Advances in Artificial Intelligence and Machine Learning; Research 5 (1) 3425-3445 Received 10-12-2024; Accepted 17-03-2025; Published 24-03-2025

A Novel Ensemble-Based Deep Learning Model with Explainable AI for Accurate Kidney Disease Diagnosis

Md. Arifuzzaman arif_cse@lus.ac.bd
Department of Computer Science and Engineering
Leading University
Sylhet, Bangladesh

Iftekhar Ahmed iftekharifat007@gmail.com
Department of Computer Science and Engineering
Leading University
Sylhet, Bangladesh

Md. Jalal Uddin Chowdhury jalal_cse@lus.ac.bd
DeepNet Research and Development Lab,
Sylhet 3100, Bangladesh

Mohammad Shoaib Rahman shoaib_cse@lus.ac.bd
Department of Computer Science and Engineering
Leading University
Sylhet, Bangladesh

Md. Ebrahim Hossain ebrahim.cse@lus.ac.bd
Department of Computer Science and Engineering
Leading University
Sylhet, Bangladesh

Apurbo Deb Nath apurbodebnath50@gmail.com
Department of Computer Science and Engineering
Leading University
Sylhet, Bangladesh

Shakib Absar sabsar42@gmail.com
Department of Computer Science and Engineering
Leading University
Sylhet, Bangladesh

Corresponding Author: Md. Arifuzzaman

Abstract

Chronic kidney disease (CKD) represents a significant global health challenge characterized by a progressive decline in renal function, leading to the accumulation of waste products and disruptions in fluid balance within the body. Given its pervasive impact on public health, there is a pressing need for effective diagnostic tools to enable timely intervention. Our study

Citation: Md. Arifuzzaman, et al. A Novel Ensemble-Based Deep Learning Model with Explainable AI for Accurate Kidney Disease Diagnosis. Advances in Artificial Intelligence and Machine Learning. 2025;5(1):196.

delves into the application of cutting-edge transfer learning models for the early detection of CKD. We carefully test the performance of several cutting-edge models, such as EfficientNetV2, InceptionNetV2, MobileNetV2, and the Vision Transformer (ViT) technique, using a large dataset that is available to the public. Remarkably, our analysis demonstrates superior accuracy rates, surpassing the 90% threshold with MobileNetV2 and achieving 91.5% accuracy with ViT. Moreover, to enhance predictive capabilities further, we integrate these individual methodologies through ensemble modeling, resulting in our ensemble model exhibiting a remarkable 96% accuracy in the early detection of CKD. This significant advancement holds immense promise for improving clinical outcomes and underscores the critical role of machine learning in addressing complex medical challenges.

Keywords: Kidney disease, Deep learning, Transfer learning, Vision transformer, Ensemble model.

1. INTRODUCTION

According to the 2010 Global Burden of Disease study, chronic renal disease jumped from 27th place in 1990 to 18th place in 2010 among all causes of death worldwide. 10% of the global population is affected by CKD, and due to a lack of awareness and affordable treatment, millions of people die from it every year. A disease is an abnormal state of an organism that interferes with normal biological functions and usually causes pain and weakness. It is usually accompanied by symptoms and signs that cause disorder, suffering, or even death [1]. The kidneys are one of the main organs of the human body. To have a healthy life, healthy kidneys are a must. The kidneys remove wastes and extra water to make urine and filter about a half cup of blood every minute in the human body [2]. Furthermore, the kidneys also maintain a balance of water, salt, and minerals by removing acid from the body [3]. When the kidneys do not function properly to filter blood into the body, that is a situation called kidney disease. Chronic kidney disease (CKD) involves a gradual loss of kidney function that leads to the accumulation of excess waste and fluids in our body [4]. The waste that has accumulated in the body can be harmful to our overall health, leading to end-stage renal disease when the kidneys stop working completely [5].

Ninety percent of the 37 million persons with CKD in the US are unaware that they have the disease. Many CKD patients remain asymptomatic until their illness progresses to more severe stages or problems develop [6]. Symptoms may include foamy urine, spitting, altered urination frequency, dry or itchy skin, fatigue and nausea, decreased appetite, and unintentional weight loss. Individuals with advanced stages of CKD may further experience impaired concentration, tingling or swelling in the legs, respiratory problems, vomiting, insomnia, and a breath stench like ammonia [7]. There are many causes of CKD, but these two are the most frequent causes of i) diabetes and ii) hypertension; by diabetes approximately 25% of cases of kidney failure are caused, while the other 33% are caused by hypertension. High blood pressure is one of the major reasons for CKD [7]. If someone has long-term conditions like hypertension, diabetes, or high blood pressure that can lead to life-threatening CKD [8]. Smoking can cause many cardiovascular diseases that can often lead to CKD. Quitting smoking, stopping drinking alcohol, eating healthy, exercising regularly, and being careful about taking painkillers can prevent CKD from happening [9]. CKD is a worldwide public health emergency. The World Health Organisation reported around 58 million worldwide deaths in 2005, with 35 million attributed to chronic diseases [10]. Unfortunately, only 2 million people get

dialysis or a kidney transplant for their survival; however, this number represents only 10% of those who require treatment [11].

To assess someone's kidneys and make the diagnosis of kidney failure, a medical professional may employ a range of renal function tests such as i)Blood test ii) Urine test, and iii) Imaging tests. Based on someone's estimated glomerular filtration presence rate in the body, there are different stages of kidney disease (eGFR) such as i) Stage I (eGFR is higher than 90 but below 100) ii) Stage II (eGFR is higher than 60 but below 89) iii) Stage III (eGFR is higher than 30 but below 59) iv) Stage IV (eGFR is higher than 15 but below 29) and v) Stage V (eGFR is below 15) [12]. Medical officials also give blood tests to discover kidney problems. If the kidneys are producing urine along with protein and blood then the kidneys are not working perfectly [13]. Then there can be two possible scenarios to solve kidney problems, one is dialysis and another one is kidney transplant [14]. Dialysis is a very physically and mentally challenging process for the patient, and the medical official must be concerned about the patient's food habits, age, gender, body movements, and how often the dialysis is being taken [15]. When the dialysis fails to improve the patient, then there is only one option left for the doctors, and that is a kidney transplant. It is a very costly and challenging process and also requires the finest medical officials [16].

In the era of Artificial Intelligence(AI) and Machine Learning(ML), researchers are coming up with many ideas to detect CKD not only using clinical diagnosis but also with the help of various algorithms and models [17]. To detect CKD medical images and physiological signals are being used by researchers in the Deep Learning (DL) techniques [18]. Various models and studies have been conducted and found that most of the researchers used the Conventional Neural Network(CNN) model, which did not perform well on multi-class image classification [19].

After studying many researches we are going to propose a model for the early detection of CKD with more accuracy. So in this study, we are trying to leverage the transfer learning models and custom Convolutional neural network(CNN) to classify CKD. We have trained 3 transfer learning models and among them, MobileNetV2 performed well with more than 90% accuracy. Here in this research, we have used state-of-the-art transfer learning models i.e., i) EfficientNetV2 ii)InceptionNetV2, and iii)MobileNetV2. We have also used the Vision Transformer (ViT) method for training our model, and we have got 91.5% accuracy. Finally, we proposed a model that used ensemble methods and we achieved 96% accuracy. We are proposing an ensemble model for the early detection of CKD so that much suffering can end and lives can be saved.

