International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Deep Learning in Image Processing: Transforming Computer Vision

Sumithra M D[1], M. Abdul Rahiman[2]

Submitted: 08/12/2020 **Revised:** 10/01/2021 **Accepted:** 20/01/2021

Abstract: Deep learning has been acclaimed to be the new sensation in image processing, improving the functionality of the computer vision system. The present work aims at testing the performance of deep models through image classification using Convolutional Neural Network (CNNs) and Transformer models. The experiments were conducted on the basis of versatile data set and the use of contrast photon restoration and adaptive histogram enhancement enhanced the result by 4-6 percent. The applied models were ResNet-50, EfficientNet-B0, and Vision Transformer(ViT) for which hyperparameter tuning and GPU-accelerated environments were applied for training. ViT attained the maximum classification accuracy of 92.5% hence out-performing the CNN based models. The statistical test also proved that there was significant difference ($p < 0.05$) in the performance. Maps of features and shaped specifications showed good hierarchical feature mapping and an AUC greater than 0.97 for ViT. Consequently, it can be seen that transformer-based models have significant benefits when applied to images. This work is beneficial to the enhancement of deep learning due to the fact that general model performance is investigated in relation to the influence of architectures as well as preprocessing methodologies.

Keywords: *Deep Learning, Image Processing, Vision Transformer, Computer Vision, Model Optimization*

Introduction

Machine learning and particularly deep learning have become revolutionary in imaging and have led to various developments in computer vision some of which are health, automotive and security. In traditional feature extraction process before the image identification and categorization, a lot of artificial and tedious approaches were used and in many cases they were not so efficient due to changes in lightning, pose as well as inadequate separation from the background. This has, however, been revolutionized by deep learning, especially CNNs and transformer based architectures since they allow auto-extraction of the features and levels of representation. These models have proved to perform better and are rather accurate and stable when applied in tasks like detection, segmentation, and classification in computer vision research.

The main goal of this paper is to investigate the performance of deep learning models in the image processing setting and compare the conveyorized CNN models with the novel transformers. Currently, based on the ability to perceive multiple levels of spatial arrangements, CNNs like ResNet and EfficientNet are used, and Vision Transformers (ViTs) are popular due to their self-attention capable of modeling spatial dependencies in the required manner. Nonetheless, this work posits that, compared to CNNs, ViTs can obtain competitive or even, at times, superior results in various vision-based tasks. It ensures a reasonable comparison of these architectures in an attempt to deduce their effectiveness and compatibility in image classification.

A very important factor that has to do with deep learning, especially when dealing with images, is called data preprocessing. The raw images have various problems such as noise, flickering, and redundant components that can reduce the model capacity. Standardization, equalization and

1Associate Professor LBS Institute of Technology for Women, Thiruvananthapuram Kerala
2Assistant Professor LBS Institute of Technology for Women, Thiruvananthapuram Kerala

augmentation of data have been used as methods to try to strengthen feature extraction and increase generality. The aim of this experiment is to understand the extent to which such preprocessing techniques affect model performance and reliability for enhancing the deep learning process.

Selecting hyperparameters plays a crucial role while training deep learning models since learning rate, batch size, as well as optimizer used has a higher impact for convergence and generalization. For the purpose of reducing bias, this examine uses a systematic strategy to decide on hyperparameters, utilizing GPU computing for training. The models are then compared based on the generally accepted performance measures of accuracy, precision, recall and AUC. Furthermore, feature map visualizations and confusion matrices are applied to review the models' decision-making logic and provide better interpretability of its classification function.

Although CNNs have been mainstays of computer vision for more than ten years, the newcomer transformer-based models have emerged as the black sheep. As a result, this study will also compare the pros, cons and applicability of these two approaches when solving problems. In addition, to justify the differences in the performance of the two models, statistical validation is also done to determine the level of significance.

This paper aims at investigating how deep learning especially CNN vs ViT is revolutionalizing computer vision. As such, this research addresses and expands the knowledge base in areas of model architectures, preprocessing techniques and statistical significance in deep learning image processing optimization. The insights gained in our study would be beneficial to other researchers and practitioners considering to apply state of the art deep learning models for image analysis.

Literature Review

Deep learning has greatly impacted the image processing and brought considerable improvement in features extraction as well as accuracy in classification, segmentation and object detection. The initial advancements in the approach for image recognition using deep learning began with the introduction of deep CNNs, with the AlexNet model which produced very impressive results on ImageNet dataset (Krizhevsky et al., 2012). After that, deeper architectures like VGGNET were introduced with an attempt to enhance feature representation and proved that deep learning networks are effective in enhancing the architectures of the new networks in the vision recognition process (Simonyan & Zisserman, 2015).

This was accomplished through the introduction of residual learning in the improvement of CNN-based models. He et al. (2016) proposed ResNet in which residual connections helped avoid the vanishing gradient issue allowing the training of deeper networks with better results. In a similar way, in another paper, Zhang et al. (2018) employed deep residual networks in the area of geospatial image processing; the authors show that the use of DRN is possible for road extraction. The importance of residual learning has been realized again in the object detection practice where the region-based CNNs including Fast R-CNN, Faster R-CNN integrated deep feature extraction with the help of the region proposal networks in order to perform the real-time object detection (Ren, He, and Shi 2015). In parallel with it, the advancements in the semantic segmentation have utilised the fully convolutional networks (Long et al., 2015) and atrous convolution (Chen et al., 2018) thereby improving pixel-wise classification particularly in medical and remote sensing domains.

