InViNet: A Dual-branch Vision Transformer based Novel Architecture for Medical Image Applications

Saksham Hooda
Department of Applied Mathematics
Delhi Technological University
Delhi, India
sakshamhooda_2k20mc116@dtu.ac.in

***Abstract*—The classification of medical images is a critical component in clinical treatment and diagnosis. While deep learning, particularly Convolutional Neural Networks (CNNs), has advanced this field, recent Vision Transformer (ViT) models have shown superior accuracy, albeit often requiring extensive pre-training. It is now understood that combining convolutions with transformers can leverage the strengths of both. Our study proposes InViNet, a novel dual-branch architecture that synergizes texture features with the benefits of a Compact Convolutional Transformer (CCT). One branch uses Central Difference Convolution (CDC) to extract fine-grained texture information, while the other learns global spatial relationships via a CCT backbone. Through comprehensive evaluation on the Malaria and BloodMNIST datasets, InViNet demonstrates robust performance, achieving an AUC of 0.9941 and 0.9933, respectively. Our model is compact, requires no pre-training, and surpasses the performance of several existing models while demanding less training time. These findings suggest InViNet could significantly advance medical image analysis, offering a practical and efficient solution for precise disease diagnosis.**

***Index Terms*—Medical Image Classification, Vision Transformer, Convolutional Neural Networks, Compact Convolutional Transformer, Texture Features, Deep Learning.**

I. Introduction

The classification of medical images is a critical task in modern healthcare for clinical diagnosis and treatment. While Convolutional Neural Networks (CNNs) have long been the state-of-the-art [16], they are now being challenged by Vision Transformers (ViTs) [6], which have shown superior accuracy on many benchmarks. However, ViTs often require massive datasets for pre-training and discard the useful inductive biases of convolutions, such as shift and scale invariance.

This has led to the development of hybrid architectures that combine the strengths of both paradigms. Models like the Convolutional Vision Transformer (CvT) [10] and the Compact Convolutional Transformer (CCT) [1] integrate convolutions into the Transformer architecture, achieving better performance with fewer parameters and less data.

In this paper, we propose InViNet, a novel dual-branch architecture for medical image classification. InViNet synergizes texture-based and spatial-based feature extraction. One branch uses Central Difference Convolution (CDC) [2] to capture fine-grained texture features, inspired by its success in face anti-spoofing. The other branch uses a CCT backbone to learn global spatial relationships. By fusing features from both branches, InViNet achieves a more comprehensive image understanding. We demonstrate that our model, trained from scratch, achieves state-of-the-art results on the Malaria and BloodMNIST datasets, outperforming existing models in Accuracy and AUC with fewer parameters and less training time.

II. Related Work

Medical image classification has evolved from traditional machine learning with handcrafted features to deep learning. CNN architectures like VGG [7], ResNet [8], and DenseNet have become standard baselines, often used with transfer learning [3], [17].

The introduction of the Vision Transformer (ViT) [6], which applied the Transformer architecture [20] directly to image patches, marked a paradigm shift. While powerful, ViTs require extensive pre-training. To address this, hybrid models emerged. The Convolutional Vision Transformer (CvT) [10] reintroduced convolutions for tokenization and in the attention block. Further pushing for efficiency, the Compact Convolutional Transformer (CCT) [1] was proposed, which is trainable from scratch on smaller datasets and uses a sequence pooling mechanism instead of a class token.

Our work builds upon these advancements. We specifically leverage the efficiency of CCT for spatial feature learning. To enhance the model's ability to discern subtle, disease-related patterns, we incorporate a dedicated texture-feature branch based on the CDCN++ architecture [2], which uses Central Difference Convolutions (CDC) to capture discriminative texture information. This dual-branch approach of combining specialized texture features with global spatial attention is a key novelty of our proposed InViNet.

III. Methodology

A. Datasets

1) Malaria: This dataset [3] contains 27,558 images of cells, balanced between 'Parasitized' and 'Uninfected' classes. All images were resized to 3 × 32 × 32 pixels. The data was split into 80% training, 10% validation, and 10% testing.

