Review: UNet-based Medical Image Segmentation

Rupak Chakraborty[1], *Shashi Kant Gupta*[2],
[1] Postdoctoral Researcher, LINCOLN UNIVERSITY COLLEGE, Malaysia; [2]Adjunct Professor, LINCOLN UNIVERSITY COLLEGE, Malaysia
Email ID pdf.rupak@lincoln.edu.my, rupak.jis@gmail.com

Abstract: Creating a deep learning framework for accurate segmentation of MRI images using robust AI models is the recent trend. It has been experienced that U-Net based models achieved higher segmentation accuracies in this scenario. It has also been surveyed that incorporating more 'skip-connections' to UNet, advanced UNet++ has been developed. Further vision transformer-based architecture has been added with UNet++ models for effective segmentation. Later, hybrid CNN–Transformer architectures like Vision Transformer (ViT) has been surveyed to leverage both local and global image features to enhancing segmentation accuracy. Some drawbacks of ViT-based models include Data Requirements, Deployment Challenges, Fixed Input Size Dependency, Weak Inductive Bias for Locality etc. As a solution of those State Space Model (SSM)-based Vision Mamba (VM) architecture has been proposed. By integrating XAI techniques such as SHAP, Grad-CAM and uncertainty quantification, the model aims to improve interpretability and reliability. Evaluations will be carried out on existing public datasets such as BraTS2020, ISIC2020, cvc clinicdb, Synapse etc. and the measurement metrics will be considered like Dice Similarity Coefficient (DSC), Intersection over Union (IoU), Hausdorff Distance, Pixel-wise Accuracy etc. The outcomes of the proposed VM-UNet++ model will be compared with UNet and UNet++, Swin Unet, Residual CNN etc.
Keywords: UNet; UNet++; ViT; SSM; VM-UNet++.

Introduction

Accurate segmentation of MRI images is a crucial task in neuroimaging and cancer diagnostics. Traditional manual annotation is time-consuming and inaccurate. Recent developments in convolutional neural networks (CNNs) and transformer-based models have undoubtedly improved segmentation performance. However, existing models often lack interpretability and robustness to noise, scanner variability, and limited data conditions. This paper addresses these gaps and proposes a hybrid model combining CNNs for spatial detail and Vision Transformers for global context understanding. The research also explores explainable AI to visualize decision-making regions and uncertainty estimation to quantify model confidence. The evolution of proposed architecture starting from CNN-based U-Net model has been presented in next section.

Related work

This section will start to investigate the work done on medical image segmentation using UNet model.

CNN-based U-Net models

To identify specific parts of medical image segmentation, U-Net architecture has been preferred for decades. In this "U-shaped" architecture, images are divided into different parts for identifying the specific target area. The performance of U-Net over small labelled medical data is promising [1-2]. The architecture consists of three parts: a) Encoder, b) Bottleneck, c) Decoder.

a) **Encoder**: Small 3×3 filters are applied to scan the image and extract its features. Then ReLU activation function has been applied to add non-linearity so that models learn better. Next max pooling (2×2 filters) is applied to shrink the image size but keeping the important information. This section is used for downsampling the image.

b) **Bottleneck**: In this area, images are reduced most but meaningful features are captured for further decoding.

c) **Decoder**: The decoder starts to upsampling the image by combining information from the encoder using "skip connections". This step helps decoder to retrieve the important information from encoder which may lost because of shrinking the image. Finally, 1×1 convoluted segmented outcome has been obtained. Figure 1 shows the model architecture.

