Journal of Information Systems Engineering and Management
2025, 10(50s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Enhancing Accuracy in Kidney Disease Prediction Using a CNN-Transformer Hybrid Model on Ultrasound Images

Mrunali Sonwalkar [1], Dr. Sharvari C. Tamane [2]

[1] Assistant Professor, Computer Science Engineering, College of Engineering Ambajogai, Beed, Maharashtra, India. Email: mrunali.sonwalkar@gmail.com

[2] Head of a Department, Information Technology, JNEC and University Department of Information and Communication Technology (UDICT), MGM University, Chatrapati Sambhaji Nagar, India. Email: hodudict@mgmu.ac.in

ARTICLE INFO

Received: 11 Mar 2025

Revised: 04 May 2025

Accepted: 13 May 2025

ABSTRACT

Kidney Disease (KD) is characterized by a gradual decline in kidney function, which can eventually lead to kidney damage or failure. As the disease progresses, diagnosis becomes more challenging. Incorporating routine clinical data to assess different stages of KD can aid in early detection and timely intervention. Advanced stages of KD are associated with a higher risk of cardiovascular complications and mortality. Ultrasound (US) imaging is widely used in clinical practice for predicting KD due to its safety, convenience, and affordability. However, manual analysis of US images is time-consuming, prone to errors, and requires highly skilled professionals. In recent years, Deep Learning (DL) has shown promising results in medical image analysis. This research introduces a hybrid DL network, Convolutional Neural Network (CNN)-Transformer, designed to predict KD from US images. To conduct the study, US images of both healthy and diseased kidneys were collected from Aadhar Diagnostic Centre, Maharashtra. The collected raw images underwent several pre-processing steps, including resizing and augmentation. The processed dataset was then split into training, validation, and test sets in a 7:2:1 ratio. The proposed hybrid network was compared with well-known DL networks, ResNet and DenseNet. All three models were trained, validated, and tested under identical conditions, including the same number of images, epochs, and hyperparameters, to ensure a fair comparison. The models were tested on 25 healthy and 25 diseased images. The results showed that DenseNet and ResNet correctly predicted 44 and 43 cases, respectively, while the proposed CNN-Transformer network achieved 49 correct predictions out of 50 samples. The proposed network attained the highest accuracy of 98%, whereas DenseNet and ResNet achieved 88% and 86%, respectively. In addition to accuracy, other evaluation metrics, including Precision, Recall, and F1-Score, were also significantly higher for the proposed network compared to the other two networks. These findings demonstrate that the proposed CNN-Transformer network delivers promising results for KD prediction using US images.

Keywords: Ultrasonic Kidney Images, Data Augmentation, Deep Learning, Transformer, Google Colaboratory, Convolutional Neural Network, Accuracy.

INTRODUCTION

KD affects more than 10% of the world's population, making it a serious public health issue [1]. Over the last few decades, the prevalence of morbidity in KD patients has skyrocketed, imposing significant medical and economic pressures on the global healthcare system. Many cases of KD progress to uremia before anyone seeks medical attention, and the disorder sometimes presents with no apparent symptoms at all [2]. However, early detection and treatment of KD allow for effective disease management and, in rare cases, reversal. One key component in the early detection and prevention of KD is accurate KD staging, which has been shown in studies to reduce the progression of kidney damage [3].

KD is defined as health-impairing structural or functional renal abnormalities that have lasted for longer than three months. KD can be divided into five stages, G1-G5, according to the estimated glomerular filtration rate. Proteinuria and serum creatinine levels are the two most commonly utilized KD screening tests today. However, biochemical testing of blood and urine is tedious and time-consuming [4]. Furthermore, many people neglect to seek out these tests during routine screenings, allowing KD to go unnoticed for a long time. Pathological studies of KD can show the extent and cause of kidney fibrosis, but they require a renal biopsy, which is invasive and can result in complications such as perirenal hematoma, arteriovenous fistula, or infection [5]. Additionally, performing a second biopsy on the same patient to track their progress and manage their therapy throughout a longitudinal study is not a viable option. In contrast, conventional US is a radiation-free, cost-effective, and noninvasive imaging technique used to diagnose KD [6]. Most kidney problems are initially detected using the US. In practice, skilled doctors can use this method to quickly detect end-stage KD with renal atrophy or significant changes in renal echogenicity. However, for the vast majority of patients with early-stage KD or transitional KD, this procedure is beyond the scope of human vision and difficult to perform in clinical settings.

Many studies have already been conducted on KD prediction using various DL models. Researchers have used different types of data, such as Computed Tomography (CT) scans, US images, Magnetic Resonance Imaging (MRI), and clinical records. Table 1 presents some of the recent and noteworthy research works on KD prediction. The table provides details on the dataset used, model advantages, and limitations of each study.

Table 1. Recent Research works on KD Prediction

Ref	Dataset	Model	Inference	Limitation
[7]	US kidney images from Kaggle	Novel Deep CNN	The Novel Deep CNN model outperforms other architectures in detecting renal cell hydronephrosis. Using ADAM optimizer, data augmentation, and transfer learning significantly enhances classification performance.	The further optimization strategies such as Stochastic Gradient Descent or Adagrad could be explored to improve model accuracy and convergence rate.
[8]	US images from the hospital	ResNet34 + texture features	The proposed model outperformed senior physicians in diagnosing Chronic KD, especially in early-stage detection.	Variability in US machines and regional differences in renal image parameters may affect performance.
[9]	US images from hospital	Multimodal DL Model	The multimodal DL model significantly outperformed single-mode DL models and clinical models in predicting early fibrosis in Chronic KD patients.	The model only focused on early fibrosis, excluding moderate and severe cases. Larger datasets are needed for robust validation.
[10]	MRI images from the Hospital	U-Net	Proposed kidney volume measurement performs comparably to medical professionals. Axial-section images yield more accurate and consistent results than coronal-section images.	Limited dataset of only 40 individuals; larger datasets are needed for validation. Measurement bias due to image orientation differences.
[11]	CT images from the Hospital	Inception-localization, DeepLab+Xception-segmentation, Decision Tree	The proposed models effectively localize, segment, and estimate kidney volume.	Challenges remain in handling complex cases with severe abnormalities such as liver cysts and significant kidney shape variations.

Ref	Dataset	Model	Inference	Limitation
[12]	Kidney CT image from Kaggle	Hybrid CNN combining ResNet101 with a custom CNN	The proposed approach significantly outperforms standalone models and provides a robust, precise, and efficient solution for automated KD diagnosis	Computational complexity and resource requirements may also limit deployment in real-time clinical settings.
[13]	Clinical dataset	Hybrid CNN-SVM	The proposed hybrid CNN-SVM model improves Chronic KD detection by addressing overfitting and class imbalance issues.	Feature selection techniques could be explored further to reduce computational complexity.
[14]	CT scans and medical records	Ant Colony Optimization (ACO) + DenseNet, + LSTM	ACO enhances feature selection, improving DL classification performance. DenseNet with ACO and LSTM achieved the highest accuracy, demonstrating superior feature extraction and classification capabilities.	The computational complexity of ACO may make real-time application challenging. The study requires validation on larger and more diverse datasets for generalizability.
[15]	MRI images from Zenodo	Transformer based EfficientNet	EfficientNet showed the highest accuracy in classifying Chronic KD and healthy images. The study highlights the importance of early Chronic KD diagnosis, and improving treatment planning.	Transformer-based models like EfficientNet performed well but may have high computational costs.
[16]	CT images from Kaggle	Ensemble Transfer Learning and Bayesian Optimized KNN	Feature extraction from multiple pre-trained DNNs combined with ML classifiers improves kidney stone detection.	The model combines the three-transfer learning model so it is highly complex. Design only to detect kidney stones, other disease cant identified

From the literature survey, it is observed that existing research has several limitations. Many models suffer from high complexity, insufficient data availability, and suboptimal accuracy due to the complex nature of US and CT kidney images. Considering these challenges, this research proposes a hybrid CNN-Transformer network to overcome these issues. The key contributions of this study are as follows:

- US kidney images are not widely available online. Therefore, US images of both healthy and diseased kidneys were collected from a diagnostic center. The collected images were minimal, and DL models generally require a large dataset for better performance. To address this limitation, geometric augmentation techniques were applied, increasing the dataset size from 100 to 300 images.
- Popular DL models like DenseNet and ResNet struggle to capture global features, leading to reduced accuracy. The integration of a Transformer module within the CNN network enables the extraction of both local and global features, enhancing prediction accuracy.
- Despite augmentation, the dataset remains relatively small (300 images). The Transformer component with pre-trained embeddings helps achieve better accuracy even with limited data.
- The proposed hybrid network is compared with existing DL models, including DenseNet and ResNet, using both positive and negative performance metrics based on confusion matrix values.

