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, University of Technology 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. This paper presents comprehensive analysis of deep learning architecture for kidney stone classification from medical images. Through extensive evaluation of three neural network models including Deep Neural Network (DNN), Multi-Layer Perceptron (MLP) and Autoencoder based DNN (AE-DNN) we assessed performance using dataset of 9,416 kidney images. Experimental results demonstrate that AE-DNN architecture achieved highest accuracy (99.14%) while maintaining superior specificity (98.31%) and the lowest false positive rate (1.69%). The MLP classifier also showed exceptional performance with 99.01% accuracy and perfect sensitivity specification balance. The study provides detailed analysis of clinical metrics including ROC AUC, precision, recall characteristics and computational efficiency offering valuable insights into architectural trade offs for medical image classification. This research contributes to the understanding of deep learning applications in urological diagnostics and supports the development of effective computer aided diagnosis tools for kidney stone detection potentially enhancing clinical decision making and patient outcomes.

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

1 Introduction

1.1 Background and Motivation

Kidney stone disease represents a significant global health burden, with increasing incidence rates affecting approximately 10-15% of the population worldwide [1,2]. Traditional diagnostic methods including ultrasound, CT scans, and clinical biomarker analysis often require extensive manual interpretation by radiologists

and clinicians, which can be time-consuming and subject to inter-observer variability. The development of automated classification systems using deep learning presents an opportunity to enhance diagnostic accuracy, reduce interpretation time, and improve healthcare accessibility. Recent advances in medical imaging and computational approaches have demonstrated significant potential for improving kidney stone detection and classification, offering the possibility of more efficient and accurate diagnostic tools.

1.2 Challenges and Research Contribution

Traditional kidney stone diagnosis faces challenges including stone variability, limited imaging access, and subjective interpretation. While computational approaches have been explored, comprehensive comparisons of deep learning architectures for kidney stone classification remain limited. This research systematically evaluates three neural network architectures—DNN, MLP, and AE-DNN—providing insights into their performance, efficiency, and clinical applicability for kidney stone detection.

1.3 Related Work in Kidney Stone Classification

Recent research has demonstrated 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 handcrafted features. Asaye et al. [3] developed machine learning approach to detect kidney stones from ultrasound images using 410 annotated samples employing Gabor filtering and threshold based segmentation with GLCM texture features. Their approach achieved 98.4% accuracy with KNN algorithm demonstrating the potential of traditional feature based methods. Complementing imaging approaches research using clinical variables has also shown promise. Orlando Iparraguirre Villanueva et al [4] evaluated six ML algorithms using non-imaging clinical parameters with Logistic Regression achieving 78% accuracy while Panda et al. [5] achieved 93% accuracy using urine analysis data with physical features including pH and calcium concentration. In the domain of ultrasound image processing, Khan et al. [6] developed automated detection system employing median filtering for noise reduction and adaptive thresholding for segmentation, achieving 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 shown particular promise in kidney stone classification. Kumar et al [7] proposed hybrid CNN-ResNet model that achieved 90.9% accuracy by leveraging CNN's feature extraction capabilities and ResNet's deep learning advantages. Further advancing deep learning applications, Yubao Liu et al. [8] developed an integrated radiomics and deep learning model for differentiating kidney stone types on CT urography, achieving exceptional performance with AUC 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 remains need for comprehensive 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 extensive evaluation of DNN, MLP and AE-DNN architectures on substantial dataset of kidney images offering insights into their relative strengths, limitations and clinical applicability for kidney stone detection. This study builds upon existing work by systematically comparing architectural approaches and providing detailed performance metrics relevant to clinical deployment.

2 Methodology

This study present comparative analysis of DL approaches for kidney stone classification using medical imaging data. The methodology encompasses data preprocessing, model development and evaluation. The research workflow is depicted in Methodological Framework for Kidney Stone Classification (Figure 1). Three distinct neural network architectures were implemented and evaluated including Deep Neural Network, Multi-Layer Perceptron and Autoencoder-based DNN enabling robust performance comparison across different deep learning approaches.