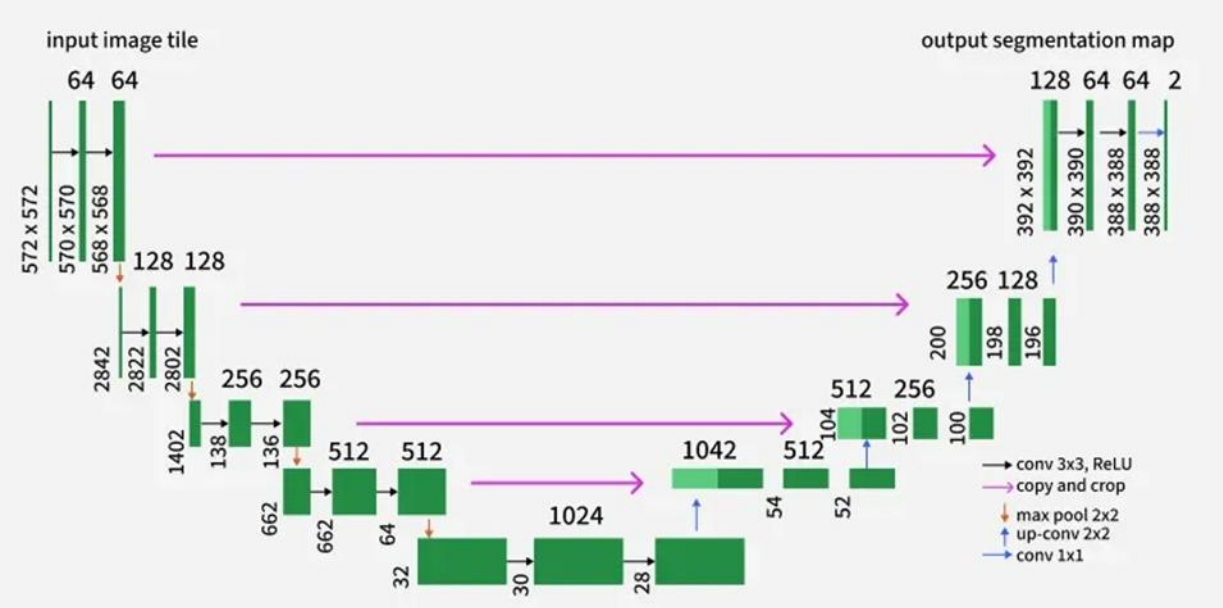

Figure 1. Detailed architecture of UNet model.

Upgradation of existing U-Net model to U-Net++ architecture

UNet++ is a nested U-Net architecture which helps to reduce the semantic gap between encoder and decoder of U-Net. Here nested skip-pathways are used to gain knowledge from both low- and high-level features at decoder end from encoder side which ultimate results in the detailed and comprehensive understanding of the image. Researchers got promising improved results in medical fields after applying U-Net++ architecture over U-Net results. Nested skip connections help to perform deep supervisions of encoder and decoder that help to improve the performance along with overfitting control while training the model [3]. Figure 2 outlines the UNet++ architecture.

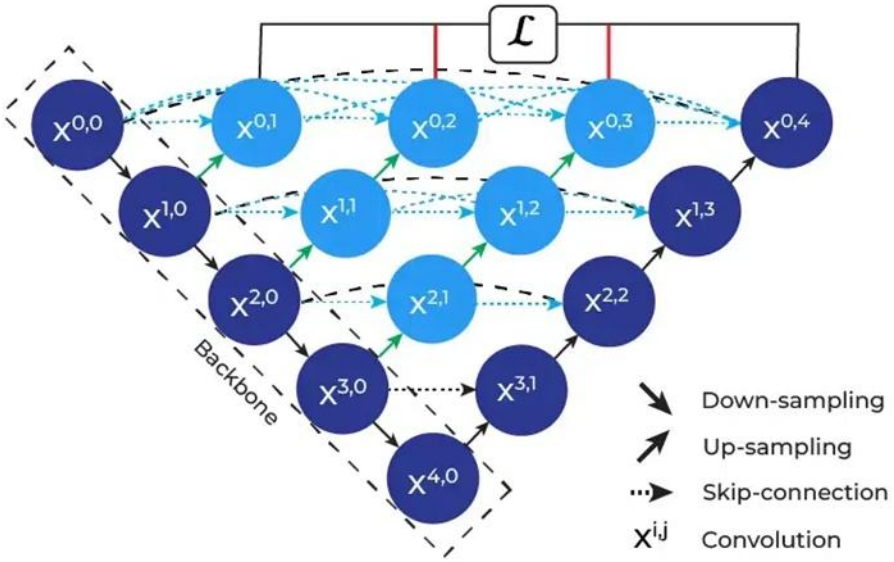

Figure 2. Detailed architecture of improved UNet++ model.