2. LITERATURE REVIEW

This section reviews some recent image-processing research works. The paper discusses several methodologies and strategies for using image processing to diagnose Kidney disease.

Singh et al. (2022) [20] proposed a framework for early prediction of chronic kidney disease using a deep neural network. The authors of this study examined the Recursive Feature Elimination technique to determine which features are crucial for making accurate predictions. For the aim of classification, they fed multiple features to machine learning models. The proposed Deep neural model achieved better than the SVM, KNN, Logistic regression, Random Forest, and Naive Bayes algorithms. The suggested model's main flaw was that it could only be validated using limited

sample sizes of data. In order to better recognize the severity of CKD, large amounts of progressively high-quality and representative CKD data will be gathered in the future.

Majid et al. (2023) [21] conducted transfer learning strategies for kidney disease classification using CT images. In order to enhance the efficiency of the training procedure, the researchers used a range of pre-processing approaches and employed image scaling methods. Within this framework, they unveiled two improved transfer learning (TL) models, ResNet-101 and DenseNet-121, for predicting kidney tumors. DenseNet-121, a transfer learning models. Another study, Sudharson et al. (2020) [22], applied an ensemble of deep neural networks using transfer learning for kidney ultrasound image classification. The process involves the combination of several pre-trained deep neural networks, including ResNet-101, ShuffleNet, and MobileNet-v2. The final predictions are made by the use of the majority voting approach. The approach that was proposed achieved a maximum classification accuracy of 96.54% when tested with quality images and 95.58% when tested with noisy images.

Kim et al. (2021) [23] applied an ANN for chronic kidney disease classification. The GLCM technique, extensively utilized in ultrasound image processing, was used to extract parameters from each ROI. The artificial neural network (ANN) has 58 input parameters, ten hidden layers, and three output layers. The concluded classification rate was 95.4% using the ANN model, and the training epoch was 38 times. They will apply the Transfer learning model to increase performance on this dataset and also will increase the dataset for the training model.

Radya et al. (2019) [24] applied data mining techniques for kidney disease prediction. These techniques included Multilayer Perceptrons(MLP), SVM, and Probabilistic Neural Networks. The PNNs algorithm has the greatest overall classification accuracy percentage of 96.7% when compared to other algorithms for identifying the stages of patients with CKD. Using four distinct algorithms, The authors used very limited datasets, consisting of just 361 instances, to forecast CKD.

Bhandari et al. (2023) [25] proposed a lightweight CNN to detect kidney disease. The training data's means and standard deviation were used by The LIME image explanation to generate a number of features and changes. LIME gave a visual description of how the model made its decisions and pointed out the parts of the image that were most important for predicting a specific category. After extensive testing, the suggested CNN model proved to be nearly perfect, with an accuracy of 98.68 percent. This study only employed a small number of CT scans. Therefore, the results may be better if the use of data augmentation. By integrating DL models with other XAI algorithms, they will increase the clarity of the outcomes.

Bhattacharjee et al. (2023) [26] proposed a computer-aided diagnostic model for kidney disease classification using a modified Xception deep neural network version, with image net weights derived via transfer learning. The model trained with these two datasets has a 99.39% success rate in this research. Due to the ensemble models' lack of depth, it is impossible to extract contextual information from adjacent slices. A 3D classification model that makes use of interc-slice context is one potential solution to this issue.

Kanwal et al. (2022) [27] proposed an automated model for the classification of kidney abnormalities. Two useful automatic models were included in the proposed study. First was Efficient-b0, and second was ResNet-18. Both of them correctly predicted problems with the kidneys. The accuracy

of the proposed ResNet-18 model was determined to be 98.1% after testing. They will gather real-time datasets in order to test and train their model as part of the next phase of their research.

Wasi et al. (2023) [28] proposed an identification model for kidney tumors using transfer learning. In order to identify kidney tumors from CT scans, a deep CNN-based transfer learning approach is proposed. Results for the 5,284-image test set showed an accuracy of 92.54% after 50 epochs of training. They only used data, including images of kidneys or tumors. Further extension activities, include enhancing the quantity and quality of datasets.

Kadir Yildirim et al. (2021) [29] proposed an automated system for detecting kidney stones in the human body using coronal computed tomography (CT) images with the help of an AI technique, namely Deep Learning. About 1800 images were used for each person's cross-sectional CT images. This system can detect small-size kidney stones with an accuracy of 96.82%. This method can be used in urology to solve many problems for clinical application because it has given great results for a larger dataset of 433 subjects.

Fuzhe Ma et al. (2020) [30] stated that chronic kidney disease (CKD) is increasing day by day, to diagnose CKD, machine learning techniques have become an essential tool in recent years. They suggested a model using a Heterogeneous Modified ANN to detect, segment, and diagnose CKD on the platform of the Internet of Medical Things (IoMT) that is described as a SVM and MLP using a Backpropagation (BP) method. During the preprocessing step, ultrasound images are used to segment the image. The proposed algorithm reduces the time and provides an accuracy of 97.5 percent.

Nicholas Heller et al. (2021) [31] reported that numerous studies have been conducted to establish a connection between the geometric and anatomical features of kidney tumors and the outcomes of oncology. Producing high-quality 3D segmentations takes a lot of time and human energy from the tumors and the kidneys that host them. Furthermore, in autonomous 3D segmentation, deep learning techniques have achieved good results, and they need a tremendous amount of training data. In this study, 90 cases were predicted based on the average Sørensen-Dice coefficient between the kidney and tumor. The Winning team set a benchmark for 3D semantic segmentation with an accuracy of 97.4% for the kidney and 85.1% for the tumor.

Swapnita Srivastava et al. (2022) [32] proposed a model to determine and diagnose CKD using computational-based methods. Another study, Navaneeth Bhaskar et al. (2019) [33] suggested a new model to find out kidney disease automatically using machine learning algorithms. To identify the disease, the salivary urea concentration is observed with a new sensing approach to monitor the urea levels in the saliva sample. A 1-dimensional DL CNN method that includes an SVM classifier is developed in this study. The accuracy of the model has been enhanced because of the CNN-SVM integrated network. The proposed model shows that it provides 98.04% accuracy.

Francesco Paolo Schena et al. (2021) [34] proposed an ANN prediction model for end-stage kidney disease (ESKD) in patients with primary immunoglobulin A nephropathy (IgAN). To predict ESKD, this study applies a two-step procedure of a classifier model and to detect the development of ESKD, a regression model is applied. A clinical decision support system (CDSS), which is easy to use, has been developed to predict ESKD in patients with IgAN with a median follow-up of 5 and 10 years. The accuracy of the classifier model is 89% for the patients with a follow-up for ten years. The proposed system gives a result of 91% to predict IgAN in a patient.

