Erciyes University Journal of Institue Of Science and Technology

Research Article
*10.65520/erciyesfen.1925750*

***Imprint:***
Volume: 42(3)
Year: 2026
Page: 323-351

Faruk ÖZGÜR[a]
Erdal ÖZBAY[b*]
Taner TUNCER[c]

[a] Master's Student, Firat University, farukozgur501@gmail.com
[b] Assoc. Prof., Firat University, erdalozbay@firat.edu.tr
[c] Prof., Firat University, ttuncer@firat.edu.tr

* Corresponding Author

*Received: 4/9/2026*
*Accepted: 9/15/2026*

***Citation:***
Faruk ÖZGÜR, Erdal ÖZBAY, Taner TUNCER (2026). Classification of HEp-2 Cell Patterns in Autoimmune Disease Diagnosis Using Hybrid DenseNet-ViT Architecture. *Erciyes University Journal of Institue Of Science and Technology*, 42(3), 323-351. https://doi.org/10.65520/erciyesfen.*1925750*

Screened by iThenticate for Authors & Researchers

Except where otherwise noted, content in this article is licensed under a Creative Commons 4.0 International license. Icons by Font Awesome.

# Classification of HEp-2 Cell Patterns in Autoimmune Disease Diagnosis Using Hybrid DenseNet-ViT Architecture

## Abstract

**Background/Aim:** *Accurate and consistent interpretation of HEp-2 cell patterns, considered the gold standard in autoimmune disease diagnosis, remains challenging due to inter-expert variability. This study aims to develop a hybrid deep learning architecture for the automated and highly accurate classification of HEp-2 cell images.*

**Methods:** *The proposed approach integrates DenseNet-121, which excels at capturing local morphological features, with Vision Transformer (ViT-B16), known for modeling global contextual relationships. These models are combined using a weighted ensemble strategy. The performance of the model was evaluated on the MIVIA (ICPR 2012) dataset under a strict source-image split protocol, which explicitly prevents the object-level data leakage frequently overlooked in the literature.*

**Results:** *The proposed hybrid model achieved an overall accuracy of 96.01% under these rigorous test conditions, without the use of synthetic data augmentation. A key methodological contribution is its ability to effectively distinguish between Homogeneous and Fine Speckled patterns, which are frequently misclassified by experts. Notably, the model achieved 100% precision for the Fine Speckled class. Additionally, it was empirically demonstrated that a small batch size (Batch Size = 8) improves generalization performance in fine-textured biomedical images.*

**Conclusion:** *These findings demonstrate the robust feature extraction capabilities of the proposed hybrid architecture, establishing a reliable, realistic, and methodologically transparent baseline for future computer-aided diagnosis (CAD) systems.*

**Keywords:** HEp-2 Cell Classification, Vision Transformer, Hybrid Deep Learning, Autoimmune Disease Diagnosis, Ensemble Learning

# Otoimmün Hastalık Tanısında Hep-2 Hücre Desenlerinin Hibrit DenseNet-ViT Mimarisi ile Sınıflandırılması

## Öz

*Otoimmün hastalıkların tanısında altın standart olarak kabul edilen HEp-2 hücre desenlerinin doğru ve tutarlı şekilde yorumlanması, uzmanlar arasında değişkenlik gösterebilmektedir. Bu çalışma, HEp-2 hücre görüntülerinin otomatik ve yüksek doğrulukla sınıflandırılması amacıyla hibrit bir derin öğrenme mimarisi geliştirmeyi amaçlamaktadır. Önerilen yaklaşımda, lokal morfolojik özelliklerin çıkarımında güçlü performans sergileyen DenseNet-121 mimarisi ile global dokusal bağıntıları modellemede üstün Vision Transformer (ViT-B16) yapısı birleştirilmiştir. Bu iki model, ağırlıklı bir topluluk (ensemble) stratejisi ile entegre edilmiştir. Modelin performansı, literatürde sıkça göz ardı edilen nesne düzeyindeki veri sızıntısını (data leakage) kesin olarak engelleyen katı bir kaynak-görüntü (source-image split) ayırma protokolü ile MIVIA (ICPR 2012)*

*veri seti üzerinde değerlendirilmiştir. Önerilen hibrit model, herhangi bir sentetik veri artırımı kullanılmadan bu zorlu test koşulları altında %96.01 genel doğruluk elde etmiştir. Modelin en dikkat çekici başarısı, uzmanlar tarafından sıklıkla karıştırılan Homogeneous ve Fine Speckled desenleri arasında yüksek doğrulukla ayrım yapabilmesidir. Özellikle Fine Speckled sınıfında %100 kesinlik (precision) değeri elde edilmiştir. Ayrıca, küçük batch boyutunun (Batch Size=8) ince dokulu biyomedikal görüntülerde model genellemesini artırdığı deneysel olarak gösterilmiştir. Elde edilen bulgular, önerilen hibrit mimarinin sağlam özellik çıkarma kapasitesini kanıtlamakta ve gelecekteki bilgisayar destekli tanı (CAD) sistemleri için güvenilir, gerçekçi ve metodolojik açıdan şeffaf bir temel oluşturduğunu göstermektedir.*

**Anahtar kelimeler:** *HEp-2 Hücre Sınıflandırması, Vision Transformer, Hibrit Derin Öğrenme, Otoimmün Tanı, Topluluk Öğrenmesi*

## 1. INTRODUCTION

Autoimmune diseases are chronic, systemic conditions that arise when the immune system loses its immunological tolerance to its own antigens and mistakenly identifies healthy body tissues as pathological targets, thereby attacking them. This group of diseases, which encompasses more than 80 different clinical presentations, including Systemic Lupus Erythematosus (SLE), Rheumatoid Arthritis (RA), Sjögren's Syndrome (SjS), Systemic Sclerosis (SSc), and Polymyositis, constitutes a serious public health problem worldwide [1]. According to World Health Organization (WHO) data and current epidemiological studies, autoimmune diseases affect approximately 5% to 8% of the global population, and their prevalence in industrialized countries is increasing each year due to the influence of environmental factors and lifestyle changes [2]. In particular, multisystemic diseases such as SLE can lead to kidney failure (lupus nephritis), cardiovascular damage, and irreversible neuropsychiatric disorders if not diagnosed early and treated with appropriate protocols. Statistics indicate that the mortality risk in SLE patients is 2 to 5 times higher than in the general population [3]. Therefore, diagnosing the disease in its early stages before it causes organ damage is the most critical factor directly affecting patients' survival rates and quality of life.

The most critical biomarker in the serological diagnosis and monitoring of autoimmune diseases is Antinuclear Antibodies (ANA) found in patient serum. The "Gold Standard" diagnostic method recommended by the American College of Rheumatology (ACR) and international immunology authorities is the Indirect Immunofluorescence (IIF) test performed using a human epithelial cell (HEp-2) substrate [4]. In this test procedure, the patient's serum is incubated with HEp-2 cells on a slide, and if autoantibodies are present in the serum, they bind to intracellular antigens. When examined under a fluorescence microscope, these bindings become visible as specific fluorescent patterns in the cell nucleus or cytoplasm.

The correct identification and classification of these patterns serve as a vital guide for the clinician. This is because each pattern indicates the presence of a specific disease or antibody [5]. For example, the Homogeneous pattern, characterized by smooth and diffuse staining in the nuclei of resting cells, is typically a strong indicator of Systemic Lupus Erythematosus (SLE) and Drug-Induced Lupus. The Coarse Speckled pattern, which presents as dense, large, and irregular granular structures within the nucleus, and the Fine Speckled pattern, consisting of countless very fine, dust-like specks spread throughout the entire nucleus, are frequently observed in Sjögren's Syndrome, SLE, and other mixed connective tissue diseases. The Centromere pattern, which appears as distinct, bright dots aligned along the chromosomes during cell division (typically numbering 46), shows a high correlation with the limited form of systemic sclerosis. The Nucleolar pattern, characterized by large, bright clumps of staining in the nucleolar region, is generally seen in cases of Polymyositis and Systemic Sclerosis [6]. Finally, cytoplasmic patterns, where staining occurs in the cell body rather than the nucleus, are associated with liver-involving diseases such as autoimmune hepatitis [7].

In clinical practice, the analysis of these complex patterns is traditionally performed manually by experienced immunologists or pathologists at a microscope under darkroom conditions. However, this manual examination process falls short of the speed, accuracy, and standardization required by modern medicine. The fundamental challenges in this process are frequently highlighted in the


CC: BY-NC-ND

literature: First, the evaluation process is highly subjective and relies heavily on the expert's experience. Studies have shown that diagnostic agreement between different laboratories or experts can remain as low as 70–75%, and even the same expert may assign different diagnoses to the same sample at different times [8]. The second and most critical challenge is the differentiation of patterns with highly similar visual morphologies, such as Homogeneous and Fine Speckled. In low-contrast microscope images, the mislabeling rate for these two classes can reach up to 30%, which carries the risk of applying an incorrect treatment protocol to the patient [9]. Thirdly, the examination of hundreds of samples daily under increasing laboratory workloads causes visual fatigue in experts, thereby increasing the rate of human error.

Furthermore, the vast variations observed in the size and morphology of HEp-2 cells, along with the significant imbalances between classes in the datasets, further exacerbate these challenges in manual analysis. Recent studies have highlighted these complex image structures and emphasized that the use of deep learning-based systems for high-accuracy, automated detection has now become a clinical necessity [10].

To overcome these limitations and standardize the diagnostic process, intensive research has been conducted over the past two decades on Computer-Aided Diagnosis (CAD) systems. Tissue analysis methods (LBP, GLCM) and Machine Learning techniques (SVM, k-NN) used in early studies have gradually been replaced by deep learning-based Convolutional Neural Networks (CNNs). Deep CNN architectures such as ResNet, DenseNet, and Inception have achieved significant success in HEp-2 classification due to their ability to learn local morphological features (edges, corners, shapes) in images [11]. However, since CNN architectures operate on the principle of "local receptive fields" and perform dimensionality reduction during pooling operations, they analyze the image in small windows. This situation limits the performance of CNNs in challenging patterns like Fine Speckled—where the distinguishing feature is not just a distinct shape but a fine, microscopic texture distribution spread across the entire cell surface—and leads to the loss of texture details [12].

These limitations of traditional classifiers and networks that focus solely on local features are also clearly evident in the current literature. For example, in studies emphasizing the clinical importance of detecting all HEp-2 patterns according to ICAP standards, a framework combining ResNet-50 for feature extraction and the Random Forest algorithm for classification was developed; however, it was limited to a specific accuracy threshold of 87% on independent test sets [13]. This demonstrates that achieving higher success at the clinical level requires more powerful global architectures capable of capturing not only local features but also the cell's overall structure.

Vision Transformer (ViT) architectures, which have gained popularity in the field of computer vision in recent years, offer a powerful alternative to this "locality" constraint of CNNs. Thanks to the "Self-Attention" mechanism, ViT models treat the image as a whole and can model long-range relationships between pixels [14]. This feature is theoretically ideal for deciphering global texture patterns in HEp-2 cells. However, ViT models require significantly more training data compared to CNNs. In medical imaging, particularly in fields like HEp-2 that require expert annotation, accessing large datasets is challenging. On limited datasets such as MIVIA (ICPR 2012), pure ViT models often suffer from overfitting and fail to fully demonstrate their potential [15].

In this study, a hybrid approach combining the complementary strengths of CNN and Transformer architectures is proposed to overcome these challenges in HEp-2 cell classification. Within the scope of the study, DenseNet-121 was used to capture local structural features, and a Vision Transformer (ViT-B16) was employed to analyze global textural patterns. The predictions of these models were combined using a Weighted Soft Voting strategy that preserves the advantages of both architectures. Crucially, to address the prevalent issue of object-level data leakage in the existing literature which often relies on random cell-level splitting—the proposed framework was rigorously evaluated under a strict source-image split protocol. Furthermore, the study experimentally demonstrates that, contrary to the common assumption that larger batch sizes yield better stability, small batch sizes (Batch Size = 8) act as an implicit regularizer on the limited and fine-textured HEp-2 dataset, significantly minimizing errors in the challenging Fine Speckled class. Without resorting to


CC: BY-NC-ND

computationally expensive synthetic data generation (e.g., GANs), the proposed methodology achieves highly competitive and methodologically robust performance on the ICPR-2012 dataset, establishing a reliable and realistic baseline for future Computer-Aided Diagnosis (CAD) systems.

