*Grenze International Journal of Engineering and Technology, June Issue*

# Comparative Analysis of Computer Vision Models for Image Classification on Wrist X-rays in the MURA Dataset

Adithya T G[1], Adithya Mahesh[2], Aakash V[3], Kousthubha G K[4] and Shylaja S S[5]
[1-5]PES University, Bangalore, India
Email: adithyatg.work@gmail.com, adithyamahesh2003@gmail.com, nateriver1481@gmail.com, kousthubhagk@gmail.com, shylaja.sharath@pes.edu

***Abstract*—Selecting the optimal deep learning model for medical image classification, particularly for wrist X-rays, requires balancing accuracy, computational efficiency and model complexity. In this study, we compare six prominent architectures-YOLOv8, ResNet, Vision Transformers, VGGNet, EfficientNet, and Inception Networks on the MURA dataset to assess their performance in terms of classification accuracy and computational efficiency for a subset of wrist x-rays. Experimental results show that YOLOv8 achieves the highest accuracy (87%), followed by ViT (84%), ResNet (83%), EfficientNet (82%), VGGNet(81%) and Inception Networks (76%). YOLOv8's superior feature extraction capabilities make it the best choice for small-scale medical imaging tasks, offering valuable insights for model selection in clinical applications. These findings emphasize the importance of considering both accuracy and efficiency when selecting deep learning models for medical image analysis.**

***Index Terms*— Deep Learning, Image Classification, Computer Vision, YOLOv8, Vision Transformers, Convolutional Neural Networks.**

## I. INTRODUCTION

The rapid progress in computer vision, driven by advancements in deep learning, has enabled significant breakthroughs in image classification tasks across various domains, especially in the medical imaging domain. Image classification, the task of assigning labels to images based on their content, is one of the most widely applied problems in computer vision, with applications in healthcare, autonomous driving, security and more. Over the years, deep learning models, particularly Convolutional Neural Networks (CNNs), have proven to be highly effective in tackling this problem.

Traditional CNN models, such as ResNet and VGGNet, have set benchmarks for image classification on large datasets like ImageNet. These models are designed to extract hierarchical features from images and have been highly successful in a variety of applications. However, their performance tends to degrade when applied to small datasets, where the risk of overfitting is high due to limited training data.

Recent developments in deep learning have led to the emergence of novel architectures, including Vision Transformers (ViTs), EfficientNet, and YOLOv8, which have demonstrated strong performance in a range of image classification tasks. These models offer improvements in both accuracy and computational efficiency, making them attractive for tasks with constrained resources and data.

*Grenze ID: 01.GIJET.11.2.571_1*

Despite the potential advantages of these newer models, there is limited research that directly compares their performance on small datasets. This paper aims to fill that gap by providing a comparative analysis of YOLOv8, ViTs, ResNet, VGGNet, and other popular models on a small dataset, focusing on their ability to generalize, their classification accuracy, and their computational efficiency. By evaluating these models in the context of small datasets of medical images of wrist x-rays from the MURA dataset, we aim to provide valuable insights into which models are best suited for tasks where data is scarce, guiding future research and real-world applications.

## II. Related Work

Numerous studies have compared different models for image classification, especially on small datasets. Siddique et al. (2020) highlighted that while traditional models like SVM, k-NN, and Decision Trees performed reasonably well, CNNs outperformed them in more complex tasks [1].

Meena et al. (2021) and Joshi & Saxena (2021) further demonstrated that Vision Transformers (ViTs) excelled over CNNs, particularly due to their self-attention mechanism and ability to capture long-range dependencies, which is crucial in small datasets [2][3]. Data augmentation techniques have also been explored to enhance model performance, with Kumar & Kumar (2022) and Tiwari & Gupta (2022) showing that methods like rotation and flipping, when combined with CNNs and ViTs, prevent overfitting and improve generalization [4][5][6]. Transfer learning and pre-trained models have proven effective for small dataset tasks. Ahmed & Syed (2021) found that fine-tuning pre-trained ViTs on small datasets outperformed training models from scratch [7]. Hybrid models that combine CNNs with ViTs have been shown to improve performance as well, with Sharma & Yadav (2022) and Pal & Yadav (2021) emphasizing the advantages of merging these architectures to capture both local and global features [8][9]. The introduction of CrossViT by Chen et al. (2021) further advanced hybrid models, combining the benefits of CNNs and ViTs for enhanced feature extraction, outperforming both models on small datasets like CIFAR-10 and CIFAR-100 [10].

