Global Scientific JOURNALS

GSJ: Volume 13, Issue 10, October 2025, Online: ISSN 2320-9186

www.globalscientificjournal.com

ENSEMBLE DEEP LEARNING FRAMEWORK FOR AUTOMATED KIDNEY DISEASE DIAGNOSIS USING CT IMAGES

Sule Haruna Sani[1], Njoku Perfect Izuchukwu[1*], Raisul Anwar[2], Nkasi Roland Hilary[3], Matavele Clesia Jamila[4], Dike Joshua Chimdiebube[5], Ifionu Onyinyechukwu Vivian[6], Ahmed Mohamed Sayed[7], Brenda Luannwi Mentan[8], Ofoefuna Chukwuemeka Isaac[9], Onodu Patrick Obinna[10], Ekweozor Ebuka Henry[11]

1 Department of Statistics, Federal University of Agriculture Abeokuta, Nigeria.

1*, 2, 4, 5, 6, & 8 Department of Public Health, Epidemiology and Evidence Base Medicine, I.M. Sechenov First Moscow State Medical University.

3 Department of Public Health, National Open University of Nigeria.

7 General Medicine, I.M. Sechenov First Moscow State Medical University

9 Department of Sciences, Federal Science and Technical College, Uromi Edo state

10 Medicine & Surgery, Enugu state university of Science and Technology

11 Early childhood primary education; Developmental psychology, Nnamdi Azikiwe University.

ABSTRACT

This study proposed an ensemble learning based model for identifying kidney diseases using two transfer learning based neural networks, which are DenseNet121 and InceptionV3. These base learners were trained using a publicly available dataset consisting of 12,446 CT-scan images using a split ratio of 80:10:10. The ensemble learning strategy utilizes a weighted strategy that uses a hyperbolic tangent function for weight allocation. The hyperbolic tangent function allocates a higher weight to the base learners with the best performance. The model performance over the test set reviews that DenseNet121 attains an accuracy score of 99.68%, a recall of 99.28%, and a precision of 99.73%, while InceptionV3 had an accuracy of 99.92%, a recall of 99.95%, and a precision of 99.82%. The ensemble learning performance over the same testing set reviews it attained an accuracy score of 99.92%, a recall of 99.82% and a precision of 99.93%. The overall results show that the ensemble learning attains the highest precision score and has a more robust generalization over the test CT-scan images. The proposed model results were compared to those of recent studies, and it was

observed to outperform the different algorithms utilized in identifying kidney diseases using CT-scan images in terms of higher accuracy, precision, and recall score.

Keyword(s): Kidney Disease, Machine learning, Ensemble learning

1.0 INTRODUCTION

Kidney disease can be described as a medical condition that leads to destabilization of renal function, and one of its common causes includes diabetes, hypertension, infections, and genetic disorders (Ayogu et al. 2025). This disease has been reported to affect 10% of the world population, hitting hard on low-income nations with limited awareness and resources (Ayogu et al. 2025). This disease has different variants, which require a different diagnostic approach to mitigate their effects, and they include cyst, stone, and tumor-like kidney disease (Obaid et al. 2025; Sasikaladevi et al. 2024). Cysts, also known as Polycystic kidney disease, are the accumulation of fluid-like substances in the kidney, while Stones or Nephrolithiasis are mineral substances that crystallize within the renal system (Suijker et al. 2025; Borah et al. 2022). Tumor-like Renal cell carcinoma affects the renal tubules, presenting significant medical challenges (Boni et al. 2019). Treatment for these variants varies from decompression surgery, lithotripsy (shock wave therapy), to radical nephrectomy (surgical removal of the kidney) in cases that deal with tumors (Fahed et al. 2024; Islam et al. 2022; Dalia et al. 2022). However, challenges in disease management are mainly related to optimizing the early detection strategy, the cost of implementation, and mitigating the side effects of treatment. In terms of early detection approaches on which this study is based, the field of Deep learning, an extension of Artificial intelligence, has shown promising performance. Various studies have shown the proficiency of Convolutional Neural Network (CNN) and Transfer learning Models pretrained using Big data in predicting kidney-related diseases using X-rays or CT-scans with high precision. This study aims to use a more innovative approach by proposing an ensemble learning algorithm that uses a weighted average strategy based on the prediction probabilities of base learners, which is a transfer learning algorithm. This study fine-tuned this transfer learning algorithm using CT scans of patients with different kidney diagnosis classes to develop an ensemble learning based framework that detects this disease type with high accuracy and precision.