2) BloodMNIST: Part of the MedMNIST v2 collection [4], this dataset has 17,092 images of individual normal blood cells across eight classes. Images were resized to 3 × 28 × 28 pixels

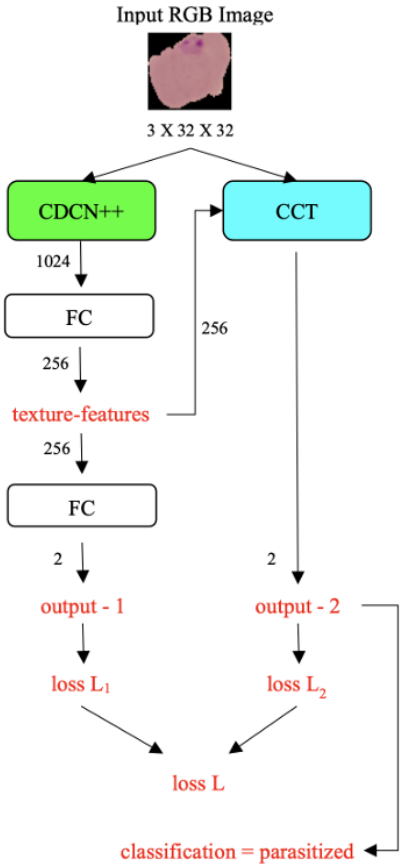

Fig. 1. Proposed Architecture of InViNet, showing the dual-branch structure with a Texture-Branch (CDCN++) and a Spatial-Branch (CCT).

and normalized. The data was split into 70% training, 10% validation, and 20% testing.

B. Proposed Architecture: InViNet

Our proposed architecture, InViNet (Intelligent Visual Network), is a dual-branch model designed to capture both texture and spatial features from RGB images, as illustrated in Fig. 1.

1) Texture-Branch: This branch is built upon the CDCN++ architecture [2] and is designed to extract detailed texture features.

- **CDC Operator:** The Central Difference Convolution (CDC) operator enhances the standard convolution by aggregating the center-oriented gradient of sampled values, improving its ability to capture texture.
- **Feature Fusion:** The CDCN++ architecture extracts low, mid, and high-level fused features. These are refined using a Multiscale Attention Fusion Module (MAFM) to learn more discriminating features.
- **Output:** The extracted texture features are passed through linear layers to produce a class-dimensional output for loss calculation (L_1) and a weighted vector for the spatial branch.

2) Spatial-Branch: This branch uses a CCT-backbone to capture global spatial information.

- **Convolutional Tokenizer:** A simple convolution block embeds local features and inductive biases. It consists of a convolution layer, ReLU activation, and a max-pooling layer.
- **Compact Transformer-Encoders:** The convolved tokens, encoded with sinusoidal positional embeddings, are passed through seven Compact Transformer-Encoder blocks to capture global context.
- **Feature Integration:** The texture-feature vector from the Texture-Branch weights the output of each token from the Transformer-Encoder blocks, enriching them with texture information.
- **Sequence Pooling:** An attention-based sequence pooling mechanism aggregates the output sequence from the transformer blocks. The pooled output is then passed to a fully connected layer for final classification, contributing to loss L_2.

Loss Calculation: The final loss L for the model is the sum of the losses from both branches: $L_{final} = L_1 + L_2$. The total parameter count is approximately 6.3 million.

C. Experimental Setup

- **Batch Size:** 128 for both datasets.
- **Positional Embedding:** Sinusoidal embedding for the CCT-backbone.
- **Loss Function:** Cross-Entropy Loss.
- **Optimizer:** AdamW with a learning rate of 0.001.
- **Epochs:** 100 for Malaria, 30 for BloodMNIST.
- **Hardware:** All models were trained on an NVIDIA A100 GPU via Google Colab Pro.

IV. Results and Discussion

We first benchmarked several standard CNN and ViT models. As shown in Table I, hybrid models like CvT and ViT outperformed traditional CNNs, motivating our direction.

