Integrating Traditional Machine Learning and Deep Learning Methods for Enhanced Wilms Tumor Detection

Anirudh Anandarao and Bhadresh Amarnath *

Posted Date: 31 December 2025

doi: 10.20944/preprints202512.2793.v1

Keywords: wilms tumor; kidney disease; machine learning; nephropathy

 doi:10.20944/preprints202512.2793.v1

Article

Integrating Traditional Machine Learning and Deep Learning Methods for Enhanced Wilms Tumor Detection

Anirudh Anandarao [1] and Bhadresh Amarnath [2,*]

[1] John Champe High School, United States of America
[2] Department of Biology, Elmira College, Elmira, NY 14901
* bamarnath29@elmira.edu

Abstract

Background/Objectives: Wilms tumor is the most common pediatric renal malignancy, and delayed or inaccurate diagnosis can significantly affect clinical outcomes. This study aimed to evaluate whether integrating traditional machine-learning and deep-learning models with computed tomography (CT) imaging could improve the accuracy of Wilms tumor detection. **Methods**: A large CT image dataset consisting of 18,205 kidney scans, including both normal and Wilms tumor cases, collected from publicly available medical sources. Images were preprocessed and resized to standardized dimensions before model training. Four supervised learning approaches: ResNet50, VGG16, XGBoost, and Random Forest, were developed and evaluated. The dataset was split into training (14,055 images) and independent testing (4,150 images) subsets. Model performance was assessed using accuracy, precision, recall, F1-score, and confusion matrix analysis. **Results**: Among the evaluated models, VGG16 demonstrated superior performance, achieving an accuracy of 99.98%, precision of 99.92%, recall of 100%, and an F1-score of 99.96%, indicating excellent sensitivity and overall classification reliability. The remaining models also performed robustly, with accuracies exceeding 94% and recall values above 90%. **Conclusions**: These findings suggest that deep-learning-based image classification, particularly using VGG16, can substantially enhance non-invasive detection of Wilms tumor from CT scans. The proposed approach has the potential to support clinical decision-making, reduce diagnostic delays, and improve early detection in pediatric oncology settings.

Keywords: wilms tumor; kidney disease; machine learning; nephropathy

1. Introduction

Wilms tumor (WT), or nephroblastoma, is a type of kidney cancer that primarily affects children, with the peak incidence around age 3 to 5. The tumor commonly affects a single kidney but in some cases can have a bilateral effect. Wilms tumor is characterized by a mixture of cell types, such as: blastemal, stromal, and epithelial components. The causes of WT are not fully known. However, it is known that genetic mutations and developmental abnormalities play a pivotal role. Wilms tumors usually arise sporadically, but certain gene mutations, like those of WT1 on chromosome 11p13, WTX, and other genes, are commonly involved in WT formation. Developmental remnants in the kidney, called nephrogenic rests, that do not mature normally, are thought to be precursors to tumors. Additionally, congenital syndromes that stifle kidney development or normal genetic regulation during embryogenesis, include syndromes in which gene regulation on chromosome 11 is disrupted, seem to also increase WT incidence. Several factors can increase the risk of developing Wilms tumor. Genetic syndromes and birth defects pose a major risk. Examples of defects include: WAGR syndrome (Wilms tumor, Aniridia, Genitourinary anomalies, and mental Retardation), Denys-Drash

 doi:10.20944/preprints202512.2793.v1

syndrome, Beckwith-Wiedemann syndrome, and other congenital disorders like hemihypertrophy or aniridia. Other than that, family history of WT can slightly increase risk, though most cases are not inherited. There are also demographic risk patterns: in the U.S., African-American children have a slightly higher incidence than Caucasian children, while Asian-Americans typically have a lower incidence. Gender plays a minor role in WT incidence but some data suggest higher incidence in female children [1].

Wilms tumor affects approximately 1 in every 10,000 children each year in North America. Overall, around 7,500 children under the age of 18 suffer from WT and about 190,000 children suffer from WT worldwide [2]. In the U.S., alone, there are approximately 650 new cases each year [1].