The study structure contains a review of recent works that integrate various deep learning algorithms in detecting kidney-related diseases in Section 2.0, Section 3.0 focuses on the methodology and implementation strategy used in this study, and Section 4.0 highlights the study result conclusion and recommendations for future works.

2.0 LITERATURE REVIEW

Fahed et al. (2024) fine-tuned a VGG16 for the identification of kidney stones. The study used X-ray kidney images sourced from a hospital in Pakistan. The fine-tuned VGG16 was trained using 9986 X-ray kidney images, and it was evaluated using 4279. The study results reveal that the model attained an accuracy score of 97.41% and a precision of 97.39%. Isil et al. (2021) investigated the efficiency of various Machine learning algorithms in predicting kidney stones using an Image dataset. The selected ML algorithms considered include Decision Trees, Random Forest, Support Vector Machine, K-Nearest Neighbor (kNN), Naive Bayes, Multilayer Perceptron, and Convolutional Neural Network. The study trained the algorithms using a K-fold size of 5 and a variety of sampling strategies to address class imbalance, ranging from SMOTE, under-sampling, and SMOTETOMEK. The study results reveal that the Decision tree was the best model for detecting Kidney stone-related illness with an F1-score of 85.3%. Qadir and Dana (2023) proposed a hybrid model that combines a DenseNet architecture with a Random Forest classifier for detecting kidney diseases using CT scans. The study results indicate that the model attained an accuracy score of 99.68% on the training set and 99.44% on the test dataset. Islam et al. (2022) utilized Vision transformers and explainable transfer learning algorithms, which include VGG16, ResNet50, EANet, CCT, Swin Transformers, and InceptionV3. The study results review that Swin Transformers attained the highest prediction accuracy (99.30%) on the test dataset, VGG16 had an accuracy score of 98.20%, CCT 96.54%, EANet 77.02%, RestNet50 73.80%, and InceptionV3 61.60%. Dalia et al. (2022) evaluated the efficiency of different deep learning models in detecting Kidney tumors. The study utilized VGG16, ResNet50, and 2D CNN; these models were trained using the CT-scans dataset. The study results showed that 2D CNN was most efficient with an accuracy score of 97%, with ResNet50 96% and VGG16 60%. Bhandari et al. (2023) explored the capacity of different lightweight CNN models in detecting Renal abnormalities. The study results indicate that the lightweight CNN models attained a maximum classification accuracy of 99.47%. Saleh et al. (2024) developed a KidneyNet model based on the gradient-weighted class activation mapping (Grad-CAM) algorithm, which enables the pinpointing of affected areas in patient CT scans. This model was implemented and compared with pretrained CNN algorithms like EfficientNetB1 & B2, Xception, and VGG19. The study results indicated that KidneyNet attained the highest classification accuracy of 99.88%. Dinesh et al. (2025) developed a Novel CNN model for identifying kidney stones and cancer growth using CT-scan images. The study results show that the proposed CNN model by the researcher outperforms other algorithms, such as EANet and ResNet50, with an accuracy score of 98.66%. The other pretrained models had an accuracy score of 83.65% and 87.92%. Megha Patel & Rajesh Patel (2025) developed a CNN algorithm for detecting different types of kidney-related diseases, ranging from kidney stones to tumors. The study results reveal that the model attained an accuracy score of 88%.

3.0 METHODOLOGY

3.1 DATA DESCRIPTION

This study chooses a public dataset that cuts across four diagnostic classes, which are tumor, cyst, normal, or stone, with relation to kidney disease. These records were collated from different hospitals in Dhaka and Bangladesh, and it was prepared using a batch of Digital Imaging and Communications in Medicine (DICOM) standardized records. These records contain both Coronal and Axial cuts from both contrast and non-contrast studies to create DICOM images of the ROI for each radiological finding. Figure 1.0 shows the CT scan images of the different CT scan diagnoses, ranging from Cyst, Normal, Stone, and Tumor. Table 1.0 shows the distribution of these CT-scan image diagnoses below.