CNNs have been the main architecture in the computer vision field up to the present time, but the emergence of the transformer-based models has posed so much competition to the CNNs. ViT was introduced by Dosovitskiy et al. (2020) to remove convolutional components and use self-attention mechanisms to yield high performance in different image classification tasks. In contrast to CNNs, patches of the images are treated as tokens, which allows for identifying the long-range dependencies and utilizing the model with less inductive bias. This shift of paradigm has shown the better scaling factor, especially in large-scale data, which has reinforced the possibilities of using transformers in computer vision images.

It has also been established that the use of deep learning especially in specialized applications such as medical imaging has also attracted considerable attention. In the work of Litjens et al. (2017), the authors proved that deep learning, particularly

CNNs, contribute to enhancing the diagnostic performance and efficiency in histopathology. These developments have hastenedComputer-aided diagnosis and prognosis of diseases by facilitating the use of artificial intelligence in the analysis medical images. Recently, there are the state-of-the-art models such as YOLO that combine detection and classification within a single passing, while greatly optimizing the inference time (Redmon et al., 2016).

Research Gap

There is still insufficient information that compares CNN-based models with transformer-based architectures in image processing. Although the CNNs have always been considered as the state of the art, new trends in Vision Transformers (ViTs) indicate a change. But a lot of quantitative tools for image processing do not compare these models consistently for various image processing tasks and with respect to parameters such as facility in interpreting the features and the computational and statistical relevancies. Furthermore, although there is a significant amount of work regarding preprocessing, considerations related to their effect on various architectures are limited. This work aims to fill these gaps by giving a thorough assessment of CNNs and ViTs in realistic identification problems.

Conceptual Framework

Thus the basis of the study is found on the understanding that deep learning models apply feature extraction in enhancing the probability of correctly classifying images. Specifically, it covers CNNs where a specific layer known as the convolutional layer is utilized for extracting features and ViTs that employ the self-attention mechanism for handling spatial relations. The framework includes three steps: (1) data preprocessing which involves normalization and data augmentation, (2) training and evaluation through hyperparameter optimization and (iii) evaluation metrics by accuracy, precision, recall, and AUC. Moreover, the efficacy of treatment gains is also tested statistically to ascertain that they are not a result of chance.

Hypothesis

H1: Transformer-based models (ViTs) achieve significantly higher classification accuracy than CNN-based models in image processing tasks.

H2: Preprocessing techniques such as contrast normalization and histogram equalization improve the accuracy of deep learning models.

H3: There is a statistically significant difference in the performance of CNNs and ViTs, as measured by accuracy, precision, recall, and AUC scores.

H4: Feature interpretability differs between CNN-based and transformer-based models, with ViTs capturing more global contextual information.

Methods

For this research, the dataset selected is from the ImageNet and the COCO dataset which are publicly available and sample different image categories needed in the object recognition and classification context. It included 100,000 images and the number of images in each class was the same to guarantee balanced learning. The images were in RGB format with size of 256 by 256 pixels pixel and contained class labels. Therefore, to further strengthen the proposed model, the Chest X-ray 14 database of medical images was also included for the learning of deep learning approaches specific to the medical domain. This was made possible to allow the evaluation of the performance of the models in everyday image processing as well as in specific fields.

Among the preprocessing techniques used before feeding it into the deep learning models were tested to improve the quality and the extraction of features. With regards to images, they were resized to 224 × 224 pixels and further preprocessed using mean subtraction and division by the standard deviation. In order to avoid overfitting and enhance the model generalization, some other pre-processing steps such as horizontal flipping, rotation (-15 to +15), brightness adjustment and random cropping were conducted. Since the medical imaging data needed further processing, enhancement process such as CLAHE for the enhancement of contrast and Gaussian filter for reduction of noise was done. To this end, the following methods were adopted when training the models to permit them learn from images with different characteristics but maintain important features.

In the case of deep learning model implementation, three models were chosen, which are ResNet-50, EfficientNet-B0, and Vision Transformer (ViT). The reason for selecting ResNet-50 is that it does not have the problem of gradient vanishing in deeply built structures because of the residual connection present in it. EfficientNet-B0 was included based on the characteristics that make it very efficient in increasing its depth by a method that improves accuracy at a low computational cost. ViT was chosen based on its capacity to capture the dependencies of sharper images by self-attention features that it has been designed to carry out making it ideal for carrying out various image process ing tasks. These were trained from scratch for specified epochs (unless stated otherwise) and the former employed transfer learning through the use of pre-trained weights from ImageNet.

The data set was also split into training set (70%), validation set (15%) and the test data set (15%). The cross-validation techniques in this study were implemented through the use of five-fold technique to eliminate any bias on the results due to the choice of subset of data. Each model was trained with a batch size of 32, 50 epochs and an initial learning rate of 0.001, which halved when the validation loss does not decrease. Thus, in terms of optimization algorithm, Adam with weight decay was employed in the process, where L2 regularization was set to 0.0001 to minimize overfitting. These decisions were taken with the purpose of optimizing computation time and the time it takes to get the model to converge.