The research paper is organized as follows: Section I discusses the significance of automated KD prediction using DL models and reviews recent research in this domain. Section II details the architecture and working principles of the DL models used in this study. Section III explains the experimental setup, data acquisition and preprocessing steps, and evaluation of DL models for KD prediction. Section IV presents the conclusion of the study and outlines future research directions.

DEEP LEARNING MODELS

For KD prediction from US images, a hybrid DL model, CNN-Transformer, is designed. To evaluate the effectiveness of the proposed model, popular DL models such as DenseNet and ResNet are used for comparison. The working of all three models are detailed in this section, along with their architectural diagrams.

A. ResNet

ResNet was created by He et al. in 2016 [17]. ResNet-50 is a sophisticated image classification model that can train on enormous datasets while producing state-of-the-art results. Figure 1 depicts the residual block. The network is straightforward to optimize, and employing this residual block allows for improved accuracy at that depth. ResNet is a kind of the CNN topology. The CNN consists of three fundamental layers: convolutional layers (CL), pooling layers (PL), and fully connected layers (FCL) [18]. Normalization, padding, and ReLU activation are some of the strategies used to improve CNN accuracy. One advantage of a residual network is its ability to mitigate the vanishing gradient problem. Residual blocks address this issue by allowing gradients to flow directly to earlier layers without excessive multiplications. These blocks include a shortcut connection that functions as an identity mapping, facilitating the effective training of multiple layers. ResNet-50 processes data through 50 layers and is structured into four major phases. In the first phase, the network comprises three residual blocks, each containing three layers.

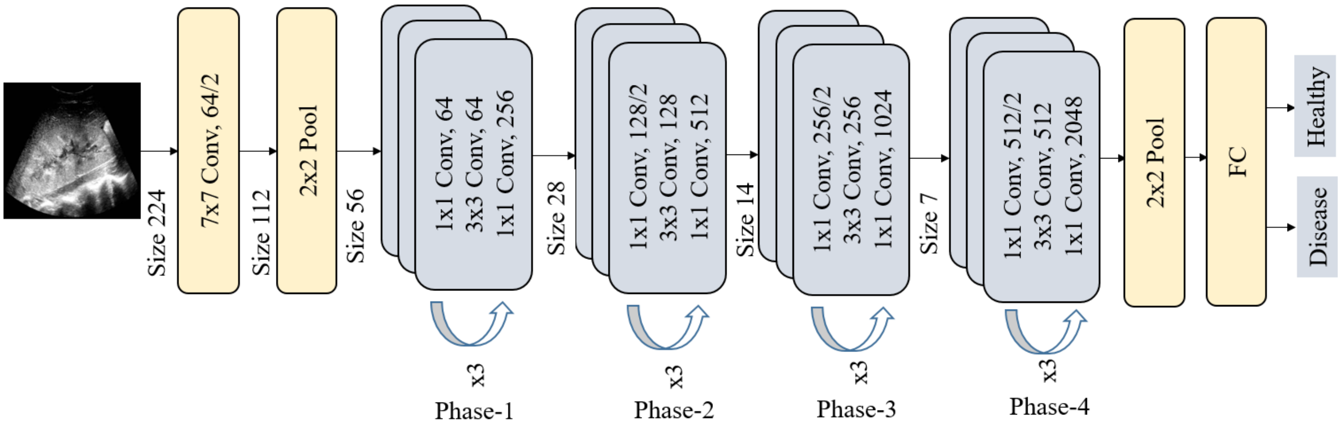


Figure 1. ResNet Architecture

The ResNet architecture requires a CL as the initial step for all inputs. In mathematics, convolution refers to the operation of applying the result of one function to another. To generate a feature map from an input image, convolution utilizes the output function. This method involves considering the values of the input image when selecting a small-number matrix, often known as a filter. Equation (1) provides a method for determining the convolution technique, where f denotes the input image and h denotes the utilized kernel. The indices for rows and columns are represented by m and n, respectively.

$$G[m,n] = (f * h)[m,n] = \sum_j \quad \sum_k \quad h[j,k]f[m-j,n-k] \qquad [1]$$

A ReLU is employed in several CLs to speed up and improve the effectiveness of model learning. The ReLU function is shown in Equation (2):

$$f(x) = max(0,x) \qquad [2]$$

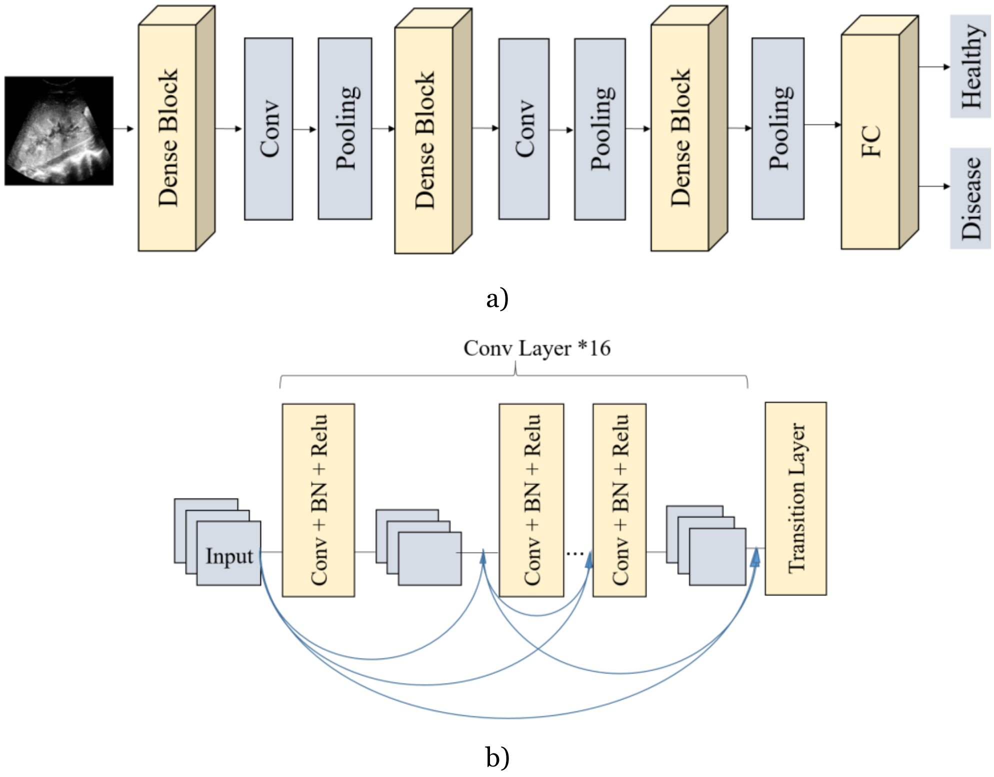
The PL follows the CL and the ReLU. This layer consists of a filter that moves across the feature map area with a predetermined stride size. The PL reduces an image's spatial dimensions and parameter count to enhance processing efficiency. Max pooling selects the highest value in the region at each filter update. Each architecture has a different number of layers at later stages. The second phase of ResNet-50 has four layers with kernel sizes of 128. The third phase of ResNet-50 has six layers with kernel sizes of 256. In the final stage, the architecture comprises three layers with 512 kernels each. An FCL with 1,000 neurons, corresponding to ImageNet output classes, follows the network's average PL. As the name suggests, average pooling reduces the image size by computing the average value of the defined region. The FCL links every activated neuron from one layer to the next. It receives the feature map generated by the CLs and PLs and produces the final output.

$$h(x) = f\left(b + \sum_i \quad w_i x_i\right) \qquad [3]$$

ResNet-50 is chosen as a bottleneck architecture. Each residual function consists of three layers, which utilize 1×1 convolutions. The 1×1 CLs first reduce the dimensions and then restore them. The 3×3 layers act as a bottleneck due to their smaller input and output sizes. In the ResNet architecture, skip connections occur every three layers. Convolutional operations in the residual blocks are performed with a stride of 2 when transitioning from one stage to the next, reducing the input dimensions (height and width) by half [21].

