



Keywords: kidney failure, improving patients, deep learning, transfer learning, CT image

Abeer ALSHIHA ¹, Abdalrahman QUBAA ^{1*}

¹ University of Mosul, Iraq, abeer.allaf@uomosul.edu.iq, abdqubaa@uomosul.edu.iq

* Corresponding author: abdqubaa@uomosul.edu.iq

Kidney disease diagnosis based on artificial intelligence/deep learning techniques

Abstract

Chronic Kidney Disease is a progressive medical ailment of growing global health importance because, in most cases, this ailment shows no symptoms during its early stages. Improving patients' outcomes and early detection are significant aspects of managing diseases. In this paper, deep learning models to classify the images of kidney diseases are presented based on a dataset of 12, 446 images collected from various renal diseases. Therefore, CNN, VGG16, MobileNet V2, DenseNet 121, and ResNet 50 were the fine-tuned and evaluated models. The training setting was the Adam optimizer, categorical cross entropy loss, and 10 epochs. Hence, the model's performance was measured using the accuracy, precision, recall, and F1-score evaluation parameters. Following that, the current evaluation illustrates that most of the examined models positively predict outstanding accuracies, with ResNet 50 having a maximal validation and test accuracy rate reaching 99.40%. At the same time, MobileNet V2 and DenseNet 121 also boast of their high efficacy. The researchers' works highlighted that deep learning algorithms are very helpful for diagnosing kidney diseases based on medical images, underlining that their application can significantly change early diagnosis and patient treatment.

1. INTRODUCTION

The medical imaging can make a significant contribution to medical practice due to providing accurate visual impressions that enable doctors to reveal the disease at the earliest stage and accurately measure its progression. The analysis of medical pictures has become more effective with the introduction of the machine learning (ML) and deep learning (DL) methods, which help to prove a more precise diagnosis and reduce the risk of human error. Such computational tools also use the big data in medical studies in order to reveal some valuable patterns, thus improving clinical decision-making (Kaddari et al., 2024).

Chronic Kidney Disease (CKD) is one of the key health issues of the entire world as it is characterized by the fall of kidney performance. It is on the rise and owes many of its developments to changes in demographics, like an aging population, and those associated with the rise in the incidence of diabetes, hypertension, and lifestyle changes (Kalantar-Zadeh et al., 2021; Delrue et al., 2024). The increased problem of CKD poses significant health care system problems globally (Charleonnann et al., 2017). The pathology of the disease is also the most problematic, given its insidious process, which usually starts with dubious and nonspecific symptoms, among which timely detection is challenging (Kamal et al., 2024). Outside the clinical implications, CKD can also be economically problematic as most long-term treatments, dialysis and kidney transplantation, in particular, consume resources and are expensive (Dritsas & Trigka, 2022). Though they are life sustaining treatments, the high financial and logistical burdens pose a strong urgency in the early detection of this kind of disease (Mohammed & Al-Hayali, 2024).

Traditional screening methods, such as serum creatinine levels and glomerular filtration rate (GFR), are not very useful in the initial stages of CKD discovery (Sanmarchi et al., 2023). This weakness highlights the necessity of superior diagnostic methods that have the potential to determine patients at the first stages of the disease at which lifestyle changes and therapeutic modifications can be most effective. In this regard, it can be stated that ML strategies are projected to be effective in filling these diagnosis gaps (Ghosh et al., 2020).

The detection and subsequent treatment of CKD in early stages is important since it can change the natural course of the condition. As soon as the disease is diagnosed, preventive measures, including lifestyle change, close surveillance, and the optimization of pharmacological management, can be taken to delay the

development of the disease and improve the outcome (Meddage et al., 2021). Nevertheless, CKD can be described as a silent killer, as its initial symptoms resemble those of other diseases, which causes many cases to be diagnosed later (Shama et al., 2016). Routinely measured standard biomarkers such as creatinine and urine albumin in serum are not suitable to consistently identify mild or moderate renal dysfunction (Iftikhar et al., 2023).

This leads to the situation that most of them only get a diagnosis when the condition has reached an end-stage, requiring end-stage renal disease (ESRD) treatment with dialysis or transplantation, which is costly and physically taxing (Swain et al., 2023). These problems highlight the paramount significance of creating new diagnostic plans. Machine learning offers a good solution, as it allows the identification of the subtle and otherwise invisible patterns in the complex data. In the context of CKD screening, the ML algorithms will be able to enhance the early warning of the at-risk group, allowing a person to intervene earlier and get better results (Meddage et al., 2020).

2. LITERATURE REVIEW

Over the past few years, researchers have paid more attention to the use of imaging, machine learning (ML), and deep learning (DL) in the process of disease diagnosis. Introduction of big data into these strategies has had a significant effect as it has allowed to decrease the rate of diagnostic error as well as increasing the rate of classification. This chapter is a literature review of earlier research, whereby the main emphasis is on kidney disease diagnosis and other biomedical uses.

Taznin et al. (2016) addressed a classification problem with the help of a decision tree classifier, which was also impressive with the accuracy of 99. The ability of decision tree to create predictive rules that are capable of successfully capturing relationship in the data can be cited to the success of the model. In the same manner, Amirgaliyev et al. (2018) used the Support Vector Machine (SVM)-based classifier and received a classification rate of 94%. SVM has traditionally been shaped as strong in addressing complicated classification problems particularly where data points cannot be separated linearly as it is able to determine the optimal separating hyperplanes in more dimensions.

Models based on neural networks also have an impressive level of performance. Yildirim (2017) used a sampling algorithm to optimize the multilayer perceptron (MLP) with the highest accuracy, precision, recall, and F1 scores of 99. The research paper demonstrates the effectiveness of neural networks to model complex trends as well as the importance of sampling methods in maintaining a balanced number of classes. In a test by Wibawa et al. (2017) on a K-nearest neighbors (KNN) classifier (K=5) the authors obtained 98% of accuracy, precision, recall, and F1-score. Their results point to the effectiveness of KNN in processing the datasets whose similarity is determined by proximity to show class membership.

SVM has been also confirmed in study by Polat et al. (2017), which showed accuracy of 98 percent showing that the model can be adapted to different classification contexts. Ensemble techniques were also responsible. Salekin and Stankovic (2016) used a Random Forest (RF) classifier with the F1-score of 99%. The high accuracy and the generalizability of the model were due to ensemble design of RF which combines the predictions made by various decision trees.

Image-based diagnostics is finding favour with deep learning Convolutional neural networks (CNNs). To have a more quantitative approach, Manonmani and Balakrishnan (2020) used a CNN to classify their dataset, resulting in an accuracy of 95% and an 96% F1 score, which proves that the algorithm has the ability to find hierarchical information in complex data. Equally, Rubini and Perumal (2020) implemented a Multi-Kernel SVM (MK SVM) and obtained 98 percent accuracy indicating that Multiple-kernel arrangement allows the model to emulate a variety of data properties compared to single-kernel arrangements.

Emon et al. (2021) once again confirmed the usefulness of the RF with 99% test accuracy, which further confirms the strength of RF when dealing with big data, as well as the decrease in the risks of overfitting. Logistic regression (LR) too has been used and achieved good results; Gupta et al. (2020) stated the accuracy of LR to be 99, a precision of 98, recall of 100, and a recognition of 82 which prove that LR is appropriate in classification tasks with high stakes and cannot allow any false positive. On the same note, Gunarathne et al. (2017) implemented an MDF classifier that achieved an accuracy of 99 percent indicating that it could be able to operate with datasets of multiple classes that possessed multiple outcome variables.

