Digital transformation in healthcare: leveraging machine learning for predictive analytics in chronic kidney diseases prevention

Eshrag Ali Refaee
Department of Engineering and Computer Sciences, Jazan University, Jazan, Saudi Arabia

Received 3 March 2025
Revised 10 May 2025
Accepted 21 May 2025

Abstract

Purpose – This study investigates the application of machine learning (ML) to enhance the early detection of Chronic Kidney Disease (CKD), addressing the limitations of traditional binary classification methods by implementing a multiclass classification approach using medical imaging data.

Design/methodology/approach – We conducted an empirical evaluation of five deep learning models – Convolutional Neural Network (CNN), ResNet50, EfficientNetB0, Vision Transformer (ViT) and a Hybrid CNN-Transformer architecture – using a publicly available dataset comprising 12,446 CT images. These images are categorised into four classes: cyst, normal, stone and tumour. Models were trained to perform a four-way classification to predict the presence and type of CKD-related conditions.

Findings – Among the tested models, ResNet50 achieved the highest performance, with an F1 score of 0.781 and an accuracy of 0.8304. The findings indicate that deep learning algorithms, particularly ResNet-based architectures, offer high diagnostic precision for multiclass CKD prediction using image data.

Practical implications – The proposed approach offers scalable solutions for healthcare systems to automate early CKD screening, potentially reducing diagnostic delays and improving patient outcomes. The implementation of such models within clinical workflows can enhance diagnostic consistency, assist overburdened practitioners and facilitate the deployment of AI-assisted decision support tools in radiology and nephrology settings.

Originality/value – This study presents a novel contribution by shifting the focus from binary to multiclass CKD classification using a large-scale image dataset. It provides a comparative analysis of modern deep learning architectures and highlights their applicability in clinical decision support. The work paves the way for more nuanced diagnostic tools and offers a replicable framework for future research in chronic disease detection.

Keywords Machine learning, Chronic kidney disease, Medical imaging, Deep learning, Digital healthcare, Classification

Paper type Research paper

1. Introduction

Chronic diseases – long-lasting health conditions that require ongoing medical care or limit daily activities – are a growing global concern. These noncommunicable diseases (NCDs), including cardiovascular disorders, diabetes, chronic kidney disease (CKD) and respiratory illnesses, now represent the leading cause of death and disability worldwide. The World Health Organization (WHO) reports that NCDs account for approximately 74% of all global deaths, totalling 41 million annually (WHO, 2024). This epidemiological shift reflects both increasing life expectancy and the widespread prevalence of modifiable risk factors such as poor diet, physical inactivity, tobacco use and excessive alcohol consumption.

The implications of this trend are both medical and economic. As the burden of chronic disease grows, so does the urgency for effective prevention strategies, including early

Disclosure statement: The authors report no conflict of interest.

Journal of Innovative Digital Transformation
Vol. 2 No. 3, 2025
pp. 217-232
Emerald Publishing Limited
e-ISSN: 2976-906X
p-ISSN: 2976-9051
DOI 10.1108/JIDT-03-2025-0012

screening and prediction tools. AI-driven predictive analytics, particularly those based on machine learning (ML), are increasingly used to analyse risk indicators and enable proactive care interventions. These approaches consider a broad array of risk factors – ranging from demographic variables like age and sex, to behavioural, clinical and genetic influences – and enable identification of at-risk individuals before the onset of severe complications (Locatelli *et al.*, 2002; Benziger *et al.*, 2016).

Among these conditions, CKD stands out as a critical public health challenge. Often asymptomatic in its early stages, CKD can progress undetected until it reaches an advanced phase where treatment becomes difficult and costly. In 2022, CKD was the ninth leading cause of death in the United States, accounting for nearly 58,000 deaths and a mortality rate of 17.4 per 100,000 people (Health Statistics, 2024). Identifying those at risk of CKD at an early stage can help mitigate disease progression and alleviate both personal and systemic burdens.

In response to this pressing issue, our study explores the potential of ML models to predict CKD risk using image-based medical data. Most existing studies focus on binary classification using structured tabular records. In contrast, this research addresses a key limitation in current literature by employing a multiclass classification task on an image-based dataset.

This study addresses a key gap in the current literature, which has focused heavily on binary classification of kidney disease using tabular patient records. By shifting focus to multiclass classification on a diverse image-based dataset, this work not only enhances diagnostic specificity but also evaluates how well advanced models like ResNet50 and Vision Transformers perform in real-world, multi-type CKD scenarios. This contributes to practical deployment in diagnostic imaging workflows.

We compare multiple state-of-the-art deep learning models to assess their effectiveness in distinguishing between four classes – cyst, normal, stone and tumour. The performance of each model provides insight into the strengths of different architectural designs for clinical application.

The theoretical contribution of this research lies in its comparative analysis of distinct deep learning architectures, revealing how model design – such as residual connections or transformer-based attention – impacts diagnostic accuracy. These findings offer guidance for theory-driven model selection in future medical imaging studies. Future investigations can refine predictive accuracy and broaden the scope of applications by incorporating additional patient features and exploring a wider range of ML algorithms. Ultimately, this research underscores the transformative potential of data-driven approaches in fostering timely and adequate medical decision-making.

2. Related work

Early detection of CKD is critical as the disease often progresses silently. CKD often remains asymptomatic in its early stages, making it challenging to diagnose without advanced analytical tools. If left undetected, it can progress to end-stage renal disease (ESRD), necessitating costly interventions such as dialysis or kidney transplantation. As a result, early detection of CKD is critical to mitigating its impact on individuals and healthcare systems. Recent advances in data science, particularly ML, have opened new avenues for automating the detection of CKD using patient data, including demographic, clinical and laboratory information (Delpino *et al.*, 2022). Several studies have proposed ML methods to improve CKD diagnosis using either structured clinical records or medical imaging data.

Mathew and Corso (2009) reviewed CKD screening frameworks in Australia, concluding that opportunistic screening through primary care is sustainable and effective. James *et al.* (2010) conducted a systematic literature review that highlighted diagnostic tools such as GFR estimation and proteinuria testing. These studies collectively stress the importance of embedding early detection in routine healthcare and tailoring it across demographics.

