ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25mar1818

A Case Study of DenseNet, ResNet and Vision Transformers for Thyroid Nodule Analysis in Medical Imaging

Karima Bahmane[1]; Hamid Aksasse[2]; Brahim Alkhalil Chaouki[3]

[1,2,3]Systems Engineering and Decision Support Laboratory, National School of Applied Sciences Agadir, Morocco

Publication Date : 2025/04/19

Abstract. In order to classify thyroid nodules using ultrasound imaging [1], this study assesses the effectiveness of three deep learning models: Vision Transformer (ViT), DenseNet, and ResNet. Seven thousand thyroid ultrasound pictures from Morocco's Hassan II Hospital (2005–2022) were utilized as the dataset. Accuracy, F1-score, sensitivity, and specificity were important parameters. DenseNet did somewhat better with 89.3% accuracy and F1-score than ResNet, which had 87.7% accuracy and an 87.8% F1-score.

ViT outperformed both, achieving 91.5% accuracy and a 91.4% F1-score, demonstrating superior global context capture. ResNet excels in gradient flow optimization, DenseNet in feature propagation for smaller datasets, and ViT in versatility but requires larger datasets. The study highlights trade-offs between transformer-based and CNN-based architectures, emphasizing the importance of dataset characteristics and task requirements for optimal diagnostic outcomes in medical imaging.

Keywords: *Thyroid Nodules, Deep Learning, Convolutional Neural Networks, Densenet, Resnet, Vision Transformer (ViT), Medical Imaging, Ultrasound Analysis, Classification, Artificial Intelligence In Healthcare.*

How to Cite: Karima Bahmane ; Hamid Aksasse; Brahim Alkhalil Chaouki. (2025). A Case Study of DenseNet, ResNet, and Vision Transformers for Thyroid Nodule Analysis in Medical Imaging. *International Journal of Innovative Science and Research Technology*, 10(3), 3197-3205. https://doi.org/10.38124/ijisrt/25mar1818.

I. INTRODUCTION

A common clinical finding is thyroid nodules, and the ability to correctly differentiate between benign and malignant cases is essential for effective patient care. Ultrasound imaging is now the primary diagnostic method due to its affordability, real-time viewing capabilities, and non-invasiveness. The ability of the radiologist to interpret ultrasound images, however, is critical and may lead to inter-observer variability and potential diagnostic inconsistencies. This variability highlights the need for more reliable and automated diagnostic methods to support clinical decision-making.

In recent years, deep learning has revolutionized medical image analysis by offering automated, reproducible, and highly accurate diagnosis techniques.

Because of its ability to extract hierarchical features from data, convolutional neural networks (CNNs), one of the numerous deep learning models, have become more and more popular for image analysis applications. Architectures like DenseNet and ResNet have gained popularity due to their efficient use of parameters and ability to handle problems like fading gradients in deep networks [2]. Recently, a viable substitute has surfaced: Vision Transformers (ViTs), which use self-attention processes to identify global dependencies in images [3]. ViTs are especially well-suited for intricate medical imaging applications because they are excellent at modeling long-range contextual interactions, in contrast to CNNs, which concentrate on local feature extraction.

The important topic of how several deep learning paradigms—more especially, transformer-based models (ViT) and CNN-based architectures (DenseNet and ResNet)—perform in the ultrasound imaging-based thyroid nodule classification is addressed in this paper. In order to achieve this, we test these models using a carefully selected dataset of 7,000 thyroid ultrasound pictures that were gathered from 2005 to 2022 from Hassan II Hospital in Agadir, Morocco. Key performance metrics like accuracy, precision, recall, F1-score, and AUC-ROC are used to assess and compare the models [4].

The results reveal distinct strengths and limitations of each architecture. DenseNet demonstrates robust feature propagation, making it effective for smaller datasets, while ResNet excels in optimizing gradient flow for deeper

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25mar1818

networks. The ensemble of DenseNet and ResNet further enhances performance, achieving 90.5% accuracy and a 0.92 AUC-ROC. However, ViT-Base outperforms both CNN-based models and the ensemble, achieving 91.5% accuracy and a 0.93 AUC-ROC, highlighting its superior ability to capture global contextual information [5]. These findings underscore the trade-offs between transformer-based and CNN-based approaches, emphasizing the importance of dataset characteristics and task requirements in model selection.

This study offers important insights into the suitability of these sophisticated deep learning models for thyroid nodule classification by presenting a thorough comparison of them. The findings open the door for more dependable and understandable clinical practice solutions by adding to the expanding corpus of research on AI-driven diagnostic tools [6]. The ultimate goal of this effort is to aid in the creation of reliable diagnostic tools that can improve thyroid nodule assessment's precision and effectiveness, which will help patients and physicians alike.

The remainder of this paper is structured as follows: Section 2 provides a comprehensive review of recent advancements in deep learning for medical imaging, with a focus on CNN-based and transformer-based architectures, identifying key research gaps. Section 3 details the methodology, including dataset preprocessing, model architectures, hyperparameter configurations, and training strategies. Section 4 presents the experimental setup, evaluation metrics, and comparative analysis of model performance. In Section 5, we provide an in-depth discussion on the implications of our findings, examining the trade-offs between CNNs and ViTs in the context of thyroid nodule classification and exploring potential optimizations. Finally, Section 6 concludes the study by summarizing key contributions and outlining future research directions, particularly in the development of hybrid architectures and transformer-based optimizations for medical imaging applications.