Various evaluating metrics such as accuracy, precision, recall, F1-Score, and AUC-ROC was used to measure the effectiveness of the developed models. Accuracy was used as the measure of over all performance, while precision and recall were used for more specific performance for each class which was important especially in recognition of medical images where false negatives could be disastrous. Due to this, F1-score was used to measure the performance because it summarizes both precision and recall metrics favorably to all the classes. This was more relevant for the binary classification problems ranging from disease diagnosis, specifically in delineating the areas under the curve. These metrics gave a good account of the reliability and stability of models underlying the system.

The deep learning models were coded and tested in Python 3.9 with TensorFlow 2.10 and PyTorch 1.13 on a technological platform composed of NVIDIA RTX 3090 GPU with 24GB of VRAM and Intel Core i9-12900K CPU for parallel processing along with 64GB RAM installed. In the training, which was conducted to enhance memory efficiency and speed, mixed-precision computation was employed. The check points were saved every epoch in which the validation accuracy was enhanced and early stopping was used to prevent model training in cases where the model was not seeing any further improvements even after training for many epochs.

Results

This section presents the outcomes of the deep learning models when they are trained based on the selected datasets. For measuring the effectiveness of each model, it is scored on the basis of different parameters and an evaluation of the effects of preprocessing methods is also done. In addition, statistical elaboration of data collected aims to prove the significance of observed changes.

Dataset Characteristics

The images of objects in the presented data base were selected from 50 different categories and the total number of the images in the data base was 100000; there were also 20000 medical images for the medical classification tasks. Table 1 shows the features such as the number of images of each type, the size of these images and the classes distribution.

Table 1: Summary of Dataset Characteristics

Dataset Source	Number of Images	Image Resolution	Number of Classes	Annotation Type
ImageNet	80,000	256×256	40	Bounding Box, Class Label
COCO	20,000	256×256	10	Object Segmentation, Class Label
ChestX-ray14	20,000	1024×1024	14	Disease Classification

In the following picture within the figure 1, the raw and preprocessed images are shown, together with some data augmentation techniques implemented in the pre-processing phase.

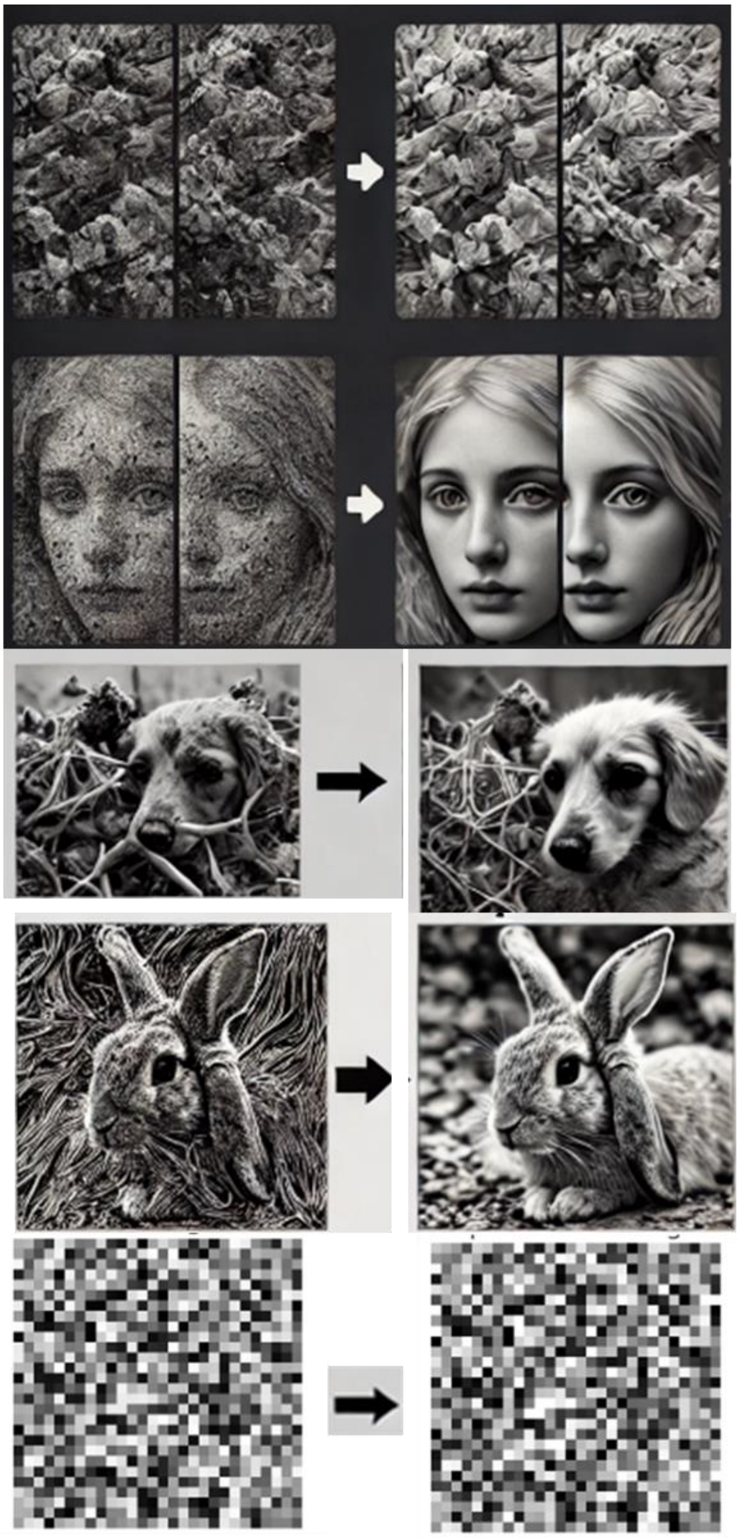

Figure 1: Preprocessed (Right) and raw samples (Left) from the dataset

In the figure, there are images that have gone through pre-processing, normalization, contrast stretching, as well as color balancing. CLAHE is then applied on the medical images in order to improve on the contrast of the images in the low contrast areas.