Structured datasets like the UCI CKD dataset have been widely used. Salekin and Stankovic (2016) applied k-NN, RF and neural networks, reporting an F1 score of 0.993 with

RF. Singh *et al.* (2022) developed a DNN architecture using preprocessing techniques like imputation, normalisation and RFE-based feature selection. Their model achieved 100% accuracy, outperforming traditional classifiers. Iftikhar *et al.* (2023) tested multiple models, including SVM, LR, DT and RF, on a local Pakistani dataset of 382 cases, with SVM reaching an F1 score above 0.93.

Delpino *et al.* (2022) compared models using AUC, observing scores between 0.74 and 1, with high performance for RF and DNNs. Nimmagadda *et al.* (2023) explored hybrid models, combining supervised and unsupervised techniques with deep learning. Their best model, RF, achieved 100% accuracy. They also experimented with saliva-based biomarkers and biosensor-based detection.

Mustafizur Rahman *et al.* (2024) used ensemble learning on the UCI dataset with SMOTE and MICE for data balancing, achieving an F1 score of 0.996. Mamatha and Terdal (2024) used CNNs on 4k Kaggle records, incorporating feature selection and scaling, though performance metrics were not disclosed.

Bahrami *et al.* (2024) built gender-specific RF and NN models using 6.8k samples from Iran, with the RF model for female records attaining 89.07% accuracy. Arifuzzaman *et al.* (2024) proposed an ensemble deep learning framework applied to a 12.4k CT image dataset split into four classes. Using EfficientNetV2, MobileNetV2 and ViT, they reported 91.5% accuracy with ViT.

Other studies leveraged imaging data. Ma *et al.* (2020) introduced a HMANN for analysing ultrasound images. Mohamad Tabish *et al.* (2024) employed deep learning models like ResNet50, MobileNet and InceptionV3, reaching 98% accuracy with ResNet50. Nagawa *et al.* (2024) used a 3D CNN on MRI data from 321 patients, reporting 86.2% accuracy and 93.6% AUC across three CKD severity levels.

Yogesh *et al.* (2025) proposed an approach to enhance the classification accuracy of CKD by integrating a hybrid feature selection method with an ensemble deep learning model. For this purpose the utilised Boruta Algorithm to rank features based on their importance in CKD classification. After that, they applied Recursive Feature Elimination (RFE) to iteratively remove the least significant features, refining the feature set obtained from Boruta. Finally, the refined feature set serves as input to the proposed ensemble deep learning classifier. The model's performance is compared against traditional classifiers like SVM, LR and RF. The best results were reported using the ensemble deep learning model, when combined with the hybrid feature selection, achieving a classification accuracy of 97.5%, outperforming the other models the used.

Table 1 presents a comparative summary of existing CKD detection studies. This comparison helps contextualise our contribution within the broader landscape of ML approaches to early CKD prediction.

While these studies collectively demonstrate strong performance for ML-based CKD detection, the majority focus on binary classification and often rely on limited or demographically constrained datasets. Our study addresses these limitations by employing multiclass classification on a large-scale image dataset and evaluating several deep learning architectures tailored for real-world diagnostic challenges. Nevertheless, there remains substantial variability in data sources, ranging from small-scale patient records to localised imaging datasets, which limits generalisability. This disparity highlights the importance of developing more standardised and scalable frameworks capable of accommodating heterogeneous data across clinical contexts, thereby advancing more inclusive and accurate CKD diagnostic systems.

3. Experimental framework

This section presents a comprehensive overview of the experimental framework employed in this study. This includes a detailed description of the dataset utilised, the architectural design of the model and the parameters configured for optimal performance. The workflow pipeline

Table 1. Summary of literature on the importance of early detection of CKD

Study	Dataset	Data instances	Methodology	Results
Mathew and Corso (2009)	Australian Institute of Health and Welfare reports	Prevalence of CKD, its stages and risk factors	Medical screening (non-ML-based)	Not reported
James *et al.* (2010)	Records from US National Health	Prevalence of CKD and risk factors	Medical screening (non-ML-based)	Not reported
Salekin and Stankovic (2016)	UCI CKD Dataset	400 patient records – India	kNN, RF and Neural Networks	F1 = 0.993 with RF
Singh *et al.* (2022)	UCI CKD Dataset	400 patient records – India	Deep NN with Adam and ReLU	Acc. = 100
Iftikhar *et al.* (2023)	Patient records from a local district in Pakistan	382 samples (240 CKD, 142 non-CKD) with 21 features	LR, PB, RF, DT, KNN, SVM	acc. = 90% and F1 = 0.93
Nimmagadda *et al.* (2023)	UCI CKD Dataset	400 patient records – India	SVM, RF, NB, DT and CNN	Acc. = 100% with RF
Mustafizur Rahman *et al.* (2024)	UCI CKD Dataset	400 patient records – India	SVM SMOTE	F1 = 0.996
Mohamad Tabish *et al.* (2024)	Screening Images Dataset	Screening images	ResNet50, MobileNet and InceptionV3	Acc. = 98% with ResNet50
Mamatha and Terdal (2024)	Chronic Kidney Disease dataset	4k patient records	CNN	Not reported
Bahrami *et al.* (2024)	Mashhad Stroke and Heart Atherosclerotic Disorders dataset	6.8k patient records – Iran	RF and NN	Acc. = 89.07% with RF
Nagawa *et al.* (2024)	Saitama Medical University Hospital dataset	MRI images from 321 patients	3D CNN based on ResNet-18 arch	Acc. = 86.2% and ROC-AUC = 93.6%
Yogesh *et al.* (2025)	UCI CKD Dataset	400 patient records – India	SVM, LR, RF and ensemble DL	Acc. = 97.5%

highlights the critical steps in data preprocessing and implementing ML algorithms. Furthermore, the parameter tuning strategies undertaken to enhance the model's accuracy and efficiency are reviewed in detail.