Other biomedical studies other than the research on kidney disease have also portrayed the creativity of ML and DL. Karpiński (2022) evaluated the vibroarthrography (VAG) as a non-invasive diagnostic tool of OA

with radial basis function (RBF) and MLP neural networks where accuracy of 90% was achieved with a sensitivity of 88.5% and specificity of 91.7%. Continuing on this work, Karpiowski et al. (2023) assessed VAG in 84 patients with cartilage lesions, where MLP was highest in terms of accuracy (89.3) and RBF in terms of balanced sensitivity (82.2) and specificity (82.1). On the contrary, the naïve Bayes classifier (NBC) was less effective (70.2%). VAG analysis was further furthered by Machrowska et al. (2024) who introduced ensemble empirical mode decomposition (EEMD) and detrended fluctuation analysis (DFA) to extract features, which were trained and learned through an artificial neural network (MLP). Their approach had 93 percent accuracy with both sensitivity and specificity to equal, and an AUC of 0.942, which makes it a potential low-cost diagnostic instrument.

In the meantime, Na and Kim (2024) created a multi-modal deep learning model, which is also called 4bay, which combines visual and motion data to categorize VR sickness. Their framework was superior to previous fusion-based approaches, which shows how a combination of multiple sources of sensory data can be used in the healthcare field in new fields of knowledge. Kaddari et al. (2024) have also examined how ChatGPT can be used to extract clinical information in French neonatal reports; they discovered that schema-based extractors were much better than zero-shot systems when handling complex medical language. Their paper brings to the fore the possibilities and constraints of large language models (LLMs) in clinical settings.

Lastly, a review of the applicability of ML in malaria control was done by James and Osubor (2025). They noticed there was a tendency towards deep methods of learning due to the increasingly rich data sets. Nevertheless, they identified the ongoing challenges, such as the absence of methodological standardization, and required real-time integration approaches and robust deployment models to leverage the input of ML to the world health endeavours to the fullest.

3. MATERIALS AND METHODS

The process begins with the preprocessing of the dataset, which includes three crucial steps. These include label encoding, class joining, and image resizing. Hot encoding is used to convert other categorical labels of images related to kidney disease into a numerical format for use in model training. Class join combines multiple classes of images from different sources or subclasses and makes them all uniform. Image resizing ensures that all images are the same size, 224 pixels by 224 pixels to be exact, to reduce the computational requirements of deep learning models.

After preprocessing, the dataset is divided into training and test subsets at a ratio of 70:31, which is ideal for the matching function. This allocation allows for sufficient data for model learning while providing a strong evaluation set. Several deep learning architectures are considered for model selection: five pre-trained models, including Convolutional Neural Network (CNN), VGG16, MobileNet V2, DenseNet 121, and ResNet 50, are used to classify the images.

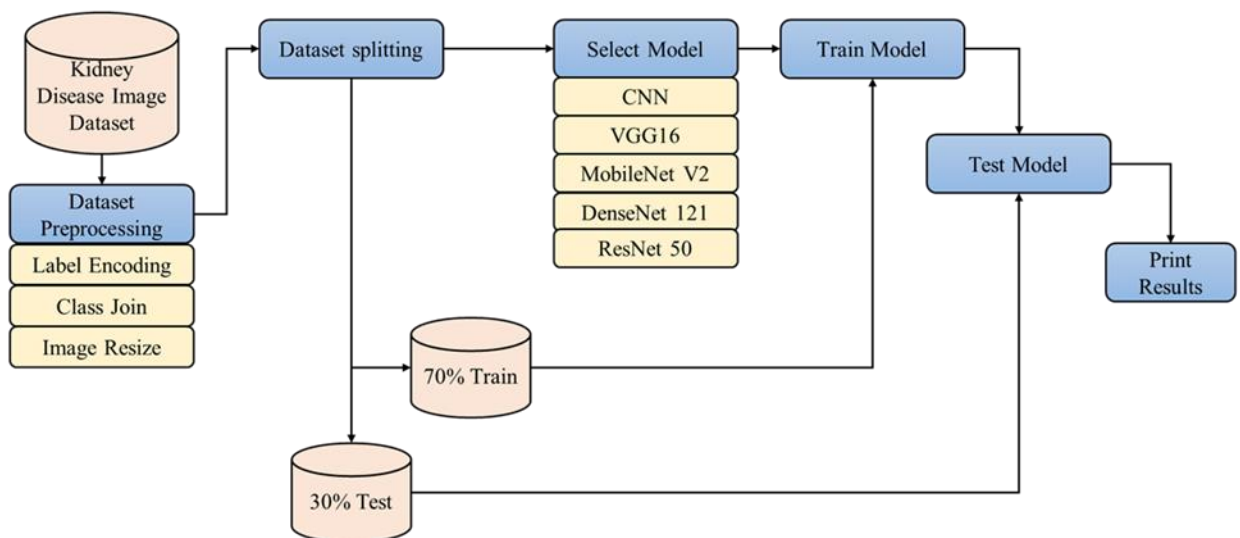


Fig. 1. Model methodology

Each model is trained on the training subset using additional data, variations, an optimal learning rate, and even stopping training before it overfits. Training is a process of adjusting the weights of the model to minimize the loss function using optimizers such as Adam, SGD, and many others. After training, the results of the models on the test subset are computed. Prediction accuracy rates or precision, recalls, and F1 score are measured to assess the overall generality of the predictions. The results are printed and discussed to compare the models for an interview on kidney disease image classification with results with an outline of the advantages and disadvantages of each model. Figure 1 describes this methodology.

3.1 Dataset description

This study used a dataset which was collected in Picture Archiving and Communication systems (PACS) of several hospitals in Dhaka, Bangladesh. The sample consisted of medical images of patients with stable renal tumors, renal cysts, with kidney stones and normal kidney results. Coronal and axial sections were also represented, with contrast-enhanced and non-contrast sections covered based on protocols on imaging of the whole abdomen and urograms. In both instances, suitable studies were chosen to create DICOM batches depicting a certain diagnosis and radiologic outcome.

The DICOM files were stripped of all identifiable data and metadata to make sure that patient confidentiality was guaranteed. The de-identified pictures were then transformed to lose-less JPG format. This was followed by a serious process of validation where a radiologist and medical technologist separately reviewed every image to ascertain accuracy and reliability. The completed dataset had 12, 446 unique images and the distribution of images among different diagnostic conditions was as follows cysts (3,709 images), normal cases (5,077 images), kidney stones(1,377 images), and tumors (2,283 images). Such filtered data set offered a strong baseline on which the deep learning models can be trained and tested about classifying kidney disease with a strong background of verified and objective medical data.

Since the dataset had a loss of class imbalance (through the underrepresentation of cases of kidney stone), the Synthetic Minority Over-sampling Technique (SMOTE) was used. SMOTE operates by artificial sampling of the minority group by interpolating existing points of data to provide more artificial samples of this population. This method minimized the biases of the majority group and enhanced the capacity of the model to perform the right identification of minority cases. The solution is that the SMOTE increased the sensitivity and accuracy of the classification system, as well as distribution in the detection of kidney stones in particular, and helped the predictive model have a greater overall effect..

3.2. Used models

3.2.1. CNN model

The CNN architecture in Figure (2) intended for the image classification of kidney disease comprises several layers arranged successively. Starting the model, it has a 2D convolutional layer with 32 filters of size 3x3 applying the ReLU activation function. The input shape of the model is 224x224x3 to accept the images resized to this size. After that, there is a max-pooling layer with a pool size of 2X2 to cut down the spatial dimension for features dominating (Al-Neama et al., 2023). Following this, another convolutional layer with 64 filters of size 3X3 and ReLU activation is added followed by another max-pool layer. These layers assist in determining extensive characteristics and relations from the images. Based on the convolutional layers' output, the data is flattened and passed to a dense layer containing 128 neurons and ReLU activation to learn the high level of representation. Finally, the output layer consists of a dense layer with several neurons equal to the number of classes (four in this case: cyst, normal, stone, tumour), and the last layer consists of a softmax activation function to predict the probability distribution of the classes. This architecture minimally encodes the spatial structure present in the input data and is well suited for classification (Al-Qubaa et al., 2014).