Current detection methods utilize clinical evaluation through medical history, physical exams, and laboratory tests of blood and urine, with a myriad of imaging studies such as ultrasound, CT, MRI, or chest imaging to check for spreading of the tumor. Yet, with these techniques, the misdiagnosis rate for Wilms tumor is approximately 62.5% and a delay of 17 days [3,4]. The gold standard for diagnosing WT is done via surgical biopsy of tumor tissue.

Previously, the use of machine learning (ML) algorithms to detect brain tumors, lung cancer, and appendicitis using CT scans have achieved accuracies above 90%. Afnaan et al. ResNet50 to achieve an accuracy above 99% to detect brain tumors from CT scans [5]. Further, Kucukakcali et al, leveraged the XGBoost model to diagnose appendicitis with an accuracy of 97.3 [6]. Nair et al. successfully used the Random Forest Classifier to detect lung cancer, from CT scans, with an accuracy of 99.6% [7]. Further, Zargar et al. have used the VGG16 model to detect lung cancer at an accuracy of 91% [8]. Through this study, we hoped to integrate traditional machine-learning and deep-learning models with computed tomography (CT) imaging to improve the accuracy of Wilms tumor detection.

 doi:10.20944/preprints202512.2793.v1

2. Materials and Methods

2.1. Dataset

This study utilized a dataset of 18,205 kidney CT scans from publicly available health data websites via web scraping, and ensured no data leakage. The dataset is composed of normal kidney CT scans and Wilms Tumor CT scans. The preprocessing included resizing them to 299 × 299-pixel dimensions to fit the model input shape while keeping important information.

2.2. Model Development

Four supervised machine-learning algorithms were developed and evaluated for WT risk prediction: ResNet50, VGG16, XGBoost, Random Forest.

ResNet50 is a deep convolutional neural network (CNN) that works by using residual layers. It allows information from earlier layers to skip ahead to later layers in order to reduce information loss and improve pattern learning. VGG16 is a CNN that works by looking at the same image repeatedly, with each layer detect basic lines and edges, then combining in later layers to detect larger shapes in order to build a complete understanding of the image. XGBoost is a gradient boosting algorithm that creates multiple decision trees, with each tree meant to correct the mistakes made by previous trees in order to slowly build prediction accuracy. Random Forest is an ensemble learning method that trains multiple decision trees at the same time, with each tree on slightly different data. The final prediction is decided by the outputs of all the trees in order to get a common answer.

Approximately 80% of the data from each dataset was used to train the model, and 20% to test the model. The final dataset included 14,055 images to train the models and 4,150 to test the models. Each model was evaluated through accuracy, precision, recall, and F1-score.

All analyses were conducted in Python 3.10, utilizing pandas (v2.1.0) for data manipulation, NumPy for numerical operations, and scikit-learn (v1.3.2) for preprocessing, classical machine-learning models, and evaluation metrics. Deep learning models (ResNet50 and VGG16) were implemented using TensorFlow/Keras (v2.x), while XGBoost models were trained using the XGBoost library.

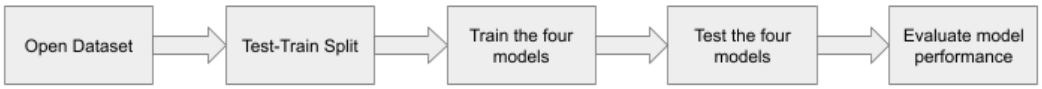

Figure 1. Schematic Diagram of Methodology.

3. Results

3.1. Overview

The results from testing the various models on the test data were considered when selecting the best model. The study found that the VGG16 model was the best model. Among the four models tested, the VGG16 model performed the best with an accuracy of 99.98%, a precision of 99.92%, a perfect recall of 100%, and a F1 Score of 91.01%.

3.2. Statistical Evaluation

The performance of four supervised machine-learning and deep-learning models: ResNet50, VGG16, XGBoost, and Random Forest were evaluated on an independent test set consisting of 4,150 kidney CT scan images. Model effectiveness was assessed using standard classification metrics, including accuracy, precision, recall, and F1-score. Accuracy measures the proportion of total predictions that are correct. Precision quantifies how many of the instances predicted as positive are truly positive. Recall measures the proportion of actual positive instances that are correctly identified. The F1-score is the mean of precision and recall. Confusion matrices were also used to determine model performances, specificities, and sensitivities. reword A confusion matrix is a tabular summary of a classification model's predictions that compares predicted labels with actual labels, showing the

 doi:10.20944/preprints202512.2793.v1

counts of true positives, true negatives, false positives, and false negatives to evaluate model performance.