3.1 Data collection

In this study, we employ a publicly available dataset to facilitate further research and ensure reproducibility of results, thereby advancing the field of chronic disease prediction. Specifically, we utilise an image-based dataset of 12.4k images with four classes (Islam *et al.*, 2022). The classes are Normal, Cyst, Tumour and Stone. Figures 1–3 show two example input images of the dataset representing the stone and tumour classes respectively [1]. We noticed the binary classification on the patient records dataset used in previous work has attained high accuracy scores up to 100% (Singh *et al.*, 2022). As such, we opted for this relatively larger image-based dataset with four-class to make the task more challenging for the models and achieve a more precise diagnosis for the patients. From this point on we will refer to the dataset we use in this work as K4DD (see Table 2).

The four classes are defined as follows.

(1) Cyst is a fluid-filled sac that forms on or inside the kidney. These cysts can be simple (harmless) or complex (potentially cancerous or linked to kidney diseases). It has the characteristics of being typically round or oval in shape and may be filled with clear fluid.

Figure 1. An example input image of *stone* class from the k4DD dataset. Source: Figure courtesy of Islam *et al.* (2022)

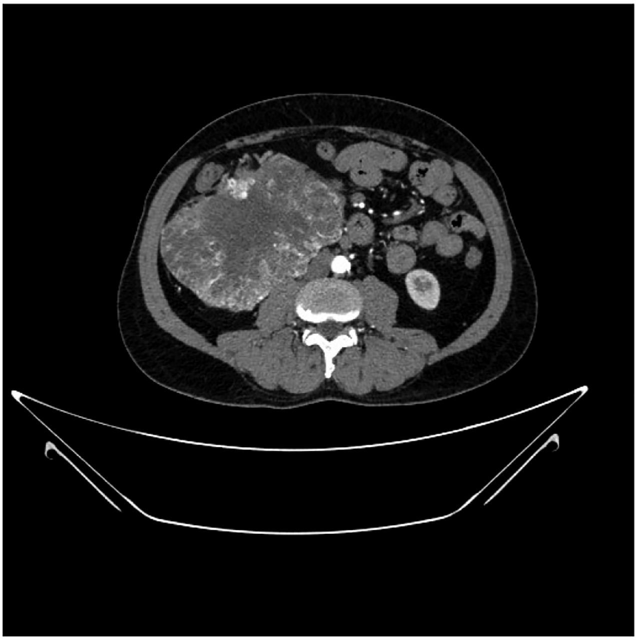

Figure 2. An example input image of *tumour* class from the k4DD dataset. Source: Figure courtesy of Islam *et al.* (2022)

(2) Normal refers to a healthy kidney functioning properly without abnormalities, obstructions or diseases. It maintains the body's fluid and electrolyte balance, filters waste from the blood and produces essential hormones such as erythropoietin (which regulates red blood cell production).

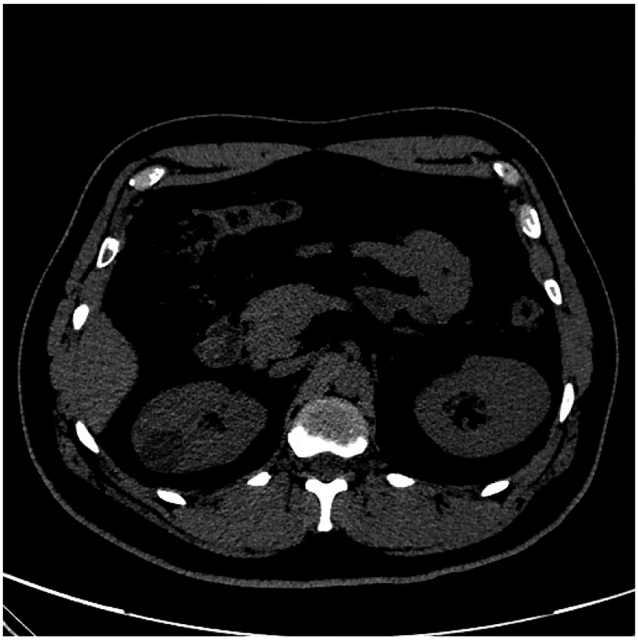

Figure 3. An example input image of *cyst* class from the k4DD dataset. Source: Figure courtesy of Islam *et al.* (2022)

Table 2. The class distribution in the KDD dataset shows the number of instances in each of the four classes

Class	No. of instances	Percentage
Cyst	3,709	29.80%
Normal	5,077	40.79%
Stone	1,377	11.06%
Tumour	2,283	18.34%
Total	*12,446*	*100%*

(3) Stone is kidney with a solid mass of crystals of minerals and salts, such as calcium oxalate, uric acid or struvite. These stones form inside the kidneys and can move through the urinary tract, sometimes causing severe pain.

(4) Tumour refers to an abnormal growth of cells in the kidney. It can be benign (non-cancerous) or malignant (cancerous). It has the characteristics of being a solid mass, unlike cysts, which are fluid-filled and can grow and spread (if malignant).

3.2 Data pre-processing

The obtained data was cleaned up using the steps utilised in previous work that involves image classification.

(1) Standardising image dimensions (height and width) to 224, 224 respectively, which is a standard size for models like ResNet and EfficientNet that we use in this work.

(2) We fixed the patch size at 32.

(3) We fixed the number of epochs at five (see Figure 4). Our preliminary experiments have shown variable performance. We experimented with up to ten epochs and based on our observations we opted to use five epochs with all experiments reported in this work.

(4) We scale the images' pixel values from 0 to 255 to 0–1, as a common practice in neural networks.

(5) The images were converted from the blue, green and red (BGR) format to the red, green and blue (RGB) format using OpenCV.

(6) The dataset was randomly split to 90% training and 10% of instances used for testing.

3.3 Machine learning models

In this section, we detail the machine models that we use in our experimental study.

3.3.1 Convolutional neural network (CNN). CNN automatically learns relevant features from images and classifies them into one of the specified categories, while incorporating regularisation techniques to enhance the model's generalisability. Using the Keras Sequential API, our system constructs and compiles a convolutional neural network (CNN) model. The function takes two parameters as input: "shape", which defines the dimensions of the input images (e.g. height, width and number of channels) and "number of classes", which specifies the number of output categories for the classification task. The model is built as a sequential stack of layers.

(1) The first convolutional layer applies 32 filters, each of size 3×3, to the input image. The ReLU activation function introduces non-linearity.

