eISSN: 2976-1190

Journal of Engineering and Sciences

Vol. 2 Issue 1 November 30, 2023

Journal homepage: ***Journal of Engineering and Sciences (nepjol.info)***

Publisher: Research Management Cell, Pashchimanchal Campus, Lamachour, Pokhara

Kidney CT Scan Image Classification Using Modified Vision Transformer

Roshan Subedi[1], Suresh Timilsina[1], Smita Adhikari[1]

[1] *Department of Electronics and Computer Engineering, IOE, Pashchimanchal Campus, Tribhuvan University, Nepal*

(Manuscript Received 10/09/2023; Revised 18/10/2023; Accepted 22/11/2023)

Abstract

With the rising number of kidney-related health issues, early and precise diagnosis is crucial. The study aims to create a reliable method for categorizing kidney CT scan images into four groups: Cyst, Normal, Tumor, and stone. Traditional approaches usually rely on typical Machine Learning (ML) and Convolution Neural Networks (CNNs). However, in this research, the potential of a novel model called Vision Transformer (ViT) is explored. ViT was initially designed for Natural Language Processing (NLP) tasks but shows promise for medical image classification. ViT's capabilities are enhanced by coupling it with Fully Connected Networks (FCN). This combination helps to merge the feature extraction capability of the ViT and the classification ability of the FCN, which ultimately helps to overcome the challenge of detecting kidney-related issues.

Keywords: CNN; Classification; CT; MLP; Vision Transformer

1. Introduction

1.1 Background

Kidney-related health issues have witnessed a surge in recent years, necessitating accurate and timely diagnosis to mitigate the associated risks. Medical imaging, particularly the analysis of kidney CT scan images, plays a pivotal role in this diagnostic process. Traditional methods for classifying such images have demonstrated limitations in achieving the required accuracy.

In light of this, our research endeavors to bridge this diagnostic gap by harnessing the potential of modern deep-learning techniques. Our approach combines a Vision Transformer (ViT) with a Fully Connected Network (FCN) to classify kidney CT images into categories: Cyst, Normal, Tumor, and Stone. This innovative fusion of computer vision and deep learning brings a fresh perspective to medical image analysis.

Previous research efforts in this domain have predominantly relied on conventional machine learning techniques and Convolutional Neural Networks (CNNs) for kidney CT image classification. While these methods have shown promise, they often fall short when dealing with the intricate patterns and subtle abnormalities in medical images. The introduction of ViT-based architectures represents a paradigm shift, potentially addressing these limitations by enabling the model to capture long-range dependencies and intricate features.

Despite these promising strides, a research gap exists in applying ViT architectures to classify kidney CT scan images. While ViT models have demonstrated remarkable success in natural language processing tasks, their adaptation to medical image analysis remains relatively unexplored. Our study aims to fill this void, exploring the untapped potential of ViT models in healthcare diagnostics.

This research contributes by developing a specialized model tailored to kidney CT image classification, leveraging ViT for feature extraction and FCN for precise categorization. We seek to demonstrate the model's effectiveness in improving diagnostic accuracy through rigorous testing and validation. This research holds the promise of enhancing healthcare diagnostics in nephrology and radiology, potentially revolutionizing how kidney-related pathologies are detected and diagnosed.

1.2 Literature Review

A new approach in image processing has been explored [1], previously used for natural language processing. They introduced a method that treats images as a sequence, like the sequence of words in a sentence. To do this, the article suggests creating an image of size 224*224, converted into small regular,

[*] Corresponding author. Tel.: +977- 9846-79-4161,
E-mail address: subedirosan98@gmail.com

non-overlapping patches of size 16x16. The transformer encoder cannot determine the position of each part in the image, so position integration is performed. The transformer can only manage 1D vectors and, therefore, patches. The position encoding is converted into a 1D vector after embedding the layer token and passed to the transformer encoder. The encoder contains attention layers that help the model learn the relationship between neighboring images and other image patches far away from them. The Multilayer Perceptron (MLP) block is used for classification tasks using GELU in the hidden layer and a sigmoid function on the output layer. It turns out that this model works similarly to modern ones.

