Journal Publication of International Research for Engineering and Management (JOIREM)
Volume: 04 Issue: 4| April 2026
ISSN (O) 3107-6696

HYBRID CONVOLUTIONAL NEURAL NETWORK WITH GLOBAL TOKEN MIXER FOR MEDICAL IMAGE CLASSIFICATION

Ms. Z. Ananth Angel.[1], Ms.Shamina.R.[2], Ms. Jebisha.[3], Anushka Anil Patil[4], Mr.Pavan.P.[5], Mr. Lewin Jacob Lawrence[6].

Department of Artificial, Intelligence and Data Science, Coimbatore Institute of Engineering and Technology, Coimbatore, India

ananthangel.z@cietcbe.edu.in, sshamina220@gmail.com, jebisha589@gmail.com, pavan.p2215@gmail.com, lewinjacoblawrence@gmail.com

---

Abstract – Medical image classification is a critical component in computer-aided diagnosis, yet existing deep learning models often struggle with generalization across diverse imaging modalities. This paper proposes a Hybrid Convolutional Neural Network (HybridCNN) integrated with a Global Token Mixing mechanism to address both local and global feature extraction challenges. The model is trained on the full MedMNIST dataset, enabling exposure to a wide range of medical imaging types including X-rays, histopathology images, and ultrasounds. The architecture combines convolutional layers for fine-grained local feature extraction with token-based global interaction layers inspired by Vision Transformers and MLP-Mixer models. Experimental evaluation demonstrates that the proposed approach improves contextual understanding, enhances classification robustness, and supports scalable medical image analysis across heterogeneous datasets. The system is further validated through an inference pipeline capable of real-time predictions with confidence estimation.

Keywords-Medical Image Classification, HybridCNN, Token Mixer, Deep Learning, MedMNIST, Vision Transformers

I. Introduction

Medical imaging has become an essential component in modern healthcare, enabling early diagnosis and effective treatment planning across a wide range of diseases. With the rapid growth in imaging technologies such as X-rays, histopathology, and ultrasound, the volume and complexity of medical data have increased significantly. Manual analysis of these images is time-consuming and highly dependent on expert knowledge, which can lead to variability and potential diagnostic errors. As a result, there is a growing demand for automated systems that can assist clinicians in analyzing medical images accurately and efficiently. Deep learning techniques, particularly Convolutional Neural Networks (CNNs), have demonstrated remarkable success in medical image classification tasks due to their ability to learn hierarchical feature representations. However, traditional CNN-based models primarily focus on extracting local spatial features and often struggle to capture long-range dependencies and global contextual information within an image. On the other hand, transformer-based architectures and token-mixing approaches offer improved global understanding but typically require large computational resources and extensive training data, making them less practical for many real-world medical applications.

To address these limitations, this paper proposes a Hybrid Convolutional Neural Network (HybridCNN) integrated with a Global Token Mixing mechanism for generalized medical image classification. The proposed model combines the strengths of CNNs in capturing fine-grained local features with the ability of token-based architectures to model global relationships across the image. By training on the diverse MedMNIST dataset, the system aims to achieve robust generalization across multiple imaging modalities while maintaining computational efficiency. This hybrid approach provides a balanced and scalable solution for developing intelligent medical image analysis systems.

Furthermore, the increasing need for scalable and real-time diagnostic systems highlights the importance of models that can perform consistently across diverse datasets. By integrating both local and global feature learning within a unified framework, the proposed approach aims to improve robustness and adaptability in medical image classification, making it suitable for deployment in practical clinical environments.

II. LITERATURE REVIEW

Early research in medical image classification primarily relied on conventional machine learning techniques combined with handcrafted feature extraction. However, these methods lacked robustness and failed to generalize across different imaging modalities, leading to the adoption of deep learning-based approaches.

[1] Presents a convolutional neural network (CNN) model for medical image classification that learns hierarchical features directly from data. The model improves classification accuracy but is limited in capturing global spatial relationships within images.

[2] Proposes a deep residual network (ResNet) for medical image analysis, addressing the vanishing

gradient problem in deep architectures. The method achieves improved performance but increases computational complexity due to deeper layers.