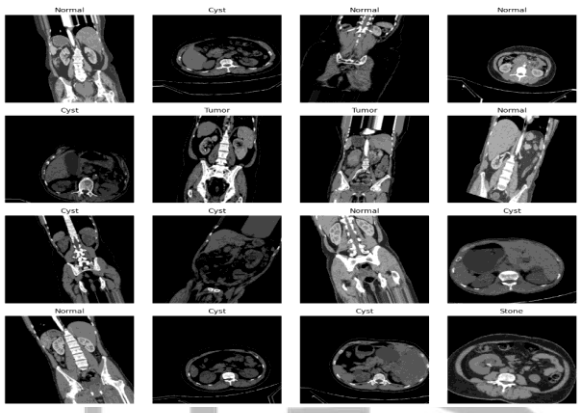

Figure 1.0: CT scans of kidney diagnosis

Table 1.0: Data distribution across train, validation and test dataset

	Train	Validation	Test	Total
Cyst	2596	556	557	3709
Normal	3553	762	762	5077
Stone	963	207	207	1377
Tumor	1598	342	343	2283

3.2 DATA PREPROCESSING

The data preprocessing techniques employed in this study include data ingestion, data splitting, image scaling, and data augmentation. The CT scans were uploaded and separated into train, validation, and test image datasets using a split ratio of 80:10:10. For the train CT scans, the images

were scaled using a scale factor of 1/255, and data augmentation was applied to boost the CNN's predictive efficiency. The augmentation implemented included image rotation, horizontal shift, shearing, flipping, and filling empty pixels. This augmentation was implemented using *the Tensorflow ImageGenerator* class. No data augmentation was applied to the validation and test datasets. This dataset was converted to a Tensor generator for the train, validation, and test datasets. This method includes stating the image directory, batch sizes, image size, and class mode. The batch size used for this study was 32, the image size was 224 by 224, and the class mode selected was categorical to align with our multiclassification task objective. The clean preprocessed dataset was fitted to a selected transfer learning model and fine-tuned to fit our study objectives.

3.3 DEEP LEARNING ALGORITHMS

3.3.1 DENSENET121

DenseNet121 architecture uses a feed-forward process that connects each layer in the network by sending feature maps from preceding layers (Sankari, 2025). This variant of the Convolution neural network supports the recycling of features, supporting gradient flow into the network (Heru et al. 2024; Sankari, 2025). This mitigates issues of vanishing gradients. The key features of this model are feature maps that use concatenation and its ability to learn with fewer parameters (Sankari, 2025; Dheeraj et al., 2025).

3.3.2 INCEPTIONV3

InceptionV3 architecture is a Convolution Neural Network variant that uses batch-normalization, alongside varying filters to capture multi-scale information (Poonam Shourie et al., 2023). The transfer learning algorithm's strength lies in its ability to extract hierarchical features efficiently (Tanishq Soni et al., 2024). Also, the concatenated layers and global average pooling give it a robust generalization over handling different classification and feature extraction tasks efficiently.

3.3.3 PROPOSED ENSEMBLE ALGORITHM

The proposed ensemble algorithm utilized in this study is based on a weighted average strategy that integrates the probability prediction scores of base learners for weight allocation. In this study, we implemented Densenet121 and InceptionV3 as our base learners. The base learner's performance metrics were evaluated in terms of recall, precision, f1-score, and Area Under Curve (AUC). These performance metrics are stored in an array-like structure $A^{(i)}$ and passed through a hyperbolic tangent function stated below for weight allocation.

$$w^{(i)} = \sum_{x \epsilon A^{(i)}} \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad 1.0$$

Where x represents the performance metrics of the base learners within the range [0,1]. The tangent function allocates a higher weight to the base learners with the highest performance metrics. These computed weights are then used to estimate the ensemble probability by multiplying the computed weight $w^{(i)}$ by the probability scores of the base learners. Equation 2.0 describes this prediction process mathematically.

$$prediction_j = argmax\left(\frac{\sum_i w^{(i)} p_j^i}{\sum_i w^{(i)}}\right) \quad 2.0$$

Where p_j^i is the probability scores of the base leaners.

