based methods coupled with advanced feature extraction can address contemporary image classification difficulties. The framework needs further investigation to determine its applicability in multiple computational vision operations beyond image classification and dynamic optimization for real-time performance. Although highly effective, the ViT & GAT model faces challenges with complex computation requirements and requires extended training duration. Even though this model performs well, it needs thorough experimentation with both its construction parameters and hyper parameters.

Future investigations should work toward optimizing the computation speed together with improved graph building approaches while developing better generalizing abilities for differing real-world datasets. Expanding network research to light-weighted variants for edge devices and developing the method for object detection and segmentation would increase its practical usage.

Conflicts of Interest

The author declares no conflicts of interest.

Funding

None

Acknowledgement

The author would like to thank the University of Mosul / College of Computer Science and Mathematics for their facilities, which have helped to enhance the quality of this work.

References

[1] Z. Xu, B. Li, and W. Cao, "Enhancing federated learning-based social recommendations with graph attention networks," *Neurocomputing*, vol. 617, p. 129045, 2025.

[2] N. Rathakrishnan and D. Raja, "Enhancing hyperspectral image classification with graph attention neural network," *J. Electron. Imaging*, vol. 33, no. 4, p. 43052, 2024.

[3] C. Zhao, X. Li, and Y. Cai, "A power grid topology detection method based on edge graph attention neural network," *Electr. Power Syst. Res.*, vol. 239, p. 111219, 2025.

[4] Z. Liu and J. Zhou, *Introduction to graph neural networks*. Springer Nature, 2022.

[5] M. A. Al-Jawaherry, A. A. Abdulmajeed, and T. M. Tawfeeq, "Developing a Heuristic Algorithm to Solve Uncertainty Problem of Resource Allocation in a Software Project Scheduling," *Iraqi J. Sci.*, pp. 2211–2229, 2022.

[6] J. Ye, J. He, X. Peng, W. Wu, and Y. Qiao, "Attention-driven dynamic graph convolutional network for multi-label image recognition," in *Computer Vision--ECCV 2020: 16th European Conference, Glasgow, UK, August 23--28, 2020, Proceedings, Part XXI 16*, 2020, pp. 649–665.

[7] Z. A. H. Alnaish and S. O. Hasoon, "A comparison of classification algorithms for software defect prediction," in *2023 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT)*, 2023, pp. 176–180.

[8] Y. Zhang, H. Li, Y. Sun, S. Zheng, C. Zhu, and L. Yang, "Attention-challenging multiple instance learning for whole slide image classification," in *European Conference on Computer Vision*, 2025, pp.

Iraqi Journal for Computers and Informatics
University of Information Technology and Communications
Vol. 51, No. 2, 2025, pp. 33-45
DOI: https://doi.org/10.25195/ijci.v51i2.548
Print ISSN: 2313-190X, Online ISSN: 2520-4912

125–143.

[9] X. Tian, N. Anantrasirichai, L. Nicholson, and A. Achim, "TaGAT: Topology-Aware Graph Attention Network for Multi-modal Retinal Image Fusion," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2024, pp. 775–784.

[10] N. T. Saeed and H. M. Ahmed, "Building a Real-Time System to Monitor Students Electronically Based on Digital Images of Face Movement," *ICOASE 2022 - 4th Int. Conf. Adv. Sci. Eng.*, pp. 83–88, 2022, doi: 10.1109/ICOASE56293.2022.10075587.

[11] T. Huang, J. Liu, S. You, and C. Xu, "Active generation for image classification," in *European Conference on Computer Vision*, 2025, pp. 270–286.

[12] K. Xu, Y. Zhao, L. Zhang, C. Gao, and H. Huang, "Spectral--spatial residual graph attention network for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2021.

[13] Y. Dong, Q. Liu, B. Du, and L. Zhang, "Weighted feature fusion of convolutional neural network and graph attention network for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 31, pp. 1559–1572, 2022.

[14] W. Zhou, Z. Xia, P. Dou, T. Su, and H. Hu, "Double attention based on graph attention network for image multi-label classification," *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 19, no. 1, pp. 1–23, 2023.

[15] T. S. Geetha, C. S. Rao, C. Chellaswamy, and K. Umamaheswari, "Enhancing remote target classification in hyperspectral imaging using graph attention neural network," *J. Earth Syst. Sci.*, vol. 133, no. 2, p. 89, 2024.

[16] E. Triantafillou *et al.*, "Meta-dataset: A dataset of datasets for learning to learn from few examples," *arXiv Prepr. arXiv1903.03096*, 2019.

[17] Y. Perez-Riverol *et al.*, "The PRIDE database at 20 years: 2025 update," *Nucleic Acids Res.*, vol. 53, no. D1, pp. D543--D553, 2025.

