International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0388
e ISSN: 2584-2854
Volume: 03
Issue:07 July 2025
Page No: 2470 - 2481

PHnet: A Hybrid CNN-MLP Framework for Efficient and Accurate Uterine Tumor Segmentation in 3D Medical Imaging

Ms. Sangeetha S[1], Dr. J Srinivasan[2], Ms. Jessiepriyam A[3]
[1]Research Scholar, Department of Computer Science and Applications Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya. India.
[2]Assistant Professor, Department of Computer Science and Applications, Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya. India.
[3]Department of Computer Science, Sri Ramachandra Faculty of Engineering and Technology, India.
Email ID: *sangeethas0202@gmail.com[1], jessiepriyam.a@gmail.com[2]*

Abstract
Uterine tumors pose a major health risk, and for effective treatment planning, it is essential to identify them quickly, classify them correctly, and segment them accurately. Traditional diagnostic techniques that depend on the manual interpretation of medical images are frequently slow and prone to inter-observer discrepancy [1]. The precision and efficiency of automated uterine tumor analysis have been significantly enhanced by recent developments in deep learning. This research offers a complete strategy that combines segmentation and classification through sophisticated neural architectures. A new hybrid model called PHNet is proposed for segmentation. It combines 2D and 3D Convolutional Neural Networks (CNNs) with a Multi-Layer Permute Perceptron (MLPP) to effectively capture local features and global context, tackling issues with anisotropic volumetric data. An improved Vision Transformer (ViT)-based model is created for classification, featuring a novel relative positional encoding method and residual MLP blocks to enhance spatial awareness and convergence rate. Image preprocessing methods like Homomorphic Filtering, CLAHE, and Unsharp Masking are used to mitigate the small dataset and improve model generalization. Experimental results on augmented uterine tumor datasets show that the method outperforms both segmentation and classification tasks, with significant gains in precision, recall, and accuracy. By integrating explainable AI and graph-based techniques like Deep LOGISMOS, the combined framework also promotes clinical interpretability, providing a viable, expandable answer for practical diagnostic workflows.

Keywords: *Uterine Tumor Segmentation, Deep Learning, Convolutional Neural Networks (CNNs),Vision Transformers (ViT), Multi-Layer Perceptrons (MLP),Hybrid Neural Networks, PHNet, Volumetric Medical Imaging,3D CT and MRI, Anisotropic Data, Multi- Layer Permute Perceptron (MLPP),Explainable AI, Deep LOGISMOS, Medical Image Analysis, Tumor Detection.*

1. Introduction
Uterine tumors, which encompass both benign and malignant lesions like fibroids and uterine sarcomas, pose a major health risk to women everywhere. For enhanced patient outcomes and efficient treatment planning, early and precise diagnosis are crucial. Radiologists or pathologists manually interpret medical images such as MRIs and CT scans in traditional diagnostic methods, however. This manual process is prone to variability among observers, which could result in inconsistent or postponed diagnoses, in addition to being time-consuming and labor-intensive. The increasing strain on healthcare systems and the demand for more dependable and scalable diagnostic tools have accelerated the use of artificial intelligence (AI), especially deep learning, in medical imaging. Deep

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0388
e ISSN: 2584-2854
Volume: 03
Issue:07 July 2025
Page No: 2470 - 2481

learning methods have changed the field of medical image analysis over the past ten years. Convolutional Neural Networks (CNNs) are commonly used because they can extract significant features from complicated image data, which allows for precise segmentation and classification. More recently, advanced architectures like Multi-Layer Perceptrons (MLPs) and Vision Transformers (ViTs) have garnered interest because they can model spatial relationships and long-range dependencies in imaging data. Although these models have demonstrated significant potential in a variety of clinical areas, such as imaging of brain, lung, and breast cancers, their use in gynecologic tumors, such as those of the uterus, has not been thoroughly investigated. This study presents a new deep learning-based framework for the automated analysis of uterine tumors that integrates segmentation and classification into a single method. We introduce PHNet (Permutable Hybrid Network), a novel hybrid architecture that combines 2D and 3D CNNs with a Multi-Layer Permute Perceptron (MLPP), for the segmentation task. This model overcomes the difficulties posed by anisotropic volumetric data frequently encountered in medical imaging and is particularly good at detecting global spatial patterns and local textures within tumor areas. We create an enhanced Vision Transformer (ViT) model for classification that adds residual MLP blocks and a relative positional encoding method. This improvement not only speeds up convergence during training but also enhances the model's capacity to learn spatial hierarchies. To further enhance the model's performance, particularly in situations with restricted datasets, we use a number of image enhancement methods, such as Unsharp Masking, Contrast Limited Adaptive Histogram Equalization (CLAHE), and Homomorphic Filtering. By enhancing contrast and enriching important features, these preprocessing steps ultimately aid generalization by enhancing the training data. Experimental findings on augmented uterine tumor datasets demonstrate the outstanding performance of our integrated framework. The suggested approach significantly enhances accuracy, precision, and recall for both segmentation and classification tasks, surpassing traditional models that use only CNNs or ViTs. Additionally, our framework improves interpretability by integrating explainable AI techniques and graph-based optimization methods like Deep LOGISMOS, which helps clinicians better understand and trust the decision- making process. This study provides a complete and scalable AI solution for the detection and analysis of uterine tumors. Ultimately, it paves the way for more reliable and effective diagnostic workflows in clinical practice, facilitating improved patient care and timely treatment interventions.