TABLE I
BASELINE MODEL PERFORMANCE ON MALARIA DATASET

Model	**Accuracy (%)**	**Training Time (s)**
VGG-16	48.80	130.55
VGG-19	50.71	151.38
ResNet50	84.59	93.14
ResNeXt50	82.10	123.58
ViT-Base	95.48	284.15
CvT-13	96.28	165.80

We then conducted experiments with different CCT variants, with and without the CDCN++ backbone, to identify the optimal configuration for InViNet. The results on the Malaria dataset are in Table II. The CCT-7/3X1 variant combined with the CDCN++ backbone yielded the best performance.

Our final InViNet architecture was compared against other state-of-the-art models on both datasets. Table III and Table IV show that InViNet, trained from scratch, achieves superior

TABLE II
PERFORMANCE OF INVINET VARIANTS ON MALARIA DATASET

CCT Variant	CDCN++ Present	Accuracy (%)	AUC
2*CCT-7/3X1	No	96.11	0.9905
	Yes (InViNet)	**96.84**	**0.9941**
2*CCT-7/3X2	No	95.64	0.9899
	Yes	96.13	0.9904
2*CCT-14/3X1	No	96.15	0.9901
	Yes	96.44	0.9903
2*CCT-14/7X2	No	95.39	0.9887
	Yes	95.60	0.9890

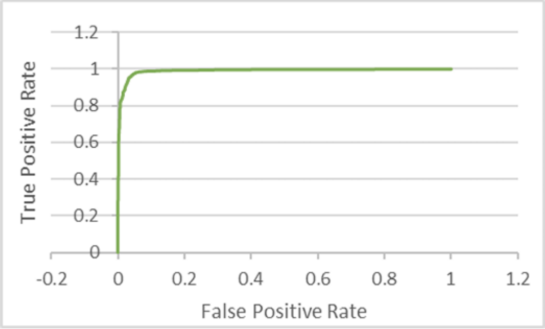

Fig. 2. AUC/ROC curve for InViNet on the Malaria test set.

or highly competitive results across all key metrics. Notably, it outperforms models that rely on transfer learning, while using fewer parameters and remaining less complex.

The high performance is further evidenced by the AUC/ROC curve for the Malaria dataset in Fig. 2. The model's discriminative power is visualized through t-SNE plots in Fig. 4, which show clear separation between classes for both datasets. Furthermore, Class Activation Maps (CAMs) from the texture-branch, shown in Fig. 3, confirm that the model focuses on relevant, discriminative regions within the cell images, validating our dual-branch design.

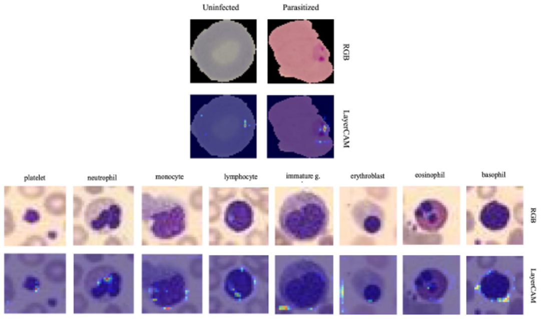

Fig. 3. Class Activation Maps (CAMs) from the texture-branch for Malaria (top) and BloodMNIST (bottom) images, highlighting discriminative regions.