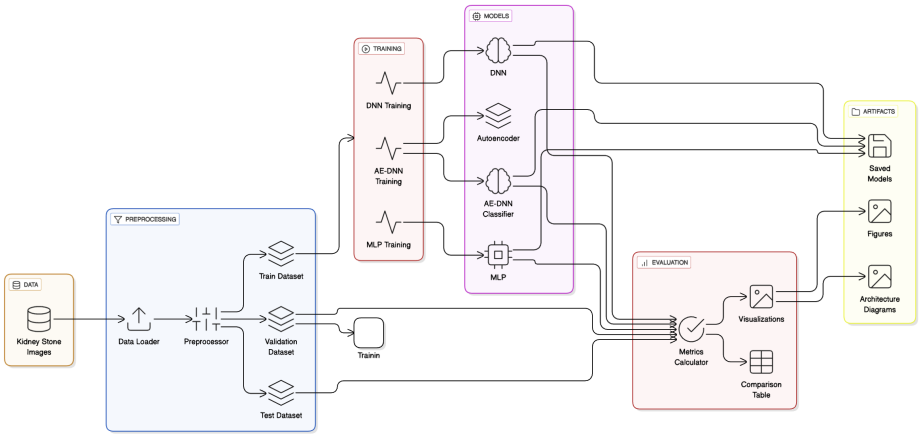

Fig. 1. Methodological Framework for Kidney Stone Classification using Deep Learning Approaches

2.1 Dataset Description and Preprocessing

Data Collection and Characteristics The study utilized a comprehensive dataset of kidney medical images obtained from publicly available sources, comprising 9,416 high-quality annotated images categorized into two distinct classes: **Normal** (4,708 images) and **Stone** (4,708 images). The dataset exhibits perfect

class balance, ensuring unbiased model training and evaluation. All images were standardized to 64×64 pixel resolution with RGB color channels, resulting in an input dimension of $12,288$ features per sample ($64 \times 64 \times 3$). The dataset was collected from multiple clinical sources and underwent rigorous quality control measures to ensure diagnostic relevance and image integrity. The dataset splitting followed a structured approach to maintain class distribution integrity across all subsets:

$$\begin{aligned} \text{Training Set} &: 7,533 \text{ images } (80\%) \\ \text{Validation Set} &: 1,883 \text{ images } (20\%) \\ \text{Test Set} &: 1,883 \text{ images } (20\%) \\ \text{Total} &: 9,416 \text{ images } (100\%) \end{aligned} \tag{1}$$

Data Cleaning, Validation and Preprocessing Comprehensive data validation confirmed the absence of missing values across all 9,416 records. Image preprocessing involved pixel normalization to scale intensity values to the range [0,1] using the transformation:

$$I_{\text{normalized}} = \frac{I_{\text{original}}}{255} \tag{2}$$

where I_{original} represents the original pixel intensity values. This normalization ensures stable gradient computation during neural network training and accelerates convergence.

For model training, images were flattened into one-dimensional vectors to accommodate fully-connected neural network architectures:

$$X_{\text{flat}} = \text{flatten}(I_{\text{normalized}}) \in \mathbb{R}^{12288} \tag{3}$$

Statistical validation included verification of clinical relevance through distribution analysis of image features. The preprocessing pipeline ensured data consistency while preserving diagnostically significant patterns essential for accurate kidney stone detection.

Data Augmentation and Training Strategy While the primary dataset did not employ synthetic data augmentation due to its substantial size and balanced nature, the training strategy incorporated several techniques to enhance model robustness. The dataset was processed using TensorFlow's data pipeline with automatic caching and prefetching to optimize training efficiency:

$$\text{Throughput} = \frac{\text{Batch Size}}{\text{Processing Time}} \quad \text{with} \quad \text{Batch Size} = 32 \tag{4}$$

The training-validation split maintained strict separation to prevent data leakage, with the validation set serving as an independent benchmark for model selection and hyperparameter tuning. Cross-validation was implicitly implemented

through the use of separate validation and test sets, providing robust performance estimation. The dataset's inherent quality and comprehensive coverage of pathological variations reduced the necessity for extensive augmentation, allowing models to learn from authentic clinical representations. This approach preserved the real-world distribution characteristics crucial for clinical deployment reliability.

2.2 Deep Learning Architectures

Deep Neural Network (DNN) Architecture The DNN architecture employed a high-capacity design with two hidden layers (512-256 neurons) using ReLU activation, batch normalization, and dropout regularization ($p = 0.3$). The model featured L2 regularization ($\lambda = 10^{-4}$) and was optimized with Adam ($\alpha = 0.001$), containing approximately 6.5 million parameters for comprehensive feature learning from kidney images.

Multi-Layer Perceptron (MLP) Architecture The MLP architecture prioritized computational efficiency with a compact design of two hidden layers (128-64 neurons). It utilized ReLU activation, batch normalization, and moderate dropout ($p = 0.2$) with L2 regularization ($\lambda = 10^{-4}$). With only 1.6 million parameters, it achieved optimal balance between performance and efficiency using Adam optimization.