3.2.2. VGG16 model

To fine-tune VGG16 for classification on kidney disease images, some changes were made to the model while using it for feature extraction. The base model VGG16 is used (Figure 3), which was trained on ImageNet and did not have the final layers (include_top=False); the input shape of the model was set to 224 x 224 x 3 = to match the image data. All layers of the base VGG16 model were made non-trainable (layer.trainable=False) to use the pre-trained weights of the network for classification (Simonyan & Zisserman, 2015).

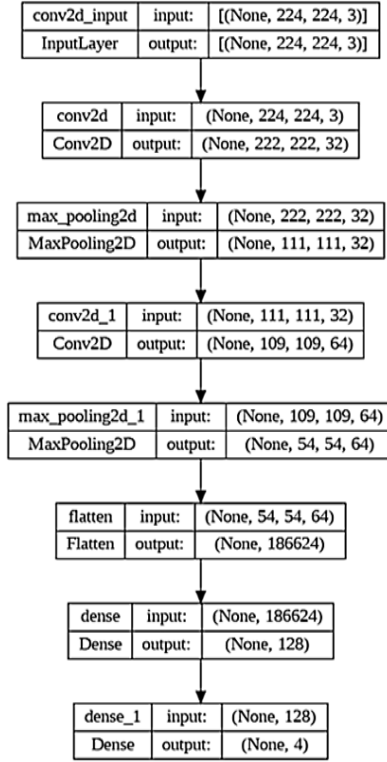


Fig. 2. CNN Model Architecture

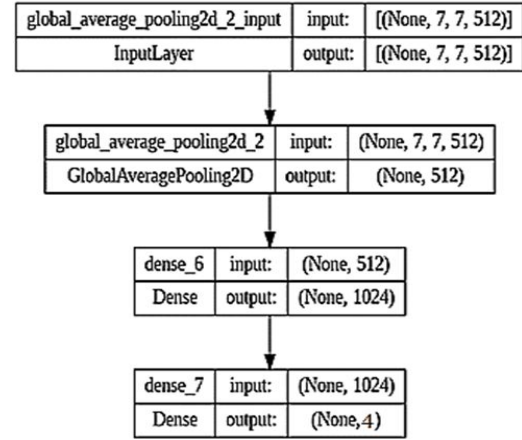


Fig. 3. Modified VGG16 Parts

This avoids distorting the learned feature extractors of VGG16, while achieving the benefits of a multimodal model. Additionally, a new fully connected classifier was created on the base model frozen during the previous steps. The output of the base model was followed by a Global Average Pooling 2D layer to reduce the size of the feature maps to avoid overfitting. The next layer was a fully connected layer with 1024 neurons and a ReLU activation function to introduce nonlinearity into the network and allow learning of complex patterns characteristic of the applied kidney disease dataset. Finally, an output layer with 7 neurons using softmax was used to obtain the probability distribution of the classes (Simonyan & Zisserman, 2015).

3.2.3. MobileNet V2 model

Therefore, in order to make the chosen pre-trained MobileNetV2 model suitable for the kidney disease image classification approach, some adjustments were made: The base MobileNetV2 model pre-trained on the ImageNet datasets is introduced, but the top layers ('include_top=False') of the model are removed and the input size of the images for the model is set to 224x224x3, which follows the standard protocol on the datasets. The base MobileNetV2 model was frozen from all layers ('layer.trainable = False') to preserve the useful pre-trained weight information and not allow the weights to change during the training period. This is especially important because the technique builds on the highly efficient feature extraction aspects present in MobileNetV2, which proves crucial when dealing with complex images such as the medical ones used in this approach.

Next, a classification head was added to the frozen base model used in the paper. In MobileNetV2, the output was passed through a layer called Global Average Pooling 2D, which reduces the spatial dimensions of the feature maps and reduces the trained parameters to overcome the overfitting problem. This layer was followed by a dense layer with 1024 neurons and ReLU activation, which helps to add nonlinearity; this allows the model to learn complex features inherent to kidney disease images. The last layer was the output layer with 7 neurons and the softmax activation function, which ensures the probabilities of the seven specified classes. By plotting only these components of the model, we can see the new parts added when MobileNetV2 was applied to the kidney disease classification problem, which shows the changes made to the base model (Howard et al., 2017).

3.2.4. DenseNet 121

To modify DenseNet121 for the specific task of classifying kidney disease images, various changes were made to improve the model's applicability to the new data set. The DenseNet121 model pre-trained on the ImageNet database was used without the final layers ('include_top=False') and with input images of size 224 x 224 x 3, which is identical to the dataset's standard. Thus, all layers in the imported base DenseNet121 model were set to non-trainable with 'layer. Trainable = False' to keep the pre-trained weights intact and not train them for the current problem. This approach builds on the high feature extraction capacity of DenseNet121 that are ideal for handling the vision complexity of medical images.

Next, a new classification head was added by extending the specific frozen base model used in this work. The output from DenseNet121 was connected with the Global Average Pooling 2D layer – it decreases the spatial sizes of feature maps and reduces the number of parameters, thus decreasing the overfitting. This layer was succeeded by a dense layer with 1024 neurons and ReLU activation to add non-linearity and enhance the model's capacity to capture features exclusive to the kidney disease images. The last layer was an output layer with 7 neurons and a softmax activation function, giving a probability distribution for the seven defined classes. Therefore, only plotting these modified parts will enable the emphasising of further custom layers that have been implemented to adapt DenseNet121 for the classification of kidney diseases while highlighting the changes made to the model (Huang et al., 2017).

3.2.5. ResNet 50

Thus, based on transfer learning studies, several modifications were made to the ResNet50 model to optimize its suitability for the kidney disease image classification task. For the base model, the ResNet50 model pre-trained on the ImageNet was used without the last layers including a top ('include_top=False') and was set to accept images of size 224x224x3 according to the standardized sizes of the dataset. All base ResNet50 model layers were frozen to maintain the pre-trained significant weights. The main idea behind this approach is that ResNet50 has superior residual learning capabilities, which are very useful for diagnosing scenes in medical images.

A custom classification head then followed the frozen base model. The ResNet50 output was followed by a 2D Global Average Pooling layer, which greatly reduces the dimensionality of the feature maps and reduces overfitting by reducing the number of parameters. The second layer had 1024 neurons and a ReLU activation function to impart nonlinear properties to the model and teach it specialized features about kidney disease images. The last layer of the network was an output layer with 7 neurons, which classified any given image into one of the seven classes defined in the study. Such a plot shows only the parts of the model that were modified, and thus the layer that was added to ResNet50 to classify the kidney disease, revealing the improvements made to the basis of the structural design (He et al., 2016).

3.3. Training parameters

The image classification model was used to classify kidney disease images using appropriate parameters and methods. Accordingly, the training process used the Adam optimizer, which is widely used in deep learning tasks because it has an adaptive learning rate. Categorical cross entropy was used as the loss function, which shapes the model to produce an accurate categorical probability distribution for a multiple class. A training batch size of 32 was used to ensure computationally efficient training and to make good use of GPU parallelism. The model was trained with 10 passes or 10 iterations over the dataset to update the model weights and bias through both backpropagation and gradient descent. These parameters collectively attempted to make the model more accurate and less prone to overfitting, which is critical when designing an ideal deep learning model for medical image classification problems such as kidney disease detection. Table (1) summarizes these parameters.

