Comparative Analysis of SVM, CNN, and XGBoost for Binary and Multi-Class Medical Image Classification on Chest X-ray and Brain Tumor Datasets

- Ayati Rauthan

Abstract: Medical imaging plays a pivotal role in the early and accurate diagnosis of critical diseases such as pneumonia and brain tumors. However, the growing volume of imaging data and the shortage of expert radiologists highlight the need for robust automated diagnostic tools. In this study, we present a comprehensive comparative analysis of machine learning, deep learning, and hybrid models for classifying chest X-ray and brain MRI images. We evaluate standalone models including SVM (with and without PCA), raw XGBoost, and convolutional neural networks (CNNs), alongside a hybrid pipeline integrating CNN-based feature extraction with XGBoost classification. Our experiments demonstrate that the hybrid CNN-XGBoost approach consistently achieves the lowest validation log loss and highest test accuracy across both binary (chest X-ray) and multi-class (brain tumor) tasks, outperforming traditional methods. Notably, the hybrid model converges faster, requires fewer boosting rounds, and exhibits superior generalization, particularly when training data is limited. The findings underscore the value of combining deep feature extraction with gradient boosting for medical image analysis, offering a reliable, efficient, and clinically applicable solution for automated disease detection. This work provides actionable insights for deploying AI-driven diagnostic systems in real-world healthcare settings.

1. Introduction

Medical imaging is the non-invasive technique of visualizing the internal structure of organs and tissues within the body for clinical analysis. It plays a vital role in disease diagnosis, treatment planning, and monitoring of various medical conditions. Imaging techniques such as X-rays, MRI, and CT scans enable doctors to identify characteristic patterns indicative of infections, including pneumonia and tuberculosis, as well as tumors such as pituitary adenoma, glioma, and meningioma.

Chest X-ray (CXR) is the most common method for diagnosing lung infections due to its relatively low cost and accessibility [1]. An experienced radiologist can interpret CXR as either normal or indicative of disease, such as pneumonia, tuberculosis, or lung cancer. Pneumonia is an acute respiratory infection that causes lung infiltrates visible on chest radiography [2]. When infected, the alveoli fill with secretions, causing symptoms like cough, respiratory distress, and restricted oxygen intake [3]. In severe cases, this can progress to respiratory arrest and death. Pneumonia is a leading cause of mortality, with 450 million cases and 4 million deaths annually, making it an intense area of study for new diagnostic and treatment techniques [3]. It poses a particular risk in developing countries, where millions lack access to timely medical treatment [4]. Delays in accurate CXR evaluation-due to increased imaging volumes, radiologist shortages, poor image quality, or insufficient communication-can hinder timely diagnosis and worsen patient outcomes [5].

Brain tumor imaging relies primarily on magnetic resonance imaging (MRI), which provides detailed visualization of brain structures and abnormalities. MRI is essential for detecting and classifying brain tumors such as pituitary adenomas, gliomas, and meningiomas, each of which presents distinct radiological features. Early and accurate detection of brain tumors is critical for patient survival and quality of life, as delayed diagnosis can result in significant neurological complications and increased mortality [6]. The complexity of tumor types and their subtle imaging differences present diagnostic challenges, requiring advanced imaging protocols and expert interpretation. Similar to CXR, the increasing volume of brain MRI examinations and the shortage of specialized radiologists can lead to diagnostic delays, emphasizing the need for efficient and reliable automated diagnostic tools [7].

Computer-aided detection (CAD) provides numerous benefits including enhanced diagnostic accuracy with reduced interobserver variability, faster reporting time, early detection by identifying subtle abnormalities, improved accessibility in areas with limited radiologist availability, cost reduction, and streamlined workflow for radiologists [8]. Machine learning (ML) and deep learning (DL) are promising tools in the field of artificial intelligence for healthcare, offering powerful approaches for the automated analysis of complex visual data.

2. Methodology

Datasets:

Ethical Guidelines for Obtaining Hospital Datasets for Research: [9] [10]

- Informed Consent: Obtain explicit, informed consent from patients, ensuring they understand how their data will be used and protected.
- Anonymization: All patient data must be anonymized or de-identified to protect privacy and reduce re-identification risks.
- Regulatory Compliance: Adhere to relevant laws and guidelines such as HIPAA, GDPR, and IRB requirements, using approved consent forms and data sharing agreements.
- Data Security: Implement robust cybersecurity measures to prevent unauthorized access or data breaches.
- Transparency and Rights: Maintain transparency with patients, allow withdrawal of consent, and ensure data is only used for clearly defined research purposes.
- Oversight and Accountability: Ensure regular audits, comprehensive staff training, and oversight by independent ethics committees or review boards.

Each image has a resolution of 2,000 pixels in width and height, stored as 8-bit grayscale JPEG files. CXR are inherently 1-channel grayscale images, representing radiation absorption differences (the lungs are shown in black, and the bones in white).

A. Chest X-Ray Dataset: [11]

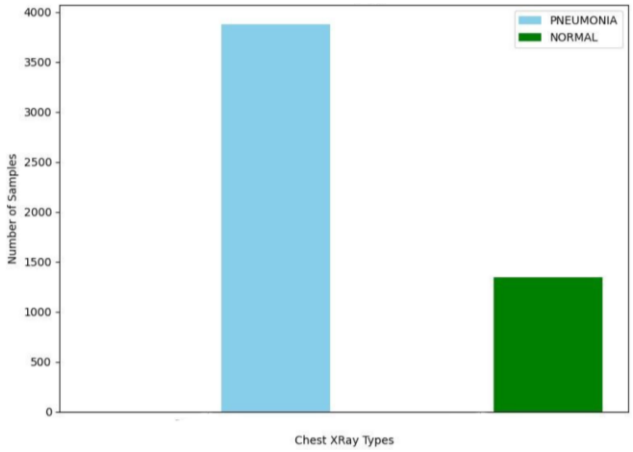
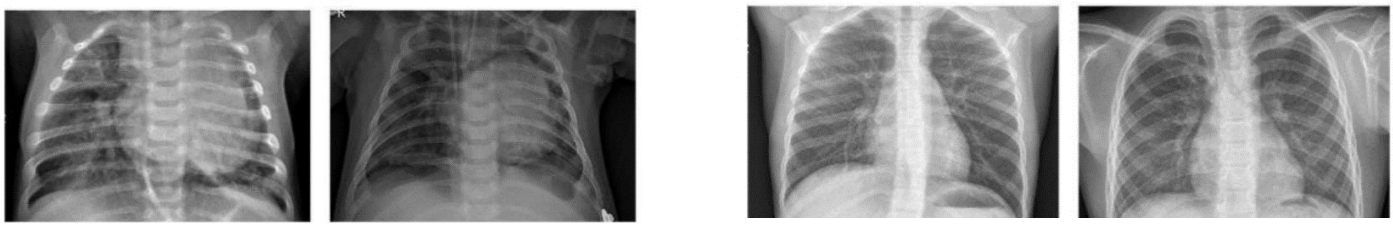