B. DenseNet

In deep networks, network characteristics are constantly subjected to linear-nonlinear synthesis computing; the more expressive the features created, the better the model predicts. However, the concatenation effect of the gradient in backpropagation during deep network training can easily cause issues such as gradient explosion or vanishing, where the gradient becomes either too large or too small, preventing the deep network from completing the training process. DenseNet [22], a densely connected network, takes a different approach to network performance enhancement compared to ResNet and Inception. It accomplishes this by significantly reducing the number of parameters, effectively resolving issues such as gradient vanishing in deep network training, and improving feature propagation through densely connected feature reuse [23].



a)

b)

Figure 2. a) Arhitecture of DenseNet b) Dense Block

DenseNet is made up of several Dense Blocks, each of which has its own structure, as seen in Figure 2.a. To ensure optimal information flow, all layers are directly connected via the Dense Block. The network in the L layer of the Dense Block contains $L(L+1)/2$ connections, whereas the L layer of a standard CNN has L connections. Assume a network with L layers propagates an image x_0. Each layer's nonlinear transformation function is H_i, which is a mixture of batch normalization, activation, and convolution functions. In this scenario, i represents the number of layers, while x_i denotes the output of the i −th layer. To facilitate information transfer between layers, layer i will be fed the feature maps of all previous layers $[x_0, x_1, \ldots, x_{i-1}]$. This means that the input is a concatenation of all preceding layers' feature maps, and the output x_i is created using the H_i nonlinear transform:

$$x_i = H_i([x_0, x_1, \ldots, x_{i-1}]) \quad [4]$$

Because the dimension of the feature map in a CNN varies, the dense connection within the Dense Block requires that it remain consistent. DenseNet is divided into many Dense Blocks, as shown in Figure 2.b, and a transition layer modulates the number and dimension of feature maps between them.

Because DenseNet is densely connected, each layer's output feature maps can be used as input by subsequent layers, boosting feature transfer and allowing the network to make better use of shallow features. Furthermore, by eliminating the gradient vanishing, this connection enhances gradient transfer efficiency and aids in the training of deeper networks. Additionally, the slower growth rate requires fewer network parameters, reducing the likelihood of the network model overfitting [24]. State-of-the-art results are obtained on a variety of datasets, with DenseNet outperforming other models with fewer parameters and computations. In trials, DenseNet may scale up to hundreds of layers without encountering optimization difficulties such as overfitting or failure to converge.

C. CNN–Transformer Hybrid Model

The CNN-Transformer hybrid DL network is proposed for KD prediction from US images. Traditional CNN-based or transfer learning models like DenseNet and ResNet primarily extract local features from images, which can lead to suboptimal classification performance. To overcome this limitation, the proposed hybrid network incorporates a Transformer module that effectively captures global features. The combination of both local and global feature extraction enhances classification accuracy, making the network more robust for KD prediction.

The architecture of the proposed hybrid CNN-Transformer is given in Figure 3. First, several CLs are utilized in the proposed network to analyze the input data and extract local features. The input feature dimensions are progressively expanded through these CLs, enhancing feature representation across different channels. ReLU activation and batch normalization ensure nonlinear feature representation and stable learning. To further optimize feature extraction, max pooling is applied to reduce the dimensionality of the feature map, improving computational efficiency. Next, a Transformer layer is introduced to capture global features using a multi-head self-attention mechanism (AM). Before passing through the Transformer encoder layers for feature extraction, the data processed by the CLs is transformed into the appropriate input format for the Transformer. To improve network stability and information flow, residual concatenation is applied to the Transformer output. Finally, a FCL processes the extracted features, and the output layer classifies the samples, distinguishing between healthy and unhealthy cases.

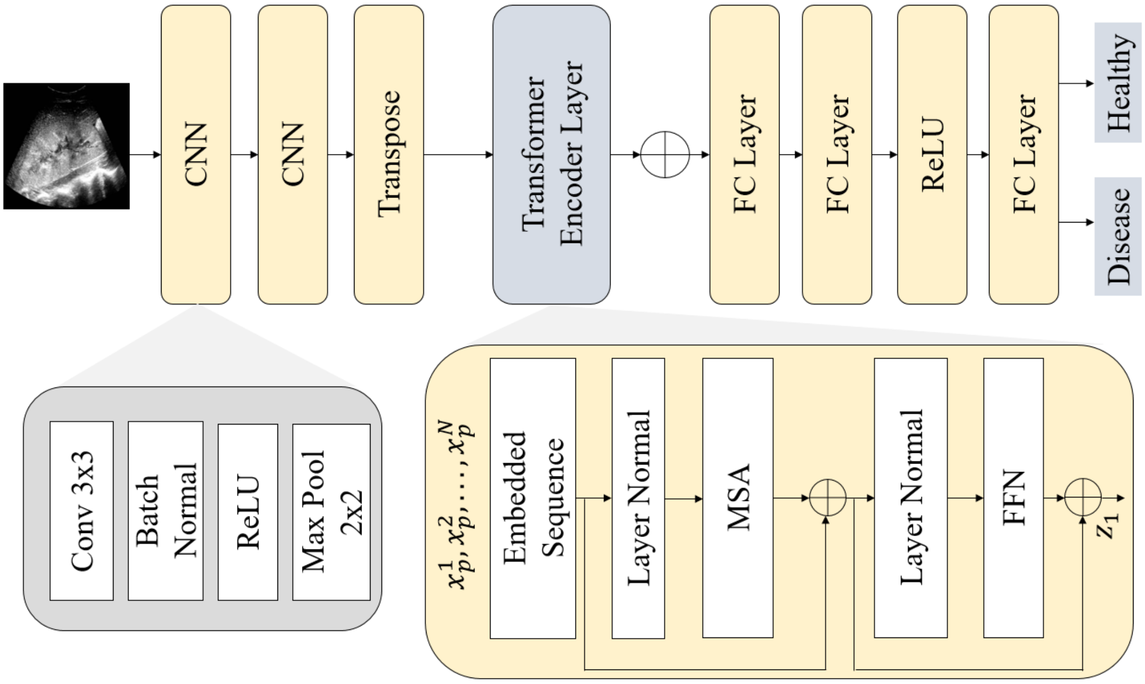


Figure 3. Architecture of the proposed Hybrid CNN-Transformer Network

CNN Feature Extraction

The CNN module extracts local features from the inputs and converts them into high-level features. The module's initial convolution operations on the image include two CLs and a max PL. Let $X \in R^{N*C*L}$ be the input sequence, where N is the batch size, C is the total channel, and L is the sequence length. The first CL extracts the key local features of the US image. Equation (5) describes the output characteristics of the l-th layer, denoted as $X^{(l)}$.

$$X^{(l)} = ReLU\left(Conv\left(X^{(l-1)}, W^{(l)}\right)\right) \qquad [5]$$

The input of the l-th layer is $W^{(l)}$, with a stride of 1, kernel dimension of 3, and a padding of 1. The network introduces nonlinearity through the ReLU activation function [27]. To generate a set of feature maps that emphasize important features such as edges and textures—essential properties for KD recognition—the network applies successive CLs to extract local information from the input image. A max-pooling technique is then used to retrieve the highest value from all the local regions, reducing the dimensions and improving the robustness of the features. The pooling function is represented in Equation (6).

$$X_{pool}^{(l)} = MaxPool\left(X^{(l)}\right) \qquad [6]$$

By reducing the feature size while preserving essential properties, the pooling technique lowers computational costs and enhances the model's generalizability. The CNN generates a feature map with dimensions $X_{cnn} \in R^{N*C*L'}$ after two layers of convolution and pooling, where F represents the total channels and L' denotes the pooled sequence length.