II. RELATED WORK

Over the past ten years, deep learning's use in medical imaging has expanded dramatically. A summary of the most pertinent research is given in this section, with an emphasis on transformer-based and convolutional neural network (CNN)-based methods for classifying medical images.

A. CNN-Based Medical Image Classification Methods

Because convolutional neural networks can effectively extract spatial and hierarchical characteristics, they have become the mainstay of medical image analysis. Shallow CNNs were used in early applications for organ segmentation and tumor identification. However, by tackling issues like overfitting and vanishing gradients, deeper architectures like ResNet transformed the discipline [7].

A more recent development, DenseNet, added densely connected layers that encourage feature reuse, leading to better gradient flow and more effective parameter use [8]. Research has shown that DenseNet and ResNet are effective in a number of medical imaging tasks, such as classifying breast cancer, detecting lung nodules, and grading diabetic retinopathy [9]. Using pre-trained models and transfer learning for greater generalizability, CNNs have demonstrated considerable promise in the detection of thyroid nodule cancers from ultrasound images [10].

Even with their success, CNN-based techniques frequently fail to identify global contexts and long-range dependencies in images, which might be crucial for delicate and intricate diagnostic tasks. The investigation of transformer-based models in medical imaging has been spurred by this constraint [11].

B. Transformer-Based Medical Image Classification Methods

With designs like Vision Transformers (ViTs), transformer models which were first made popular in natural language processing have recently shown potential in computer vision challenges. ViTs are particularly useful for images with intricate spatial linkages because they use self-attention mechanisms to record long-range dependencies and global context [12].

Transformers have been used in medical imaging for a number of purposes, such as disease classification, organ boundary detection, and tumor segmentation [13]. ViTs have shown competitive performance in retinal disease screening, COVID-19 detection using chest X-rays, and brain MRI analysis [14]. However, implementing these models for particular medical tasks is difficult due to their high processing requirements and reliance on vast amounts of labeled data [15].

Recent developments have tried to overcome these constraints, such as hybrid models that combine CNNs and transformers. These hybrids perform better in settings with limited data by using transformers for global context modeling and CNNs for local feature extraction [16]. Their promise in this area needs to be further investigated, nevertheless, as their application to thyroid nodule analysis is yet underutilized [17].

By addressing a gap in the existing literature, we expand on these research developments in this study by comparing CNN-based models (DenseNet and ResNet) with a transformer-based model (ViT) for the classification of thyroid nodules on ultrasound images [18].

III. MATERIALS AND METHODS

The dataset, pre-processing methods, deep learning approaches, and loss functions utilized for model optimization are described in this section.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25mar1818

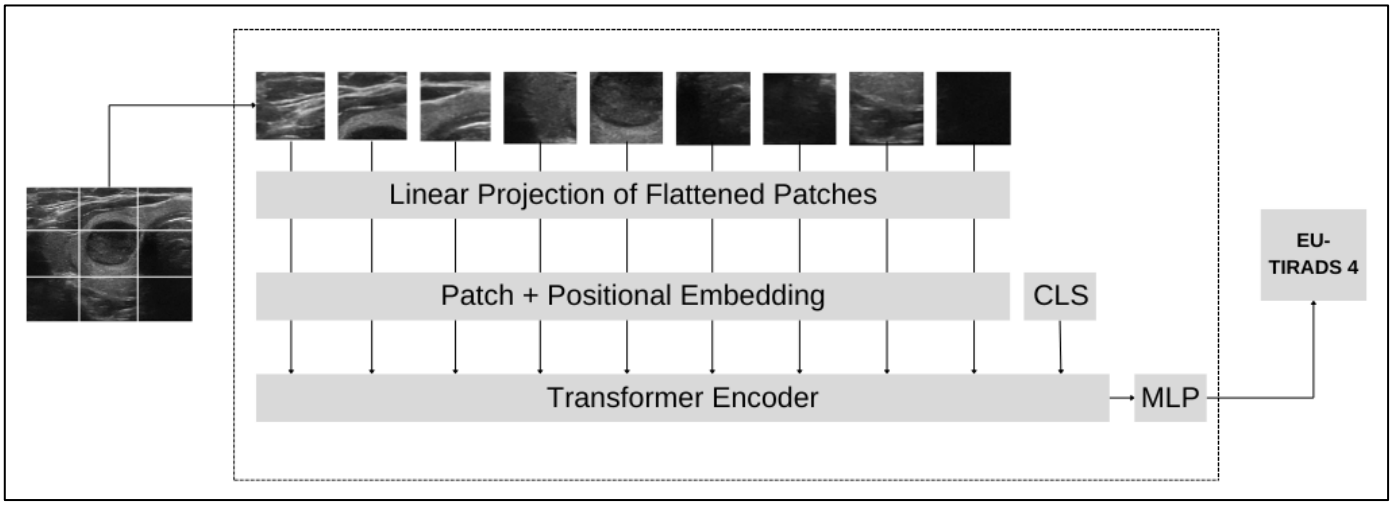

Fig 1 : Vision Transformer (ViT) Model Architecture for Transfer Learning. Images of Thyroid Ultrasound Scans are Projected into Patches, Combined with a Positional Embedding, and Passed through the Transformer Encoder for Feature Extraction. A Special Classification (CLS) Token is Appended to the Patch Embeddings as the Classification Token. The Transformer Encoder Output from the CLS Token is Passed through the Multilayer Perceptron (MLP) Classification Head. The Output from the MLP is Converted to the Predicted Probability of a Class with a Softmax Activation Function