3.4 MODEL PERFORMANCE METRICS

The performance metrics we used for evaluating the proficiency of our proposed models are accuracy, recall, precision, F1-score, and Area Under Curve (AUC). Accuracy metrics evaluate the overall correctness of the model in predicting the image classes. Precision measures the correctness proportion of elements belonging to a particular class. Recall measures the proportion of instances that belong to a particular class captured by the algorithm. F1-score measures the balance between the precision and recall for a particular class. AUC measures the algorithm's capacity to identify the patterns between different classes effectively.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad 3.0$$

$$Precision = \frac{TP}{TP + FP} \quad 4.0$$

$$Recall = \frac{TP}{TP + FN} \quad 5.0$$

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall} \quad 6.0$$

4.0 RESULTS FINDINGS

The sections focus on the results metrics of the pretrained base learners and the ensemble learning models. The base learners were trained using a batch size of 32, an input image size of 224 by 224, a ReLu activation function, and a SoftMax output function using an epoch size of 30. Before training the model, we initiated a checkpoint for training our base learners by saving the model weights that attained the highest accuracy and lowest loss. The base learners and the ensemble learning model performance over the test CT-scans are displayed below:

Table 2.0: DenseNet121 model performance

Classes	**Precision**	**Recall**	**F1-score**
Cyst	98.93%	100%	99.46%
Normal	100%	100%	100%
Stone	100%	97.10%	98.53%
Tumor	100%	100%	100%

Table 3.0: InceptionV3 model performance

Classes	**Precision**	**Recall**	**F1-score**
Cyst	100%	100%	100%
Normal	100%	99.80%	99.90 %
Stone	99.28%	100%	99.64%
Tumor	100%	100%	100%

Table 4.0: Ensemble model performance

Classes	**Precision**	**Recall**	**F1-score**
Cyst	99.73%	100%	99.87%
Normal	100%	100%	100%
Stone	100%	99.28%	99.64%
Tumor	100%	100%	100%

Table 5.0: Model summary report

Model(s)	**Accuracy**	**Precision**	**Recall**	**F1-score**	**AUC**
DenseNet121	99.68%	99.73%	99.28%	99.50%	100%
InceptionV3	99.92%	99.82%	99.95%	99.89%	100%
Ensemble Model	**99.92%**	**99.93%**	**99.82%**	**99.88%**	**100%**