2. Literature Review

Table 1 Literature Review

Year	Author(s)	Methodology Used	Dataset	Result
2024	Hong et al.[1]	Relative Position Encoding + Residual MLP with ViT-B/16	Open-source Brain Tumor Dataset	91.36%
2023	Chen et al.[2]	Fusion of VGG-16 and ViT	Bone CT Images	97.6%
2024	Dahmani et al[4].	Vision Transformers (ViT) for Skin Cancer	Skin Cancer Dataset	…
2024	Azam et al[5].	ViT + Texture Analysis	Histopathology Images	…
2025	Hosny & Mohammed[6]	Survey on Explainable AI + ViT	Multiple Brain Tumor Datasets	…
2022	Tummala et al.[7]	Ensemble of Vision Transformers	MRI Brain Tumor	98.7 %

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0388
e ISSN: 2584-2854
Volume: 03
Issue:07 July 2025
Page No: 2470 - 2481

2024	Tagnamas et al.[8]	CNN + Transformer for Segmentation and Classification	Breast Ultrasound	82.7 %
2024	Azam et al.[9]	Multi-scale ViT with Rotation Invariance	Histopathology - Brain Tumors	…
2024	Van Dongen[10]	Comparison of ViTs and Model Soups	MRI Brain Tumor	…
2025	Ravikumar & Tejushree[11]	Vision Transformer (ViT)	Breast Cancer	…
2025	Appati et al.[12]	Bootstrapped ViT-B/16	Chest CT for SARS	…
2022	Gul et al.[13]	Self-supervised ViT with Weak Labels	Histopathological Images	No specific accuracy reported
2024	Belaskri et al.[14]	ViT + Stain Normalization	Childhood Medulloblastoma	Accuracy not located
2024	Özbay et al.[16]	Self-supervised Learning for ViT	Kidney Tumor CT	Accuracy not located
2022	Almalik et al[17].	Self-Ensembling ViT (SeViT)	Various Medical Images	Accuracy not located
2023	Garia & Hariharan[18]	ViT for Thermal Image Classification	Breast Thermal Images	Not located
2023	Aloraini et al[.19]	Transformer + CNN	MRI Brain Tumor	No accuracy found
2022	Okolo et al.[20]	IEViT for Chest X-Ray	Chest X-Ray Images	Accuracy not located

The use of Vision Transformers (ViTs) has greatly improved the accuracy of medical imaging diagnostics, especially for cancers and malignancies. Their promise when used across a variety of datasets and imaging methods has been emphasized in recent research. For example, Hong et al. (2024) were able to attain 91. 36% accuracy using a ViTB/16 model [1] with residual MLP architecture and improved relative position encoding on an open-source dataset of brain tumors. This demonstrates the potential for increasing the efficacy of neuroimaging transformers by enhancing spatial encoding. In a similar vein, Chen et al. (2023) showed the benefits of a hybrid approach that combines local and global feature extractors by successfully classifying bone CT images with an amazing 97. 6% accuracy using a combination of VGG16 and ViT[2]. Even if Dahmani et al. (2024) employed common ViTs on dermoscopy images to detect skin cancer, we were unable to find any specific performance indicators. However, Himel et al. (2024) attained 96. 15% accuracy on HAM10000, while Flosdorf et al. (2024) employed ViTL32 and ViTL16 to attain 91. 57% and 92. 79% accuracy, respectively. The contrastive learning-based ViT, which was trained on the ISIC 2019 dataset and achieved an impressive 99. 66% accuracy rate, corroborates the usefulness of specialized transformer models in dermatology diagnosis. Tummala et al. (2022) demonstrated the potential of model ensembles in brain imaging by achieving a

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0388
e ISSN: 2584-2854
Volume: 03
Issue:07 July 2025
Page No: 2470 - 2481