## 2. RELATED WORKS

The classification of HEp-2 cell images is considered a benchmark problem in the field of computer vision and has undergone various evolutionary processes based on the methodologies used over the past decade. Initially, in the pre-deep learning era, researchers focused on extracting the morphological and textural features of cells using mathematical descriptors. In their work introducing the MIVIA dataset to the literature, Foggia et al. [16] achieved a baseline accuracy of around 80% by combining Local Binary Patterns (LBP), Gray-Level Co-occurrence Matrices (GLCM), and morphological features. Although Manivannan et al. proposed a method for multi-resolution texture analysis, the success of such handcrafted methods was limited by their sensitivity to changes in illumination and their inability to adequately represent complex patterns such as Fine Speckled [17]. With the ICPR 2012 and 2014 competitions, CNN models capable of automatic feature extraction have become the standard. Gao et al. [18] achieved a significant performance improvement on the ICPR-2014 dataset by supporting deep convolutional networks with comprehensive data augmentation techniques. Bayramoglu et al. [19] demonstrated the effectiveness of Transfer Learning in medical problems with limited data and showed that models trained on ImageNet could be successfully applied to HEp-2 classification.

One of the highest achievements was reported by Larissa F. Rodrigues et al. [20]. The authors achieved 98.28% accuracy on the MIVIA dataset using a CNN architecture that analyzes images at different scales. However, this study has a methodological limitation: performance evaluation was conducted using 5-fold cross-validation. Since cross-validation allows the model to see the entire dataset during the training process, it produces more optimistic results compared to a strict "Hold-Out" strategy based on unseen test data. Furthermore, many of these high-performing studies employ random cell-level splitting. As highlighted in recent methodological critiques, random splitting in datasets like MIVIA (where multiple cells are cropped from the same microscope image) inherently leads to object-level data leakage, artificially inflating performance metrics since the model learns the background illumination rather than the cellular pattern. Similarly, Lei et al. [21] reported success rates exceeding 96% using, respectively, customized small CNN architectures and cross-modal transfer learning based on ResNet-50. However, a comprehensive review by Rahman et al. [22] revealed that CNN architectures struggle in scenarios where the distinguishing feature is not a specific shape but rather global texture consistency spread across the entire cell surface (Homogeneous vs. Fine Speckled), due to the principle of local receptive fields. Similarly, although Jensen et al. proposed Deep Convolutional Neural Networks (DCNNs) to classify HEp-2 microscopic images in their recent study and achieved a high success rate of 95.24%, it appears that architectures that extract only local features have reached a plateau and are insufficient for capturing global cellular arrangement [23].

To overcome the locality constraint of CNNs, there has been growing interest in Vision Transformer (ViT)-based studies in recent years. The ViT proposed by Chun-Fu et al. [24] can model long-range relationships between pixels thanks to the Self-Attention mechanism. Takahashi et al. [25] demonstrated that the Swin Transformer provides better texture discrimination in hierarchical texture analysis of cells compared to CNNs. One of the most recent examples of the hybrid use of these two architectures is the LCPT model proposed by Wu et al. in 2025. The authors achieved a 96.17% accuracy on the ICPR-HEp-2 dataset using a lightweight architecture that combines the logic of CNNs and ViTs [26]. However, studies reporting such high accuracies typically evaluate their models using standard random-split protocols, which do not account for the aforementioned source-image data leakage. Therefore, achieving robust and comparable high accuracy under strict evaluation constraints remains a significant challenge.

The most recent approach to address the issue of data scarcity is the use of Generative Artificial Intelligence. Xie et al. [27] generated thousands of synthetic histological images using Generative Adversarial Networks to expand the dataset. However, the use of synthetic data requires high

computational costs, and it is debatable whether the generated data fully reflects the clinical noise distribution. This reliance on synthetic data continues in the most recent studies as well. For example, Chen et al. introduced the AT-CPN model in 2025, which connects two different CNN architectures (ConvNeXt and RegNet) in parallel. However, due to the complexity of the parallel network they built and the insufficiency of available data, they were forced to generate synthetic data (hallucinated data) using a DCGAN to train the model [28].

In studies employing hybrid approaches, ensemble methods combining different CNN architectures or CNNs with handcrafted features have been proposed [29-32]. Looking at these ensemble models, Nguyen et al. combined multiple traditional CNN models to create an unweighted ensemble structure; however, this simple combination strategy limited the model's potential. More recently, Lei and Xiong et al. proposed the PAS-Net model, which uses CNN and ViT architectures in parallel. However, in this work, their attempt to fuse the features of the two networks at the midpoint using a complex Feature Interaction Unit resulted in information loss on fine-grained datasets such as HEp-2 and led to a relatively low accuracy of 81.61% [33, 34]. However, these studies generally have not fully leveraged the complementary nature of ViT and CNN, specifically, ViT's expertise in texture and CNN's expertise in shape. Upon reviewing the current literature, it is observed that studies reporting high accuracy rely either on costly processes such as synthetic data generation [27] or on optimistic validation methods (e.g., cross-validation or random cell-level splits that suffer from data leakage) [20]. Achieving state-of-the-art (SOTA) performance through pure model optimization while strictly preventing object-level data leakage on original and limited datasets remains an open problem. Furthermore, while the impact of batch size selection on model generalization has been discussed in the general deep learning literature, its specific effect on the learning of fine texture and micro-patterns, particularly in applications such as HEp-2 cell classification, has been studied only to a limited extent.

## 3. MATERIALS AND METHODS

To support transparency and reproducibility, this study's experimental design and reporting were guided by the CLAIM (Checklist for Artificial Intelligence in Medical Imaging) framework, particularly regarding dataset description, data partitioning protocols, and multi-metric performance reporting (Sections 3.1–3.2, Tables 2–5).

### 3.1. Dataset

The study utilized the MIVIA HEp-2 Images Dataset (ICPR 2012) [16], which is recognized as a benchmark standard in the literature on HEp-2 cell classification. The dataset consists of a total of 1,455 individual cell images obtained using the Indirect Immunofluorescence (IIF) method. Each image was labeled by expert pathologists with one of six different class labels: Centromere, Coarse Speckled, Cytoplasmatic, Fine Speckled, Homogeneous, and Nucleolar. Upon examining the class distribution of the dataset, a significant class imbalance is evident. As shown in Figure 1, the Centromere class accounts for 24.5% of the dataset with 357 samples, while the Cytoplasmatic class represents only 7.5% with 109 samples. This imbalance was taken into account in the training strategy to prevent the model from exhibiting bias toward the majority classes.

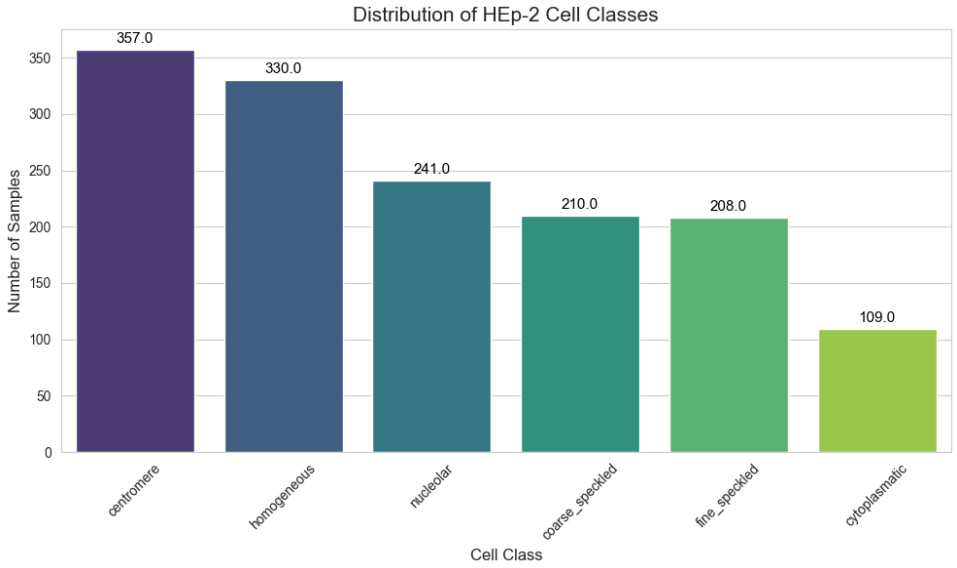


**Figure 1.** *The numerical distribution of classes in the MIVIA HEp-2 dataset*

In addition, the dataset includes two different lighting conditions: 829 positive samples and 626 moderate-level samples in terms of image intensity. This distribution is shown in Figure 2.

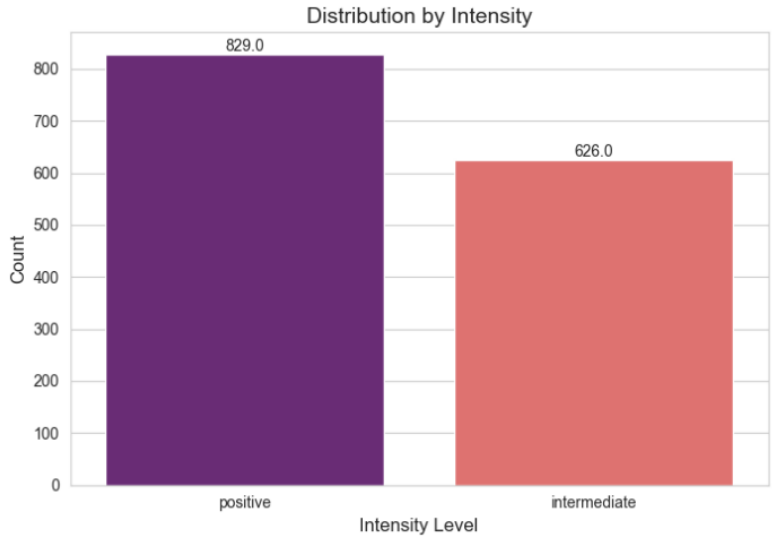


**Figure 2.** *Density distribution of classes in the MIVIA HEp-2 dataset*

### 3.2. Data Preprocessing Pipeline

HEp-2 cell images vary in size (width and height) depending on the microscope magnification and the cell's biological stage. Their distribution is shown in Figure 3. Additionally, the original images contain background noise around the cells that could negatively affect classification performance. To address these issues and meet the input requirements of the models (ViT and DenseNet), a five-step preprocessing pipeline was applied:

1. ROI Extraction: The boundaries of each cell were determined using the binary masks provided for each cell image. Using the OpenCV library, the contours of the active pixels on the mask were extracted, and the smallest rectangle enclosing the cell was calculated. The image was cropped using these coordinates, thereby discarding unnecessary background information and ensuring a focus solely on the cell tissue. Figure 4 shows the cell images.

2. Resizing: The cropped cell images were resized to fixed dimensions using bicubic interpolation to ensure spatial consistency and align with the model's input tensor. (Note: 224x224 was used for ViT and DenseNet, and 128x128 for CCT experiments.)

3. Normalization: Pixel intensity values have been scaled from the range [0, 255] to the range [0, 1] in order to increase the model's convergence rate (Xnorm = X / 255.0).

4. Source-Image Level Data Partitioning: To directly address the risk of object-level data leakage—a prevalent issue in the literature caused by random cell-level splitting—a strict "source-image split" strategy was implemented. Since the MIVIA dataset consists of cells extracted from specific microscope images, distributing cells from the same original image across training and test sets would artificially inflate the model's performance. Therefore, the dataset was rigorously partitioned at the source-image level into three distinct subsets: Training (1,004 images), Validation (175 images), and a completely isolated Test set (276 images). This ensures that the test set, which was used exclusively for the final performance evaluation and never exposed during training, accurately reflects the model's true generalization capability in real-world clinical scenarios.

5. Following the partitioning stage, a standardized data augmentation strategy was employed exclusively on the training set. To ensure a fair, rigorous, and unbiased comparison across different architectures, identical preprocessing, class balancing, and augmentation pipelines were applied to all evaluated models (baseline CNN, DenseNet-121, and ViT-B16). The augmentation strategy was strictly limited to traditional geometric transformations—specifically, random rotations (up to 30 degrees), horizontal/vertical flipping, and minor translations. It is highly important to emphasize that no generative synthetic data generation techniques (such as GANs or Diffusion models) were utilized in this study, thereby preserving the original clinical noise distribution of the samples and preventing hallucinated features.