(2) The pooling layer reduces the feature maps' spatial dimensions (height and width) by taking the maximum value in each 2×2 window, which helps reduce computational complexity and control overfitting.

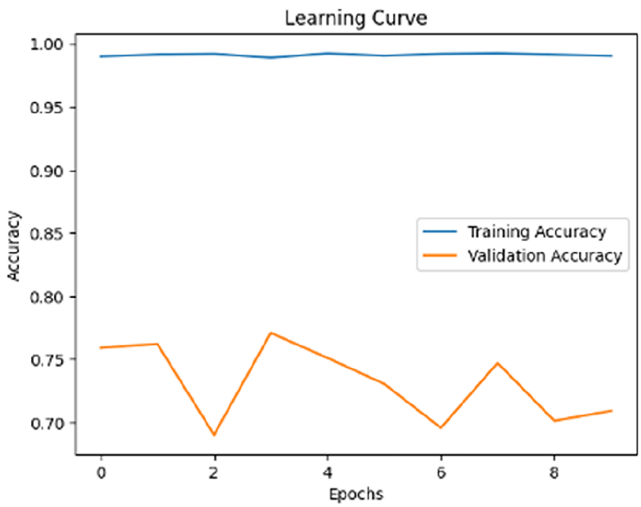

Figure 4. The overall accuracy performance concerning the number of epochs. Source: Authors' own work

(3) The dropout layer with a dropout rate of 25%, this layer randomly disables 25% of the neurons during training, reducing the likelihood of overfitting.

(4) The second convolutional layer increases the number of filters to 64, allowing the model to capture more complex features from the previous layer's outputs.

(5) The second pooling layer downsamples the feature maps. The second dropout layer regularises the network by dropping out 25% of the neurons.

(6) The flattening layer converts the 2D feature maps into a 1D vector, making the data suitable for the fully connected (dense) layers.

(7) The dense layer has 128 neurons and introduces non-linearity using the ReLU activation function. It acts as a hidden layer to learn high-level features.

(8) The third dropout layer applies 50% of dropout rate, which is more aggressive to prevent overfitting before the final classification.

(9) The final output layer, wherein the softmax activation function is used to output a probability distribution over the classes, making it suitable for multi-class classification, we use in this work. Adam Optimiser is used for gradient descent optimization, which adapts the learning rate during training.

3.3.2 ResNet50. ResNet50 has shown to be effective in previous work that involves image classification. Our system uses a ResNet50 model with customized parameters for image classification. The model is created without pre-trained weights, meaning it starts with randomly initialized weights and will be trained from scratch. The input shape parameter is set to the predefined dimension discussed earlier in the data preprocessing section. With ResNet50 we specified that the images have three colour channels (RGB). The classes parameter is assigned the value train generator, which dynamically sets the number of output neurons in the final classification layer to match the number of classes in the training dataset. We use a 50-layer deep CNN (ResNet50) that utilises residual connections to alleviate the vanishing gradient problem, ensuring efficient training even in very deep architectures.

3.3.3 EffecientNetB0. The EfficientNetB0 model we use is configured for an image classification task. The EfficientNetB0 constructor is initialised with random weights. The input shape parameter specifies the dimensions of the input images, where height and width are variables representing the height and width of the images, while the images have three colour channels (red, green and blue).

3.3.4 Vision Transformer (ViT) model. Our system constructs a Vision Transformer (ViT) model using TensorFlow and Keras.

(1) The input layer accepts images with the specified shape.

(2) The model then divides each image into non-overlapping patches by setting a patch size of 16 pixels. It calculates the total number of patches as the product of the image dimensions divided by the patch size.

(3) A convolutional layer (Conv2D) is employed with a kernel size and stride equal to the patch size, effectively extracting patches and projecting each into a 64-dimensional vector. The output is reshaped into a two-dimensional tensor of shape, where each row represents a patch (see Figure 5).

(4) The model incorporates positional embeddings by creating a trainable TensorFlow variable with dimensions. This variable is initialised with values from a normal distribution and added element-wise to the patch embeddings, thereby encoding spatial information.

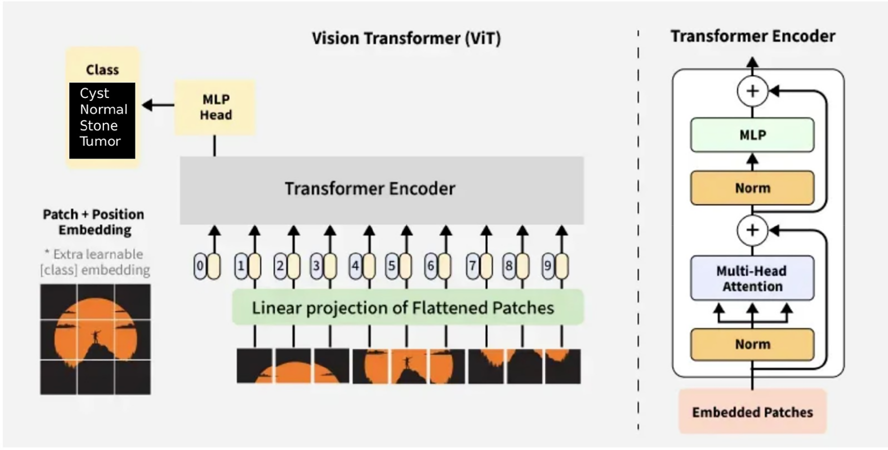

Figure 5. The Vision Transformer (ViT) model architecture. Source: Figure courtesy of GeeksForGeeks (2025)

(5) The core of the model consists of four transformer encoder layers. In each layer, a multi-head self-attention mechanism (with four heads and a key dimension equal to projection dimension) computes attention scores over the patches. The attention output is normalised and added to the original input via a residual connection.

$$\text{Attention}\left(Q, K, V\right) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V$$

Where Q (query), K (key) and V (value) are learned linear projections of the input patch embeddings, the dot product between queries and keys determines the attention score and softmax normalizes it. The weighted sum of values determines the output.

(1) A feed-forward network follows, consisting of a Dense layer with 128 neurons and ReLU activation.