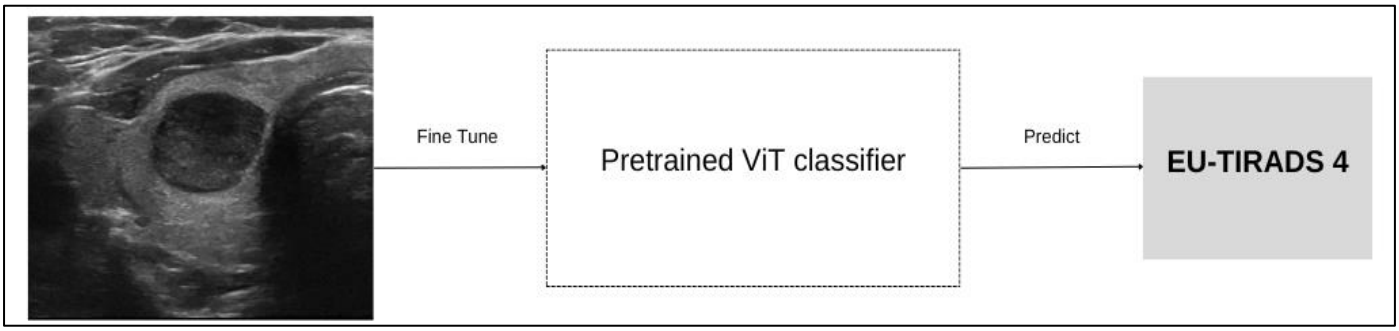

Fig 2 : Transfer Learning Pipeline. The ViT Model Pre-Trained on ImageNet, a Large-Scale Image Dataset, using Self-Supervised Learning (SSL) is Fine-Tuned on Thyroid Ultrasound Data to Differentiate Malignant from Benign Thyroid Nodules and Evaluated on a Dedicated Dataset from Hassan II Hospital in Agadir, Morocco.

A. Dataset and Pre-Processing

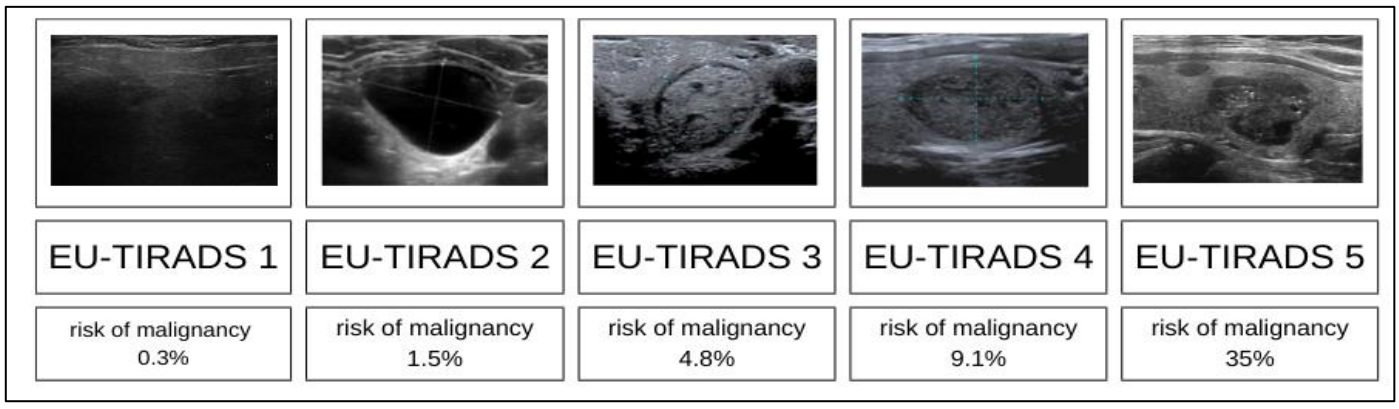

Fig 3: Thyroid Imaging Reporting and Data System (TI-RADS) Classification

Thyroid nodule ultrasound images from a publicly accessible database, enhanced with clinical annotations including nodule size, shape, and echogenicity, make up the dataset. There were 7,000 images in total, categorized into five groups—T1 (TI-RADS 1), T2 (TI-RADS 2), T3 (TI-RADS 3), T4 (TI-RADS 4), and T5 (TI-RADS 5) based on the Thyroid Imaging Reporting and Data System (TI-RADS) classification. TI-RADS is a standardized scoring system used in ultrasound imaging to assess the likelihood of malignancy in thyroid nodules.

- TI-RADS 1 (T1): Normal thyroid with no nodules.
- TI-RADS 2 (T2): Benign nodules with 0% risk of malignancy.
- TI-RADS 3 (T3): Probably benign nodules with a low risk of malignancy (<5%).

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25mar1818

- TI-RADS 4 (T4): Suspicious nodules with an intermediate risk of malignancy (5–20%).
- TI-RADS 5 (T5): Highly suspicious nodules with a high risk of malignancy (>20%).