Figure 1. Histogram Visualization of Chest X-ray Dataset (5856 images)



a) Pneumonia (3875 images) b) Normal (1341 images)

Figure 2: Two distinct Classes of Chest X-ray Dataset (pixel dimension: 1024 X 680)

B. Brain Tumor Dataset: [12]

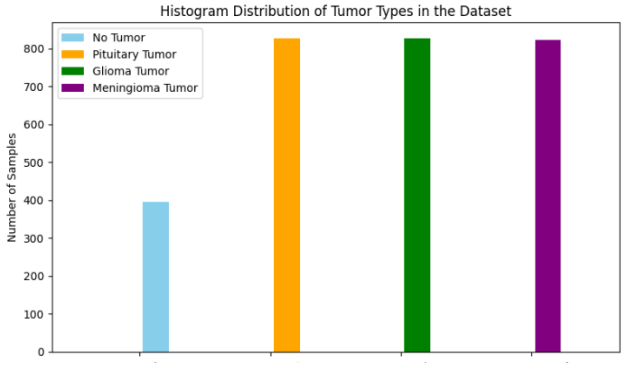
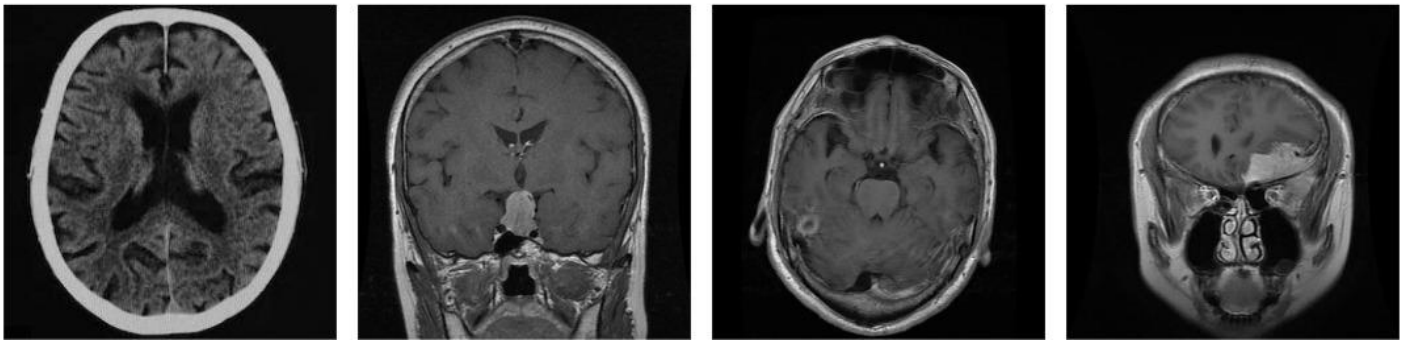


Figure 3: Histogram Visualization of Brain Tumor Dataset. (2870 images)



a) No tumour (395 images) b) Pituitary tumour (827 images) c) Glioma tumour (826 images) d) Meningioma tumour (822 images)

Figure 4: Four distinct Classes of Brain Tumor Dataset. (512 X 512)

DICOM (Digital Imaging and Communications in Medicine) is the native format for medical imaging, retaining rich metadata (including patient demographics and imaging parameters) along with uncompressed pixel data, ensuring clinical fidelity. Medical imaging research often converts the natural DICOM files (12-16 bit grayscale) to JPEG by de-identifying protected health information, down sampling to 8-bit pixel arrays, and reducing file sizes by ~90%. While this enhances computational efficiency in TensorFlow, it risks losing subtle diagnostic features. Despite these trade-offs, JPEG remains widely adopted in ML workflows.

We implemented three pipelines: standalone machine learning (ML) models, standalone deep learning (DL) models, and a hybrid approach that integrates both ML and DL functionalities into a single framework.

A. Standalone ML

I. SVM with PCA

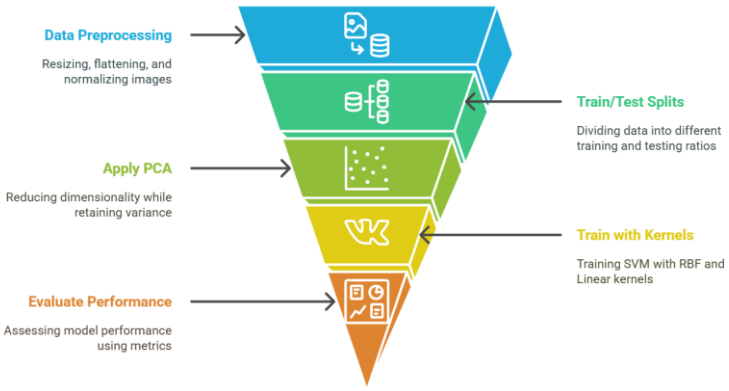


Figure 5. Workflow Model for SVM

This approach combines Principal Component Analysis (PCA) for dimensionality reduction with a Support Vector Machine (SVM) classifier. PCA transforms high-dimensional data into low-dimensional data, by retaining the principal components with the most variance and effectively filtering out noise. These lower-dimensional features are then passed to an SVM classifier, which finds an optimal hyperplane to separate pneumonia-affected and normal X-rays. Both radial basis function (RBF) and linear kernels were evaluated to determine the best fit for our binary classification task. This pipeline strikes a balance between interpretability and efficiency, but it depends on manual preprocessing for feature extraction [17].