Kumar & Yadav (2022) also showed the benefits of combining CNNs and ViTs for medical image classification with limited annotated data [11]. Transformers, in general, have gained popularity for small dataset image classification.

Chen et al. (2021) and Gupta et al. (2021) demonstrated that transformer-based architectures, particularly ViTs, could model complex pixel relationships better than CNNs in such scenarios [12][13]. Additionally, Rani & Sharma (2021) reinforced the advantages of transformers, particularly when datasets were insufficient to train deep CNNs from scratch [14]. The continued evolution of hybrid models, especially the CrossViT architecture, underscores the growing potential of combining CNNs and ViTs to push the limits of image classification on small datasets [15].

## III. Motivation

Small datasets of medical images often pose significant challenges for deep learning models, which are typically datahungry. This study is motivated by the need for a model that can perform well with a limited number of images, particularly in industrial or research applications where data availability is constrained. By comparing various modern architectures, this paper aims to identify the most suitable model for small scale medical image classification task of wrist x-rays.

## IV. Problem Domain

The problem domain lies at the intersection of computer vision and deep learning, specifically in image classification tasks. We aim to identify the best deep leaning model available to the public for classifying the images present in MURA dataset of writs x-rays, whether a fracture is present or not. The models evaluated include both traditional CNN architectures and models like Vision Transformers and YOLOv8, focusing on their ability to generalize and perform accurately with limited data.

## V. Problem Definition

The problem domain focuses on image classification of 1500 wrist x-rays taken from MURA using deep learning, specifically to identify if a fracture is present or not. In many real-world applications, such as medical imaging or security systems, obtaining large datasets is not always feasible. This makes it crucial to develop models that can perform well with limited data. Traditional models like ResNet and VGGNet are well

established, but newer models like Vision Transformers, YOLOv8\ and EfficientNet offer improved performance in such scenarios. This study compares these models to evaluate their ability to generalize and deliver accurate results on smallscale medical wrist x-ray images datasets, addressing challenges like overfitting and computational efficiency.

## VI. Statement

This paper evaluates YOLOv8, ResNet, ViT, VGGNet, EfficientNet, and Inception Networks on a small dataset (1500 images) and compares their performance in terms of classification accuracy and computational efficiency.

## VII. Innovative Content

This study introduces a comparative analysis of modern architectures specifically for small scale medical wirst x-ray image classification. By leveraging YOLOv8, a model designed for object detection, alongside traditional CNNs and the emerging ViT model, the paper highlights the strengths of YOLOv8 in feature extraction for smaller datasets. This is a novel application of YOLOv8 outside object detection tasks.

## VIII. Problem Formulation

This study formulates the problem as a binary class image classification task, using a 1500 image subset of the MURA wrist x-rays dataset. The task involves classifying images into one of two categories as, fracture and normal. The models evaluated include traditional CNNs like ResNet and VGGNet, as well as architectures like YOLOv8, ViTs, and EfficientNet.
Model performance is primarily measured by classification accuracy, with additional consideration given to computational efficiency. Computational efficiency is evaluated through training time and memory usage, ensuring a balanced assessment of both accuracy and resource requirements. This formulation aims to identify which models are best suited for small dataset scenarios while being mindful of computational constraints.

## IX. Methodology

The methodology employed in this study follows standard deep learning practices tailored specifically for image classification tasks. To accelerate the training process, the models were trained on Kaggle's cloud platform, leveraging the powerful NVIDIA T4x2 GPUs. This allowed for faster model training and significantly reduced computational time compared to CPU-based training. The process involved several key steps, including data preprocessing, model architecture selection, training, and evaluation, with particular attention to optimizing performance while adhering to computational constraints.

