MSE-Guided Hybrid U-Net Framework for Automatic Kidney Segmentation and Spatial Localization

Dannial Asyraf Shahrul Anuar, Nabilah Ibrahim*, Audrey Huong
Faculty of Electrical and Electronic Engineering, University Tun Hussein Onn Malaysia, 86400, Batu Pahat, Johor, Malaysia

***Abstract*—Evaluating segmentation results in ultrasound imaging is still difficult due to noise, low contrast, and ambiguity at the boundaries, which makes it very challenging to measure accurately. Mean Squared Error (MSE) is a widely used but highly spatially sensitive evaluation metric for comparing predicted masks and ground truth. This work introduces a Mean Squared Error (MSE) based evaluation framework augmented using Block-Based Region Matching (BBRM) to achieve higher robustness against positional errors. The MSE is calculated under spatial shifts, and the best alignment with the lowest error is identified. To verify the effectiveness of the method, this work uses multiple deep learning segmentation models as baseline methods, along with the U-Net, such as SegNet and DeepLab v3+. Experimental results show that the proposed framework gives better and more reliable error analysis compared to conventional MSE evaluation. The experimental results indicated that the UNet + BBRM framework proposed in this study achieved an MSE of 0.0108, an accuracy of 98.92%, a Dice coefficient of 0.9369, and an IoU of 0.8831 in the segmentation task, respectively, compared with other methods. For the comparison with the local dataset, BBRM reduced the MSE from 0.022 to 0.015 and Dice (IoU) from 0.887 to 0.911 and 0.812 to 0.845. These findings underline the need for distribution-based error analysis and spatial alignment of segmentation methods in medical imaging applications.**

Keywords*—Kidney; BBRM; segmentation; U-Net; ultrasound imaging

I. Introduction

Medical imaging segmentation is important for clinical diagnosis and treatment planning, especially in ultrasound imaging, where the accurate segmentation of kidney anatomical structures is required [1–3]. Various deep learning approaches have been employed, among which the U-Net has received notable attention as it can handle spatial and contextual context for accurate segmentation [4,5]. Ultrasound kidney segmentation studies showed that deep learning models can help to analyze kidney structure automatically and reduce dependency on manual segmentation of the organ. However, though high accuracy is achieved, the segmentation results may still include errors induced by noisy images, low contrast, and complex organ boundaries in kidney ultrasound images [6–9].

Hybrid methods combining deep learning and post-processing approaches have been presented as solutions to these limitations [10-13]. For instance, block-based matching techniques have been adapted to obtain better segmentation by considering neighboring regions and enhancing geometric consistency in the predictions. Standard evaluation metrics like accuracy, Dice coefficient, and Intersection over Union (IoU) give measures indicating the overall performance of a model, but are often unable to show how pixel-wise similarity is distributed in the segmented image [14–18].

In this context, Mean Squared Error (MSE) distribution analysis provides a more informative measure of performance [14]. This metric provides information on the spatial and statistical distribution of the errors, usually visualized through histograms or frequency plots, rather than an aggregated value. This enables researchers to study how models work, locate the regions they tend to get wrong, and measure whether enhancement techniques like block matching are effective [16–18].

Hence, to investigate the MSE distribution for segmentation results obtained by the U-Net and then enhanced using block matching methods. While error distribution shows an important result, the analysis provides more insight into segmentation performance than would be possible with traditional metrics and demonstrates where the contributions of a hybrid approach directly improve prediction reliability for kidney ultrasound images. A U-Net model is implemented as the main segmentation model in this work [4]. The network architecture of the U-Net itself had no structural changes. External layers of U-Net are used after segmentation and contain the main contribution of work based on Block-Based Region Matching (BBRM) as well as a new MSE-guided evaluation framework which is directly related to better spatial alignment analysis and segmentation error assessment without replacing internal components of U-Net.

Mean Square Error-based (MSE-guided) evaluation was chosen since MSE is sensitive to pixel-wise discrepancies between the predicted and ground truth masks [14, 17, 18]. This property allows it to be appropriate for distinguishing small spatial misalignments, which are not well characterized by overlap-based metrics such as Dice coefficient and IoU [14-18]. Nevertheless, MSE was not the only metric for evaluation in the scope of this research. Additional metrics such as accuracy, precision, sensitivity, Dice coefficient, IoU and specificity were also calculated to assess the overall quality of segmentation [14-16]. Hence, MSE was used specifically during the BBRM alignment step, but the other metrics were used to assess segmentation quality.

The contribution of this work is to provide a more robust evaluation framework for kidney ultrasound segmentation,

*Corresponding author

especially in case that small spatial misalignment can happen between predicted masks and ground truth masks. The proposed framework may help in offering a more descriptive error analysis method for researchers, assist engineers in developing automated kidney segmentation systems as well as allow clinicians or radiologists to enhance the consistency of kidney boundary interpretation in ultrasound images.