Incorporating Transformer-based architecture to Medical Image Segmentation

After continuous success of CNN-based UNet models in medical image segmentation field, researchers also tried to consider the transformer-based architecture and compared the outcomes. The limitations of U-Net based architecture include susceptibility to class imbalance, sensitivity to image quality variations, and potential for inaccurate boundary delineation. So, Vision Transformer (ViT) architecture was proposed where images are represented as a set of patches which are non-overlapping blocks of image. These blocks are formed by vector embeddings of pixel information. ViT is built on transformer architecture where self-attention mechanisms are applied on patches to form relationships between blocks. This model is pre-trained on large dataset where knowledge is transferred to custom dataset to get good results [4]. The steps included in this architecture are a) Patch Embeddings, b) Self-attention, c) Positional Encoding, d) Scalability.

a) Patch Embeddings: Image is divided into fixed size patches, and those patches are embedded into linear vector.

b) Self-Attention: Self-attention mechanisms capture long-range dependencies across the full image, allowing for effective modeling of spatial relationships.

c) Positional Encoding: Spatial information of images is retained when Positional encodings are added to the patch embeddings.

d) Scalability: the scalability of the model is achieved easily as this architecture is already trained on large datasets like ImageNet21K and transfer the knowledge to enhance accuracy.

Figure 3 details the stepwise architecture.