Guozhen Chen et al. 2020 [35] proposed a method to detect Chronic Kidney Disease (CKD) in the early stage efficiently and effectively by using various deep learning methods, and Adaptive hybridized Deep Convolutional Neural Networks (AHDCNN). This study stated that to achieve high accuracy, an algorithm model has been developed using a Conventional Neural Network (CNN) to classify the dataset properly applying feature dimension. The Internet of Medical Things platform (IoMT) concluded that using machine learning techniques helped to produce the solution to kidney disease as well as other diseases too. The proposed system can achieve an accuracy of 97.3% in detecting CKD.

Md Nazmul Islam et al. 2022 [36], to diagnose kidney disease properly at the earliest time, an AI-based system needs to be developed. After analyzing the data, the study found that the images had the same type of mean color distribution from all of the classes. To find the best result, six machine learning models were used, namely EANet, CCT, and Swin transformers, Resnet, VGG16, and Inception v3. The results show that the VGG16 and CCT models provide a decent output in terms of accuracy, but the swin transformer gives an accuracy of 99.30%.

Chin-Chi Kuo, et al. (2019) [37] proposed a deep learning approach for automatically determining the estimated glomerular filtration rate (eGFR) and Chronic Kidney Disease (CKD) status. In this study, to predict kidney function, they used 4,505 kidney ultrasound images labeled using eGFRs. This study demonstrated the use of a neural network architecture for the transfer learning approach in conjunction with the ResNet model on the ImageNet dataset. This work used kidney length annotations to exclude the peripheral regions of the kidneys and employed diverse data augmentation techniques to provide supplementary data variations for enhanced information extraction from the ultrasound pictures. The proposed AI-GFR model provides an accuracy of 85.6%, which shows that this model can be applied to detect CKD status in clinical practice.

3. METHODOLOGY

3.1 Dataset

3.1.1 Data collection

The dataset, aptly named ”CT KIDNEY DATASET: Normal-Cyst-Tumor and Stone” [38], was meticulously assembled from diverse medical sources, specifically various hospitals in Dhaka, Bangladesh. Patients within this dataset had previously received diagnoses related to kidney conditions, covering an extensive array of scenarios. The dataset is notably comprehensive, featuring 12,446 unique instances. These cases include 3,709 instances of cysts, 5,077 normal cases, 1,377 instances of stones, and 2,283 tumor cases. Figure 1 includes some sample images to help understand the datasets. Notably, both contrast and non-contrast studies, as well as Coronal and axial cuts, contribute to the dataset’s richness and representativeness of diverse kidney pathologies.

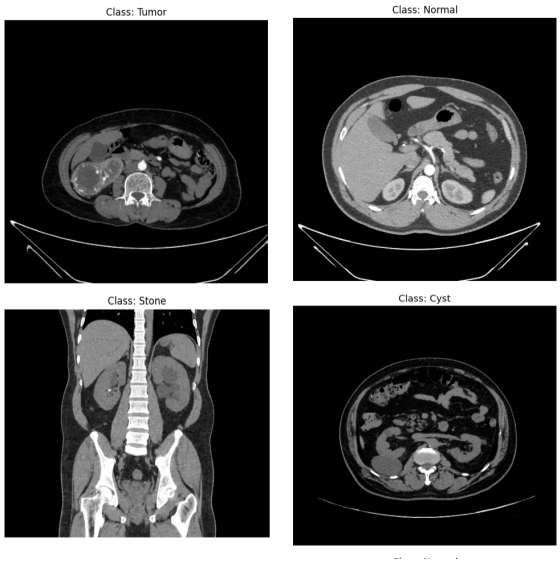

Figure 1: Random Images From Dataset

3.1.2 Dataset analysis

The following is a breakdown of the class distribution and class imbalance for the four classes. It shows that normal class contains (40.72%) of the data while cyst class contains (29.74%) then tumor class contains (18.30%) and the least representation in stone class (11.04%) which added in FIGURE 2. This imbalance can result in bias during model training, as DNNs have a tendency to prefer classes with more training samples while leading to poorer classification performance for classes with fewer training samples. As the dataset is collected from multiple hospitals and information regarding demographic factors such as age, gender and the socioeconomic background of the patients is not available.

The dataset annotations were validated by both a radiologist and a medical technologist to ensure accuracy. However, inter-observer variability remains a concern. Implementing consensus-based techniques, such as majority voting, could enhance annotation reliability. Additionally, analyzing factors like pixel intensity, resolution variations, and imaging modalities can refine preprocessing steps and improve model robustness.

3.2 Data Pre-processing

3.2.1 Image augmentation

A pivotal aspect of this research involves the strategic application of data augmentation techniques, realized through the utilization of the ‘ImageDataGenerator‘ class. The augmentation strategy encompasses a range of transformations, including rescaling (224x224), rotation, zooming, horizontal and vertical flipping, and shifting. These augmentations serve the dual purpose of diversifying

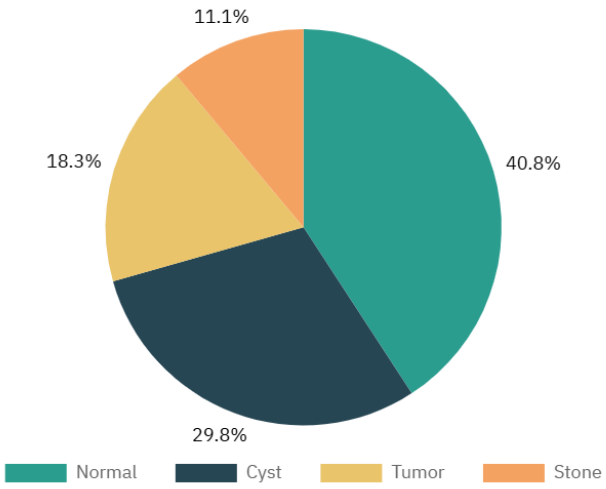

Figure 2: Class Distribution of the CT Kidney Dataset

the dataset, enriching it with varied instances, and enhancing the model's resilience to the inherent variations in kidney images.

3.2.2 Train-Test split

To meticulously assess the generalization capabilities of the models, an 80-20 train-test split was implemented. This careful partitioning ensures that the models are trained on a substantial dataset while retaining a sufficiently independent test set for rigorous evaluation. The adoption of an 80-20 split aims to strike a balance between model convergence and the prevention of overfitting.

3.2.3 Label encoding

The process of assigning numeric labels to distinct classes was carried out through the application of the 'LabelEncoder' from the scikit-learn library. Following this, a crucial step involved the conversion of these labels into one-hot encoding. This categorical representation is fundamental for training the models, enabling them to accurately discern and classify the diverse array of kidney conditions present in the dataset.

3.2.4 Dataset creation

The creation of datasets for both the training and validation phases was meticulously executed through the 'flow_from_dataframe' method from TensorFlow. The training dataset was intentionally enriched with augmentations, thereby exposing the model to an even broader array of representations of kidney conditions. In contrast, the validation dataset remained unaltered. ensuring that the model's performance could be evaluated on authentic, real-world, and unaltered data.