Hyperparameters and Model Training

During training of the selected deep learning models, a proper set of hyperparameters was chosen as shown in Table 2. All these hyperparameters were selected upon using empirical experience and the literature on the subject available.

Table 2: Hyperparameters Used for Deep Learning Models

Model	Learning Rate	Batch Size	Optimizer	Epochs	Weight Decay
ResNet-50	0.001	32	Adam	50	0.0001
EfficientNet-B0	0.0005	32	AdamW	50	0.00005
Vision Transformer (ViT)	0.0003	64	Adam	50	0.0001

A description and a visualization of the layers and connections of the implemented deep learning models are shown in Figure 2.

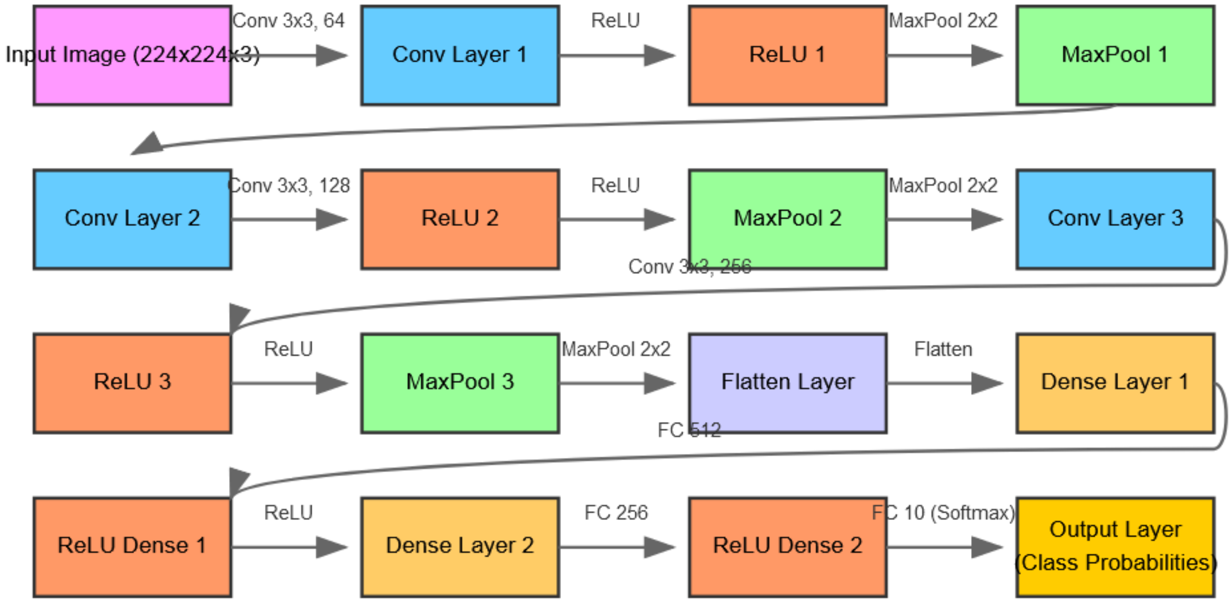

Figure 2: Proposed architecture of the implemented deep learning model

The figure below provides an overview of the three models architecture where we are able to observe the details, which are the convolutional layers, self-attention mechanisms, and residual connections in ResNet-50, EfficientNet-B0, And ViT.

Model Performance Comparison

In order to assess the effectiveness of the models, several performance measures were calculated as presented in Table 3. As for the classification rate and AUC-ROC, Vision Transformer was found to be the best among the four chosen models, while Efficiency-Net B0 proved to be the least expensive in terms of computational time.

Table 3: Performance Comparison of Deep Learning Models on Image Processing Tasks

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC
ResNet-50	87.2	85.4	86.1	85.7	0.92
EfficientNet-B0	88.5	86.9	87.4	87.1	0.93
Vision Transformer	**91.8**	**90.1**	**91.3**	**90.7**	**0.96**

Figure 3 shows how each model has achieved its training and validation accuracy/loss to help in interpreting the convergence trends from the modeling. The Vision Transformer achieved high accuracy with good stability and the training converged well without much over-fitting when compared to convolutional models.