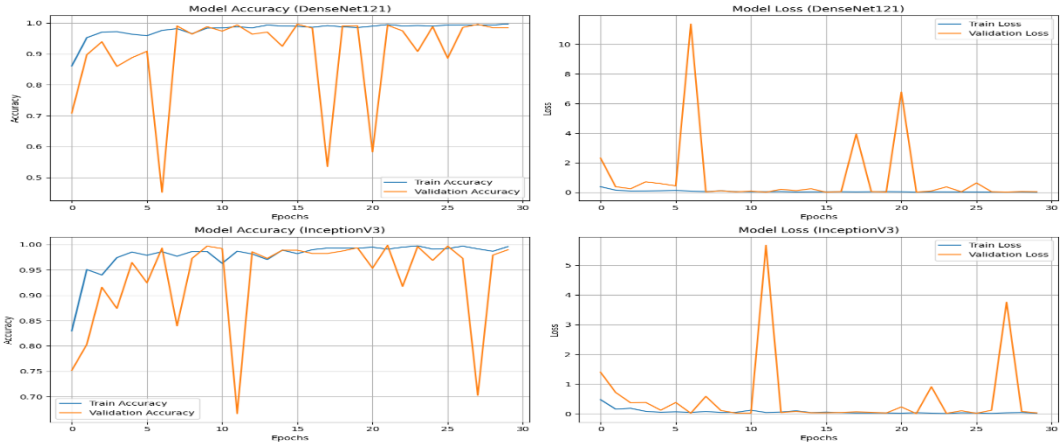

Figure 2.0: Model accuracy and loss

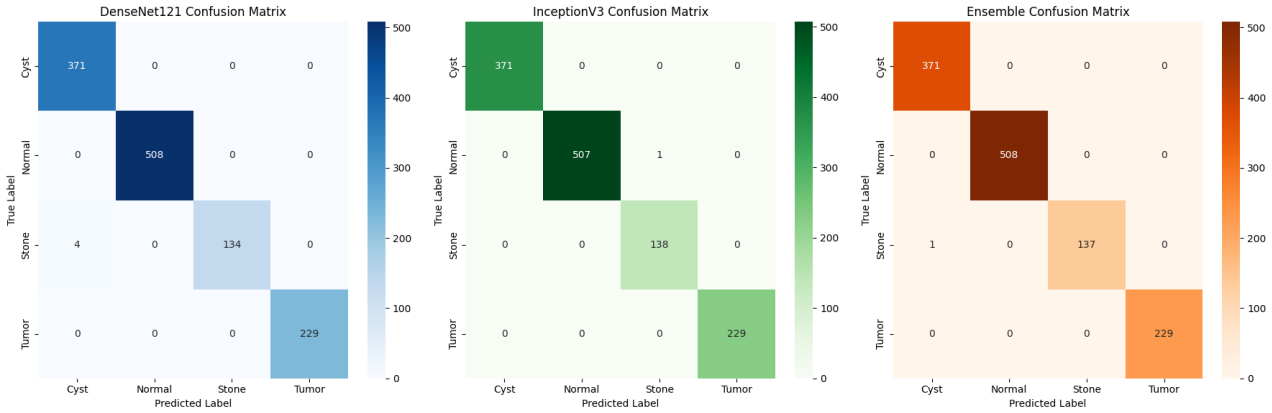

Figure 3.0: Model Confusion Matrix

Tables 2.0 and 3.0 show the base learners' evaluation results across the four distinct diagnosis classes. The results review that the DenseNet121 model was capable of a minimum of 97.10% of the kidney diagnosis (Tumor) and a maximum of 100% for (Cyst, Normal, and Stone). The CT-scans classes were captured with a corresponding precision of 98.93% for Cyst and 100% for Normal, Stone, and Tumor, respectively. DenseNet121 attained a maximum f1-score of 100% (Normal) and a minimum of 98.53% (Stone). InceptionV3 was able to capture a minimum recall score of 99.80% of the image classes (Normal), with 100% of (Cyst, Stone, and Tumor), respectively. These image classes were predicted with a minimum precision of 99.28% for Stone and 100% for Cyst, Normal, and Tumor, respectively. InceptionV3 had a minimum f1-score value of 99.64% (Stone) and a maximum of 100% (Cyst and Tumor), respectively. These results indicate that our base learner is well able to attain a good generalization of the test CT scans with high precision. Table 4.0 shows the performance metrics of the proposed ensemble learning model. The proposed model attained a maximum recall score of 100% for the Cyst, Normal, and Tumor classes, and a minimum recall score for Stone of 99.28%. The model attained a higher precision when compared to the base learners, with a minimum value of 99.73% for the Cyst class and 100% for Normal, Stone, and Tumor, respectively. It attained a

minimum f1-score of 99.64% (Stone) and a maximum of 100% (Normal and Tumor). Table 5.0 shows the average values for the corresponding models and the accuracy attained over the test set. The results show that the ensemble learning model and InceptionV3 outperform DenseNet121 in terms of accuracy (99.92%) and other metrics. Also, the ensemble learning model outperforms the InceptionV3 by attaining a higher precision value of 99.93%. The three models attained a 100% in terms of AUC, indicating these models are capable of distinguishing effectively between the different diagnosis classes using the CT-scans images.

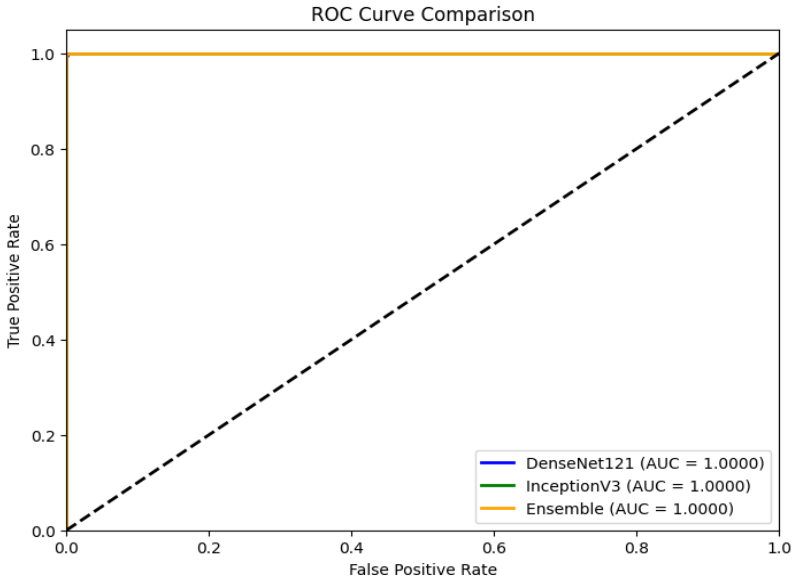

Figure 4.0: ROC-Curve Comparison

Table 6.0: Summary report from recent studies that used Transfer learning for Kidney detection