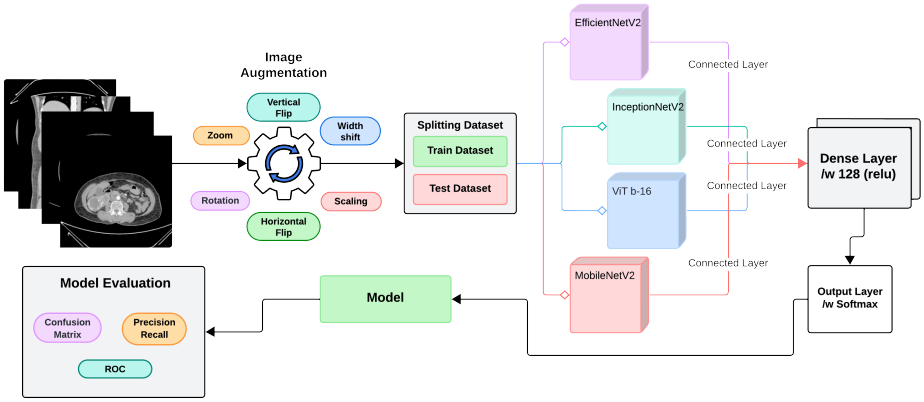

Figure 3: Methodology of Proposed Architecture

3.3 Transfer Learning Models

3.3.1 EfficientNetV2

EfficientNetV2 [39], a family of convolutional neural network (CNN) models, introduces a novel approach known as compound scaling to systematically enhance the model's depth, width, and resolution in tandem, resulting in improved overall performance. This architecture features a sequence of mobile inverted bottleneck blocks similar to those found in MobileNetV2 but with increased depth and width. Through intelligent employment of compound scaling, EfficientNetV2 achieves a fine balance between accuracy and computational efficiency. In our research, we leverage EfficientNetV2 as a state-of-the-art model for image classification tasks, utilizing it as a benchmark to evaluate its classification capabilities and to compare it against other contemporary architectural paradigms.

3.3.2 InceptionNetV2

InceptionNetV2 [40] stands as a noteworthy family of convolutional neural network (CNN) models, extending the original InceptionNet with innovative architectural enhancements. InceptionNetV2 captures features at different resolutions thanks to its unique inception modules that use multi-scale convolutional filters. The model further integrates factorized convolutions to manage computational complexity without compromising representational power. Unlike EfficientNetV2's compound scaling, InceptionNetV2 relies on meticulously designed modules and architectural refinements to strike a balance between model intricacy and performance. Featuring sophisticated blocks, including Inception and reduction blocks, InceptionNetV2 excels in effective feature extraction. Our comparison study uses InceptionNetV2 as a key part because it lets us compare its image classification performance to other modern, state-of-the-art models. This shows how model design, complexity, and classification accuracy are all connected in the world of image recognition tasks.

3.3.3 MobileNetV2

MobileNetV2 [41] architecture caters to mobile and embedded devices characterized by limited computational resources. It strategically decomposes the conventional convolution operation into depthwise and pointwise convolutions, effectively reducing computational overhead. Notably, MobileNetV2 employs depth-wise separable convolutions, a key innovation contributing to its efficiency. Through the utilization of inverted residual blocks with linear bottleneck layers, the architecture strikes an optimal balance between model compactness and accuracy. These blocks facilitate effective feature extraction while simultaneously minimizing the parameter count. In our study, MobileNetV2 serves as a comparative benchmark, enabling us to assess its performance relative to other contemporary state-of-the-art models.

3.3.4 Vision transformer

The Vision Transformer (ViT) represents a paradigm shift in image classification, moving away from traditional convolutional neural networks (CNNs) towards transformer-based architectures. Pioneered by Dosovitskiy et al. 2021 [42], ViT applies self-attention mechanisms and transformer architecture originally developed for natural language processing tasks to image recognition. By breaking down input images into fixed-size patches, ViT facilitates direct processing through transformer layers, capturing global context information effectively. Despite the lack of hierarchical feature extraction inherent in CNNs, ViT compensates with its ability to learn long-range dependencies through self-attention, enabling robust feature representation. In our research, ViT serves as a crucial benchmark for evaluating its classification capabilities alongside other contemporary state-of-the-art models. The goal of this comparison study is to show how transformer-based architectures, computational efficiency, and classification accuracy all work together in complex ways. This will help the field of image recognition research as it grows.

3.4 Model Configuration

The training of each transfer learning model was conducted with meticulous care over 50 epochs. This iterative process sought a harmonious convergence of the model while guarding against the potential pitfalls of overfitting. FIGURE 3 represents the Workflow diagram where all steps for this study. The training was executed leveraging the augmented training dataset, with subsequent validation on an unaltered validation dataset. Softmax as the activation function and Adam as the optimizer were used. Evaluation metrics, including categorical cross-entropy loss and accuracy, were employed to critically assess model performance. The early stopping feature, along with the patience of 5 epochs, was also very important for improving the efficiency of training and reducing the chance of overfitting, which in turn made the models more robust.

3.5 Ensemble model

The ensemble model is an ensemble architecture that amalgamates multiple state-of-the-art image classification models, including convolutional neural networks (CNNs) and the Vision Transformer