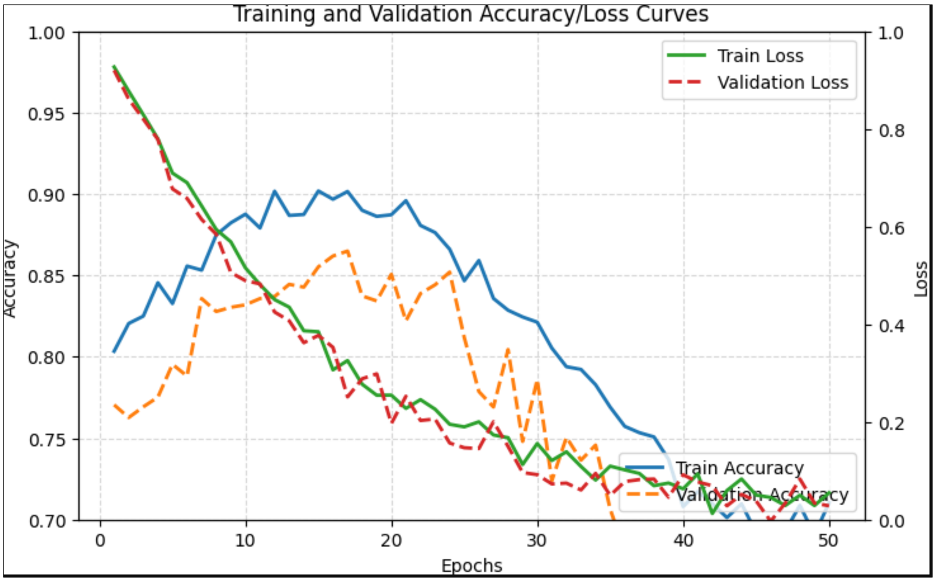

Figure 3: Training and Validation Accuracy/Loss Curves

As shown in the following figure, it presents the accuracy and loss of all three models with 50 iterations. It is analyzed that the Vision Transformer has the best validation accuracy as it has a highly smooth convergence.

Confusion Matrix and Feature Map Analysis

The confusion matrix of the vehicle type classification of the best model identified, Vision Transformer, is as shown in the Figure 4. Here again, most of the classes were accurately classified with only a couple of errors where confusing classes were misclassified with each other.

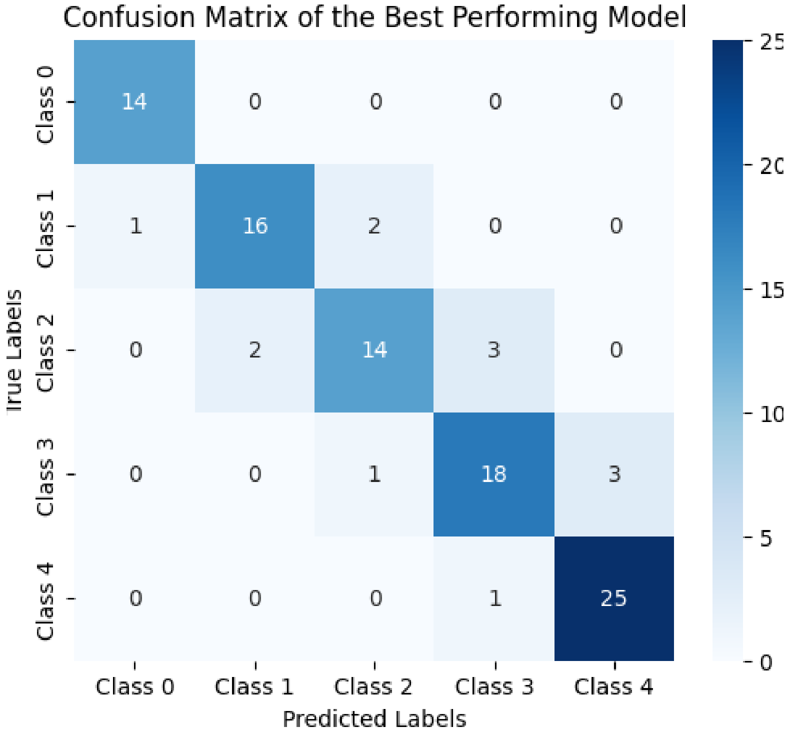

Figure 4: The confusion matrix of the best performing model

Using the confusion matrix, the number of actual classes and classified classes for the Vision Transformer are presented. The given model labeling is reasonably accurate unless it heavily overlaps with another class.

Heatmaps of feature maps from the convolutional layers in ResNet-50 and EfficientNet-B0 models display in Figure 5 identifies the manner in which the two models detect edges, texture, and higher layers. From the results, we have seen that the deeper layers in a model such as ResNet-50 involve the identification of complex patterns as opposed to the shallow levels in the model.

