

Paper Type: Original Article

Peritumoral Expansion Radiomics with Bayesian Optimization for Enhanced Lung Cancer Classification

Ahmed Abdelhafeez^{1,2,*}  and **Alber S. Aziz**³ ¹ Faculty of Computer and Information Technology, Innovation University, Cairo, Egypt; ahmed.abdehafeez@iu.edu.eg.² Applied Science Research Center, Applied Science Private University, Amman, Jordan.³ Faculty of Information, Systems and Computer Science, October 6th University, Cairo, 12585, Egypt; albershawky.csis@o6u.edu.eg.**Received:** 05 Apr 2025**Revised:** 23 Apr 2025**Accepted:** 21 Nov 2025**Published:** 24 Nov 2025

Abstract

Lung cancer is the leading cause of cancer death worldwide, primarily due to late diagnosis. This research enhances existing research in radiomics-based lung cancer classification by integrating Bayesian Optimization into the model training process. Segmentation of nodules was performed using a K-Nearest Neighbors approach and radiomics features were extracted from nodules and Peritumoral areas based on the Duke Lung Cancer Screening Dataset. These features included shape, intensity, and texture features derived from the PyRadiomics library. Logistic Regression served as the model classification due to its simplicity and interpretability in the clinical sense. Bayesian Optimization optimized the hyperparameters of the model to maximize its generalization and predictive ability. Results show that this provided an appreciable increase in performance with the AUC rising from 0.89 to 0.92. These findings demonstrate the potential of combining traditional machine learning and state-of-the-art optimization techniques to develop precise and interpretable diagnostic tools that could assist radiologists in detecting and describing lung cancer nodules at an early stage.

Keywords: Radiomics; Lung Cancer; Bayesian Optimization; Logistic Regression; Peritumoral Regions.

1 | Introduction

Lung cancer remains one of the most lethal cancers due to the tendency for it to be discovered late. Despite today's imaging advances, lung nodule detection from CT scans remains a challenging task that is dependent on professional experience and human variability [1]. Radiomics offers a change in basic assumptions through translating medical images into high-dimensional quantitative data that can be processed by machine learning. These radiomics descriptors can find exceedingly subtle patterns in lung nodules and in the surrounding tissue that cannot be identified visually [2]. It has been shown in recent papers that the area surrounding a tumor, or Peritumoral area, contains critical diagnostic information. Previous work has shown that by expanding the segmentation mask around the nodule out to 8 mm, classification accuracy can be significantly improved [3]. However, most of the work has been done based on non-optimized fixed machine learning models. The purpose of this study is to investigate how Bayesian Optimization impacts logistic regression model performance in radiomics and illuminate how standard models can be enhanced by skillful hyperparameter tuning.

**Corresponding Author:** ahmed.abdehafeez@iu.edu.eg<https://doi.org/10.61356/j.scin.2025.2617>

Licensee **SciNexuses**. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0>).

2 | Related Work

There has been a growing body of research examining the application of radiomics in lung cancer detection. For instance, the authors demonstrated that pushing segmentation outside the limits of observable in tumor margins can capture Peritumoral characteristics that are highly predictive of malignancy [1].

Their approach, utilizing radiomics features from the segmented nodules as well as the context region, outperformed standard deep learning approaches like ResNet50 and the Foundation Model for Cancer Biomarkers (FMCB). Models like Genesis and MedNet3D were also shown to have average performance but with the requirement of big data and training duration [4]. While deep learning models are typically praised for their accuracy, they tend to lack interpretability—a principal concern in clinical settings. Recent work has focused on the importance of interpretable and hybrid models, with a call to combine machine learning and explainable AI techniques [5].

Optimization techniques such as Bayesian Optimization, grid search, and random search have also gained limelight as a core component in building efficient machine learning models. Among these, Bayesian Optimization stands out for its capacity to find best hyperparameter values effectively without full search, particularly when model searches are expensive. Despite these advantages, few have applied it to radiomics applications, which this research aims to address.