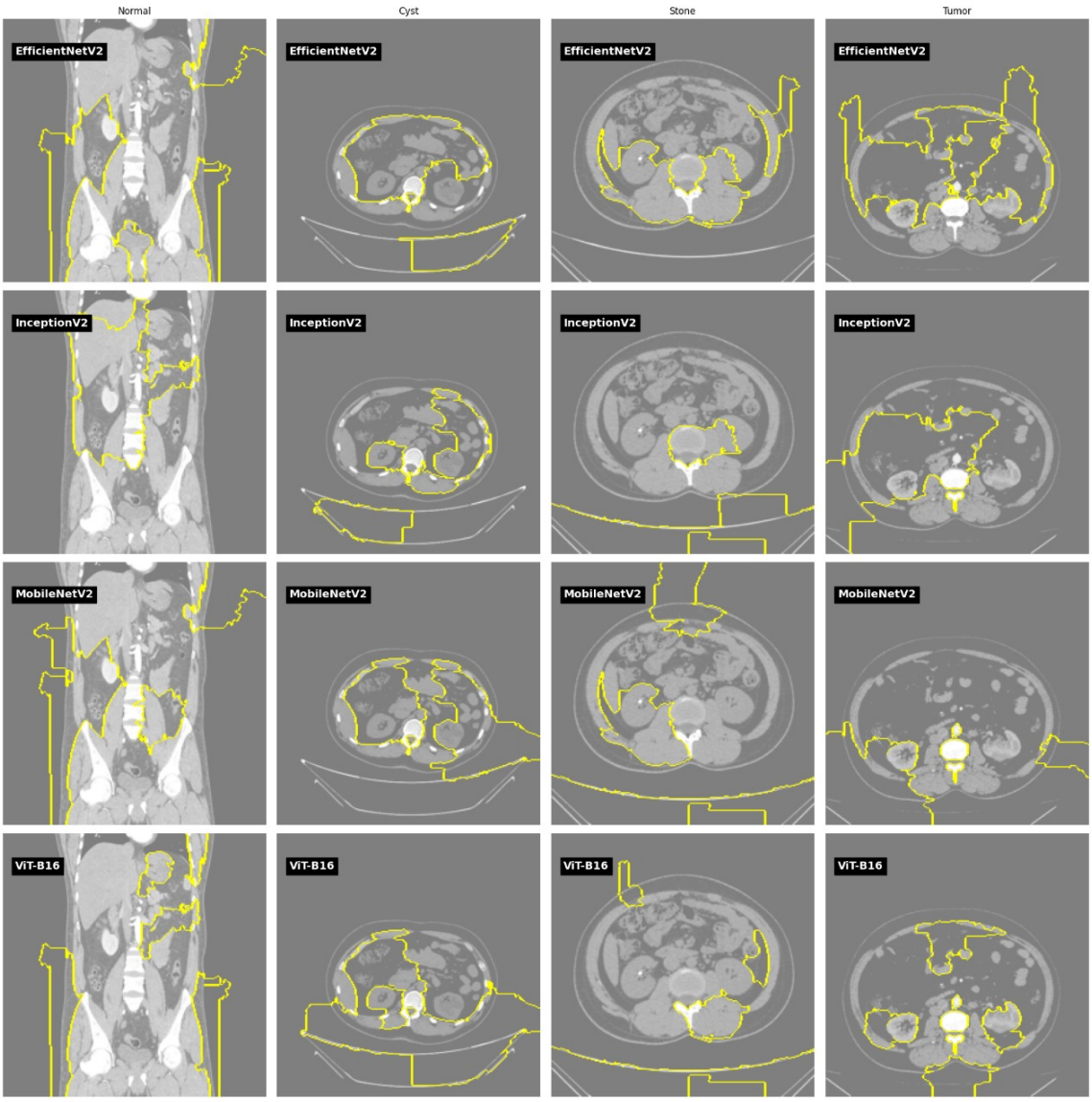

Figure 4: Methodology of Proposed Architecture

(ViT), in a harmonious manner. This ensemble approach is a strategic endeavor to harness the distinct strengths of diverse architectural paradigms and capture multifaceted feature representations, thereby potentially enhancing the overall classification performance. The model's design is a symphony of parallel branches, each composed of a pre-trained model architecture, that converges through a concatenation layer, allowing for the seamless fusion of hierarchical and global representations. The subsequent dense layers and the output layer orchestrate the amalgamated features, culminating in a robust ensemble model poised to deliver superior classification accuracy.

3.6 Explainable AI (XAI) Using LIME

In order to improve the comprehensibility of our deep learning model for detecting kidney disease, we utilized Explainable AI (XAI) methods, specifically employing Local Interpretable Model-agnostic Explanations (LIME). LIME offers clear and explicit explanations for the predictions provided by complicated models. It achieves this by generating simpler and more interpretable models that approximate the behavior of the 11 complex models for specific cases. This study employed LIME to detect and emphasize areas in medical images that had a significant effect on the decision-making process of the model. This provided crucial insights into the primary features that drove the model's predictions. FIGURE 4 illustrates the visual outcomes of implementing LIME on four different deep learning models: InceptionNet-V2, MobileNet-V2, EfficientNetV2, and Vision Transformer (ViT-B16). Each model was assigned the duty of classifying images under four different categories: Normal, Cyst, Stone, and Tumor. The yellow highlighted por tions in the figure indicate the areas of the images that had the most influence on the model's predictions. After implementing LIME, it was seen that certain models had exceptional performance for specific classes, as depicted in the figure. EfficientNetV2 showed a robust capability in accurately detecting areas that indicated the presence of cysts and tumors, whereas MobileNet-V2 proved to be more efficient in classifying normal and stone classes. In order to enhance the overall performance in terms of both robustness and accuracy, we combined these four models into an ensemble, focusing on their distinct advantages across various classes. The use of this ensemble technique guaranteed that the best possible model was not only precise but also dependable for all classifications of kidney disease. The incorporation of XAI using LIME pro vided a lucid comprehension of the decision-making process employed by the models, hence bolstering the reliability and interpretability of our ensemble model in clinical environments.

4. PERFORMANCE ANALYSIS

To establish a baseline before presenting the results of the deep learning models, we also performed a comparative analysis of traditional machine learning models, such as Logistic Regression, k-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Random Forest. We performed comparison so as to comprehensively examine our proposed deep learning ensemble model against traditional methods. Therefore, traditional approaches yielded low accuracy and low precision in classifications, which is strong evidence for the benefit of applying advanced deep learning architectures for this task. Through comparison of these models, we emphasize the unique benefits of our method in addressing the intricacies of the problem, resulting in greater accuracy, precision, and explainability.

4.1 Performance of Transfer Learning Models

The research studies applied pre-trained InceptionV3, EfficientNet, and MobileNet models and also applied ViT to determine the most significantly effective model for identifying and classifying Kidney disease using the CT kidney dataset. Our models were designed utilizing 128x128 images as input data. To fine-tune the hyperparameters, a batch size of 32 and 50 epochs was employed during the training phase. Given the multi-class nature of the dataset, the softmax activation function

was applied in the output layer. The model was compiled using the Adam optimizer and the categorical_crossentropy loss function. We achieved the ability to obtain an average accuracy for three different models, such as MobileNet-V2, achieving 87.25% accuracy, EfficientNet-V2, achieving 86.75% accuracy, InceptionNet-V2, achieving 83.25% accuracy, and ViT, achieving 91.5% accuracy.

Five performance indicators - accuracy (ACC), precision (PPR), recall or sensitivity (Sen), F1 score, and Area under the ROC curve (AUC) score - have been adapted for use with every analysis dataset in order to evaluate the proposed kidney disease classifier. On average, the mobileNet-V2 model obtained a score of 87.25% in terms of precision, 87.5% in terms of recall, 87.25% in terms of f1-score, and 92% in terms of AUC Score. For the efficientNet-V2 model, the average score obtained was 86.75% accuracy, 85.5% recall, 88.5% f1-score, and 90.75% AUC Score. For the InceptionNet-V2 model, the average scores were 83.25% precision, 79.75% recall, 80.75% f1-score, and 87.25% AUC score. For Transformer ViT, the average scores were 91.5% precision, 89.25% recall, 90% f1-score, and 93.25% AUC score. The performance indicators for each class are briefly presented in TABLE 1, for all models.

Table 1: Performance Analysis of Different Models

Model (Class)	Precision	Recall	F1-Score	AUC
MobileNET-V2				
Tumor	0.87	0.98	0.92	0.96
Cyst	0.96	0.89	0.92	0.93
Normal	0.79	0.80	0.80	0.89
Stone	0.87	0.83	0.85	0.90
EfficientNet-V2				
Tumor	0.94	0.89	0.91	0.93
Cyst	0.90	0.95	0.93	0.94
Normal	0.82	0.72	0.87	0.85
Stone	0.81	0.86	0.83	0.91
InceptionNet-V2				
Tumor	0.90	0.91	0.90	0.93
Cyst	0.86	0.91	0.89	0.91
Normal	0.77	0.54	0.63	0.76
Stone	0.80	0.83	0.81	0.89
ViT				
Tumor	0.90	0.98	0.94	0.97
Cyst	0.95	0.95	0.95	0.96
Normal	0.87	0.82	0.84	0.90
Stone	0.94	0.82	0.87	0.90

The confusion matrix for attentiveness models, namely mobileNet-V2, EfficientNet-V2, InceptionNet-V2, and ViT, is shown in FIGURE 5, respectively. The more efficient performance of the three deep learning models is evident when applied to CT Kidney datasets, with the method of attention of these models proving to be particularly effective. It is apparent that the classifier successfully classified a significant portion of the instances.