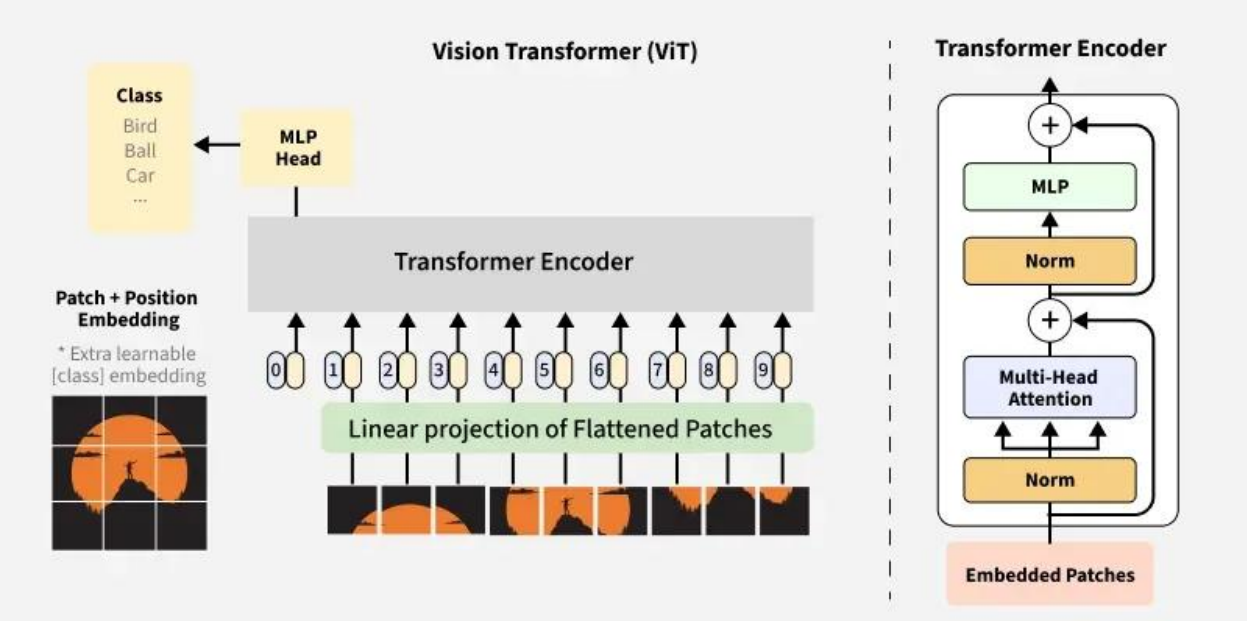

Figure 3. Detailed architecture of ViT model.

Advancing to Vision-Mamba based model

Transformer-based models can capture global information, but computational complexities of those are extremely high and require heavy computing power support. Some drawbacks of ViT-based models include Data Requirements, Deployment Challenges, Fixed Input Size Dependency, Weak Inductive Bias for Locality etc. As a solution of those State Space Model (SSM)-based Vision Mamba (VM) architecture has been proposed. The key features are as given following:

a) **State Space Models (SSMs):** SSMs are used to handle sequences by modeling the hidden states over time. Vision Mamba extends these models to visual data, incorporating bidirectional sequence modeling.

b) **Efficient Representation**: VM is designed to be more computationally efficient than traditional transformers, addressing memory and computation constraints, especially for high-resolution images.

c) **Bidirectional Processing:** Vision Mamba processes token sequences in both backward and forward directions, enhancing the ability of model to capture complex dependencies within the image.

d) **Patch-Based Tokenization**: Like ViTs, Vim splits images into patches and projects these patches into tokens. However, it leverages the bidirectional state space approach for more effective encoding.

The main advantages of vision Mamba over ViT architecture are unlike ViT, Mamba is more computationally efficient, suitable for high-resolution images, perform well with limited computational resources and work as bidirectional sequence modeling. This VM architecture has been combined with UNet architecture and is of recent research interest [5-6].

Figure 4 has shown the steps of VM-UNet architecture along with VSS block for image segmentation.

