A Comprehensive Survey of Radiomics and Machine Learning in Medical Image Processing

Dr. Ashutosh kumar singh
Thesis Concepts
ashutosh@thesisconcepts.com

Abstract: Radiomics, defined as the high-throughput extraction of quantitative features from medical images, represents a paradigm shift in diagnostic and prognostic medicine by enabling the discovery of imaging biomarkers beyond human perception. This review synthesizes existing literature on the integration of advanced image processing techniques with machine learning (ML) and deep learning (DL) methods in radiomics. It outlines the standard workflow—including image acquisition, segmentation, feature extraction, feature selection, and model development—while examining both its potential and technical challenges. Applications across pulmonary disease analysis, oncology, and cardiac risk prediction highlight the state-of-the-art. Key issues such as feature reproducibility, model interpretability, data heterogeneity, and the need for robust validation are critically discussed. The paper concludes by identifying future directions, including standardized protocols, explainable AI, multimodal data fusion, and the ethical deployment of radiomics in clinical practice.

Keywords: Radiomics, Machine Learning, Deep Learning, Image Processing, Feature Extraction, Predictive Modeling, Quantitative Imaging, Medical Imaging Analytics.

1. Introduction

The medical imaging and artificial intelligence (AI) combination has led to the emergence of radiomics, a subdiscipline that is dedicated to the automated extraction of a huge number of quantitative features of radiographic images. These characteristics that extract texture, shape, intensity and wavelet patterns give a complete radiomic signature that can help identify intra-tumoral heterogeneity, disease subtypes and prognostic data not visible to the naked eye (Amudala Puchakayala, 2023). The premise of the radiomics is that these data-driven signatures have the potential to be used as non-invasive biomarkers in the diagnosis, prediction of treatment response and prognosis of a broad range of ailments, such as cancer, chronic obstructive lung disease (COPD) and cardiovascular ailments.

Radiomics pipeline is a highly interdisciplinary process since it unites both the techniques of advanced image processing and the advanced technologies of ML/DL. The present review proposed is expected to give a synthesized account of this pipeline based on and through the incorporation of the findings of a curated collection of the recent studies. We are going to consider the utilization of traditional machine learning systems, such as Support Vector Machines (SVM) and ensemble models, and the new deep learning frameworks. In addition, we will place radiomics in the framework of general trends in AI, including the issues of data quality, model transparency and ethical use as discussed in related literature on supervised

learning, multimodal learning, and responsible AI (Ghule et al., 2024; Sardesai et al., 2025; Puchakayala, 2022).

2. The Radiomics Pipeline: From Images to Insights

2.1. Image Acquisition and Preprocessing

Any radiomics study is based on non-irregular and good image data. The acquisition parameters (e.g. scanner type, slice thickness, reconstruction kernel) may cause very high amounts of noise and bias on the extracted features, which poses a threat to reproducibility (Puchakayala, 2024; Ghori, 2019). Research states how vital it is to have standardized imaging protocols (Ghori, 2021; Puchakayala, 2022). As an example, the studies of COPD detection showed that the model can perform with great results with both low-dose and standard-dose CT scans, but the most predictive types of features (e.g., parenchymal texture vs. lung shape) differ across doses and suggest the effects of acquisition settings on the radiomic feature space (Saha et al., 2025; Amudala Puchakayala et al., 2024). All these variabilities should be countered by preprocessing, including resampling of images, normalization of images (e.g., Z-score, histogram matching) and reduction of noise, which will make features comparable across datasets (Puchakayala, 2022; Shalini et al., 2023).

2.2. Tumor or Region of Interest (ROI) Segmentation

Proper definition of the ROI, be it a tumor, be it an organ, be it a particular pathological region is the priority as the values of the features are computed directly based on these volumes (Puchakayala, 2024; Sardesai et al., 2025). Segmentation may be carried out in a manual, semi-automatic or in an automatic manner. Manual segmentation is the best, the gold standard though it is time-consuming and has an inter-observer variation (Ghori, 2018). The use of tools that use deep learning to perform automatic segmentation based on the architectures such as CNNs and U-Nets has been motivated by the rising trend and has been highly successful in computer vision application areas including content-based image retrieval and hand gesture recognition (Marathe et al., 2022; Ghori, 2019). These models are capable of enhancing uniformity and size in radiomics research-intensive studies (Shalini et al., 2023).