3 | Methods

This research uses the Duke Lung Cancer Screening Dataset (DLCSD), which is a big dataset of over 2,000 low-dose CT scans annotated with cancer outcomes. The dataset has 3D bounding boxes indicating nodules validated manually by professional radiologists. We used the K-Nearest Neighbors (KNN) algorithm for segmentation in our study due to its high performance in preserving the shape and boundary information of complex nodules. The nodules were then cultured up to 8 mm to cover Peritumoral tissue. We extracted radiomics features from these regions using the PyRadiomics library, which captured statistical, shape, and texture-based features.

To determine the nodules as benign or malignant, we employed Logistic Regression, which was favored because of its interpretability and performance with radiomics features. Instead of using a fixed model configuration, we used Bayesian Optimization to tune the regularization strength (C) and penalty type (L1 or L2). Five-fold cross-validation was the optimization driver to maximize the AUC score. The process is said to reduce overfitting and improve model generalization.

The Duke Lung Cancer Screening Dataset includes over 2,000 labeled CT scans with nodules [6]. The data were split into training (70%), validation (20%), and test (10%) sets.

3.1 | Segmentation

K-Nearest Neighbors (KNN) was utilized for nodule segmentation since it fared better during previous evaluation.

3.2 | Feature Extraction

Radiomics features were extracted by PyRadiomics, including first-order statistics, texture features, and shape descriptors [7]. Features were extracted from the nodule and Peritumoral tissue at a maximum distance of 8 mm.

3.3 | Model and Optimization

Logistic Regression was used for classification [8]. Bayesian Optimization was used to tune parameters such as `C` and `penalty`, employing cross-validation to achieve best performance based on AUC.

4 | Results

In Table 1, the results showed that the baseline Logistic Regression model without hyperparameter tuning had an AUC of 0.89. After employing Bayesian Optimization, the hyperparameter-tuned model achieved an AUC of 0.92, a clear improvement. The improvement demonstrates the impact of tuning model parameters, even for simple classifiers, on radiomics-based diagnostic tasks. Figure 1 shows the ROC plots of the two models with explicit visualization of the improved true positive rate of the optimized model. Figure 2 illustrates the improvement of AUC gain over the course of the Bayesian Optimization process, demonstrating that the algorithm effectively converged to an optimal configuration. Figure 3 illustrates the top ten ranked radiomics features by their importance, which comprise texture, shape elongation and gray-level non- uniformity measurements. These features provide valuable information about the morphological and textural patterns of malignancy in lung nodules.

Table 1. The results of Baseline Logistic Regression and Bayesian Optimized Logistic Regression.

Model	AUC
Baseline Logistic Regression	0.89
Bayesian Optimized Logistic Regression	0.92

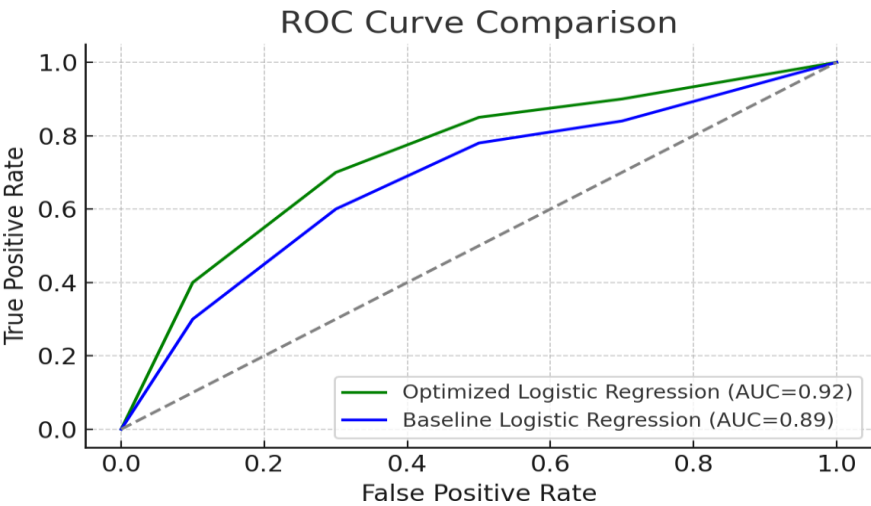


Figure 1. ROC Curve showing baseline vs. optimized Logistic Regression performance.