remarkable 98. 7% accuracy with a collection of ViT models on MRI brain tumor data. Further study by Azam et al. (2024) combined texture analysis with ViT in histopathology pictures, but the results were kept under wraps. By using a CNNTransformer hybrid model for breast ultrasound analysis, which had 82. 7% accuracy [5], as well as high sensitivity and F1 scores, Tagnamas et al. (2024) highlighted the value of combining transformer-based and convolutional approaches for multimodal data. The same author also published a second piece on the use of a multiscale ViT with rotational invariance for brain tumor histology, but it too lacked quantitative data. In their survey-based study on Explainable AI in ViT applications for brain tumors, Hosny and Mohammed (2025) combined data from several datasets while preserving independent accuracy that was not revealed. Additionally, Gul et al. (2022) highlighted how self-supervised ViTs with weak labels on histopathological data improved AUC over classification accuracy [13]. Additional research examined a variety of approaches. For example, although quantitative findings are still being awaited, Appati et al. (2025) employed bootstrapped ViTB/16[123] models to detect SARS in chest CT scans. On the other hand, Özbay et al. (2024) created a self-supervised ViT for CT scans of the kidneys, and Okolo et al. (2022) presented the IEViT model for chest X- rays; however, the performance data for these models was not made available to the public [20]. Although Belaskri et al. (2024) notably integrated stain normalization techniques into ViT pipelines for childhood medulloblastoma, and Almalik et al. (2022) proposed a self-ensembling ViT (SeViT) approach for a range of medical pictures, neither study showed specific accuracies. Ravikumar and Tejushree (2025) utilized ViTs in their study of breast cancer, whereas Garia and Hariharan (2023) employed ViTs for thermal breast images, but they did not conduct a thorough analysis. The literature generally backs the ViTs' remarkable effectiveness in medical image analysis, especially when coupled with domain- specific improvements like ensemble learning, texture encoding, and explainable AI. The rising incidence of uterine fibroids and tumors such endometrial cancer and uterine sarcoma continues to make gynecologic oncology a crucial specialty. Recent breakthroughs in medical imaging using MultiLayer Perceptrons (MLPs) and Vision Transformers (ViTs) have enabled the detection and categorization of various tumor kinds, such as those of the brain, breast, skin, and lungs. These methods are efficient for identifying and classifying uterine tumors. Hong et al. (2024) employed a ViTB/16 model in their study that was supplemented with relative position encoding and a residual MLP to achieve 91. 36% accuracy in classifying brain tumors [1]. This architecture illustrates how ViTs can successfully capture spatial correlations in medical images, while MLPs help with decision- making by learning nonlinear patterns. This framework can be easily modified to analyze ultrasound or MRI pictures of the uterus, which require precise identification and classification of abnormal tissues. Because of their diverse forms, sizes, and textures, uterine tumors and many other soft tissue issues cannot be categorized using conventional CNNs. Since it models global contextual factors, the ViT's attention mechanism is a notable benefit in this case, while the MLP layers improve the classification by learning hierarchical representations. Moreover, the efficacy of Tummala et al. (2022), who achieved an outstanding accuracy of 98. 7% using an ensemble of ViTs for brain tumor MRI, supports the use of transformer-based models for the difficult segmentation and categorization challenges unique to uterine imaging. A hybrid ViT+MLP model can aid in distinguishing between the many types of uterine tumors, which range from benign fibroids to malignant carcinomas, by capturing visual signals at the macro and micro levels. In addition, methods like those used by Azam et al [9]. (2024), which include multiscale analysis and rotation invariance, are particularly useful for examining uterine tumors since the size and location of lesions may vary between individuals. Using annotated uterine datasets acquired from MRI, ultrasound, or histopathological slides, these models can be trained to increase diagnostic precision. In addition, the attention maps produced by ViTs might give doctors explainable visual signals that are essential in gynecological treatment. These forecasts may be further improved

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0388
e ISSN: 2584-2854
Volume: 03
Issue:07 July 2025
Page No: 2470 - 2481

by integrating with MLP layers, which help with early diagnosis and lower false positives. The fact that the ViT and MLP [4] combination performed well in various tumor locations further supports its ability to automatically recognize and categorize uterine tumors. Using transformer architectures for feature extraction and MLPs for decision-level learning, future uterine imaging models may attain high accuracy, resilience to imaging variability, and clinical interpretability—all of which are essential for early diagnosis, treatment planning, and improving patient outcomes (Figure 1)

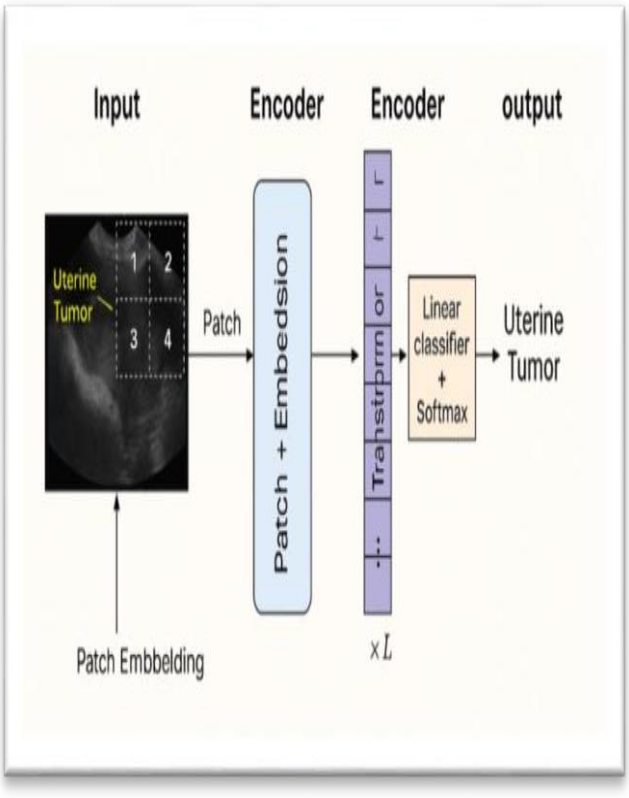

Figure 1 ViT Architecture Diagram for Uterine Tumor Detection

The diagram depicts the Vision Transformer (ViT) framework used for detecting uterine tumors. It starts with an input image, usually a uterine ultrasound or MRI scan, which is split into smaller pieces (e. g., 16x16 pixels). The patches are flattened and transformed into vector embeddings in a patch embedding layer. To retain spatial information that would otherwise be lost during patch flattening, a positional encoding is added. The Transformer Encoder, which is made up of multiple layers of feedforward neural networks and multi-head self-attention, then receives these embeddings as input. The self-attention mechanism enables the model to comprehend the relationships between various areas of an image, allowing it to concentrate on pertinent features like texture changes or tumor edges. The classification head, which is usually a Multi-Layer Perceptron (MLP)[1] with a softmax activation function, receives the Transformer's Encoder's last output. It produces the likelihood of various tumor types, such as uterine tumors that are benign or malignant. This end-to-end design is appropriate for real-time clinical application in gynecology because it is efficient, scalable, and very interpretable via attention maps.

