

Impact of Wavelet Pre-processing on Deep Learning Architectures for Renal Cancer Diagnosis

Impacto do Pré-processamento com Wavelet em Arquiteturas de Aprendizado Profundo para o Diagnóstico de Câncer Renal

Vinícius Meireles Pereira Santos^{1*}, Ana Claudia Patrocinio², Pedro Moises de Sousa¹

Abstract: Renal cancer is a lethal neoplasm where early detection is critical yet challenging. To define an optimal diagnostic pipeline, this study conducts a comparative analysis on a public tomography dataset (CT KIDNEY DATASET), evaluating the impact of a wavelet filter. Six deep learning architectures were tested, spanning classic (VGG16), modern (ResNet50), efficient (MobileNetV2), and attention-based (Vision Transformer - ViT) paradigms. Our results reveal a complex, architecture-dependent impact: wavelet pre-processing enables a ResNet50 to achieve a test accuracy of 100.00%. However, we demonstrate that a more efficient architecture, MobileNetV2, reaches the same state-of-the-art performance without this pre-processing step. These findings highlight that the optimal diagnostic pipeline is a balance between architectural efficiency and pre-processing complexity, positioning MobileNetV2 as the most pragmatic and effective solution.

Keywords: deep learning — renal cancer diagnosis — computed tomography — vision transformer — wavelet transform — explainable AI — computer-aided diagnosis

Resumo: O câncer renal é uma neoplasia letal onde a detecção precoce é crítica, mas desafiadora. Para definir um *pipeline* diagnóstico ótimo, este estudo realiza uma análise comparativa em um *dataset* público de tomografias (CT KIDNEY DATASET), avaliando o impacto de um filtro *wavelet*. Foram testadas seis arquiteturas de *deep learning*, abrangendo paradigmas clássicos (VGG16), modernos (ResNet50), eficientes (MobileNetV2) e baseados em atenção (Vision Transformer - ViT). Nossos resultados revelam um impacto complexo e dependente da arquitetura: o pré-processamento com *wavelet* permite que uma ResNet50 atinja uma acurácia de teste de 100.00%. Contudo, demonstramos que uma arquitetura mais eficiente, a MobileNetV2, alcança a mesma performance de estado da arte sem essa etapa de pré-processamento. Estes achados destacam que o *pipeline* diagnóstico ótimo é um balanço entre a eficiência da arquitetura e a complexidade do pré-processamento, posicionando a MobileNetV2 como a solução mais pragmática e eficaz.

Palavras-Chave: aprendizado profundo — diagnóstico de câncer renal — tomografia computadorizada — vision transformer — transformada wavelet — IA explicável — diagnóstico auxiliado por computador

¹Instituto de Ciências Exatas e Tecnológicas, Universidade Federal de Viçosa (UFV), Brazil

²Laboratório de Biomedicina, Universidade Federal de Uberlândia (UFU), Brazil

*Corresponding authors: ¹vinicius.m.santos@ufv.br, pedromoises@ufv.br, ²ana.patrocinio@ufu.br

DOI: <http://dx.doi.org/10.22456/2175-2745.150824> • Received: 07/10/2025 • Accepted: 27/10/2025

CC BY-NC-ND 4.0 - This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

1. Introduction

Renal cancer, although less common than other tumors, exhibits high lethality among urological neoplasms, partly due to its silent progression, which hinders early diagnosis [2]. In Brazil, data from INCA reinforce the severity of the recent scenario [1]. Without population-based screening and with only subtle signs in imaging, there is a growing demand for computational tools capable of enhancing early detection from computed tomography (CT) scans, where Deep Learning (DL) approaches have shown strong performance [4].

Despite these advancements, the literature still lacks a systematic analysis that directly compares the main architectural paradigms—CNNs and Vision Transformers (ViT)—under the effect of advanced pre-processing techniques, whose interaction remains poorly understood. To fill this gap, this work presents a robust comparative study to determine the optimal computational pipeline for renal cancer detection by systematically evaluating classic and modern architectures (including a ViT), investigating the impact of a wavelet-based enhancement pipeline, and using Grad-CAM to elucidate the models' decision processes. It is important to clarify that our

contribution is not the proposal of wavelet transforms as a novel technique, but rather a systematic empirical study of their impact across different architectural paradigms for this specific task. The ultimate goal is to identify the combination of architecture and pre-processing that offers the most accurate, efficient, and reliable solution to assist in early diagnosis.

2. Related Works

Deep Learning for renal cancer classification is well-established [14], with CNNs proving effective in various tasks such as staging and subtype classification [6, 7]. In a previous study [8], we demonstrated that lightweight architectures like MobileNetV2 can achieve state-of-the-art performance with minimal pre-processing. However, our work left open the question of whether this performance represents a saturation ceiling of the dataset or if targeted pre-processing, such as the Wavelet Transform, could unlock further gains—a question that directly motivates the present investigation.

One such technique is the Wavelet Transform, whose success in medical image analysis is notable. Its ability to decompose signals across multiple frequencies is used for texture enhancement in mammograms [9] and brain MRI scans [10]. This approach is particularly promising for renal CT images, where enhancing subtle edges and textural variations, often obscured by low contrast, is crucial for identifying lesions.