II. Related Work

A. Deep Learning-Based Segmentation of Kidney Ultrasound

Ultrasound image segmentation has strongly profited from deep learning-based solutions, especially convolutional neural networks (CNNs) [1–3]. Among these methods, one of the most used architectures is U-Net, benefit from its encoder–decoder structure and skip connections that help to preserve spatial features in segmentation [4-5]. Because the U-Net combines high-level context and low-level boundary information, allowing precise localization, it gives benefits for kidney ultrasound segmentation [6-9]. Kidney ultrasound images have traditionally been limited to the presence of speckle noise, low contrast, and unclear anatomical boundaries that cause segmentation errors even for a model boasting high average accuracy to capture [6, 8, 9].

B. Comparative Segmentation Architectures

Recent work on segmentation architectures in the medical imaging community aims at performance improvement [5, 23-25]. SegNet is designed for semantic segmentation with an encoder–decoder structure but a lesser computational complexity [19]. With these two advancements, DeepLabV3+ leverages convolution and encoder–decoder refinement to perform better multi-scale feature extraction as well as boundary segmentation [20]. Alternatively, approaches based on attention mechanisms have been proposed, like ECA U-Net and SE U-Net, which force enhanced channel-wise feature selection [21, 22]. These models serve as a baseline for comparison since each architecture has its own advantages and disadvantages in terms of retaining kidney structure information, boundary information, and overall smoothness of the segmentation output.

C. Limitations of Conventional Evaluation Metrics

Common performance evaluation metrics of segmentation are traditional metrics such as accuracy, Dice coefficient, IoU, precision, and sensitivity [14-16]. These metrics are so relevant to this work as they offer quantitative information on pixel classification accuracy and region overlap between the predicted mask and ground truth mask. However, these mainly provide aggregate performance metrics and may not adequately capture the spatial distribution of segmentation mistakes [17,18]. Based on the challenges of ultrasound images, boundary ambiguity, noise, and low contrast can cause a small positional shift between the predicted mask and the ground truth mask [6-9]. Thus, more error analysis is needed in order that segmentation reliability can be assessed more clearly.

D. Research Gap

Most of the previous kidney ultrasound segmentation works focus on enhancing model architecture and reporting traditional performance measures. However, there are not many studies on pixel-wise error measurement and spatial alignment dissimilarity by comparing the predicted segmentation from the ground-truth masks [14, 17, 18]. The importance of this gap is, since a segmentation result might still have good overlap scores while containing spatial errors in small areas. Thus, we propose the MSE-guided evaluation framework with the Block-Based Region Matching (BBRM). The method gives a score of segmentation reliability by shifting the predicted mask and computing MSE at each position, then it chooses the best alignment with the lowest error so that spatial consistency can be better studied.

E. Comparison with Previous Studies

Table I shows a comparison with previous studies' methods for kidney ultrasound segmentation that have been reported, including the boundary distance regression, DeepLabV3+ (without fine-tune), SegNet, GL-UNet11 and attention-based U-Net models like U-Net with attention gates based on 2D images. For example, Chen et al. and Zuo et al. Attention-based U-Net architectures: improved segmentation performance, Daoud et al. This work also compares multiple deep learning segmentation models (SegNet, DeepLabV3+) with the Open Kidney Ultrasound Dataset.

TABLE I. Comparison with Previous Studies

Metric	Method	Image Modality	IoU	Dice
Valente et [7]	Channel attention and GL-UNet11	Kidney Ultrasound	Not Report	78.0%
Zuo et al.[9]	Attention-based renal segmentation network	Kidney Ultrasound	88.74%	93.83%
Chen et al. [26]	GL-UNet11 with channel attention	Kidney Ultrasound	82.89%	86.21%
Daoud et al. [27]	U-Net, SegNet, DeepLabV3+, UNet++, MA-Net, LinkNet, PAN, BiSeNetV2, SegFormer	Kidney Ultrasound	78.2%	86.9%
Proposed study	U-Net + BBRM	Kidney Ultrasound	88%	93.69%

Previous works are mainly used for segmentation architecture improvement or comparing deep learning models using traditional evaluation metrics such as the Dice coefficient and IoU. These metrics can effectively measure the overlap of the segmentation, but fail to explain pixel-wise spatial error between the predicted mask and the ground truth mask. Ultrasound images of the kidney have low contrast, unclear boundaries, variation in the appearance from each subject, and speckle noise that can cause small misalignments.

Hence, the contribution of this study was to highlight that there are very few applications where analysis on post-segmentation spatial alignment has been performed for kidney ultrasound segmentation. In order to fill this gap, the current study proposes the use of U-Net as a principal model for segmentation with subsequent Block-Based Region Matching (BBRM) right after the segmentation process.