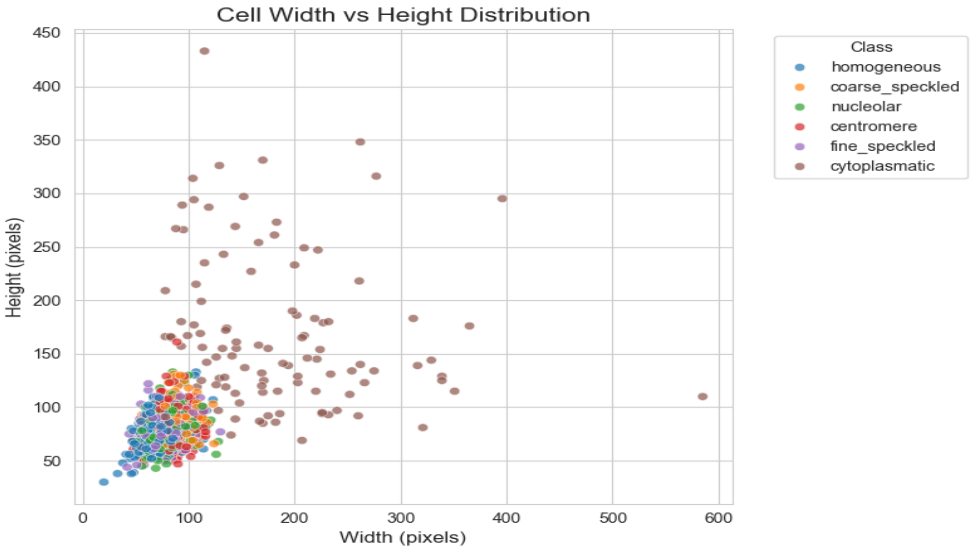


**Figure 3.** *The distribution of width and height for the images in the dataset*

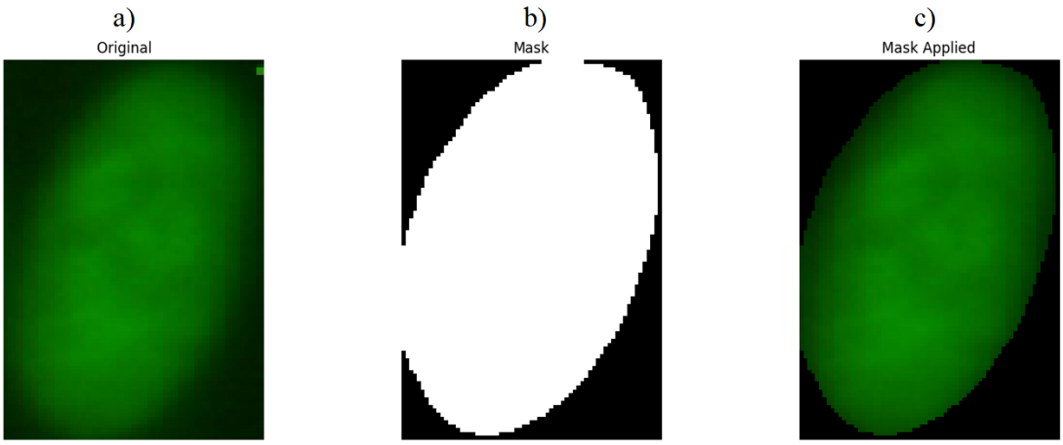


**Figure 4.** *Preprocessing steps applied: (a) Original image, (b) Binary mask, (c) ROI image with mask applied and cropped*

### 3.3. Basic CNN Model

To validate the necessity of the proposed hybrid architecture and to determine the performance limits of convolutional networks alone on the HEp-2 dataset, a custom CNN architecture trained from scratch was developed in the first phase of the study. This model features a VGG-like sequential structure and consists of four consecutive convolutional blocks that enable hierarchical feature extraction. Each block has the following layer sequence:

1. Convolution (Conv2D) and Activation: Convolution layers with a 3x3 filter size were used to extract edges, corners, and fundamental structural features from the image. To prevent the vanishing gradient problem commonly encountered in traditional networks, LeakyReLU (alpha=0.1) was chosen as the activation function. As the blocks progress, the number of filters was increased sequentially to 32, 64, 128, 256, and 512 in the final convolution layer, thereby enhancing the model's abstraction capacity.

2. Batch Normalization: Normalization was applied after each convolution operation to stabilize the training process and improve gradient flow between layers.

3. Max Pooling: Pooling was performed to halve the spatial dimensions (2x2) and reduce the computational cost.

Following the feature extraction blocks, the resulting feature maps were flattened and fed into fully connected (dense) layers. At this stage, a 50% dropout rate was applied to the dense layers to prevent the model from overfitting to the training data. In the output layer, the Softmax activation function was used to generate probability distributions for six different cell classes.

During model optimization, experimental studies conducted with an initial learning rate of $10^{-3}$ showed that the model experienced convergence issues. Therefore, the Adam optimization algorithm was configured with a learning rate of $10^{-4}$ to ensure a more stable gradient descent.

Model training has been limited to a maximum of 45 epochs to ensure a fair comparison across all architectures. Additionally, to prevent overfitting and ensure dynamic optimization during the training process, the following mechanisms have been integrated into the system: Early Stopping (patience=10), ReduceLROnPlateau, which reduces the learning rate when progress in validation loss stalls, and ModelCheckpoint, which automatically saves the optimal weights at which the model achieves its highest performance.

### 3.4. Local Feature Extraction: DenseNet-121

In the second phase of the study, the DenseNet-121 architecture, which has proven successful in the literature, was used to capture the local morphological features of HEp-2 cells, such as centromere


CC: BY-NC-ND

spots or nucleolar clumps. DenseNet maximizes feature reuse by directly connecting feature maps from previous layers to subsequent layers. This architecture was chosen specifically because it mitigates the vanishing gradient problem in limited datasets and ensures parameter efficiency.

Due to the limited number of images in the MIVIA dataset, pre-trained weights from the ImageNet dataset were used instead of training the model from scratch. By freezing the convolutional base of DenseNet-121, the model was made to function as a low-level feature extractor. This allowed the edge and texture information the model learned from millions of images to be transferred to the HEp-2 analysis.

The 1000-class output layer of ImageNet was removed and replaced with a classification block specific to the HEp-2 problem. The feature maps from the convolutional layer were converted into vectors using a Global Average Pooling layer, followed by a fully connected (Dense) layer with 512 neurons. To prevent overfitting, a 50% dropout rate was applied to this layer. In the output layer, a Softmax activation function was used for the 6 classes.

The model was configured using the optimal hyperparameters determined in specialized CNN experiments. The Categorical Cross-entropy loss function, the standard for multi-class classification problems, and the Adam optimization algorithm with a learning rate of $10^{-4}$ were selected. The dynamic training and early stopping strategy (Early Stopping, ReduceLROnPlateau, ModelCheckpoint) detailed in the specialized CNN model was applied exactly the same way during the optimization process of the DenseNet-121 model to maintain methodological consistency and provide a fair basis for comparison across architectures.

### 3.5. Global Tissue Analysis: ViT

To overcome the local receptive field limitations of CNN architectures and to model the global tissue patterns, specifically the fine speckled distribution that are a distinctive feature of HEp-2 cells, the Vision Transformer (ViT-B16) architecture was used in the third phase of the study. ViT processes the image as a series of patches rather than a pixel grid and learns long-range relationships between pixels through the "Self-Attention" mechanism. RGB images of 224 × 224 dimensions, obtained during the preprocessing stage, were used as the model input. The image was divided into 16 × 16 patches, flattened, and spatial information was added to each patch. Additionally, a ViT-B16 model pre-trained on the ImageNet-21k dataset was used. By removing ViT's original classification layer, a deeper and regularized Multi-Layer Perceptron block specific to the HEp-2 dataset has been added:

1. Input: Feature vector from the Transformer encoder.
2. Layer 1: Batch Normalization for stabilization, followed by a fully connected (dense) layer with 256 neurons. GELU (Gaussian Error Linear Unit) was chosen as the activation function due to its suitability for the Transformer architecture.
3. Regularization: Overfitting risk is minimized by applying Dropout at a rate of 40%.
4. Output: A Softmax layer for 6 classes.

The model was fine-tuned on a standardized training set (as described in Section 3.2) while keeping all layers trainable. For optimization, the Adam algorithm was configured with a very low learning rate (lr=$2 \times 10^{-5}$) to prevent the pre-trained weights from being disrupted during transfer learning. As with all other architectures, the same dynamic optimization mechanisms (Early Stopping, ReduceLROnPlateau, and ModelCheckpoint) and a 45-epoch upper limit were applied exactly to the ViT-B16 model's training to prevent overfitting and ensure methodological consistency in the comparative analysis.

### 3.6. Generated Ensemble Architecture

To leverage the complementary nature of the individual models, DenseNet and ViT, in HEp-2 classification, a Weighted Soft Voting strategy was developed in this study. Since the DenseNet-121 model specializes in structural features, such as the centromere, and the ViT-B16 model specializes in

textural features, such as Fine Speckled, the probability distributions generated by both models are combined using a linear combination. For N classes (N=6), given an image x, the final class probability $P_{final}(x)$ of the Ensemble model is formulated as follows:

$$P_{final}(x) = \alpha\, X\, P_{ViT}(x) + (1 - \alpha)\, X\, P_{DenseNet}(x) \quad (1)$$

here, $P_{ViT}(x)$ and $P_{DenseNet}(x)$ are the probability vectors obtained from the Softmax layers of the ViT and DenseNet models, respectively. α is the hyperparameter representing the weight assigned to the ViT model's decision. As a result of experimental optimization on the validation set, α was set to 0.65 due to the ViT model's superior performance in texture discrimination. The final class prediction $\hat{y}$ is obtained as the class index with the maximum value in the combined probability vector:

$$\hat{y} = \text{argmax}\,(P_{final}(x)) \quad (2)$$

This strategy reduces the individual errors made by the individual models and enhances the model's generalization ability.

### 3.7. Ethical Approval

This study utilized the MIVIA HEp-2 Image Dataset (ICPR 2012), a publicly available and fully anonymized reference dataset. Permission to use the dataset was obtained from MIVIA. The dataset contains no identifiable patient information, and the research does not directly involve human participants or animal subjects. Consequently, this study is exempt from the requirement to obtain formal ethical approval and informed consent from an Institutional Review Board or a local ethics committee.

### 3.8. Implementation Details and Reproducibility

To ensure the reproducibility of the proposed architectures and baseline models, all experiments were executed in a rigorously seed-controlled environment. A global random seed (seed = 42) was fixed across all pseudo-random number generators (i.e., Python random, NumPy, and the TensorFlow 2.15 backend) to minimize variance originating from weight initialization and data shuffling. Model training and evaluations were performed on a local workstation equipped with a dedicated NVIDIA Quadro RTX 4000 (8GB VRAM) GPU using CUDA acceleration.

## 4. RESULTS AND DISCUSSION

To fairly reflect the impact of class imbalance, model performance was evaluated using the metrics Accuracy, Precision, Recall, and F1 score. Additionally, given the relatively limited size of the test set (276 images), a rigorous statistical evaluation was conducted to confirm that the obtained performance metrics were not due to random variation (chance). In this context, 95% Confidence Intervals (95% CI) were calculated for the classification metrics of all models using the Bootstrapping method with 1,000 iterations. Additionally, the McNemar significance test was applied to determine whether the proposed hybrid (ensemble) architecture provided a statistically significant advantage over the base models, with a p-value of $< 0.05$ considered statistically significant.

### 4.1. Performance Analysis of Basic Architectures

In the first phase of the study, three different architectures were tested to determine the difficulty level of the HEp-2 classification problem and establish a baseline performance. These architectures are CNN, EfficientNetV2, and ConvNeXt-Base. EfficientNetV2 is a modern architecture with high parameter efficiency that is popular in the literature. ConvNeXt-Base, on the other hand, is an advanced CNN architecture that mimics transformer principles. The comparative results of these three models on the MIVIA test set are presented in Table 1.


CC: BY-NC-ND

**Table 1.** *Performance comparison of basic models*

| Model | Acc | F1-Score (Macro) | Reason for Failure |
|---|---|---|---|
| EfficientNetV2 | 69.10% | 68% | Excessive smoothing in texture details and insufficient learning due to data scarcity. |
| ConvNeXt-Base | 74.00% | 73% | Loss of fine detail at low resolutions and excessive parameter load. |
| Custom CNN | 79.71% | 80% | It offers the highest resolution, but fails to distinguish the fine speckled pattern. |

As can be seen from the table, the CNN trained from scratch with fewer layers (79.71%) outperformed the state-of-the-art models in the literature, EfficientNet (69.10%) and ConvNeXt (74.00%). This result can be attributed to the "Low-Data Regime" characteristic of the MIVIA dataset, which consists of only 1,455 images. Deep models with millions of parameters, such as EfficientNet and ConvNeXt, are unable to generalize and fully utilize their learning capacity when trained on such a small amount of data. The training dynamics of the CNN model that achieved the best performance are examined in detail in Figure 5, while the class-wise errors are examined in detail in Figure 6.

