Deep Learning Architectures for Urolithiasis Classification: A Comparative Analysis of DNN, MLP, and Autoencoder-based Models

Ibtasam Ur Rehman[1], Abdulraqeb Alhammadi[2], and Jibran K. Yousafzai[3]

[1] [1] Ho Chi Minh City University of Technology (HCMUT), Ho Chi Minh City, Vietnam
ribtrasam.sdh231@hcmut.edu.vn
[2] [2] Faculty of Artificial Intelligence, Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia
abdulraqeb.alhammadi@utm.my
[3] [3] College of Engineering and Applied Sciences, American University of Kuwait, Kuwait
jyousafzai@auk.edu.kw

Abstract. Renal calculus is a common urological condition that necessitates an accurate and effective diagnostic methodology. This study examines three deep learning architectures applied to the classification of kidney stone images: the Deep Neural Network (DNN), the Multi-Layer Perceptron (MLP), and the Autoencoder-based Deep Neural Network (AE-DNN). A rigorous training and assessment system utilizing a publicly available dataset of 9,416 photos is presented. This study offers a comprehensive comparison of performance and clinical measures, emphasizing the trade-off between model complexity and efficiency. The results indicate that all models achieve outstanding performance, with the MLP model being the most efficient and precise, attaining a validation accuracy of 99.67%, precision of 99.67%, and specificity of 99.44%. AE-DNN exhibits a competitive performance with an accuracy of 99.47%. A key finding is the enhanced computational efficiency of the MLP, which trains significantly faster than alternative architectures. Supplementing the promising results, this work acknowledges limitations stemming from the binary classification task and image down-sampling, as well as the necessity for external validation to establish clinical usefulness.

Keywords: Kidney Stone Classification · Deep Learning · Medical Imaging · Neural Networks · Computer-Aided Diagnosis

1 Introduction

1.1 Background and Motivation

Kidney stone disease is a significant global health issue, with incidence rates rising and impacting approximately 10-15% of the worldwide population [1,2].

Traditional diagnostic techniques, such as ultrasound scans, standardized clinical biomarker analysis, and manual interpretation by radiologists, can be time-consuming and prone to interobserver variability. Deep learning-based automated systems offer the opportunity to enhance diagnostic accuracy, thereby reducing interpretation time and increasing access to healthcare. Recent research in medical imaging and computational techniques has shown great promise for enhancing kidney stone categorization and detection, potentially leading to the development of effective and precise diagnostic instruments.

1.2 Challenges and Research Contribution

Classical kidney stone diagnosis presents challenges, including stone variability, limited access to imaging, and subjective interpretation. In contrast, computational approaches have been explored, but comprehensive comparisons of deep learning architectures for kidney stone classification remain limited. This research evaluates the three neural network architectures (DNN, MLP, and AE-DNN) and provides insight into their performance and efficiency for kidney stone detection.

1.3 Related Work in Kidney Stone Classification

Recent advancements have shown the effectiveness of machine learning and deep learning approaches for kidney stone classification across multiple data modalities and imaging techniques. Several studies have focused on traditional machine learning approaches using custom features. Asaye et al. [3] utilized a machine learning approach to detect kidney stones from ultrasound images using 410 labeled samples, employing Gabor filtering and threshold-based segmentation with GLCM texture features. Their approach obtained 98.4% accuracy with the KNN algorithm, illustrating the potential of traditional feature-based methods. Iparraguirre-Villanueva et al. [4] evaluated multiple ML algorithms using non-imaging clinical parameters, with Logistic Regression achieving 78% accuracy. Furthermore, Panda et al. [5] reported 93% accuracy using urine analysis data, which included physical features such as pH and calcium concentration. Khan et al. [6] developed an automated detection system employing median filtering for noise reduction and adaptive thresholding for segmentation, which achieved 96.82% accuracy and 92.16% sensitivity. Their method demonstrated computational efficiency while overcoming challenges like speckle noise and low contrast in ultrasound images. Deep learning approaches have also shown promise in kidney stone classification. Studies like Kumar et al. [7] proposed a hybrid CNN-ResNet model that achieved 90.9% accuracy by utilizing CNN feature extraction capabilities and ResNet deep learning advantages. Further advancing deep learning applications Yubao Liu et al. [8] developed integrated radiomics and deep learning model for classifying kidney stone types on CT urography achieving exceptional performance withAUC of 0.95. Jacob et al. [9] compared three deep learning models for kidney stone identification from CT scans finding that ResNet50 outperformed both VGG16 and standard CNN by 7.7% and 4.5%

respectively. Despite these advancements there is need that remains for comparative studies of different neural network architectures specifically optimized for kidney stone classification using clinical biomarkers and imaging data. Our research addresses this gap by providing evaluation of DNN, MLP and AE-DNN architectures on dataset of kidney images offering information into their relative strengths, limitations and clinical applicability for kidney stone detection. [10]