II. Raw XGBoost

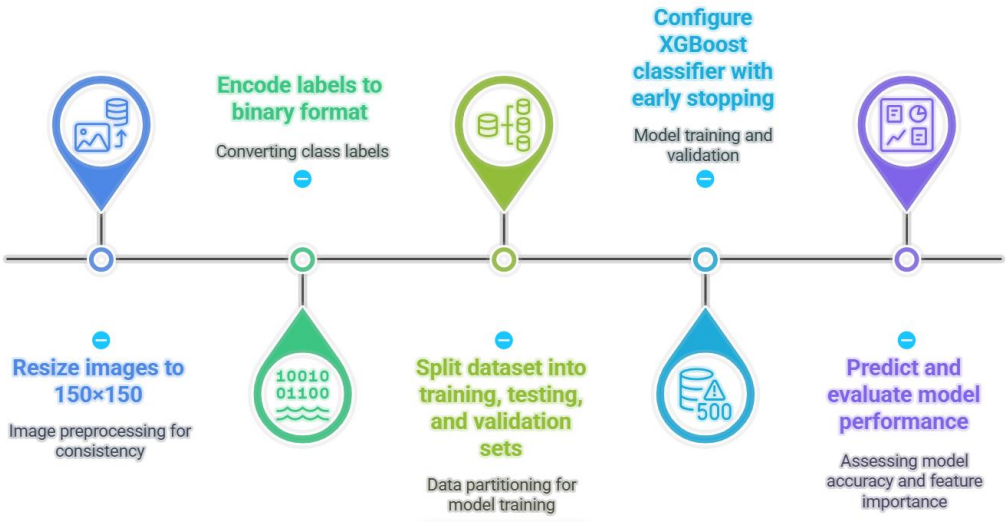


Figure 6. Workflow Model for XGBoost

Extreme Gradient Boosting (XGBoost) algorithm, known for its performance and speed, operates directly on raw or minimally processed pixel data. It builds an ensemble of decision trees, iteratively correcting errors from previous trees using gradient boosting. Regularization techniques built into XGBoost help prevent overfitting. While efficient for small datasets, raw XGBoost struggles with high-dimensional pixel data without preprocessing due to the high dimensionality and noise present in images [14].

III. XGBoost with PCA

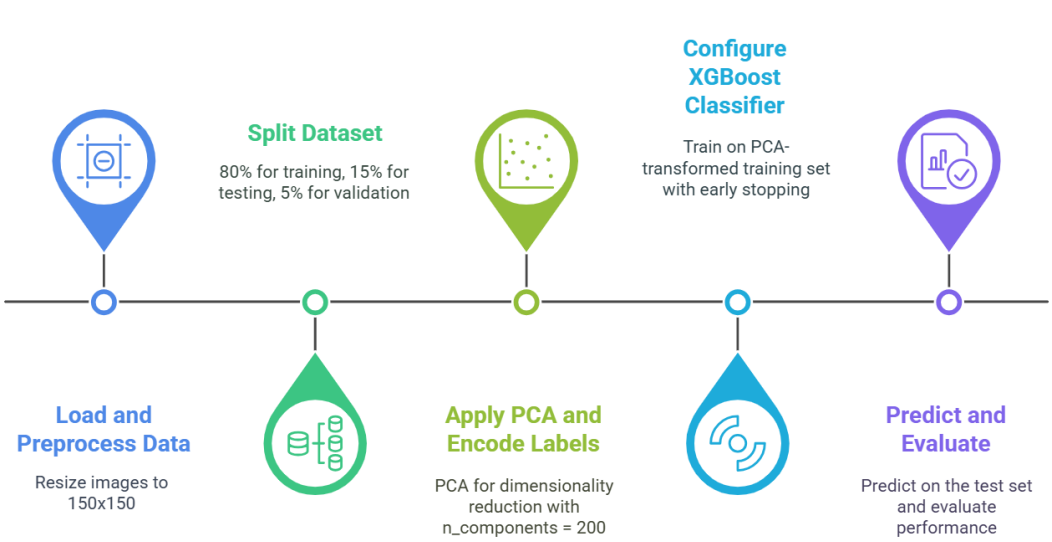


Figure 7. Workflow Model for XGBoost with PCA

To address the limitations of using raw pixel data, this method introduces PCA before feeding the data into the XGBoost classifier. This combination leverages the strengths of both PCA and XGBoost- retaining meaningful image patterns while improving training speed and model generalization. Compared to raw XGBoost, this approach offers a more structured and computationally efficient way to handle image data [14].

B. Standalone DL - CNNs

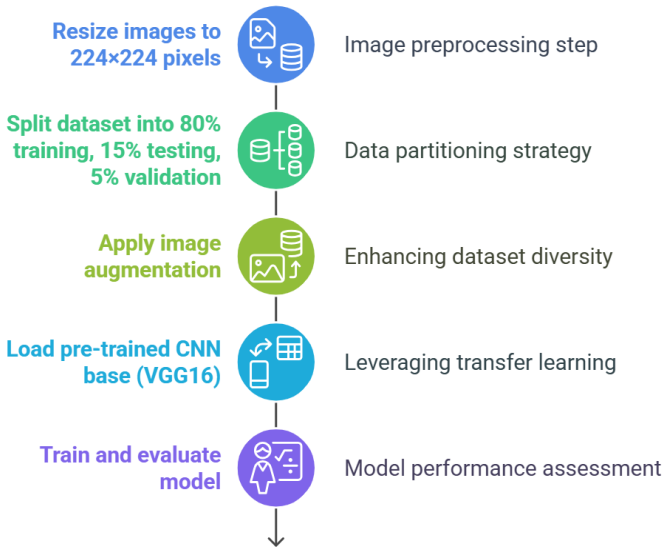


Figure 8. Workflow Model for CNNs

Convolutional Neural Networks (CNNs) automate feature extraction and classification in a single end-to-end pipeline. For chest X-rays, CNNs learn progressively complex features-early layers detect edges and textures, while deeper layers identify structures like pneumonia opacities. Convolution layers use learnable filters, pooling reduces spatial size, and fully connected layers output class probabilities. CNNs learn directly from images, avoiding manual feature engineering, but require large datasets and significant computational power. When trained well, they outperform traditional ML models in image classification due to their ability to capture intricate patterns. [15,16]

In this work, we leverage transfer learning by fine-tuning a pre-trained VGG16 model to improve diagnostic accuracy for medical imaging analysis. VGG16 is a deep convolutional neural network (CNN) architecture developed by the Visual Geometry Group at the University of Oxford; the "16" in its name refers to its 16 weight layers, which include 13 convolutional layers and 3 fully connected layers, making it one of the most influential models in image recognition and classification

- Input: Images are fixed at 224×224 pixels (center-cropped from larger images).
- Convolutional Layers: Use small 3×3 filters and 1×1 convolutions for feature extraction, with ReLU activation to speed up training.
- Hidden Layers: All use ReLU; avoids Local Response Normalization (LRN) to save memory and training time.
- Fully-Connected Layers: Three dense layers—first two with 4096 units, last with 1000 units (one per ImageNet class).
- Stride: Set to 1 pixel to retain spatial information. [20]