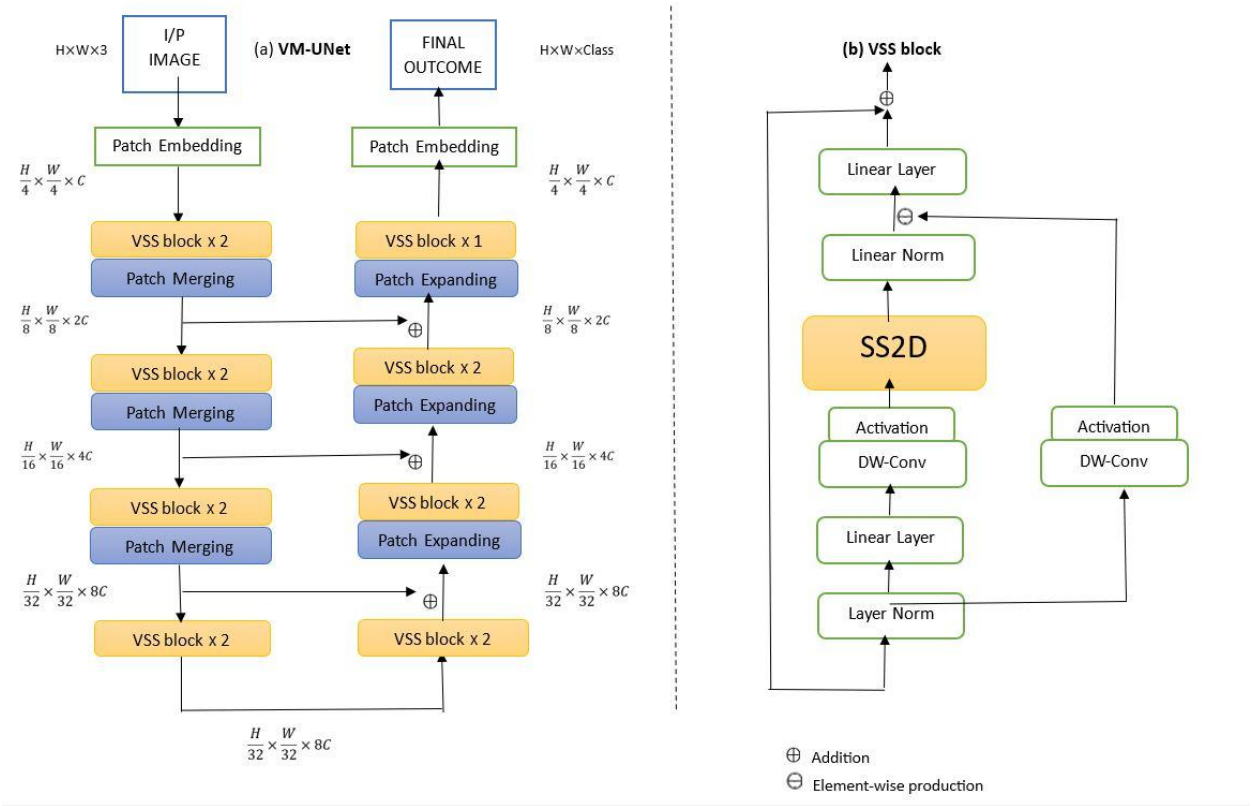

Figure 4. VM-UNet architecture with VSS block for segmentation

Next section describes the research objectives in detail.

Research Objectives

- To design a hybrid CNN–Transformer architecture for segmentation of MRI images.
- Apply the simulated model to existing public datasets and then extend to collected datasets.
- To implement explainability modules (Grad-CAM, attention maps) to visualize model decisions.
- To evaluate robustness under domain shifts, scanner differences, and imaging noise.
- To benchmark the suggested model against existing state-of-the-art approaches.

Next section details the proposed steps of methodology along with the descriptions of the datasets.

Methodology

From the literature review section, it has been shown that VM-based architecture may overcome the drawback the transformer-based models. Combining VM architecture with UNet++ may produce the best segmentation results. So, in the next subsection VM-UNet++ architecture along with XAI has been discussed.

Proposed VM-UNet++ model with XAI for visualizing results

VM-UNet passes through the steps like skip connections, Patch Embedding layer, an encoder, a decoder, a Final Projection layer. The Patch Embedding layer is responsible for dividing the input image into non-overlapping patches of fixed size. The resultant embedded image is normalized using Layer Normalization before feeding it into the encoder for extraction of features. Similarly, the decoder is organized into four stages where in last three stages, a patch expanding operation is carried out to decrease the number of

feature channels so that the height and width can be increased. At the decoder end, patch expanding is carried out using the concept of 'skip-connections' of U-Net architecture. Now in continuation with that, instead of single skip connection, we are going to use nested skip-connections of U-Net++ model that will help to up-sample the decoding results more accurately [7-8]. Further to add, while training the model basic Binary Cross-Entropy and Dice (BceDice) loss will be considered is the addition of L_{Bce} and L_{Dice}. Further to visualize the segmented outcome, a sophisticated method called Gradient-weighted Class Activation Mapping (Grad-CAM)/SHAP/LIME. This technique highlights the significant areas of an image that helped with the prediction by creating a coarse localization map using the gradients that flow into the final convolutional layer. To improve the visual quality, the advanced version in the form of Grad-CAM++ will also be taken care of [9-10]. The proposed model of VM-UNet++ has been shown in Figure 5.