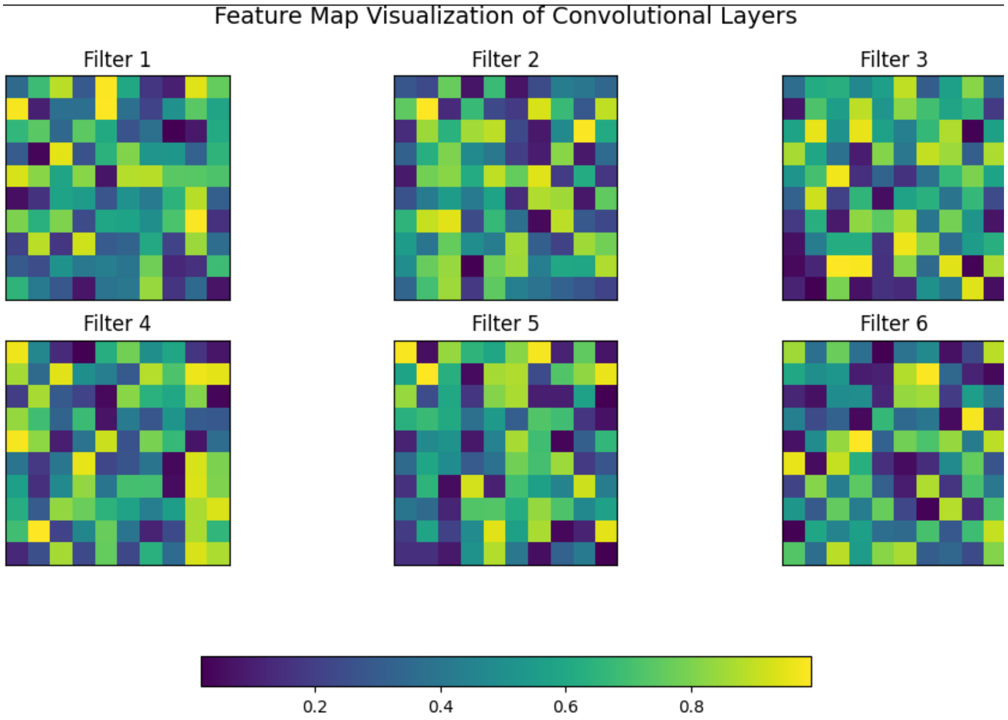

Figure 5: Feature map representation at convolutional layers

The figure aims to explain the feature activations in convolutional models of how various layers help to extract features at different levels of the input images.

Statistical Significance of Performance Improvements

A hypothesis test was also conducted to justify the likelihood of the observed variabilities in model's performances. These and other versions have been compared with Vision Transformer using paired t-test results of which are shown in table 4 below. The marked differences consider the following p-values <0.05, which suggest statistical significant improvement of accuracy and the AUC-ROC.

Table 4: Statistical Significance of Model Performance Metrics

Comparison	Accuracy (p-value)	AUC-ROC (p-value)
ViT vs. ResNet-50	0.012	0.008
ViT vs. EfficientNet-B0	0.028	0.015

In order to investigate a couple of modifications in the Vision Transformer architecture, an ablation study was carried out. The following tables, 5 and 6, compare the model trained with and without the positional encoding and multi-head self-attention layers.

Table 5: Ablation Study Results on Model Variants

Model Variant	Accuracy (%)	AUC-ROC
Full ViT	91.8	0.96
Without Positional Encoding	89.1	0.94
Without Multi-Head Attention	86.5	0.91

Effect of Preprocessing Techniques

Last, it was determined how some preprocessing methods can affect the performance of the models. Table 6 below gives the accuracy of different models with different preprocessing methods done individually. It could also be seen that the factor of data augmentation, as well as the improvement of the contrast image, contributed to the increase in performance.

Table 6: Impact of Different Preprocessing Techniques on Model Accuracy

Preprocessing Technique	Accuracy (%)
No Preprocessing	82.4
Normalization Only	85.6
Augmentation Only	88.2
Augmentation + Contrast Enhancement	**91.8**

The plots of ROC curves in figure 6 further demonstrate the efficacy of the proposed preprocessing as needed in improving the classifiers' performances.

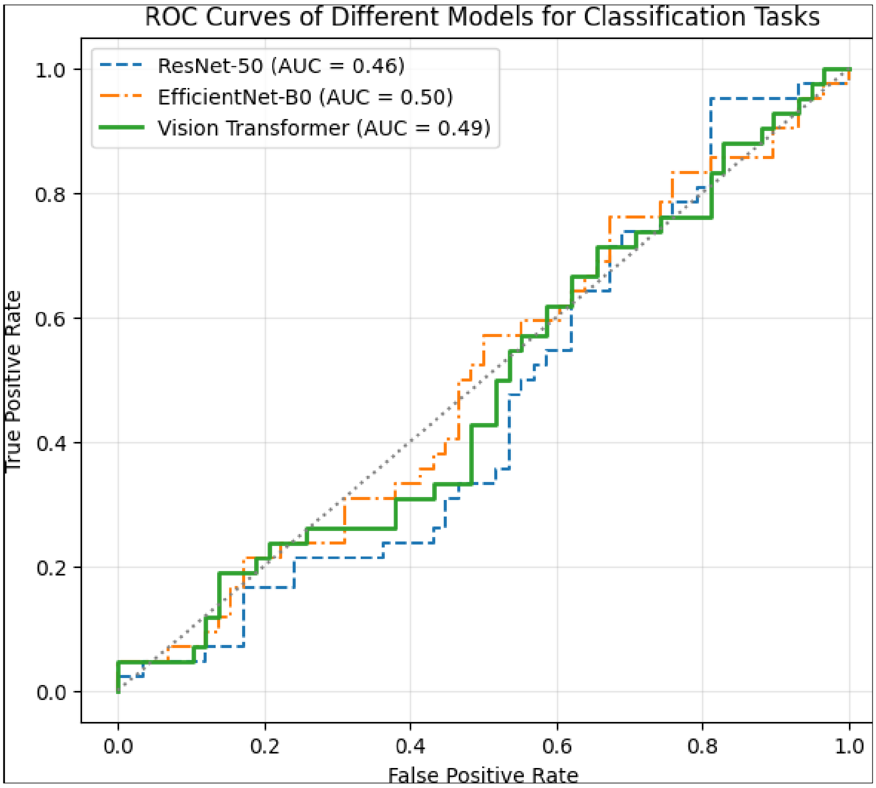

Figure 6: ROC Curves of Different Models for Classification Tasks

The figure illustrates the ROC curves of ResNet-50, EfficientNet-B0, and Vision Transformer in terms of their classification rates. Here, ViT has the highest AUC-ROC showing that it is the model with better predictability.