Convolution neural network (CNN) and different variations of vision transformers (ViT) were compared, and it was found that the ViT models are at least as resilient as the ResNet counterparts on a wide range of perturbations. Transformers can withstand the removal of nearly any single layer. Despite the fact that activation from later layers is highly correlated with one another, they are still crucial for categorization. Additionally, the model performs better by expanding the training dataset or reducing the patch size, resulting in more calculations. Additionally, research has shown that the MLP removal causes less pain than the attention layer removal in the transformer encoder [2].

Comparative analysis of the Deep Convolution Neural Networks and the Vision Transforms on the categorization of X-ray pictures and the results revealed that the Vision Transforms perform similarly to the CNN and are a competitive alternative. [3] Additionally, pretrained ViT models perform better than CNN-based models. Transfer learning on them, which have fewer medical picture datasets, can be used to achieve comparable or superior outcomes. Regarding classification accuracy, recall, and precision, the vit-based models are far better.

According to a comparison of the various CNN-based models and transformer-based models [4], Swin Transformer from the transformer and VGG16 from the CNN-based model perform better than ResNet, Inception V3 CNN-based models, EANet, and Compact Convolution Transformer. The model used a dataset of 12446 CT scan pictures divided into four classes. The 98.20% and 99.30% accuracy of the VGG16 model and swing transformer, respectively, improve performance. Comparable networks cannot match the models' recall and precision. The outcome indicates that a transformer-based design only needs a small number of features to predict classes.

The proposed work uses the pre-trained vision transformers for the feature extraction and FCN for classification tasks, as removing the MLP layer on the Transformer encoder doesn’t significantly affect the performance and the robustness of the ViT.

1.3 Contribution

It is discovered from the research mentioned above that CNN dominates picture classification jobs. However, a classification of X-ray images using the vision transformer revealed that it beat Deep Neural Networks [3]. ViT performs better when transfer learning is used. Additionally, VIT performs better with smaller patch sizes and more internal layers, but the complexity and training time are increased due to the increased layer count. Similarly, a comparison of the performance of CNN with Multilayer Perceptron [8] on several datasets revealed that CNN outperforms the latter. Regarding PSNR, OCR, and MSE, CNN performs superior to the MLP. Here, the model consists of transfer learning using vision transformers and FCN.

2. Methodology

2.1 Transformer

Transformer is the first transduction model relying entirely on self-attention to compute representations of its input and output without using sequence-aligned Recurrent Neural Networks (RNNs) or convolution [9]. Transformers consist of a multiheaded attention layer, a fully connected feed-forward network, and a normalization layer. These layers are stacked upon one another in a sequence, allowing the lower layer to inherit the feature learned by the previous layer. The core component of the transformers is its Multiheaded attention layer. Attention boosts how fast the model can translate from one sequence to another. They are based on attention mechanisms. An attention function can be described as mapping a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key [9].

The pretrained vision transformer model *Vit-base-16-224* was used for the feature extraction process. The model was pre-trained on ImageNet-21k, a dataset comprising 14 million images of 21K classes and tuned on ImageNet, consisting of 1 million images and 10000 classes at the resolution of 224×224 and the patch size of 16×16 [5]. Images are represented as a sequence of fixed-sized patches with a resolution of 16×16, which are linearly embedded.

Position and the class token are also embedded in the sequence before feeding to the stack of the transformer encoders. The pre-trained model learns the