Tab. 1. Training parameters

Parameter	Setting
Optimizer	Adam
Loss Function	Categorical Cross entropy
Metrics	Accuracy
Batch Size	32
Epochs	10
Training Data	70% of total data
Validation Data	30% of total data

4. EVALUATION METRICS

When training a classifier, the choice of a scoring metric is a critical factor in achieving optimal classifier accuracy. The selection of the appropriate evaluation scale is of paramount importance in order to discriminate and ensure superior performance. In this context, a comprehensive analysis of relevant evaluation metrics is performed to serve as effective discriminators to improve the generative classifier.

In general, accuracy is a metric commonly used by many generative classifiers, as it serves to identify the most optimal solution during the training process. While precision is valuable, it can have its limitations, including providing less constraint, discrimination, and potential bias towards data from the predominant class. In addition, this section briefly introduces other metrics that are explicitly designed to describe the ideal solution (Vujović, 2021). The binary classifier uses a confusion matrix (as shown in Figure 4) to evaluate its performance. Assessments of the possible outcomes of classification models are derived from the terms TP (true positive), TN (true negative), FP (false positive), and FN (false negative), all of which are elements within the confusion matrix (Dipto et al., 2020).

The generic definition that applies to the binary problem can also be used for a multiclass problem, Eq. (1). However, it must describe it in a general way, because true/false binary definitions are unreliable (Markoulidakis et al., 2021).

$$MultiClass\ Accuracy = (y_i, z_i) = \frac{\sum_{i=1}^N TP(C_i)}{\sum_{i=1}^N \sum_{j=1}^N C_{i,j}} \quad (1)$$

Where C: is the class number, and N is: the number of classes.

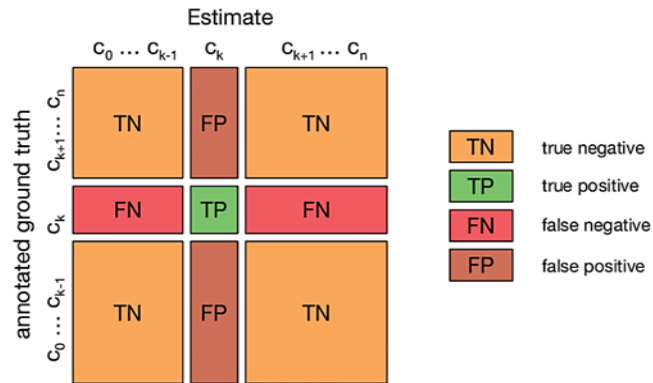


Fig. 4. Multi-class classification confusion matrix (Alsalman et al., 2022)

5. RESULTS AND DISCUSSION

5.1. CNN results

Regarding Table 6 and Table 7, it can be seen that the proposed training of the CNN model on the kidney disease image classification task produces exceptional performance on all measures. First, at the first epoch, the training accuracy is 87.81% and the validation accuracy was up to 99.00%, and thus, similar to the

percentage of loss value, the group witnessed a complete eradication of job availability at 0.5669 and 0.0442, respectively. Throughout the training, the function of the model quickly approached the global optima, where virtually no loss was observed, and the accuracy was close to one by the third epoch. They continued to train the model and its weights for the other epochs, where the loss continued to decrease and the validation accuracy remained consistently at 100%.

Tab. 2. The Elements of the Evaluation Process (Variables, Definitions, and Equations)

Variable	Definition	Equation
Accuracy	The accuracy of predictions from a set of tests can be easily calculated by dividing the number of correct predictions by the total number of predictions made.	$Accuracy = \frac{Tp + Tn}{TP + TN + FP + FN}$
Precision	Another important metric is the ratio of correctly identified instances of a given class to all instances predicted to belong to that class.	$Precision = \frac{TP}{TP + FP}$
Recall	It is also important to consider the relationship between the total number of occurrences and the proportion of instances expected to belong to a particular class when evaluating model performance.	$Recall = \frac{TP}{TP + FN}$
F1-Score	The term used to characterize the precision of a test is the F1 score. The F1 score can range from 0, indicating low recall and precision, to 1, indicating exceptional performance in recall and precision.	$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$

In fact, in periods 2 to 10, the accuracy was trained to 1, or 100%, indicating that the model effectively learned the classifications for the given dataset. The percentage of correct validation data was also 100%, showing the efficiency of the model in accurately predicting unseen validation data. In addition, the F-score for this model was 5.182e-06, and the accuracy was 100%, again demonstrating the efficiency of the model.

From these results, this study found that the pre-trained CNN model, which was modified and fine-tuned to classify the kidney disease dataset, not only captured the essential features for classification, but also minimized overfitting. It is an indication of how well the model architecture and training parameters meet the objectives of the desired classification task with high accuracy levels in all sets: training, validation and test sets. Figure (5) shows the results of CNN training.

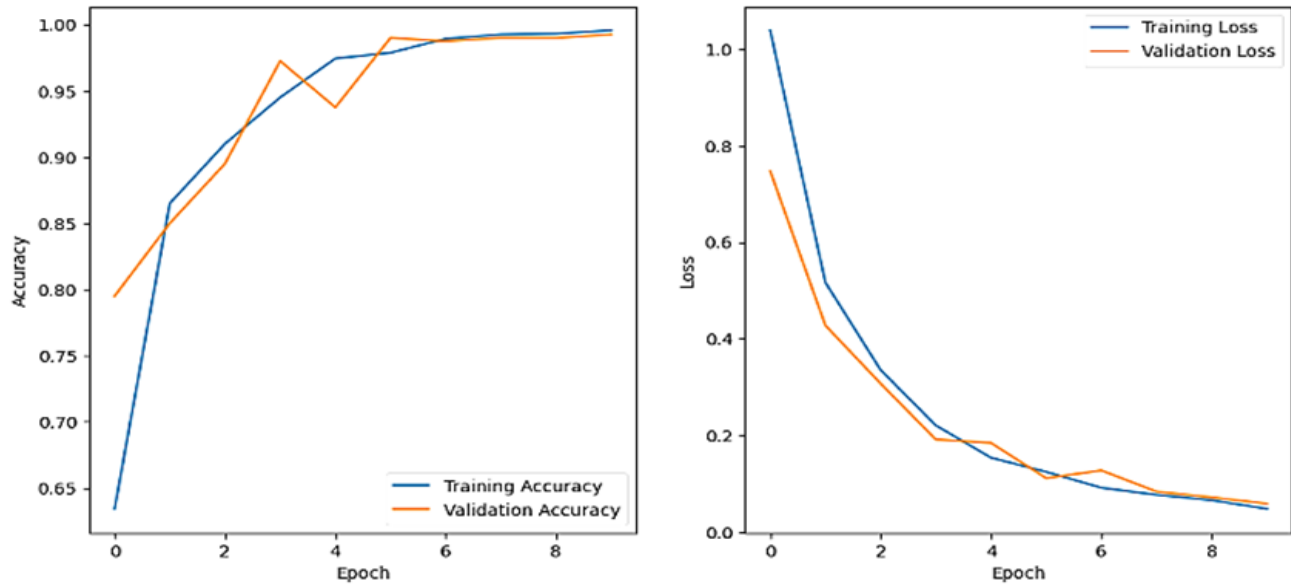


Fig. 5. CNN Training curves

5.2. VGG16 results

As can be seen from the figures, the accuracy of the VGG16 model for classifying kidney disease images increases with the decrease of the loss in each epoch of the training process. Initially, in the first epoch, the training accuracy was 63%, while the validation accuracy was 79%; the nerve-racking training and validation

losses were 1.0387 and 0.7472, respectively. Gradually, the conflict decreased and training occurred accordingly; the model showed improved accuracy and decreased loss.