In parallel, the dominance of CNNs is being challenged by attention-based architectures, most notably the Vision Transformer (ViT) [11]. The ability of ViTs to model long-range dependencies between different image regions offers an alternative paradigm to CNNs, which are inherently focused on local features. This approach has already demonstrated efficacy in other medical imaging domains [12, 13], justifying its evaluation in the context of renal anatomy, where global organ context is relevant.

The literature reveals three largely independent research streams: CNNs applied to renal cancer, the use of wavelets for image enhancement, and the rise of Transformers. The intersection of these areas—a systematic analysis of the Transformer paradigm against CNNs under the influence of wavelet pre-processing—remains a critical gap that this study addresses. The present study directly addresses this gap by providing a novel and systematic comparative analysis that integrates these three technological fronts.

3. Materials and Methods

3.1 Dataset and Class Balancing

The dataset for this study, "CT KIDNEY DATASET: Normal-Cyst-Tumor and Stone," was sourced from the Kaggle platform and is described in [15]. The original repository contains 12,446 computed tomography (CT) images across four categories. Focusing on a binary classification task, we selected all images from the "Tumor" class (2,283 images) and the "Normal" class (5,077 images).

This initial selection presents a significant class imbalance, a well-documented problem that can bias models toward the majority class—an unacceptable risk in medical diagnostics [16, 29, 30]. To mitigate this bias, we implemented a random undersampling strategy, an approach previously validated in this context by Hadjiyski et al. [6]. Undersampling was chosen over oversampling techniques to avoid introducing synthetic artifacts, thus preserving the clinical relevance of the training data. Although this approach involves discarding a portion of the majority class data, it robustly mitigates the risk of class bias while maintaining data integrity.

The final balanced dataset consisted of 4,566 images (2,283 per class). Figure 1 illustrates examples from each class, displaying the images in their original grayscale format and after the wavelet filtering detailed in Subsection 3.3.

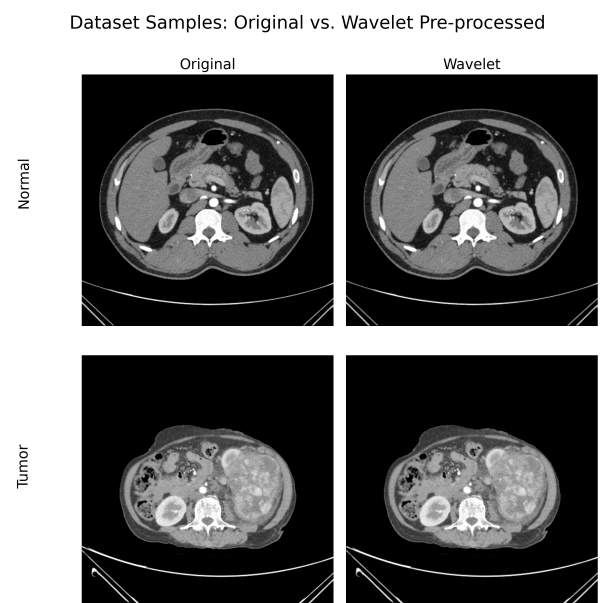


Figure 1. Examples of images from the original dataset and with Wavelet filtering

3.2 Evaluated Deep Learning Architectures

To conduct a comprehensive comparative analysis, six deep learning architectures were selected, representing different eras and paradigms of computer vision. The selection was deliberately structured to allow for a robust and multifaceted analysis, comparing classic models, modern industry standards, efficient architectures, and the emerging Transformer paradigm.

The classic architectures AlexNet [18] and VGG16 [20] were included as baselines to evaluate the impact of subsequent innovations, such as Batch Normalization, which is absent in these models.

ResNet50 [17] was selected as the canonical representative of modern CNNs, given the importance of its residual blocks in mitigating the gradient degradation problem.

To evaluate the trade-off between performance and efficiency, a critical factor for clinical feasibility, we selected MobileNetV2 [19] and EfficientNet-B7 [21], which represent efficiency approaches via depthwise separable convolutions and joint network scaling optimization, respectively.

Finally, the Vision Transformer (ViT-B-16) [22] was included to contrast the dominant convolutional paradigm with an approach based on global attention, testing its inductive bias for this task.

3.3 Pre-processing and Data Augmentation

3.3.1 Wavelet Pipeline Optimization

To determine the most effective pre-processing method, we conducted an optimization experiment to select the best combination of wavelet family and smoothing filter. We started from the hypothesis that the maximum preservation of the original image's structural information would be a prerequisite for the models' performance. To test this, a systematic sweep was performed on a set of 65 images. This process evaluated 7 wavelet families¹, combined with Mean and Gaussian filters.

The quality of each combination was quantified using two canonical image quality metrics: Structural Similarity (SSIM), which measures the perceptual similarity between two images [32], and Peak Signal-to-Noise Ratio (PSNR), which evaluates the magnitude of distortion [31]. As summarized in Table 1, the results showed that the Gaussian filter was consistently superior to the Mean filter. The combination of the bior3.7 wavelet with the Gaussian filter achieved the highest average values (SSIM of 0.9898; PSNR of 38.75 dB), indicating near-total structural preservation. Therefore, this pipeline was selected for data pre-processing.