Transformer Module

The transformer module employs a CL to extract local features, which are then used to detect global dependencies in US images. Initially, the input sequence is passed through embedding and positional encoding layers. This phase is crucial for accurately capturing the temporal variation patterns in US images, forming a solid foundation for subsequent feature extraction and dependent modeling. The Transformer encoder, composed of multiple stacked encoder layers, processes the images after they have been embedded and positionally encoded. Each encoder layer consists of two primary components: a feed-forward neural network (FFNN) and a self-attention mechanism (self-

AM). The Self-AM captures interactions within the input sequence, while the FFNN refines these interactions, improving the network's ability to retrieve nonlinear relationships. To optimize gradient flow, accelerate training, and ensure efficient information transfer throughout the deep network, Layer Normalization (LN) and residual connections (RC) are incorporated after each encoder layer. Notably, the input and output dimensions of each encoder layer remain consistent, minimizing issues related to missing data or dimensional mismatches. The key components of the transformer encoder include the AM, FFNN, RCs, and LN. The specific process consists of three steps:

Step 1: The transformer is centered around the self-AM, which is designed to identify global features [31]. Given a feature dimension, let $X_{cnn} = \left[x_p^1, x_p^2, \ldots., x_p^N\right]$, where x_p^i represents the input feature vector at the i-th time step. At first, the query (Q), key (K), and value (V) matrices are generated through the following linear transformations:

$$Q = X_{cnn} . W_Q \qquad [7]$$

$$K = X_{cnn} . W_K \qquad [8]$$

$$V = X_{cnn} . W_V \qquad [9]$$

The learned parameter matrices W_K, W_Q, and W_V represent linear mappings of K, Q, and V, respectively. The self-AM outcome is calculated by the scaled dot-product attention network [32]. Before dividing the outcome by $\sqrt{d}$, where d is the query and key vector's dimension, the dot product of the Q and K matrices is computed. This scaling factor helps to ensure gradient stability. The attention weights are then computed by applying the softmax to the scaled results. Finally, as shown in Equation (10), the outcome of the self-AM is calculated by multiplying the weights by the V.

$$Attention(Q, K, V) = softmax\left(\frac{Q.K^T}{\sqrt{d}}\right)V \qquad [10]$$

The transformer network utilizes a multi-head AM to extract features from multiple subspaces of the input sequence. The network can collect data from various subspaces by processing multiple attention heads simultaneously, and improve its learning capability. The formula of multi-head AM is given in Equation (11):

$$Multi - Head\ AM(X_{cnn}) = Concat(head_1, head_2, \ldots, head_h). W_O \qquad [11]$$

Where, W_O represents the linear projection matrix, and $head_i$ represents the computation result of the ith head, specifically

$$head_i = Attention(Q_i, K_i, V_i) \qquad [12]$$

Step 2: The multi-head attention self-output is followed by RCs and LN. By adding the X_{cnn} to the outcome of the multi-head self-AM, the vanishing gradient issue is successfully mitigated. LN, which ensures consistent feature distribution and improves network training stability [33]. The transformer encoder module frequently employs this RC design to enable efficient gradient propagation. The specific process is described in Equation (13):

$$Z_{Multi-Head\ AM} = LayerNorm\left(X_{cnn} + Multi - Head\ AM(X_{cnn})\right) \qquad [13]$$

Step 3: To enhance the representation of non-linear features, the next component is the FFN, which has two FCLs and a ReLU. The formula for computing the FFN is given in Equation (14):

$$FFN(Z_{Multi-Head\ AM}) = ReLU(Z_{Multi-Head\ AM}. W_1 + b_1). W_2 + b_2 \qquad [14]$$

The bias terms are b_1 and b_2, while the weight matrices are W_1 and W_2. The outcome Z of the encoder layer is obtained by passing the FFN's output through residual connectivity and LN, as shown in Equation (15):

$$Z = LayerNorm\left(Z_{Multi-Head\ AM} + FFN(Z_{Multi-Head\ AM})\right) \qquad [15]$$

The output Z is then fed into the classification module or the next encoder layer. The encoder composed of multiple layers, efficiently encodes the global, temporal features of the input image in the context of KD detection [34]. To differentiate between healthy and unhealthy samples, the classification module must extract rich feature representations.

Fully Connected Network (FCN)

The FCN [35] is responsible for the prediction of the features obtained through transformer and convolution processes. As shown in Figure 3, this module consists of several FCLs, denoted as fc, $fc1$, and $fc2$. The FCN is a common neural network structure composed of multiple FCLs. The fc generates a feature representation of size 128 by linearly transforming the transformer's output features. This layer efficiently fuses the features acquired by the transformer at each time step using weighted combinations, and retrieves the high-level features. The feature dimension is then reduced to 64 by passing the resulting feature vector through $fc1$. This compression preserves important information while reducing extraneous noise or redundant features. A ReLU is applied after the $fc1$ layer, enhancing the network's ability to encode complex patterns and improving its learning of non-linear relationships. After the $fc2$, the features are transformed to match the dimensions of the target category, resulting in a binary outcome: Disease (1) or Healthy (0).

RESULT AND DISCUSSION

A. Experimental Setup

The classification of kidney images as healthy and diseased was performed on Google Colaboratory [36]. The Python programming language was used to access NVIDIA's T4 GPUs [37]. The GPU has a clock rate of 1.59 GHz, which helps to accelerate the processing. It consists of 40 cores, enabling parallel computation, along with 16GB memory for handling large image datasets and a bandwidth of 300GB/sec. The selection of this GPU overall reduces the training time, helps in processing image datasets, and decreases computational complexity. The collected kidney images from the hospital were stored in Google Drive. Google Colab accesses the Drive data for pre-processing and classification.

B. Data Acquisition and Processing

The US kidney images were collected from Aadhar Diagnostic Centre in Maharashtra and annotated by Dr. Nitin Rajaram Potdar, MBBS, DMRD, Consultant Radiologist, who has 18 years of US experience and 5 years of experience with the Army Medical Corps. A total of 50 images from diseased kidneys and 50 images from healthy kidneys were collected. The raw US images are not suitable for direct input into an AI model and require preprocessing. The collected images vary in dimensions. In this research, ResNet, DenseNet, and the proposed network are used, which accept input dimensions of 224×224, 224×224, and 299×299, respectively [38]. Figure 4 shows the actual images and their resized versions with dimensions of 224×224.

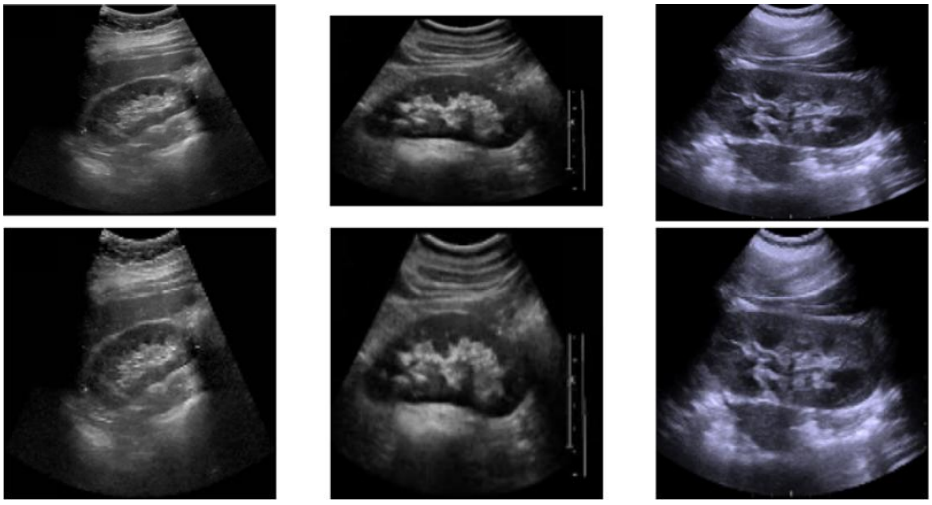


Figure 4. Actual images versus resized images

The first pre-processing step is resizing. Since the number of collected images is minimal, data augmentation is performed to enhance the dataset. The augmentation helps to increase the DL network accuracy [39]. The augmentation steps include rotating by 30 degrees, width and height shifts of 0.2, shear and zoom range of 0.2, and horizontal flipping. The collected 100 images are augmented to 600 images, comprising 300 healthy and 300 diseased images. Figure 5 presents some sample augmented images. Table 2 provides the detailed distribution of the kidney dataset before and after augmentation. It also includes the number of images used for training, validation, and testing.