2 Methodology

This study demonstrate contrastive analysis of deep learning approaches for binary classification of kidney stone images and methodology consist of data preprocessing, model development and comprehensive evaluation strategy. Three distinct neural network architectures are implemented and evaluated Deep Neural Network, Multi-Layer Perceptron and Autoencoder-based DNN (AE-DNN). The experimental workflow, illustrating the data pipeline and model training process is shown in Figure 1. To make sure reproducibility all random seeds were fixed to 123 at the beginning of the execution.

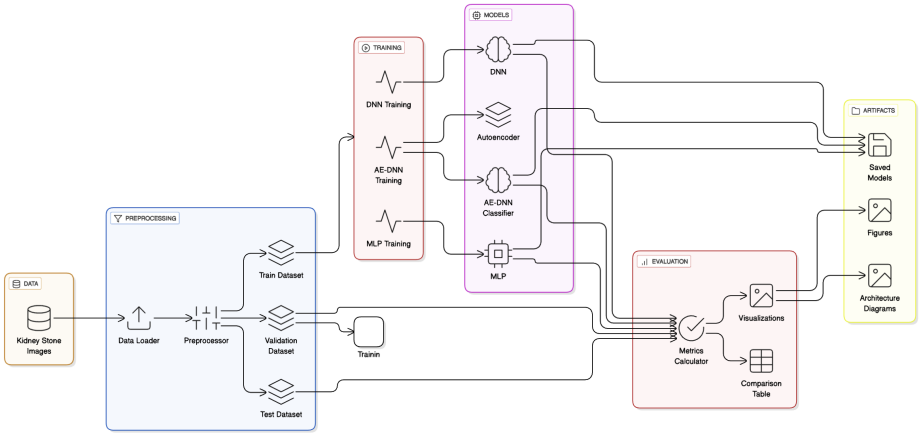

Fig. 1. Methodological Framework for Kidney Stone Classification using Deep Learning Approaches

2.1 Dataset Description and Preprocessing

Data Collection and Characteristics The study utilized publicly available Kidney Stone Classification dataset from Kaggle [11] consisting of 9,416 high quality labeled images categorized into two distinct classes **Normal** (4,708 images) and **Stone** (4,708 images) and dataset exhibit precisely balanced class distribution to prevent class imbalance bias during model training and to make

sure robust performance evaluation furthuremore images are systematically collected from multiple clinical sources under consistent acquisition conditions. All images are standardized to resolution of 64×64 pixels with RGB color channels resulting in input dimension of $12,288$ features per sample ($64 \times 64 \times 3$). The 64×64 resolution was selected after analysis confirmed that the key stone characteristics remained discernible at this scale while ensuring computational efficiency for architectural comparisons.

$$\begin{aligned} \text{Training Set} \quad & 6,026 \text{ images} \approx 64\% \\ \text{Validation Set} \quad & 1,507 \text{ images} \approx 16\% \\ \text{Test Set} \quad & 1,883 \text{ images} \approx 20\% \end{aligned}$$

$$\text{Total} = 6,026 + 1,507 + 1,883 = \mathbf{9,416}$$

Data Cleaning, Validation and Preprocessing A comprehensive data validation confirmed no missing images or corrupted files with in the dataset and preprocessing pipeline include pixel normalization and scalingintensity values to range [0, 1] using the transformation:

$$I_{\text{normalized}} = \frac{I_{\text{original}}}{255} \tag{1}$$

This step make sure that the stable gradient computation during training and for compatibility with the fully connected architectures DNN, MLP, AE-DNN the normalized images were flattened into one dimensional vectors:

$$X_{\text{flat}} = \text{flatten}(I_{\text{normalized}}) \in \mathbb{R}^{12288} \tag{2}$$

Data Augmentation and Training Strategy Given the size and inherent balance of the dataset, explicit data augmentation techniques were not employed and this decision was made to allow the models to learn from authentic clinical representations without synthetic alterations however the training strategy utilize TensorFlow data pipeline for efficiency, utilizing automatic caching and prefetching. The dataset splits were created with fixed random seed (123) to ensure consistency across experiments and the validation set was used for model selection and hyperparameter tuning while the final held out test set provided unbiased estimate of model performance.