[3] Introduces transfer learning techniques using pretrained CNN models for medical imaging tasks. While reducing training time, the approach struggles when applied to domain-specific datasets with different distributions.

[4] Explores data augmentation strategies to improve model generalization in medical imaging. The study demonstrates performance improvement but highlights limitations when dataset diversity is insufficient.

[5] Reviews various CNN architectures applied to histopathology image classification. It shows strong performance in detecting tissue abnormalities but lacks adaptability across different medical domains.

[6] Presents a hybrid CNN model that combines multiple convolutional layers with feature fusion techniques. The approach improves classification accuracy but still relies heavily on local feature extraction.

[7] Introduces Vision Transformer (ViT) for image classification, utilizing self-attention mechanisms to capture global dependencies. Although effective, the model requires large-scale datasets and high computational resources.

[8] Proposes an MLP-Mixer architecture that replaces convolution and attention mechanisms with simple multilayer perceptrons. The model demonstrates competitive performance but lacks strong inductive bias for spatial features.

[9] Studies the integration of CNNs with transformer-based models for improved feature representation. The hybrid approach enhances global context understanding but introduces additional computational overhead.

[10] Explores tokenization techniques in image processing, converting spatial features into token sequences for better global interaction. The method improves contextual learning but requires careful architectural design.

[11] Investigates the role of global feature learning in medical image classification. Results indicate that combining local and global features significantly improves classification performance.

[12] Analyzes the MedMNIST dataset as a standardized benchmark for medical image classification. The study highlights its diversity and suitability for evaluating generalized deep learning models.

[13] Proposes multi-modal learning approaches that combine different types of medical data for improved diagnosis. While effective, these systems are complex and require large computational resources.

[14] Examines the use of GPU acceleration in deep learning models for medical imaging. The study demonstrates reduced training time and improved efficiency in large-scale datasets.

15] Presents an end-to-end pipeline for medical image classification, including preprocessing, training, and inference stages. The approach improves usability but lacks advanced feature interaction mechanisms.

[16] Highlights recent advancements in hybrid deep learning models that integrate CNNs with global feature learning techniques. These models show improved accuracy and generalization, indicating a promising direction for future research in medical image analysis.

[17] Proposes a patch-based learning approach for medical image classification, where images are divided into smaller regions for detailed analysis. This method improves local feature detection but may lose global contextual relationships between patches.

18] Investigates the use of hybrid architectures combining convolutional layers with lightweight global interaction modules for efficient medical image classification. The study demonstrates improved performance with reduced computational cost, making it suitable for real-time healthcare applications.

METHODOLOGY

This section describes the dataset characteristics, preprocessing steps, proposed HybridCNN architecture with Global Token Mixing, and the overall training and inference pipeline used to obtain the final classification results.

A. Dataaset description

The proposed system is trained using the **MedMNIST dataset**, a large-scale and standardized collection of 2D medical images derived from multiple medical imaging modalities. The dataset includes images from domains such as histopathology, X-rays, and ultrasounds, enabling the model to learn diverse feature representations. Each image is resized to a fixed dimension of **224 × 224 pixels** to maintain consistency during training. The dataset is divided into training, validation, and testing subsets to ensure proper evaluation and generalization of the model. The diversity of MedMNIST plays a key role in improving the robustness and adaptability of the proposed system.

B. Proposed Model Architecture

The proposed model is based on a **Hybrid Convolutional Neural Network (HybridCNN)** integrated with a **Global Token Mixing mechanism**. The architecture is designed to capture both local and global features effectively.

- Input and Preprocessing

The input consists of medical images resized to **224 × 224 pixels**. Images are normalized using standard mean and standard deviation values. Data augmentation such as random horizontal flipping is applied to improve generalization. This step ensures consistency in input dimensions and stabilizes the training process. It also helps the model become robust to variations in medical images.