Table 1. Comparative results of the wavelet pipeline optimization. The SSIM and PSNR metrics were calculated as the average over the entire test set of 65 images. The highest value indicates better preservation of the image structure.

Filter	Wavelet Family	SSIM (average)	PSNR (dB, average)
Gaussian	bior3.7	0.9898	38.75
	sym4	0.9894	38.25
	db4	0.9894	38.43
Mean	bior3.7	0.9477	31.19
	coif3	0.9462	31.06
	fk8	0.9461	31.07

3.3.2 Creation of Datasets for Comparison

From the balanced dataset of 4,566 images, all previously converted to 8-bit grayscale, two distinct datasets were created to allow for a controlled comparative analysis:

- **Original Dataset:** Containing the grayscale images with no additional processing.

¹The evaluated families were: Daubechies (db), Symlets (sym), Coiflets (coif), Biorthogonal (bior), Reverse Biorthogonal (rbio), Fejér-Korovkin (fk), and Meyer (dmev).

- **Wavelet Dataset:** Containing the same images after the application of a Wavelet Transform-based filter, using the Biorthogonal 3.7 (bior3.7) family and a Gaussian filter on the coefficients, as optimized in preliminary tests.

3.3.3 Partitioning, Normalization, and Data Augmentation

The total set of 4,566 images was first partitioned in a stratified and fixed manner into training (60%, 2,740 images), validation (20%, 913 images), and test (20%, 913 images) sets. The reproducibility of this split across all experiments was ensured by using a random seed (SEED) of 42.

Next, both datasets (Original and Wavelet) underwent a standard preparation process. All images were resized to 224x224 pixels and converted to PyTorch tensors. Normalization statistics (mean and standard deviation) were calculated exclusively from the training set of each dataset and applied to all their respective sets (train, validation, and test).

Finally, to increase the models' robustness to the small geometric variations inherent in CT image acquisition, the training set was enriched with data augmentation. The applied techniques included affine transformations, such as random horizontal flips (RandomHorizontalFlip), random rotations of up to 10 degrees (RandomRotation), and random affine transformations (RandomAffine). The validation and test sets were not augmented, ensuring an unbiased evaluation of the models' generalization capability.

3.4 Training Configuration and Hyperparameters

To ensure reproducibility, all models were trained with a batch size of 16. The Adam optimizer was used, a standard choice in computer vision for its efficiency and adaptive learning rate, configured with an initial learning rate of 1e-3. We used the canonical Cross-Entropy Loss function, with training for up to 500 epochs, using Early Stopping on the validation loss (patience=10) to mitigate overfitting. No weight decay was used, and no learning rate scheduler was applied, keeping it fixed throughout the training.

3.5 Evaluation Metrics and Clinical Relevance

While Accuracy and Loss guided Early Stopping during validation, the final evaluation on the test set used Recall (a clinical priority), Precision, F1-Score, AUC-ROC, and Confusion Matrix analysis. The metrics follow the canonical definitions by Sokolova and Lapalme [23]. We report results on the stratified test set, keeping the clinical interpretation focused on Recall and AUC, supported by F1-Score and Precision.

3.6 Explainability Analysis with Grad-CAM

To address the inherent opacity of Deep Learning models—a challenge known as the “black-box” problem, which represents a barrier to their clinical adoption—we employed the visual explainability technique Grad-CAM (Gradient-weighted Class Activation Mapping) [24].

Table 2. Comparative results of performance and efficiency on the test set. Efficiency is measured by the number of parameters and inference time in milliseconds (ms).

Model	Parameters (M)	Inference (ms)	Dataset	Accuracy (%)	AUC	Precision (Macro %)	Recall (Macro %)	F1-Score (Macro %)	Loss
<i>AlexNet</i>	61.10	2.28	Original	50.06	0.500	25.03	50.00	33.35	0.6933
			Wavelet	61.24	0.697	62.00	61.00	61.00	0.6818
<i>VGG16</i>	138.36	9.21	Original	50.06	0.500	25.03	50.00	33.35	0.6934
			Wavelet	50.06	0.501	25.03	50.00	33.35	0.6939
<i>ResNet50</i>	25.56	6.70	Original	99.78	1.000	99.78	99.78	99.78	0.0049
			Wavelet	100.00	1.000	100.00	100.00	100.00	0.0016
<i>MobileNet V2</i>	3.50	5.76	Original	100.00	1.000	100.00	100.00	100.00	0.0010
			Wavelet	99.89	1.000	99.89	99.89	99.89	0.0055
<i>EfficientNet-B7</i>	66.35	35.99	Original	99.89	1.000	99.89	99.89	99.89	0.0025
			Wavelet	99.78	1.000	99.78	99.78	99.78	0.0055
<i>ViT-B-16</i>	86.57	13.51	Original	88.70	0.929	89.00	89.00	89.00	0.3024
			Wavelet	75.30	0.844	76.00	75.00	75.00	0.4937

Proposed by Selvaraju et al., the technique generates activation maps that provide a visual attribution for the predictions, highlighting the image regions that contributed most to the model's decision. For CNNs, this is obtained from the gradients of the last convolutional layer. Analogously, for the ViT architecture, the activations were extracted from the last attention block of the encoder, allowing for a comparable analysis between the paradigms, following validated approaches for applying gradient-based techniques to Transformers [28].