(2) Another Dense layer that projects back to projection dimension is used. The output of the feed-forward network is again added through a skip connection and normalised.

(3) The model applies GlobalAveragePooling1D to aggregate the patch features into a single vector, which is then passed through a Dense layer with softmax activation to produce a probability distribution over the number of classes. The model is compiled using the Adam optimizer, and categorical crossentropy as the loss function.

3.3.5 Hybrid CNN-transformer model. Our system constructs and compiles a deep learning model by integrating a CNN backbone with a transformer head. The model accepts two parameters as input: "shape" which specifies the dimensions of the input data (e.g. height, width and channels), and "number of classes", which indicates the number of output classes for classification. The model comprises the following.

(1) The input layer is an input tensor using the Input layer with the specified input shape.

(2) CNN Backbone model first extracts features using a series of convolutional operations, involving the following:

- A Conv2D layer applies 32 size filters (3, 3) to the inputs with the ReLU activation function.

- A MaxPooling2D layer with a pool size of (2, 2) reduces the spatial dimensions, thus downsampling the feature maps.
- A second Conv2D layer applies 64 filters of size (3, 3) with ReLU activation to further extract features.
- Another MaxPooling2D layer again reduces the spatial dimensions.
- The Flatten layer converts the resulting feature maps into a one-dimensional vector, preparing the data for subsequent processing.

(3) The Transformer Head: the model reshapes the flattened output to prepare it for transformer-style processing:

- The Reshape layer reorganises the data into a sequence format, assuming 64 channels per token.
- A LayerNormalization layer with a small epsilon value (1e$-$6) is applied to stabilise the learning process by normalising the activations.
- The MultiHeadAttention layer performs self-attention over the reshaped data with four attention heads and a key dimension of 64. This mechanism enables the model to capture long-range dependencies and contextual relationships within the data.
- A residual connection returns the attention output to the original input, and another LayerNormalization is applied to the summed output.
- A GlobalAveragePooling1D layer aggregates the sequence of features by computing their average, resulting in a fixed-length vector.

(4) A Dense layer with softmax activation produces the output probabilities for each class, matching the number of classes specified by the number of classes.

(5) The model is compiled with the Adam optimizer, categorical crossentropy as the loss function.

3.4 Performance evaluation

Following prior work on image classification (Refaee *et al.*, 2024), we assess model performance using four standard metrics: accuracy, precision, recall and F1 score.

Accuracy refers to the proportion of correctly classified instances out of the total number of instances and is calculated as:

$$\text{Accuracy} = \frac{N_{\text{correct}}}{N_{\text{total}}} \tag{1}$$

The F1 score is the harmonic mean of precision and recall. A control parameter β can be used to adjust the emphasis between precision and recall. When $\beta = 1$, the F1 score is calculated as:

$$F_1 = 2 \times \frac{P \times R}{P + R} \tag{2}$$

Precision (P) and recall (R) are given by:

$$P = \frac{A}{A + C} \tag{3}$$

$$R = \frac{A}{A + B} \tag{4}$$

In these equations:

(1) A is the number of true positives,

(2) B is the number of false negatives,

(3) C is the number of false positives,

(4) N_{correct} is the number of correct predictions,

(5) N_{total} is the total number of predictions.

3.5 Machine specifications

All experiments in this are run using the Google Colab environment. The system operates on a Linux platform (version 6.1.85+), utilising glibc 2.35 on an ×86 64 architecture. It features an Intel® Xeon® CPU running at 2.00 GHz, configured with one socket comprising a single core that supports two threads per core. The processor accommodates 32-bit and 64-bit operations and has a layered cache architecture, including 32 KiB L1 data and instruction caches, a 1 MiB L2 cache and a 38.5 MiB L3 cache. Despite its extensive instruction set and performance capabilities, the CPU remains susceptible to specific security vulnerabilities, including Meltdown and various Spectre variants, even though some mitigative measures are in place. The system has 12.67 GB of memory, with 5.68 GB currently available. Additionally, it incorporates a Tesla T4 GPU, which operates under driver version 550.54.15 and CUDA version 12.4, maintaining a temperature of 63°C while utilising nearly the entire 15,360 MiB of its memory capacity.

3.6 Results

This section presents the main findings of our experimental investigations. Table 3 summarises the results. We report the performance of five deep learning architectures (discussed in section 3.3) applied to a four-way classification task: cyst, normal, stone, tumour. All models are trained over five epochs, with an average time of an hour per epoch. Figure 6 compares the overall performance of the five models over the four-way classification task: cyst vs. normal vs. stone, tumour.

The conventional CNN achieves an average validation accuracy of 61.33% and an F1 score of 0.6051. Its per-class precision is recorded as 0.4941, 0.8923, 0.4696 and 0.6308, while the corresponding recall values are 0.6765, 0.5609, 0.7552 and 0.5395. In contrast, the ResNet50 model demonstrates a markedly superior performance, attaining an average accuracy of 83.04% and an F1 score of 0.7806, with per-class precision values of 0.9375, 0.8868, 0.8235 and 0.6091, and recall values of 0.9626, 0.8337, 0.4895 and 0.8202.

The EfficientNetB0 model exhibits moderate performance, achieving an average accuracy of 74.01% and an F1 score of 0.6611. Its per-class precision is 0.8575, 0.95, 0.8 and 0.4729, corresponding recall values of 0.9652, 0.6248, 0.2238 and 0.9561. The VisionTransformer (ViT) model records an average accuracy of 63.87% and an F1 score of 0.5724, with per-class precision figures of 0.9424, 0.8538, 0.3452 and 0.3638, and recall values of 0.9626, 0.4294,

Table 3. The results of the five different classification algorithms for the four-way classification of the cyst vs. normal vs. stone vs. tumour classes

ML algorithm	Avg. Precision	Avg. Recall	Avg. F1	Avg. Accuracy
CNN	0.6649	0.6602	0.6423	0.6601
ResNet50	*0.8142*	*0.7765*	*0.7806*	*0.8304*
EffecientNetB0	0.7701	0.6925	0.6611	0.7401
ViT	0.6263	0.6125	0.5724	0.6387
Hybrid	0.7163	0.7599	0.6841	0.6926