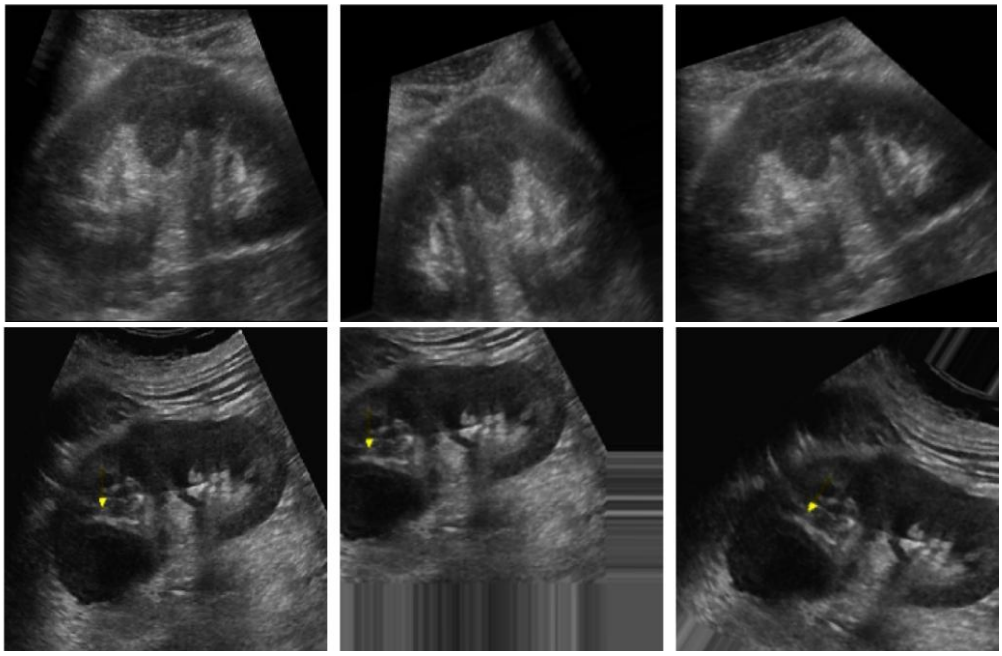
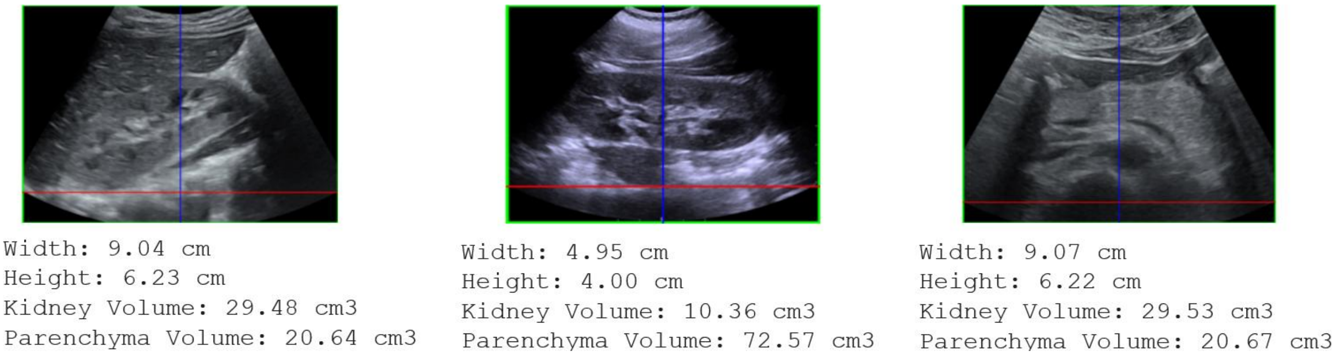


Figure 5. The outcome of augmented images

Table 2. Distribution of Ultrasonic Kidney Image Dataset

Ultrasonic Images	Actual Data	Augmented Data	Train Data	Validation Data	Test Data
Healthy Images	50	300	175	50	25
Diseased Images	50	300	175	50	25

Next, the kidney and parenchyma volumes are detected. The kidney is identified using a bounding box, from which the height and width are determined. These measurements are then used to calculate the kidney and parenchyma volume. Figure 6 illustrates the extracted height and width (in cm) along with the kidney and parenchyma volume (in cm³) from the US images.



```
Width: 9.04 cm
Height: 6.23 cm
Kidney Volume: 29.48 cm3
Parenchyma Volume: 20.64 cm3
```

```
Width: 4.95 cm
Height: 4.00 cm
Kidney Volume: 10.36 cm3
Parenchyma Volume: 72.57 cm3
```

```
Width: 9.07 cm
Height: 6.22 cm
Kidney Volume: 29.53 cm3
Parenchyma Volume: 20.67 cm3
```

Figure 6. Kidney and parenchyma volume detection

Journal of Information Systems Engineering and Management
2025, 10(50s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
Research Article

C. Experimental Outcome

The proposed hybrid CNN-Transformer network is designed for classifying kidney images. For comparison, two other standard DL models, ResNet and DenseNet, were also implemented. All three models were trained and validated using 175 healthy and diseased images and tested on 50 healthy and diseased images. The learning rate was fixed at 0.001, the loss function used was cross-entropy, the performance metric was accuracy, and hyperparameters were kept the same for all models for comparison purposes.

Figure 7 illustrates the performance of the DenseNet during training and validation. The accuracy and loss metrics were chosen to evaluate the performance. In the figure, the training metrics are represented by a dotted line, while the validation metrics are shown as a solid line. The maximum accuracy achieved in training and validation was 0.89 and 0.83, respectively, while the corresponding loss values were 0.28 and 0.63.

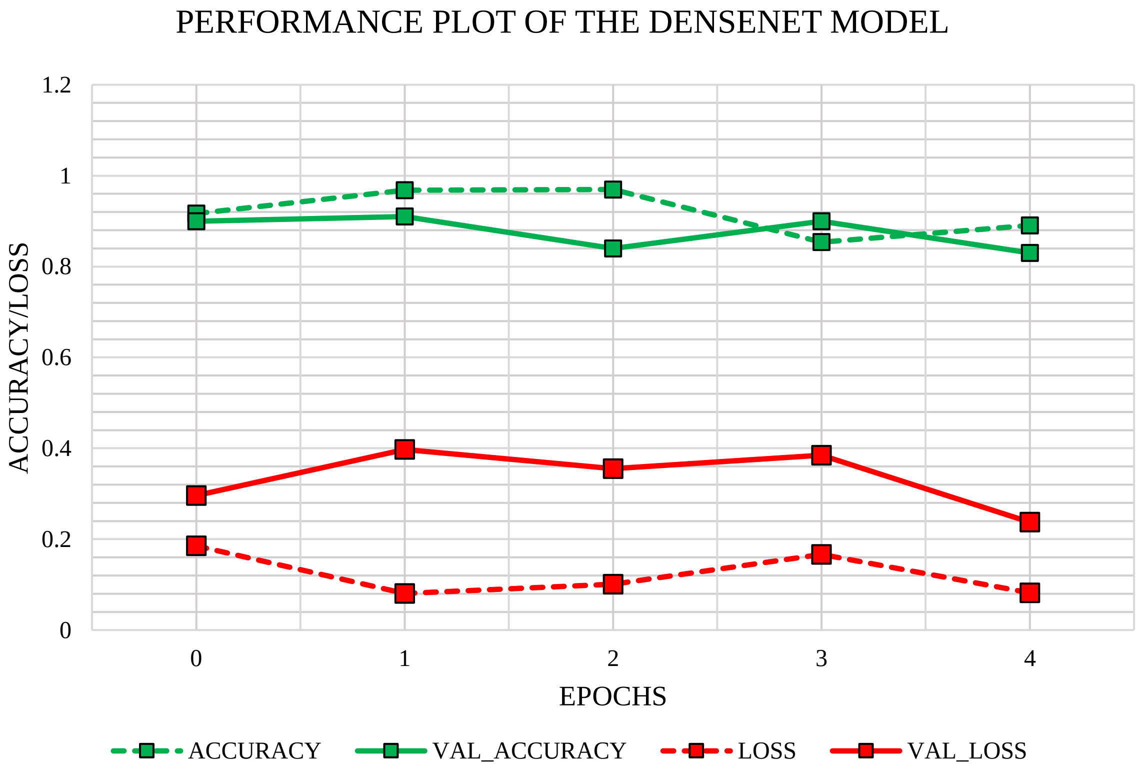


Figure 7. Performance plot of the DenseNet model during the training and validation phases.