2.2 Deep Learning Architectures

Three neural network architectures were designed with varying complexities to facilitate robust comparison and all models were implemented using TensorFlow/Keras and shared common experimental setup: an input dimension of

12,288 (flattened 64×64×3 images), ReLU activation functions for hidden layers and a softmax output layer for the 2 class classification, Adam optimizer, fixed training schedule of 10 epochs, batch size of 32 and random seed of 123 for reproducibility.

Deep Neural Network (DNN) Architecture The DNN was designed as high capacity model with two hidden layers comprising 512 and 256 neurons and each layer employed ReLU activation which was followed by Batch Normalization for stable training and utilized Dropout regularization ($p = 0.3$) to mitigate overfitting. L2 weight regularization ($\lambda = 10^{-4}$) was also applied to all layers and the model was compiled with Adam optimizer and sparse categorical cross-entropy loss, $\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda\|\mathbf{W}\|_2^2$ resulting in total of approximately 6.5 million trainable parameters and this architecture aimed to use its depth and breadth to learn complex, hierarchical features directly from the pixel data.

Multi-Layer Perceptron (MLP) Architecture The MLP was designed as more efficient and compact architecture because it featured two hidden layers with 128 and 64 neurons and it maintained use of ReLU, Batch Normalization and L2 regularization ($\lambda = 10^{-4}$) but utilized lower Dropout rate ($p = 0.2$) reflective of it reduced capacity and with approximately 1.6 million parameters. This model aimed to achieve favorable balance between performance and computational cost, testing whether simpler network could match or exceed the performance.

Autoencoder-based DNN (AE-DNN) Architecture The AE-DNN employed hybrid two stage training process. Firstly, symmetric autoencoder was trained in unsupervised manner to reconstruct its input, learning compressed latent representation and the encoder consisted of layers with 512, 256 and 128 neurons (ReLU activation) with the 128-neuron layer serving as the latent space and the decoder mirrored this structure. The autoencoder was trained using Mean Squared Error (MSE) loss: $\mathcal{L}_{\text{recon}} = \text{MSE}(\mathbf{x}, \hat{\mathbf{x}})$. Subsequently a classifier,which shared the encoder weights was attached to latent space and trained using the labeled data with sparse categorical cross-entropy loss, $\mathcal{L}_{\text{class}} = \mathcal{L}_{\text{CE}}$. Total parameter count for full AE-DNN system was approximately 7.2 million and this approach investigated whether feature learned through unsupervised reconstruction could enhance the performance of downstream classifier.

2.3 Experimental Setup

Training Configuration and Hyperparameters All models are trained for fixed number of 10 epochs using batch size of 32 and Adam optimizer was used with it default parameters (learning rate $\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$) however the primary loss function for classification tasks (DNN, MLP, and the

second stage of AE-DNN) was sparse categorical cross-entropy:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{2} y_{i,c}\log(\hat{y}_{i,c}) \tag{3}$$

For autoencoder pretraining stage of AE-DNN model and Mean Squared Error loss was used:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_2^2 \tag{4}$$

Evaluation Metrics and Validation Strategy A comprehensive set of metrics was used to assess model performance from multiple perspectives. The primary metrics included accuracy, precision, recall and F1 score calculated as follows:

1. **Performance Metrics:**
   - **Accuracy:** The proportion of correctly classified samples
   $$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$
   - **Area Under the ROC Curve (AUC):** The area under Receiver Operating Characteristic curve measuring the model ability to distinguish between classes across all classification thresholds.
2. **Clinical Metrics:**
   - **Sensitivity (Recall):** The true positive rate, measuring the model ability to correctly identify positive cases
   $$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{6}$$
   - **Specificity:** The true negative rate, measuring the model ability to correctly identify negative cases
   $$\text{Specificity} = \frac{TN}{TN + FP} \tag{7}$$
   - **Precision:** The positive predictive value measuring the proportion of true positives among all predicted positives
   $$\text{Precision} = \frac{TP}{TP + FP} \tag{8}$$
   - **F1-Score:** The harmonic mean of precision and recall,providing balanced measure
   $$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{9}$$