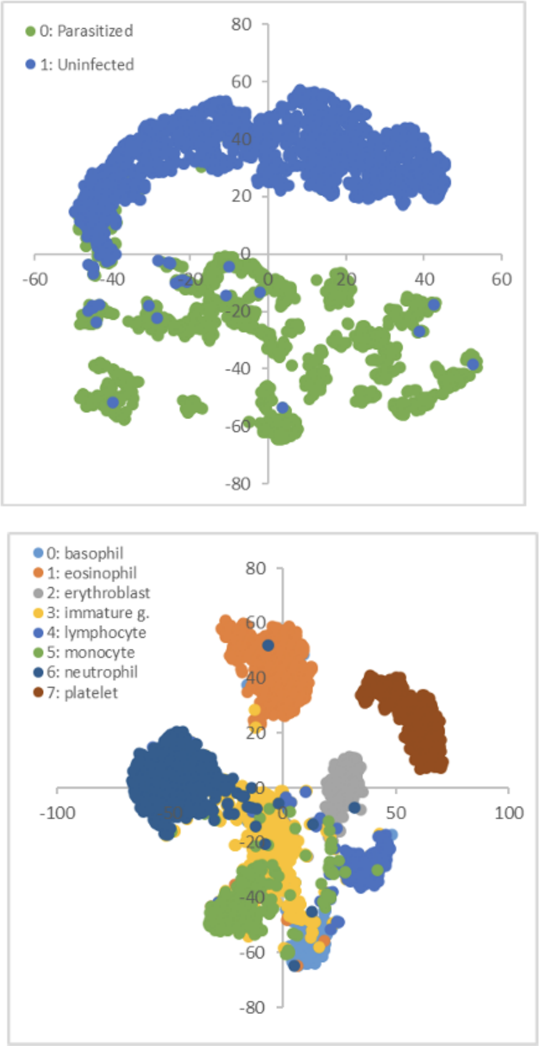

Fig. 4. t-SNE visualization of features from the test set for Malaria (top) and BloodMNIST (bottom), showing clear class separation.

TABLE III
COMPARISON OF INVINET WITH EXISTING ARCHITECTURES ON THE MALARIA DATASET

Model	Accuracy (%)	Precision	Recall	F1-Score	AUC
kEffNet-B0 [23]	96.70	—	—	—	—
Seq. CNN Model [24]	96.00	—	—	—	—
DenseNet 201 [25]	93.39	0.9549	0.9104	0.9321	—
Custom CNN [26]	96.29	0.9804	0.9234	0.9495	0.9116
InViNet (Ours)	**96.84**	**0.9649**	**0.9713**	**0.9681**	**0.9941**

TABLE IV
COMPARISON OF INVINET WITH EXISTING ARCHITECTURES ON THE BLOODMNIST DATASET

Model	Accuracy (%)	AUC
Vision Transformer [6]	88.80	0.985
OrthoFNN [28]	82.00	0.972
auto-sklearn [29]	87.80	0.984
MonoNet [19]	88.10	—
InViNet (Ours)	**92.75**	**0.9933**

V. Conclusion

In this project, we introduced InViNet, a novel dual-branch architecture that effectively combines texture-based feature extraction using Central Difference Convolutions with the global spatial reasoning of a Compact Convolutional Transformer. Our results show that InViNet achieves state-of-the-art performance on both the Malaria and BloodMNIST datasets, outperforming many existing models in key metrics like accuracy and AUC.

Crucially, InViNet is compact, with only 6.3 million parameters, and achieves these results without relying on large-scale pre-training or transfer learning. This makes it a practical and efficient solution for medical image classification, especially in scenarios with limited data. The success of our architecture underscores the value of hybrid models that are thoughtfully designed to capture complementary image features.

VI. Future Work

The InViNet architecture shows great promise, and several avenues for future work exist. To improve usability on resource-constrained hardware, adaptive or quantized transformers could be explored. The model's robustness could be tested on a wider range of medical imaging datasets, such as those for brain tumor segmentation (BraTS). Additionally, the architecture could be extended to handle multi-modal data by integrating other data types alongside image features. Exploring the use of ViTs within the discriminator of a Generative Adversarial Network (GAN) for data augmentation is another promising direction.

References

[1] A. Hassani et al., "Escaping the Big Data Paradigm with Compact Transformers," *ArXiv*, 2021. [Online]. Available: https://arxiv.org/abs/2104.05704

[2] Z. Yu et al., "Searching Central Difference Convolutional Networks for Face Anti-Spoofing," *ArXiv*, 2020. [Online]. Available: https://arxiv.org/abs/2003.04092

[3] S. Rajaraman et al., "Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images," *PeerJ*, vol. 6, e4568, 2018.

[4] J. Yang et al., "MedMNIST v2– A large-scale light-weight benchmark for 2D and 3D biomedical image classification," *Scientific Data*, vol. 9, no. 1, p. 652, 2022.