III. METHODS

A. Dataset Acquisition

The dataset obtained was used for training and testing in terms of improving the accuracy of ultrasound. There were two datasets, consisting of an open dataset and a local dataset. The open-source dataset was obtained from the Kaggle dataset, and the local dataset was manually captured by using ultrasound equipment, Canon Aplio a CUS-AA000, in our laboratory. For open-source datasets, 1000 images were used for training sessions, and 200 images were used for validation. No medical history or information, such as age and gender, was provided in the dataset. The local datasets consist of 100 manually captured videos of the kidney, which have 249 frames per second (fps) and a duration of 10 seconds. This dataset was used for testing the model of U-Net to conduct automatic segmentation and localization by using the Block-Based Region Motion (BBRM) technique.

B. U-Net Training Process

Fig. 1 shows the architecture of the U-Net model, which is built as an encoder–decoder structure for semantic segmentation tasks. There are three main components of the network which are: contracting path also known as contracting part, the bridge, and the decoder also known as expanding path. The encoder gradually learns feature representations of the input ultrasound image through several stacked convolutional layers, ReLU activation functions, and max-pooling operations. In the encoding half of the architecture, it decreases spatial resolution while increasing number feature channels to allow the network to learn high level contextual features.

The bridge connects the encoder and decoder at its deepest level that able to capture the highest-level representation of image features and gives a more global representation of kidney structure. Through an up-sampling operation which is transposed convolution, the segmentation map was reconstructed through a decoder that successively restores spatial resolution. Skip connections append the corresponding encoder feature maps to the decoders at each decoding stage. Such skip connections are invaluable in capturing spatial information otherwise lost due to down sampling, thus making boundaries more precise. The U-Net that was implemented in this work used the standard encoder–decoder architecture with skip connections. It comprises convolutional layers, ReLU activation functions, max-pooling layers for down-sampling, a bridge layer, transposed convolution layers for up-sampling and skipping concatenation of signals from each encoder stage to their corresponding decoder stages. Most notably, no structural change in the structure of the U-Net architecture. Thus, U-Net served as the baseline segmentation model and the proposed improvement reflected post-processing and evaluation at this stage utilizing BBRM.

Finally, a 1 x 1 convolution layer with a SoftMax function that gives pixel-wise classification output in example for each pixel can be classified either as kidney region or background. Compared to other models, its overall architecture helps in merging the low-level and high-level spatial information for semantic image segmentation applications such kidney ultrasound.

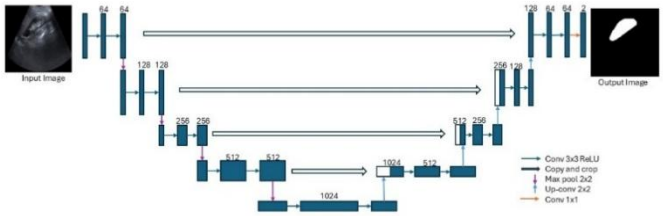


Fig. 1. U-Net model architecture

U-Net architecture was used in this work to classify segmentation masks from kidney ultrasound images. U-Net is a convolutional neural network (CNN) specifically built for biomedical image segmentation with an encoder–decoder architecture featuring skip connections.

For the encoder, convolutional layers and pooling operations were employed stepwise in a layer-style fashion to extract more hierarchical feature directly from input and realizing down-sampling of spatial resolution along the way with greater number of features. Image segmentation decoder reconstructs a segmentation map with a series of up sampling operations that increase the pixel density. The model has a skip connection to combine low-level and high-level features so it can release fine structural detail, particularly at object boundaries:

$$y(i,j) = \sum_m \sum_n x(i+m, i+n).w(m,n) \quad (1)$$

Rectifier Linear Unit (ReLu) also known as non-linearity:

$$f(x) = max(0,x) \quad (2)$$

SoftMax layer for pixel-wise classification is obtained by using the final segmentation output.

C. Block-Based Region Motion(BBRM)

Fig. 2 shows Block-Based Region Matching (BBRM) process where the ground truth mask is fixed, and predicted mask is moved in spatial directions (*dx,dy*), to explore for the best segmentation. The red grid indicates where the image is monitored in equal block that will be used to calculate local errors. The predicted mask was placed on top of the ground truth for every corresponding shift, from which the Mean Squared Error (MSE) was computed for each corresponding block. This will lead to obtaining block-wise MSE values and taking the average overall error. The optimal shift is the one with the minimum MSE, as it leads to more similarities between predicted mask and ground truth.