Author(s)	**ML algorithms**	**Accuracy**
K Basava Raju et al. (2025)	ResNet18, MobileNetV2, and EfficientNet	ResNet18: 99.60%, MobileNetV2: 99.62%, EfficientNet: 98.67%
Qadir and Abd (2023)	Densenet-201 model and RF	99.44%
Bingol et al. (2023)	CNN	99.37%
Bhandari et al. (2023)	CNN	99.47%
Saleh et al. (2024)	CNN	99.88%
Proposed Ensemble Algorithm	**Ensemble Model (DenseNet121+InceptionV3)**	**Accuracy: 99.92%**

5.0 CONCLUSION

This study demonstrates the potential of an ensemble learning based model in mitigating the early detection of kidney-based disease with relation to Cyst, Stone, and tumor-related infections. The study results using a diverse kidney disease dataset trained to transfer learning base learners, which are InceptionV3 and DenseNet121, using a split ratio of 80:10:10. The base learners attained a classification accuracy of 99.68% (DenseNet121) and 99.92% (InceptionV3), with an average recall score of 99.28% and 99.95%. This shows the model's capacity to capture significant instances of kidney diagnostic classes efficiently with high precision scores of 99.73% and 99.82%. These base learners' performance metrics were then used to develop our proposed weighted average strategy that was deployed using the hyperbolic tangent function for weight allocation to identify the model with the highest performance. These in turn were used to create the ensemble learning based prediction that was evaluated, and it attained a classification accuracy of 99.92%, matching the InceptionV3 model, however, with a higher precision score of 99.93%. The models' AUC for both the ensemble learning models and the base learners shows the model's ability to recognize different diagnostic classes efficiently. The ensemble learning model was compared with previous algorithms used by researchers in predicting kidney disease diagnosis, developed using a similar dataset. The results review our ensemble learning attained the highest accuracy when compared with that of K Basava Raju et al. (2025); Qadir and Abd (2023); Bingol et al. (2023); Bhandari et al. (2023); and Saleh et al. (2024). This further shows the potential of ensemble learning based strategies in optimizing healthcare practices for dealing with the detection of kidney diseases. We recommend future studies to extend the use of our proposed model to detect different variants of kidney diseases, and also employ the use of Big Data in medical analysis for the development of sophisticated algorithms using a hybrid-ensemble framework to enhance clinical applicability and diagnostic reliability.

6.0 REFERENCES

Dheeraj, A., & Chand, S. (2024). LWDN: lightweight DenseNet model for plant disease diagnosis. Journal of Plant Diseases and Protection, 131(3), 1043–1059. https://doi.org/10.1007/s41348-024-00915-z

K Basava Raju, B Venu Gopal, Shishir, S Sai Shiva (2025). Fine-Tuning CNN Models for Accurate Kidney Condition Classification from CT scans. Retrieved from: https://restpublisher.com/wp-content/uploads/2025/04/Fine-Tuning-CNN-Models-for-Accurate-Kidney-Condition-Classification-from-CT-Scans.pdf

Bingol, H.; Yildirim, M.; Yildirim, K.; Alatas, B. Automatic classification of kidney CT images with relief based novel hybrid deep model. Peer JComput. Sci. 2023, 9, e1717

Qadir, A.M.; Abd, D.F. Kidney diseases classification using hybrid transfer-learning densenet201-based and random forest classifier. Kurd. J. Appl. Res. 2023, 7, 131–144

Bhandari, M.; Yogarajah, P.; Kavitha, M.S.; 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.