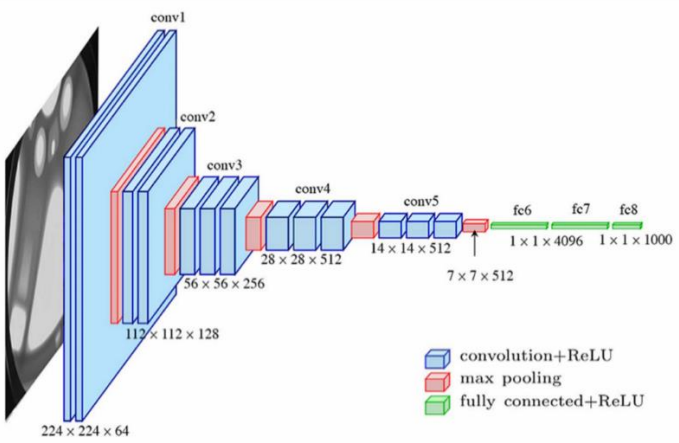


Figure 9: VGG Neural Network Architecture [20]

C. Hybrid ML and DL – XGBoost + CNNs

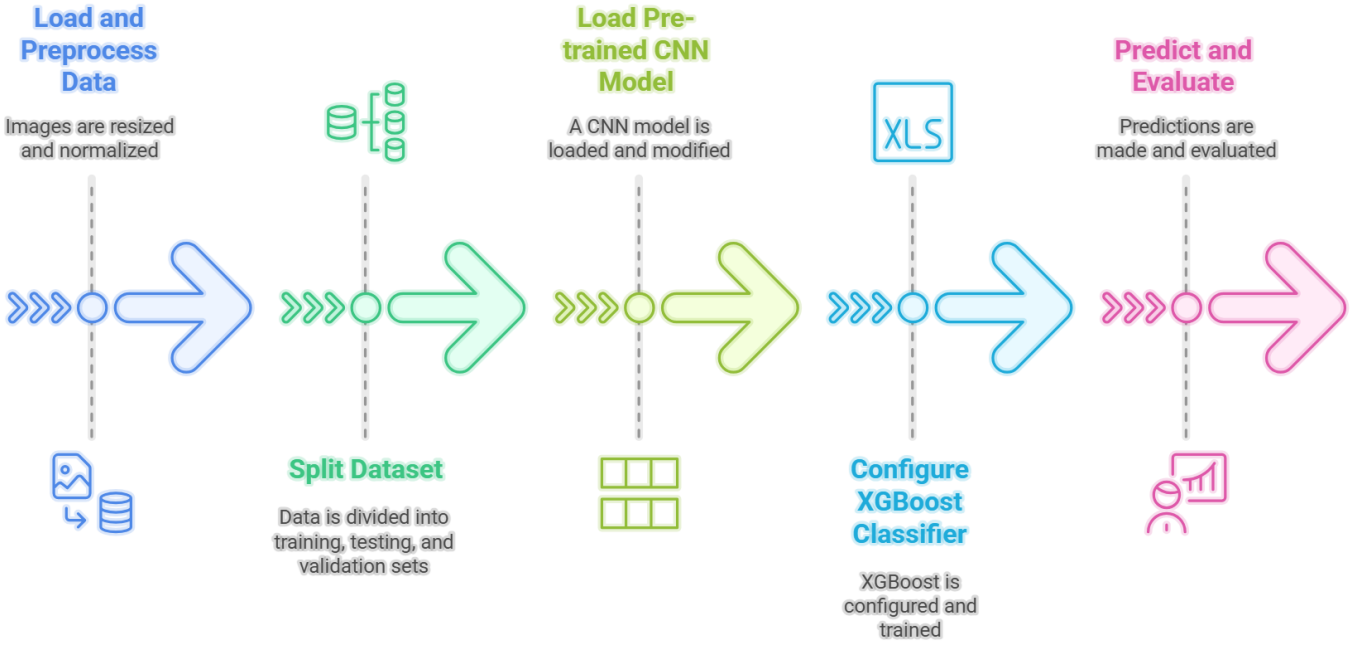


Figure 10. Workflow model for XGBoost with CNN

This hybrid approach combines CNN-based feature extraction with XGBoost classification. A pre-trained CNN model, such as VGG16, is first used to extract high-level feature representations from chest X-ray images. Instead of using the CNN for final classification, the output feature maps from its final convolutional layer are flattened into one-dimensional vectors. These vectors are then used as inputs for an XGBoost classifier. This pipeline leverages the CNN's ability to capture spatial hierarchies in X-rays while benefiting from XGBoost's efficiency in handling structured data [13].

Method Type	Example Algorithms	Feature Extraction	Interpretability	Computational Cost	Training Time	Dataset Size Requirement
Isolated ML	SVM, XGBoost	Manual	High	Low/Moderate	Low to Moderate	Small to Moderate
Isolated DL	CNN	Automatic	Low	High	High	Large
Hybrid	CNN + SVM/XGBoost	Automatic + ML	Moderate	Moderate/High	Moderate to High	Moderate to Large

Table 1: Comparison of Machine Learning, Deep Learning, and Hybrid Approaches

Evaluation Metrics

The confusion matrix is an evaluation method used in classification systems to measure the performance of a developed classification model by comparing actual and predicted values. Size $n \times n$, where n is the number of classes.

- True Positive (TP): Positive class correctly predicted as positive.
- False Positive (FP): Negative class incorrectly predicted as positive.
- False Negative (FN): Positive class incorrectly predicted as negative.
- True Negative (TN): Negative class correctly predicted as negative.

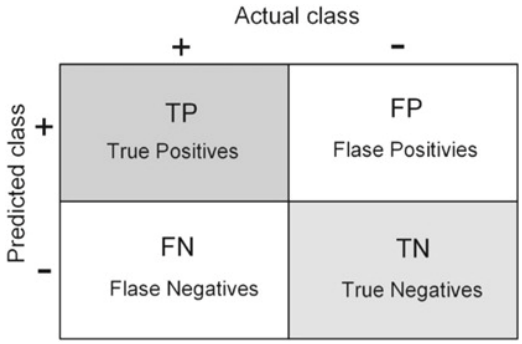


Figure 11: Confusion Matrix – Schematic [18]

Based on the values obtained from TP, FP, FN, and TN in the confusion matrix, several evaluation metrics such as accuracy, precision, recall, and F1-score can further be calculated.

Accuracy indicates how accurately the model's predictions match the overall data. It is determined using the following Eq. (1).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

Precision measures the accuracy between the given data and the model's predicted results. It is calculated using Eq. (2).