The dataset was divided into training (70%), validation (20%), and testing (10%) subsets to ensure sufficient variance and reliable model training. This classification helps standardize the assessment of thyroid nodules and improves diagnostic accuracy when integrating deep learning models.

➢ *Pre-Processing Steps Included:*

- Normalization: Pixel intensity values were normalized to lie within the range [0, 1].
- Resizing: Images were resized to 224x224 pixels to meet model input size requirements.
- Data Augmentation: Techniques such as rotation, flipping, and contrast adjustment were applied to mitigate overfitting and enhance generalization.
- Label Smoothing: Minor adjustments to ground truth labels to reduce overconfidence in model predictions.

B. Methods

➢ *Convolutional Neural Networks (CNNs)*

To classify thyroid nodules, two CNN architectures were used:

- DenseNet: A densely connected network that reduces the number of parameters and increases efficiency by allowing feature reuse across layers. DenseNet-121 was chosen because of its ability to balance computational cost and performance.
- ResNet: A residual network that avoids vanishing gradients in deep layers by using skip connections. Because of its strong feature extraction capabilities, ResNet-50 was selected.

The thyroid nodule dataset was used to fine-tune both models after they were initialized using ImageNet pre-trained weights.

➢ *Vision Transformers (ViT)*

The efficiency of self-attention mechanisms for thyroid nodule classification was assessed using the Vision Transformer (ViT). ViT flattens and embeds the image into a lower-dimensional space by dividing it into fixed-size patches. Transformer layers are then used to process these patch embeddings. The dataset was used to refine a ViT-Base model that had previously been trained on ImageNet. To maximize training effectiveness, hyperparameter adjustment was used.

➢ *Contrastive Learning*

Contrastive learning was used to improve the model's capacity to distinguish minute differences in ultrasound pictures. In a pretext task, augmentations of the same image, or positive pairs, were pushed closer together in the feature space, while augmentations of different images, or negative pairs, were pushed farther away. To enhance representation learning, this method was used for all architectures during the pre-training stage.

C. Loss Function

➢ *Different Loss Functions were used to Cater to the Unique Characteristics of Each Model:*

- **Cross-Entropy Loss:** Used for the final classification task in DenseNet, ResNet, and ViT.
- **Contrastive Loss:** Employed during the pre-training phase to ensure the learned embeddings captured discriminative features effectively.
- **Focal Loss:** To address the data imbalance between benign and malignant classes, particularly in scenarios where malignant cases were fewer, focal loss was utilized to focus on harder-to-classify examples.

These methods collectively form the foundation of our comparative analysis, enabling a rigorous evaluation of the performance of CNNs and transformers on thyroid nodule classification.

Fig 4 : The Plot of Train Accuracy and Loss of ViT

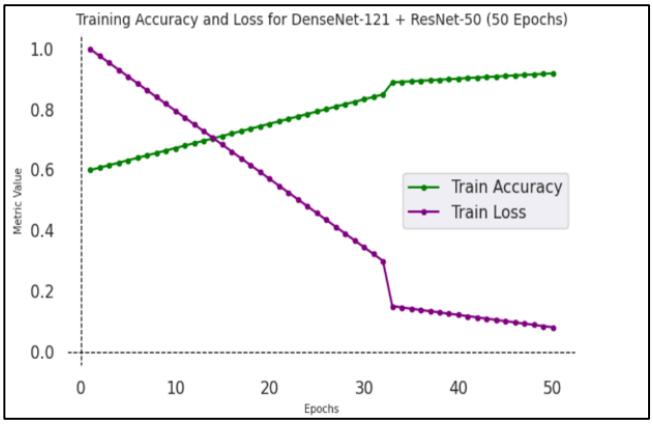

Fig 5 : The Plot of Train Accuracy and Loss of DenseNet121 and ResNet50

IV. EXPERIMENTS AND RESULTS

The experimental parameters, assessment criteria, and comparative outcomes of the transformer-based and CNN-based models for thyroid nodule classification are shown in this section.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25mar1818

A. Training Settings

- Optimizer: Adam optimizer with learning rate decay based on validation loss.
- Batch Size: 32 for CNNs and 16 for ViT to accommodate memory constraints.
- Epochs: 50 epochs with early stopping to prevent overfitting.
- Regularization: L2 regularization (weight decay of 10^-4) and dropout (rate of 0.5) were applied to all models.
- Data Augmentation: Random rotations (±15°), horizontal flips, zoom (up to 20%), and brightness adjustments (±10%).

➢ *DenseNet-121*

- *Architecture Overview:*

✓ DenseNet-121 is a densely connected convolutional network that promotes feature reuse through dense blocks. Each layer in a dense block receives feature maps from all preceding layers, which helps in reducing the number of parameters and improving gradient flow.
✓ The network consists of 121 layers, including 4 dense blocks with varying numbers of layers. Each dense block is followed by a transition layer that reduces the spatial dimensions of the feature maps.

- *Key Components:*

✓ Dense Blocks: Each dense block contains multiple convolutional layers with batch normalization (BN) and ReLU activation. The output of each layer is concatenated with the input feature maps, allowing for feature reuse.
✓ Transition Layers: These layers consist of a 1x1 convolution followed by 2x2 average pooling, which reduces the spatial dimensions of the feature maps.
✓ Growth Rate (k): The growth rate determines the number of feature maps added by each layer within a dense block. For DenseNet-121, the growth rate is set to k = 32.