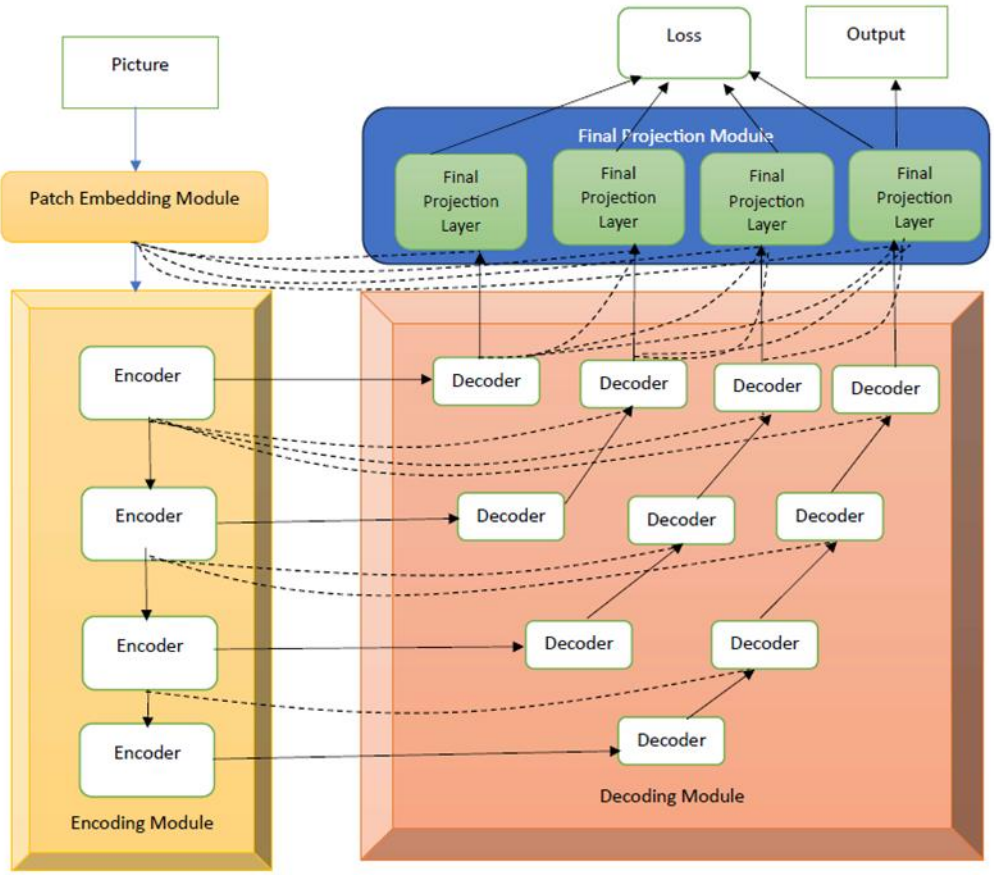

Figure 5 Proposed Model architecture of VM-UNet++

Databases for conducting research

The below mentioned publicly available datasets below have been chosen for the experiment.

- **ISIC 2017 Dataset: -** https://challenge.isic-archive.com/data/#2017

This dataset contains 2000 lesion images in JPEG format, 2000 binary mask images in PNG format, and 2000 corresponding super pixel masks in PNG format.

- **ISIC 2018 Dataset:** https://challenge.isic-archive.com/data/#2018

This dataset contains 2594 images and 12970 corresponding ground truth response masks (5 for each image).

- **CVC-ClinicDB :-**https://www.kaggle.com/datasets/balraj98/cvcclinicdb

CVC-ClinicDB is a dataset for semantic segmentation tasks. The dataset consists of 612 images with 612 labeled objects.

- **Synapse Dataset: -**https://www.synapse.org/Synapse:syn3193805/wiki/217753

Abdominal images were acquired from CT scanners across the Vanderbilt University Medical Center (VUMC); the cervix images were acquired from CT scanners at the Erasmus Medical Center (EMC) Cancer Institute in Rotterdam.