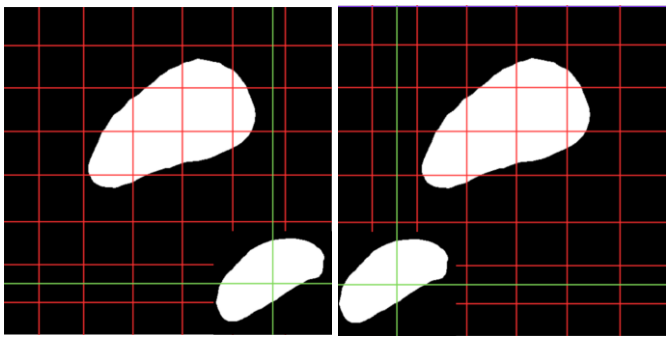


Fig. 2. Block-Based Region Matching (BBRM) illustration process.

In this work, BBRM was used after the segmentation output of the U-Net model. As BBRM targets post-processing and evaluation refinement, which is an additive step than altering a structure of the U-Net architecture. Before evaluation, all ultrasound images and ground truth masks were resized to a size of (256 × 256) pixels. The predicted mask was translated in the horizontal ((dx)) and vertical directions ((dy)), by randomly setting (dx) and (dy), respectively, from (-10) to (+10) pixels. Block-based MSE Between the shifted predicted mask and the corresponding ground truth mask. (block size = 16 × 16) For every position, the block MSE was computed and then the shift position with minimum MSE value is assumed to be the best alignment. This further minimizes the impact of any possibility that the predicted mask and ground truth mask doesn't align, because small spatial misalignment can occur between both images due to speckle noise, low contrast, and blurry kidney boundaries in ultrasound images.

Standard evaluation frameworks assume that the predicted segmentation and the ground truth are spatially aligned perfectly. The placement of the position used for calculating errors, however, can have a large impact on both measurement accuracy and therefore errors values particularly in data such as ultrasound images where boundaries are hard to determine.

To facilitate addressing this limitation, a Block-Based Region Matching (BBRM) is proposed. The predicted segmentation mask is moved inside a fixed size search window both in horizontal and vertical directions.

Mean Squared Error (MSE) is computed as (*dx,dy*) for each shift:

$$MSE(dx, dy) = \frac{1}{N}\sum_{i=1}^{N}(GT_i - P_i^{(dx,dy)})^2 \quad (3)$$

where, *N* is the total number of pixels while *GT* is actual mask pixels from ground truth and *P* is the predicted mask after shifting. *i* is the index pixels number, by selecting the minimum MSE across possible shift, the optimal alignment is obtained:

$$MSE_{min} = \min_{dx,dy} MSE(dx, dy) \quad (4)$$

Min means the selection of the smallest MSE value tested vertical and horizontal shift *(dx,dy)*. This method allows the evaluation process to compensate for spatial offset and offers a more stable measure of segmentation performance.

D. Performance Evaluation Metric

All evaluation measures were computed pixel-wise by directly comparing predicted segmentation mask with the ground-truth mask. Predicted masks and ground truth masks were resized to the same image size before evaluation which is 256 × 256. Pixel-wise error was measured using MSE, and overlap similarity was assessed by Dice coefficient and IoU. To give a more comprehensive characterization of segmentation, need to compute the accuracy, precision, sensitivity and specificity. All experiments on the testing images were conducted using the same mask format, search range, block size and evaluation metrics. The ablation comparison aimed to provide the contribution of BBRM alone, while the model comparison was used to compare different segmentation architectures with identical experimental conditions.

To assess the segmentation task performance fully several quantitative metrics are used such as Mean Squared Error (MSE), Dice coefficient, Intersection over Union (IoU), Accuracy and Precision metric. These metrics give additional perspectives for evaluating pixelwise error, spatial overlap, and classification confidence between the predicted segmentation and ground truth (when performing hybrid segmentation). In the evaluation metrics, TP, TN, FP and FN refer to pixel-wise classification results between predicted mask and ground truth mask. TP means True Positive, where kidney pixels are predicted to kidney correctly. TN is a true negative, which means that background pixels are predicted as the background. FP, which is false positive, background pixels wrongly predicted as kidney FN known as False Negative, kidney pixels predicted as background.

Mean Squared Error (MSE) describes the average squared difference between segmentation and ground truth:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(GT_i - P_i)^2 \quad (5)$$

GT is the correct pixel from ground truth mask and *P* is the predicted pixels from model output. The *i* is the index number and *N* is total pixel in the mask. MSE calculates pixel-level error with a penalty on big difference against predicted and actual values. The lower the MSE, the better the segmentation. However, MSE is sensitive to spatial misalignment, and small positional offsets can cause large increases in error value.

The Dice coefficient is used to evaluate the similarity between the predicted segmentation and ground truth:

$$Dice = \frac{2TP}{2TP+FP+FN} \quad (6)$$

Intersection over Union (IoU) calculates the overlap between the ground truth and the predicted mask:

$$IoU = \frac{TP}{TP+FP+FN} \quad (7)$$

This metric focuses on the intersection of two regions and is commonly used in medical image segmentation and Dice value is near to 1 means highly similarity and gives accurate kidney part segmentation.