Note(s): Italic values denote the best performance in each column

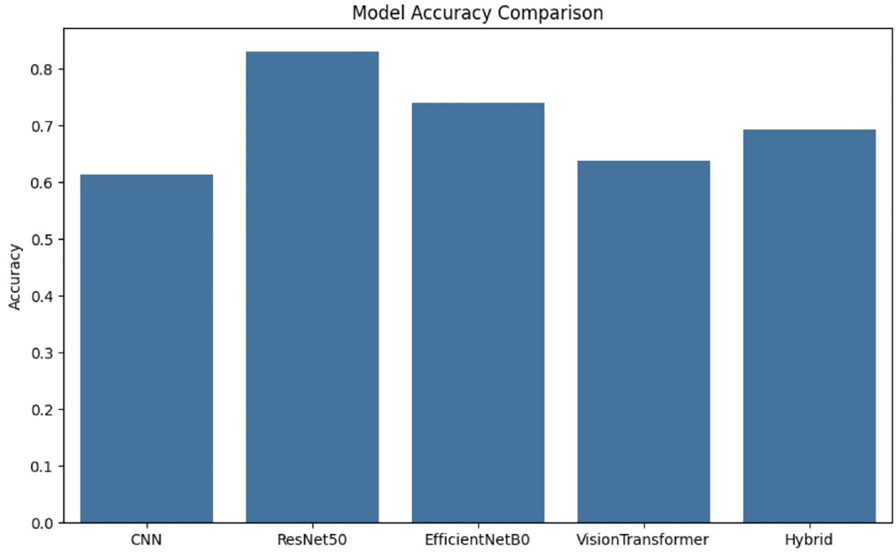

Figure 6. A diagram showing the performance of the five models: CNN, ResNet50, EfficientNetB0, ViT, Hybrid on four way classification: cyst vs. normal, stone, tumour. Source: Authors' own work

0.2028 and 0.8553. Finally, the Hybrid model achieves an average accuracy of 69.26% and an F1 score of 0.6841, with per-class precision values of 0.9953, 0.8934, 0.4751 and 0.5013, and corresponding recall values of 0.5695, 0.6325, 1.0 and 0.8377. The Hybrid model's average precision and recall across classes are 0.7163 and 0.7599, respectively. Figure 7 compares the overall performance across the five models using the ROC curve.

The ROC curve shows that the ResNet50 model performs best, indicating its potential superiority for image classification. Our results align with the findings of Mohamad Tabish *et al*. (2024) wherein the authors experimented with several models to identify which model performed better with image classification. These results highlight the significant impact of model architecture on classification accuracy and underscore the importance of careful model selection in addressing domain-specific challenges. As the best performing model, Figure 8 shows the confusion matrix of ResNet50 on the four-way image classification task. It can be observed that the *stone* class is the most difficult to distinguish, which is aligned with the per-class performance reported in table . One possible explanation is that the *stone* class is the class with the least number of instances in the dataset (see Table 2), which can be further improved if the model gets exposed to more training instances of the *stone* class.

Table 4 shows a closer look at the per-class performance. We can see that CNN seems to be the least performing in general, while the other models have shown variable performances across the four different classes. It is interesting to see that the Hybrid model can achieve a perfect recall rate on the *stone* class despite having the least number of instances, which probably indicates a distinctive ability of the model to fetch the data instances even being exposed to more minor training data. The average precision and recall scores are superior to the ResNet50 on this image classification task. We anticipate further dataset expansion would allow the model to experience more variant features and attain a more robust performance.

Among the evaluated models, ResNet50 demonstrated the highest overall performance, achieving the best average accuracy (83.04%) and F1 score (0.7806). Its residual connections enable deeper network training by alleviating the vanishing gradient problem, making it more effective at capturing complex patterns in medical images. This finding underscores the

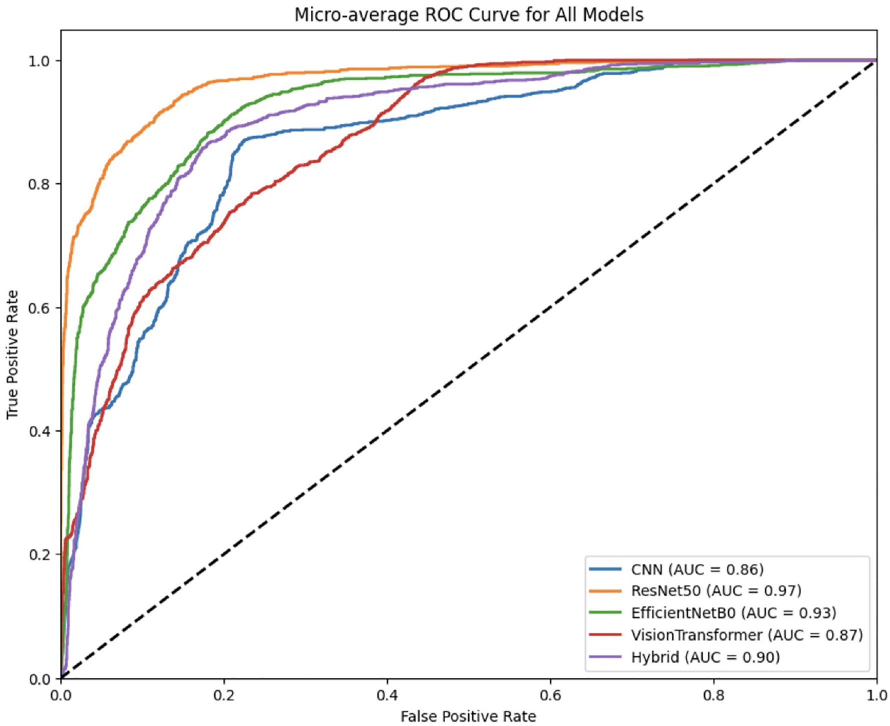

Figure 7. The ROC curve comparing the performance of the five models: CNN ResNet50, EfficientNetB0, ViT and Hybrid. Source: Authors' own work

suitability of ResNet50 for high-stakes diagnostic tasks where subtle visual distinctions are crucial [2].