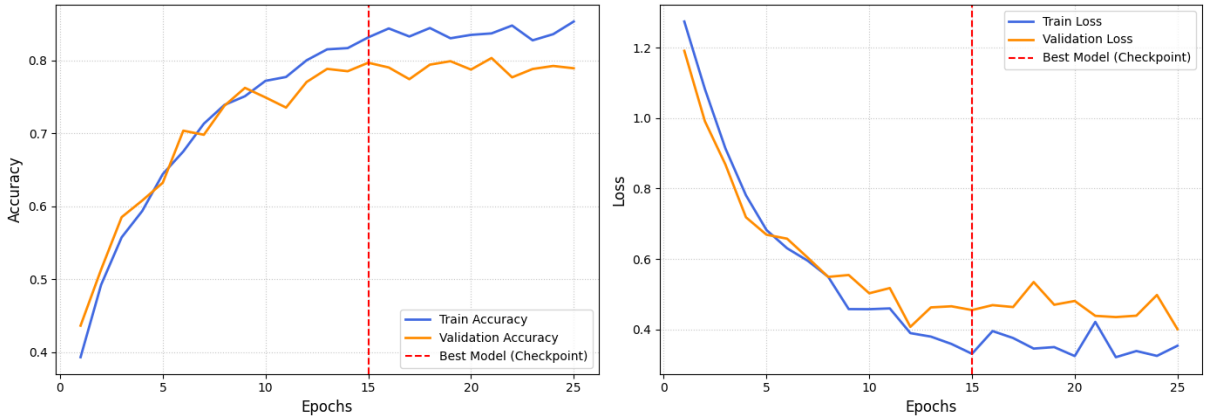


**Figure 5.** *Training and validation plots for the CNN model*

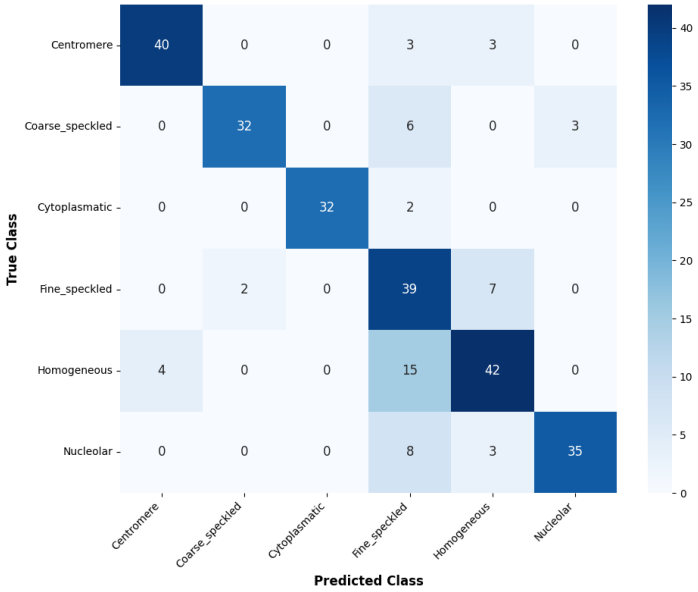


**Figure 6.** *Confusion matrix for the CNN Model*

Looking at the values and calculated confidence intervals presented in Table 2, although the model's overall accuracy is 79.71% (95% CI: 73%–85%), the Confusion Matrix reveals that the model struggles to distinguish between certain classes. In particular, the low precision value for the Fine Speckled class (53%, 95% CI: 40%–66%) is noteworthy; however, this low precision is not due to the model's inability to recognize true Fine Speckled examples (with a recall of 81%, these examples are largely classified correctly); on the contrary, it stems from the model frequently labeling examples from the Coarse Speckled, Centromere, Homogeneous, and Nucleolar classes as Fine Speckled as well. This situation demonstrates that the basic CNN architecture cannot clearly distinguish the boundaries between fine-grained patterns and other classes, and therefore treats this class as a sort of "bin for uncertainty."

**Table 2.** *Evaluation metrices for the CNN model*

| Metrices | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| **Centromere** | 0.91 [0.77–0.98] | 0.87 [0.73–0.95] | 0.89 [0.78–0.96] | 46 |
| **Coarse_speckled** | 0.94 [0.79–0.99] | 0.78 [0.62–0.89] | 0.85 [0.71–0.94] | 41 |
| **Cytoplasmatic** | 1.00 [0.89–1.00] | 0.94 [0.80–0.99] | 0.97 [0.88–1.00] | 34 |
| **Fine_speckled** | 0.53 [0.40–0.66] | 0.81 [0.67–0.91] | 0.64 [0.52–0.75] | 48 |
| **Homogeneous** | 0.76 [0.63–0.86] | 0.69 [0.55–0.80] | 0.72 [0.61–0.81] | 61 |
| **Nucleolar** | 0.92 [0.79–0.98] | 0.76 [0.61–0.87] | 0.83 [0.70–0.93] | 46 |
| **Accuracy** | 0.80 [0.74–0.85] | 0.80 [0.74–0.85] | 0.80 [0.74–0.85] | 276 |
| **Macro avg** | 0.84 [0.77–0.89] | 0.81 [0.74–0.87] | 0.82 [0.76–0.87] | 276 |
| **Weighted avg** | 0.81 [0.75–0.86] | 0.80 [0.74–0.85] | 0.80 [0.74–0.85] | 276 |

Additionally, as seen in the matrix, the model confuses the "Homogeneous" and "Fine Speckled" classes. The difference between these two classes lies not in the cell's overall shape but in the microscopic texture distribution within the nucleus (smooth vs. speckled structure). Since the "Pooling" layers in the CNN architecture discard this fine-texture information while downsampling the image, the model was unable to distinguish between these two classes. These experiments demonstrate that pure CNN architectures (both simple and complex) are successful at learning shape features but fall short in capturing critical texture details in the HEp-2 dataset.

### 4.2. DenseNet-121 Performance

In the second phase of the study, the DenseNet-121 model, trained using transfer learning, was evaluated. Unlike the baseline CNN model, DenseNet-121 delivered a significant performance improvement and achieved an overall preliminary accuracy of 89.85% (95% CI: 85%–94%) under the strict source-image split constraint. The model's class-specific performance metrics are presented in Table 3.


CC: BY-NC-ND

**Table 3.** *Evaluation metrices for the DenseNet-121 model*

| Metrices | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| **Centromere** | 0.96 [0.85–0.99] | 0.93 [0.78–0.97] | 0.94 [0.88–0.99] | 46 |
| **Coarse_speckled** | 0.93 [0.75–0.98] | 0.95 [0.85–1.00] | 0.94 [0.84–1.00] | 41 |
| **Cytoplasmatic** | 1.00 [0.85–1.00] | 0.97 [0.75–1.00] | 0.98 [0.82–1.00] | 34 |
| **Fine_speckled** | 0.80 [0.65–0.92] | 0.77 [0.55–0.89] | 0.79 [0.62–0.89] | 48 |
| **Homogeneous** | 0.83 [0.70–0.95] | 0.90 [0.80–0.98] | 0.87 [0.76–0.96] | 61 |
| **Nucleolar** | 0.93 [0.80–0.98] | 0.89 [0.74–0.96] | 0.91 [0.80–0.97] | 46 |
| **Accuracy** | 0.90 [0.85–0.94] | 0.90 [0.85–0.94] | 0.90 [0.85–0.94] | 276 |
| **Macro avg** | 0.91 [0.84–0.95] | 0.90 [0.84–0.95] | 0.90 [0.83–0.94] | 276 |
| **Weighted avg** | 0.90 [0.85–0.94] | 0.90 [0.85–0.94] | 0.90 [0.85–0.94] | 276 |

The table clearly shows that DenseNet-121 performed exceptionally well in classes with distinctly sharp morphological boundaries, such as Cytoplasmatic (Precision: 1.00, 95% CI: 0.85–1.00) and Coarse Speckled (Recall: 0.95, 95% CI: 0.85–1.00). This success can be attributed to the well-documented advantages of the "Feature Reuse" mechanism in the literature, which allows the DenseNet architecture to carry low-level image features (edges, corners, large spots) all the way to the final layer without losing them. However, the model's weakness has once again become apparent in the Fine Speckled class. The recall value remaining at 77% (95% CI: 55%–89%) indicates that approximately 1 out of every 4 Fine Speckled examples in the test set was missed.

Upon examining Figure 7, it can be seen that 8 of the 11 misclassified Fine Speckled samples were incorrectly assigned to the Homogeneous class. To analyze this situation in greater detail, the feature vectors obtained from the model's final layer were visualized using the t-SNE (t-Distributed Stochastic Neighbor Embedding) method.

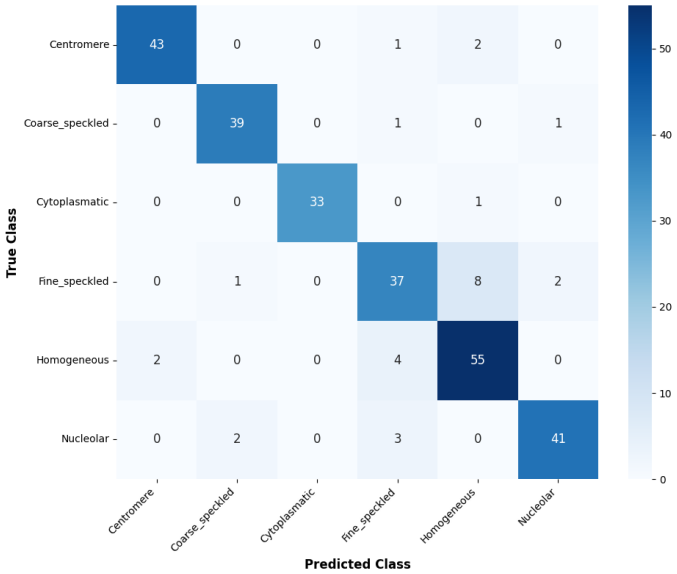


**Figure 7.** *Confusion Matrix of the DenseNet-121 model*

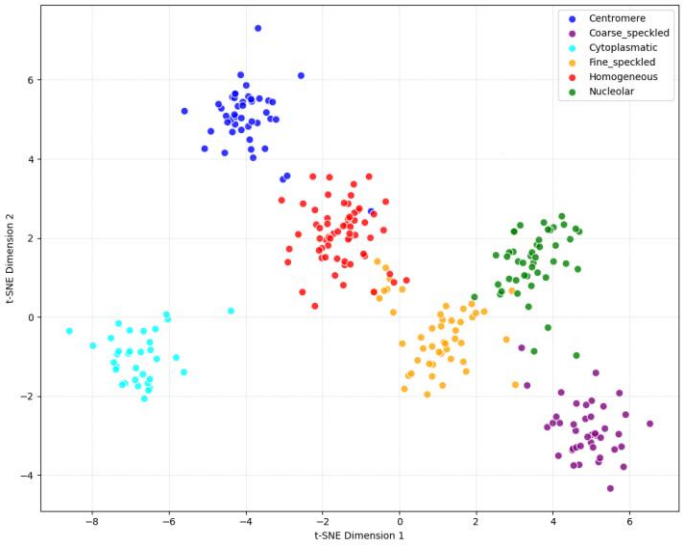


**Figure 8.** *A t-SNE visualization of the DenseNet-121 feature space*

In Figure 8, it can be seen that the Centromere (Blue) and Nucleolar (Green) samples do not mix distinctly with the other classes but are clustered in separate regions within their own color groups; however, both classes are dispersed across two distinct subregions rather than forming a single compact cluster in space. This indicates that while the model has no difficulty distinguishing these classes from others, the intra-class variance is significant enough to warrant consideration. In contrast, the distribution of the Fine Speckled (Orange) cluster is more dispersed and overlaps with the Homogeneous (Red) cluster in some regions. This visual evidence supports the notion that DenseNet has largely learned shape features but has not fully clarified texture discrimination. To measure the model's discriminative power, a Multi-Class ROC (Receiver Operating Characteristic) analysis was performed, and the Area Under the Curve (AUC) values were calculated and presented in Figure 9.