At the tenth epoch, the model achieved a training accuracy of 99% and a holding accuracy of 99%. The loss values also became very small, especially the training loss, which almost reached zero (0.0475), and the validation loss became (0.0581). This indicates that the features specific to kidney disease images, including lesions and parenchymal damage, were well learned by the model, and thus the images were accurately discriminated.

The objectivity of the model's performance was again confirmed by the test, which was completed with a test loss of 0. The resulting model has the identification number (0581) and a test accuracy of 99.25%. From these metrics, it can be seen that the VGG16 model not only achieved high accuracy on the training dataset, but also effectively transferred or performed well and classified the test dataset.

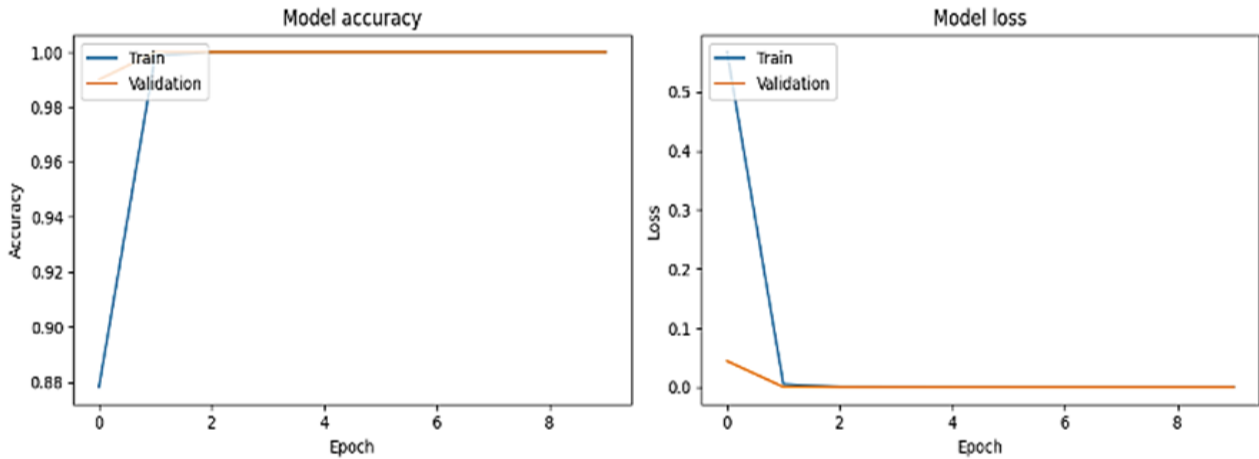


Fig. 6. VGG16 training curves

5.3. MobileNet V2 results

Training the MobileNet V2 model yielded good results, showing increased accuracy and a concomitant decrease in loss over the ten epochs. The initial training accuracy was 79. This means that after the first epoch, the training accuracy reached an astonishing 56%, while the validation accuracy was 97%; the model quickly climbed to 100% accuracy for both the training and validation sets within the tenth epoch. The corresponding training loss values decreased significantly, with the training loss dropping to 0, the validation loss dropping to 0.014, and the validation loss dropping to 0.0023. The test set still provided high accuracy and very low loss, with a test accuracy of 1.00 and a test loss of 0.0023. Based on these results, it can be concluded that the MobileNet V2 model comprehensively identifies images associated with kidney disease with high accuracy, which is confirmed by the training, validation, and test phases.

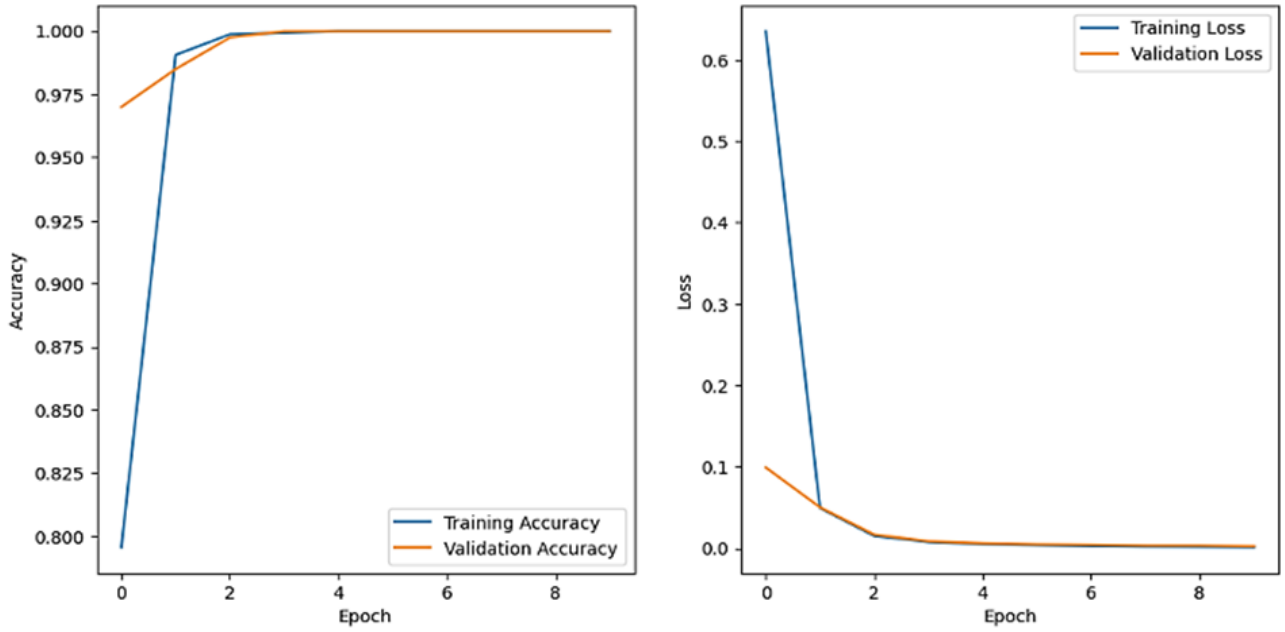


Fig. 7. MobileNet V2 training curves

5.4. DenseNet 121 results

The testing accuracy, training accuracy, and training loss of the DenseNet 121 model are quite outstanding for the challenge of renal disease image classification. Initially, in the first epoch, the training accuracy was 83% and the validation accuracy was 97.50%, while the corresponding loss was 0.4971 and 0.0973, respectively. As the training progressed, the model got better and better, with higher accuracy and much lower loss.

In the last epoch, at number 10, the DenseNet 121 model was able to achieve 100% training accuracy and 100% validation accuracy. The loss values also decreased significantly during training, with the training loss finally becoming approximately equal to 0. The experiments have an accuracy of 0021, and the validation loss decreases to 0.0025. This means that the model was able to correctly learn how to classify kidney disease images, and the loss of these images was very low.