2.3. Feature Extraction

The step entails the calculation of a vast number of quantitative characteristics on the segmented ROI (Puchakayala, 2024; Ghori, 2020). The features are normally classified into:

- ***First-Order Statistics:*** Characterize the frequency of the voxel intensities (e.g. mean, median, kurtosis, entropy) (Ghori, 2018).
- ***Shape-Based Features:*** Characterize three-dimensional geometry of the ROI (e.g., volume, sphericity, surface area) (Puchakayala, 2024).
- ***Second-Order/Texture Features:*** Characterize the spatial relationships between voxels, based on matrices such as Gray-Level Co-occurrence Matrix (GLCM) and Gray-Level Run-Length Matrix (GLRLM), and Gray-Level Size Zone Matrix (GLSZM) (Sardesai et al., 2025).
- ***Higher-Order Features:*** These are functions of filter transforms (or Laplacian of Gaussian) which describe patterns of different scale and orientation (Ghori, 2019).

More complex advanced signal processing methods, e.g., Quantum Wavelet Transform - QWT, Empirical Mode Decomposition - EMD are being applied in radiomics to do more elaborate

multi-resolution texture analysis and artifact removal (Sardesai & Gedam, 2025; Sheela & Jadagerimath, 2022).

2.4. Feature Selection and Dimensionality Reduction

The first feature set can effortlessly be in the thousands, and this causes dimensionality curse and significant chances of overfitting the model (Ghori, 2019; Puchakayala, 2022). Mainly, it is important to have a strong feature selection. Methods include:

- ***Univariate Statistics:*** Alternative features (e.g., t-tests, ANOVA) that have very high differences among the groups (Ghori, 2018).
- ***Multivariate Methods:*** Applying such techniques as Least Absolut Shrinkage and Selection Operator (LASSO) regression or tree-based importance scores to get a parsimonious, non-redundant set of features (Ghule et al., 2024; Puchakayala, 2024).
- ***Principal Component Analysis (PCA):*** A transformation to a lower dimensional space of components that have no correlation (Ghori, 2020).

The objective will be to establish a stable disease-relevant radiomic signature, which is predictable in other cohorts (Sheela & Shalini, 2024).

3. Machine Learning and Deep Learning in Radiomics Modeling

3.1. Traditional Machine Learning Models

After a fine set of features is acquired, the predictive or prognostic model is developed by training in an ML classifier (Ghori, 2018; Ghule et al., 2024). Algorithms that have been found in the application of educational and medical prediction are common, and they include:

- ***Support Vector Machines (SVM):*** It works well with high-dimensional data and can be combined with other optimization methods to learn hyper parameters such as Bayesian Optimization (BO-SVM) as seen in high accuracy in signal classification tasks (Sardesai & Gedam, 2025; Ghori, 2019).
- ***Ensemble Methods (Random Forests, and Gradience Boosting):*** Radiomics in the literature find Ensemble Methods more attractive: That is, they are resistant to noise and can warrant non-linear relationships as well as they typically provide perceptivities on features in priori. In COPD dynamic, the identification of changes in air density and the identification of fast progressors were effectively conducted with the help of gradient boosting models (i.e., CatBoost) (Saha et al., 2025). The use of similar methods is promoted in predicting soil quality and detecting financial anomalies, and they also have a broad range (Saha et al., 2025; Kumar et al., 2023; Ghori, 2018; Puchakayala, 2024).
- ***Logistic Regression:*** Due to its tendency to be frequently used as a benchmark, it is most frequently used when the feature coefficients are desirable to be interpretable (Ghule et al., 2024).

These strengths and challenges of these models, which are directly relevant to the radiomics context, are echoed in the systematic review of supervised learning by Ghule et al. (2024) as an aid in performance prediction.

3.2. Deep Learning Approaches

DL proposes end-to-end substitute, which learns to extract feature representations in hierarchy by using image patches or complete images, and does not require a step in which features are extracted manually (Ghori, 2019; Puchakayala, 2024).