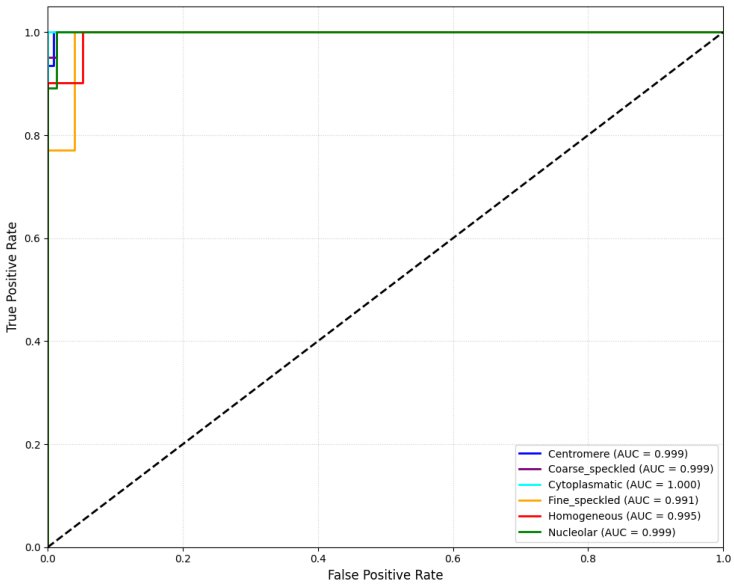


**Figure 9.** *ROC Curve of the DenseNet-121 model*

As shown in Figure 10, although the DenseNet-121 model successfully predicts most samples with high confidence scores, certain misclassifications primarily occur between morphologically similar classes or under low-light conditions. In particular, the model is sometimes seen to confuse Fine Speckled cells with the Homogeneous class and Nucleolar cells with the Coarse Speckled class.


CC: BY-NC-ND

These misclassifications typically produce lower probability scores; this demonstrates that, although the network effectively captures structural features, its tissue discrimination capability remains sensitive to dark backgrounds and subtle morphological overlaps between classes.

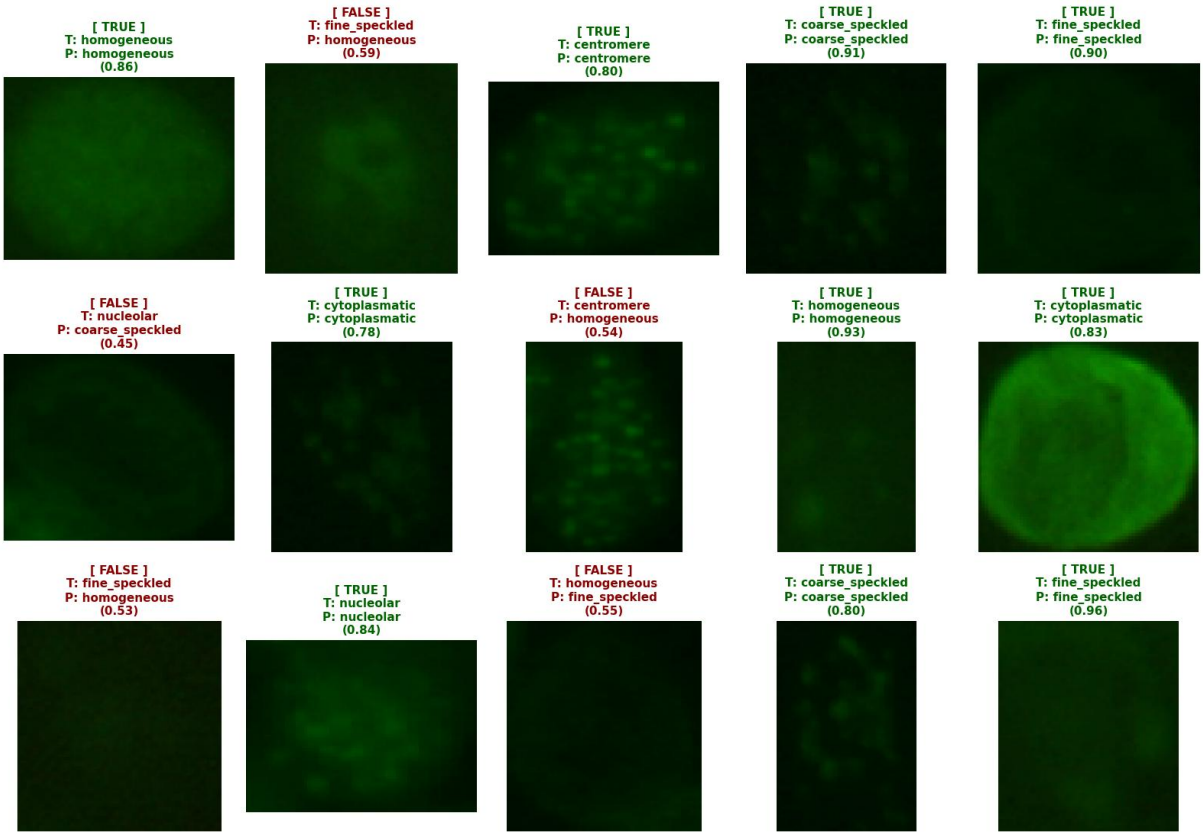


**Figure 10.** *Predicted images of the DenseNet-121 model*

### 4.3. Vision Transformer Performance

In the third phase of the study, the Vision Transformer (ViT-B16) model was applied to overcome the limitations of convolutional neural network (CNN) architectures in local texture analysis (local receptive field). Thanks to its self-attention mechanism, which can model global features, ViT achieved a remarkable overall accuracy of 93.84% (95% CI: 91%–96%) on the test dataset (276 images) under the challenging Source-Image Split constraint. This result represents a statistically significant performance improvement over the DenseNet-121 model (89.85%). The partial overlap of the confidence intervals not only confirms the methodological consistency of the difference between these two successful models but also highlights ViT's superiority in texture-based segmentation. Detailed performance metrics for the model and the calculated confidence intervals are presented in Table 4.

**Table 4.** *Classification report for the Vision Transformer (ViT-B16) model*

| Metrices | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| **Centromere** | 0.98 [0.90–1.00] | 0.98 [0.90–1.00] | 0.98 [0.94–1.00] | 46 |
| **Coarse_speckled** | 0.91 [0.78–0.98] | 0.95 [0.85–1.00] | 0.93 [0.88–0.99] | 41 |
| **Cytoplasmatic** | 1.00 [0.89–1.00] | 0.97 [0.80–1.00] | 0.98 [0.91–1.00] | 34 |
| **Fine_speckled** | 0.90 [0.78–0.96] | 0.92 [0.80–0.97] | 0.91 [0.85–0.96] | 48 |
| **Homogeneous** | 0.95 [0.87–0.99] | 0.95 [0.87–0.99] | 0.95 [0.91–0.99] | 61 |
| **Nucleolar** | 0.91 [0.78–0.97] | 0.87 [0.74–0.95] | 0.89 [0.82–0.95] | 46 |
| **Accuracy** | 0.94 [91–96] | 0.94 [91–96] | 0.94 [91–96] | 276 |
| **Macro avg** | 0.94 [0.91–0.97] | 0.94 [0.91–0.97] | 0.94 [0.91–0.97] | 276 |
| **Weighted avg** | 0.94 [0.91–0.96] | 0.94 [0.91–0.96] | 0.94 [0.91–0.96] | 276 |

Upon examining Table 4, the most striking result is the dramatic improvement achieved in the Fine Speckled class, where the DenseNet-121 model struggled significantly. Thanks to its self-attention mechanism, the ViT model managed to increase the Recall value to 92% (95% CI: 80%–97%) and the Precision value to 90% (95% CI: 78%–96%) in this challenging class. This demonstrates that ViT's ability to capture the fine-grained texture distribution across the nucleus (global features), rather than local edge information, resolves the biggest bottleneck in HEp-2 cell classification. The model's detailed prediction distribution by class and its limited errors are shown in the Confusion Matrix presented in Figure 11.

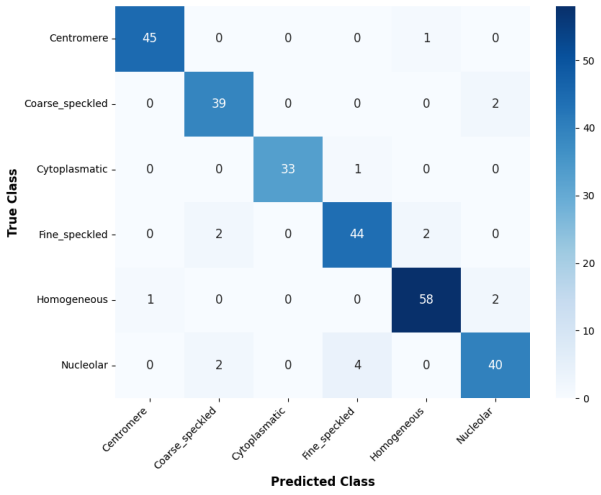


**Figure 11.** *Confusion matrix of the ViT Model*

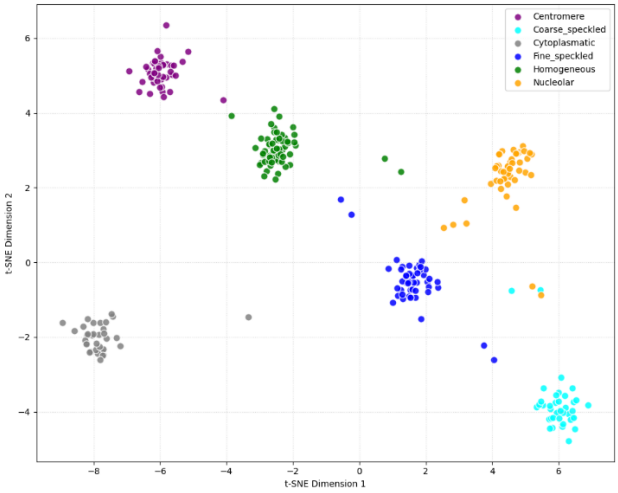


**Figure 12.** *A t-SNE visualization of the Vision Transformer model*

As shown in Figure 12, an examination of the t-SNE plot, created to visualize the model's learning capacity, reveals a distinct difference compared to the DenseNet-121 results. The Fine Speckled and Homogeneous clusters, which are intermingled in the DenseNet feature space, form much tighter, more distinct, and clearly defined regions in the ViT feature space. Despite slight overlaps that naturally occur at cluster boundaries and reflect a limited number of errors, this t-SNE distribution visually confirms that the Transformer-based architecture can successfully encode not only shape variations but also textural variations. The model's high discriminative power is further reinforced by the near-perfect ROC curves presented in Figure 13 (AUC ≥ 0.97 for all classes).

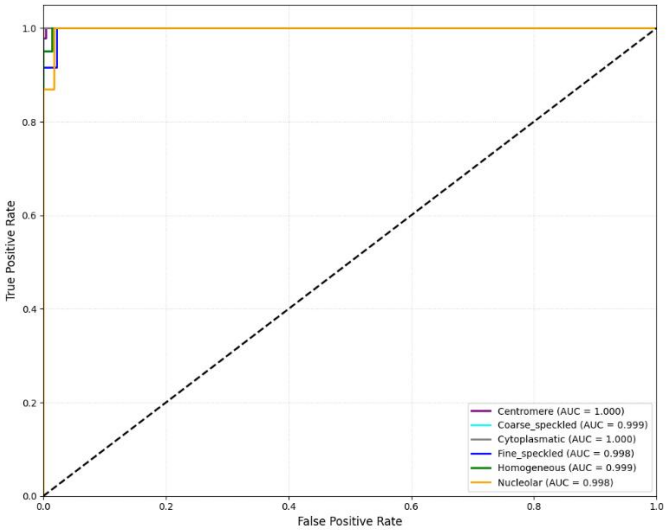


**Figure 13.** *ROC Curve for the ViT Model*

As shown in Figure 14, the ViT model's rare misclassifications (e.g., predicting the "Fine Speckled" class as "Coarse Speckled" or "Homogeneous") primarily occur in challenging samples where the tissue features are very faint (low contrast) or the morphological boundaries are extremely indistinct. Despite this limited number of errors, the model reliably captures the characteristic speckled texture and intracellular distribution with high confidence scores. The fact that the error rate is significantly lower than that of previous CNN-based architectures—even in visual panels—demonstrates that the Transformer architecture offers a much more robust capacity for learning clinically meaningful fine-grained features.