The ROC Curve plots for three transformed deep learning architectures, namely mobileNet-V2, EfficientNet-V2, InceptionNet-V2, and ViT, are shown in FIGURE 6, respectively. The models effectively distinguished between all positive and negative classes with a high degree of accuracy,

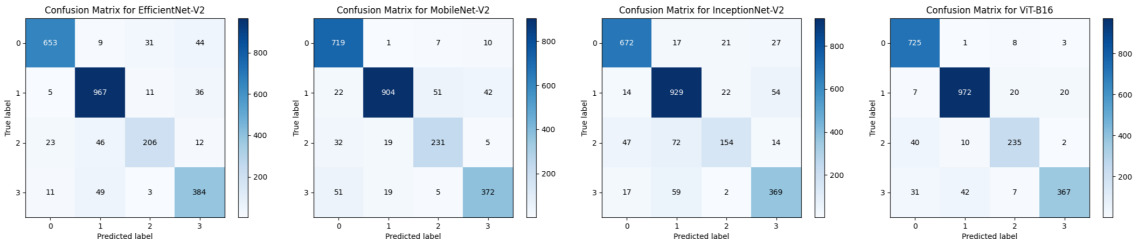

Figure 5: Confusion matrix for four models

as shown by the considerable AUC for all anomalies. Based on the test data, the ViT model predicted a very high true-positive rate(TPR), achieving 97% of the AUC values for tumor detection, which was important since the ROC curve was reliant on the TPR and the FPR and average AUC score is 93.25%. MobileNetV2 model also performed the highest average score from transfer learning models. Therefore, even for classes with non-uniform sample distributions, these findings indicated that the Transformer ViT model was more robust and consistent.

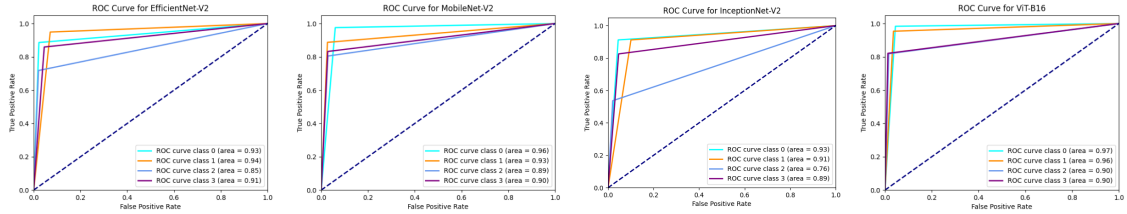

Figure 6: ROC Curve for four models

Precision is a measure of accuracy that evaluates the proportion of accurately identified positive samples (True Positive) out of the total number of identified positive samples. It serves to analyze the validity of the machine learning model in classifying the model as positive. The percentage of positive samples that are accurately identified as positive samples to the total number of accurate positive samples is called recall. The effectiveness of the model in identifying positive samples can be measured by the recall. When it comes to producing a great machine learning model that generates outcomes that are more precise and accurate, these values are essential. Following the completion of the measurement, we were able to achieve the result shown in FIGURE 7 for precision compared to recall. When compared to other transfer learning models, ViT obtained a much higher score, which had an average score of 96.75%, than other models. ViT model achieved significant precision-recall scores for tumors and cysts, 99% and 99%, respectively.

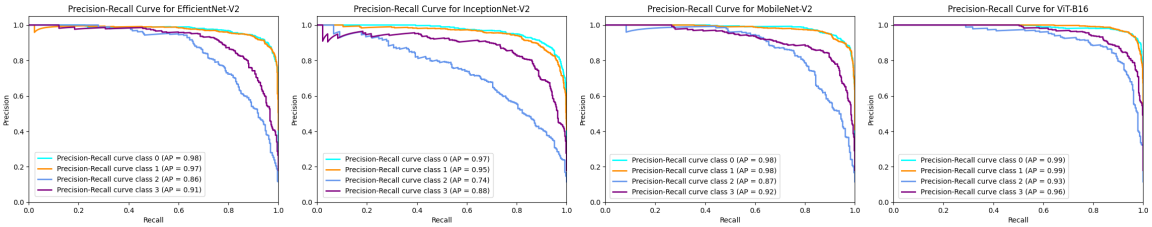

Figure 7: Precision vs Recall for four models

After analyzing all models, we show that the ViT model achieved the highest score for precision, recall, f1-score, AUC score, and precision-recall score. When compared to all transfer learning models, the mobileNetV2 model also performed significantly in detecting Kidney disease identification.

4.2 Performance of Ensemble Model

We used the ensemble method for the proposed work, which actually based on ViT and a pre-trained model. After training our ensemble model, we get outperforming results that accurately predict kidney disease. We achieved significant results from the ensemble model accuracy is 96%, recall and f1-score are 97% and 96.5%, respectively. When we compare our ensemble model performance from other baseline models where we got precision value for the tumor is 98%, the cyst is 100%, 92%, and 94%, which is a satisfactory score than baseline MobileNet-v2, EfficientNet-V2, InceptionNet-V2, and ViT models which we see to TABLE 1. We show in TABLE 2, that other performance indicators recall, f1-score, and AUC score, are impressive for our ensemble model than baseline models to compare TABLE 1.

In comparison to the curve, the ensemble model Performed significantly better for individual's ROC and Precision-recall curve than other distinct models. The ensemble model achieved a 98% average score from the ROC curve and a 99.5% score from the precision-recall score. We decided that after all performances, the ensemble model performed superior and robustly, which will contribute identification of CKD in the medical sector for Confusion matrix, precision-recall and ROC Curve are FIGURE 8–FIGURE 10, respectively.