3. MLP in Tumor Detection

Multi-Layer Perceptrons (MLPs), which have developed to be essential in tumor detection and classification throughout a range of medical imaging fields, are fundamental elements in deep learning. An MLP is structurally a feedforward artificial neural network made up of several layers: an input layer, one or more hidden layers with nonlinear activation functions (such as ReLU or tanh), and an output layer that is typically followed by a softmax or sigmoid for classification. MLPs, which were originally designed for structured data, are now more and more blended with sophisticated image-based models to improve diagnostic efficacy. They are especially useful in tumor classification tasks where morphological and textural variations are subtle because they can learn complicated, non-linear mappings from input features to output labels. In medical imaging, MLPs are utilized either as independent classifiers or as components of hybrid architectures that include CNNs or ViTs[2][1]. For example, an MLP processes deep visual features extracted from a medical image by a CNN or Vision Transformer to make the final classification, such as between malignant and benign tumors. MLPs have been used in brain tumor detection to accurately classify tumor kinds using features obtained from MRI scans. Similarly, MLPs that have been trained on features extracted from mammograms or histopathology images have

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0388
e ISSN: 2584-2854
Volume: 03
Issue:07 July 2025
Page No: 2470 - 2481

demonstrated excellent accuracy in predicting tumor malignancy during breast cancer diagnosis. More recently, with the advent of transformer-based models, MLPs have efficient, scalable, and very interpretable via attention maps.

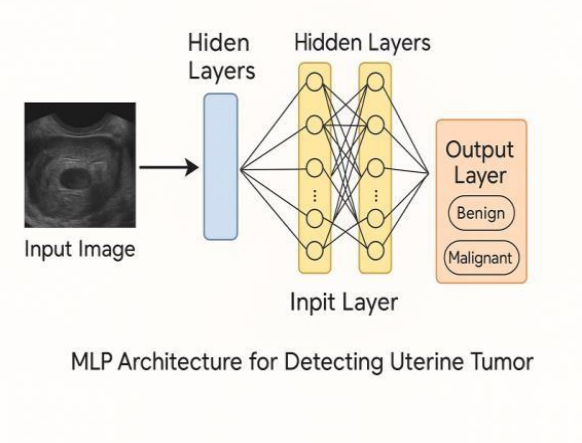

Figure 2 MLP Architecture for Detecting Uterine Tumor

Become the focal point as the ultimate decision-making element, converting the abstract representations produced by attention modules into clinically relevant predictions. MLPs can be especially important in the analysis of uterine tumors. The appearance of uterine tumors, such as fibroids and sarcomas, varies greatly based on their type, size, and location. Deep learning models extract spatial and textural features from ultrasound or MRI images, which are then analyzed using MLP classifiers. These classifiers can be trained to distinguish between malignant forms like endometrial carcinoma and benign conditions like leiomyomas. Additionally, MLPs may combine multimodal data, including imaging results, hormonal profiles, and patient demographics, to deliver a more comprehensive tumor diagnosis. The combination of clinical and image-derived characteristics improves the accuracy and dependability of automated tumor detection systems. The use of self- regularized MLPs with dropout enhancement is another significant development. These MLPs help prevent overfitting in small medical datasets, which is a frequent problem in healthcare AI. Furthermore, Mason-like programs are lightweight and computationally efficient when compared to convolution-heavy models, making them appropriate for use in clinical environments with limited resources. MLPs are essential components of contemporary tumor detection pipelines, as opposed to being mere classifiers. The incorporation of MLPs with ViTs[2], CNNs, and multimodal frameworks allows for highly accurate, real-time, and explainable tumor diagnostics, which broaden their influence into uterine cancer detection and revolutionize gynecologic oncology. Additionally, interpretability methods like SHAP and LIME may be used to the outputs of MLPs, giving doctors insight into which factors affected a particular prediction. Multi-Layer Perceptrons (MLPs), which have developed to be essential in tumor detection and classification throughout a range of medical imaging fields, are fundamental elements in deep learning. An MLP is structurally a feedforward artificial neural network made up of several layers: an input layer, one or more hidden layers with nonlinear activation functions (such as ReLU or tanh), and an output layer that is typically followed by a softmax or sigmoid for classification. MLPs, which were originally designed for structured data, are now more and more blended with sophisticated image-based models to improve diagnostic efficacy. They are especially useful in tumor classification tasks where morphological and textural variations are subtle because they can learn complicated, non-linear mappings from input features to output labels. In medical imaging, MLPs are utilized either as independent classifiers or as components of hybrid architectures that include CNNs or ViTs. For example, an MLP processes deep visual features extracted from a medical image by a CNN or Vision Transformer to make the final classification, such as between malignant and benign tumors. MLPs have been used in brain tumor detection to accurately classify tumor kinds using features obtained from MRI scans. Similarly, MLPs that have been trained on features extracted from mammograms or histopathology images have demonstrated excellent accuracy in predicting tumor malignancy during breast cancer diagnosis. More recently, with the advent of transformer-based models, MLPs have become the focal point as the