Figure 8 presents the performance plot of the ResNet, showing accuracy and loss during training and validation. Similar to the previous figure, the dotted and solid lines represent the training and validation metrics, respectively. The maximum accuracy reached by ResNet was 0.85 for validation and 0.58 for training, while the loss values were 0.60 and 0.58.

Journal of Information Systems Engineering and Management
2025, 10(50s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

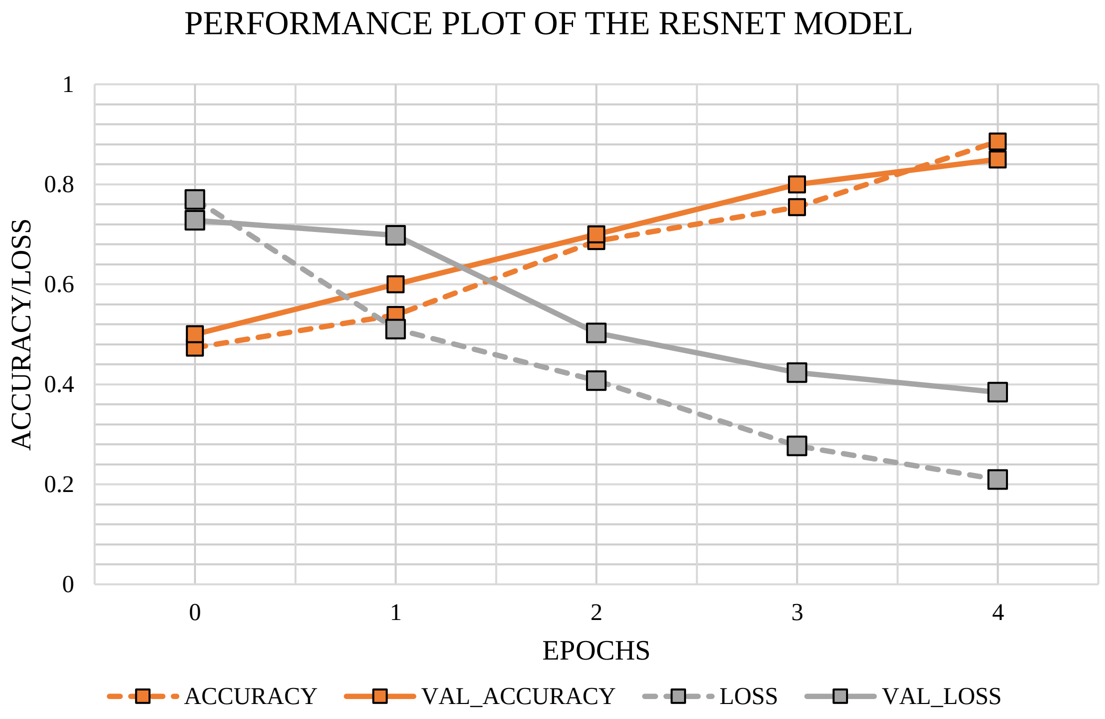


Figure 8. Performance plot of the ResNet model during the training and validation phases.

Figure 9 displays the performance plot of the proposed hybrid model. The proposed network achieved an accuracy of 0.98, with a loss of less than 0.3. When comparing performance plots, the proposed network consistently maintains stable accuracy values, whereas the other two models show fluctuations. This clearly demonstrates the efficiency of the proposed network.

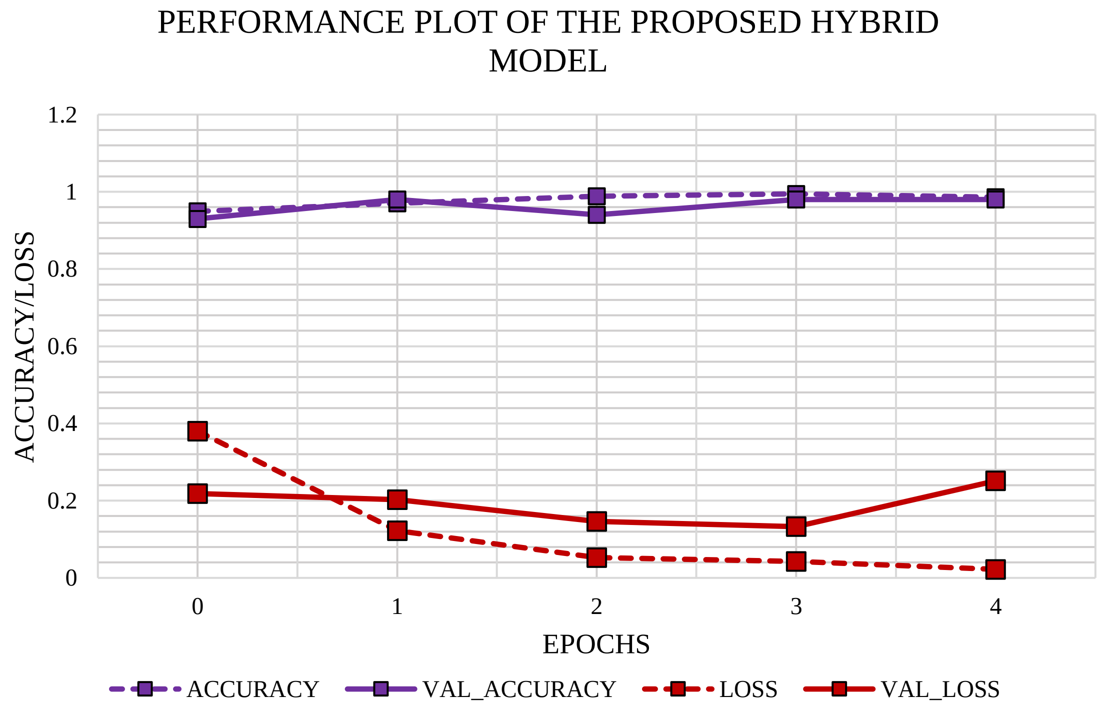
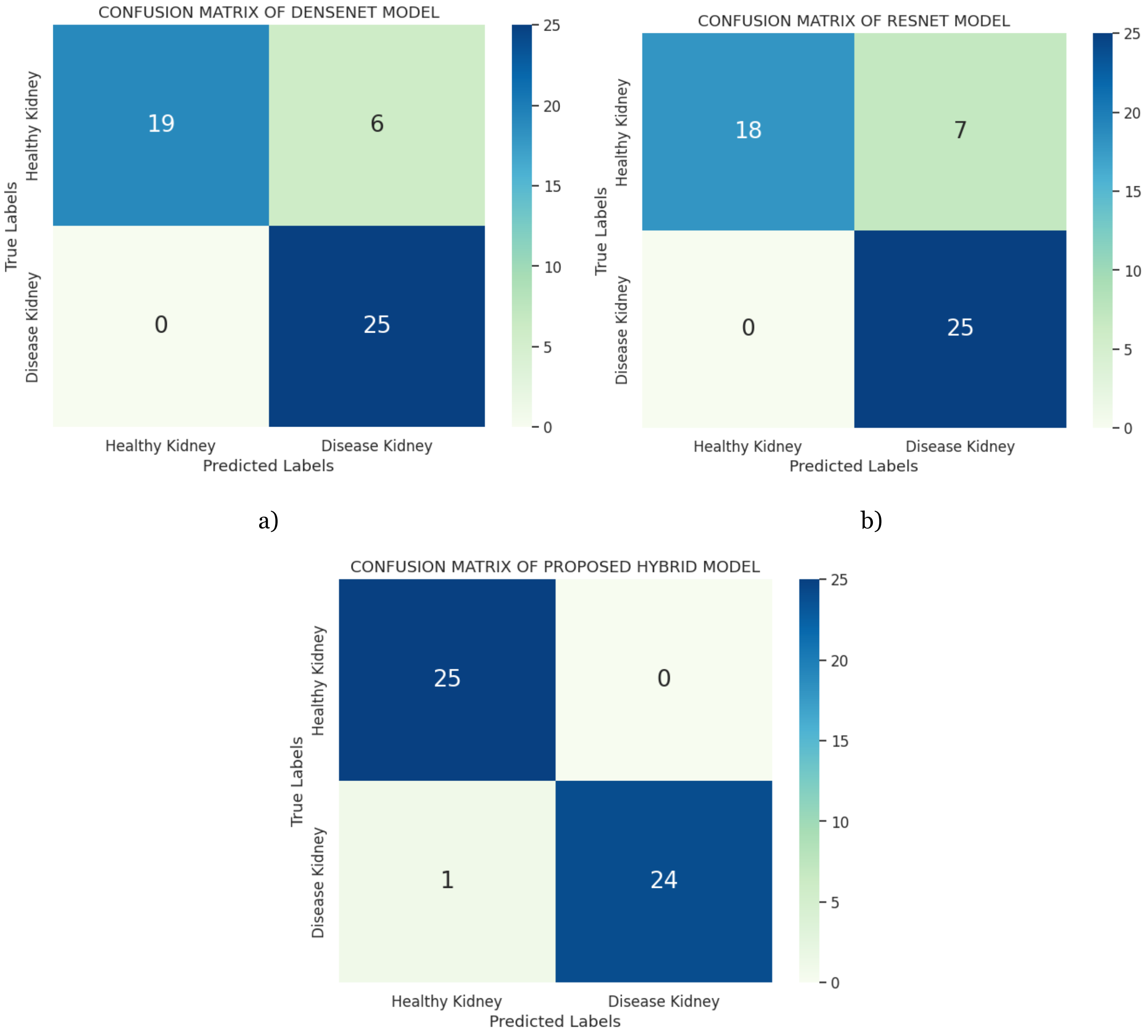