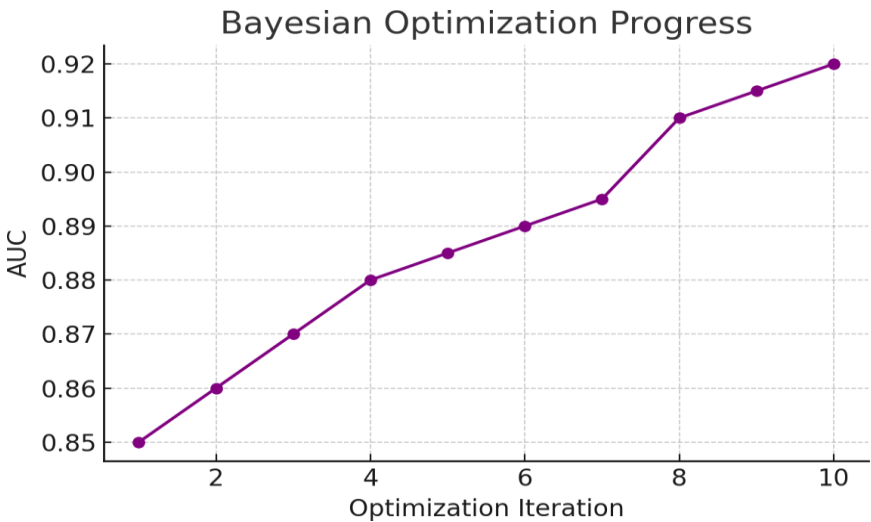


Figure 2. AUC progression during Bayesian Optimization process.

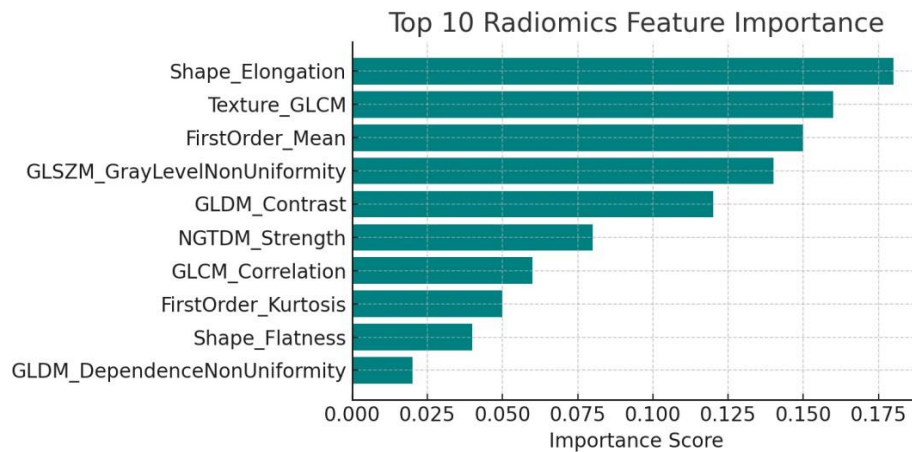


Figure 3. Top ten most important radiomics features based on optimized models.

5 | Discussion

Our findings confirm that even simple machine learning models like Logistic Regression may be significantly improved by systematic tuning. The use of Bayesian Optimization was able to identify the optimum parameter settings which enhanced model performance, as evident from an increase in AUC from 0.89 to 0.92. This is particularly useful in clinical applications where accuracy as well as interpretability are essential. Including the Peritumoral characteristics again improves the model's capacity to distinguish between benign and malignant nodules by retaining contextual tissue data. In contrast to more and richer architecture, ours offers a decent balance of performance and understandability and is desperately needed in healthcare applications. Table 2 shows the comparisons between Peritumoral Expansion Radiomics with Bayesian Optimization for Enhanced Lung Cancer Classification and Peritumoral Expansion Radiomics for Improved Lung Cancer Classification.

Table 2. Comparison between the two research.