- **BraTS 2020: -**https://www.kaggle.com/datasets/awsaf49/brats20-dataset-training-validation

BraTS 2020 utilizes multi-institutional pre-operative MRI scans and primarily focuses on the segmentation.

Evaluation Metrics:

Dice coefficient: This metric is used to measure the segmentation accuracy between predicted (p) and ground truth mask (g). The formula is given below:

$$\frac{2 \times |p \cap g|}{|p| + |g|}$$

where **|p**∩ g**|**denotes the number of pixels overlapped and **|p|+|g|** counts the total number of pixels.

Intersection over Union (IoU): It is very often used for image segmentation purposes as this metric measure's separation of objects from its background. It is evaluated using the formula below:

$$\frac{TP}{(TP + FP + FN)}$$

where TP, FP, and FN represent true positive, false positive and false negative.

Hausdorff Distance: Dissimilarity between two sets of boundary points (predicted and ground truth) is measured here. The formula is given below:

$$D_H(a,b) = \max(h(a,b), h(b,a))$$

This formula finds the directed HD **h(a,b)** from set **a** to **b** (max distance from any point in **a** to its nearest in **b** and from **b** to **a**).

So, the proposed methodology in nutshell consists of data acquisition, preprocessing, model design, training, evaluation, and explainability integration:

- Data: Public datasets such as BraTS (Brain Tumor Segmentation Challenge), ISIC2018, ISIC2020, Synapse etc. will be chosen.
- Preprocessing: Skull stripping, normalization, and data augmentation.
- Model Design: Hybrid CNN–Transformer segmentation network.
- Training: Multi-loss optimization with Dice and cross-entropy losses.
- Explainability: Grad-CAM and attention visualization modules.
- Evaluation: Dice coefficient, IoU, sensitivity, specificity, Hausdorff distance

The following Figure 6 has summarized the steps of entire process visually.