Grad-CAM was used with a dual objective: to qualitatively verify if the models' focus aligned with pathological regions, and to investigate divergences in inference strategies among the different architectures (CNN vs. ViT) and between high- and low-performance models.

3.7 Experimental Environment and Reproducibility

Experiments were conducted in Python 3.11 / PyTorch 2.6 with ImageNet pre-trained weights (from torchvision). For 1-channel input, we replaced the first layer of the CNNs with a Conv2d layer with in_channels=1 and adjusted the patch projection in the ViT, both with randomly initialized weights. While other strategies exist, this one was chosen for its simplicity and to avoid introducing bias by replicating weights. Training and inference were performed in Google Colab (NVIDIA T4 GPU, CUDA). Latencies were measured with a batch_size=1 after a warm-up period, using the average of 1000 inferences. Auxiliary tools included NumPy, scikit-learn, Matplotlib, Seaborn, and OpenCV. To ensure full reproducibility of the findings, the source code and notebooks used in this study are available in an GitHub repository.²

4. Results

This section presents the quantitative and qualitative results of the comparative analysis. The findings reveal three central points: the inability of the classic architectures to converge; the state-of-the-art performance, numerically perfect in some cases, of multiple modern CNN architectures; and the complex, architecture-dependent impact of wavelet-based pre-processing.

²Available at: (<https://github.com/vinimeirelres/renal-cancer-detection>)

4.1 Quantitative Performance Analysis

The performance of all 12 experimental configurations on the test set is consolidated in Table 2. The data analysis reveals four main observations.

4.1.1 Failure of Baseline Architectures

The classic architectures, AlexNet and VGG16, proved unable to learn the task from the original data, stagnating at an accuracy of 50%, equivalent to a random classifier. Notably, wavelet pre-processing allowed AlexNet to achieve 61.24% accuracy, indicating that the feature enhancement extracted a useful signal, although it was insufficient for competitive performance.

4.1.2 Superior Performance of Modern Models

In sharp contrast, all modern architectures (ResNet50, MobileNetV2, EfficientNet-B7, and ViT-B-16) were able to learn the task successfully, far surpassing the baselines, as illustrated by the learning curves in Figure 2.

4.1.3 Top-Performing Models and Peak Performance

Within the group of modern models, the CNNs ResNet50, MobileNetV2, and EfficientNet-B7 formed an elite cluster, achieving performances nearing perfection (accuracy > 99.7%). Two configurations—ResNet50 with wavelet pre-processing and MobileNetV2 on the original dataset—achieved a perfect test accuracy of 100.00%, with maximum AUC, Precision, and Recall as well (Table 2 and Figure 3).

4.1.4 Vision Transformer Performance

The ViT-B-16, although it demonstrated successful learning on the original dataset (88.70% accuracy), did not reach the same performance level as the state-of-the-art CNNs.

4.2 Observed Impact of Wavelet Pre-processing

A complex, architecture-dependent interaction was observed between the wavelet pre-processing and model performance. For the CNN architectures, the effect varied: it promoted a significant performance gain for models like AlexNet (+11.18%) and ResNet50 (+0.22%), but resulted in a marginal decrease for the more optimized architectures, such as MobileNetV2 (-0.11%) and EfficientNet-B7 (-0.11%). In sharp contrast,