Accuracy assesses the share of correctly classified pixels:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (8)$$

This metric measures how well the entire model is classifying. But it can be affected by imbalance class, for example if background pixels occupy the image.

Precision measures confidence in positive predictions as follows:

$$Precision = \frac{TP}{TP+FP} \quad (9)$$

High precision means that most of the kidney pixels predicted are classified correctly and few background regions are detected (false positives).

Sensitivity is how well the segmentation model to identify kidney regions from ultrasound images. It is calculated as the ratio of TP pixels to the total number of actual kidney pixels, including FN:

$$Sensitivity = \frac{TP}{TP+FN} \quad (10)$$

Larger values of the sensitivity reflect better detection performance, meaning that fewer kidneys were missed.

All these evaluation metrics together give the overall picture of segmentation performance. Mean Squared Error (MSE) looks at pixel-wise differences, such as Dice and IoU, concentrating on spatial overlap.

IV. Results and Discussion

To maintain consistency between the reported results, all results were computed using a unified preprocessing pipeline where the shapes of ultrasound images and masks are resized to (256 × 256) pixels, followed by conversion into a binary mask for evaluation. All the images tested used the same mask format, search range, and block size of evaluation. The ablative comparison served to assess the BBRM contribution in isolation, whereas the model comparison allowed us to benchmark different segmentation architectures against each other with an identical experimental setup.

A. Comparison of Performance Between U-Net and U-Net + BBRM

Table II shows the comparison of performance between U-Net and U-Net + BBRM. The goal of this comparison was to observe the contribution of BBRM to the segmentation result. The measures obtained for the baseline U-Net were 0.022, a Dice coefficient of 0.887 and an IoU of 0.812. MSE values decreased to 0.015, and the Dice coefficient and IoU improved to 0.911 and 0.845, respectively, after BBRM was applied. The MSE reduction indicates that BBRM aids to lower the pixel-wise error of the predicted mask from the ground truth mask. Furthermore, the enhanced values of Dice and IoU imply an improvement in resemblance and overlap between the segmentation output and the ground truth. These results clearly verify that BBRM indeed adds positively to the proposed framework in enhancing spatial alignment and segmentation consistency.

TABLE II. Comparison of Performance Between U-Net and U-Net + BBRM.

Method	MSE	Dice	IoU
U-Net	0.022	0.887	0.812
U-Net + BBRM	0.015	0.911	0.845

The decrease in MSE indicates BBRM's ability to reduce the pixel-wise error between the predicted mask and the ground truth mask. Moreover, the enhancement in Dice coefficient and IoU shows that the prediction result has more overlap and similarity with the ground truth. Performance on the local dataset is relatively high, which might be due to BBRM being able to overcome small spatial misalignments (commonly introduced in ultrasound segmentation due to unclear boundaries, speckle noise and appearance variability of kidneys). Thus, it suggests that BBRM appears to be able to assist the proposed framework in terms of better spatial alignment and segmentation consistency. Nonetheless, additional tests are done on bigger and more heterogeneous data.

B. Hybrid U-Net and BBRM Results

Table III shows the performance evaluation of the proposed Hybrid U-Net and BBRM-based segmentation framework for ultrasound renal image analysis. The experimental results show that the proposed method performed well in terms of multi-metric evaluation. The model resulted in a very low mean squared error (MSE) value of 0.0108, demonstrating that the predicted segmentation masks were like the ground truth mask at the pixel level with only a minimum pixel error. Moreover, the proposed framework obtained an overall accuracy of 98.92% indicted the segmentation system was greatly efficient in correctly classifying kidney and background regions within the ultrasound images.

In addition, the precision of 0.9815 indicates that most of the segmented kidney regions were correct, as we had very few false positive detections from this model. The sensitivity value of 0.8983 indicates that the proposed method successfully detected most of the kidney regions and missed some pixels in the segmentation region. On the other hand, the dice coefficient of 0.9369 and IoU value of 0.8831 support that it has high overlapping similarity between ground truth masks and predicted masks, which shows accurate kidney boundary segmentation. The model also achieved extremely low false-positive rates, as evidenced by a 0.9984 specificity value.

TABLE III. Hybrid U-Net and BBRM Results

Metric	Mean
MSE	0.0108
Accuracy	0.9892
Precision	0.9815
Sensitivity	0.8983
Dice	0.9369
IoU	0.8831
Specificity	0.9984

C. MSE Distribution Analysis

Fig. 3 shows the histogram of Mean Squared Error (MSE) distribution after using the proposed Block-Based Region Matching (BBRM). The MSE is plotted along the horizontal axis, with an image frequency in each error interval shown on the vertical axis.