Figure 9. Performance plot of the Proposed network during the training and validation phases.

After training and validation, the DL models were tested using 25 healthy and 25 diseased kidney images. The correct and incorrect classifications of kidney images by DenseNet, ResNet, and the proposed hybrid network are represented in the confusion matrix [40] in Figure 10.

The correctly identified healthy images by DenseNet, ResNet, and the proposed network were 19, 18, and 25, respectively, representing true positives (TP). The correctly identified diseased images were 25, 25, and 24, respectively, representing true negatives (TN). The false positive (FP) and false negative (FN) counts for DenseNet were 6 and 0, for ResNet 7 and 0, and for the proposed network 0 and 1.



a)

b)

c)

Figure 10. Confusion Matrix a) DenseNet b) ResNet c) Proposed Model

Using the values from the confusion matrix (TP, TN, FP, FN), performance metrics such as accuracy, precision, recall, F1-score, false positive rate (FPR), and false negative rate (FNR) were calculated [41]. Table 3 presents the performance metric values obtained by each model along with the formulas used for computation. The accuracy of the proposed network was significantly high, reaching 0.98, while DenseNet and ResNet achieved 0.88 and 0.86, respectively. The precision and recall of the proposed network were 1.00 and 0.96, while its F1-score was 0.97.

DenseNet and ResNet attained F1-scores of 0.89 and 0.87, respectively. In positive performance metrics, the proposed network achieved the highest values, while in negative metrics, it achieved FPR = 0 and FNR = 0.04. The performance evaluation demonstrates the efficiency of the proposed network in identifying KD with high accuracy and reliability.

Table 3. Comparison of the Proposed Hybrid Network Performance Metrics with the Existing Model

Model	Accuracy	Precision	Recall	F1-Score	FPR	FNR
Formula	$\frac{TP+TN}{TP+TN+FP+F}$	$\frac{TP}{TP+FP}$	$\frac{TP}{TP+FN}$	$2*\frac{Precision*Recall}{Precision+Recall}$	$\frac{FP}{TN+FP}$	$\frac{FN}{TP+FN}$
DenseNet	0.88	0.8065	1.00	0.8929	0.24	0.0
ResNet	0.86	0.7812	1.00	0.8772	0.28	0.0
Proposed Hybrid Model	0.98	1.00	0.96	0.9796	0.0	0.04

CONCLUSION

The research aims to develop a reliable DL model for KD prediction from US images. While numerous research has been conducted in this field, achieving the expected level of accuracy remains a challenge. Existing models struggle to capture all essential features from medical images, leading to suboptimal performance. To address this limitation, the study integrates CNN with a Transformer model. CNN effectively captures local features, while the Transformer extracts global features from US images. The combination of both local and global features significantly improves the accuracy of KD prediction. To assess the effectiveness of the proposed model, it is compared with widely used DL models in medical imaging, such as DenseNet and ResNet. All three models are trained, validated, and tested on US images for KD prediction. The proposed network achieves an accuracy of 98%, whereas DenseNet and ResNet obtain 88% and 86%, respectively. This 10% improvement highlights the superiority of the hybrid network in KD prediction and underscores its potential for real-time deployment.

One of the key limitations in existing research is the scarcity of publicly available US datasets. Due to this, the study collects its own dataset from clinical sources. However, due to medical restrictions, only 100 images are acquired. As a result, generalizability testing of the proposed network is limited. This is a crucial factor for real-time deployment. Future research will focus on collecting data from multiple medical centers to ensure diversity across age, gender, and geographic regions. Additionally, network complexity will be analyzed, and a user-friendly website will be developed for real-world applications. The proposed web-based system will allow users to upload US images with a single click, providing immediate kidney health status predictions. This will benefit both patients and medical professionals by facilitating early detection and intervention for KD.

REFERENCE

[1] Kovesdy, Csaba P. "Epidemiology of chronic kidney disease: an update 2022." *Kidney international supplements* 12, no. 1 (2022): 7-11.

[2] Peracha, Javeria, and Smeeta Sinha. "Clinical assessment of renal disease and identification of kidney disease in the community." *Medicine* 51, no. 2 (2023): 89-97.

[3] Yan, Ming-Tso, Chia-Ter Chao, and Shih-Hua Lin. "Chronic kidney disease: strategies to retard progression." *International journal of molecular sciences* 22, no. 18 (2021): 10084.

[4] Kumahor, Elikem Kwami. "The biochemical basis of renal diseases." In *Current Trends in the Diagnosis and Management of Metabolic Disorders*, pp. 185-200. CRC Press, 2024.

[5] Schnuelle, Peter. "Renal biopsy for diagnosis in kidney disease: indication, technique, and safety." *Journal of clinical medicine* 12, no. 19 (2023): 6424.

[6] Aggarwal, Ankita, Chandan J. Das, and Sanjay Sharma. "Recent advances in imaging techniques of renal masses." *World Journal of Radiology* 14, no. 6 (2022): 137.

[7] Islam, Umar, Abdullah A. Al-Atawi, Hathal Salamah Alwageed, Gulzar Mehmood, Faheem Khan, and Nisreen Innab. "Detection of renal cell hydronephrosis in ultrasound kidney images: a study on the efficacy of deep convolutional neural networks." *PeerJ Computer Science* 10 (2024): e1797.

[8] Tian, Shuyuan, Yonghong Yu, Kangjian Shi, Yunwen Jiang, Huachun Song, Yuting Wang, Xiaoqian Yan, Yu Zhong, and Guoliang Shao. "Deep learning radiomics based on ultrasound images for the assisted diagnosis of chronic kidney disease." *Nephrology* 29, no. 11 (2024): 748-757.