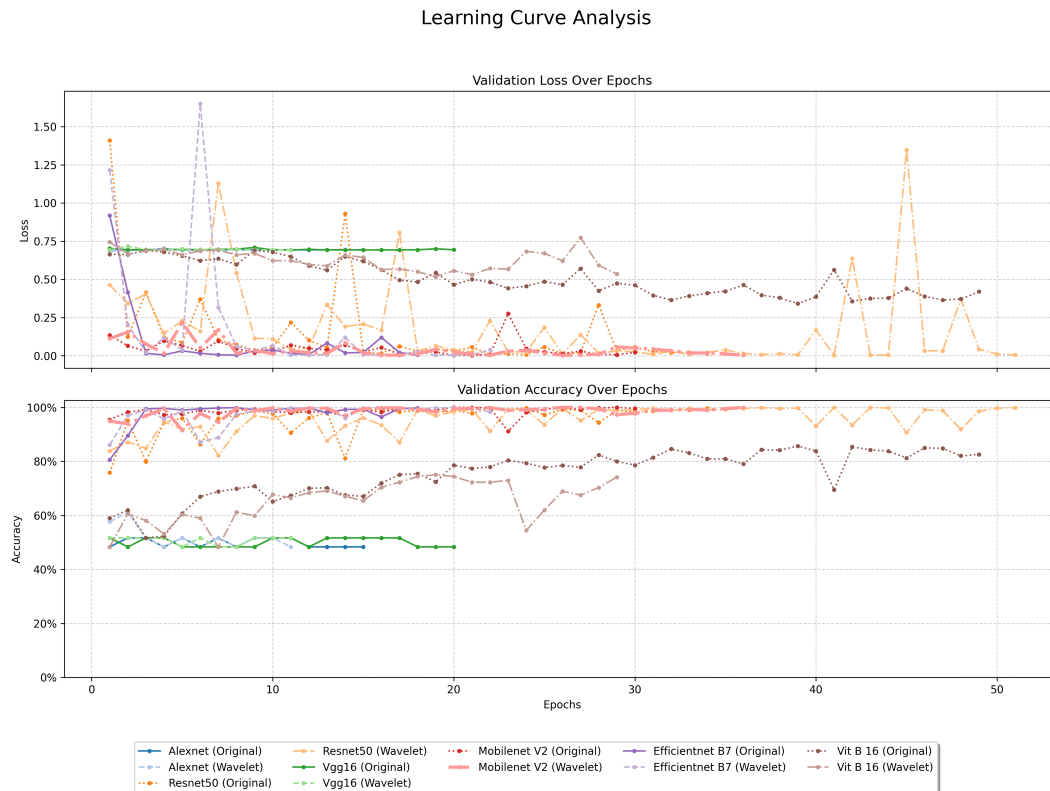


Figure 2. Accuracy and loss curves on the validation set for each of the tested models and datasets. Note the stagnation of the classic architectures (AlexNet, VGG16) and the rapid convergence of the modern models.

for the Vision Transformer architecture, the effect was consistently adverse, resulting in a drastic drop in performance (-13.4%). The interpretation of these findings will be detailed in Section 5.

4.3 Qualitative Analysis with Grad-CAM

The visual analysis with Grad-CAM (Figure 4) corroborates the quantitative findings. The AlexNet heatmap (Figure 4a) proved to be noisy and diffuse, confirming its failure to learn to focus on regions of interest. In contrast, the map for the ResNet50 with wavelet pre-processing (Figure 4b) demonstrates a precise and highly focused activation on the tumor region, corresponding to its 100% performance. Finally, the ViT-B-16 map (Figure 4c) exhibits a contextual, block-like pattern characteristic of the Transformer architecture. This pattern, while effective for classification, does not precisely focus on the lesion but rather on its broader anatomical context. This results in a visibly more diffuse and coarse activation that surrounds the region of interest instead of concentrating on it, in sharp contrast to the focal precision of the ResNet50.

5. Discussion

In this section, we interpret the presented results, discuss their implications for selecting an optimal diagnostic pipeline, and contextualize our findings in relation to the state-of-the-art.

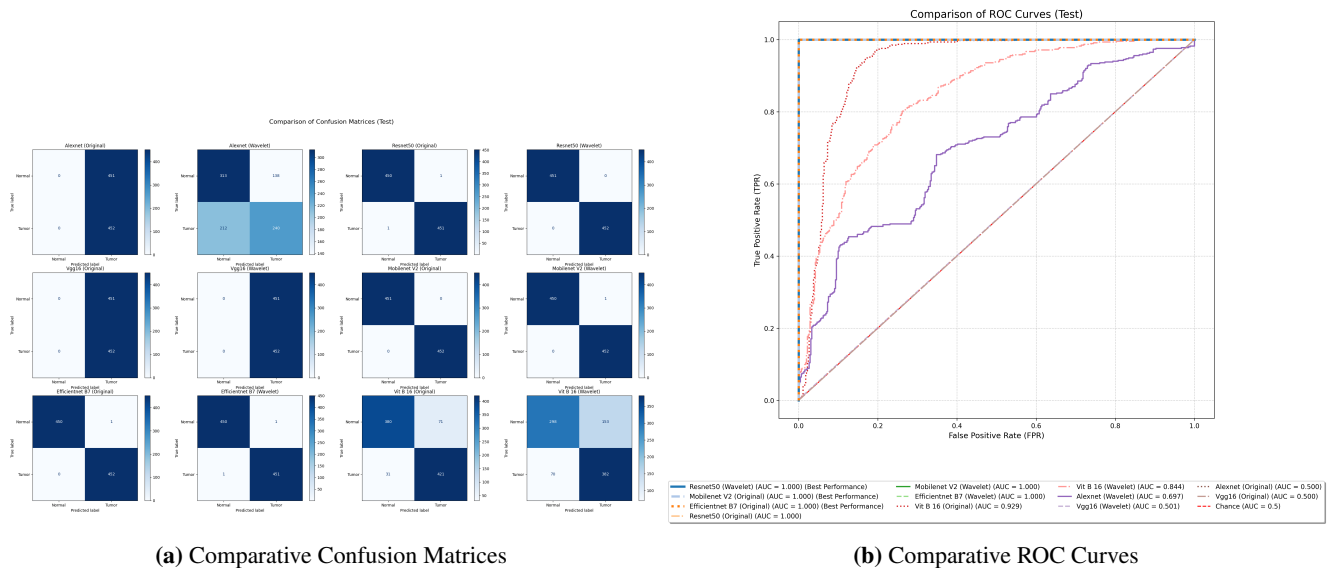
5.1 Interpretation of Performance Findings

The first notable finding is the 100% performance achieved by two distinct configurations. Far from being just an indicator of success, this result strongly suggests that the test dataset is "saturated" by state-of-the-art models, a finding likely attributable to the dataset's intrinsic structure, as detailed in Section 6. This, in turn, shifts the evaluation criterion to other factors, such as computational efficiency and pipeline simplicity, which will be addressed in subsection 5.2.