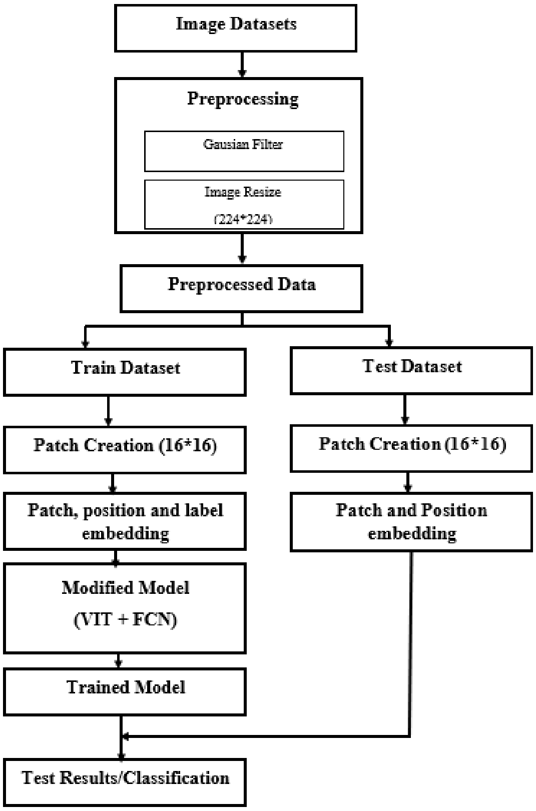

Figure 1: Model Architecture

inner representation of the images used to extract features useful for the downstream. A fully connected Network (FCN) is used at the top of the pretrained transformer for classification tasks. The sigmoid function is used on the output layer of the FCN to classify the images at the top of the different classes.

The proposed model workflow diagram is shown in Figure 1. A modified version of the Vision Transformer model, where the FCN replaces the MLP layer on the Transformer Encoder Layer to perform the final classification after the Vision Transformer has retrieved the features. Figure 1 provides a simple diagram for the suggested paradigm. The diagram provides a summary of how the operations are carried out. The images on the dataset were passed through the Gaussian filter to remove Gaussian noise and were resized to 224×224 patch of each image was created and the position of each patch in the images was embedded. The patch was fed to vision transformers. The vision transformer consists of several layers, each layer performing separate tasks. The normalization layer normalizes the patches using min-max normalization and is fed to the multi-headed attention layer. The multi-headed attention layer captures the relationships between each patch in the images and captures the features essential for image classification. After passing through the normalization layer, the extracted features are then fed to the FCN layer for classification purposes.

2.2 Dataset

The dataset includes 12446 images of the kidney from CT scans. Each class of photos has a name with four subfolders. The photographs were gathered from the publicly accessible online source Kaggle [6], which contains pictures of 5077 normal, 3709 cystic, 2283 tumorous, and 1377 stone-like structures. The photos were scaled to 224×224 [1] to remove Gaussian noise and passed through a Gaussian filter [7].

2.3 Model Architecture

The model introduces transfer learning using a vision transformer for feature extraction and a fully connected network (FCN) for classification. The model employs a transformer's encoder section, consisting of a normalization layer, multiheaded attention layer, and normalization layer followed by FCN. Here, the initial Multilayer perceptron previously used for classification was replaced by the FCN. Using the min-max normalization technique, the normalization layer normalizes the pixel value between 0-1. The multiheaded attention layer is an important layer in this model which is responsible for extracting features for classification tasks. The multiheaded attention layer finds the relation between each patch by using the cross product of query, key and score. Patches having higher scores bear large amounts of information. The feature vector then created is passed to the normalization layer for value normalization between 0-1 using min-max normalization. The data are fed to the FCN network, which consists of a dense network of hidden layers that perform classification tasks.

2.4 Training

The processed dataset was split into different combinations of train-test ratios. The training and testing datasets were broken down into smaller patches of 16*16 pixels. The position of each patch is embedded and generates a one-dimensional vector matrix. The labels are embedded in the vector. Now, the model was trained using various combinations of test-train split ratios, and the results obtained were tabulated and visualized. The model was trained with a learning rate of 0.001, batch size of 32, patch size of 16×16 and patch count of 196 for each image.

2.5 Loss Function

Cross-entropy loss function, also known as log function is used as a loss function. It establishes the relation between the predicted class and the true label of the data and is commonly used for multiclass classifications.

$$Cross - Entropy\ loss = -\sum y_i * \log(p_i) \quad (1)$$

Where, y_i = true label for the class 'i',
p_i = predicted probability for class 'i'.

This loss function measures how well the predicted class probabilities align with the true class labels. If the loss tends to 0, the predicted class is more likely to be the true class of an image.