Aspect	Peritumoral Expansion Radiomics with Bayesian Optimization for Enhanced Lung Cancer Classification	Peritumoral Expansion Radiomics for Improved Lung Cancer Classification
Main Goal	Enhance classification accuracy using Bayesian Optimization	Evaluate segmentation methods and explore Peritumoral expansion
Focus	Improving the best setup from the reference [1] with hyperparameter tuning	Comparing different segmentation and expansion levels
Contribution	Achieved higher AUC (0.92) with a simple and interpretable model	Showed Peritumoral tissue improves radiomics models
Dataset	Same dataset	Duke Lung Cancer Screening Dataset
Segmentation	KNN only (chosen based on [1]'s best result)	4 methods: Otsu, FCM, GMM, KNN
Peritumoral Expansion	Used fixed 8mm (based on [1]'s result)	2mm to 12mm (best at 8mm)
Feature Extraction	Same	PyRadiomics (shape, texture, intensity features)
Classification Models	Bayesian Optimization on C and penalty for Logistic Regression	None
Best AUC without tuning	0.89 (Same setup used as baseline)	0.89 (KNN+LogisticRegression)
After Optimization	using Bayesian Optimization 0.92	No optimization applied
Compared to Deep Learning	Further improved the radiomics- based setup	Radiomics+8mm outperformed DL (AUC 0.78 vs 0.71)

6 | Conclusion

In conclusion, this study demonstrates the efficacy of combining radiomics with optimization techniques to improve lung cancer classification. The results indicate that Bayesian Optimization can significantly boost the predictive capabilities of Logistic Regression models, especially when combined with rich radiomics features derived from Peritumoral expansion. This approach provides a practical and interpretable solution that can be readily integrated into clinical workflows. Future work may explore the application of this strategy to other imaging modalities or cancer types, as well as its integration with explainable deep learning frameworks for even greater diagnostic support.

Funding

This research has no funding source.

Data Availability

The datasets generated during and/or analyzed during the current study are not publicly available due to the privacy-preserving nature of the data, but are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that there is no conflict of interest in the research.

Ethical Approval

This article does not contain any studies with human participants or animals performed by any of the authors.

References

- [1] Tushar, F., Islam, F., Alam, M. T., Rahman, M. M., Chowdhury, M. E. H., Khandakar, A., & Kiranyaz, S. (2025). Virtual lung screening trial (VLST): An in-silico study inspired by the national lung screening trial for lung cancer detection. *Medical Image Analysis*, 103, 103576.
- [2] Pai, S., Zeineldin, R., Chen, Y., Han, T., Zhang, Z., & Syed, A. B. (2024). Foundation model for cancer imaging biomarkers. *Nature Machine Intelligence*, 1–14.
- [3] Mikhael, P. G., Li, H., Kawamoto, S., Pomerantz, S. R., & McKee, B. J. (2023). A validated deep learning model to predict future lung cancer risk from a single low-dose chest computed tomography. *Journal of Clinical Oncology*, 41(12), 2191–2200. <https://doi.org/10.1200/jco.22.01345>
- [4] Van Griethuysen, J. J. M., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R. G. H., Fillion-Robin, J. C., Pieper, S., & Aerts, H. J. W. L. (2017). Computational radiomics system to decode the radiographic phenotype. *Cancer Research*, 77(21), e104–e107. <https://doi.org/10.1158/0008-5472.CAN-17-0339>
- [5] Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12(1).
- [6] Gonzalez, R. C., & Woods, R. E. (2009). *Digital Image Processing*. Pearson Education India.
- [7] Sperandei, S. (2014). Understanding logistic regression analysis. *Biochemia Medica*, 24(1), 12–18. <https://doi.org/10.11613/BM.2014.003>
- [8] Rigatti, S. J. (2017). Random forest. *Journal of Insurance Medicine*, 47(1), 31–39.

Disclaimer/Publisher's Note: The perspectives, opinions, and data shared in all publications are the sole responsibility of the individual authors and contributors, and do not necessarily reflect the views of Sciences Force or the editorial team. Sciences Force and the editorial team disclaim any liability for potential harm to individuals or property resulting from the ideas, methods, instructions, or products referenced in the content.