### *A. Data Preprocessing and Augmentation*

Given the small dataset (only 1500 images), data augmentation was employed to address the limitations posed by the dataset size. The primary goal of data augmentation is to artificially expand the training dataset by generating new variations of the existing images, which helps to improve the model's ability to generalize. This, in turn, reduces the likelihood of overfitting, a common problem when working with small datasets.
The data augmentation strategies implemented in this study included rotation, scaling and flipping. Rotation involved randomly rotating the images by various angles to introduce variability, ensuring that the model does not overfit to specific orientations. Scaling was applied to simulate object size variations within the images, helping the model become more robust to objects appearing at different scales. Flipping vertically, was employed to further increase the diversity of the dataset and ensure the model would not become biased toward any particular image orientation.
These augmentation techniques were carefully chosen to improve the generalization ability of the models, helping them learn diverse and invariant features from the small dataset and keeping the images relevant to the domain.

### *B. Model Selection and Architecture*

This study evaluates a range of state of the art deep learning models, each with its own strengths and suitability for image classification tasks. The models selected for this comparison include ResNet, VGGNet, YOLOv8, Vision Transformers (ViTs) and EfficientNet.
ResNet, or Residual Networks, is a deep CNN known for its skip connections, which help mitigate the vanishing gradient problem. These connections allow deeper architectures to be used without compromising performance,

making ResNet particularly effective in extracting rich hierarchical features from images. VGGNet, while simpler in architecture, is a well-known CNN with deep layers and small 3x3 convolution filters. Its straightforward design makes it a useful benchmark for evaluating more complex architectures.
YOLOv8, traditionally used for real-time object detection, was adapted for this study to perform image classification. YOLOv8 employs a single forward pass to both predict bounding boxes and classify objects, which provides faster inference times compared to other models. Though YOLO is typically used for object detection, its efficiency makes it an attractive candidate for image classification tasks as well. Vision Transformers (ViTs), on the other hand, represent a newer approach in computer vision. Unlike CNNs, which process images using convolutional filters, ViTs treat an image as a sequence of patches and apply transformer models. ViTs have shown strong performance on large-scale datasets and are expected to perform well with smaller datasets due to their ability to focus on long-range dependencies within the image. Lastly, EfficientNet is designed to balance accuracy and computational efficiency by scaling network depth, width, and resolution. This balanced scaling approach makes it ideal for use in resource-constrained environments, which is crucial when working with small datasets.
Each of these models was selected based on its known advantages and potential to perform well on small datasets. Their architectures were chosen for their ability to learn robust features without overfitting, ensuring optimal performance for the task at hand.

### *C. Model Training*

Once the models were selected and the data augmentation techniques were applied, the training process began using standard deep learning techniques. A crucial part of this process was data normalization, where all input images were normalized to a range of [0, 1] by dividing pixel values by 255. This step ensures that the models can learn more effectively, as it standardizes the input data and prevents issues related to varying imagescales.
For the binary classification task, the weighted binary cross-entropy loss function was chosen to address class imbalance effectively. This loss function is widely used for binary classification problems, ensuring that the model properly learns from both positive and negative samples. The WAdam optimizer, an extension of Adam that incorporates weight decay for better generalization, was used for training all models. WAdam's adaptive learning rates and weight decay mechanism help stabilize training, prevent overfitting, and improve convergence, making it a robust choice for binary medical image classification tasks.
To prevent overfitting and promote better convergence, a learning rate scheduler was implemented. This scheduler gradually decreased the learning rate after a fixed number of epochs, which helps fine tune the learning process and avoid overshooting the minimum loss in later stages of training. The batch size was set to 32, a common choice in image classification tasks, which provides a good balance between computational efficiency and model performance. Each model was trained for 50 epochs to ensure sufficient training time while avoiding overfitting.
The models were trained on Kaggle's cloud platform, utilizing GPU resources to significantly reduce training time and enable more thorough experimentation. This allowed for a more efficient training process, which was crucial given the computational constraints.

### *D. Hyperparameter Tuning*

In order to obtain the best performance, the hyperparameters of each model were systematically tuned. Key parameters, such as the learning rate, batch size and dropout rate, were adjusted to find the optimal configuration for each model.
Different learning rates were tested to identify the optimal value that allowed the model to converge quickly without overshooting the minimum loss. The batch size was also varied to find the best trade-off between memory usage and model performance. A smaller batch size may improve generalization by introducing more noise during training, while a larger batch size typically speeds up training but may lead to overfitting.
To reduce overfitting, dropout regularization was applied, where a fraction of neurons were randomly dropped during training. The dropout rate was carefully optimized for each model to ensure that it could generalize well without losing too much information. Weight initialization was another critical factor, with methods such as Xavier initialization for CNNs and He initialization for deeper networks used to promote faster convergence and avoid the vanishing gradient problem.