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

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

$$F1\ Score = \frac{TP + TN}{TP + \frac{1}{2}(FP + FN)}$$

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

3.3. Data

Among the four machine learning algorithms evaluated, the VGG16 model demonstrated the strongest overall performance, achieving the highest mean accuracy (0.9998), precision (0.9992), recall (1.0000), and F1 Score (0.9996). The ResNet50 model also performed well, though it was limited by a substantially lower recall relative to its precision. All four models performed fairly well, with accuracies consistently above 0.94 and recall scores above 0.90.

Table 1. Statistical evaluation of the four models tested: ResNet50, VGG16, XGBoost, and Random Forest.

Model Name	ResNet50	VGG16	XGBoost	Random Forest
Accuracy	0.9581	0.9998	0.9424	0.9477
Precision	0.9282	0.9992	0.8352	0.8480
Recall	0.9282	1.0000	1.0000	1.0000
F1 Score	0.9282	0.9996	0.9101	0.9178

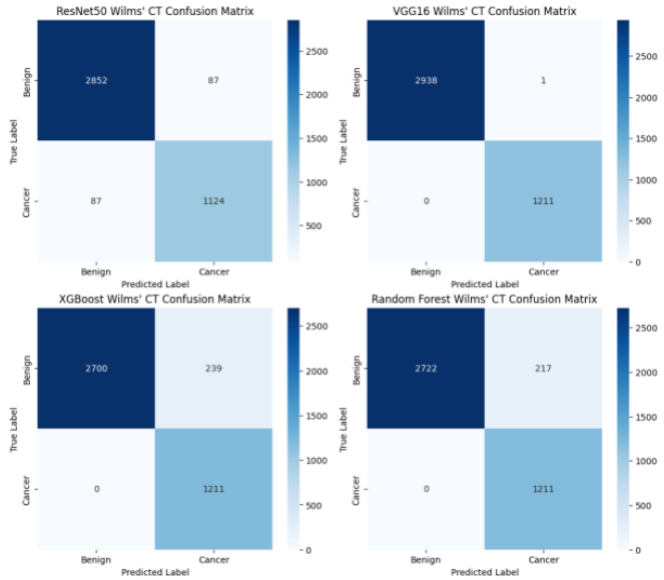

Figure 2. Confusion matrices of the four models tested: ResNet50, VGG16, XGBoost, and Random Forest.

4. Discussion

4.1. Interpretation of Results

Among the four models evaluated, the VGG16 model demonstrated the highest accuracy along with a perfect recall (1.0000). This indicates that the VGG16 model was able to classify all patients with Wilms tumor in the dataset. While other metrics such as accuracy, precision, and F1 score are important, in the context of diagnosing tumors, the ability of a model to have a high sensitivity and specificity is paramount in dictating the diagnosis of Wilms Tumor.

 doi:10.20944/preprints202512.2793.v1

4.2. Comparison of Results with Previous Studies

In the meta-analysis conducted by Neves et al., the included studies primarily relied on radiomics-based and conventional machine learning models, most commonly support vector machines (SVMs), random forest classifiers, and logistic regression, applied to handcrafted CT features for Wilms Tumor detection. Across these models, pooled recall was reported as 63.9%, while accuracy, precision, and F1-score were not consistently reported across studies [9]. Ma et al. employed a radiomics-based SVM model, achieving an accuracy of 0.79 and recall of 0.87 for Wilms Tumor detection from CT scans. Precision and F1-score were not explicitly reported [10]. In contrast, our study evaluated both deep learning and ensemble-based models, including ResNet50, VGG16, XGBoost, and Random Forest, and demonstrated substantially stronger performance across all four evaluated metrics. The VGG16 model achieved a mean accuracy of 0.9998, precision of 0.9992, recall of 1.0000, and F1-score of 0.9996, with all models maintaining accuracies above 0.94 and recall values above 0.90. These findings indicate a clear improvement in recall and balanced classification performance relative to prior radiomics-based approaches, likely driven by the use of deep convolutional neural networks and a substantially larger dataset, enabling more robust feature learning and improved generalizability.