[9] Qin, Xiachuan, Xiaoling Liu, Linlin Xia, Qi Luo, and Chaoxue Zhang. "Multimodal ultrasound deep learning to detect fibrosis in early chronic kidney disease." *Renal Failure* 46, no. 2 (2024): 2417740.

[10] Hsu, Jia-Lien, Anandakumar Singaravelan, Chih-Yun Lai, Zhi-Lin Li, Chia-Nan Lin, Wen-Shuo Wu, Tze-Wah Kao, and Pei-Lun Chu. "Applying a Deep Learning Model for Total Kidney Volume Measurement in Autosomal Dominant Polycystic Kidney Disease." *Bioengineering* 11, no. 10 (2024): 963.

[11] Sheng, Ting-Wen, Djeane Debora Onthoni, Pushpanjali Gupta, Tsong-Hai Lee, and Prasan Kumar Sahoo. "Segmentation of ADPKD Computed Tomography Images with Deep Learning Approach for Predicting Total Kidney Volume." *Biomedicines* 13, no. 2 (2025): 263.

[12] Sharma, Kiran, Ziya Uddin, Adarsh Wadal, and Dhruv Gupta. "Hybrid Deep Learning Framework for Classification of Kidney CT Images: Diagnosis of Stones, Cysts, and Tumors." *arXiv preprint arXiv: 2502.04367* (2025).

[13] Ramu, K., Sridhar Patthi, Yogendra Narayan Prajapati, Janjhyam Venkata Naga Ramesh, Sudipta Banerjee, KBV Brahma Rao, and Saleh I. Alzahrani. "Hybrid CNN-SVM model for enhanced early detection of Chronic kidney disease." *Biomedical Signal Processing and Control* 100 (2025): 107084.

[14] Singh, Jagendra, Deepak Sharma, Ch Bhavani, Satyakam Rahul, Manoj Rana, and Jai Prakash Mishra. "Integrating Ant Colony Optimization With Deep Learning for Improved Kidney Disease Diagnosis and Prognosis." In *Nature-Inspired Optimization Algorithms for Cyber-Physical Systems*, pp. 175-192. IGI Global Scientific Publishing, 2025.

[15] Islam, Md Sakib Bin, Md Shaheenur Islam Sumon, Rusab Sarmun, Enamul H. Bhuiyan, and Muhammad EH Chowdhury. "Classification and segmentation of kidney MRI images for chronic kidney disease detection." *Computers and Electrical Engineering* 119 (2024): 109613.

[16] Chaki, Jyotismita, and Ayşegül Uçar. "An efficient and robust approach using inductive transfer-based ensemble deep neural networks for kidney stone detection." *IEEE Access* 12 (2024): 32894-32910.

[17] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770-778. 2016.

[18] Wani, M. Arif, Farooq Ahmad Bhat, Saduf Afzal, and Asif Iqbal Khan. *Advances in deep learning*. Springer, 2020.

[19] Hu, Yuhuang, Adrian Huber, Jithendar Anumula, and Shih-Chii Liu. "Overcoming the vanishing gradient problem in plain recurrent networks." *arXiv preprint arXiv:1801.06105* (2018).

[20] Koonce, Brett, and Brett Koonce. "ResNet 50." *Convolutional neural networks with swift for tensorflow: image recognition and dataset categorization* (2021): 63-72.

[21] Mascarenhas, Sheldon, and Mukul Agarwal. "A comparison between VGG16, VGG19 and ResNet50 architecture frameworks for Image Classification." In *2021 International conference on disruptive technologies for multi-disciplinary research and applications (CENTCON)*, vol. 1, pp. 96-99. IEEE, 2021.

[22] Huang, Gao, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. "Densely connected convolutional networks." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700-4708. 2017.

[23] Li, Guoqing, Meng Zhang, Jiaojie Li, Feng Lv, and Guodong Tong. "Efficient densely connected convolutional neural networks." *Pattern Recognition* 109 (2021): 107610.

[24] Kuang, Ping, Tingsong Ma, Ziwei Chen, and Fan Li. "Image super-resolution with densely connected convolutional networks." *Applied Intelligence* 49 (2019): 125-136.

[25] Zafar, Afia, Muhammad Aamir, Nazri Mohd Nawi, Ali Arshad, Saman Riaz, Abdulrahman Alruban, Ashit Kumar Dutta, and Sultan Almotairi. "A comparison of pooling methods for convolutional neural networks." *Applied Sciences* 12, no. 17 (2022): 8643.

[26] Zhao, Feng, Fan Feng, Shixin Ye, Yanyan Mao, Xiaobo Chen, Yuan Li, Mao Ning, and MingLi Zhang. "Multi-head self-attention mechanism-based global feature learning model for ASD diagnosis." *Biomedical Signal Processing and Control* 91 (2024): 106090.

[27] Banerjee, Chaity, Tathagata Mukherjee, and Eduardo Pasiliao Jr. "An empirical study on generalizations of the ReLU activation function." In *Proceedings of the 2019 ACM Southeast Conference*, pp. 164-167. 2019.

[28] Antipova, Kateryna, and Hlib Horban. "Positional encoding for transformers." *Publishing House "Baltija Publishing"* (2024).

[29] Parmar, Niki, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. "Image transformer." In *International conference on machine learning*, pp. 4055-4064. PMLR, 2018.

[30] Huang, Xiao Shi, Felipe Perez, Jimmy Ba, and Maksims Volkovs. "Improving transformer optimization through better initialization." In *International Conference on Machine Learning*, pp. 4475-4483. PMLR, 2020.

[31] Yang, Jianwei, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. "Focal self-attention for local-global interactions in vision transformers." *arXiv preprint arXiv:2107.00641* (2021).

[32] Zhao, Hengshuang, Jiaya Jia, and Vladlen Koltun. "Exploring self-attention for image recognition." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10076-10085. 2020.

[33] Huang, Lei, Jie Qin, Yi Zhou, Fan Zhu, Li Liu, and Ling Shao. "Normalization techniques in training dnns: Methodology, analysis and application." *IEEE transactions on pattern analysis and machine intelligence* 45, no. 8 (2023): 10173-10196.

[34] Garnot, Vivien Sainte Fare, Loic Landrieu, Sebastien Giordano, and Nesrine Chehata. "Satellite image time series classification with pixel-set encoders and temporal self-attention." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12325-12334. 2020.

[35] Basha, SH Shabbeer, Shiv Ram Dubey, Viswanath Pulabaigari, and Snehasis Mukherjee. "Impact of fully connected layers on performance of convolutional neural networks for image classification." *Neurocomputing* 378 (2020): 112-119.

[36] Sukhdeve, Dr Shitalkumar R., and Sandika S. Sukhdeve. "Google colaboratory." In *Google Cloud Platform for Data Science: A Crash Course on Big Data, Machine Learning, and Data Analytics Services*, pp. 11-34. Berkeley, CA: Apress, 2023.

[37] Voon, Yong Shing, Yunze Wu, Xinzhi Lin, and Kamran Siddique. "Performance analysis of cpu, gpu and tpu for deep learning applications." *Professor Ka Lok Man, Xi'an Jiaotong-Liverpool University, China Professor Young B. Park, Dankook University, Korea Chairs of CICET* 16, no. 12 (2021): 2021.

[38] Saponara, Sergio, and Abdussalam Elhanashi. "Impact of image resizing on deep learning detectors for training time and model performance." In *International Conference on Applications in Electronics Pervading Industry, Environment and Society*, pp. 10-17. Cham: Springer International Publishing, 2021.

[39] Wang, Jason, and Luis Perez. "The effectiveness of data augmentation in image classification using deep learning." *Convolutional Neural Networks Vis. Recognit* 11, no. 2017 (2017): 1-8.

[40] Amin, Fahmy, and M. Mahmoud. "Confusion matrix in binary classification problems: A step-by-step tutorial." *Journal of Engineering Research* 6, no. 5 (2022).

[41] Sathyanarayanan, S., and B. Roopashri Tantri. "Confusion matrix-based performance evaluation metrics." *African Journal of Biomedical Research* (2024): 4023-4031.