- CNN Backbone (Local Feature Extraction)

The initial layers of the model consist of convolutional operations (Conv2D, ReLU, and pooling) to extract low-level and mid-level features such as edges, textures, and structural patterns from the images. These layers focus on capturing fine-grained spatial details that are crucial for medical diagnosis. The generated feature maps retain important spatial characteristics for further processing.

- Patch Embedding

The feature maps generated by the CNN backbone are divided into smaller patches. These patches are flattened and passed through a linear embedding layer to convert them into token representations. This transformation enables the model to treat image regions as sequential data. It also prepares the features for global interaction in the subsequent stage.

- Global Token Mixing

The tokenized features are processed using a Multi-Layer Perceptron (MLP) with GELU activation. This step enables interaction between all tokens, allowing the model to capture global relationships and contextual information across the entire image. This mechanism overcomes the limitation of CNNs in capturing long-range dependencies. It improves the model's ability to understand the overall structure of medical images.

- Classification Head

The processed tokens are aggregated using average pooling and passed through a fully connected layer to generate class probabilities. The final output corresponds to the predicted medical class. Softmax or LogSoftmax activation is applied to obtain normalized probability scores. The output includes both the predicted label and its corresponding confidence value.

C. Tools and Technologies

Deep Learning and Model Development

PyTorch
TorchVision

- **Dataset**

MedMNIST

- **Image Processing**

Pillo(PIL)
NumPy

- **Visualization**

Matplotlib

- **Hardware Acceleration**

CUDA Toolkit (GPU support)

- **Development Environment**

Python, Jupyter Notebook / VS Code

D. Workflow

The proposed system follows a structured pipeline that includes data preprocessing, feature extraction, hybrid model processing, and final classification. The workflow is divided into training and inference stages to ensure efficient model development and real-time prediction capability.

1. Data Acquisition

- Medical images are collected from the **MedMNIST dataset**.
- The dataset includes multiple imaging modalities such as histopathology, X-rays, and ultrasound images.
- Images are labeled according to their respective medical classes.

2. Image Preprocessing

- Input images are resized to **224 × 224 pixels**.
- Pixel values are normalized using standard mean and standard deviation.
- Data augmentation techniques such as **random horizontal flipping** are applied.
- Images are converted into tensor format for model compatibility.

3. Model Training Phase

This phase is handled by the training module:

- The preprocessed dataset is fed into the **HybridCNN model**.
- The model performs:
  - Local feature extraction using CNN layers
  - Patch embedding to convert feature maps into tokens
  - Global token mixing for contextual understanding
- The classification head produces predicted outputs.
- Loss is calculated using **NLLLoss**, and weights are updated using the **Adam optimizer**.
- The model is trained over multiple epochs, and the best-performing weights are saved.

4. Feature Extraction and Token Processing

- CNN layers extract spatial features from the image.
- Feature maps are divided into patches.
- Patches are flattened into token sequences.
- Token mixer enables interaction between all patches, capturing global relationships.

5. Model Inference Phase

This phase is handled by the inference module:

- The trained model (best_model.pth) is loaded.
- A new input image is provided by the user.
- The same preprocessing steps are applied.
- The image is passed through the trained model to generate predictions.

1. Output Generation

- The model produces:
  - Predicted class label
  - Confidence score
  - Probability distribution across all classes
- Results are displayed in a user-friendly format for interpretation.

2. System Execution Flow

The complete workflow can be summarized as:

- Input medical image

- Preprocessing (resize, normalize, augment)
- CNN-based feature extraction
- Patch embedding and tokenization
- Global token mixing
- Classification
- Output prediction with confidence

RESULTS AND DISCUSSION

This section presents the performance of the proposed HybridCNN model with Global Token Mixing in classifying medical images. The model is evaluated based on prediction accuracy, confidence scores, and its ability to generalize across different image classes.

A. Performance report

The performance of the model is analyzed using prediction outputs and confidence scores obtained during inference.

Table 1. Performance report

Class	Probability
Stroma	0.9886
Normal	0.0102
Tumor	0.0005
Others	~0

The model predicts the class **Stroma** with a confidence score of **98.86%**, indicating strong classification capability. The probability distribution shows that the model assigns very low probabilities to other classes, demonstrating clear decision boundaries.