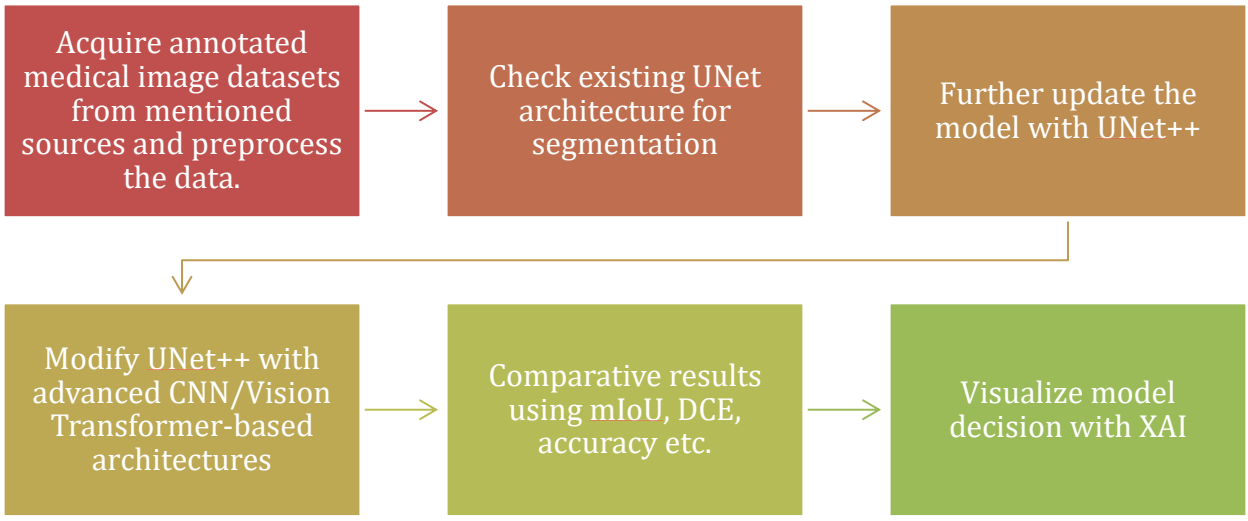

Figure 6 Steps of the proposed system

Conclusion

This review paper surveys various segmentation techniques of medical data. Starting from UNet architecture to its advanced versions have been surveyed here. The drawback of UNet model has been overcome by adding additional skip connections between the encoder and decoder blocks at multiple resolutions. Then research survey focused on transformer-based architecture for improved and accurate outcomes. Survey continued and found some major drawbacks of transformer-based models like computational cost, complexities, high resource power etc. So, 2D-Selective-Scan algorithm-based VSS block has been found which effectively did segmentation with the help of Vision Mamba (VM) architecture. The working of UNet and UNet++ in combination with VM have been finalized to carry forward in this field. In future, the final model will be tested on publicly available popular databases for best results.

References

1. S. Sangui, T. Iqbal, P.C. Chandra, S.K. Ghosh, and A. Ghosh, "3D MRI Segmentation using U-Net Architecture for the detection of Brain Tumor", *Procedia Computer Science*, vol. 218, pp.542-553,2023. https://doi.org/10.1016/j.procs.2023.01.036
2. X. Wang, Z. Hu, S. Shi, M. Hou, L. Xu, and X. Zhang, "A deep learning method for optimizing semantic segmentation accuracy of remote sensing images based on improved Unet", *Scientific reports*, vol. 13, pp.7600, 2023. https://doi.org/10.1038/s41598-023-34379-2
3. R. Singh, S. Gupta, H.G. Mohamed, S. Bharany, A.U. Rehman, Y.Y. Ghadi, and S. Hussen, "Advancing prenatal healthcare by explainable AI enhanced fetal ultrasound image segmentation using U-Net++ with attention mechanisms", *Scientific Reports*, vol. 15, pp.19612, 2025 https://doi.org/10.1038/s41598-025-04631-y

4. C. Chen, L. Yu, S. Min, and S. Wang, “Msvm-unet: Multi-scale vision mamba unet for medical image segmentation”, *In 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* pp. 3111-3114, 2024. https://doi.org/10.1109/BIBM62325.2024.10821761
5. J. Ruan, J. Li, and S. Xiang, “Vm-unet: Vision mamba unet for medical image segmentation”, *ACM Transactions on Multimedia Computing, Communications and Applications*, 2025. https://doi.org/10.1145/3767748
6. X. Zhong, G. Lu, and H. Li, “Vision Mamba and xLSTM-UNet for medical image segmentation”, *Scientific reports*, vol. 15, pp.8163, 2025. https://doi.org/10.1038/s41598-025-88967-5
7. Y. Lei, and D. Yin, “VM-UNet++: Advanced Nested Vision Mamba UNet for Precise Medical Image Segmentation”, *In 2024 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)*, pp. 1012-1016, 2024. https://doi.org/10.1109/ICICML63543.2024.10957912
8. W. Tang, Z. Wu, W. Wang, Y. Pan, and W. Gan, “VM-UNet++ research on crack image segmentation based on improved VM-Unet”, *Scientific Reports*, vol. 15, pp.8938, 2025 https://doi.org/10.1038/s41598-025-92994-7
9. D.J. Mala, M. Chattopadhyay, P. Mukhopadhyay, and R. Sinha, “Visualizing UNet Decisions: An Explainable AI Perspective for Brain MRI Segmentation”, *IEEE Access,* vol.13, pp. 133869 – 133881, 2025. https://doi.org/10.1109/ACCESS.2025.3592239
10. S. Kanrar, R. Piyush, Q. Razi, D. Chakraborty, V. Hassija, and G. S. S. Chalapathi, “Medical Image Segmentation Using Advanced Unet: VMSE-Unet and VM-Unet CBAM+”, arXiv preprint arXiv:2507.00511, 2025. https://doi.org/10.48550/arXiv.2507.00511