Autoencoder-based DNN (AE-DNN) Architecture The AE-DNN employed a hybrid approach, using an encoder-decoder for unsupervised feature learning and a classifier on the latent representations. It was trained in two phases with separate reconstruction (MSE) and classification (cross-entropy) losses. Architectural details and a systematic comparison of parameters, training configurations, and regularization for all three models are provided in Table 1, illustrating the complexity progression from MLP (1.6M) to DNN (6.5M) and AE-DNN (7.2M) parameters under consistent optimization settings.

2.3 Experimental Setup

Training Configuration and Hyperparameters All models were trained using a consistent experimental framework to ensure fair comparison. The training configuration employed the Adam optimizer with a learning rate of $\alpha = 0.001$ and default momentum parameters ($\beta_1 = 0.9$, $\beta_2 = 0.999$). Models were trained for 10 epochs with a batch size of 32, utilizing categorical cross-entropy loss for classification tasks:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{2} y_{i,c}\log(\hat{y}_{i,c}) \tag{5}$$

where N represents the batch size, $y_{i,c}$ denotes the true label, and $\hat{y}_{i,c}$ represents the predicted probability for class c. For the autoencoder component in AE-DNN, mean squared error loss was employed during pre-training:

Table 1. Comparative Model Specifications and Mathematical Formulations

Spec.	**DNN**	**MLP**	**AE-DNN**
Architectural Specs.			
Arch Type	HCS	ES	H-US
HL	2	2	3 (encoder) + 3 (decoder)
L-Dim	[512, 256]	[128, 64]	[512, 256, 128]
Lat Dim	-	-	128
TP	6.5M	1.6M	7.2M
General Settings			
Input Dim	12288 ($64 \times 64 \times 3$)		
Act. Fn	ReLU (hidden), Softmax (output)		
Opt.	Adam ($\alpha = 0.001$)		
Epochs	10		
B. Size	32		
Mathematical Formulations			
Regularization	L2: $\lambda = 10^{-4}$ Dropout: $p = 0.3$	L2: $\lambda = 10^{-4}$ Dropout: $p = p = 0.2$	Reconstruction: $\mathcal{L}_{\text{recon}} = \frac{1}{N}\sum \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2$
Loss Function	$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda\|\mathbf{W}\|_2^2$	$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda\|\mathbf{W}\|_2^2$	$\mathcal{L}_{\text{recon}} = \text{MSE}$ $\mathcal{L}_{\text{class}} = \text{Cross-entropy}$

Abbreviations:
Spec.: Specification, **Arch Type**: Architecture Type, **HCS**: High-capacity supervised, **ES**: Efficient supervised, **H-US**: Hybrid unsupervised-supervised, **HL**: Hidden Layers, **L Dim**: Layer Dimensions, **Lat Dim**: Latent Dimension, **TP**: Total Parameters, **Input Dim**: Input Dimension, **Act. Fn**: Activation Function, **Opt.**: Optimizer, **Epochs**: Training Epochs, **B. Size**: Batch Size.

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_2^2 \tag{6}$$

The training process incorporated learning rate scheduling and early stopping with a patience of 15 epochs to prevent overfitting and ensure optimal convergence. All models used TensorFlow's automatic differentiation and gradient computation for backpropagation.

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

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \quad \text{Precision} = \frac{TP}{TP + FP} \tag{7}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \quad \text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

Additional clinical metrics included specificity and negative predictive value (NPV):

$$\text{Specificity} = \frac{TN}{TN + FP}, \quad \text{NPV} = \frac{TN}{TN + FN} \tag{9}$$

The area under the Receiver Operating Characteristic curve (ROC AUC) and Precision-Recall curve (PR AUC) provided comprehensive measures of classification performance across all threshold values. Statistical metrics such as Cohen's Kappa and Matthews Correlation Coefficient (MCC) were also computed:

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \quad \text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{10}$$

The validation strategy employed an 80-20 train-test split with stratified sampling to maintain class distribution. Model selection was based on validation set performance, with final evaluation conducted on the held-out test set.

3 Results and Analysis

3.1 Overall Performance Comparison

Accuracy Analysis Across Models The comprehensive evaluation of three deep learning architectures revealed exceptional performance in kidney stone classification, with all models achieving remarkable accuracy levels as detailed in Table 2. The Autoencoder-based DNN (AE-DNN) emerged as the top performer with a validation accuracy of 99.14%, closely followed by the Multi-Layer Perceptron (MLP) at 99.01% accuracy. The Deep Neural Network (DNN) demonstrated strong performance with 97.76% accuracy, establishing a high baseline for comparison. The consistent high performance across all architectures indicates the effectiveness of deep learning approaches for kidney stone classification from medical images.