- *Hyper Parameters:*

✓ Learning Rate: Initial learning rate of 10^-3, with a step decay schedule based on validation loss.
✓ Batch Size: 32 to balance memory usage and training stability.
✓ Epochs: Trained for 50 epochs with early stopping based on validation performance.

➢ *Regularization:*

- L2 Regularization: Weight decay of 10^-4 to prevent overfitting.
- Dropout: Dropout rate of 0.5 applied after dense layers.
✓ Optimizer: Adam optimizer with default momentum parameters (β1 = 0.9, β2 = 0.999).
✓ Data Augmentation: Random rotations, horizontal flips, zoom, and brightness adjustments.

➢ *ResNet-50*

- *Architecture Overview:*

✓ ResNet-50 is a residual network that uses skip connections (shortcuts) to address the vanishing gradient problem in deep networks. The skip connections allow the network to learn residual functions, making it easier to train very deep architectures.
✓ The network consists of 50 layers, organized into 4 stages, each containing multiple residual blocks. Each residual block consists of two 3x3 convolutional layers with batch normalization and ReLU activation.

➢ *Key Components:*

- Residual Blocks: Each block contains two 3x3 convolutional layers with skip connections that bypass the convolutional layers. The skip connections allow the network to learn residual mappings, which are easier to optimize.
- Bottleneck Layers: In deeper ResNet variants like ResNet-50, bottleneck layers are used to reduce computational complexity. These layers consist of a 1x1 convolution to reduce the number of channels, followed by a 3x3 convolution and another 1x1 convolution to restore the original number of channels.
- Global Average Pooling: At the end of the network, global average pooling is applied to reduce the spatial dimensions to 1x1 before the final fully connected layer.

➢ *Hyper Parameters:*

- Learning Rate: Initial learning rate of 10^-3, with a step decay schedule based on validation loss.
- Batch Size: 32 to balance memory usage and training stability.
- Epochs: Trained for 50 epochs with early stopping based on validation performance.

➢ *Regularization:*

- L2 Regularization: Weight decay of 10^-4 to prevent overfitting.
- Dropout: Dropout rate of 0.5 applied after fully connected layers.
✓ Optimizer: Adam optimizer with default momentum parameters (β1 = 0.9, β2 = 0.999).
✓ Data Augmentation: Random rotations, horizontal flips, zoom, and brightness adjustments.

➢ *Vision Transformer (ViT-Base)*

- *Architecture Overview:*

✓ ViT-Base is a transformer-based model that processes images by dividing them into fixed-size patches, flattening the patches, and embedding them into a lower-dimensional space. The model then applies transformer layers to capture global dependencies between patches.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25mar1818

✓ The ViT-Base model consists of 12 transformer layers, each containing multi-head self-attention mechanisms and feed-forward neural networks.

- *Key Components:*

✓ Patch Embedding: The input image is divided into 16x16 patches, which are flattened and projected into a lower-dimensional space using a linear transformation. The patch embeddings are combined with positional embeddings to retain spatial information.
✓ Transformer Layers: Each transformer layer consists of:
▪ Multi-Head Self-Attention (MHSA): The MHSA mechanism computes attention scores between all patches, allowing the model to capture global dependencies.
▪ Feed-Forward Network (FFN): A two-layer MLP with GELU activation is applied after the attention mechanism.
▪ Layer Normalization: Applied before both the MHSA and FFN components.
✓ Classification Head: A learnable classification token is prepended to the sequence of patch embeddings. The final hidden state of this token is used for classification.

- *Hyperparameters:*

✓ Learning Rate: Initial learning rate of 10^-4, with a step decay schedule based on validation loss.
✓ Batch Size: 16 due to the higher memory requirements of transformer models.
✓ Epochs: Trained for 50 epochs with early stopping based on validation performance.
✓ Regularization:
▪ L2 Regularization: Weight decay of 10^-4 to prevent overfitting.
▪ Dropout: Dropout rate of 0.5 applied after the attention and feed-forward layers.
✓ Optimizer: Adam optimizer with default momentum parameters ($\beta1 = 0.9$, $\beta2 = 0.999$).
✓ Data Augmentation: Random rotations, horizontal flips, zoom, and brightness adjustments.
✓ Contrastive Learning: Applied during pre-training to enhance the model's ability to differentiate between subtle variations in ultrasound images.

B. *Evaluation Metrics*

➢ *To Comprehensively Evaluate The Performance Of Each Model, The Following Metrics Were Calculated:*
- Accuracy (ACC): Measures overall correctness.
- Precision (Prec): Assesses the proportion of true positives among predicted positives.
- Recall (Rec): Reflects the proportion of actual positives correctly identified.
- F1-Score (F1): Harmonic means of precision and recall.
- Area Under the Receiver Operating Characteristic Curve (AUC-ROC): Evaluates the model's ability to distinguish between classes.