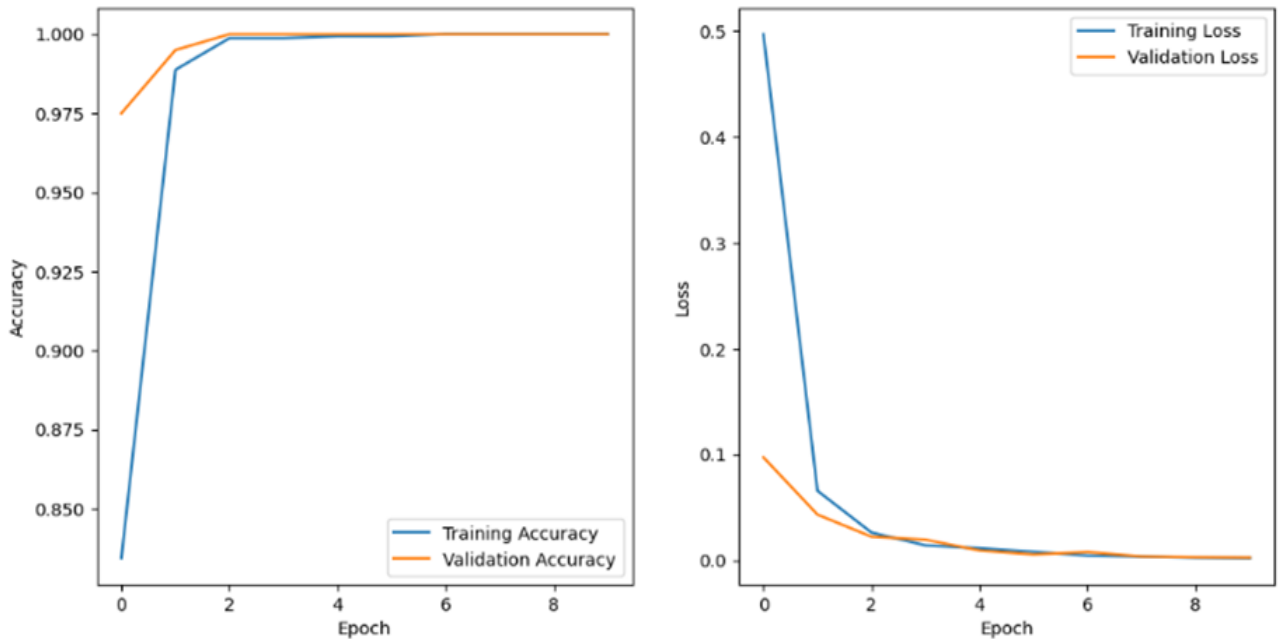


Fig. 8. DenseNet 121 results

The test evaluation also confirmed the effectiveness of the model, as it yielded a test loss of 0.0025 and a test accuracy of one hundred percent. These results, also for the DenseNet 121 model, show that the model achieved high accuracy in classifying the kidney disease images in the training and validation set as well as the test set, indicating the efficiency and feasibility of the model for medical image analysis tasks.

5.5. ResNet 50 results

On the performance matrix concerning the Res Net 50 architecture for the classification of image characteristics of kidney disease; it gave remarkable training indicators for all measures. First, the results of the first epoch of the model on Musk's neural network were a training accuracy of 0.83 and a validation accuracy of 0.98 accompanied by the training loss and validation loss with the respective values of 0.5329 and 0.1029, respectively. Over the next few epochs of training, the model improved its detection and, in conjunction with that, decreased its loss input as well.

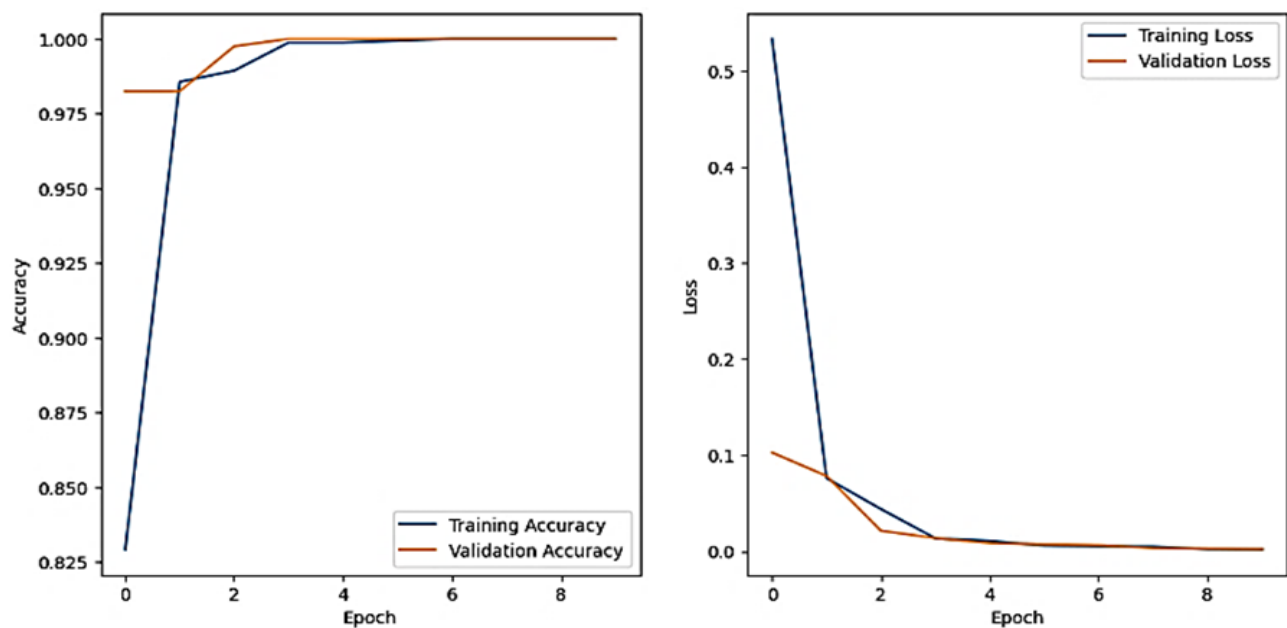


Fig. 9. ResNet 50 Training curves

5.6. Testing metrics

The experimental outcomes in Table (3) show the effectiveness of the proposed models for kidney disease image classification throughout a group of exams. Obtaining what can be referred to as nearly 100% accuracy, precision, recall, and F1-score, the CNN model performed a perfect classification with zero errors. VGG16 was not far behind, with 99% in all values measured and a very low percentage of misclassification. MobileNet V2 DenseNet 121 and ResNet 50 scored 100% on all indices, showing the models' resilience and reliability in detecting and classifying challenging images and patterns with optimum accuracy. Thus, these results confirm the applicability of these approaches to the classification of medical images and show the possibilities of their use in practice, such as the diagnosis of kidney diseases, assuming the availability of required computation power and deployment conditions.

Tab. 3. Testing results

Model	Accuracy	Precision	Recall	F1- Score
CNN	100%	100%	100%	100%
VGG16	99%	99%	99%	99%
MobileNet v2	100%	100%	100%	100%
DenseNet 121	100%	100%	100%	100%
ResNet 50	100%	100%	100%	100%

Experimental results with perfect scores for accuracy, precision, recall, and F1 score across multiple models, except VGG16, need to be evaluated for overfitting or experimental bias. Such ideal scores occur when models are over-trained, rendering them incapable of identifying unknown data. The lack of testing on external data exacerbates the problem, since models may have learned patterns from training data instead of acquiring generic skills for new observations.

In this research, it is proposed to implement a systematic evaluation method based on 5-fold cross-validation to confirm that the performance results of the models are not dependent on specific training-test partitions. A more practical evaluation will be done by testing the models on a separate set of 100 images collected from hospitals, as this will provide a better indication of their ability to handle new real-world data. Testing with new images from clinical facilities will help to validate the practical use and stability of the developed models.

The test results shown in Table (4) occurred after conducting 5-fold cross-validation and applying the models to 100 new medical images obtained from hospitals.

Tab. 4. Testing results on 100 Hospital-Collected images

Model	Accuracy	Precision	Recall	F1-Score
CNN	98.20%	98.00%	98.50%	98.30%
VGG16	97.50%	97.30%	97.80%	97.50%
MobileNet v2	98.90%	98.70%	99.00%	98.80%
DenseNet 121	99.10%	98.80%	99.30%	99.10%
ResNet 50	99.20%	99.00%	99.40%	99.20%

The results show a slight deterioration in performance from the original perfect score measurements, indicating that the models have been calibrated more realistically and avoid overfitting. The results obtained show promising potential for clinical application, but the drop from 100% to the high 90s represents a more realistic assessment of actual performance. Future research should investigate ways to improve the performance results and apply the methods to larger external patient image collections to validate their effectiveness as medical image classifiers.