3. Results and Discussion

3.1 Results

Gaussian noises are reduced when we pass the image into the Gaussian filter. The image is free from noises and some blurred edges, as shown in Figure 2. The filtered image is then broken into smaller pieces known as the patches of order 16×16. Hence, each image is now a group of 196 image patches, which are shown in Figure 3. Now, the images are linearly embedded into linear vectors after embedding the position and the class token on the unidirectional positional vector. The linear vector is now fed to the linear stack of the Transformer encoder which consists of attention mechanisms to find the relation of each image tile. The trained Vision Transformer model extracts 1000 features from the given image along with the class token and is now fed to the Fully Connected Network with 1000 neurons on the input layers and 4 neurons at the output layer with a sigmoid function as an activation function. The model was trained on the various combinations of train and test ratios and the results obtained afterward are shown in Figures 4, 5, and 6.

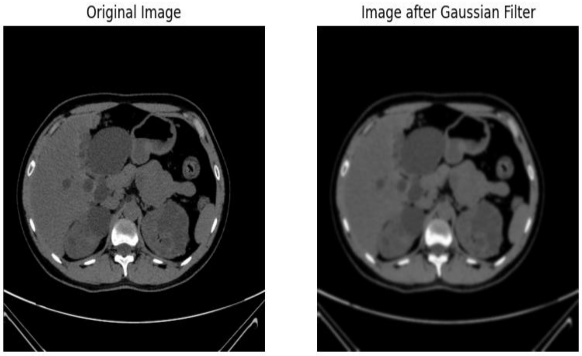

Figure 2: Image after passing from Gaussian Filter.

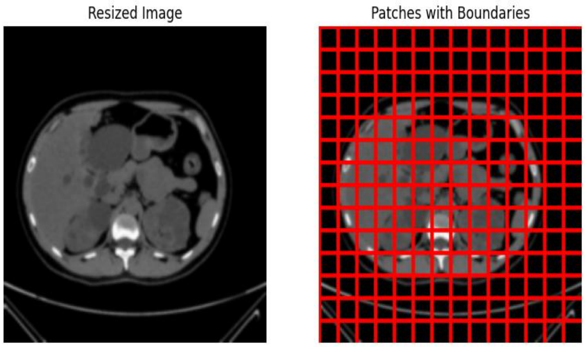

Figure 3: Resized image along with patches.