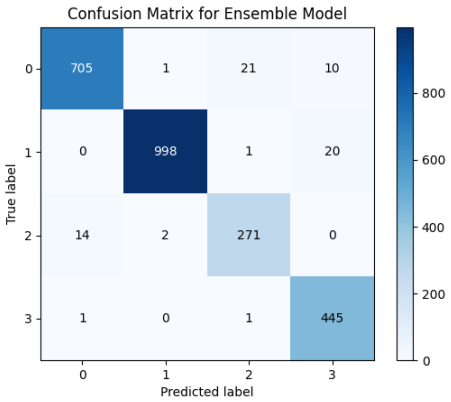

Figure 8: Ensemble Model Confusion Matrix

4.3 Comparison with Existing Models

The top-performing models from both the most recent literature and the proposed study are compared and analyzed based on their evaluation accuracy, as shown in TABLE 3. The findings show that in order to move the field forward and become better at identifying images and related tasks, it's crucial to look at and compare various models and approaches. The ensemble model's most

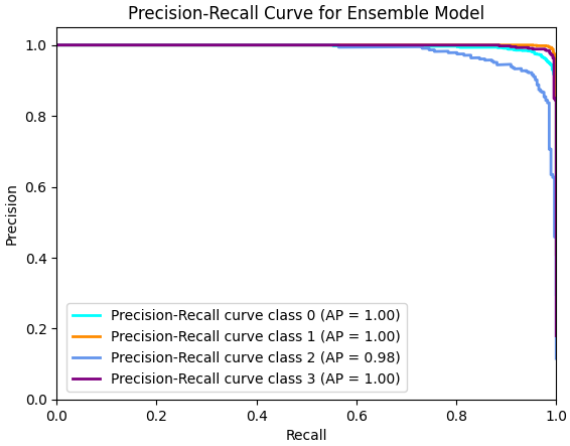

Figure 9: Ensemble Model Precision Recall

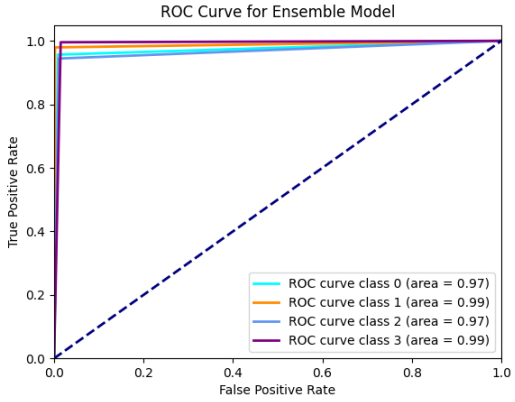

Figure 10: Ensemble Model ROC

incredible accuracy of 90% and AUC of 95% demonstrated the model's effectiveness in assisting clinical decision-making about the prognosis of kidney disease.

Table 2: Performance Analysis of the ensemble model

Classes	Precision	Recall	F1-Score	AUC
Tumor	0.98	0.96	0.97	0.97
Cyst	1.0	0.98	0.99	0.99
Normal	0.92	0.94	0.93	0.97
Stone	0.94	1.0	0.97	0.99

Table 3: Comparison with Existing Works

References	Dataset Size	Model(s)	Accuracy
Zabihollahy et al. (2020) [43]	315	Convolutional neural network	83.75%
Akgun et al. (2021) [44]	460	MobileNet ResNet50	86.42% 82.06%
Kalkan et al. (2021) [45]	5000	ResNet152V2 MobileNetV2	89.58% 88.80%
Jee et al. (2022)[45]	1596	Inception V3 MobileNet	74.3% 72.37%
This Work (2021) [46]	12466	Ensemble Method	96%

5. COMPUTATIONAL EFFICIENCY AND DEPLOYMENT

The model was trained and evaluated using TensorFlow on Kaggle, leveraging an NVIDIA Tesla P100 GPU with 16GB RAM. A batch size of 32 was used to optimize GPU memory utilization, and the model was trained for 100 epochs on a kidney disease dataset containing 12,466 images. Each epoch took approximately 30 minutes, resulting in a total training duration of 50 hours. Training Vision Transformers (ViT) and EfficientNetV2 is computationally expensive due to their high parameter count and complex attention mechanisms. While the P100 GPU provided adequate computational power, deploying the model in real-world clinical settings presents challenges such as hardware constraints, inference latency, and regulatory compliance. The average inference time was X milliseconds per image, ensuring near-real-time performance on GPUs but potentially increasing on CPU-based systems, which are more common in hospitals. Techniques such as quantization, pruning, and TensorRT acceleration can help optimize inference speed and reduce computational demands. Additionally, regulatory approvals like FDA and CE marking require extensive validation and explainability, while seamless integration with Electronic Health Records (EHRs) and existing imaging platforms is crucial for adoption. Future work should focus on model compression (quantization-aware training, knowledge distillation), inference optimization (TensorRT, ONNX Runtime), and validation on CPU-based systems to ensure accessibility and efficiency. Addressing these factors will enable effective deployment, making AI-driven diagnostics more practical for real-world clinical environments.

6. CONCLUSION

Early detection and classification of kidney disease may save human lives. Detection procedures that are done manually are often laborious and dependent on the knowledge of medical professionals. The advancement of automated classification systems is thus highly encouraging, as they provide reliable and quick outcomes. Improved accuracy and reliability in kidney disease detection may be achieved by utilizing deep learning methods such as pre-trained models and ViT. According to our findings, ViT had the highest accuracy (90%) when compared to all of the fine-tuned models, then MobileNet-V2, EffecientNet-V2, and InceptionNet-V3.However, The ensemble model achieved a robust performance that is 96% which is more accurate than baseline models. The research also conducted a manual error analysis to improve the performance of the pre-trained models. This might result in more accurate diagnoses and improved treatment choices for individuals with kidney diseases.

For future studies, it would be prudent to experiment with the MobileNet-V2, EffecientNet-V2, and InceptionNet-V2 optimized Tranfer learning models and ensemble models on other datasets to see how well they hold up in different types of clinical situations. To make the proposed strategy more practical, it is essential to study other datasets with diverse demographics, patient groups, and quality imaging levels. An in-depth review of the model's abilities and shortcomings will reveal any adjustments that may be necessary for its successful implementation in clinical practice. Finally, by tackling these potential future research directions, we may create a kidney tumor identification approach that is more comprehensive and flexible, making it applicable and successful in more clinical circumstances.

References

[1] https://www.biologyonline.com/dictionary/disease

[2] https://my.clevelandclinic.org/health/body/21824-kidney

[3] https://www.niddk.nih.gov/health-information/kidney-disease/kidneys-how-they-work

[4] https://www.mayoclinic.org/diseases-conditions/chronic-kidney-disease/symptoms-causes/syc-20354521

[5] https://www.kidneyfund.org/all-about-kidneys/types-kidney-diseases

[6] https://www.cdc.gov/kidneydisease/basics.html

[7] https://www.kidney.org/atoz/content/about-chronic-kidney-disease#causes

[8] https://www.niddk.nih.gov/health-information/kidney-disease/high-blood-pressure

[9] https://ncdalliance.org/why-ncds/ncds/chronic-kidney-disease

[10] https://www.kidney.org/kidneydisease/global-facts-about-kidney-disease

[11] Arifuzzaman M, Chowdhury MJ, Ahmed I, Siddiky MN, Rashid D. Heart Disease Prediction Through Enhanced Machine Learning and Diverse Feature Selection Approaches. In: 10th International Conference on Smart Instrumentation Measurement and Applications (ICSIMA). IEEE. 2024:119-124.

[12] https://my.clevelandclinic.org/health/diseases/17689-kidney-failure

[13] Levey AS, Cattran D, Friedman A, Miller WG, Sedor J, et. al. Proteinuria as a Surrogate Outcome in Ckd: Report of a Scientific Workshop Sponsored by the National Kidney Foundation and the US Food and Drug Administration. Am J Kidney Dis. 2009;54:205-226.