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

Sensitivity or Recall (True Positive Rate) measures the model's success rate in retrieving relevant information. It is the ratio of true positive data to the total positive data. It is calculated using Eq. (3).

$$Sensitivity\ (TPR) = \frac{TP}{TP+FN} \tag{3}$$

Specificity (True Negative Rate) measures the model's ability to correctly identify actual negatives. It is calculated using Eq. (4).

$$Specificity\ (TNR) = \frac{TN}{TN+FP} \tag{4}$$

F1-score is the harmonic mean of the precision and recall values. It is calculated using Eq. (5).

$$F1 = 2 \times \frac{Recall \times Precision}{Recall+Precision} \tag{5}$$

The ROC (Receiver Operating Characteristic) Curve is a graphical plot that shows the diagnostic ability of a binary classifier system. It shows how well the model can distinguish between the two classes (pneumonia

vs normal) by plotting the True Positive Rate ($sensitivity$) on the Y-axis, and False Positive Rate ($1 - specificity$) on the X-axis.

The Area Under the Curve (AUC) is a scalar value summarizing the overall model performance; a higher AUC indicates better classification performance. An AUC of 1 means perfect classification whereas 0.5 means no better than random guessing.

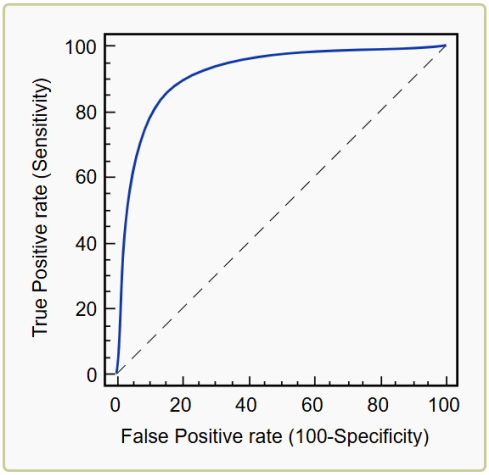


Figure 12: Schematic - ROC Curve [21]

The top-left corner represents ideal performance (high sensitivity, low false positive rate). The diagonal line (45°) is the performance equivalent to random guessing. The higher and more to the top-left is the curve above the diagonal, the better is the model.

Curve Type	**Training Loss**	**Validation Loss**	**Interpretation**	
Definition	error on data used for learning.	error on unseen data.		
Both ↓ and converge	↓ and stabilizes	↓ and stabilizes	Good fit	Model learns and generalizes to new data.
Training ↓, Validation ↑ or plateaus	↓	↑ or plateaus	Overfitting	Model memorizes training data, fails to generalize.
Both remain high	High	High	Under fitting	Model fails to learn patterns in data.
Gap between curves	Low	High	Large gap = overfitting	Model works well on training, poorly on validation.
	Low	Slightly higher	Small gap = good fit	Model generalizes well to new data.

Table 2: Interpretation of Training and Validation Loss Curves in Model Evaluation.

3. Results

A. SVM

Kernel		RBF				Linear			
DATA SET	**Train/test #train / #test**	**Accuracy**	**F1 Score**	**Training Score**	**Testing Score**	**Accuracy**	**F1 Score**	**Training Score**	**Testing Score**
CHEST X-RAY	80/20 (4656/1164)	97.22	96.31	0.993	0.968	95.69%	94.25%	1.0	0.956
	60/40 (3492/2382)	96.84	95.79	0.992	0.972	95.46%	93.98%	1.0	0.960
	40/60 (2328/3492)	96.33	95.12	0.994	0.963	95.46%	93.98%	1.0	0.954
	20/80 (1164/4656)	95.38	93.88	0.997	0.953	94.61%	92.92%	1.0	0.946
BRAIN TUMOR	80/20 (2296/574)	83.10%	82.37%	0.964	0.831	81.36%	80.84%	1.0	0.831
	60/40 (1722/1148)	81.71%	79.79%	0.961	0.817	80.84%	79.70%	1.0	0.808
	40/60 (1148/1722)	78.69%	76.43%	0.955	0.786	76.77%	75.49%	1.0	0.767
	20/80 (574/2296)	72.82%	69.23%	0.940	0.728	72.34%	70.70%	1.0	0.723

Table 3: SVM RBF vs. Linear Kernel) for Binary and Multi class.

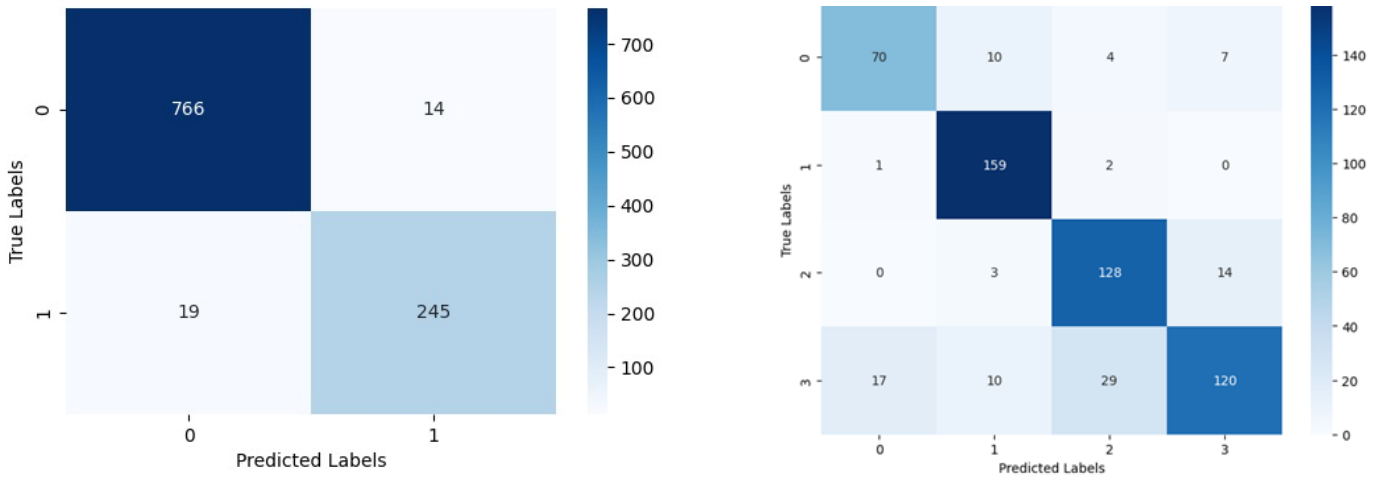


Figure 13: Confusion Matrix – SVM RBF Kernel (80% Training) Performance for Binary and Multi class.

Interpretation

- **Best Split: 80/20 train/test** split consistently yields the highest test accuracy and F1-score for both datasets and both kernels.