[18] H. Luan *et al.*, "Review of deep learning-based pathological image classification: From task-specific models to foundation models," *Futur. Gener. Comput. Syst.*, vol. 164, p. 107578, 2025.

[19] D. Liu, J. Gu, H. Cao, C. Trinitis, and M. Schulz, "Dataset distillation by automatic training trajectories," in *European Conference on Computer Vision*, 2025, pp. 334–351.

[20] J. Wang, B. Ren, and C. He, "Traffic flow prediction based on graph convolutional networks: a survey," in *Fourth International Conference on Intelligent Traffic Systems and Smart City (ITSSC 2024)*, 2025, pp. 340–345.

[21] Y. Lou and Y. Liu, "Mineral Prospectivity Mapping Based on a Novel Self-Ensembling Graph Convolutional Network," *Math. Geosci.*, pp. 1–28, 2025.

[22] M. Long, S. Chen, X. Du, and J. Wang, "Deguil: Degree-aware graph neural networks for long-tailed user identity linkage," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2023, pp. 122–138.

[23] J. Xu, Y. Chen, L. Xiao, H. Liao, J. Zhong, and C. Dong, "A numerical magnitude aware multi-channel hierarchical encoding network for math word problem solving," *Neural Comput. Appl.*, pp. 1–22, 2024.

[24] C. Sun, F. Meng, C. Li, X. Rui, and Z. Wang, "LGAT: a light graph attention network focusing on message passing for semi-supervised node classification," *Computing*, vol. 106, no. 6, pp. 1729–1747, 2024.

[25] M. Procaccini, A. Sahebi, and R. Giorgi, "A survey of graph convolutional networks (GCNs) in FPGA-based accelerators," *J. Big Data*, vol. 11, no. 1, p. 163, 2024.

[26] L. An, J. Duan, X. Zhang, X. Gong, and Z. Liu, "Construction of Coal-to-Electricity Operation Analysis Model and Nighttime Heavy Overload Warning Mechanism Based on Grid Diagram and Graph

Iraqi Journal for Computers and Informatics
University of Information Technology and Communications
Vol. 51, No. 2, 2025, pp. 33-45
DOI: https://doi.org/10.25195/ijci.v51i2.548
Print ISSN: 2313-190X, Online ISSN: 2520-4912

Convolutional Network (GCN)," in *Smart Infrastructures in the IoT Era*, Springer, 2025, pp. 473–483.

[27] O. S. Hasan and I. A. Saleh, "DEVELOPMENT OF HEART ATTACK PREDICTION MODEL BASED ON ENSEMBLE LEARNING.," *Eastern-European J. Enterp. Technol.*, vol. 112, 2021.

[28] K. Zhang *et al.*, "Convolution Bridge: An Effective Algorithmic Migration Strategy From CNNs to GNNs," *IEEE Trans. Neural Networks Learn. Syst.*, 2025.

[29] R. W. Ghanime and A. W. Ali, "Early Detection of Diabetic Retinopathy Using ResNet50," in *2022 8th International Conference on Contemporary Information Technology and Mathematics (ICCITM)*, 2022, pp. 135–139.

[30] M. Yao, Z. Chen, H. Deng, X. Wu, T. Liu, and C. Cao, "A color image compression and encryption algorithm combining compressed sensing, Sudoku matrix, and hyperchaotic map," *Nonlinear Dyn.*, vol. 113, no. 3, pp. 2831–2865, 2025.

[31] Z. Xue *et al.*, "Multimodal self-supervised learning for remote sensing data land cover classification," *Pattern Recognit.*, vol. 157, p. 110959, 2025.

[32] R. Xu, X. Wang, T. Wang, Y. Chen, J. Pang, and D. Lin, "Pointllm: Empowering large language models to understand point clouds," in *European Conference on Computer Vision*, 2025, pp. 131–147.

[33] H. N. Thanh *et al.*, "Forest cover change mapping based on Deep Neuron Network, GIS, and High-resolution Imagery," *Vietnam J. Earth Sci.*, pp. 151–175, 2025.

[34] L. Aliyeva and N. Abdullayev, "Hybrid Deep Learning Approach Towards Smart Grid Stability Prediction," in *2024 IEEE 8th Energy Conference (ENERGYCON)*, 2024, pp. 1–5.

[35] M. A. SAEED and N. M. Edan, "Design and implement video and chat application using Mesh network," *Int. J. Appl. Sci. Technol.*, vol. 4, no. 4, 2022.

[36] N. A. AL-Saati, "The exploration of wild horse optimization in reliability redundancy allocation problems," *Int. J. Intell. Eng. Syst.*, vol. 15, no. 4, pp. 198–207, 2022.