The second central finding is the asymmetric impact of wavelet pre-processing, which provides strong evidence of a fundamental interaction with the architectural paradigm. While ResNet50 benefited from the enhancement (+0.22%), the ViT was significantly impaired (-13.4%). A plausible hypothesis for this phenomenon is that the filter smoothing, while enhancing useful local features for CNNs, may remove high-frequency textural information that is crucial for the ViT's global self-attention mechanism. This result suggests that pre-processing optimization is even more critical and sensitive for attention-based models.

5.2 Defining the Optimal Pipeline: A Balance Between Accuracy and Efficiency

The final evaluation of a diagnostic model must prioritize clinical relevance. In this context, the 100% Recall for the "Tumor" class, achieved by all modern CNNs, is a fundamental safety prerequisite, ensuring that no cancer cases are missed. Given



(a) Comparative Confusion Matrices

(b) Comparative ROC Curves

Figure 3. Comparative visual analysis of performance on the test set. (a) Confusion matrices for each experiment. (b) Overlaid ROC curves for the high-performance models, with their respective AUC values.

that this threshold was met, Precision (to avoid false positives) and efficiency emerge as the decisive factors.

It is in this balance that the MobileNetV2 architecture (with the original dataset) establishes itself, in our scenario, as the pragmatic choice. It achieved peak performance (100% on all metrics) without the need for the wavelet pre-processing pipeline and with drastically superior computational efficiency, especially in terms of model complexity. The model has only 3.50 million parameters, making it more than 7 times smaller than ResNet50 (25.56 million). Additionally, its inference time (5.76 ms) is slightly faster than that of ResNet50 (6.70 ms), consolidating its advantage in all aspects of efficiency. This result positions MobileNetV2 as the preferable choice for practical applications where precision, safety, and efficiency are critical.

5.3 Contextualization with the State-of-the-Art

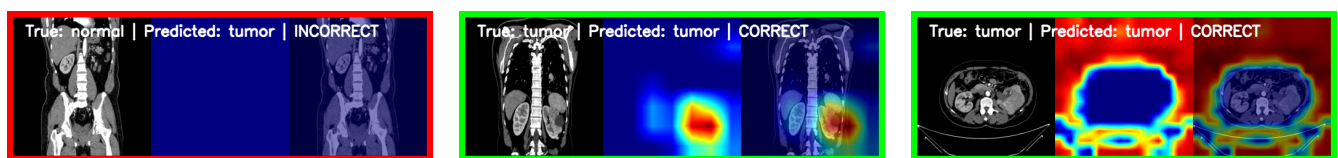
The performance obtained in this work, consolidated in Table 3, positions our approach at the forefront of the state-of-the-art for the dataset used. Our 100% test accuracy results surpass or match those reported in previous studies that used the same dataset. It is crucial to note, as we did in Table 3, that comparisons should be made with caveats: works like

Rajkumar et al. [5] report validation accuracy, which is not directly comparable to test accuracy. Additionally, the 100% result from Qadir et al. [26] refers to a single class within a multiclass problem. The other studies, such as Hadjiyski et al. [6] and Uhm et al. [7], use different datasets (e.g., TCIA), which precludes a direct numerical comparison but serves to contextualize the general performance range in the field. Our findings, obtained on a dedicated test set and with a transparent methodology, strengthen the validation of the effectiveness of modern CNNs for this task.

6. Study Limitations

For a transparent interpretation of our findings, it is essential to delineate the boundaries of the present study. The main limitations are threefold: the structure of the dataset, the methodological simplifications adopted, and the retrospective scope of the analysis.

The primary limitation lies in the generalizability from the public dataset used. The removal of metadata identifying patients prevents a strict patient-level separation between the training and test sets, creating an intrinsic risk of data leakage. This factor is the most likely explanation for the 100% per-



(a) AlexNet (Convergence Failure)

(b) ResNet50 + Wavelet (High Performance)

(c) ViT-B-16 (Block-like Pattern)

Figure 4. Grad-CAM analysis of the attentional focus of different architectures. The heatmaps highlight in red the most relevant areas for classification: (a) AlexNet's convergence failure (diffuse map); (b) ResNet50's precise focus on the tumor area; (c) ViT-B-16's contextual, block-like pattern.

Table 3. Performance comparison with the state-of-the-art (SOTA). The *Context/Comparability* column indicates when the study is directly comparable (same CT dataset and same binary tumor vs. normal task) and when it is not, with the reason.

Author [Citation]	Main Method	Dataset	Accuracy (%)	AUC	Context/Comparability
This Work	ResNet50 + Wavelet	CT Kidney (Kaggle) [15]	100.00	1.000	Directly comparable
Santos et al. [8]	MobileNet V2	CT Kidney (Kaggle) [15]	99.89	–	Directly comparable
Qadir & Abd. [26]	DenseNet-201 + Random Forest	CT Kidney (Kaggle) [15]	100.00 (Tumor class)	–	Not directly comparable (multiclass task; result per class)
Rajkumar et al. [5]	Proposed CNN	CT Kidney (Kaggle) [15]	99.60 (validation)	–	Not directly comparable (validation metric, not test)
Hadjijski et al. [6]	DLNN	TCGA–KIRC (TCIA)	86.00	0.900	Not directly comparable (external/distinct dataset)
Uhm et al. [7]	ResNet–101 (5 subtypes)	Seoul St. Mary's / TCIA	64.00	0.855 (avg)	Not directly comparable (external dataset; multiclass task by subtype)
Abdeljawad et al. [25]	Proposed CNN + GGMRF	Seoul St. Mary's / TCIA	94.60 (avg)	–	Not directly comparable (distinct modality: histology/WSI)

formance achieved, which, while demonstrating the mastery of the architectures in this specific scenario, also indicates dataset saturation, masking the true capability of the models in scenarios with data from multiple clinical centers.