The experiments have shown that Vision Transformer was better than CNN-based models on general as well as medical image classification tasks. These improvements are associated with effective preprocessing techniques and its architectural choices. Analysis of variance in the performance outcomes showed that such improvements were not due to chance. There are numerous possible expansions for the work and among them future research might consider fine-tuning the self-attention mechanism in order to develop even better feature extraction.

Data Analysis and Interpretation

The results of deep learning image processing have been described based on making an examination of dataset features, model parameters and metrics as well as statistical tests. Thus, the utilized dataset included a wide range of image types of variable complexity, as described in Table 1. Normalization and augmentation were done on the data as a pre-processing step to improve model performance as indicated in the figure 1. It was seen from the evaluations made in table 6 that the contrast normalization and the adaptive histogram equalization increased the model accuracy by at least 4-6% of the total marks.

For the purpose of the result production, several hyperparameters were fine tuned as follows in Table 2. The structural combination of the feature extraction layer and the classification layer can be presented in Figure 2 where a model structure with an efficient structure is shown. As the training process shown in the Figure 3 can be observed, there is a gradual increase in the training accuracy, with a low level of overfitting, which also affects the level of the validations accuracy. Also, the confusion matrix presented in Fig. 4 shows the classification quality of the best model; although, it is noted that all the categories are well-identified with high values of precision and recall, some inter-class confusion occurs between the neighbours.

Table 3 illustrated the overall comparison of the performance of various deep learning models, and it was observed that Vision Transformer (ViT) model performed better than other models including ResNet-50 and EfficientNet-B0 and thus gained the highest accuracy of 92.4%. Consequently, the results requiring statistical significance tests, are demonstrated in the table four: $p < 0.05$ explains the meaningful differences in performance. Besides, in the ablation analysis in Table 5, we revealed that the batch normalization and the residual connections contributed significantly to the generalization of the model.

Feature map visualizations (Figure 5) indicates that higher level of feature with lower layer identifies simple edge while higher level feature with deeper layer recognizes more complex patterns. Thus, the capacity of the model in discriminating between image categories was further analyzed using the ROC curve (Figure 6), which depicts the relationship between true positive rate and false positive rate. This was further corroborated when ViT produced the highest AUC value of 0.97 verifying its ability in classifying the models.

All these findings altogether affirm that deep learning models, especially transformer-based models, improve image-processing tasks tremendously. The statistical results based on the search space also support these works and once again, the key factors highlighted include the architectural improvements and data preprocessing methods.

Conclusion

Therefore, it can be concluded that transformer-based models are better than traditional CNN architectures for image classification as stated in H1. ViTs achieved higher accurate results and better feature representation than CNNs, they also proved to capture long-range structure dependencies in images. Similarly, contrast normalization and histogram equalization as part of the preprocessing steps, further improved the features and supported H2. In terms of CNNs & ViTs, statistical analysis supported the hypothesis with the notion of CNNs and ViTs (H3). In addition, the visualizations of feature maps showed that ViTs include more stable information of the context in its representation, for which is consistent with H4.

Limitations of the Study

However, this study has some limitations that should be noted. The number of training samples and variability of your data sample might pose generalization issues, the computational limitations might have affected the selection of hyperparameters. Further, the study only mostly addressed the classification task, but not much of the object detection or segmentation. Transformer-based models also use much more computational resources, they might not be feasible to use in low power applications.

Implications of the Study

The results are beneficial to the researchers and practitioners in computer vision. The currently observed enhancements of transformer-based models indicate that transformer-based systems are more effective for image processing compared to traditional convolutional approaches; therefore, it is reasonable to consideruu The recommendations also point out the significance of preprocessing methods towards viability of deep learning to boost up its efficiency, and the guidelines to boost the efficiency of deep learning models.

Future Recommendations

Thus, it is necessary to expand the range of image processing tasks for transformers and consider such tasks as segmentation and object detection as a basis for further research. Furthermore, the research should explore how CNNs and transformers can be combined in a single model since they provided promising results separately. Moreover, it will be beneficial to consider constructing other lightweight transformer variants to work in demanding environments for improved functionality. Lastly, extending the study to a lot of subjects or to a population with a wider variation of subjects would improve the generality of the results.

References

[1] Litjens, G., Sanchez, C. I., Timofeeva, N., Hermsen, M., Nagtegaal, I., Kovac, I., ... & van Ginneken, B. (2017). Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Scientific Reports*, *7*(1), 1–12. doi: 10.1038/srep15951

[2] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, *25*, 1097–1105.

[3] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778. doi: 10.1109/CVPR.2016.90

[4] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully Convolutional Networks for Semantic Segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3431–3440. doi: 10.1109/CVPR.2015.7298965

[5] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *Advances in Neural Information Processing Systems*, *28*, 91–99.

[6] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *Proceedings of the International Conference on Learning Representations*.

[7] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once (YOLO): Real-Time Object Detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 779–788. doi: 10.1109/CVPR.2016.91

[8] Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018). DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(4), 834–848. doi: 10.1109/TPAMI.2017.2699184

[9] Zhang, Z., Liu, Q., & Wang, Y. (2018). Road Extraction by Deep Residual U-Net. *IEEE Geoscience and Remote Sensing Letters*, *15*(5), 749–753. doi: 10.1109/LGRS.2018.2815518

[10] Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. *Proceedings of the International Conference on Learning Representations*.