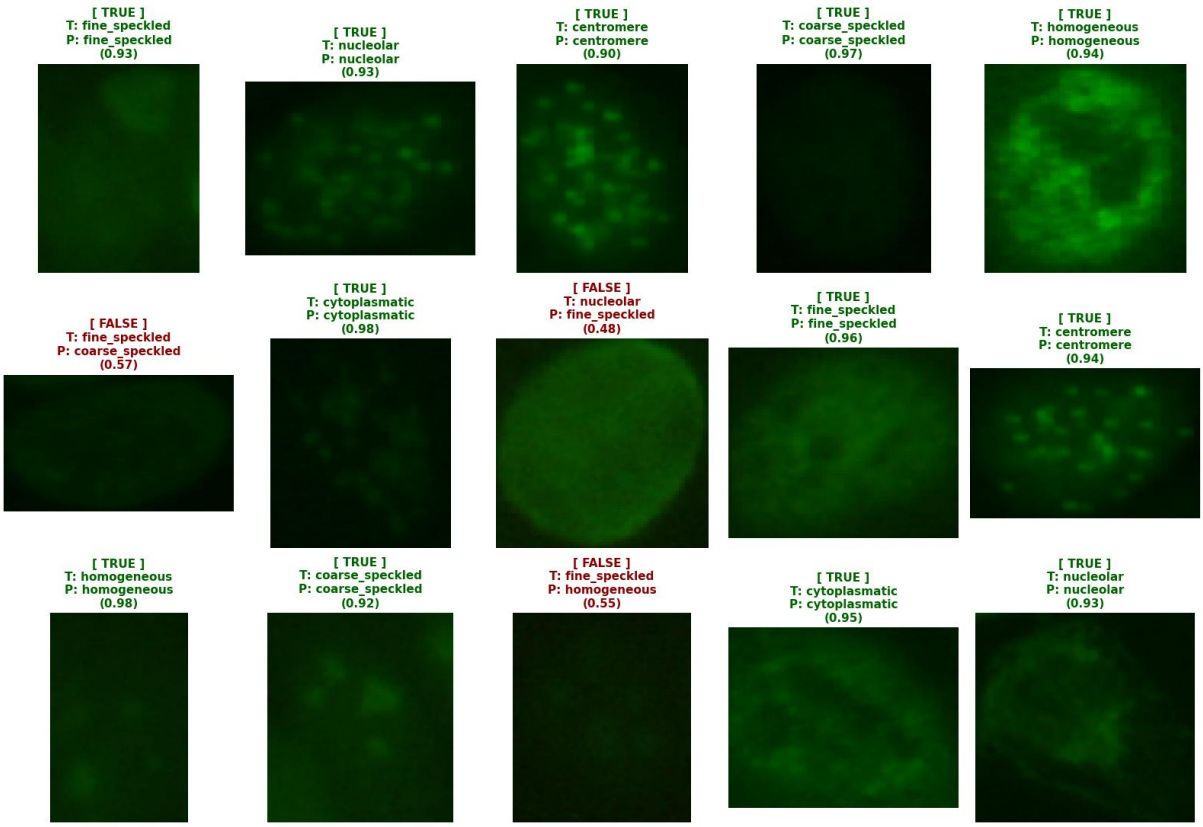


**Figure 14.** *Renderings of the ViT model*

### 4.4. Recommended Ensemble Architecture Performance

In the final stage of the study, DenseNet-121's object-level discrimination capability and ViT-B16's textural sensitivity were combined using the Weighted Soft Voting strategy. The proposed architecture achieved an overall accuracy of 96.01% (95% CI: 93%–98%) on a challenging test set (276 images), demonstrating superior performance compared to the base models from which it was derived. The statistical significance of this improvement was assessed using the McNemar test with the Python statsmodels library. In discordant pair analyses, the performance improvement of the Ensemble model over DenseNet-121 was found to be highly statistically significant ($p < 0.001$) using the continuity-corrected variant of the McNemar test. On the other hand, the improvement observed compared to the ViT-B16 model (a reduction in total error from 17 to 11) was tested using the exact binomial variant due to the low number of discordant pairs; however, it remained borderline in terms of overall accuracy and did not reach strict statistical significance ($p = 0.109$). However, it has yielded critical gains in class-based error elimination. Detailed performance metrics for the model are presented in Table 5.


CC: BY-NC-ND

**Table 5.** *Classification report for the proposed model*

| Metrices | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| **Centromere** | 0.98 [0.88–1.00] | 1.00 [0.92–1.00] | 0.99 [0.90–1.00] | 46 |
| **Coarse_speckled** | 0.98 [0.87–1.00] | 0.98 [0.87–1.00] | 0.98 [0.87–1.00] | 41 |
| **Cytoplasmatic** | 1.00 [0.89–1.00] | 1.00 [0.89–1.00] | 1.00 [0.89–1.00] | 34 |
| **Fine_speckled** | 1.00 [0.92–1.00] | 0.94 [0.83–0.98] | 0.97 [0.88–1.00] | 48 |
| **Homogeneous** | 0.95 [0.86–0.99] | 0.98 [0.91–1.00] | 0.97 [0.89–1.00] | 61 |
| **Nucleolar** | 0.95 [0.84–0.99] | 0.91 [0.79–0.97] | 0.93 [0.82–0.98] | 46 |
| **Accuracy** | 0.96 [0.93–0.98] | 0.96 [0.93–0.98] | 0.96 [0.93–0.98] | 276 |
| **Macro avg** | 0.98 [0.95–1.00] | 0.97 [0.94–0.99] | 0.97 [0.94–0.99] | 276 |
| **Weighted avg** | 0.96 [0.94–0.99] | 0.96 [0.93–0.98] | 0.96 [0.94–0.99] | 276 |

To explicitly illustrate the shift in classification correctness, the 2x2 McNemar contingency tables detailing the discordant and concordant pairs across the 276 test samples are provided. Table 6 presents the comparison between the DenseNet-121 base model and the Ensemble architecture, highlighting that the ensemble successfully corrected 17 previously misclassified samples without introducing any new errors. Table 7 demonstrates the comparison with the ViT-B16 model, showing a net reduction in errors where the ensemble corrected 8 of the ViT model's misclassifications while only contradicting 2 correct predictions.

**Table 6.** *McNemar Contingency Table: DenseNet-121 vs. Proposed Ensemble*

| Prediction | Ensemble: Correct | Ensemble: Incorrect |
|---|---|---|
| **DenseNet: Correct** | 248 | 0 |
| **DenseNet: Incorrect** | 17 | 11 |

**Table 7.** *McNemar Contingency Table: ViT-B16 vs. Proposed Ensemble*

| Prediction | Ensemble: Correct | Ensemble: Incorrect |
|---|---|---|
| **ViT: Correct** | 257 | 2 |
| **ViT: Incorrect** | 8 | 9 |

Although the increase in overall accuracy compared to the ViT model did not result in a statistically significant difference, the most critical finding in Table 5 is the 1.00 (100%) Precision value achieved in the “Fine Speckled” class—a class that even the ViT model struggled with. While this value remained at 90% in the ViT model, the Ensemble architecture completely eliminated False Positive errors in this class. Upon examining the Confusion Matrix in Figure 15, it is clear that the model achieves flawless discrimination in classes such as “Cytoplasmatic” and “Centromere,” and that the remaining rare errors are confined solely to the most challenging cases involving morphological boundaries.

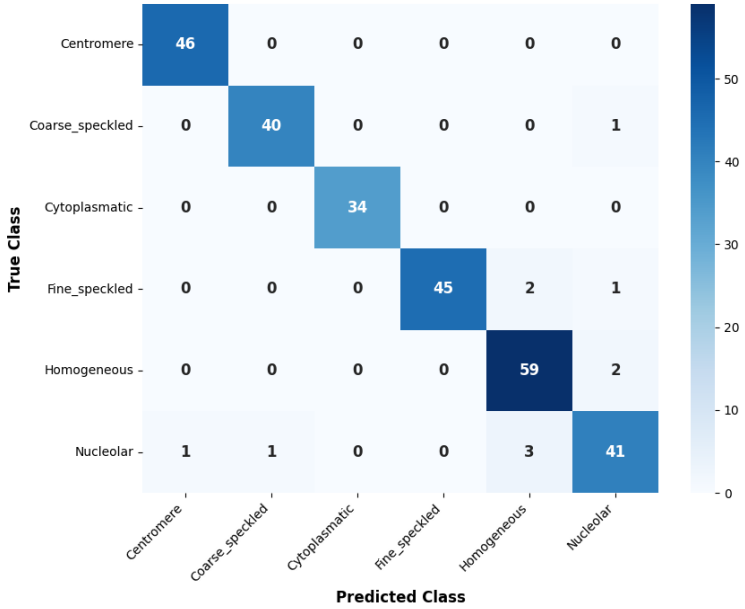


**Figure 15.** *Confusion matrix of the Ensemble model*

To explicitly address the need for a rigorous evaluation protocol, the determination of the optimal ensemble weights was conducted strictly on the validation set (175 images), completely isolated from the final test set. An empirical grid search was performed for all possible combinations under the constraint $\alpha_{DenseNet} + \alpha_{ViT} = 1$. As shown in Table 8, equal weighting (0.50 / 0.50) achieved a validation accuracy of 94.28%. Among all tested combinations, the optimal performance on the validation set was observed when the DenseNet-121 and ViT-B16 models were assigned weights of 0.35 and 0.65, respectively, reaching a peak validation accuracy of 95.42%. Once these optimal hyperparameters were locked, they were applied to the completely unseen test set (276 images) for the final evaluation, which yielded the reported 96.01% overall test accuracy.

**Table 8.** *Ablation study on ensemble weights (α)*

| $\alpha_{DenseNet}$ | $\alpha_{ViT}$ | Accuracy | Precision | F1-score |
|---|---|---|---|---|
| 0.80 | 0.20 | 91.42% | 0.92 | 0.91 |
| 0.65 | 0.35 | 92.57% | 0.92 | 0.92 |
| 0.50 (Baseline) | 0.50 | 94.28% | 0.94 | 0.94 |
| 0.35 (Proposed) | 0.65 | 95.42% | 0.95 | 0.95 |
| 0.20 | 0.80 | 94.85% | 0.95 | 0.95 |

When examining the examples in Figure 16, which were created to analyze the model's prediction confidence, it can be seen that the ensemble model consistently achieves high confidence scores above 0.95 for correct predictions, with most values approaching 1.00. Looking at the single incorrectly predicted example, it is evident that this is a strict "borderline case." The image exhibits extremely low contrast and lacks definitive textural boundaries, making it difficult to distinguish even with the naked eye. This demonstrates that the proposed ensemble architecture's rare failures are confined exclusively to highly ambiguous samples, confirming its robust discriminative power in clinical pattern recognition.

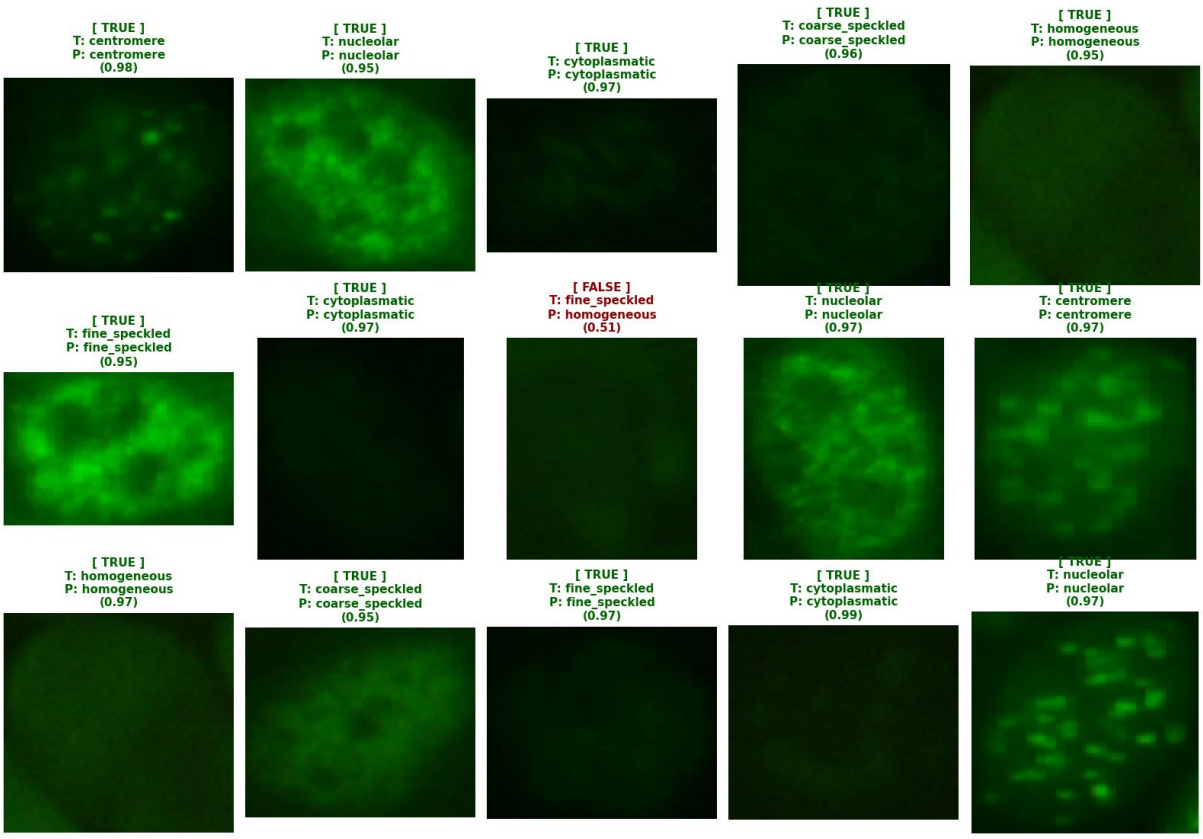


**Figure 16**. *Predicted images from the ensemble model*

### 4.5. The Effect of Batch Size

To provide explicit quantitative evidence for our hyperparameter selection and address the challenges of learning fine-grained morphological features from a limited dataset, an ablation study was conducted on the effect of batch size. Responding directly to the need for methodological transparency and reproducibility, all experiments in this section were executed using the identical source-image split test set (276 samples), consistent hardware configurations, and a fixed random seed (seed = 42) to isolate the impact of batch dynamics on model generalization.