Recent studies provide evidence of computed tomography (CT) as a valuable non-invasive procedure in the evaluation of Wilms tumor, particularly for initial diagnosis and disease characterization. CT imaging enables clinicians to visualize renal masses, determine tumor location, size, and local extent, and assess relationships to surrounding biological structures, which are critical for clinical decision-making and patient treatment [11]. In parallel, emerging work highlights the growing use of artificial intelligence with medical imaging as a noninvasive diagnostic support tool, illustrating its capability in assisting clinicians in tumor identification and classification using acquired CT data [12]. Together, these findings support CT-based, AI-assisted approaches as a means of improving diagnostic accuracy while reducing reliance on invasive procedures, an especially important consideration in pediatric oncology.

4.3. Implications of findings

Given the documented challenges of delayed and inaccurate diagnosis in Wilms tumor, with misdiagnosis rate for Wilms tumor being approximately 62.5% and delay in diagnosis being 17 days [3,4], the use of automated CT-based machine learning models provides a clear advantage by improving both diagnostic speed and reliability. Automated analysis of CT scans enables rapid, standardized tumor classification, reducing the time required for clinical interpretation and supporting quicker diagnostic decision-making. By increasing classification accuracy and recall, this approach can significantly reduce misdiagnosis rates, ensuring a greater proportion of patients are correctly diagnosed during an initial clinical examination. Earlier and more accurate diagnosis can streamline clinical workflows, minimize unnecessary follow-up testing, and allow clinicians to initiate appropriate treatment sooner. Collectively, these benefits suggest that high-performing, CT-based machine learning models can meaningfully enhance diagnostic efficiency while addressing key limitations of current diagnostic pathways.

4.4. Future Directions

Future research may explore ensemble learning approaches that combine the strengths of multiple machine-learning and deep-learning models to further enhance predictive performance and statistical reliability. By integrating CNN and tree-based algorithms, ensemble methods can reduce model bias and variance, leading to improved generalization on independent datasets. This was demonstrated in the study conducted by Alam et. al. The four models tested had accuracies of: 0.9766, 0.8763, 0.8617, and 0.8556. Following this, the team used ensemble methods to combine the two models in various ways and the best-performing combination of models resulted in an accuracy of 0.9998, which is significantly better than the individual models' performances [13]. Similar techniques could be applied to the models created in our study to determine if combinations of these models result in a more effective classification.

 doi:10.20944/preprints202512.2793.v1

5. Conclusions

The use of machine learning models to predict the tumorigenesis of Wilms Tumor via medical imaging makes diagnosis more non-invasive, rapid, and cost-effective for the families of patients exhibiting symptoms.

The VGG16 model is more likely to predict the tumorigenesis of Wilms Tumor through CT scans than are the other three ML algorithms. By leveraging Machine Learning algorithms, we hope to contribute to early diagnostic techniques in pediatric oncology. Future research should continue to validate such models in diverse populations and clinical environments to enhance their impact.

Author Contributions: Conceptualization, A.A. and B.A.; methodology, B.A.; software, B.A.; validation, A.A and B.A.; formal analysis, A.A. and B.A.; investigation, A.A. and B.A.; resources, B.A.; data curation, B.A.; writing—original draft preparation, A.A. and B.A.; writing—review and editing, B.A.; visualization, A.A.; supervision, B.A.; project administration, A.A. and B.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used in this study, Amarnath, B. Wilms tumor, is publicly available through Kaggle. The dataset includes CT results of patients, and it was utilized for the purpose of diagnosing Wilms Tumor using machine learning models. The dataset can be reached at "https://www.kaggle.com/datasets/bhadresha/wilms-tumor/".