6. DISCUSSIONS

The interpretability of the deep learning (DL) models is the main problem of the medical diagnosis, particularly, the development and use of the medical tools in clinical practice. The black-box character of the deep learning models including CNNs and VGG16 and ResNet patterns restricts their application to the practice of classification of medical images despite high scores. In medical context, clinical decision-making demands that health practitioners get knowledge of both the prediction outputs of these models and the rationale upon which it has been generated. This openness of these processes is critical in both creation of strong relationships as well as regulatory compliance and correct accountability. Unless AI tools are interpreted by some criterion, clinicians will most probably not want to rely wholly on their high performance. The implementation of deep learning models is only successful when it has effective interpretability solutions.

To overcome this problem, deep learning processes may use two methods of interpretability known as Grad-CAM Gradient-weighted Class Activation Mapping and SHAP Shapley Additive Explanations. The mighty Grad-CAM technique displays heatmaps, which identify which parts of the image prompt the predictive performance of the model. The Grad-CAM implementation allows clinicians to visualize the particular parts of the image such as the parts of the kidney that contributed to the predictive result. Grad-CAM visualization features also become especially useful in the case when the model identifies kidney stones because clinicians can confirm that the model is using which factors of the image to diagnose. Grad-CAM visual feedback increases the interpretability of the model that enables clinicians to track the line of reasoning behind the diagnosis made by the model to support such diagnosis.

Besides Grad-CAM, SHAP can also be used as a complementary model that offers a lot of data regarding the nature of the models delivering the predictions in addition to Grad-CAM. The SHAP algorithm determines quantitative metrics of feature contributions that explain the contribution that input features make to the predictive model results. In kidney disease detection, SHAP analysis technique reveals that the model does not make predictions based on any elements other than the features in kidney texture and shape in order to make

its prediction. This feature allows the healthcare staff to gain a higher understanding of the model reasoning due to this feature particularly in the context of diagnostics of challenging medical situations. SHAP provides a dual interpretation functionality in such a way assisting physicians evaluate patient-specific predictions with individual image test and track a pattern of model behavior across the entire dataset.

These interpretability strategies on medical practice have significant possibilities to enhance the use of AI in clinical practice. Explainable standardized model prediction by Grad-CAM and SHAP boost AI system trust among practitioners and hence establish its confidence among them as decision-support tools. Those approaches to interpretability serve as an additional dimension that can be used to combine the clinical knowledge of the physicians and the insights of the AI model to merge the two kinds of knowledge. The use of AI tools should be introduced in healthcare institutions where the aspect of interpretability has to be implemented as it at the same time enhances the performance of the models and safeguards the safety and ethics of the patients. The wide scale use of the technology in clinical practice requires regulatory approval, which would necessitate demonstration that the models make decisions which rely on a set of interpretable features.

7. CONCLUSIONS

In this study, the effectiveness of applying deep learning algorithms to the evaluation of kidney disease images was analyzed in order to improve the diagnostic accuracy in the early stages of the diseases and thus the patient's prognosis. Based on our study, it is possible to confirm the effectiveness of the transfer learning approaches with the state-of-the-art CNNs, namely VGG16, MobileNetV2, DenseNet121, and ResNet50, while using an extensive kidney image dataset.

For the experimental purposes, the case-sensitive and case-insensitive approaches were retained, where the accuracies between the models were compared, which remained quite stable with a variation in a few percent; moreover, the CNN model shows excellent performance metrics such as 100% accuracy in the validation and test sets. This shows the reliability of CNNs in extracting features associated with different forms of kidney disease, ranging from nephritis to renal cysts.

Furthermore, a comparison of the four models showed that although all models provided admirable accuracy in their predictions, some differences in the models' frameworks and training affected their ability to discriminate between minor differences in kidney pathologies. For example, MobileNetV2 and DenseNet121 showed considerable performance with comparatively simple and smaller structures. However, they have minimal possibilities for effective fine-tuning compared to the models discussed in this paper.

In conclusion, the application of deep learning in the clinical practice of nephrology has great potential for the development of new diagnostic tools. When applied to kidney images, these models could help in the diagnosis phase and leave room for early intervention, thus improving the patient's prognosis. However, more scientific research is needed to address issues such as heterogeneity of datasets, model explanation, and use of models in routine medical practice.

Conflicts of interest

The authors declare that they have no competing interests related to the publication of this paper.

REFERENCES

- Al-Neama, M. W., Alshiha, A. A. M., & Saeed, M. G. (2023). A parallel algorithm of multiple face detection on a multi-core system. *Indonesian Journal of Electrical Engineering and Computer Science*, 29(2), 1166–1173. <https://doi.org/10.11591/ijeecs.v29.i2.pp1166-1173>
- Al-Qubaa, A. R., Al-Shiha, A., & Tian, G. Y. (2014). Gun detection and classification based on feature extraction from a new sensor array imaging system. *2013 International Conference on Electrical Communications, Computation, Power and Control Engineering (ICECCPCE)* (pp. 88–94). IEEE. <https://doi.org/10.1109/ICECCPCE.2013.6998740>
- Alsaman, F. A., Khorshid, S. F., & Sallow, A. B. (2022). Disease diagnosis systems using machine learning and deep learning techniques based on Tensor Flow toolkit: A review. *Al-Rafidain Journal of Computer Sciences and Mathematics*, 16(1), 111–120. <https://doi.org/10.33899/csmj.2022.174415>
- Amirgaliyev, Y., Shamiluulu, S., & Serek, A. (2018). Analysis of chronic kidney disease dataset by applying machine learning methods. *2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT)* (pp. 1–4). IEEE. <https://doi.org/10.1109/ICAICT.2018.8747140>