3. **Error Analysis:**

   – **Error Rate:** The proportion of incorrectly classified samples

$$\text{Error Rate} = \frac{FP + FN}{TP + TN + FP + FN} \tag{10}$$

   – **False Negative Rate (FNR):** The proportion of actual positives incorrectly classified as negative

$$\text{FNR} = \frac{FN}{TP + FN} \tag{11}$$

   – **False Positive Rate (FPR):** The proportion of actual negatives incorrectly classified as positive

$$\text{FPR} = \frac{FP}{TN + FP} \tag{12}$$

Supplementary metrics such as Cohen's Kappa and Matthews Correlation Coefficient were also computed and Cohen's Kappa (κ) measures inter rater agreement for categorical items accounting for chance agreement:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{13}$$

where p_o is observed agreement and p_e is expected agreement by chance and Matthews Correlation Coefficient provides balanced measure even with imbalanced classes:

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{14}$$

The validation strategy was defined as the model were trained on training set, model selection and tuning were based on validation set performance and final reported results in the comparative analysis are from this validation set as is standard practice. The test set was used as a final, unbiased check, with results e.g. Test Accuracy: MLP 99.31% confirming models generalizability within the dataset.

3 Results and Analysis

3.1 Overall Performance and Computational Efficiency

The results shows that our all three deep learning models achieved high performance on kidney stone classification task and comprehensive summary of results including performance, clinical and error metrics calculated on validation set and that is presented in Table 1. The Multi Layer Perceptron emerged

Table 1. Comprehensive Kidney Stone Classification Performance Analysis and Clinical Evaluation Metrics

Model	Performance Metrics		Clinical Metrics					Error Analysis	
	Acc (%)	Time (s)	Sen (%)	Spe (%)	Pre (%)	F1 (%)	Err (%)	FN (%)	FP (%)
AE-DNN	99.47	4.2	99.47	98.87	99.48	99.47	0.53	0.53	1.13
MLP	**99.67**	**1.5**	**99.67**	**99.44**	**99.67**	**99.67**	**0.33**	**0.33**	**0.56**
DNN	98.95	5.3	98.95	97.75	98.97	98.94	1.05	1.05	2.25

as the top performing model, achieving highest validation accuracy of 99.67% and it was closely followed by Autoencoder based DNN at 99.47% accuracy while Deep Neural Network established strong baseline with 98.95% accuracy. This schema was consistent across all primary clinical metrics with MLP also attaining highest sensitivity (99.67%) specificity (99.44%) precision (99.67%) and F1-score (99.67%).Critical differentiator was computational efficiency and the MLP compact architecture resulted in dramatically faster training time of 0.58 seconds per epoch which is order of magnitude more efficient than DNN (6.34 s/epoch) and AE-DNN (8.42 s/epoch)

3.2 Comprehensive Model Evaluation and Clinical Relevance

The confusion matrix presented in Figure 2 provides view of classification performance across all models and the matrix shows average normalized confusion values and absolute counts (in parentheses) confirming high accuracy and low error rates reported in Table 1. The minimal off diagonal values visually corroborate low false negative and false positive rates achieved by all models particularly the MLP. High sensitivity (Recall) of all models (98.95-99.67%) is essential as it minimizes false negatives critical factor where missing kidney stone diagnosis could have serious consequences for the patient. The specificity (97.75-99.44%) is equally important reducing false positives that could lead to unnecessary follow up procedures and patient anxiety. The MLP balanced performance with near perfect sensitivity of 99.67% and highest specificity of 99.44% positions it the most clinically reliable model in this study

3.3 Limitations and Statistical Considerations

While results demonstrates strong performance several limitations warrant consideration. The perfectly balanced 50:50 class distribution while methodologically appropriate for comparative analysis does not reflect real world clinical prevalence rate, potentially affecting predictive values in practical deployment and the decision to flatten images and employ fully connected networks though intentional for architectural comparison discarded valuable spatial information that the convolutional networks typically exploit. The consistent high performance across all model while promising raises questions about potential ceiling