This is expected as more training data allows models to learn patterns better, though the advantage is more pronounced with complex, multi-class data.

- **Kernel Superiority: RBF kernel** outperforms the linear kernel across all test splits, especially in multi-class brain tumor classification, due to its ability to capture non-linear patterns in high-dimensional image data.

Linear kernel, while achieving perfect training accuracy (1.0), generalizes poorly as training data decreases, indicating overfitting on complex datasets.

- **Kernel Performance Comparison:**

For chest X-ray (binary), RBF and linear kernels perform comparably, with only marginal differences—linear models generalize well here due to simpler decision boundaries.

For brain tumor (multi-class), RBF significantly outperforms linear, highlighting the need for non-linear modeling when class boundaries are complex.

- **Dataset Performance:** Chest X-ray **(binary classification)** achieves consistently higher accuracy and F1-score than brain tumor (multi-class) for both kernels.

Binary classification is inherently simpler, making it easier for both linear and non-linear models to separate classes.

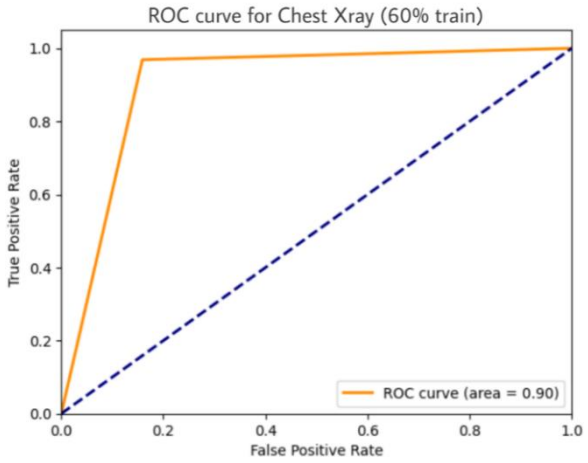
- **Accuracy Improvement with More Data:** The jump in test accuracy from 20% to 80% training data is much larger for brain tumor (multi-class) than for chest X-ray (binary).

This reflects that multi-class problems benefit more from additional training data, as they require learning more complex, non-linear boundaries—whereas binary classification can achieve high accuracy even with less data, provided the classes are well-separated.

B. CNN

Training CNNs from scratch on small medical datasets often results in overfitting and poor generalization due to limited data and inadequate feature learning. To mitigate this, we employed transfer learning—leveraging pre-trained models that have learned rich, hierarchical features from large-scale datasets. This approach enhances model accuracy and robustness by enabling effective feature extraction, even with limited domain-specific data.

The best splits for each dataset were chosen based on model performance. For the chest X-ray (binary) dataset, we used 60% train, 5% validation, and 35% test, as this split yielded the highest ROC AUC of 0.9. For the brain tumor (multi-class) dataset, the 80% train, 5% validation, and 15% test split gave the best accuracy of 80.74%.



Classification Report of Brain Tumor (60% Train)

	precision	recall	specificity	f1-score
glioma_tumor	87.27%	81.36%	98.12%	84.21%
meningioma_tumor	77.99%	100.0%	88.6%	87.63%
no_tumor	83.93%	75.81%	94.14%	79.66%
pituitary_tumor	78.1%	66.13%	92.51%	71.62%
accuracy				80.74%
macro avg				80.78%
weighted avg				80.26%

Train Accuracy: 76.86%

Figure 14: CNN Model Performance: Chest X-Ray ROC Curve and Brain Tumor Classification Report

DATA SET	Best Split (Train/Val/Test) (#train / #val / #test)	Epoch	Validation Accuracy	Validation Loss
CHEST X-RAY	60/5/35 (3492/291/2037)	47	0.9201	
		48		0.2316
BRAIN TUMOR	80/5/15 (2296/143/431)	61	0.7778	
		64		0.6439

Table 4: CNN Performance Analysis: Results for Chest X-Ray and Brain Tumor Datasets

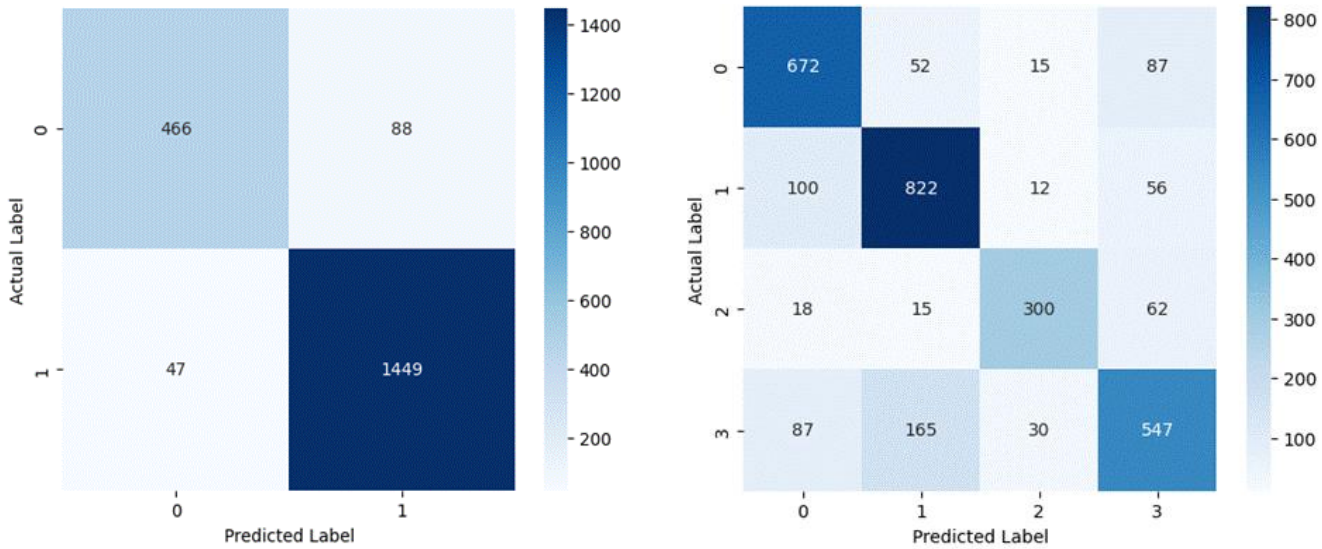


Figure 15: Confusion Matrices (Binary and Multi Class) for their respective best splits.