- ***Convolutional Neural Networks (CNNs):*** The STREAM of imaging DL. Model CNNs (ex: ResNet, Inception) may be either applied as feature extractors or can be brought to selected radiomics tasks, similar to those utilized in content-based image retrieval (Marathe et al., 2022; Shalini et al., 2023).
- ***Hybrid Models:*** Radiomic features with deep learning features- Use both methods together to take advantage of the other ones. Moreover, generative models such as Generative Adversarial Networks (GANs) are also under investigation of data augmentation to boost small sample sizes or even highly complex missing data imputation, which is common in healthcare data (Bansal et al., 2025; Sardesai et al., 2025; Puchakayala, 2024).

4. Key Applications and Empirical Findings

Some of the interesting applications are brought out in the literature:

- ***Pulmonary Disease:*** COPD has exhibited a great potential with Radiomics. Models that used radiomics characteristics basing on inspiratory CT scan measurements were found to have a better diagnostic quality (AUC of approximately 0.90) than standard measures such as the percentage of emphysema. What is more, these features are likely to predict the emphysema progression in the future and detect so-called rapid progressors with the high level of accuracy (AUC =0.74) and possible early intervention (Amudala Puchakayala et al., 2024; Saha et al., 2025; Ghori, 2021).
- ***Oncology:*** This is not the main point of the given citations, but a huge amount of literature utilizes radiomics to classify tumors, grade them, and predict their response to any given treatment (e.g., chemotherapy, radiotherapy) and survival, regardless of the type of cancer (Puchakayala, 2024; Sheela & Jadagerimath, 2022).
- ***Cardiovascular Risk:*** Predicting risks with radiomics and clinical information, similar to how the conceptual framework of predicting cardiac arrest in diabetic patients work, is an emerging field where the ML models can synthesize complex, multi-modal data to make personalized risk (Sheela & Shalini, 2024; Ghule et al., 2024).

5. Critical Challenges and Future Directions

Radiomics has enormous challenges to clinical translation, significant as it might be:

1. **Reproducibility and Standardization:** The biggest obstacle is related to lack of standardization in the entire pipeline. Activities such as the Image Biomarker Standardisation Initiative (IBSI) are very important.
2. **Data Quality and Quantity:** Radiomics is data-hungry. There are common problems of missing data, imbalance in the classes and small samples in comparison to the size of features. The Spatio-Convolutional GAIN of imputation (Bansal et al., 2025) and the GANs used as generators of synthetic data are a promising solution.
3. **Model Interpretability and "Explainable AI" (XAI):** Complex ML/DL models possess a black-box, which is a critical issue to clinical adoption. There is an urgent demand to achieve made measures of model decisions transparent and reliable, which is also subject to behavioral economics AI applications (Ghule) and discussions on responsible AI (Puchakayala, 2022).

4. **Multimodal Integration:** Multimodal machine learning is the way forward, which involves radiomic data along with genomic (radiogenomics), clinical, pathologic, and other data stream combined in constructing more holistic models. Sardesai et al. (2025) sum up the significance and difficulties of this combination in their systematic review of ML.
5. **Moral and Equitable Implementation:** Like any AI in medicine, the problem of algorithmic biases, privacy of information, and equal access should be presupposed. Regulatory guidelines and ethics must be put in place to make sure that such technologies have advantages to all patients without discrimination (Puchakayala, 2022).

6. Conclusion

Radiomics is the state-of-the-art presence in the hand of precision medicine, which is driven by machine learning and sophisticated image processing. This is a synthesized review of the current developments in interconnected fields that show its ability to release the prognostic and diagnostic data that is concealed in medical images. One of the examples of how this may be transformative is in pulmonary disease. The way to strong models that are strategically, clinically implemented, however, is fraught with significant hurdles to do with standardization, validation, interpretability and ethics. The way forward in future studies should be to come up with reproducible pipelines, encouraging multimodal data integration, promoting explainable AI, and integrating ethical concerns upfront. Surmounting these obstacles, radiomics might become a promising research instrument to a trusted part of clinical decision-support systems, which will eventually result in better patient care and patient outcomes.