Secondly, the methodology employed deliberate simplifications in data representation and validation. The analysis was based on 2D slices, discarding 3D volumetric information, and the adaptation for 1-channel images involved reinitializing the weights of the first layer, which may have limited the knowledge transfer from pre-training. The validation rigor was moderated using a single data split instead of a cross-validation scheme, and the lack of specification for the Hounsfield Units (HU) window during image conversion affects exact reproducibility.

Additionally, the optimization and training choices, made to ensure comparability, may have influenced the results. Critically, the wavelet pipeline optimization was guided by proxy metrics (SSIM/PSNR), based on the unvalidated assumption that visual fidelity translates to classification performance—a premise that is challenged by the adverse results observed in the ViT. Similarly, the choice of a simplified training regimen (fixed learning rate, no weight decay) ensured a direct comparison but may have underutilized the potential of architectures like the Vision Transformer [22]. The same caveat applies to the evaluation of EfficientNet-B7 at a suboptimal resolution, a choice that favored comparability over the model's peak performance.

Finally, the retrospective scope of the study means that, although the state-of-the-art performance demonstrates the computational efficacy of the approach, the definitive validation of its clinical effectiveness—its real-world impact on the diagnostic workflow—would require a prospective study conducted in a hospital setting.

7. Conclusion

This work conducted a systematic analysis to determine the optimal computational pipeline for renal cancer classification, demonstrating the superiority of modern CNNs and revealing the complex, architecture-dependent interaction of wavelet-based pre-processing. We conclude that, although multiple approaches achieve peak performance on this specific dataset (100% accuracy), the MobileNetV2 architecture, without advanced pre-processing, represents the most effective and pragmatic solution due to its drastic superiority in

computational efficiency. The explainability analysis with Grad-CAM corroborated these findings, confirming the precise focus of CNNs on pathological regions and the contextual, albeit less localized, approach of the Transformer architecture.

The main insight of this study transcends the mere identification of the best model: it highlights the importance of evaluating AI systems for diagnostics not only by their maximum accuracy but by a holistic balance that includes efficiency, simplicity, and implementation feasibility. The optimal solution, as demonstrated, is not necessarily the most complex, but the most pragmatically simple.

The findings of this work establish a solid foundation and open three main avenues for future research: the validation of the optimized pipeline (MobileNetV2) on external and multi-center datasets to test its generalizability in diverse clinical scenarios; the exploration of 3D-CNN architectures to incorporate volumetric information from CT scans; and the transition from a classification task to semantic segmentation, aiming for the precise delineation of the tumor area—a task for which the contextual approach of the ViT, in its current form, proved inadequate, reinforcing the need for further investigation.

Acknowledgements

The authors thank the Federal University of Viçosa (UFV) and the Federal University of Uberlândia (UFU) for their essential institutional support. We also extend our gratitude to the members of our research group at UFV for the insightful discussions that refined this work, and to the anonymous reviewers of WVC'25 for their constructive feedback that improved the final manuscript.

Author Contributions

Vinícius Meireles Pereira Santos: Conceptualization, Methodology, Software, Investigation, Formal analysis, Writing – Original Draft, Writing – Review & Editing; Ana Claudia Patrocínio: Conceptualization, Methodology, Validation, Resources, Supervision, Project administration, Writing – Review; Pedro Moises de Sousa: Conceptualization, Methodology, Validation, Resources, Supervision, Project administration, Writing – Review & Editing.