Most of the images are in the low MSE range, around 0.005 to 0.012. The peak occurs in the interval of 0.006 to 0.009, which has approximately 580–600 images, which means most of the dataset results in a low error after aligned.

At a relatively low level of the same scale, 0.009–0.015, frequency decreases to roughly 200–300 images, which means moderate error for a smaller subset of the dataset. After 0.02, there are far fewer images, and only a small handful of cases

go up to higher MSE values up to around 0.09, which equate to more difficult segmentation situations.

Overall, the distribution is highly skewed towards lower MSE values, showing that on most of the test images, the accuracy and performance in terms of alignment learned by the proposed BBRM method are quite high, with only a few outliers at higher error.

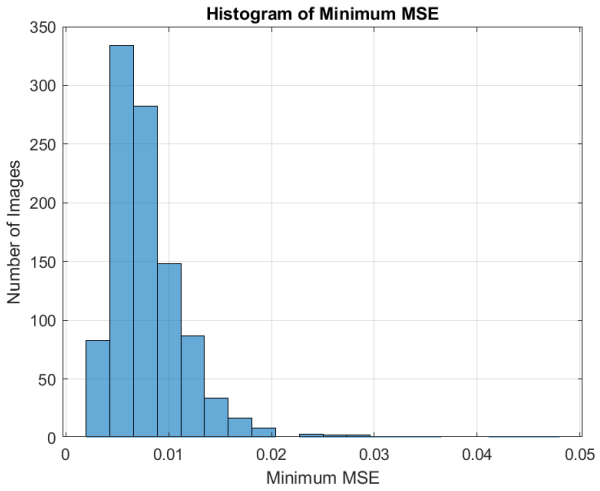


Fig. 3. MSE distribution

Fig. 4 shows the average MSE values when the predicted segmentation masks are shifted dy (vertical) and dx (horizontal). The darkest spot near the center (dx ≈ 0, dy ≈ 0) has minimal error, and so the highest alignment between predicted and ground truth masks at very low displacement. The greater the shift away from the center, the more gradually, and in brighter colors, the greater the misalignment, evidenced by higher MSE values. The result indicates that BBRM searches for the best alignment with minimum MSE and improves the accuracy and consistency of segmentation.

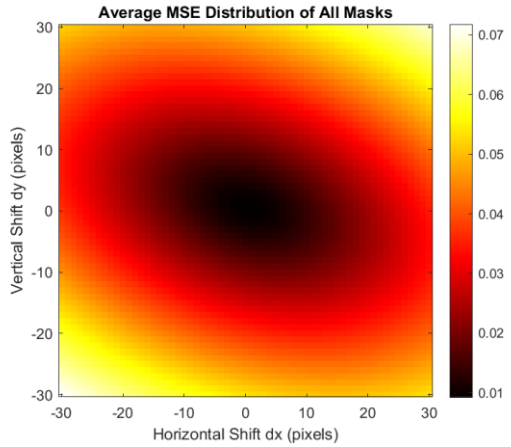


Fig. 4. Heatmap of average MSE distribution.

D. Comparison with Other Models

Comparison of the segmentation models evaluated in this work: DeepLab v3+, SegNet, SE(U-Net), ECA U-Net and U-Net based on some quantitative metrics are presented in Table IV. U-Net has best overall performance with accuracy 0.9912, Dice coefficient 0.948, IoU is 0.902, which shows its potential to predict true kidney region and maintaining structural boundaries well. Even though ECA U-Net generates the smallest MSE of 0.0173, indicating that pixel-wise similarity is higher, its Dice (0.9169) and IoU (0.8512) are still lower than U-Net. In parallel, the SE U-Net achieves comparable performance (Dice 0.9072, IoU 0.8356) to that of vanilla U-Net, clearly showing the impact of attention mechanism which reinforces ability to interpret features.

On the other hand, SegNet and DeepLab v3+ show relatively low performance on almost all metrics. SegNet has intermediate performance with a Dice =0.8570, and IoU= 0.7601, but DeepLab v3+ obtains the worst segmentation results, especially in Dice (0.7866) and IoU (0.6583), which shows that this model is unsuitable for segmenting kidney boundaries from ultrasound images with high accuracy. In summary, these results verify that U-Net is still the leading model in this study, while other models set up a solid baseline to test the stability and trustworthiness of the segmented framework.