References

1. Amudala Puchakayala, P. R., Sthanam, V. L., Nakhmani, A., Chaudhary, M. F., Kizhakke Puliyakote, A., Reinhardt, J. M., & Bodduluri, S. (2023). Radiomics for improved detection of chronic obstructive pulmonary disease in low-dose and standard-dose chest CT scans. Radiology, 307(5), e222998.
2. Bansal, A., Puchakayala, P. R. A., Suddala, S., Bansal, R., & Singhal, A. (2025, May). Missing Value Imputation using Spatio-Convolutional Generative Adversarial Imputation Network. In 2025 3rd International Conference on Data Science and Information System (ICDSIS) (pp. 1-6). IEEE.
3. Ghori, P. (2018). Anomaly detection in financial data using deep learning models. International Journal Of Engineering Sciences & Research Technology, 7(11), 192-203.
4. Ghori, P. (2019). Advancements in Machine Learning Techniques for Multivariate Time Series Forecasting in Electricity Demand. International Journal of New Practices in Management and Engineering, 8(01), 25-37. Retrieved from https://ijnpme.org/index.php/IJNPME/article/view/220
5. Ghori, P. (2021). Enhancing disaster management in India through artificial intelligence: A strategic approach. International Journal of Engineering Sciences & Research Technology, 10(10), 40–54.
6. Ghori, P. (2023). LLM-based fraud detection in financial transactions: A defense framework against adversarial attacks. International Journal of Engineering Sciences & Research Technology, 12(11), 42–50.

7. Ghule, P. A., Sardesai, S., & Walhekar, R. (2024, February). An Extensive Investigation of Supervised Machine Learning (SML) Procedures Aimed at Learners' Performance Forecast with Learning Analytics. In International Conference on Current Advancements in Machine Learning (pp. 63-81). Cham: Springer Nature Switzerland.
8. Puchakayala, P. R. A. (2022). Responsible AI Ensuring Ethical, Transparent, and Accountable Artificial Intelligence Systems. Journal of Computational Analysis and Applications, 30(1).
9. Puchakayala, P. R. A. (2024). Generative Artificial Intelligence Applications in Banking and Finance Sector. Master's thesis, University of California, Berkeley, CA, USA.
10. Saha, P., Bodduluri, S., Nakhmani, A., Chaudhary, M. F., Amudala Puchakayala, P. R., Sthanam, V., & Bhatt, S. P. (2025). Computed tomography radiomics features predict change in lung density and rate of emphysema progression. Annals of the American Thoracic Society, 22(1), 83-92.
11. Sardesai, S., & Gedam, R. (2025, February). Hybrid EEG Signal Processing Framework for Driver Drowsiness Detection Using QWT, EMD, and Bayesian Optimized SVM. In 2025 3rd International Conference on Integrated Circuits and Communication Systems (ICICACS) (pp. 1-6). IEEE.
12. Sardesai, S., Kirange, Y. K., Ghori, P., & Mahalaxmi, U. S. B. K. (2025). Secure and intelligent financial data analysis using machine learning, fuzzy logic, and cryptography. Journal of Discrete Mathematical Sciences and Cryptography, 28(5-B), 2163–2173.
13. Shalini, S., Abhishek, S., Bhavyashree, P., Gunashree, C., & Rohan, K. S. (2023, May). An Effective Counterfeit Medicine Authentication System Using Blockchain and IoT. In 2023 4th International Conference for Emerging Technology (INCET) (pp. 1-5). IEEE.
14. Sheela, S., & Jadagerimath, A. N. (2022). Assistive communication system for physically challenged people. International Journal of Creative Research Thoughts, 10(2), 878–884. ISSN 2320-2882.
15. Sheela, S., Harshith, D., Jyothi, S., Reshma, D. S., Ravindranath, R. C., Sharmila, N., & Mallikarjunaswamy, S. (2025, July). Securing Pharmaceutical Supply Chains using AI-Integrated Blockchain Technology. In 2025 International Conference on Innovations in Intelligent Systems: Advancements in Computing, Communication, and Cybersecurity (ISAC3) (pp. 1-6). IEEE.