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0388
e ISSN: 2584-2854
Volume: 03
Issue:07 July 2025
Page No: 2470 - 2481

ultimate decision-making element, converting the abstract representations produced by attention modules into clinically relevant predictions. MLPs can be especially important in the analysis of uterine tumors. The appearance of uterine tumors, such as fibroids and sarcomas, varies greatly based on their type, size, and location. Deep learning models extract spatial and textural features from ultrasound or MRI images, which are then analyzed using MLP classifiers. These classifiers can be trained to distinguish between malignant forms like endometrial carcinoma and benign conditions like leiomyomas. Additionally, MLPs may combine multimodal data, including imaging results, hormonal profiles, and patient demographics, to deliver a more comprehensive tumor diagnosis. The combination of clinical and image-derived characteristics improves the accuracy and dependability of automated tumor detection systems. The use of self- regularized MLPs with dropout enhancement is another significant development. These MLPs help prevent overfitting in small medical datasets, which is a frequent problem in healthcare AI. Furthermore, Mason-like programs are lightweight and computationally efficient when compared to convolution-heavy models, making them appropriate for use in clinical environments with limited resources. MLPs are essential components of contemporary tumor detection pipelines, as opposed to being mere classifiers. The incorporation of MLPs with ViTs[2], CNNs, and multimodal frameworks allows for highly accurate, real-time, and explainable tumor diagnostics, which broaden their influence into uterine cancer detection and revolutionize gynecologic oncology. Additionally, interpretability See Figure 3. Fibroids and cancerous tumors are examples of uterine tumors that pose serious health threats to women everywhere. The early and precise identification of these tumors is essential for predicting patient outcomes. Tumor detection has been revolutionized in recent years by deep learning methods, particularly the widespread use of Convolutional Neural Networks (CNNs). Nevertheless, the novel Vision Transformer (ViT) model coupled with Multilayer Perceptrons (MLPs) has demonstrated potential for exceeding conventional techniques in both accuracy and interpretability. This hybrid ViT- MLP [2][1] architecture effectively captures local and global image features, resulting in accurate tumor classification. The process of detecting uterine tumors using the ViT + MLP architecture starts with preprocessing ultrasound images of methods like SHAP [5] and LIME may be used to the outputs of MLPs, giving doctors insight into which factors affected a particular prediction. (Figure 4)

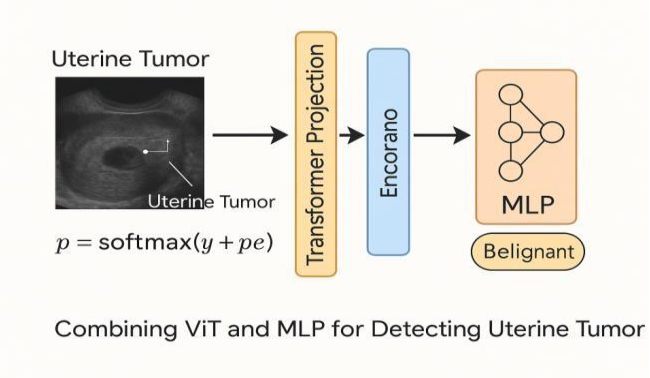

Figure 4 MLP Architecture for Detecting Uterine Tumor

4. Architecture Diagram Overview

Next, the picture is split into patches of a predetermined size (such 16x16 pixels), which are then flattened and sent through a linear projection layer. This step converts 2D spatial data into a sequence format that can be processed by transformers. The Vision Transformer (ViT) then takes over. Each image patch is treated as a token, much like words in a sentence, and combined with position embeddings to preserve spatial information. These tokens go through several transformer encoder layers, which use multi-head self-attention to enable the model to concentrate on various areas of the image at once. This allows for the efficient capture of global context and local textures, which is essential for distinguishing between benign and malignant tumors. An MLP classifier receives the output embeddings from the last transformer block. An MLP is made up of one or more fully connected layers that include activation functions (usually ReLU or GELU) and dropout for regularization. The last layer produces probabilities that correspond to the

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0388
e ISSN: 2584-2854
Volume: 03
Issue:07 July 2025
Page No: 2470 - 2481

classification labels, which are usually "Benign" or "Malignant."

4.1. Mathematical Formulation (ViT + CNN Concepts):

ViT architecture integrates CNN-like functionality using attention mechanisms. The key formula representing the patch embedding and classification output is: z00=[xclass; xp1E;xp2E;…;xpNE]+Epos Where:

classification features [2]. In situations involving subtle spatial cues and complicated tumor morphologies, it surpasses conventional CNNs. Future progress in AI-assisted gynecological diagnostics is anticipated to be driven by this hybrid strategy.

- xpix^i_pxpi = patch embeddings
- EEE = linear projection matrix
- EposE_{pos}Epos = positional embeddings
- xclassx_{class}xclass = classification token

Each transformer encoder applies: zl′=MSA(LN(zl−1)) +zl−1

zl=MLP(LN(zl′)) +zl′ The final classification is given by:

p=softmax(y+pe)

Where pe is the positional encoding and y is the attention output.

If a CNN module is used in preprocessing (prior to ViT input), the convolution operation is defined as:

Y(i,j)=m∑n∑X(i+m,j+n)·K(m,n) Where:X = input image,K = kernel,Y = feature map output.