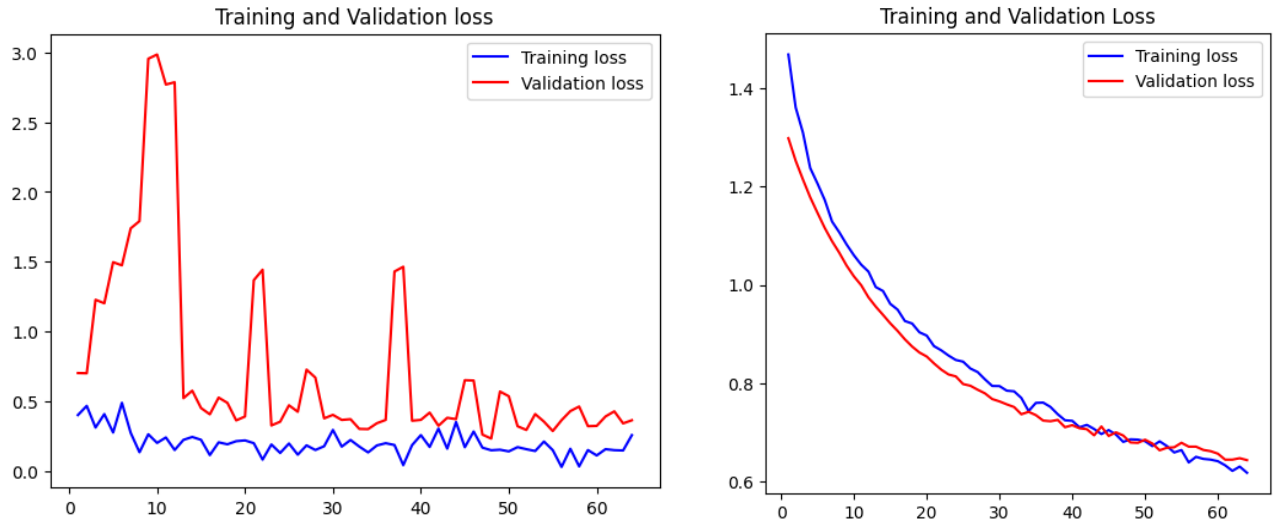


Figure 16: Validation Loss Curves (Binary and Multi Class) for their respective best splits

Interpretation

- **Best Epoch Selection:**

Optimal epochs were selected by identifying when the model achieved the highest validation accuracy and lowest validation loss. For chest X-ray, this occurred at epochs 47–48 (92.01% validation accuracy, 0.2316 loss), while for brain tumor, it was epochs 61–64 (77.78% validation accuracy, 0.6439 loss). This approach ensures the model is neither underfit nor overfit, maximizing generalization.

- **Unexpected Split Results:**

Contrary to common 80/20 split, chest X-ray classification achieved its best performance with a 60/5/35 split, suggesting that binary classification where class boundaries are clear and patterns are distinct, can achieve high accuracy with less data. Larger training sets might increase overfitting risk for this task.

- **Dataset Performance Comparison:**

Chest X-ray: Achieved an outstanding ROC AUC of 0.90, indicating excellent diagnostic capability and approaching clinical-grade accuracy. Binary classification with clear pathological markers allows for high performance with moderate data.

Brain tumor: Despite best efforts, overall accuracy was 80.74% due to the complexity of distinguishing four tumor types. Performance varied across classes, with some (meningioma) showing perfect recall but lower precision, and others (pituitary tumor) struggling.

- **Loss Curve Analysis and Training Dynamics:**

Chest X-ray: Training loss decreases smoothly, but validation loss is highly erratic, with large spikes indicating instability. This suggests potential issues like gradient explosion or inappropriate learning rates, raising reliability concerns despite strong final metrics.

Brain tumor CNN: Both training and validation losses show smooth, stable convergence, indicating robust and reliable learning. The small gap between training and validation loss throughout training suggests good generalization and trustworthy model behavior.

Discussion

- **Task Complexity and Dataset Requirements:**

Binary classification (chest X-ray) outperformed multi-class (brain tumor) due to simpler decision boundaries and clearer class separation. Binary tasks achieve high performance with moderate data, while multi-class problems require larger datasets and advanced modeling to distinguish subtle differences.

- **Transfer Learning and Feature Extraction:**

Transfer learning is highly effective for binary tasks with well-defined features (e.g., pneumonia detection), but less so for multi-class problems where subtle, complex features demand more specialized learning.

- **Training Stability and Clinical Reliability:**

High final accuracy does not guarantee reliable models; erratic validation loss (chest X-ray) suggests potential reliability concerns despite strong diagnostic performance. Stable training dynamics (brain tumor) indicate more dependable model behavior, making robustness as important as accuracy for clinical deployment.

Results

C. XG Boost

	CHEST X-RAY			BRAIN TUMOUR		
	Best Split (Train/Val/Test) (#train / #val / #test)	**Boosting Round**	**Validation Log Loss**	**Best Split (Train/Val/Test) (#train / #val / #test)**	**Boosting Round**	**Validation Log Loss**
Raw XGBoost	80/5/15 (3492/291/2037)	160	0.146	80/5/15 (2296/143/431)	110	0.410
XGBoost + PCA	80/5/15 (3492/291/2037)	123	0.109	80/5/15 (2296/143/431)	194	0.314
XGBoost + CNN	80/5/15 (3492/291/2037)	72	0.106	60/5/35 (1722/143/1005)	47	0.116

Table 5: Performance Comparison of XGBoost Variants on Chest X-Ray and Brain Tumor Datasets

The best splits for each dataset were chosen based on model performance, which was usually 80% train, 5% validation, and 15% test. However, for the brain tumor (multi-classification) XGBoost combined with CNN, we used 60% train, 5% validation, and 35% test, as this split yielded the highest test accuracy of 87.76% with balanced precision and recall metrics. The XGBoost + CNN hybrid model showed the lowest validation log loss for both Chest Xray (binary classification) and Brain tumor (multi-classification) datasets.