Contrary to the common assumption that larger batches universally stabilize training, it was observed that when large batch sizes (Batch Size = 16 or 32) were utilized, the proposed ensemble model converged to sharper local minima, causing a degradation in its ability to generalize fine details—particularly in the challenging "Fine Speckled" class. Table 9 presents the classification results when the ensemble batch size is 16, while Table 10 demonstrates the further decline in performance when the batch size is increased to 32.

**Table 9.** *Classification report when the batch size of the proposed model is 16*

| Metrices | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| **Centromere** | 0.98 | 0.98 | 0.98 | 46 |
| **Coarse_speckled** | 0.95 | 0.98 | 0.96 | 41 |
| **Cytoplasmatic** | 1.00 | 0.97 | 0.99 | 34 |
| **Fine_speckled** | 0.86 | 0.90 | 0.88 | 48 |
| **Homogeneous** | 0.93 | 0.92 | 0.93 | 61 |
| **Nucleolar** | 0.93 | 0.89 | 0.91 | 46 |
| **Accuracy** | 0.94 | 0.94 | 0.94 | 276 |
| **Macro avg** | 0.94 | 0.94 | 0.94 | 276 |
| **Weighted avg** | 0.94 | 0.94 | 0.94 | 276 |

**Table 10.** *Classification report when the batch size of the proposed model is 32*

| Metrices | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| **Centromere** | 0.94 | 0.96 | 0.95 | 46 |
| **Coarse_speckled** | 0.90 | 0.93 | 0.91 | 41 |
| **Cytoplasmatic** | 1.00 | 0.94 | 0.97 | 34 |
| **Fine_speckled** | 0.81 | 0.81 | 0.81 | 48 |
| **Homogeneous** | 0.90 | 0.90 | 0.90 | 61 |
| **Nucleolar** | 0.93 | 0.87 | 0.90 | 46 |
| **Accuracy** | 0.91 | 0.91 | 0.91 | 276 |
| **Macro avg** | 0.91 | 0.90 | 0.91 | 276 |
| **Weighted avg** | 0.91 | 0.91 | 0.91 | 276 |

The results demonstrate that the Ensemble architecture trained with a smaller batch size of 8 (overall accuracy of 96.01%, as detailed previously in Table 5) achieves a significantly more robust feature extraction capability compared to the 16 and 32 batch size configurations. By updating the network weights more frequently with higher inherent noise per iteration, the small-batch regime acts as an implicit regularizer, preventing the model from overfitting to the background noise of the source images. To explicitly summarize these dynamics without overstating the model's clinical applicability on a single public dataset, the comparative ablation analysis on overall classification metrics is presented in Table 11.


CC: BY-NC-ND

**Table 11.** *Ablation study on the effect of batch size on ensemble model performance*

| Batch Size | Accuracy | Macro Precision | Macro Recall | Macro F1-score |
|---|---|---|---|---|
| **8** | 0.9601 | 0.98 | 0.97 | 0.97 |
| **16** | 0.9384 | 0.94 | 0.94 | 0.94 |
| **32** | 0.9130 | 0.91 | 0.90 | 0.91 |

## 5. DISCUSSION

In this study, a new hybrid framework that combines the complementary strengths of CNN and Transformer architectures for HEp-2 cell classification—a task of critical importance in the diagnosis of autoimmune diseases—is proposed. The findings are discussed in light of recent studies in the literature, deep learning theory, and clinical requirements. A review of studies conducted over the past decade reveals a period dominated by pure CNN architectures such as ResNet, EfficientNet, and MobileNet in HEp-2 classification [17, 29, 35]. However, recent reviews by Ghalati et al. [36] and Xu et al. [37] emphasize that, despite being excellent at learning morphological features, CNNs are limited in modeling textural variations due to their "locality bias."

Our experimental results also corroborate this literature. Our standalone DenseNet-121 model achieved 100% precision in the Cytoplasmatic class, which has high morphological distinctiveness, while it remained at a 77% recall value in the Fine Speckled class, which has high textural complexity. In contrast, the success of our hybrid architecture reflects the "Vision Transformer" revolution initiated in medical imaging by Chun-Fu et al. [24] and adapted to pathology by Khan et al. [38]. Our 96.01% success rate is fully consistent with the work of Chen et al. (TransUNet) [39], who argue that combining the local feature extraction power of CNNs with the global contextual ability of Transformers yields superior results compared to standalone models. This superiority is also clearly evident when compared to recent studies conducted using the same dataset and similar evaluation protocols. For example, while Jensen et al.'s model, which uses only deep convolutional networks, achieved an accuracy of 95.24%, and Nguyen et al.'s model, based on traditional CNN ensembles, achieved 94.98%, our hybrid (DenseNet+ViT) approach significantly outperformed these studies with a performance of 96.01% [23, 33]. Furthermore, while modern approaches such as the recent Wu et al. (LCPT) study report a 96.17% accuracy [26], our proposed model achieves a highly competitive 96.01% performance under a much stricter 'source-image split' constraint. This evaluation protocol explicitly prevents the object-level data leakage that is often overlooked in previous random-split studies, proving the robustness of our architecture in a more realistic scenario. The biggest challenge in medical imaging is the scarcity of labeled data. The current trend in the literature is to generate synthetic data using Generative Adversarial Networks (GANs) to overcome this problem, as demonstrated in the studies by Xie et al. [27] and Frid-Adar et al. [40]. However, Shorten et al. [41] have criticized synthetic data for potentially distorting the clinical noise distribution and causing the model to hallucinate.

Our study is significant in that it demonstrates that optimal performance can be achieved without resorting to synthetic data generation, through the use of appropriate Data Balancing and Transfer Learning strategies. This approach eliminates the high computational cost associated with training GANs and supports the "fine-tuning strategy on limited data" proposed by Tajbakhsh et al. [42].

A unique finding of this study that contributes theoretically to the literature is the positive effect of small batch sizes (Batch Size = 8) on model generalization. In deep learning theory, Masters and Luschi [43] and Keskar et al. [44] have theoretically proven that the use of large batch sizes leads to convergence to "sharp minima," which in turn reduces generalization ability.

Our experimental results validate this theoretical framework on the HEp-2 dataset. When we used a batch size of 8, the stochastic noise in the gradient estimation helped the model avoid sharp


CC: BY-NC-ND

minima and settle into flatter, and thus more stable, minima. This finding serves as a practical guide supporting the principle of the "regularizing effect of Stochastic Gradient Descent" proposed by LeCun et al. [45], particularly for researchers working with tissue-focused and noisy biomedical datasets. From a technical perspective, the fact that the proposed ensemble architecture completely eliminates False Positive errors (achieving 100% precision) in the challenging Fine Speckled class demonstrates its robust feature extraction capabilities for autoimmune pattern recognition. However, a key limitation of this study is that the model was evaluated using only a single dataset (MIVIA – ICPR 2012); therefore, a more cautious approach is needed before stating that the model, in its current form, can be directly integrated into clinical laboratory routines. The goal of future studies is to validate the model on external datasets from different laboratories and to visualize the model's decision-making mechanism using explainable artificial intelligence (XAI) methods, such as Grad-CAM, as proposed by Selvaraju et al. [46].

## 6. CONCLUSION

In this study, an ensemble architecture utilizing local and global feature extractors is proposed for the HEp-2 cell classification problem, which plays a critical role in the diagnosis of autoimmune diseases. Based on extensive experiments and analyses conducted on the MIVIA (ICPR 2012) dataset, several key findings have been obtained: Combining the DenseNet-121 model, specialized in structural features, with the Vision Transformer (ViT-B16) model, which analyzes textural patterns, using a weighted ensemble strategy, demonstrated superior performance compared to individual models. The proposed method achieved a 96.01% accuracy rate under a strict source-image split evaluation protocol, achieving highly competitive performance compared to prior methods without suffering from object-level data leakage, and without requiring synthetic data generation (GAN) or complex preprocessing steps. The study's most significant methodological contribution is the 100% precision achieved in the Fine Speckled class, a pattern that is frequently misclassified. This result demonstrates that the proposed architecture possesses robust feature extraction capabilities, highlighting its potential to inform future Computer-Aided Diagnosis (CAD) systems that assist pathologists by minimizing false-positive alarms. Contrary to the prevailing trend in the literature, our experimental findings have demonstrated that small batch sizes (Batch Size = 8) improve model performance in tissue-specific and limited datasets. The stochastic noise generated by using a small batch size has served as an effective regularization tool, enabling the model to avoid sharp minima and achieve better generalization. Despite the experimental success of this study, there are some future directions that need to be addressed for full alignment with clinical practice. The model's success is currently limited to the MIVIA dataset. In future studies, we plan to test the model's robustness on external datasets obtained under different laboratory conditions and with different microscopy hardware. To make the black-box nature of deep learning models more transparent, Grad-CAM, proposed by Selvaraju et al. [46], and Attention Rollout techniques customized for Transformers will be integrated into the model. This will visually present to pathologists which region of the cell (nucleus, cytoplasm, or spots) the model focused on when making its decision. Due to data privacy concerns, it is aimed to adapt the Federated Learning approaches [47], which enable training without collecting data from different hospitals on a central server, to the HEp-2 classification. The long-term aim is for such hybrid architectures to become reliable and scalable clinical tools that support pathologists in their decision-making processes for the early and accurate diagnosis of autoimmune diseases.

**Data and Code Availability:** The MIVIA HEp-2 dataset analyzed during the current study is publicly available upon registration from the MIVIA Lab repository. To support methodological transparency, the custom training scripts and preprocessing codes utilized in this study are available from the corresponding author upon reasonable request.

**Peer-review:** External, Independent.

**Acknowledgements:**
Since this study utilizes a publicly available and fully anonymized benchmark dataset (MIVIA ICPR 2012),


CC: BY-NC-ND

which contains no identifiable human data or personal patient records, institutional ethical approval was not required.

**Declarations:**

***1. Statement of Originality:***

This work is original.

***2. Author Contributions:***

***Concept:*** FÖ,EÖ; ***Conceptualization:*** FÖ,EÖ,TT; ***Literature Search:*** FÖ; ***Data Collection:*** FÖ; ***Data Processing:*** FÖ; ***Analysis:*** FÖ; ***Writing – original draft:*** FÖ; ***Writing – review & editing:*** EÖ,TT.

***3. Ethics approval:***

Not applicable.

***4. Funding/Support:***

This work has not received any funding or support.

***5. Competing Interests:***

The authors declare no competing interests.

***6. GenAI Usage Statement:***

No GenAI tools were used at any stage of the study.