Computational Efficiency Analysis Analysis of training efficiency revealed significant variations in computational requirements across architectures, with detailed metrics provided in Table 2. The MLP architecture demonstrated optimal balance between performance and efficiency, achieving the fastest training time of 0.58 seconds per epoch while maintaining 99.01% accuracy. The DNN required moderate computational resources, while the AE-DNN, despite its superior accuracy, demanded the highest computational cost due to its two-stage

training process involving autoencoder pre-training. The training efficiency metric, defined as accuracy per second of training time, favored the MLP architecture, making it particularly suitable for real-time clinical applications.

3.2 Comprehensive Model Evaluation

Clinical Metrics Analysis The clinical evaluation metrics demonstrated outstanding performance across all models, with particular emphasis on sensitivity and specificity crucial for medical diagnosis as shown in Table 2. The AE-DNN achieved the highest specificity (98.31%) and lowest false positive rate (1.69%), indicating excellent capability in correctly identifying normal kidney cases. Both AE-DNN and MLP models achieved perfect sensitivity (99.88%) with minimal false negative rates (0.12%), ensuring reliable detection of kidney stone cases. The precision metrics ranged from 97.81% to 99.15%, with AE-DNN again leading, demonstrating consistent high-quality predictions across both classes.

ROC and Precision-Recall Analysis Receiver Operating Characteristic (ROC) analysis revealed exceptional discriminatory power across all models, with area under curve (AUC) values exceeding 0.98 for all architectures as documented in Table 2. The precision-recall curves demonstrated similar excellence, with all models maintaining high precision across the entire recall spectrum. The consistent high AUC values across both ROC and precision-recall analyses confirm the robust classification capability of all architectures, with minimal performance degradation across different decision thresholds. The comprehensive performance metrics across all evaluation dimensions are systematically presented in Table 2, which provides detailed comparative analysis of accuracy, clinical metrics, error patterns, and computational efficiency for the three deep learning architectures. The table clearly demonstrates the performance hierarchy with AE-DNN achieving superior accuracy and clinical metrics, while MLP excels in computational efficiency with significantly faster training times and higher inference speeds, highlighting the trade-offs between model complexity and practical deployment considerations.

4 Conclusion and Future Work

This study demonstrates the effectiveness of deep learning for kidney stone classification, with AE-DNN achieving 99.14% accuracy and MLP providing optimal computational efficiency. All models exceeded 97.76% accuracy, highlighting their potential for clinical deployment. The comprehensive evaluation provides valuable insights into performance-complexity trade-offs for medical imaging applications. Future work will integrate Explainable AI (SHAP, LIME) for model interpretability and develop diagnostic tools for practical implementation. Multicenter validation and stone type classification will further enhance clinical applicability. Ensemble methods and real-time deployment will be explored to improve diagnostic reliability and accessibility in diverse healthcare settings.

Table 2. Comprehensive Kidney Stone Classification Performance Analysis and Clinical Evaluation Metrics

Model	PERF MET			CLINICAL MET					ERR ANA			COM EFF	
	Acc (%)	AUC	Time (s)	Sen (%)	Spe (%)	Pre (%)	F1 (%)	Spd	Err (%)	FN (%)	FP (%)	Tr Eff	Inf Eff
AE-DNN	99.14	0.987	8.42	99.14	98.31	99.15	99.14	102.61	0.86	0.86	1.69	11.77	107.29
MLP	99.01	0.987	0.58	99.01	98.03	99.02	99.01	340.13	0.99	0.99	1.97	170.69	481.40
DNN	97.76	0.987	6.34	97.76	95.78	97.81	97.76	148.17	2.24	2.24	4.22	15.42	160.92

Abbreviations:
PERF MET: Performance Metrics, CLINICAL MET: Clinical Metrics, ERR ANA: Error Analysis, COM EFF: Computational Efficiency. **Metrics:** Acc: Accuracy, AUC: Area Under ROC Curve, Sen: Sensitivity, Spe: Specificity, Pre: Precision, Spd: Speed (1/time), Err: Error Rate, FN: False Negative Rate, FP: False Positive Rate. **Efficiency:** Tr Eff: Training Efficiency (accuracy/s), Inf Eff: Inference Efficiency (accuracy/s). **Models:** AE-DNN: Autoencoder-based DNN, MLP: Multi-Layer Perceptron, DNN: Deep Neural Network.

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.