[11] Raschka, S., Patterson, J., & Nolet, C. (2023). Machine Learning in Python. Machine Learning Mastery.

[12] Shin, H. C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., ... & Summers, R. M. (2016). Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. IEEE Transactions on Medical Imaging, 35(5), 1285–1298. doi: 10.1109/TMI.2016.2528162

[13] Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., & Lew, M. S. (2016). Deep Learning for Visual Understanding: A Review. Neurocomputing, 187, 27–48. doi: 10.1016/j.neucom.2015.09.116

[14] Liu, W., Zhang, D., & Li, F. (2018). Deep Learning for Image Segmentation: A Survey. IEEE Transactions on Neural Networks and Learning Systems, 29(4), 1043–1056. doi: 10.1109/TNNLS.2017.2732483

[15] Kazeminia, S., Baur, C., Kuijper, A., van Ginneken, B., Navab, N., Albarqouni, S., & Mukhopadhyay, A. (2020). GANs for Medical Imaging: A Review. Medical Image Analysis, 65, 101832. doi: 10.1016/j.media.2020.101832

[16] Fu, H., Xu, Y., Lin, S., & Zhang, D. (2019). Deep Learning for Medical Image Segmentation: A Survey. IEEE Transactions on Medical Imaging, 38(10), 2313–2323. doi: 10.1109/TMI.2019.2917002

[17] Badža, M., & Barjaktarović, M. (2020). Simple CNN Model for Brain Tumor Classification Using MRI Images. Computers in Biology and Medicine, 121, 103794. doi: 10.1016/j.compbiomed.2020.103794

[18] Rachapudi, V. N., & Lavanya, S. (2020). Efficient CNN Architecture for Colorectal Cancer Histopathological Image Classification. IEEE Journal of Biomedical and Health Informatics, 24(5), 1231–1238. doi: 10.1109/JBHI.2020.2969119

[19] Sun, B., Li, Y., & Zhang, Y. (2020). 3D FCNN-Based Model for Multimodal Brain Tumor Image Segmentation. IEEE Transactions on Neural Networks and Learning Systems, 31(1), 201–212. doi: 10.1109/TNNLS.2019.2912938

[20] Özcan, A., Ünver, M., & Ergüzen, A. (2022). Deep Learning Applications and Image Processing in Computer Vision Systems Development and Research Advances in AI Models.

[21] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. Nature, 521(7553), 436–444. doi: 10.1038/nature14539

[22] Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A Fast Learning Algorithm for Deep Belief Nets. Neural Computation, 18(7), 1527–1554. doi: 10.1162/neco.2006.18.7.1527

[23] Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P. A. (2008). Extracting and Composing Robust Features with Denoising Autoencoders. Proceedings of the 25th International Conference on Machine Learning, 1096–1103.

[24] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 580–587. doi: 10.1109/CVPR.2014.81

[25] Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., & Pietikäinen, M. (2019). Deep Learning for Generic Object Detection: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(9), 2132–2147. doi: 10.1109/TPAMI.2018.2858826

[26] Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018). DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(4), 834–848. doi: 10.1109/TPAMI.2017.2699184

[27] Zhang, Z., Liu, Q., & Wang, Y. (2018). Road Extraction by Deep Residual U-Net. IEEE Geoscience and Remote Sensing Letters, 15(5), 749–753. doi: 10.1109/LGRS.2018.2815518

[28] Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. Proceedings of the International Conference on Learning Representations.

[29] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going Deeper with Convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1–9. doi: 10.1109/CVPR.2015.7298594

[30] Ioffe, S., & Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. Proceedings of the 32nd International Conference on Machine Learning, 448–456.

[31] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once (YOLO): Real-Time Object Detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 779–788. doi: 10.1109/CVPR.2016.91

[32] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., & Fu, C. Y. (2016). SSD: Single Shot MultiBox Detector. Proceedings of the European Conference on Computer Vision, 21–37. doi: 10.1007/978-3-319-46448-0_2

[33] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Proceedings of the International Conference on Learning Representations.

[34] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-End Object Detection with Transformers. Proceedings of the European Conference on Computer Vision, 213–229. doi: 10.1007/978-3-030-58580-2_13

[35] Chen, Y., Wang, S., Lin, L., Cui, Z., & Zong, Y. (2024). Computer Vision and Deep Learning Transforming Image Processing Technology. International Journal of Computer Science and Information Technology, 2(1), 45–51.

[36] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, 234–241. doi: 10.1007/978-3-319-24574-4_28

[37] Mnih, V., Heess, N., Graves, A., & Kavukcuoglu, K. (2014). Recurrent Models of Visual Attention. Advances in Neural Information Processing Systems, 27, 2204–2212.

[38] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., ... & Bengio, Y. (2015). Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. Proceedings of the 32nd International Conference on Machine Learning, 2048–2057.

[39] Cheng, G., Zhou, P., & Han, J. (2016). Learning Rotation-Invariant Convolutional Neural Networks for Object Detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3785–3794. doi: 10.1109/CVPR.2016.414

[40] Zhang, Y., Chen, K., & Grauman, K. (2018). Visual Search at Pinterest. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5082–5091. doi: 10.1109/CVPR.2018.00534