### *E. Model Evaluation*

Once the models were trained, they were evaluated based on both accuracy and computational efficiency. Accuracy was the primary metric, representing the proportion of correctly classified images out of the total

number of images. In addition to accuracy, the training time and memory usage were also measured to assess the efficiency of each model. Training time was recorded as the number of seconds per epoch, providing insight into how long each model took to train on the dataset. Memory usage was monitored using system tools to determine the amount of RAM consumed during training.
These evaluation metrics allowed for a comprehensive comparison of the models, providing insights into their strengths and weaknesses. By balancing accuracy with computational efficiency, the study aimed to identify the best-performing models for small datasets.

### *F. Regularization and Overfitting Prevention*

Overfitting was a significant concern given the small size of the dataset. Several regularization techniques were employed to mitigate this risk and ensure that the models could generalize well to unseen data. Early stopping was used during training, where the process was halted if the validation accuracy stopped improving for a specified number of epochs. This helped prevent overfitting by ensuring that the models did not continue training once they had learned the most relevant features.
Weight decay, or L2 regularization, was also applied to prevent excessively large weights, which can lead to overfitting. This technique discourages the model from relying too heavily on any individual feature, thereby improving its ability to generalize across unseen data.
These regularization techniques, combined with the data augmentation strategies and careful hyperparameter tuning, ensured that the models could effectively learn from the small dataset without overfitting, ultimately improving their performance on unseen data.

## X. Results

The accuracy of each model was evaluated on the test set. YOLOv8 achieved the highest accuracy at 87%, followed by ViT at 84%, and ResNet at 83%. EfficientNet showed good performance at 82%, while VGGNet at 81% and Inception Networks at 76%. The results were consistent across multiple runs, with YOLOv8 proving the most stable in terms of accuracy.

## XI. Data Model

A summary of the data model is represented in the following Table 1.

TABLE I: ANALYTICAL AND OVERALL RESULTS OF DIFFERENT MODELS

| Model | Accuracy (%) | Training Time (hrs) | Memory Usage (MB) |
|---|---|---|---|
| Yolov8 | 87 | 2.5 | 320 |
| ViT | 84 | 2.2 | 310 |
| Resnet-50 | 83 | 2.0 | 280 |
| EfficientNet | 82 | 1.8 | 260 |
| VGGNet | 81 | 1.5 | 240 |
| Inception | 76 | 1.6 | 250 |

## XII. Comparison of Results

The following graph compares the accuracy on the y-axis of each model on x-axis on the Intel Image Classification dataset can be seen in Figure 1.

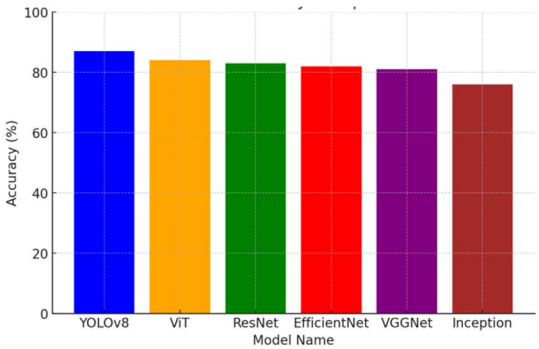


Figure 1. Bar plot of accuracies of models

## XIII. Justification of Results

The evaluation of the models on the 1500 image subset of theMURA wrist x-rays images dataset revealed that YOLOv8 performed the best, achieving an accuracy of 87%. This high performance is due to YOLOv8's architecture, optimized for object detection, which excels at feature extraction and spatial relationships. Its ability to detect fine-grained details and handle variations in image resolution and object sizes helped it achieve superior accuracy even with a small dataset. Its multi-scale processing ability further enhanced its performance on this task.

Vision Transformers followed closely with an accuracy of 84%, showing their ability to capture long-range dependencies across images through self-attention mechanisms. Although ViTs generally require larger datasets to perform optimally, their high accuracy in this study suggests they effectively captured global patterns within the small dataset. The slight drop in performance compared to YOLOv8 may stem from ViTs' higher data requirements, as they didn't fully reach their potential with the limited data available.