4. Conclusion and future work

The prevalence of chronic diseases has risen significantly in recent years, imposing substantial economic and healthcare burdens worldwide. The World Health Assembly called for the development of a roadmap to accelerate progress in preventing NCDs. This trend underscores the urgent need for advanced tools to enhance the prediction and, consequently, the early prevention of chronic illnesses. ML present a highly effective approach to analysing medical records and forecasting the likelihood of disease onset by identifying key contributing factors. CKD is one of the most influential NCDs and comes with catastrophic consequences once diagnosed late. With the increasing number of adults diagnosed with CKD worldwide, algorithm-based research has been interested in helping the early discovery of CKD.

This work uses a large dataset of 12.4k images classified into four categories by technicians: cyst, normal, stone or tumour. We utilised five models: CNN, ResNet50, EfficientNetB0, ViT and Hybrid. The models have shown variable performance across the different classes, with the ResNetB0 being the best performing on average, attaining an 83% accuracy score. We believe that this work has advanced research instead of merely focusing on CDK detection, we further classify whether the detected CKD is cyst, stone or tumour. We anticipate that such advancements can significantly impact the effectiveness of how practitioners diagnose NCDs.

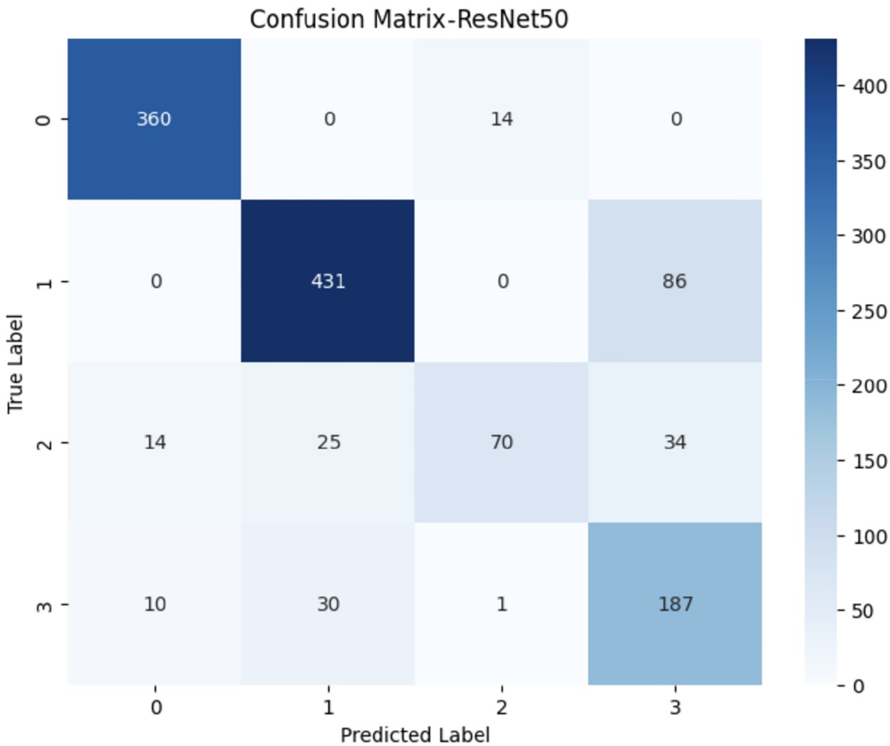

Figure 8. A heatmap showing the confusion matrix of the ResNet50 model. Class mapping (name: code): Cyst: 0, Normal: 1, Stone: 2, Tumour: 3. Source: Authors' own work

Table 4. The four-way classification's per-class precision and recall scores across the five models

	Cyst Precision	Recall	Normal Precision	Recall	Stone Precision	Recall	Tumour Precision	Recall
CNN	0.4941	0.6764	0.8923	0.5609	0.4695	0.7552	*0.6307*	0.5394
ResNet50	0.9375	0.9625	0.8868	*0.8336*	*0.8235*	0.4895	0.6091	0.8201
EffecientNetB0	0.8574	*0.9652*	*0.9500*	0.6247	0.8000	0.2237	0.4728	*0.9561*
ViT	0.9424	0.9625	0.8538	0.4294	0.3452	0.2027	0.3638	0.8552
Hybrid	*0.9953*	0.5695	0.8934	0.6324	0.4750	*1.0000*	0.5013	0.8377

Note(s): Italic values denote the best performance in each column

The future direction of this work can focus on multiple directions, including the utilisation of more extensive and diverse datasets. For instance, the US Chronic Disease Indicators (CDI)-2023 dataset comprises over 1 million patient records with 34 attributes, enabling a more comprehensive analysis encompassing diverse demographics and a wider range of patient cases. This can play a significant role in maximising the trained models' ability to generalise. Models can be trained to utilise features other than CT images, including the patient's age, blood pressure, blood sugar levels, appetite and haemoglobin, to predict the likelihood of a patient developing CDK. Medical centres should also make more datasets publicly available for researchers to provide a robust foundation to refine methodologies and propose further advancements in chronic disease detection.

Another promising direction would be the implementation of explainable AI (XAI) techniques to increase the interpretability of model predictions, allowing clinicians to better understand and trust the model outputs. Additionally, longitudinal studies could be conducted to assess the models' performance on real-time progression tracking of CKD in patients, which may support adaptive monitoring and personalised treatment plans. Future work should also explore integrating hybrid models with electronic health records (EHRs) and real-world clinical workflows to evaluate their practical applicability and user acceptability in hospital environments.

Notes

1. The dataset is publicly available at: https://www.kaggle.com/datasets/nazmul0087/ct-kidney-dataset-normal-cyst-tumor-and-stone

2. The source code can be accessed at: https://colab.research.google.com/drive/1Q50zY6rYM_FIKpb4UvJcBtLQJQZaBMgV?usp=sharing

References

Arifuzzaman, Md., Ahmed, I., Chowdhury, M.J.U., Sakib, S., Rahman, M.S., Hossain, M.E. and Absar, S. (2024), "A novel ensemble-based deep learning model with explainable AI for accurate kidney disease diagnosis", *Advances in Artificial Intelligence and Machine Learning*, Vol. 5 No. 1, pp. 3425-3445.