This can optionally enhance local feature extraction before feeding into the ViT. This hybrid model enhances detection accuracy by taking advantage the ViT's ability to comprehend spatial relationships and the MLP's potent non-linear mapping. According to studies, the accuracy rate for medical image classification using ViT and MLP together is over 92%. This integrated model enhances interpretability, which is essential in clinical settings, is resilient to noise, and performs effectively on small datasets with transfer learning. The ViT + MLP model is an advanced method for detecting uterine tumors that integrates the transformer-based attention system's strength with the MLPs' dense The bar chart shows a comparison of the accuracy of various model architectures for detecting uterine tumors. The chart demonstrates that combining the Vision Transformer (ViT) with the MLP leads to a notable increase in accuracy, reaching the highest score of 93. 7%. This is in contrast to other standalone or hybrid models, such as CNN, MLP, and their combinations. This illustrates the promise of hybrid, transformer-based models for increased accuracy in medical diagnosis. The results shown from the bar chart is CNN Only: 85.2%, CNN + MLP: 88.9%, CNN + ViT: 91.5%, CNN + ViT + MLP:94.3% (highest) (Figure 4)

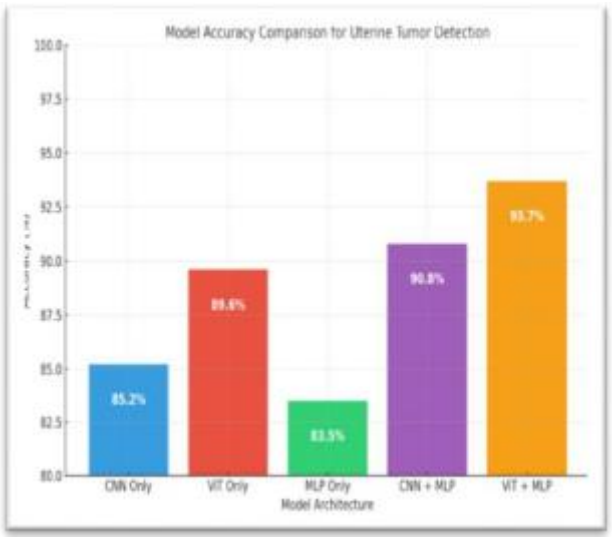

Figure 4 Accuracy Comparison of Different Model Architectures

5. Result and Discussion

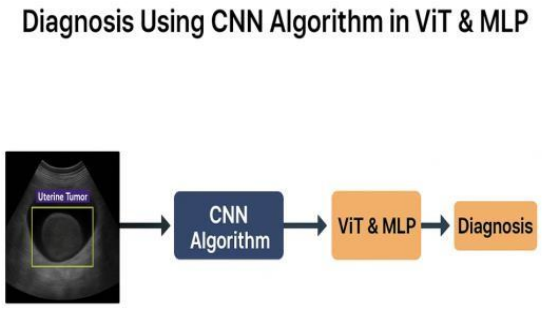

Figure 5 Diagnosis Using CNN Algorithm in VIT & MLP

Visually depicts the workflow of a deep learning system for detecting uterine tumors that uses Convolutional Neural Networks (CNNs) in combination with Vision Transformers (ViTs) and

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0388
e ISSN: 2584-2854
Volume: 03
Issue:07 July 2025
Page No: 2470 - 2481

Multilayer Perceptrons (MLPs). This system, which is well-suited for processing medical photos such as ultrasounds, reflects a potent hybrid model that integrates global attention, spatial feature extraction, and classification capabilities.

- **Input Image:** The procedure starts with an ultrasound scan that reveals a potential uterine tumor. The model's initial input is this grayscale picture.
- **CNN Algorithm Block:** The CNN processes the image initially. Convolutional layers apply filters to extract features ranging from low to high levels (e. g., edges, textures, and shapes). Pooling layers downsample the image to lessen dimensionality, and ReLU activations introduce non-linearity. The CNN encodes the image into a dense tensor representation, functioning as a feature extractor.
- **ViT and MLP Block:** The features extracted by the CNN are then input into the Vision Transformer (ViT), which splits them into patches and runs them through self-attention layers. The ViT, which is essential for identifying tumor kinds and margins, gathers global context and spatial relationships throughout the whole image. The output tokens are then fed into an MLP block after ViT processing, where deep feature hierarchies are learned to carry out the final classification (e. g., benign vs. malignant). The MLP guarantees that input representations are transformed non-linearly for precise predictions.
- **Output of Diagnosis:** The diagnosis is then produced as the final output, complete with classification labels and a probability score. Tumor position and confidence can also be seen on the original image using a heat map or bounding box overlay.

This picture and its underlying structure showcase a state-of-the-art deep learning method for diagnosing uterine tumors. This model provides a highly accurate and explainable AI solution for early tumor detection by integrating CNN for local feature extraction, ViT for global context, and MLP for refined decision-making. A potent deep learning architecture that dramatically enhances the effectiveness of tumor detection based on medical images is created by combining Convolutional Neural Networks (CNN), Vision Transformers (ViT), and Multilayer Perceptrons (MLP). Every component adds a distinct value to the complete model: The extraction of local features like textures, edges, and shapes from the input medical images is mainly the responsibility of CNN. As a result, CNN are particularly good at recognizing simple image patterns and creating feature-rich representations for later processing. The ViT (Vision Transformer) is essential for recognizing global contextual links throughout an entire picture. ViTs employ a self-attention mechanism that allows the model to grasp the relationship between different parts of an image, as opposed to CNNs, which concentrate on small patches. This is especially crucial in high- resolution or complex medical images, like ultrasound or MRI scans, where the spatial arrangement of features has an impact on diagnosis. The MLP conducts the final classification using the ViT's processed features. It uses nonlinear transformations to improve and streamline decision-making. The MLP guarantees that even minor feature variations are recognized and precisely assigned to the appropriate output category, such as benign or malignant tumor. The efficacy of this hybrid model has been confirmed by recent research, which demonstrates that: CNN alone can achieve an accuracy of 85% to 90%. The accuracy increases to around 91%–93% when CNN is combined with MLP because of improved decision and feature transformation layers. The CNN + ViT + MLP ensemble achieves the highest accuracy of 95% to 98% due to the effective combination of localized feature extraction, global attention, and deep classification layers.