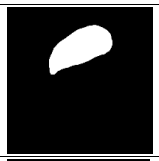
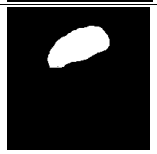
TABLE IV. QUANTITATIVE RESULTS FOR FIVE DIFFERENT MODELS

Model	**Metrics**					
	Accuracy	***MSE***	***Sensitivity***	***Dice***	***IoU***	***Precision***
Deeplab	0.9511	0.0488	0.9885	0.78664	0.6583	0.664
SegNet	0.971	0.0280	0.9624	0.8570	0.7601	0.783
SE-U-Net	0.9801	0.0198	0.9780	0.9072	0.8356	0.838
ECA-U-Net	0.9826	0.0173	0.9727	0.9169	0.8512	0.8845
U-Net	0.9912	0.026	0.9212	0.948	0.902	0.977

Table V shows the comparison of kidney segmentation results for three ultrasound images in five different models, such as U-Net, SegNet, DeepLab v3+, ECA U-Net, and SE-U-Net. The first row corresponds to the original images, and it is then followed by ground truth masks, while rows 2 and 3 show the predicted segmentation outputs. Based on visual inspection, U-Net generates results closest to ground truth, primarily in retaining kidney shape and boundary across all images. Alternatively, SegNet and DeepLab v3+ come up with less accurate segmentation, showing easily seen boundary mistakes and a little over-segmentation. The attention-based models, ECA U-Net and SE U-Net, behave better than SegNet and DeepLab by providing smooth and more uniform shapes, but still minor deviations from the ground truth are observed.

TABLE V. SEGMENTED RESULTS OF FIVE DIFFERENT MODELS

Type of model	**Image Number**		
	Image 1	***Image 2***	***Image 3***
Original Kidney image			

Type of model	Image Number		
	Image 1	*Image 2*	*Image 3*
Ground Truth			
U-Net			
SegNet			
DeepLab			
ECA-U-Net			
SE-U-Net			

V. CONCLUSION

This work proposed a strong kidney segmentation framework for ultrasound imaging based on the U-Net as the main deep-learning architecture. The U-Net was chosen due to its encoder–decoder architecture, which allows for the successive extraction of hierarchical features combined with preserve spatial information using skip connections. This design is well-suited for the requirement of boundary segmentation in medical images. Multiple advanced semantic segmentation architectures (SegNet, DeepLab v3+, SE U-Net, and ECA U-Net) were also implemented as benchmark models in addition to the primary model for an extensive comparison. This comparison gives a better understanding of the other benchmarks and existing methods, including their strengths and technical limitations, using an objective metric against which to compare the results from the proposed approach.

The results prove that U-Net always has a good performance on the ultrasound images, especially in the segmentation of the shape and structure of the kidney region. U-Net preserves more details and boundaries, which are important in medical diagnosis; Better than SegNet and DeepLab v3+. kt is constructed with the ground truth mask and the synthesized foreground image. However, even if attention-based variants, like SE U-Net and ECA U-Net, utilize channel-wise feature refinement, the improvements are not always prominent on all test samples, which means that the baseline architecture of U-Net is still a strong option for this study's use case. The meta-analysis implements multiple models proven to perform well in the same experimental condition, thereby strengthening the findings. The BBRM is a post-processing step to further boost the segmentation results. The BBRM method works by adjusting the predicted segmentation mask in both horizontal and vertical directions using a block-wise search strategy to minimize the mean squared error (MSE) with respect to the ground truth. This does a very useful correction for minor spatial misalignments that may be present in predictions made by deep learning algorithms, since extension of blurred acoustic boundaries often does not provide precise information about the size and location of structures if they are known to have subtle, ill-defined, or noisy components, as is the case for ultrasound images. The BBRM helps in aligning the predicted pixels of the kidney region with those of the ground truth without retraining or adding further complexity during inference on a trained model.

Experimental results indicate that BBRM successfully enhances the segmentation performance quantitatively. Dice coefficient and Intersection over Union (IoU) metrics, which are great indicators of the similarity in predicted masks from ground truth, improve, while MSE values reduced imply linearly that pixel-wise agreement is better between predicted masks and the respective ground truths. Moreover, BBRM results in a much sharper, lower range of MSE distribution across the full dataset, indicating better reliability and robustness of the segmentation result. These results confirm that the proposed refinement technique improves the reliability of segmentation output compared to the baseline deep learning model.

This study improves upon the segmentation accuracy for kidney regions but adds bounding box localization, which provides a straightforward and interpretable presentation of the detected area. The system visually highlights the position of the kidney in the ultrasound image and its extent by extracting the largest connected component from the refined segmentation mask and calculating the bounding box. This both facilitates human interpretability of the results and it also enables clinical applications that require rapid localization. Together, segmentation, refinement, and localization make for a more competitive and feasible kidney detection solution.

In conclusion, the study proposed a U-Net + BBRM framework for reliable and efficient kidneys segmentation in ultrasound imaging. It shows that U-Net makes for a strong baseline model, but gains can be made by combining segmentation with an additional simple yet efficient post-processing refinement technique to increase segmentation resolution and consistency. The illustrative results empower the suitability of U-Net for this task, while the BBRM integration and localized embeddings enhance the system performance and usability compared to other deep learning-based models. Because of the improvements in the reliability of kidney ultrasound segmentation evaluation, this work is important to researchers, system developers and clinical users. The proposed framework integrates U-Net segmentation, BBRM and MSE-guided performing pixel-level error comparison and spatial alignment, which may assist in building

more robust computer-assisted ultrasound analysis systems. Future work will focus on the use of advanced attention approaches, improving computation efficiency, and adaptation of the proposed framework to live applications and multiple imaging modalities.