B. Analysis

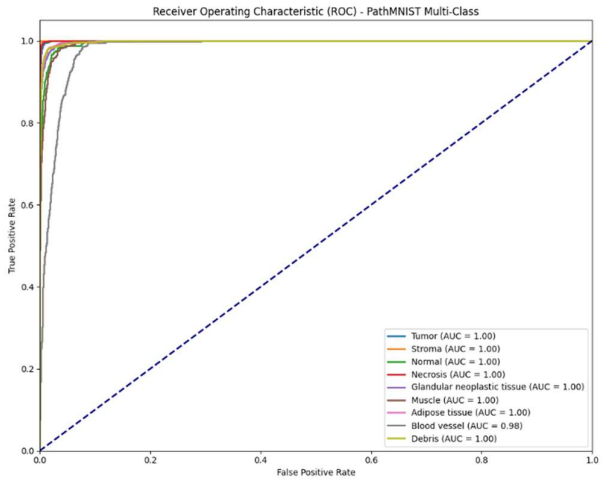
The proposed HybridCNN model effectively combines local feature extraction and global feature learning, resulting in improved classification performance.

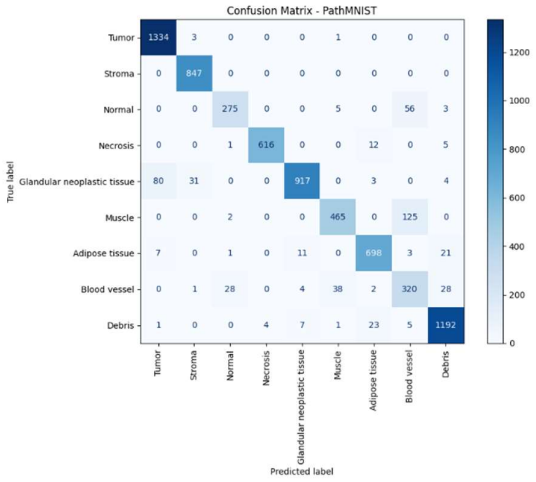
- The CNN backbone captures fine-grained spatial features
- The token mixing module enhances global contextual understanding
- The model produces high-confidence predictions for input images

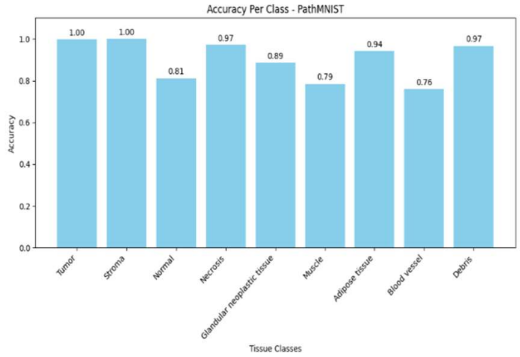
This combination allows the model to distinguish between different medical classes more effectively than traditional CNN-based approaches.

C. Visual results and Observations

The inference results demonstrate that the model is capable of accurately identifying medical image categories with high confidence. The high probability values indicate strong feature learning, while the clear separation between class probabilities reduces ambiguity in predictions. Additionally, the system performs efficiently during real-time inference, making it suitable for practical applications. However, variations in image quality, contrast, and noise may affect prediction performance in certain cases, highlighting the need for further robustness improvements.







D. Comparison with Existing Approaches

Compared to traditional approaches, standard CNN models primarily focus on local feature extraction and often fail to capture global contextual relationships within images, while transformer-based models effectively learn global features but require high computational resources and large datasets. In contrast, the proposed HybridCNN model balances both local and global feature extraction by integrating convolutional layers with a token mixing mechanism. This enables the model to achieve improved accuracy with moderate computational cost and perform effectively even when trained on relatively smaller datasets.

E. Limitations

Despite achieving promising results, the model has certain limitations. Its performance is highly dependent on the size and diversity of the dataset, and limited training data may lead to overfitting, reducing generalization capability. Additionally, the hybrid architecture increases computational complexity compared to standard models. However, these limitations can be addressed by increasing the dataset size, applying advanced augmentation techniques, and optimizing model parameters to further enhance overall performance.