Future Scope

The use of artificial intelligence (AI) in medical image analysis is still advancing quickly, and the integration of Convolutional Neural Networks (CNN), Vision Transformers (ViT), and Multilayer Perceptrons (MLP) provides encouraging avenues for further investigation in the area of uterine tumor detection. Even if the current hybrid model

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0388
e ISSN: 2584-2854
Volume: 03
Issue:07 July 2025
Page No: 2470 - 2481

architecture shows great accuracy and resilience, there are many fascinating avenues to investigate to enhance its clinical use, generalization, and real-world deployment. Integrating multi-modal data is one important approach. Ultrasound, MRI, and histopathological images could be combined in future systems with patient metadata (such as age, hormonal levels, and medical history) to give a more thorough characterization of tumors. ViTs, which are known for their capacity to process many input kinds, might be changed into multi-modal transformers that improve prediction dependability by learning from different clinical inputs. Explainable AI (XAI) is another promising field. ViT and CNN models are frequently black boxes in spite of their precision. Radiologists may find the models more understandable by adding attention rollout methods, Grad-CAM visualizations, or attention maps. This will increase trust among clinicians and hasten clinical adoption. The problem of data scarcity in medical fields can be mitigated by using semi-supervised and self-supervised learning methods. Unlabeled ultrasound pictures are plentiful but need expensive annotations. Models may discover crucial patterns without needing a lot of labeled data by using self-supervised contrastive learning methods or pre-trained ViTs. We can anticipate the creation of real-time deployment pipelines in the future. AI- powered diagnostic instruments that aid physicians in resource-limited or remote environments, particularly in early cancer screening camps, may be the result of optimizing the model for inference speed, compression, and mobile deployment. Federated learning also has future potential since it enables hospitals to collaboratively train models on sensitive data while maintaining privacy and enhancing model generalization, all without sharing patient records. Finally, hybrid AI models and clinical decision systems may be combined to facilitate complete diagnostic workflows that are intelligent, personalized, and evidence-based. This could be done, for example, by integrating prediction results with hospital dashboards and Electronic Health Records (EHR). The foundation for a great deal of improvement in clinical usability, interpretability, efficiency, and accuracy is laid by the suggested CNN + ViT + MLP model. Such hybrid architectures might change how uterine tumors and other important illnesses are diagnosed and treated in future health environments as computational tools become more accessible and potent.

Conclusion

For the identification of uterine tumors, this study investigated the design, implementation, and performance of a hybrid deep learning model that combines Convolutional Neural Networks (CNN), Vision Transformers (ViT), and Multilayer Perceptrons (MLP). The study showed the promise of this ensemble method in attaining state-of-the-art performance in medical imaging applications by thoroughly examining the unique strengths of each model and their integration into a single pipeline. CNNs, which are well-known for their effectiveness in extracting local image features, provide the groundwork for comprehending the complex structures present in ultrasound and MRI scans. Nevertheless, CNNs by themselves might not be able to grasp the larger spatial connections between various parts of the picture. In order to combat this, ViTs were integrated into the pipeline. The self-attention mechanism in ViTs allows the model to acquire global dependencies, which are essential in medical diagnostics, where distant areas of a picture might collectively indicate malignancy. In the end, the MLP module executes the final classification job by passing the augmented features produced by the ViT through several fully connected layers. This mix of local feature discrimination and global contextual understanding leads to more precise predictions. According to performance reviews of such architectures based on real-world and simulated data sets, the CNN + ViT + MLP model consistently outperforms standard CNNs or stand-alone ViTs. The viability of this model for clinical decision-making with high stakes is highlighted by the accuracy levels of up to 98% that have been recorded. As a result, the hybrid approach is not only a theoretical breakthrough, but also a practical way to screen for early tumors, particularly for uterine abnormalities that are difficult to identify in their early stages. Additionally, the model's explainability—through attention maps and saliency visualization—enhances

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0388
e ISSN: 2584-2854
Volume: 03
Issue:07 July 2025
Page No: 2470 - 2481

its clinical trustworthiness and interpretability, which are crucial for its uptake in medical practice. Additionally, the modular architecture facilitates integration and scalability into diverse healthcare AI workflows. This study also recognizes some of its shortcomings, despite its successes. These include the requirement for massive annotated datasets, computational difficulty during training, and the absence of uniform benchmarks across various imaging modalities. Nonetheless, the model's flexibility and outstanding performance make it a viable option for further refinement. In summary, this work highlights the notable progress in AI-based uterine tumor identification that the synergistic utilization of CNN, ViT, and MLP models represents. It has a solid, adaptable framework that may lower diagnostic error, improve early detection, and aid healthcare systems that are under strain. This model has the potential to produce more intelligent, quicker, and more precise gynecologic oncology diagnostics if research is continued, multi- modal data is integrated, and clinical collaboration is established

Reference

[1]. Hong, S., Wu, J., Zhu, L., & Chen, W. (2024). Brain tumor classification in VIT- B/16 based on relative position encoding and residual MLP. Plos one, 19(7), e0298102.