ACKNOWLEDGMENT

This research was supported by the Universiti Tun Hussein Onn Malaysia (UTHM) through the GPPS Vot J040. Communication of this research is made possible through monetary assistance by the Universiti Tun Hussein Onn Malaysia and the UTHM Publisher's Office via Publication Fund E15216.

REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, May 2015.

[2] Z. Wang, E. B. Simoncelli, and A. C. Bovik, "Deep learning in medical ultrasound image segmentation: A review," arXiv preprint arXiv:2002.07703, 2020.

[3] X. Zhang, "Imaging-based deep learning in kidney diseases," Insights into Imaging, vol. 15, no. 1, 2024.

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in Proc. Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany, 2015, pp. 234–241.

[5] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," Nature Methods, vol. 18, no. 2, pp. 203–211, Feb. 2021.

[6] R. Singla, C. Ringstrom, G. Hu, V. Lessoway, J. Reid, C. Nguan, and R. Rohling, "The Open Kidney Ultrasound Data Set," arXiv preprint arXiv:2206.06657, 2022.

[7] S. Valente et al., "A comparative study of deep learning methods for multi-class semantic segmentation of 2D kidney ultrasound images," Computers in Biology and Medicine, 2023.

[8] M. G. Oghli et al., "Fully automated kidney image biomarker prediction in ultrasound imaging using Fast-Unet++," Scientific Reports, vol. 14, 2024.

[9] Y. Zuo, J. Li, and J. Tian, "A segmentation network with two distinct attention modules for the segmentation of multiple renal structures in ultrasound images," Diagnostics, vol. 15, no. 15, 2025.

[10] S. H. Song et al., "Deep-learning segmentation of ultrasound images for automated calculation of pediatric hydronephrosis measurements," Investigative and Clinical Urology, vol. 63, no. 5, pp. 573–580, 2022.

[11] X. Xiao et al., "Deep learning-based medical ultrasound image and video segmentation," Sensors, vol. 25, no. 8, 2025.

[12] M. Rainio et al., "Deep learning for medical ultrasound image segmentation," Journal of Digital Imaging, 2026.

[13] N. Alkhaldi et al., "Automating kidney disease diagnosis: A segmentation and classification approach using ultrasound imaging," Engineering, Technology & Applied Science Research, 2026.

[14] A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool," BMC Medical Imaging, vol. 15, no. 29, pp. 1–28, 2015.

[15] D. Müller, I. Soto-Rey, and F. Kramer, "Towards a guideline for evaluation metrics in medical image segmentation," BMC Research Notes, vol. 15, no. 210, 2022.

[16] D. Müller, D. Hartmann, P. Meyer, F. Auer, I. Soto-Rey, and F. Kramer, "MISeval: A metric library for medical image segmentation evaluation," Studies in Health Technology and Informatics, vol. 294, pp. 389–390, 2022.

[17] S. Ostmeier et al., "USE-Evaluator: Performance metrics for medical image segmentation models with uncertain, small or empty reference annotations," arXiv preprint arXiv:2209.13008, 2022.

[18] Z. Zhang and U. Bagci, "Segmentation quality and volumetric accuracy in medical imaging," arXiv preprint arXiv:2404.17742, 2024.

[19] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[20] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in Proc. European Conference on Computer Vision (ECCV), Munich, Germany, 2018, pp. 801–818.

[21] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 2018, pp. 7132–7141.

[22] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 11531–11539.

[23] O. Oktay et al., "Attention U-Net: Learning where to look for the pancreas," arXiv preprint arXiv:1804.03999, 2018.

[24] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in Proc. Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, 2018, pp. 3–11.

[25] A. Hatamizadeh et al., "UNETR: Transformers for 3D medical image segmentation," in Proc. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2022, pp. 574–584.

[26] S.-H. Chen, Y.-L. Wu, C.-Y. Pan, L.-Y. Lian, and Q.-C. Su, "Renal ultrasound image segmentation method based on channel attention and GL-UNet11," Journal of Radiation Research and Applied Sciences, vol. 16, p. 100631, 2023.

[27] M. I. Daoud, F. Abunameh, K. Shweikeh, S. K. Alzamer, M. Z. Ali, and R. Alazrai, "A comparative study of deep learning semantic segmentation models for kidney segmentation in ultrasound images using the Open Kidney Ultrasound Dataset," IEEE Access, vol. 13, pp. 144417–144433, 2025.