[14] Chapman JR. What Are the Key Challenges We Face in Kidney Transplantation Today? Transplant Res. 2013;2:1-7.

[15] Gerogianni S, Babatsikou F, Gerogianni G, Grapsa E, Vasilopoulos G, et. al. Concerns of Patients on Dialysis: A Research Study. Health Science Journal. 2014;8:423.

[16] Hernández D, Caballero A. Kidney transplant in the next decade: Strategies, challenges and vision of the future. Nefrología. 2023;43:281-292.

[17] Ma F, Sun T, Liu L, Jing H. Detection and Diagnosis of Chronic Kidney Disease Using Deep Learning-Based Heterogeneous Modified Artificial Neural Network. Future Gener Comput Syst. 2020;111:17-26.

[18] MHesamian MH, Jia W, He X, Kennedy P. Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges. J Digit Imaging. 2019;32:582-596.

[19] Chowdhury MJU, Kibria S. Performance Analysis for Convolutional Neural Network Architectures Using Brain Tumour Datasets: A Proposed System. In 2023 International workshop on Artificial Intelligence and Image Processing (IWAIIP). 2023:195–199.

[20] Singh V, Asari VK, Rajasekaran R. A Deep Neural Network for Early Detection and prediction of Chronic Kidney Disease. Diagnostics. 2022;12:116.

[21] Majid M, Gulzar Y, Ayoub S, Khan F, Reegu FA, et al. Enhanced Transfer Learning Strategies for Effective Kidney Tumor Classification With CT Imaging. IJACSA. 2023;14.

[22] Sudharson S, Kokil P. An Ensemble of Deep Neural Networks for Kidney Ultrasound Image Classification. Comput Methods Programs Biomed. 2020;197:105709.

[23] Kim DH, Ye SY. Classification of Chronic Kidney Disease in Sonography Using the Glcm and Artificial Neural Network. Diagnostics. 2021;11:864.

[24] Rady EH, Anwar AS. Prediction of Kidney Disease Stages Using Data Mining Algorithms. Inform Med Unlocked. 2019;15:100178.

[25] Bhandari M, Yogarajah P, Kavitha MS, Condell J. Exploring the Capabilities of a Lightweight CNN Model in Accurately Identifying Renal Abnormalities: Cysts Stones and Tumors Using Lime and Shap. Appl Sci. 2023;13:3125.

[26] Bhattacharjee A, Rabea S, Bhattacharjee A, Elkaeed EB, Murugan R, et. al. A Multi-Class Deep Learning Model for Early Lung Cancer and Chronic Kidney Disease Detection Using Computed Tomography Images. Front Oncol. 2023;13:1193746.

[27] Kanwal S, Khan MA, Fatima A, Al-Sakhnini MM, Sattar O, et. al. Ia2skabs: Intelligent Automated and Accurate System for Classification of Kidney Abnormalities. International Conference on Cyber Resilience (ICCR). IEEE. 2022:1–10.

[28] Wasi S, Alam SB, Rahman R, Amin MA, Kobashi S. Kidney Tumor Recognition From Abdominal CT Images Using Transfer Learning. In: IEEE 53rd International Symposium on Multiple-Valued Logic (ISMVL). IEEE. 2023:54-58.

[29] Yildirim K, Bozdag PG, Talo M, Yildirim O, Karabatak M, et. al. Deep Learning Model for Automated Kidney Stone Detection Using Coronal CT Images. Comput Biol Med. 2021;135:104569.

[30] Ma F, Sun T, Liu L, Jing H. Detection and diagnosis of chronic kidney disease using deep learning-based heterogeneous modified artificial neural network. Future Gener Comput Syst. 2020;111:17-26.

[31] Heller N, Isensee F, Maier-Hein KH, Hou X, Xie C, et. al. The State of the Art in Kidney and Kidney Tumor Segmentation in Contrast-Enhanced CT Imaging: Results of the KITS19 Challenge. Med Image Anal. 2021;67:101821.

[32] Srivastava S, Yadav RK, Narayan V, Mall PK. An Ensemble Learning Approach for Chronic Kidney Disease Classification. J Pharm Neg Results. 2022;13:2401-2409.

[33] Bhaskar N, Manikandan S. A Deep-Learning-Based System for Automated Sensing of Chronic Kidney Disease. IEEE Sens Lett. 2019;3:1-4.

[34] Schena FP, Anelli VW, Trotta J, Di Noia T, Manno C, et. al. Development and Testing of an Artificial Intelligence Tool for Predicting End-Stage Kidney Disease in Patients With Immunoglobulin a Nephropathy. Kidney Int. 2021;99:1179-1188.

[35] Chen G, Ding C, Li Y, Hu X, Li X, et. al. Prediction of Chronic Kidney Disease Using Adaptive Hybridized Deep Convolutional Neural Network on the Internet of Medical Things Platform. IEEE Access. 2020;8:100497–100508.

[36] Islam MN, Hasan M, Hossain MK, Alam MG, Uddin MZ, et. al. Vision Transformer and Explainable Transfer Learning Models for Auto Detection of Kidney Cyst Stone and Tumor From Ct-Radiography. Sci Rep. 2022;12:11440.

[37] Kuo CC, Chang CM, Liu KT, Lin WK, Chiang HY, et. al. Automation of the Kidney Function Prediction and Classification Through Ultrasound-Based Kidney Imaging Using Deep Learning. NPJ Digit Med. 2019;2:29.

[38] Pande SD, Agarwal R. Multi-class kidney abnormalities detecting novel system through computed tomography. IEEE Access. 2024;12:21147-21155.

[39] Tan M, Le Q. Efficientnetv2: Smaller models and faster training. In International conference on machine learning. PMLR.2021:10096-10106.

[40] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception Architecture for Computer Vision. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016:2818-2826.

[41] Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. MobileNetV2: Inverted Residuals and Linear Bottlenecks. in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2018:4510-4520.

[42] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. 2021. Arxiv Preprint: https://arxiv.org/pdf/2010.11929

[43] Zabihollahy F, Schieda N, Krishna S, Ukwatta E. Automated Classification of Solid Renal Masses on Contrast-Enhanced Computed Tomography Images Using Convolutional Neural Network With Decision Fusion. Eur Radiol. 2020;30:5183-5190.

[44] Akgün D, KABAKUŞ AT, ŞENTÜRK ZK, ŞENTÜRK A, Küçükkülahli E. A Transfer Learning-Based Deep Learning Approach for Automated Covid-19 DIAGNOSIS With Audio Data. Turk J Electr Eng Comput Sci. 2021;29:2807-2823.

[45] Kalkan M, Bostancı GE, Güzel MS, Kalkan B, Özsarı Ş, et .al. Cloudy/Clear Weather Classification Using Deep Learning Techniques With Cloud Images. Comput Electr Eng. 2022;102:108271.

[46] Lee SW. Novel Classification Method of Plastic Wastes With Optimal Hyperparameter Tuning of Inception_resnetv2. 2021:274-279.