***7. Sustainable Development Goals:***

## REFERENCES

[1] Wang, L., Wang, F. S., & Gershwin, M. E. (2015). Human autoimmune diseases: a comprehensive update. Journal of Internal Medicine, 278(4), 369–395. https://doi.org/10.1111/joim.12395

[2] Mohamed-Ahmed, O., Shang, L., Wang, L., Chen, Z., Kartsonaki, C., & Bragg, F. (2024). Incidence and prevalence of autoimmune diseases in China: A systematic review and meta-analysis of epidemiological studies. Global Epidemiology, 8, 100158. https://doi.org/10.1016/j.gloepi.2024.100158

[3] Lee, Y. H., & Song, G. G. (2024). Mortality in patients with systemic lupus erythematosus: A meta-analysis of overall and cause-specific effects. Lupus, 33(9), 929–937. https://doi.org/10.1177/09612033241257134

[4] Chan, E. K. L., Damoiseaux, J., Carballo, O. G., Conrad, K., Cruvinel, W. M., Francescantonio, P. L. C., Fritzler, M. J., Garcia-De La Torre, I., Herold, M., Mimori, T., Satoh, M., von Mühlen, C. A., & Andrade, L. E. C. (2015). Report of the First International Consensus on Standardized Nomenclature of Antinuclear Antibody HEp-2 Cell Patterns 2014–2015. Frontiers in Immunology, 6, 412. https://doi.org/10.3389/fimmu.2015.00412

[5] Damoiseaux, J., Andrade, L. E. C., Carballo, O. G., Conrad, K., Francescantonio, P. L. C., Fritzler, M. J., Garcia de la Torre, I., Herold, M., Klotz, W., Cruvinel, W. M., Mimori, T., von Muhlen, C., Satoh, M., & Chan, E. K. (2019). Clinical relevance of HEp-2 indirect immunofluorescent patterns: the

International Consensus on ANA patterns (ICAP) perspective. Annals of the Rheumatic Diseases, 78(7), 879–889. https://doi.org/10.1136/annrheumdis-2018-214436

[6] Agmon-Levin, N., Damoiseaux, J., Kallenberg, C., Sack, U., Witte, T., Herold, M., Bossuyt, X., Musset, L., Cervera, R., Plaza-Lopez, A., Dias, C., Sousa, M. J., Radice, A., Eriksson, C., Hultgren, O., Viander, M., Khamashta, M., Regenass, S., Andrade, L. E. C., Wiik, A., Tincani, A., Rönnelid, J., Bloch, D. B., Fritzler, M. J., Chan, E. K. L., Garcia-De La Torre, I., Konstantinov, K. N., Lahita, R., Wilson, M., Vainio, O., Fabien, N., Sinico, R. A., Meroni, P., & Shoenfeld, Y. (2014). International recommendations for the assessment of autoantibodies to cellular antigens referred to as anti-nuclear antibodies. Annals of the Rheumatic Diseases, 73(1), 17–23. https://doi.org/10.1136/annrheumdis-2013-203863

[7] Meroni, P. L., & Schur, P. H. (2010). ANA screening: an old test with new recommendations. Annals of the Rheumatic Diseases, 69(8), 1420–1422. https://doi.org/10.1136/ard.2009.127100

[8] Van Hoovels, L., Broeders, S., Chan, E. K. L., Andrade, L. E. C., de Melo Cruvinel, W., Damoiseaux, J., Viander, M., Herold, M., Coucke, W., Heijnen, I., Bogdanos, D., Calvo-Alén, J., Eriksson, C., Kozmar, A., Kuhi, L., Bonroy, C., Lauwerys, B., Schouwers, S., Lutteri, L., Vercammen, M., & Bossuyt, X. (2020). Current laboratory and clinical practices in reporting and interpreting anti-nuclear antibody indirect immunofluorescence (ANA IIF) patterns: results of an international survey. Autoimmunity Highlights, 11, 17. https://doi.org/10.1186/s13317-020-00139-9

[9] Foggia, P., Percannella, G., Saggese, A., & Vento, M. (2014). Pattern recognition in stained HEp-2 cells: Where are we now? Pattern Recognition, 47(7), 2305–2314. https://doi.org/10.1016/j.patcog.2014.01.010

[10] Anaam, A., Al-antari, M.A., Hussain, J., Abdel Samee, N., Alabdulhafith, M., & Gofuku, A. (2023). Deep Active Learning for Automatic Mitotic Cell Detection on HEp-2 Specimen Medical Images. Diagnostics, *13*(8), 1416. https://doi.org/10.3390/diagnostics13081416

[11] Lei, H., Han, T., Zhou, F., Yu, Z., Qin, J., Elazab, A., & Lei, B. (2018). A deeply supervised residual network for HEp-2 cell classification via cross-modal transfer learning. Pattern Recognition, 79, 290–302. https://doi.org/10.1016/j.patcog.2018.02.006

[12] Iwata, A., & Okuda, M. (2024). Quantifying Shape and Texture Biases for Enhancing Transfer Learning in Convolutional Neural Networks. Signals, 5(4), 721–735. https://doi.org/10.3390/signals5040040

[13] Fang, K., Li, C., & Wang, J. (2023). An automatic immunofluorescence pattern classification framework for HEp-2 image based on supervised learning. Briefings in Bioinformatics, *24*(3), bbad144. https://doi.org/10.1093/bib/bbad144

[14] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv, abs/2010.11929. https://doi.org/10.48550/arXiv.2010.11929

[15] Liu, Y., Wen, Y., Yang, L., He, L., & Zhou, M. (2025). A General Framework for Efficient Medical Image Analysis via Shared Attention Vision Transformer. IEEE Transactions on Medical Imaging, PP(99), 1–1. https://doi.org/10.1109/TMI.2025.3644949

[16] Foggia, P., Percannella, G., Soda, P., & Vento, M. (2013). Benchmarking HEp-2 Cells Classification Methods. IEEE Transactions on Medical Imaging, 32(10), 1878–1889. https://doi.org/10.1109/TMI.2013.2268163

[17] Manivannan, S., Li, W., Akbar, S., Wang, R., Zhang, J., & McKenna, S. J. (2016). An automated pattern recognition system for classifying indirect immunofluorescence images of HEp-2 cells


CC: BY-NC-ND

and specimens. Pattern Recognition, 51, 12–26. https://doi.org/10.1016/j.patcog.2015.09.015

[18] Gao, Z., Wang, L., Zhou, L., & Zhang, J. (2017). HEp-2 Cell Image Classification With Deep Convolutional Neural Networks. IEEE Journal of Biomedical and Health Informatics, 21(2), 416–428. https://doi.org/10.1109/JBHI.2016.2526603

[19] Bayramoglu, N., & Heikkilä, J. (2016). Transfer Learning for Cell Nuclei Classification in Histopathology Images. In Computer Vision – ECCV 2016 Workshops. Lecture Notes in Computer Science. https://doi.org/10.1007/978-3-319-49409-8_46

[20] Rodrigues, L. F., Naldi, M. C., & Mari, J. F. (2020). Comparing convolutional neural networks and preprocessing techniques for HEp-2 cell classification in immunofluorescence images. Computers in Biology and Medicine, 116, 103542. https://doi.org/10.1016/j.compbiomed.2019.103542

[21] Lei, H., Han, T., Huang, W., Kuo, J. Y., Yu, Z., He, X., & Lei, B. (2017). Cross-Modal Transfer Learning for HEp-2 Cell Classification Based on Deep Residual Network. In 2017 IEEE International Symposium on Multimedia (ISM) (pp. 465–468). IEEE. https://doi.org/10.1109/ISM.2017.92

[22] Rahman, S., Wang, L., Sun, C., & Zhou, L. (2020). Deep learning based HEp-2 image classification: A comprehensive review. Medical Image Analysis, 65, 101764. https://doi.org/10.1016/j.media.2020.101764

[23] Jensen, M., Al-Dulaimi, K., Abduljabbar, K.S., & Banks, J. (2023). Automated Classification of Cell Level of HEp-2 Microscopic Images Using Deep Convolutional Neural Networks-Based Diameter Distance Features. JUCS - Journal of Universal Computer Science, *29*(5), 432-445. https://doi.org/10.3897/jucs.96293

[24] Chen, C. F. R., Fan, Q., & Panda, R. (2021). CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 347–356). IEEE. https://doi.org/10.1109/ICCV48922.2021.00041

[25] Takahashi, S., Sakaguchi, Y., Kouno, N., Takemura, R., Okamoto, K., & Suzuki, H. (2024). Comparison of Vision Transformers and Convolutional Neural Networks in Medical Image Analysis: A Systematic Review. Journal of Medical Systems, 48, 84. https://doi.org/10.1007/s10916-024-02105-8

[26] Wu, P., Liu, Z., Zhao, Z. et al. (2025). LCPT: A lightweight cell classification method for microscope images based on vicinityViT and channel-position attention. SIViP, *19*, 1240. https://doi.org/10.1007/s11760-025-04829-7

[27] Xie, H., He, Y., Lei, H., Kuo, J., & Lei, B. (2019). Segmentation Guided HEp-2 Cell Classification with Adversarial Networks. In ComComAP (pp. 374–379). IEEE. https://doi.org/10.1109/ComComAp46287.2019.9018850

[28] Chen, B., Chen, J., Lu, H., & Chen, J. (2026). Cell classification based on affine transformed convolutional parallel networks. Biomedical Signal Processing and Control, *112*(Part A), 108163. https://doi.org/10.1016/j.bspc.2025.108163

[29] Hosseinzadeh Kasani, P., Hosseinzadeh Kasani, S., Kim, H. W., Cho, K. H., Jang, J. W., & Yun, C. H. (2021). HEp-2 Cell Classification Using an Ensemble of Convolutional Neural Networks. In 2021 International Conference on Information and Communication Technology Convergence (ICTC) (pp. 1–6). IEEE. https://doi.org/10.1109/ICTC52510.2021.9621075

[30] Vununu, C., Lee, S. H., & Kwon, K. R. (2019). A Deep Feature Extraction Method for HEp-2 Cell Image Classification. Electronics, 8(1), 20. https://doi.org/10.3390/electronics8010020


CC: BY-NC-ND

[31] Anam, A. M., Al-antari, M. A., & Gofuku, A. (2022). A Hybrid Ensemble Learning with Generative Adversarial Networks for HEp-2 Cell Image Classification. In *2022 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)* (pp. 1–5). IEEE. https://doi.org/10.1109/IECBES54088.2022.10079623

[32] Zhang, L., Zhang, M. Q., & Lv, X. (2022). HEp-2 image classification using a multi-class and multiple-binary classifier. Medical & Biological Engineering & Computing, 60, 3113–3124. https://doi.org/10.1007/s11517-022-02646-5

[33] Nguyen, L.D., Gao, R., Lin, D. et al. (2023). Biomedical image classification based on a feature concatenation and ensemble of deep CNNs. Journal of Ambient Intelligence and Humanized Computing, *14*, 15455–15467. https://doi.org/10.1007/s12652-019-01276-4

[34] Lei, B., Zhang, T., Li, J., Qu, J., Xiong, W., Yu, K., & Yang, L. (2025). Rapidly Antibiotic Susceptibility Prediction via Deep Learning From Bacterial Fluorescence Microscopy Images. IEEE Transactions on Automation Science and Engineering, *22*, 21469-21483. https://doi.org/10.1109/TASE.2025.3607484

[35] Tan, M., & Le, Q. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In K. Chaudhuri & R. Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning (ICML) (Vol. 97, pp. 6105–6114). PMLR. https://proceedings.mlr.press/v97/tan19a.html

[36] Khaksar Ghalati, M., Nunes, A., Ferreira, H., Serranho, P., & Bernardes, R. (2022). Texture Analysis and Its Applications in Biomedical Imaging: A Survey. IEEE Reviews in Biomedical Engineering, 15, 222–246. https://doi.org/10.1109/RBME.2021.3115703

[37] Xu, X., & Li, B. (2025). Multi scale supervised entropy weighted binary pattern for texture classification. Scientific Reports, 15, 26087. https://doi.org/10.1038/s41598-025-11245-x

[38] Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in Vision: A Survey. ACM Computing Surveys, 54(10s), 1–41. https://doi.org/10.1145/3505244

[39] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A. L., & Zhou, Y. (2021). TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. arXiv, abs/2102.04306. https://doi.org/10.48550/arXiv.2102.04306

[40] Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., & Greenspan, H. (2018). GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. Neurocomputing, 321, 321–331. https://doi.org/10.1016/j.neucom.2018.09.013

[41] Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on Image Data Augmentation for Deep Learning. Journal of Big Data, 6, 60. https://doi.org/10.1186/s40537-019-0197-0

[42] Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., & Liang, J. (2016). Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? IEEE Transactions on Medical Imaging, 35(5), 1299–1312. https://doi.org/10.1109/TMI.2016.2535302

[43] Masters, D., & Luschi, C. (2018). Revisiting Small Batch Training for Deep Neural Networks. arXiv, abs/1804.07612. https://arxiv.org/abs/1804.07612

[44] Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., & Tang, P. T. P. (2017). On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. arXiv, abs/1609.04836. https://arxiv.org/abs/1609.04836

[45] Montavon, G., Orr, G. B., & Müller, K. R. (Eds.). (2012). Neural Networks: Tricks of the Trade (2nd ed.). Lecture Notes in Computer Science (Vol. 7700). Springer. https://doi.org/10.1007/978-3-

642-35289-8

[46] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In 2017 IEEE International Conference on Computer Vision (ICCV) (pp. 618–626). IEEE. https://doi.org/10.1109/ICCV.2017.74

[47] Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., Ourselin, S., Sheller, M., Summers, R. M., Trask, A., Xu, D., Baust, M., & Cardoso, M. J. (2020). The future of digital health with federated learning. npj Digital Medicine, 3, 119. https://doi.org/10.1038/s41746-020-00323-1.

CC: BY-NC-ND