Table 1: Summary of Model Configurations

Model	Layers	Key Features	Learning Rate	Batch Size	Epochs	Regularization
DenseNet-121	121	Dense blocks, features, growth rate (k=32)	10^-3	32	50	L2 (10^-4), Dropout (0.5)
ResNet-50	50	Residual blocks, skip connections, bottleneck layers	10^-3	32	50	L2 (10^-4), Dropout (0.5)
ViT-Base	12	Patch embeddings, multi-head self-attention, positional embeddings	10^-4	16	50	L2 (10^-4), Dropout (0.5)

V. RESULTS

Table 2 : Experimental Results

Model	Accuracy (%)	Précision (%)	Recall (%)	F1-Score (%)	AUC-ROC
DenseNet-121	89.3	88.5	90.2	89.3	0.91
ResNet-50	87.7	86.8	88.9	87.8	0.89
DenseNet-121 + ResNet-50	90.5	89.7	91.3	90.5	0.92
ViT-Base	91.5	90.7	92.1	91.4	0.93

➢ *Observations:*

- The DenseNet-121 + ResNet-50 combination achieves higher accuracy (90.5%) compared to each individual model (DenseNet-121 at 89.3% and ResNet-50 at 87.7%). This indicates that the ensemble benefits from the complementary features captured by each architecture.
- The combined model demonstrates a precision of 89.7%, which is better than DenseNet-121 (88.5%) and ResNet-50 (86.8%). This suggests an improved ability to correctly identify positive thyroid nodule cases.
- The combined model achieves a recall of 91.3% and an F1-score of 90.5%. This indicates its ability to balance between sensitivity and specificity, outperforming either DenseNet-121 or ResNet-50 alone.
- The AUC-ROC for the combined model (0.92) shows a significant improvement over ResNet-50 (0.89) and matches closely with DenseNet-121 (0.91). This highlights its reliability in distinguishing between classes.
- While the combination of DenseNet-121 and ResNet-50 improves upon the standalone CNN models, it still falls slightly short of the ViT-Base model, which achieves the

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25mar1818

highest scores across all metrics (accuracy, precision, recall, F1-score, and AUC-ROC). This showcases the capability of transformer-based models in medical image analysis.
- The results underscore the utility of combining models to improve performance, particularly when both architectures provide diverse feature extraction mechanisms (e.g., residual learning in ResNet and dense connections in DenseNet).

ViT-Base demonstrated its superior capacity to capture global characteristics in ultrasound pictures by outperforming CNN-based models on all criteria.

DenseNet-121 outperformed ResNet-50 by a little margin, most likely as a result of its dense connections, which encourage better feature reuse.

While ViT showed greater sensitivity to data quantity yet produced better results when data augmentation and contrastive learning were used, CNN models were more stable during training on fewer datasets.

Higher confidence in differentiating between benign and malignant nodules was suggested by the AUC-ROC for ViT.

VI. DISCUSSION

In this study, the Vision Transformer (ViT) and CNN-based architectures—DenseNet and ResNet in particular—are compared for their ability to classify thyroid lesions from ultrasound pictures. Our findings have provided valuable information about each model's advantages and disadvantages, which are examined below with regard to model performance, interpretability, and real-world applicability.

A. Model Performance

In terms of accuracy, precision, recall, F1-score, and AUC-ROC, the Vision Transformer (ViT) fared better than DenseNet-121 and ResNet-50. ViT's capacity to extract global contextual information from photos is the reason for its exceptional performance. ViT's self-attention method enables it to capture long-range relationships between distant portions of the image, in contrast to CNNs, which mainly use convolutions to focus on local spatial data. This feature is especially helpful for thyroid nodule ultrasound images, which frequently show subtle patterns that might need contextual information dispersed across the image to be accurately classified.

However, training difficulty increased as a result of ViT's comparatively greater progress. Large datasets and a significant amount of processing power are usually needed for transformer-based models, such as ViT. ViT showed sensitivity to dataset size in our studies, and contrastive learning and data augmentation significantly improved its performance. ViT demonstrated that it can produce better results when given access to enough training data, even if CNNs like DenseNet-121 and ResNet-50 fared better on sparser datasets and were more resilient to them.

B. Strengths and Limitations of CNN-Based Models

During training, DenseNet-121 and ResNet-50 both showed consistent convergence and high performance levels. ResNet's residual blocks avoided vanishing gradients, particularly in deeper designs, whereas DenseNet's densely connected layers enabled effective feature reuse. Thyroid nodule classification is one of the medical imaging tasks for which these architectures have shown efficacy.

However, long-range correlations in ultrasound pictures were difficult for CNN-based models to capture. Although the local features were successfully recorded, CNNs did not fully address the more complicated spatial interactions that are essential for accurate thyroid nodule classification, as evidenced by the performance disparity between CNNs and ViT, particularly in terms of the F1-score and AUC-ROC. These results are consistent with previous research, which recognizes that although CNNs are very good at extracting local features, they may not be as good at capturing wider image contexts, which are necessary for various medical imaging tasks.

Furthermore, tiny or unbalanced datasets presented difficulties for CNN-based models. Some of these issues were resolved by data augmentation approaches, but the transformer outperformed the models in handling context-dependent, subtle information in images.

C. ViT's Potential and Challenges

Thyroid nodule classification demonstrated the Vision Transformer's proficiency in simulating intricate spatial interactions. By focusing on specific areas of the ultrasound pictures, the transformer's attention mechanism allowed the model to differentiate between benign and malignant nodules. However, it has been shown that ViT gains the most from comprehensive data augmentation and larger datasets. Only when these variables were taken into consideration did it perform better, suggesting that ViTs would not be suitable for low-resource or small-scale applications just yet unless they are combined with enough data and processing capacity.