[5] A. Acevedo et al., "A dataset of microscopic peripheral blood cell images for development of automatic recognition systems," *Data in Brief*, vol. 30, 105474, 2020.

[6] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *ArXiv*, 2020. [Online]. Available: https://arxiv.org/abs/2010.11929

[7] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *ArXiv*, 2014. [Online]. Available: https://arxiv.org/abs/1409.1556

[8] K. He et al., "Deep Residual Learning for Image Recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[9] S. Xie et al., "Aggregated Residual Transformations for Deep Neural Networks," in *Proc. CVPR*, 2017, pp. 1492-1500.

[10] H. Wu et al., "CvT: Introducing Convolutions to Vision Transformers," in *Proc. ICCV*, 2021, pp. 22-31.

[11] R. Paredes et al., "Classification of Medical Images Using Local Representations," in *Bildverarbeitung für die Medizin 2002*, 2002, pp. 166-170.

[12] N. Parveen and M. M. Sathik, "Detection of Pneumonia in Chest X-ray Images," *IJCSET*, vol. 1, no. 10, pp. 627-631, 2011.

[13] J. C. Caicedo et al., "Histopathology image classification using bag of features and kernel functions," in *Proc. AIME*, 2009, pp. 126–135.

[14] E. Rublee et al., "ORB: An efficient alternative to SIFT or SURF," in *Proc. ICCV*, 2011, pp. 2564–2571.

[15] O. Russakovsky et al., "ImageNet Large Scale Visual Recognition Challenge," *Int J Comput Vis*, vol. 115, pp. 211–252, 2015.

[16] Q. Li et al., "Medical image classification with convolutional neural network," in *Proc. ICARCV*, 2014, pp. 844-848.

[17] D. S. Kermany et al., "Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning," *Cell*, vol. 172, no. 5, pp. 1122-1131, 2018.

[18] X. Wang et al., "ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases," in *Proc. CVPR*, 2017, pp. 2097-2106.

[19] A. Nguyen et al., "MonoNet: Enhancing interpretability in neural networks via Monotonic Features," *Bioinformatics Advances*, vol. 3, no. 1, vbad016, 2023.

[20] A. Vaswani et al., "Attention Is All You Need," in *Proc. NeurIPS*, 2017, pp. 5998-6008.

[21] J. Yang et al., "MTD-Net: Learning to Detect Deepfakes Images by Multi-Scale Texture Difference," *IEEE T-IFS*, vol. 16, pp. 4234-4245, 2021.

[22] N. Park and S. Kim, "How Do Vision Transformers Work?" *ArXiv*, 2022. [Online]. Available: https://arxiv.org/abs/2202.06709

[23] S. Sinha et al., "Performance assessment of Deep Learning procedures on Malaria dataset," *Journal of Robotics and Control (JRC)*, vol. 2, no. 1, pp. 12-18, 2021.

[24] A. Mohammed et al., "Malaria Parasite Identification from Red Blood Cell Images Using Transfer Learning Models," in *Proc. ICAC*, 2022.

[25] A. Rahman et al., "Improving Malaria Parasite Detection from Red Blood Cell using Deep Convolutional Neural Networks," *ArXiv*, 2019. [Online]. Available: https://arxiv.org/abs/1907.10418

[26] J. P. Schwarz Schuler et al., "An Enhanced Scheme for Reducing the Complexity of Pointwise Convolutions in CNNs for Image Classification," *Entropy*, vol. 24, no. 9, 1264, 2022.

[27] E. A. Cherrat et al., "Quantum Vision Transformers," *ArXiv*, 2022. [Online]. Available: https://arxiv.org/abs/2209.08167

[28] I. Kerenidis et al., "Classical and Quantum Algorithms for Orthogonal Neural Networks," *ArXiv*, 2021. [Online]. Available: https://arxiv.org/abs/2106.07198

[29] M. Feurer et al., "Auto-sklearn: Efficient and Robust Automated Machine Learning," in *Automated Machine Learning*, 2019, pp. 113-134.