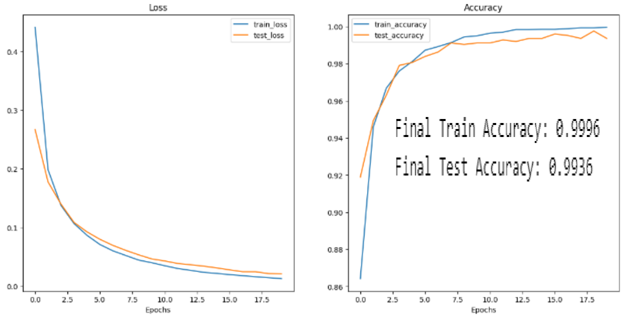

Figure 4: Loss and accuracy curve at a train-test ratio of 90:10

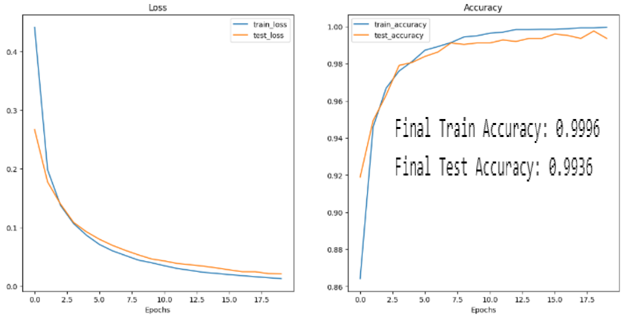

Figure 5: Loss and accuracy curve at a train-test ratio of 80:20

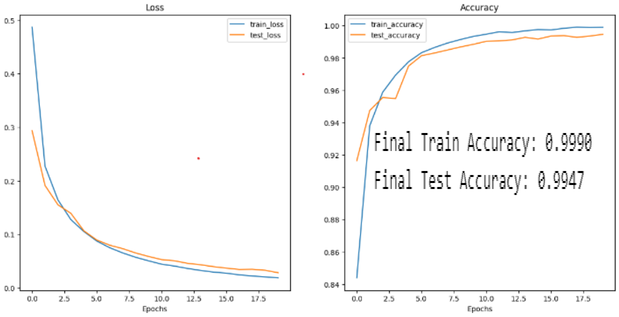

Figure 6: Loss and accuracy curve at a train-test ratio of 70:30

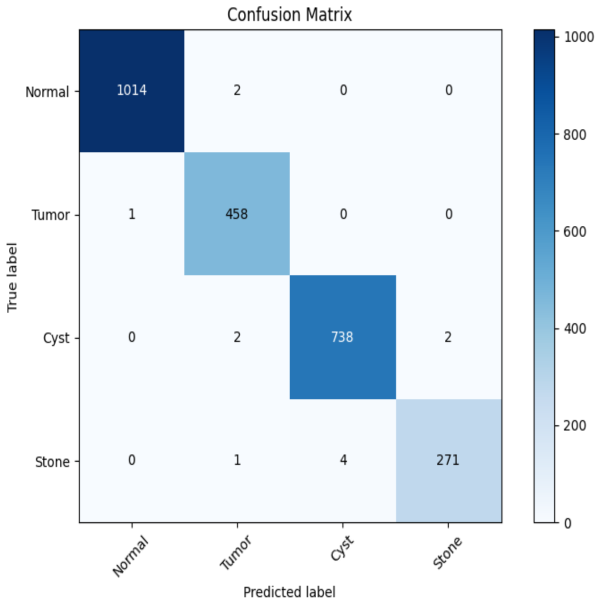

Figure 7: Confusion Matrix on the train-test ratio of 80:20

Figures 4, 5, and 6 show the promising results of the model in kidney CT scan image classification. The model training loss is in decreasing order, which seems to be much less than the 1% on each set of training and testing, which symbolizes the model is learning with the lowest error possible. Furthermore, the model performs best on the training and testing approach and classification at the 90:10 set, but the testing accuracy is better at the 80:20 set. Analyzing the facts, further analysis is done on the train test ratio of the dataset. The results obtained thereafter are shown in Figure 7 and Table 1. Table 1 depicts the class-wise performance of the model in terms of recall and precision. The model recall and precisions of the class Cyst and Normal are more compared to the other class as the number of images in the class is of varying nature; that is, the dataset is imbalanced. Still, the model shows at least similar or more than the other variants and popular CNN networks.

3.2 Discussion

Table 1. Classification report on train-test ratio of 80:20

Class	Precision	Recall	Fl-Score
Cyst	0.9990	0.9980	0.9985
Normal	0.9892	0.9978	0.9935
Tumor	0.9946	0.9946	0.9946
Stone	0.9927	0.9819	0.9872

The comparative analysis of the performance of the proposed model, along with other state-of-the-art**,** is shown in Table 2.

Table 2. Comparative Analysis

Model	Accuracy	Class	Recall	Precision	F1-Score
Resnet50	73.80%	Cyst	0.641	0.735	0.685
		Normal	0.79	0.77	0.78
		Tumor	0.827	0.706	0.762
		Stone	0.692	0.745	0.717
CCT	96.54%	Cyst	0.923	0.968	0.945
		Normal	0.975	0.989	0.982
		Tumor	0.964	0.964	0.964
		Stone	1	0.94	0.969
VGG16	98.20%	Cyst	0.968	0.996	0.982
		Normal	0.973	0.985	0.979
		Tumor	0.996	0.982	0.989
		Stone	0.988	0.966	0.977
Swin Transformer	99.30%	Cyst	0.996	0.996	0.996
		Normal	0.981	0.996	0.988
		Tumor	1	0.993	0.996
		Stone	0.989	0.981	0.985
Proposed model	99.64%	Cyst	0.998	0.999	0.9985
		Normal	0.9978	0.9892	0.9935
		Tumor	0.9846	0.9946	0.9946
		Stone	0.9819	0.9927	0.9872

The model's performance is better regarding recall, precisions, and F1-Score. From the comparative analysis of the different state of art, it is clear that the model is best among different classification models.