Acknowledgments: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.:

Abbreviations

The following abbreviations are used in this manuscript:

WT	wilms tumor
CT	computed tomography
CNN	convolutional neural network

References

1. Leslie, S.W.; Sajjad, H.; Murphy, P.B. Wilms tumor (nephroblastoma). StatPearls Publishing 2023. Available online: https://www.ncbi.nlm.nih.gov/books/NBK442004/ (accessed on 29 May 2025).
2. Wilms Foundation. Kidney cancer awareness | Wilms Foundation. 2016. Available online: https://www.wilmsfoundation.org/wilmsbythenumbers (accessed on 29 May 2025).
3. Yimenu, G.; Wassie, M.; Wodajo, S.; Giza, M.; Ayalew, M.; Sewale, Y.; Feleke, Z.; Dessie, M.T. Delay in diagnosis and associated factors among children with cancer admitted at pediatric oncology ward, University of Gondar comprehensive specialized hospital, Ethiopia: A retrospective cross-sectional study. BMC Cancer 2023, 23, 10873. https://doi.org/10.1186/s12885-023-10873-8.
4. Garcés-Visier, C.; Maruszewski, P.; Luis-Huertas, A.L.; Borrego-Jimenez, P.; Azorín, D.; Martín-Vega, A.; Espinoza-Vega, M.; Herrero-Velasco, B.; Alonso-Calderón, J.L. Non-Wilms renal tumors: Twenty years experience in a referral center. J. Pediatr. Surg. Open 2024, 8, 100151. https://doi.org/10.1016/j.yjpso.2024.100151.
5. Afnaan, K.; Arunbalaji, C.G.; Singh, T.; Kumar, R.; Naik, G.R. Boosting brain tumor detection with an optimized ResNet and explainability via Grad-CAM and LIME. Brain Informatics 2025, 12, 33. https://doi.org/10.1186/s40708-025-00279-6.
6. Kucukakcali, Z.; Akbulut, S.; Colak, C. Evaluating ensemble-based machine learning models for diagnosing pediatric acute appendicitis: Insights from a retrospective observational study. J. Clin. Med. 2025, 14, 4264. https://doi.org/10.3390/jcm14124264.

 doi:10.20944/preprints202512.2793.v1

7. Nair, S.S.; Meena Devi, V.N.; Bhasi, S. Enhanced lung cancer detection: Integrating improved random walker segmentation with artificial neural network and random forest classifier. Heliyon 2024, 10, e29032. https://doi.org/10.1016/j.heliyon.2024.e29032.
8. Zargar, H.H.; Zargar, S.H.; Mehri, R.; Tajidini, F. Using VGG16 algorithms for classification of lung cancer in CT scans. arXiv 2023, arXiv:2305.18367. Available online: https://arxiv.org/abs/2305.18367 (accessed on 29 May 2025).
9. Neves, H.; Lima, R.V.; Filipe, J.; Ohannesian, V.A.; Eulálio, E.C.; Vianna, P.; Macêdo, L.; Feitosa, I.D.; Bezerra, G. Artificial intelligence computed tomography models for the discrimination of Wilms versus non-Wilms tumors: Systematic review and meta-analysis. Braz. J. Nephrol. 2025, 48. https://doi.org/10.1590/2175-8239-jbn-2025-0010en.
10. Ma, X.-H.; Shu, L.; Jia, X.; Zhou, H.-C.; Liu, T.-T.; Liang, J.-W.; Ding, Y.; He, M.; Shu, Q. Machine learning-based CT radiomics method for identifying the stage of Wilms tumor in children. Front. Pediatr. 2022, 10, 873035. https://doi.org/10.3389/fped.2022.873035.
11. van der Kamp, A.; de Bel, T.; van Alst, L.; Rutgers, J.; van den Heuvel-Eibrink, M.M.; Mavinkurve-Groothuis, A.M.C.; van der Laak, J.; de Krijger, R.R. Automated deep learning-based classification of Wilms tumor histopathology. Cancers 2023, 15, 2656. https://doi.org/10.3390/cancers15092656.
12. Huang, J.; Li, Y.; Pan, X.; Wei, J.; Xu, Q.; Zheng, Y.; Chen, P.; Chen, J. Construction of a Wilms tumor risk model based on machine learning and identification of cuproptosis-related clusters. BMC Med. Inform. Decis. Mak. 2024, 24, 325. https://doi.org/10.1186/s12911-024-02716-8.
13. Nur-A-Alam, M.; Nasir, M.K.; Ahsan, M.; Based, M.A.; Haider, J.; Kowalski, M. Ensemble classification of integrated CT scan datasets in detecting COVID-19 using feature fusion from contourlet transform and CNN. Sci. Rep. 2023, 13, 20063. https://doi.org/10.1038/s41598-023-47183-9.