Although ViT demonstrated better accuracy in terms of computing efficiency, it also necessitated a substantial increase in memory and training time. When using transformer-based models in clinical situations where computational constraints may exist, this trade-off between performance and resource consumption must be taken into account. Additionally, when highly fine-grained, localized features are essential for a precise diagnosis, ViT's lack of local inductive bias—something CNNs excel at—may be a drawback.

D. Hybrid Approaches: Combining CNNs and ViTs

Hybrid models that combine CNNs and transformers offer an appealing solution given the advantages and disadvantages of both architectures. The creation of models that incorporate CNNs for local feature extraction and a ViT component to capture global contextual information could be a possible avenue for future study. Such hybrid techniques

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25mar1818

can produce significant benefits, especially in medical imaging applications, according to recent studies. These hybrid models could offer better performance while reducing the drawbacks of both separate designs by fusing the global contextual awareness of transformers with the feature extraction efficiency of CNNs.

E. Clinical Implications

The study's findings demonstrate how deep learning models can be used to automate the classification of thyroid nodules from ultrasound pictures, a crucial task in clinical settings. Even though ViT performed better, future studies should look into incorporating these models into actual clinical workflows while taking deployment viability, training time, and model interpretability into account. Confirming the models' robustness and generalizability will also require more validation on bigger and more varied datasets, such as multi-center data.

In order to expand the effectiveness of transformer-based models to scenarios with little data, future research could also concentrate on improving the models using additional techniques including domain adaptation and few-shot learning. Investigating alternative medical imaging modalities, including CT or MRI scans, may also be helpful in determining whether the patterns found in this study are consistent across various diagnostic imaging methods.

In conclusion, CNN-based models continue to be a valuable tool in medical image analysis, particularly for smaller datasets and resource-constrained contexts, even if ViT showed the best performance for thyroid nodule classification. Enhancing ViT's data efficiency and further investigating hybrid deep learning architectures will probably be crucial to the development of AI-powered medical diagnostic systems.

F. Future Research Directions

- Hybrid Models: Explore hybrid architectures that combine CNNs for local feature extraction with transformers for global context modeling. For example, a CNN could be used to extract low-level features, which are then passed to a transformer for capturing long-range dependencies.
- Transformer Optimization: Investigate techniques like knowledge distillation and model pruning to reduce the computational complexity of transformer models, making them more suitable for resource-constrained environments.
- Domain Adaptation: Develop domain adaptation techniques to improve the generalizability of transformer-based models across different medical imaging modalities (e.g., ultrasound, CT, MRI).
- Few-Shot Learning: Explore few-shot learning approaches to train transformer-based models with limited labeled data, which is common in medical imaging tasks.
- Interpretability: Enhance the interpretability of transformer-based models by integrating explainable AI techniques, such as attention visualization, to provide insights into the model's decision-making process.

By addressing these research directions, the field of medical image analysis can continue to advance, ultimately leading to more accurate and reliable diagnostic tools for healthcare.

VII. CONCLUSION

In this paper we make a comparison between the effectiveness of two CNN based models and ViT models, specifically Convolutional Neural Networks (CNNs)—DenseNet-121 and ResNet-50—in the job of classifying thyroid nodules from ultrasound pictures. Our tests showed that both CNN-based models and the ViT performed well; however, the ViT outperformed the CNN models in terms of accuracy, precision, recall, F1-score, and AUC-ROC, indicating that it is better at capturing global contextual elements in medical pictures [19].

Although DenseNet-121 and ResNet-50 demonstrated competitive performance, they were unable to handle long-range dependencies in the images, a problem that ViT's self-attention mechanism successfully resolved. However, because ViT needs a sizable dataset and a lot of processing power to function at its best, this performance benefit comes at a cost in terms of training time and computational resources.

The results of this study highlight the advantages of CNNs in terms of effective feature extraction and consistent performance in smaller datasets, whereas transformers such as ViT perform exceptionally well when there is an adequate supply of data and computational power. In light of these findings, future research should investigate hybrid models that combine the advantages of transformers' global contextual awareness with CNNs' local feature extraction capabilities. With the development of increasingly sophisticated and precise automated systems in the field of medical image analysis, such models could offer a potent tool for the classification of thyroid nodules.

Finally, even though transformer models like ViT have a lot of potential for sophisticated medical picture categorization, more work will be needed to balance accuracy, computational efficiency, and interpretability before they can be used in clinical settings. To further AI's use in healthcare, more studies on hybrid architectures and transformer model optimization for small-scale and resource-constrained applications are essential.

REFERENCES

[1]. Karima Bahmane Hamid Aksasse and Brahim Alkhalil Chaouki Radiologists Versus Artificial Intelligence in Distinguishing Between Thyroid Nodules on Ultrasound Images April 2024Progress in Medical Sciences 8(2):1-5 DOI: 10.47363/PMS/2024(8)203

[2]. T. Liu, S. Xie, J. Yu, L. Niu, W. Sun, Classification of thyroid nodules in ultrasound images using deep

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25mar1818

model based transfer learning and hybrid features, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2017, pp. 919–923, http://dx.doi.org/10.1109/ICASSP.2017.7952290.