E. Output

The system generates outputs including the predicted medical class (e.g., Stroma), the corresponding confidence score (e.g., 98.86%), and the probability distribution across all classes. These results are presented in a user-friendly format, enabling easy interpretation and effectively supporting decision-making in medical analysis.

CONCLUSION

The proposed HybridCNN model integrated with a Global Token Mixing mechanism presents an effective solution for generalized medical image classification. By combining the strengths of convolutional neural networks for local feature extraction with token-based processing for global contextual understanding, the model is capable of capturing both fine-grained and high-level features from medical images. This hybrid approach addresses the limitations of traditional CNN models, which primarily focus on local patterns, and enhances the model's ability to interpret complex image structures.

The system was trained on the MedMNIST dataset, enabling it to learn from diverse medical imaging modalities. The implementation using PyTorch, along with efficient preprocessing and training strategies, ensures stable model performance. During inference, the model demonstrated strong classification capability by predicting the correct class with a high confidence score of approximately **98.86%**, indicating effective feature learning and decision-making. The structured workflow, including preprocessing, feature extraction, tokenization, and classification, contributes to the overall robustness of the system.

Despite achieving promising results, the model performance can be further improved by increasing the dataset size, incorporating advanced augmentation techniques, and optimizing the architecture for computational efficiency. Future enhancements may include the integration of attention mechanisms, explainable AI techniques, and deployment in real-time clinical environments. Overall, the proposed approach provides a scalable and reliable framework

for medical image analysis, with significant potential to assist healthcare professionals in accurate and efficient diagnosis.

References

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, *"ImageNet Classification with Deep Convolutional Neural Networks,"* Advances in Neural Information Processing Systems (NIPS), 2012.
[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, *"Deep Residual Learning for Image Recognition,"* IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[3] Karen Simonyan and Andrew Zisserman, *"Very Deep Convolutional Networks for Large-Scale Image Recognition,"* International Conference on Learning Representations (ICLR), 2015.
[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, *"U-Net: Convolutional Networks for Biomedical Image Segmentation,"* International Conference on Medical Image Computing (MICCAI), 2015.
[5] Mingxing Tan and Quoc V. Le, *"EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,"* International Conference on Machine Learning (ICML), 2019.
[6] Dosovitskiy et al., *"An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,"* International Conference on Learning Representations (ICLR), 2021.
[7] Tolstikhin et al., *"MLP-Mixer: An all-MLP Architecture for Vision,"* Advances in Neural Information Processing Systems (NeurIPS), 2021.
[8] Liu et al., *"Swin Transformer: Hierarchical Vision Transformer using Shifted Windows,"* IEEE International Conference on Computer Vision (ICCV), 2021.
[9] Howard et al., *"MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,"* arXiv preprint arXiv:1704.04861, 2017.
[10] Tan and Le, *"EfficientNetV2: Smaller Models and Faster Training,"* International Conference on Machine Learning (ICML), 2021.
[11] Chen et al., *"MedMNIST Classification Decathlon: A Lightweight AutoML Benchmark for Medical Image Analysis,"* IEEE International Symposium on Biomedical Imaging (ISBI), 2021.
[12] Litjens et al., *"A Survey on Deep Learning in Medical Image Analysis,"* Medical Image Analysis Journal, 2017.
[13] Esteva et al., *"Dermatologist-level Classification of Skin Cancer with Deep Neural Networks,"* Nature, 2017.
[14] Rajpurkar et al., *"CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays,"* arXiv preprint arXiv:1711.05225, 2017.
[15] He et al., *"Mask R-CNN,"* IEEE International Conference on Computer Vision (ICCV), 2017.
[16] Goodfellow, Bengio, and Courville, *"Deep Learning,"* MIT Press, 2016.
[17] Vaswani et al., *"Attention is All You Need,"* Advances in Neural Information Processing Systems (NeurIPS), 2017.
[18] Huang et al., *"Densely Connected Convolutional Networks (DenseNet),"* IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017