Bahrami, P., Tanbakuchi, D., Afzalaghaee, M., Ghayour-Mobarhan, M. and Esmaily, H. (2024), "Development of risk models for early detection and prediction of chronic kidney disease in clinical settings", *Scientific Reports*, Vol. 14 No. 1, 32136, doi: 10.1038/s41598-024-83973-5.

Benziger, C.P., Roth, G.A. and Moran, A.E. (2016), "The global burden of disease study and the preventable burden of NCD", *Global heart*, Vol. 11 No. 4, pp. 393-397, doi: 10.1016/j.gheart.2016.10.024.

Delpino, F.M., Costa, Â., Farias, S., Chiavegatto Filho, A., Arcêncio, R. and Nunes, B. (2022), "Machine learning for predicting chronic diseases: a systematic review", *Public Health*, Vol. 205, pp. 14-25, doi: 10.1016/j.puhe.2022.01.007.

GeeksForGeeks (2025), "Vision transformer (ViT) architecture", available at: https://www.geeksforgeeks.org/vision-transformer-vit-architecture/?ref=ml_lbp (accessed 8 February 2025).

Health Statistics (2024), "National Center for Kidney disease", available at: https://www.cdc.gov/nchs/fastats/kidney-disease.htm?utm_source=chatgpt.com (accessed 8 January 2025).

Iftikhar, H., Khan, M., Khan, Z., Khan, F., Alshanbari, H.M. and Ahmad, Z. (2023), "A comparative analysis of machine learning models: a case study in predicting chronic kidney disease", *Sustainability*, Vol. 15 No. 3, p. 2754, doi: 10.3390/su15032754.

Islam, Md N., Hasan, M., Hossain, M.K., Alam, M.G.R., Uddin, M.Z. and Soylu, A. (2022), "Vision transformer and explainable transfer learning models for auto detection of kidney cyst, stone and tumor from CT-radiography", *Scientific Reports*, Vol. 12 No. 1, pp. 1-14, doi: 10.1038/s41598-022-15634-4.

James, M.T., Hemmelgarn, B.R. and Tonelli, M. (2010), "Early recognition and prevention of chronic kidney disease", *The Lancet*, Vol. 375 No. 9722, pp. 1296-1309, doi: 10.1016/s0140-6736(09)62004-3.

Locatelli, F., Vecchio, L.D. and Pozzoni, P. (2002), "The importance of early detection of chronic kidney disease", *Nephrology Dialysis Transplantation*, Vol. 17 No. suppl 11, pp. 2-7, doi: 10.1093/ndt/17.suppl_11.2.

Ma, F., Sun, T., Liu, L. and Jing, H. (2020), "Detection and diagnosis of chronic kidney disease using deep learning-based heterogeneous modified artificial neural network", *Future Generation Computer Systems*, Vol. 111, pp. 17-26, issn: 0167-739X, doi: 10.1016/j.future.2020.04.036, available at: https://www.sciencedirect.com/science/article/pii/S0167739X20308128

Mamatha, B. and Terdal, S.P. (2024), "A review on early detection of chronic kidney disease", *Journal of Scientific Research and Technology*, pp. 35-43, doi: 10.61808/jsrt96.

Mathew, T. and Corso, O. (2009), "Early detection of chronic kidney disease in Australia: which way to go?", *Nephrology*, Vol. 14 No. 4, pp. 367-373, doi: 10.1111/j.1440-1797.2009.01113.x.

Mohamad Tabish, H.A., Arbaz, S.M., Agarwal, M. and Sinha, A. (2024), "Early prediction and risk identification of chronic kidney disease using deep learning technique", *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pp. 1-9, doi: 10.1109/ICCCNT61001.2024.10725032.

Mustafizur Rahman, Md., Al-Amin, Md. and Hossain, J. (2024), "Machine learning models for chronic kidney disease diagnosis and prediction", *Biomedical Signal Processing and Control*, Vol. 87, p. 105, issn: 1746-8094, doi: 10.1016/j.bspc.2023.105368, available at: https://www.sciencedirect.com/science/article/pii/S1746809423008017

Nagawa, K., Hara, Y., Inoue, K., Yamagishi, Y., Koyama, M., Shimizu, H., Matsuura, K., Osawa, I., Inoue, T., Okada, H., Kobayashi, N. and Kozawa, E. (2024), "Three-dimensional convolutional neural network-based classification of chronic kidney disease severity using kidney MRI", *Scientific Reports*, Vol. 14 No. 1, p. 157, doi: 10.1038/s41598-024-66814-3.

Nimmagadda, S.M., Agasthi, S.S., Shai, A., Khandavalli, D.K.R. and Vatti, J.R. (2023), "Kidney failure detection and predictive analytics for CKD using machine learning procedures", *Archives of Computational Methods in Engineering*, Vol. 30 No. 4, pp. 2341-2354, doi: 10.1007/s11831-022-09866-w.

Refaee, E.A., Sheneamer, A. and Assiri, B. (2024), "A deep-learning-based approach to the classification of fire types", *Applied Sciences*, Vol. 14 No. 17, p. 7862, issn: 2076-3417, doi: 10.3390/app14177862, available at: https://www.mdpi.com/2076-3417/14/17/7862

Salekin, A. and Stankovic, J. (2016), "Detection of chronic kidney disease and selecting important predictive attributes", *2016 IEEE International Conference on Healthcare Informatics (ICHI)*, IEEE, pp. 262-270.

Singh, V., Asari, V.K. and Rajasekaran, R. (2022), "A deep neural network for early detection and prediction of chronic kidney disease", *Diagnostics*, Vol. 12 No. 1, p. 116, doi: 10.3390/diagnostics12010116.

WHO (2024), "Noncommunicable diseases", available at: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases (accessed 6 Aug 2024).

Yogesh, N., Shrinivasacharya, P., Naik, N. and Vikranth, B. (2025), "Chronic kidney disease classification through hybrid feature selection and ensemble deep learning", *International Journal of Statistics in Medical Research*, Vol. 14, pp. 109-117, doi: 10.6000/1929-6029.2025.14.11.

Corresponding author
Eshrag Ali Refaee can be contacted at: erefaie@jazanu.edu.sa