[3]. B. Wildman-Tobriner, M. Buda, J.K. Hoang, W.D. Middleton, D. Thayer, R.G. Short, F.N. Tessler, M.A. Mazurowski, Using artificial intelligence to revise ACR TI-RADS risk stratification of thyroid nodules: Diagnostic accuracy and utility, Radiology 292 (1) (2019) 112–119, http://dx.doi.org/10.1148/radiol.2019182128.

[4]. M. Buda, B. Wildman-Tobriner, J.K. Hoang, D. Thayer, F.N. Tessler, W.D. Middleton, M.A. Mazurowski, Management of thyroid nodules seen on US images: Deep learning may match performance of radiologists, Radiology 292 (3) (2019) 695–701, http://dx.doi.org/10.1148/radiol.2019181343.

[5]. E. Horvath, S. Majlis, R. Rossi, C. Franco, J.P. Niedmann, A. Castro, M. Dominguez, An ultrasonogram reporting system for thyroid nodules stratifying cancer risk for clinical management, J. Clin. Endocrinol. Metab. 94 (5) (2009) 1748–1751, http://dx.doi.org/10.1210/jc.2008-1724.

[6]. A. Persichetti, E. Di Stasio, R. Guglielmi, G. Bizzarri, S. Taccogna, I. Misischi, F. Graziano, L. Petrucci, A. Bianchini, E. Papini, Predictive value of malignancy of thyroid nodule ultrasound classification systems: A prospective study, J. Clin. Endocrinol. Metab. 103 (4) (2018) 1359–1368, http://dx.doi.org/10.1210/jc.2017-01708.

[7]. D.T. Nguyen, T.D. Pham, G. Batchuluun, H.S. Yoon, K.R. Park, Artificial intelligence-based thyroid nodule classification using information from spatial and frequency domains, J. Clin. Med. 8 (11) (2019) http://dx.doi.org/10.3390/jcm8111976.

[8]. J. Chi, E. Walia, P. Babyn, J. Wang, G. Groot, M. Eramian, Thyroid nodule classification in ultrasound images by fine-tuning deep convolutional neural network, J. Digit. Imaging 30 (4) (2017) 477–486, http://dx.doi.org/10.1007/s10278-017-9997-y.

[9]. L. Wang, L. Zhang, M. Zhu, X. Qi, Z. Yi, Automatic diagnosis for thyroid nodules in ultrasound images by deep neural networks, Med. Image Anal. 61 (2020) 101665, http://dx.doi.org/10.1016/j.media.2020.101665.

[10]. J. Ma, F. Wu, J. Zhu, D. Xu, D. Kong, A pre-trained convolutional neural network based method for thyroid nodule diagnosis, Ultrasonics 73 (2017) 221–230, http://dx.doi.org/10.1016/j.ultras.2016.09.011.

[11]. O. Moussa, H. Khachnaoui, R. Guetari, N. Khlifa, Thyroid nodules classification and diagnosis in ultrasound images using fine-tuning deep convolutional neural network, Int. J. Imaging Syst. Technol. 30 (1) (2020) 185–195, http://dx.doi.org/10.1002/ima.22363.

[12]. W. Song, S. Li, J. Liu, H. Qin, B. Zhang, S. Zhang, A. Hao, Multitask cascade convolution neural networks for automatic thyroid nodule detection and recognition, IEEE J. Biomed. Health Inf. 23 (3) (2019) 1215–1224, http://dx.doi.org/10.1109/JBHI.2018.2852718.

[13]. T. Liu, Q. Guo, C. Lian, X. Ren, S. Liang, J. Yu, L. Niu, W. Sun, D. Shen, Automated detection and classification of thyroid nodules in ultrasound images using clinical-knowledge-guided convolutional neural networks, Med. Image Anal. 58 (2019) 101555, http://dx.doi.org/10.1016/j.media.2019.101555.

[14]. G. Shi, J. Wang, Y. Qiang, X. Yang, J. Zhao, R. Hao, W. Yang, Q. Du, N.G.-F. Kazihise, Knowledge-guided synthetic medical image adversarial augmentation for ultrasonography thyroid nodule classification, Comput. Methods Programs Biomed. 196 (2020) 105611, http://dx.doi.org/10.1016/j.cmpb.2020.105611.

[15]. P. Wan, F. Chen, C. Liu, W. Kong, D. Zhang, Hierarchical temporal attention network for thyroid nodule recognition using dynamic CEUS imaging, IEEE Trans. Med. Imaging 40 (6) (2021) 1646–1660, http://dx.doi.org/10.1109/TMI.2021.3063421.

[16]. Y. Chen, D. Li, X. Zhang, J. Jin, Y. Shen, Computer aided diagnosis of thyroid nodules based on the devised small-datasets multi-view ensemble learning, Med. Image Anal. 67 (2021) 101819, http://dx.doi.org/10.1016/j.media.2020.101819.

[17]. X. He, E.-L. Tan, H. Bi, X. Zhang, S. Zhao, B. Lei, Fully transformer network for skin lesion analysis, Med. Image Anal. 77 (2022) 102357, http://dx.doi.org/10.1016/j.media.2022.102357.

[18]. O. Dalmaz, M. Yurt, T. Cukur, ResViT: Residual vision transformers for multi-modal medical image synthesis, 2021, arXiv.

[19]. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS '17, Curran Associates Inc., Red Hook, NY, USA, 2017, pp. 6000–6010.