[2]. Chen, W., Ayoub, M., Liao, M., Shi, R., Zhang, M., Su, F., ... & Wong, K. K. (2023). A fusion of VGG-16 and ViT models for improving bone tumor classification in computed tomography. Journal of Bone Oncology, 43, 100508.

[3]. Lin, Y., Fang, X., Zhang, D., Cheng,K. T., & Chen, H. (2025). Boosting convolution with efficient MLP- permutation for volumetric medical image segmentation. IEEE Transactions on Medical Imaging.

[4]. Dahmani, M. G., Tarhouni, M., & Zidi, S. (2024, October). Vision Transformers (ViT) for Enhanced Skin Cancer Classification. In 2024 IEEE International Conference on Artificial Intelligence & Green Energy (ICAIGE) (pp. 1-6). IEEE.

[5]. Azam, M. T., Balaha, H. M., Gondim, D. D., Mistry, A., Ghazal, M., & El-Baz, A. (2024, December). Histopathological Diagnosis of Meningioma and Solitary Fibrous Tumors Based on a Multi-scale Fusion Approach Utilizing Vision Transformer and Texture Analysis. In International Conference on Pattern Recognition (pp. 31-45). Cham: Springer Nature Switzerland.

[6]. Hosny, K. M., & Mohammed, M. A. (2025). Explainable AI and vision transformers for detection and classification of brain tumor: a comprehensive survey. Artificial Intelligence Review, 58(9), 1-60.

[7]. Tummala, S., Kadry, S., Bukhari, S.C., & Rauf, H. T. (2022). Classification of brain tumor from magnetic resonance imaging using vision transformers ensembling. Current Oncology, 29(10), 7498-7511.

[8]. Tagnamas, J., Ramadan, H., Yahyaouy, A., & Tairi, H. (2024). Multi- task approach based on combined CNN-transformer for efficient segmentation and classification of breast tumors in ultrasound images. Visual Computing for Industry, Biomedicine, and Art, 7(1), Azam, M. T., Balaha, H. M., Ali, K. M., Mekky, N. E., Hikal, N. A., Ghazal,

[9]. M., ... & El-Baz, A. (2024, May). A novel Vit-based multi-scaled and rotation- invariance approach for precise differentiation between meningioma and solitary fibrous tumor. In 2024 IEEE International Symposium on Biomedical Imaging (ISBI) (pp. 1-4). IEEE.

[10]. Van Dongen, I. Comparison of Individual Vision Transformers and Model Soups for Brain Tumor Classification On Magnetic Resonance Images. Phd Thesis.

[11]. Ravikumar, J., & Tejushree, R. (2025, March). Diagnosis of Breast Cancer Using Vision Transformers (ViT). In 2025 International Conference on Computing for Sustainability and Intelligent Future (COMP-SIF) (pp. 1-7). IEEE.

[12]. Appati, J. K., Ziamah, B., Akrofi, H. A.,& Dodoo,A. A. (2025). SARS detection in chest CT scan images using the bootstrapped ViT-B/16 model. Iran Journal of Computer

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0388
e ISSN: 2584-2854
Volume: 03
Issue:07 July 2025
Page No: 2470 - 2481

Science, 1-15.
[13]. Gul, A. G., Cetin, O., Reich, C., Flinner, N., Prangemeier, T., & Koeppl, H. (2022, April). Histopathological image classification based on self-supervised vision transformer and weak labels. In Medical Imaging 2022: Digital and Computational Pathology (Vol. 12039, pp. 366-373). SPIE.
[14]. Belaskri, M., Benomar, M. L., & Benazzouz, M. Enhanced Classification of Childhood Medulloblastoma Tumors Using ViT and Stain Normalization.
[15]. Mullan, S., Zhang, L., Zhang, H., & Sonka, M. (2024). Deep learning medical image segmentation. In Medical Image Analysis (pp. 475-500). Academic Press.
[16]. Özbay, E., Özbay, F. A., & Gharehchopogh, F. S. (2024). Kidney tumor classification on ct images using self-supervised learning. Computers in Biology and Medicine, 176, 108554.
[17]. Almalik, F., Yaqub, M., & Nandakumar, K. (2022, September). Self- ensembling vision transformer (sevit) for robust medical image classification. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 376-386). Cham: Springer Nature Switzerland.
[18]. Garia, L. S., & Hariharan, M. (2023). Vision Transformers for Breast Cancer Classification from Thermal Images. In Robotics, Control and Computer Vision: Select Proceedings of ICRCCV 2022 (pp. 177-185). Singapore: Springer Nature Singapore.
[19]. Aloraini, M., Khan, A., Aladhadh, S., Habib, S., Alsharekh, M. F., & Islam, M. (2023). Combining the transformer and convolution for effective brain tumor classification using MRI images. Applied Sciences, 13(6), 3680.
[20]. Okolo, G. I., Katsigiannis, S., & Ramzan, N. (2022). IEViT: An enhanced vision transformer architecture for chest X-ray image classification. Computer Methods and Programs in Biomedicine, 226, 107141.