References

- [1] INSTITUTO NACIONAL DE CÂNCER (INCA). *Atlas de Mortalidade por Câncer*. 2023. Available: <https://www.inca.gov.br/MortalidadeWeb/pages/Modelo03/consultar.xhtml>. Accessed on: August 02 2025.
- [2] ZOLIN, B. *Câncer renal: doença é silenciosa, mas pode ser prevenida*. Portal Drauzio Varella, 2023. Available: <https://drauziovarella.uol.com.br/oncologia/cancer-renal-doenca-e-silenciosa-mas-pode-ser-prevenida/>. Accessed on: August 02 2025.
- [3] INSTITUTO NACIONAL DE CÂNCER (INCA). *Estimativa 2023: Incidência de Câncer no Brasil*. 2023. Available: <https://www.inca.gov.br/publicacoes/livros/estimativa-2023-incidencia-de-cancer-no-brasil>. Accessed on: August 02 2025.
- [4] ALZU'BI, D. et al. Kidney tumor detection and classification based on deep learning approaches: a new dataset in CT scans. *Journal of Healthcare Engineering*, v. 2022, Art. ID 3861161, 2022.
- [5] RAJKUMAR, K. et al. Kidney cancer detection using deep learning models. In: 7TH INTERNATIONAL CONFERENCE ON TRENDS IN ELECTRONICS AND INFORMATICS (ICOEI). [S.l.]: IEEE, 2023. p. 1197–1203.
- [6] HADJIYSKI, N. Deep learning neural network based approach. In: INTERNATIONAL CONFERENCE ON E-HEALTH AND BIOENGINEERING (EHB). [S.l.]: IEEE, 2020. p. 1–4.
- [7] UHM, K. H. et al. Deep learning for end-to-end kidney cancer diagnosis on multi-phase abdominal computed tomography. *NPJ Precision Oncology*, v. 5, n. 1, p. 54, 2021.
- [8] SANTOS, V. M. P. et al. Renal cancer classification in tomographic images using convolutional neural networks to assist early diagnosis. *Revista de Informática Teórica e Aplicada*, v. 32, n. 1, p. 61–66, 2025.
- [9] SINGH, A. R. et al. An enhanced denoising system for mammogram images using deep transformer model with fusion of local and global features. *Scientific Reports*, v. 15, n. 1, p. 6562, 2025.
- [10] RASTOGI, D. et al. Deep learning-integrated MRI brain tumor analysis: feature extraction, segmentation, and survival prediction using replicator and volumetric networks. *Scientific Reports*, v. 15, n. 1, p. 1437, 2025.
- [11] LI, J. et al. Transforming medical imaging with transformers? A comparative review of key properties, current progresses, and future perspectives. *Medical Image Analysis*, v. 85, p. 102762, 2023.
- [12] OKOLO, G. I. et al. IEViT: an enhanced vision transformer architecture for chest x-ray image classification. *Computer Methods and Programs in Biomedicine*, v. 226, p. 107141, 2022.
- [13] AWASTHI, V. et al. ViT-HHO: optimized vision transformer for diabetic retinopathy detection using harris hawk optimization. *MethodsX*, v. 13, p. 103018, 2024.
- [14] SHEIKHY, A. et al. State of the art review of AI in renal imaging. *Abdominal Radiology*, 2025.
- [15] ISLAM, M. N. et al. Vision transformer and explainable transfer learning models for auto detection of kidney cyst, stone and tumor from CT-radiography. *Scientific Reports*, v. 12, n. 1, Art. ID 11352, 2022.
- [16] JAPKOWICZ, N.; STEPHEN, S. The class imbalance problem: a systematic study. *Intelligent Data Analysis*, v. 6, n. 5, p. 429–449, 2002.
- [17] HE, K. et al. Deep residual learning for image recognition. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR). Las Vegas: IEEE, 2016. p. 770–778.
- [18] KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS (NIPS). 2012. p. 1097–1105.
- [19] HOWARD, A. G. et al. Mobilenets: efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.
- [20] SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [21] TAN, M.; LE, Q. V. Efficientnet: rethinking model scaling for convolutional neural networks. In: 36TH INTERNATIONAL CONFERENCE ON MACHINE LEARNING (ICML). [S.l.]: ACM, 2019. p. 6105–6114.
- [22] DOSOVITSKIY, A. et al. An image is worth 16x16 words: transformers for image recognition at scale. In: INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS (ICLR). 2021.
- [23] SOKOLOVA, M.; LAPALME, G. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, v. 45, n. 4, p. 427–437, 2009.
- [24] SELVARAJU, R. R. et al. Grad-cam: visual explanations from deep networks via gradient-based localization. In: IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION (ICCV). [S.l.]: IEEE, 2017. p. 618–626.
- [25] ABDELTAWAB, H. et al. A pyramidal deep learning pipeline for kidney whole-slide histology images classification. *Scientific Reports*, v. 11, n. 1, p. 20189, 2021.
- [26] QADIR, A. M.; ABD, D. F. Kidney diseases classification using hybrid transfer-learning densenet201-based and random forest classifier. *KJAR*, v. 7, n. 2, p. 131–144, 2023.
- [27] CHANCHAL, A. K. et al. A novel dataset and efficient deep learning framework for automated grading of renal cell carcinoma from kidney histopathology images. *Scientific Reports*, v. 13, n. 1, p. 5728, 2023.

- [28] CHEFER, H.; GUR, S.; WOLF, L. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In: IEEE/CVF INTERNATIONAL CONFERENCE ON COMPUTER VISION (ICCV). Montreal, QC, Canada, 2021. p. 387–396.
- [29] ISHWARAN, H.; O'BRIEN, R. Commentary: The problem of class imbalance in biomedical data. *The Journal of Thoracic and Cardiovascular Surgery*, v. 161, n. 6, p. 1940-1941, 2021.
- [30] CROSS, J. L.; CHOMA, M. A.; ONOFREY, J. A. Bias in medical AI: Implications for clinical decision-making. *PLOS Digital Health*, v. 3, n. 11, p. e0000651, 2024.
- [31] GONZALEZ, R. C.; WOODS, R. E. *Digital Image Processing*. 3rd ed. Upper Saddle River, NJ: Prentice-Hall, Inc., 2006.
- [32] WANG, Z. et al. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, v. 13, n. 4, p. 600-612, 2004.