4. Conclusions

In this study, a comprehensive approach for the classification of CT scan images, aiming to distinguish between four clinically significant categories: Cyst, Normal, Tumor and Stone. The primary objective was to leverage the capabilities of a modified Vision Transformer (ViT) model, enhancing its ability to handle medical image data effectively.

The results obtained from experiments, as outlined in the adjacent Table 2, demonstrate the effectiveness of this proposed approach. A pretrained ViT model was successfully employed for feature extraction, followed by replacing its Multi-Layer Perceptron (MLP) head with a custom-designed Fully Connected Network (FCN) for classification tasks. The modification allowed us to harness the power of deep learning and transfer learning, resulting in notable improvements in accuracy, recall, and precision compared to traditional methods.

This comprehensive evaluation of the model on a diverse dataset of kidney CT scans yields promising results. Specifically, the obtained recall, precision, and F1-score value outperformed the existing approaches, underscoring the potential clinical relevance of the proposed model. These improvements are of utmost importance in the medical field, where accurate classification of kidney conditions can greatly assist healthcare practitioners in diagnosis and treatment planning.

This research has significantly contributed to the medical image analysis field. Incorporating a modified ViT model into the classification pipeline has demonstrated the viability of using cutting-edge deep learning techniques for Kidney CT scan classification tasks. This innovative approach extends the boundaries of traditional methods and showcases the potential for further advancement in the domain.

The significance of this task goes beyond the scope of this investigation. Our experiments' improved recall and precision have practical implications for healthcare providers, as accurate kidney condition classification can lead to more informed clinical decisions. Future research avenues may include further refinement of the model architecture, optimization of hyperparameters, and exploration of larger and more diverse datasets to enhance the model's generalizability.

Acknowledgment

The Department of Electronics and Computer Engineering, Pashchimanchal Campus, Pokhara, Nepal, has supported this work.

References

[1] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale, *ICLR 2021 Conference,* (2021).

[2] Bhojanapaili, S., Chakrabarti, A., Glasner, D., Li, D., Unterthiner, T., and Veit, A. Understanding robustness of transformers for image classification. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2021)

[3] Uparkar, O., Bharti, J., Pateriya, R. K., Gupta, R. K. and Sharma, A. Vision transformer outperforms deep convolutional neural network-based model in classifying X-ray images *Procedia Computer Science,* 218, (2023) 2338–2349.

[4] Islam, M. N., Hasan, M., Hossian, M. K., Alam, M. G. R., Uddin, M. Z., and Soylu, A. Vision transformer and explainable transfer learning models for auto detection of kidney cyst, stone and tumor from CT-radiography. *Scientific Reports,*12(1) (2022).

[5] Google/VIT-base-patch16-224 Hugging Face, see https://huggingface.co/google/vit-base-patch16-224 (last accessed Sep. 29, 2023).

[6] Islam, M. N. CT kidney dataset: Normal-cyst-tumor and stone, *Kaggle*, see https://www.kaggle.com/datasets/nazmul0087/ct-kidney-dataset-normal-cyst-tumor-and-stone (last accessed October 18, 2023).

[7] Shah, T. and Kadge, S. Analysis and identification of renal calculi in computed tomography images, *2019 International Conference on Nascent Technologies in Engineering (ICNTE),* (2019).

[8] Peyrard, C. Mamalet, F. and Garcia, C. A comparison between multi-layer perceptrons and convolutional neural networks for text image super-resolution, *Proceedings of the 10th International Conference on Computer Vision Theory and Applications*, (2015).

[9] Vaswani, A., Shazeer, N., Parmar, N, Uzskoreit, J., Jones, L., Gomez, A. N., Kaiser, L. and Polosukhin, I. Attention is all you need, *31st Conference on Neural Information Processing Systems,* (2017).