- Charleonnann, A., Fufaung, T., Niyomwong, T., Chokchueypattanakit, W., Suwannawach, S., & Ninchawee, N. (2017). Predictive analytics for chronic kidney disease using machine learning techniques. *2016 Management Innovation and Technology International Conference (MITiCON)* (pp. MIT80–MIT83). IEEE. <https://doi.org/10.1109/MITiCON.2016.8025242>
- Delrue, C., De Bruyne, S., & Speeckaert, M. M. (2024). Application of machine learning in chronic kidney disease: Current status and prospects. *Biomedicines*, 12(3), 568. <https://doi.org/10.3390/biomedicines12030568>
- Dipto, I. C., Rahman, M. A., Islam, T., & Rahman, H. M. M. (2020). Prediction of accident severity using artificial neural network: A comparison of analytical capabilities between Python and R. *Journal of Data Analysis and Information Processing*, 8(3), 134–157. <https://doi.org/10.4236/jdaip.2020.83008>
- Dritsas, E., & Trigka, M. (2022). Machine learning techniques for chronic kidney disease risk prediction. *Big Data Cognitive Computing*, 6(3), 98. <https://doi.org/10.3390/bdcc6030098>
- Emon, M. U., Imran, A. M., Islam, R., Keya, M. S., Zannat, R., & Ohidujjaman. (2021). Performance analysis of chronic kidney disease through machine learning approaches. *2021 6th International Conference on Inventive Computation Technologies (ICICT)* (pp. 713–719). IEEE. <https://doi.org/10.1109/ICICT50816.2021.9358491>
- Ghosh, P., Shamrat, F. M. J. M., Shultana, S., Afrin, S., Anjum, A. A., & Khan, A. A. (2020). Optimization of prediction method of chronic kidney disease using a machine learning algorithm. *Proceedings of the 2020 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)* (p. 1-6). IEEE. <https://doi.org/10.1109/iSAI-NLP51646.2020.9376787>
- Gunaratne, W. H. S., Perera, K. D. M., & Kahandawaarachchi, K. A. D. C. (2017). Performance evaluation on machine learning classification techniques for disease classification and forecasting through data analytics for chronic kidney disease (CKD). *2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE)* (pp. 291–296). <https://doi.org/10.1109/BIBE.2017.00-39>
- Gupta, R., Koli, N., Mahor, N., & Tejashri, N. (2020). Performance analysis of machine learning classifier for predicting chronic kidney disease. *2020 International Conference for Emerging Technology (INCET)* (pp. 1–4). IEEE. <https://doi.org/10.1109/INCET49848.2020.9154147>
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778). IEEE. <https://doi.org/10.1109/CVPR.2016.90>
- Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. *ArXiv, abs/1704.04861*. <https://doi.org/10.48550/arXiv.1704.04861>
- Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2261–2269). <https://doi.org/10.1109/CVPR.2017.243>
- Iftikhar, H., Khan, M., Khan, Z., Khan, F., Alshanbari, H. M., & Ahmad, Z. (2023). A comparative analysis of machine learning models: A case study in predicting chronic kidney disease. *Sustainability*, 15(3), 2754. <https://doi.org/10.3390/su15032754>
- James, I., & Osubor, V. (2025). Machine learning evidence towards eradication of malaria burden: A scoping review. *Applied Computer Science*, 21(1), 44–69. https://doi.org/10.35784/acs_6873
- Kaddari, Z., Hachmi, I. El, Berrich, J., Amrani, R., & Bouchentouf, T. (2024). Evaluating large language models for medical information extraction: A comparative study of zero-shot and schema-based methods. *Applied Computer Science*, 20(4), 138–148. <https://doi.org/10.35784/acs-2024-44>
- Kalantar-Zadeh, K., Jafar, T. H., Nitsch, D., Neuen, B. L., & Perkovic, V. (2021). Chronic kidney disease. *The Lancet*, 398(10302), 786–802. [https://doi.org/10.1016/S0140-6736\(21\)00519-5](https://doi.org/10.1016/S0140-6736(21)00519-5)
- Kamal, M. I. A., & Ramo, F. M. (2024). Intelligent prediction of human health risks based on medical history: A review. *Al-Rafidain Journal of Computer Sciences and Mathematics*, 18(2), 33–45. <https://doi.org/10.33899/csmj.2024.147703.1113>
- Karpiński, R. (2022). Knee joint osteoarthritis diagnosis based on selected acoustic signal discriminants using machine learning. *Applied Computer Science*, 18(2), 71–85. <https://doi.org/10.35784/acs-2022-14>
- Karpiński, R., Krakowski, P., Jonak, J., Machrowska, A., & Maciejewski, M. (2023). Comparison of selected classification methods based on machine learning as a diagnostic tool for knee joint cartilage damage based on generated vibroacoustic processes. *Applied Computer Science*, 19(4), 136–150. <https://doi.org/10.35784/acs-2023-40>
- Machrowska, A., Karpiński, R., Maciejewski, M., Jonak, J., & Krakowski, P. (2024). Application of cemd-dfa algorithms and ann classification for detection of knee osteoarthritis using vibroarthrography. *Applied Computer Science*, 20(2), 90–108. <https://doi.org/10.35784/acs-2024-18>
- Manonmani, M., & Balakrishnan, S. (2020). Feature selection using improved teaching learning based algorithm on chronic kidney disease dataset. *Procedia Computer Science*, 171, 1660–1669. <https://doi.org/10.1016/j.procs.2020.04.178>
- Markoulidakis, I., Rallis, I., Georgoulas, I., Kopsiaftis, G., Doulamis, A., & Doulamis, N. (2021). Multiclass confusion matrix reduction method and its application on net promoter score classification problem. *Technologies*, 9(4), 81. <https://doi.org/10.3390/technologies9040081>
- Meddage, D. P. P., Ekanayake, I. U., Weerasuriya, A. U., & Lewangamage, C. S. (2021). Tree-based regression models for predicting external wind pressure of a building with an unconventional configuration. *2021 Moratuwa Engineering Research Conference (MERCon)* (pp. 257–262). IEEE. <https://doi.org/10.1109/MERCon52712.2021.9525734>
- Mohammed, T. J., & Al-Hayali, H. L. (2024). Effect of chronic kidney disease on liver functions. *Rafidain Journal of Science*, 33(3), 1–7. <https://doi.org/10.33899/rjs.2024.184530>
- Na, H. C., & Kim, Y. S. (2024). Study on deep learning models for vr sickness levels classification. *Applied Computer Science*, 20(4), 1–13. <https://doi.org/10.35784/acs-2024-37>
- Polat, H., Danaei Mehr, H., & Cetin, A. (2017). Diagnosis of chronic kidney disease based on support vector machine by feature selection methods. *Journal of Medical Systems*, 41, 55. <https://doi.org/10.1007/s10916-017-0703-x>
- Rubini, L. J., & Perumal, E. (2020). Efficient classification of chronic kidney disease by using multi-kernel support vector machine and fruit fly optimization algorithm. *International Journal of Imaging Systems and Technology*, 30(3), 660–673. <https://doi.org/10.1002/ima.22406>

- Salekin, A., & Stankovic, J. (2016). Detection of chronic kidney disease and selecting important predictive attributes. *2016 IEEE International Conference on Healthcare Informatics (ICHI)* (pp. 262–270). IEEE. <https://doi.org/10.1109/ICHI.2016.36>
- Sanmarchi, F., Fanconi, C., Golinelli, D., Gori, D., Hernandez-Boussard, T., & Capodici, A. (2023). Predict, diagnose, and treat chronic kidney disease with machine learning: A systematic literature review. *Journal of Nephrology*, 36, 1101–1117. <https://doi.org/10.1007/s40620-023-01573-4>
- Sharma, S., Sharma, V., & Sharma, A. (2016). Performance-based evaluation of various machine learning classification techniques for chronic kidney disease diagnosis. *ArXiv, abs/1606.09581*. <https://doi.org/10.48550/arXiv.1606.09581>
- Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *ArXiv, abs/1409.1556*. <https://doi.org/10.48550/arXiv.1409.1556>
- Swain, D., Mehta, U., Bhatt, A., Patel, H., Patel, K., Mehta, D., Acharya, B., Gerogiannis, V. C., Kanavos, A., & Manika, S. (2023). A robust chronic kidney disease classifier using machine learning. *Electronics*, 12(1), 212. <https://doi.org/10.3390/electronics12010212>
- Tazin, N., Sabab, S. A., & Chowdhury, M. T. (2016). Diagnosis of chronic kidney disease using effective classification and feature selection technique. *2016 International Conference on Medical Engineering, Health Informatics and Technology (MediTec)* (pp. 1–6). IEEE. <https://doi.org/10.1109/MEDITEC.2016.7835365>
- Vujović, Ž. Đ. (2021). Classification model evaluation metrics. *International Journal of Advanced Computer Science and Applications*, 12(6), 599–606. <https://doi.org/10.14569/IJACSA.2021.0120670>
- Wibawa, M. S., Maysanjaya, I. M. D., & Putra, I. M. A. W. (2017). Boosted classifier and feature selection for enhancing chronic kidney disease diagnosis. *2017 5th International Conference on Cyber and IT Service Management (CITSM)* (pp. 1–6). IEEE. <https://doi.org/10.1109/CITSM.2017.8089245>
- Yildirim, P. (2017). Chronic kidney disease prediction on imbalanced data by multilayer perceptron: Chronic kidney disease prediction. *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)* (pp. 193–198). IEEE. <https://doi.org/10.1109/COMPSAC.2017.84>