Almuayqil, S.N.; Abd El-Ghany, S.; Abd El-Aziz, A.A.; Elmogy, M. KidneyNet: A Novel CNN-Based Technique for the Automated Diagnosis of Chronic Kidney Diseases from CT scans. Electronics 2024, 13,4981. https://doi.org/10.3390/electronics13244981

Fahad Ahmed, Sagheer Abbas, Atifa Athar, Tariq Shahzad, Wasim Ahmad Khan, Meshal Alharbi, Muhammad Adnan Khan & Arfan Ahmed (2024). Identification of kidney stones in KUB X-ray images using VGG16 empowered with explainable artificial intelligence. Retrieved from: Identification of kidney stones in KUB X-ray images using VGG16 empowered with explainable artificial intelligence

Işıl Karabey Aksakallı, Sibel Kaçdıoğlu and Yusuf Sinan Hanay (2021). Kidney X-ray Images Classification using Machine Learning and Deep Learning Methods. Retrieved from: dergipark.org.tr/en/download/article-file/1569382

M. Islam, M. Hasan, M. Hossain, M. Alam, M. Uddin, A. Soylu, "Vision transformer and explainable transfer learning models for auto detection of kidney cyst, stone and tumor from CT-radiography," vol. 12, no. 1, pp. 1-14, 2022. Retrieved from: https://www.nature.com/articles/s41598-022-15634-4.pdf

D. Alzu'bi, M. Abdullah, I. Hmeidi, R. AlAzab, M. Gharaibeh, M. El-Heis, K. Almotairi, A. Forestiero, A. Hussein, L. Abualigah, "Kidney Tumor Detection and Classification Based on Deep Learning Approaches: A New Dataset in CT Scans," vol. 2022, 2022. Retrieved from: (PDF) Kidney Tumor Detection and Classification Based on Deep Learning Approaches: A New Dataset in CT Scans

Ayogu, I. I., Daniel, C. F., Ayogu, B. A., Odii, J. N., Okpalla, C. L., & Nwokorie, E. C. (2025). Investigation of ensembles of deep learning models for improved chronic kidney diseases detection in CT scan images. *Franklin Open*, *11*, 100298. https://doi.org/10.1016/j.fraope.2025.100298

Boni, A., Cochetti, G., Sidoni, A., Bellezza, G., Lepri, E., Giglio, A. D., Turco, M., Vermandois, J. A. R. D., Zingaro, M. D., Cirocchi, R., & Mearini, E. (2019). Primary angiosarcoma of the kidney: case report and comprehensive literature review. *Open Medicine*, *14*(1), 443–455. https://doi.org/10.1515/med-2019-0048

Borah, S., & Pandit, A. V. (2022). Prediction Methodologies to Detect Kidney Stones using Deep Learning. *Research Journal of Computer Systems and Engineering*, *3*(2), 46–53. https://doi.org/10.52710/rjcse.55

Sasikaladevi, N., & Revathi, A. (2024). Digital twin of renal system with CT-radiography for the early diagnosis of chronic kidney diseases. *Biomedical Signal Processing and Control*, *88*, 105632. https://doi.org/10.1016/j.bspc.2023.105632

Suijker, C. A., van Mazijk, C., & Roemeling, S. (2025). Kidney stone disease: phenomenological perspectives. *Medicine, Health Care and Philosophy*. https://doi.org/10.1007/s11019-025-10301-7

Obaid, W., Hussain, A., Rabie, T., Abd, D. H., & Mansoor, W. (2025). Multi-model deep learning approach for the classification of kidney diseases using medical images. *Informatics in Medicine Unlocked*, *57*, 101663. https://doi.org/10.1016/j.imu.2025.101663

P. Dinesh, Dr. D. Deepa, S. Chanakya (2025). Novel Deep Learning Method for Automated Diagnosis of Kidney Disease from Medical Image using CNN. Retrieved from: https://www.ijsat.org/papers/2025/2/3114.pdf

Ms. Megha Patel1, Dr. Rajesh Patel (2025). Image Based Chronic Renal Disease Diagnosis Using Convolution Neural Network Deep Learning Approach.

Sankari, C. (2025). Optimized Deep Learning Framework Utilizing DenseNet121 for High-Accuracy Image Classification with Improved Computational Efficiency and Feature Learning. *2025 International Conference on Data Science, Agents &Amp; Artificial Intelligence (ICDSAAI)*, 1–6. https://doi.org/10.1109/icdsaai65575.2025.11011711

Heru Agus Santoso, Brylian Fandhi Safsalta, Nanang Febrianto, Galuh Wilujeng Saraswati, Su-Cheng Haw (2024). Comparative analysis of convolutional neural network and DenseNet121 transfer learning in agriculture focusing on crop leaf disease identification.