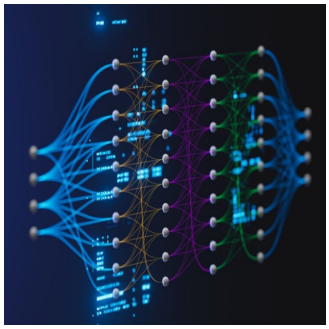
HealthManagement
Promoting Management and Leadership

DL Outperforms Radiomics in Chest X-Ray Classification



Artificial intelligence is reshaping chest radiography by automating complex image interpretation tasks and supporting multi-class diagnosis. Two prominent strategies are compared side by side: radiomics, which extracts handcrafted quantitative features, and deep learning, which learns hierarchical representations directly from images using convolutional architectures. A controlled evaluation across training cohorts ranging from very small to large assessed detection of COVID-19, lung opacity and viral pneumonia alongside normal findings. Preprocessing, validation and metrics were aligned to ensure comparability, and statistical testing examined the effects of model type and data volume. The results highlight clear advantages for deep learning as data scale increases, while radiomics remains pertinent where datasets and compute are constrained.

Design, Data and Evaluation

Analyses drew on 21,165 chest x-ray images partitioned into four classes: 3616 COVID-19, 6012 lung opacity, 1345 viral pneumonia and 10,192 normal. Images were 299 × 299 pixels in PNG format with lung segmentation masks in the posteroanterior view. A stratified test set reserved 344 samples per class. The remaining images formed stratified training subsets of 24, 48, 100, 248, 500, 1000, 2000 and 4000 samples, each evaluated under fivefold cross-validation with 80% training and 20% validation per fold.

Two pipelines were implemented. The radiomics workflow standardised images and masks, rescaled intensities to 0–255 and extracted intensity and texture features with PyRadiomics. Feature selection used ANOVA F-scores via SelectKBest, inputs were normalised, and classifiers included Decision Tree, Random Forest, Gradient Boosting and Support Vector Machine with default configurations, plus a multi-layer perceptron using 64- and 32-unit dense layers with dropout and softmax output trained for 100 epochs. For each classifier, the feature subset yielding the maximum F1 score was retained.

Must Read: ***Radiomics and Deep Learning in EGFR Prediction***

The deep learning workflow resized images to 256 × 256, normalised pixels to [0, 1] and applied augmentation with rescaling, shear, zoom and horizontal flips. Three ImageNet-pretrained architectures were fine-tuned end to end with a consistent classification head: ConvNeXtXLarge, EfficientNetL and InceptionV3. Models used the Adam optimiser with learning rate 0.0001, batch size 16 and categorical cross-entropy over 100 epochs without scheduling or early stopping.

All models produced class probabilities for normal, COVID-19, viral pneumonia and lung opacity. Performance was reported from the best epoch on the validation set and averaged across folds and runs. Metrics included F1 score, area under the receiver operating characteristic curve, accuracy, sensitivity and specificity. To examine the effects of model type and data volume, a Scheirer–Ray–Hare test on ranked scores assessed main and interaction effects, and pairwise Mann–Whitney U tests with Bonferroni correction probed differences between model families and between sample sizes.

Performance Across Sample Sizes

Performance improved as sample size increased, with deep learning showing the steepest gains. At 24 samples, EfficientNetL led with an area under the curve of 0.839, surpassing Support Vector Machine at 0.762. InceptionV3 reached 0.804 at this size, while ConvNeXtXLarge posted 0.691. By 4000 samples, InceptionV3 achieved the top area under the curve of 0.996 with accuracy of 0.960. EfficientNetL closely followed at 0.994, and ConvNeXtXLarge reached 0.978. Among radiomics-based approaches, Random Forest and Support Vector Machine improved steadily but plateaued below deep learning at larger scales, reaching area under the curve values around 0.885 and 0.881 respectively at 4000 samples.

Variability was greatest at the smallest training sizes and declined as cohorts grew, particularly for deep learning. EfficientNetL's F1 standard deviation fell from ± 0.042 at 24 samples to ± 0.006 at 4000 samples, indicating more stable learning with scale. Accuracy and sensitivity followed similar trajectories. Radiomics models displayed modest variance reduction yet remained less competitive at higher data volumes. Decision Tree and Gradient Boosting consistently underperformed relative to Support Vector Machine and Random Forest, especially in small-sample settings where instability and lower mean scores were pronounced.

Statistical analyses corroborated these trends. The Scheirer–Ray–Hare test identified significant main effects of model and sample size across area under the curve, F1 score, accuracy, sensitivity and specificity. Significant interaction effects indicated that performance gains from additional data depended on the model family. Pairwise Mann–Whitney U tests showed highly significant differences for many comparisons, including consistent gaps between deep learning models and several radiomics classifiers. Some deep learning pairings, notably InceptionV3 versus EfficientNetL, did not show statistically significant differences under the corrected tests, reflecting broadly similar performance at larger scales. Sample size comparisons showed progressively smaller p-values as gaps widened, with very small groups differing markedly from larger sets, underscoring the central role of data availability.

Implications for Model Selection

The comparative picture is clear. When sufficient data is available, deep learning models deliver the highest area under the curve, accuracy and sensitivity, and their variability diminishes with scale. InceptionV3 topped the leaderboard at 4000 samples, while EfficientNetL was consistently strong from small to medium cohorts and remained competitive at larger sizes. ConvNeXtXLarge improved steadily, aligning with the other top performers as data increased.

Radiomics-based models retain practical value in constrained settings. Support Vector Machine and Random Forest provided dependable, if lower, performance, offering advantages in transparency through feature-level outputs and in reduced computational demands that suit environments without specialised hardware. The multi-layer perceptron, although trained on radiomics features, benefited from neural network capacity and showed gains with larger radiomics feature sets, particularly for F1 score and area under the curve in the 1000–4000 sample range. Conversely, Decision Tree and Gradient Boosting were comparatively unreliable in this task, especially where data were scarce.

Operational considerations complement these results. Deep learning requires GPU resources and longer training times, and learned features are less interpretable. Radiomics pipelines can be deployed with modest compute and offer feature-level explanations that aid clinical acceptance, albeit with lower peak accuracy. The significant interaction between model architecture and sample size indicates that model choice should be closely tied to data volume, stability requirements and infrastructure.

Across balanced, multi-class chest x-ray classification of normal, COVID-19, viral pneumonia and lung opacity, deep learning outperformed radiomics approaches as training data increased, culminating in InceptionV3 reaching 0.996 area under the curve and 0.960 accuracy at 4000 samples. EfficientNetL was particularly strong at small and medium sizes and remained competitive at scale. Support Vector Machine and Random Forest offered stable performance for low-data or low-compute contexts, while Decision Tree and Gradient Boosting were least reliable. Statistical testing confirmed significant effects of model type, sample size and their interaction. Aligning model selection with available data and infrastructure can maximise accuracy, stability and feasibility in diagnostic AI deployment.

Source: **Journal of Imaging Informatics in Medicine**

Image Credit: **iStock**

Published on : Sun, 5 Oct 2025