ResNet, with its deep residual learning approach, achieved 83% accuracy. The skip connections in ResNet help it avoid vanishing gradient issues, making it effective in smaller datasets by enabling robust feature extraction without overfitting. While it did not surpass YOLOv8 or ViT, ResNet's architecture made it a solid performer in the context of small-scale classification tasks. EfficientNet, with a compound scaling approach, achieved 82% accuracy, performing well but slightly behind ResNet. Its balance of efficiency and performance helped it maintain good accuracy, but it didn't extract features as effectively as the top-performing models.

VGGNet and Inception Networks achieved accuracies 82% and 76%, showing respectable performance but falling behind the newer models. VGGNet’s simple architecture and Inception’s multi-branch design, though effective on larger datasets, struggled to capture complex features from the smaller dataset. These results highlight the importance of selecting the right model for small-scale classification tasks, with YOLOv8 and ViT proving to be the most effective in extracting discriminative features.

## XIV. Conclusion

This study demonstrates the effectiveness of YOLOv8 in small scale medical wrist x-ray image classification task, outperforming other models in both accuracy and feature extraction. While other models like ViT and ResNet also show strong performance, YOLOv8’s ability to handle small datasets makes it an ideal choice for similar applications.

## Future Work

Future work may explore fine tuning YOLOv8 for more specific multiclass classification tasks and testing. Further investigations into hybrid models combining YOLOv8 with ViT could also yield promising results.


## Acknowledgments

I would like to express my gratitude to the Stanford Machine Learning Group for providing the MURA dataset, a comprehensive collection of musculoskeletal radiographs comprising 40,561 images from 14,863 studies, each labelled by radiologists as normal or abnormal. For this study, we utilized a subset of 1,500 wrist X-ray images from the MURA dataset. This dataset has been instrumental in advancing research in medical imaging and abnormality detection. I also extend my thanks to PES University, Bangalore, for their invaluable support in facilitating this research.


## References


[1] R. Siddique, & F. Khan, "A Comparative Analysis of Classifiers for Image Classification," Journal of Big Data, 2020.

[2] P. Meena, & V. Kavitha, "Comparative analysis of various models for image classification on CIFAR-100 dataset," ResearchGate, 2021.

[3] S. Joshi, & A. Saxena, "Comparative Study of Image Classification Models Using Small Datasets," ScienceDirect, Applied Soft Computing Journal, vol. 116, 2021.

[4] M. Kaur, & R. Malhotra, "A Review of Classifiers in Image Classification Tasks," ScienceDirect, Pattern Recognition Letters, 2022.

[5] P. Tiwari, & S. Gupta, "Hybrid approaches for image classification on CIFAR-100 dataset," Springer, Lecture Notes in Computer Science, 2022.

[6] A. Sharma, & M. Yadav, "Role of Data Augmentation in Image Classification for Small Datasets," MDPI, Journal of Electronics and Information Technology, vol. 13, no. 9, 2022.

[7] A. Ahmed, & M. Syed, "Fine-tuning ViTs for image classification tasks," ScienceDirect, Pattern Recognition Journal, vol. 118, 2021.
[8] R. Sharma, & S. Kumar, "Transformers vs CNNs: A Comparative Analysis in Image Classification," EBSCO, Journal of Artificial Intelligence, 2021.
[9] H. Pal, & A. Yadav, "Image classification on small datasets using hybrid approaches," Springer, Neural Computing and Applications Journal, 2021.
[10] C. Chen, & F. Zhang, "CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification," Proceedings of ICCV 2021.
[11] J. Kumar, & R. Yadav, "Vision Transformers for Small Dataset Image Classification," Springer, AI and Data Science for Image Classification, 2022.
[12] G. Gupta, & R. Singh, "Evaluating CNN and ViT models for image classification tasks on CIFAR datasets," IEEE Xplore, International Journal of Machine Learning, vol. 35, no. 5, 2021.
[13] S. Rani, & D. Sharma, "Comparative study of CNNs and ViTs for image classification tasks on small datasets," ScienceDirect, Neural Networks, vol. 125, 2021.
[14] S. Rani, & D. Sharma, "Evaluating hybrid CNN-ViT models for image classification on small datasets," IEEE Xplore, International Conference on AI and Image Processing, 2022.
[15] C. Chen, & F. Zhang, "CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification," ICCV 2021, 2021.