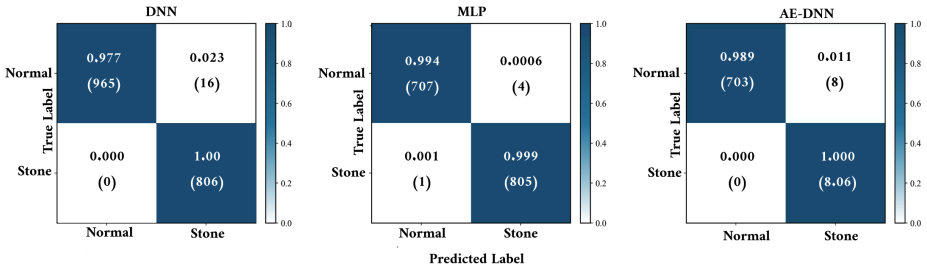

Fig. 2. Confusion Matrix for All Models

effects in specific dataset and the perfect scores could indicate exceptional model performance or binary classification task may be relatively straightforward for deep learning models given the current data characteristics.

4 Conclusion and Future Work

This study demonstrated effectiveness of three deep learning architectures for kidney stone classification with the MLP achieving the highest accuracy (99.67%) while maintaining the superior computational efficiency and the consistent high performance across all models highlights their potential for clinical deployment. Future work will focus on the developing of an integrated diagnostic tool implementing Explainable AI techniques including SHAP and LIME for model interpretability conducting multi center validation and expanding to multi class classification for different stone types and this research contributes to AI powered diagnostic tools that can enhance precision and efficiency in urological practice.

References

1. W. Kittanamongkolchai, L. E. Vaughan, F. T. Enders, T. Dhondup, R. A. Mehta, A. E. Krambeck, C. H. McCollough, T. J. Vrtiska, J. C. Lieske, and A. D. Rule, "The changing incidence and presentation of urinary stones over 3 decades," in *Mayo Clinic Proceedings*, vol. 93, no. 3. Elsevier, 2018, pp. 291–299.
2. A. Chewcharat and G. Curhan, "Trends in the prevalence of kidney stones in the united states from 2007 to 2016," *Urolithiasis*, vol. 49, no. 1, pp. 27–39, 2021.
3. Y. A. Asaye, P. Annamalai, and L. G. Ayalew, "Detection of kidney stone from ultrasound images using machine learning algorithms," *Scientific African*, vol. 28, p. e02618, 2025.
4. O. Iparraguirre-Villanueva, G. Paucar-Palomino, and C. Paulino-Moreno, "From data to diagnosis: evaluation of machine learning models in predicting kidney stones," *Neural Computing and Applications*, pp. 1–14, 2025.
5. A. R. Panda, J. Tripathy, M. K. Mishra, L. Mohanty, J. J. Jena, and M. K. Gourisaria, "Kidney stone prediction based on urine analysis: A comprehensive study of machine learning models," in *2025 International Conference on Emerging Systems and Intelligent Computing (ESIC)*, 2025, pp. 748–753.

6. A. Khan, R. Das, and M. Parameshwara, "Detection of kidney stone using digital image processing: a holistic approach," *Engineering Research Express*, vol. 4, no. 3, p. 035040, 2022.
7. P. Kumar, D. Singh, and J. S. Samagh, "A hybrid model for kidney stone detection using deep learning," *IJSTM*, vol. 13, pp. 65–85, 2024.
8. Y. Liu, H. Song, D. Luo, R. Xu, Z. Xu, B. Wang, W. Hu, B. Xiao, G. Zhang, and J. Li, "Integrated radiomics and deep learning model for identifying medullary sponge kidney stones," *Frontiers in Medicine*, vol. 12, p. 1623850, 2025.
9. Y. J. Jacob, B. Janney *et al.*, "Optimised hybrid deep learning classification model for kidney stone diagnosis," *Results in Engineering*, p. 105221, 2025.
10. I. Rehman and H.-A. Pham, "Cortex vision: Detection of ophthalmic disease using machine learning algorithm," in *International Conference on Smart Objects and Technologies for Social Good*. Springer, 2024, pp. 138–149.
11. Kaggle, "Kidney stone classification dataset," https://www.kaggle.com/datasets/nazmul0087/ct-kidney-dataset-normal-cyst-tumor-stone, 2024, accessed: 2024-10-27.