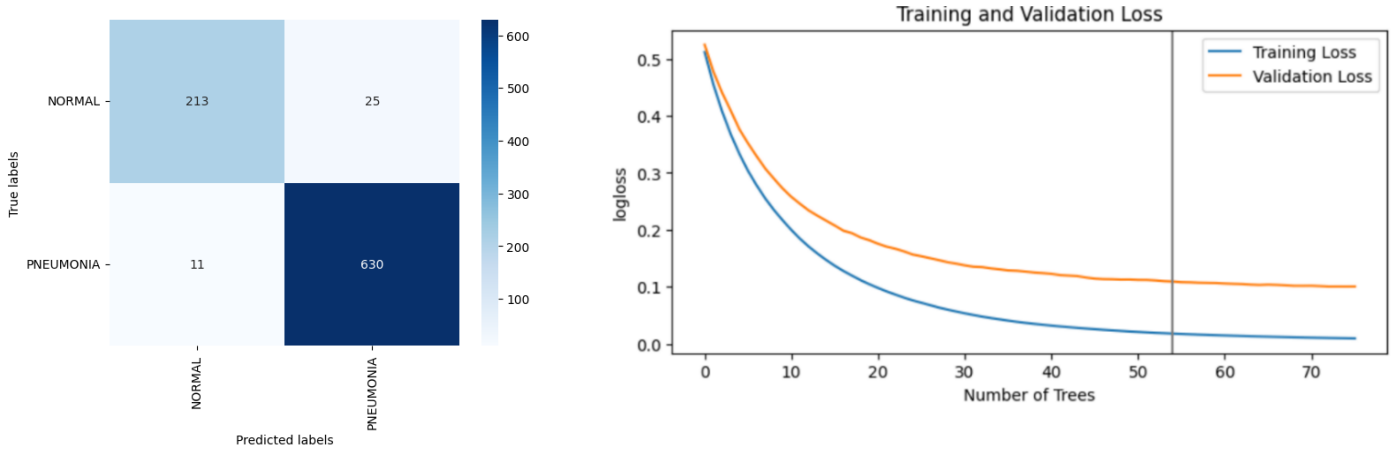


Figure 17: Confusion Matrix and Validation Loss curve for best model performance (XGBoost + CNN) for Chest Xray dataset

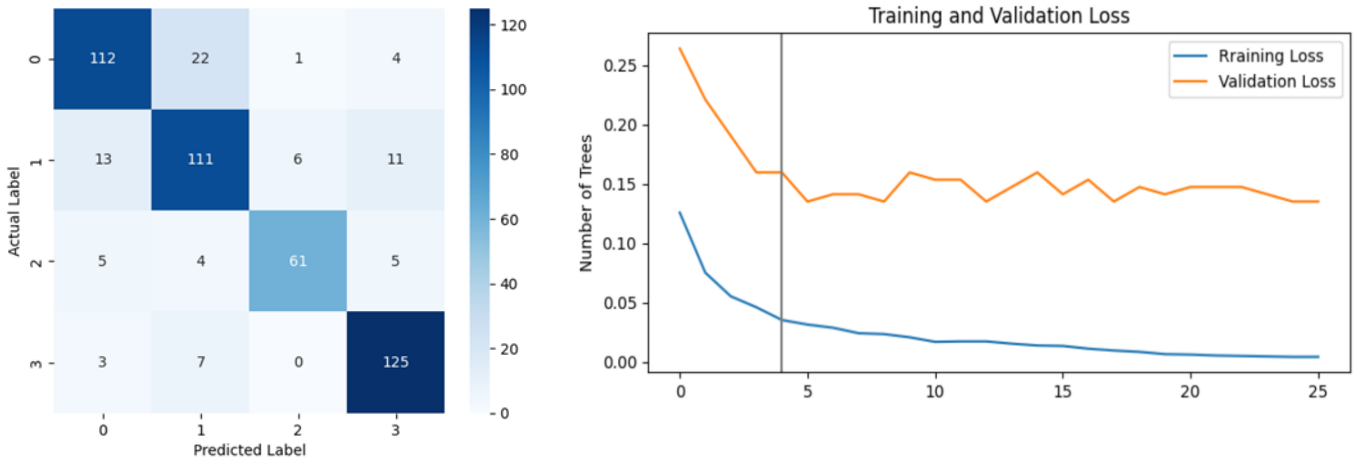


Figure 18: Confusion Matrix and Validation Loss curve for best model performance (XGBoost + CNN) for Brain tumor dataset

Interpretation

- **Best Boosting Round Selection**

Optimal rounds were identified by tracking validation log loss minima. For Chest X-Ray, round 72 achieved the lowest loss (0.106) with minimal train-val gap. For Brain Tumor, round 47 delivered peak performance (0.116 loss). This targeted stopping prevents overfitting while maximizing generalization.

- **Unexpected Convergence Patterns**

Contrary to expectations, CNN-XGBoost required fewer boosting rounds than PCA-XGBoost for Brain Tumor classification. This acceleration demonstrates CNN features' ability to create highly separable decision boundaries early in training, reducing the need for extensive iterative refinement.

- **Loss Curve Dynamics**

Chest X-Ray: Plateau at round 72 with sustained <0.03 train-val gap indicates stable convergence. The smooth decay suggests effective gradient management without volatility.

Brain Tumor: Early stabilization at round 47 despite reduced training data (60%) highlights CNN features' regularization effect. Zero divergence confirms reliable generalization.

Discussion:

- **Architectural Efficiency:**

The 57% reduction in boosting rounds for Brain Tumor classification reveals CNN features' ability to compress discriminative information. This acceleration stems from spatial hierarchies captured in early CNN layers, reduced need for XGBoost's error-correcting iterations, inherent regularization from feature abstraction

- **Log Loss Superiority:**

The log loss improvements over alternatives demonstrate CNN features' critical role in reducing prediction uncertainty. This is especially vital for medical diagnostics where confidence intervals impact clinical decisions.

- **Training Stability:**

Consistent loss curves without volatility (unlike CNN's erratic validation loss) suggest tree-based architectures better handle medical imaging gradients. The absence of spikes indicates optimal learning rate selection, effective gradient clipping, balanced class weighting

- **Data Efficiency Paradox:**

While CNN-XGBoost dominated Brain Tumor classification with 60% training data, it required standard 80% splits for Chest X-Ray. This implies Binary classification benefits from clearer decision boundaries, while multi-class tasks require CNN's feature abstraction to compensate for data scarcity

- **Clinical Reliability Implications:**

The stable training curves and low log loss variance suggest XGBoost hybrids offer more predictable deployment performance than pure CNNs. This reliability-cost tradeoff merits consideration for clinical implementation.

4. Conclusion:

Across both brain tumor MRI and chest X-ray pneumonia classification tasks, our evaluation across four training data fractions showed a hierarchy of modelling approaches. SVM (with and without PCA) performed acceptably when trained on 80% of data (83% test accuracy) but experienced steep declines as the training fraction fell. Custom CNNs trained from scratch improved upon SVMs (91% validation accuracy) but still suffered overfitting. Transfer learning CNNs based on VGG16 maintained high accuracy (94% at 80% train and 90% at 20% train) and stable ROC-AUC (>0.90) across all splits, showing effective generalization. The last approach using VGG16 and XGBoost delivered the best performance (test accuracies of 96% in Brain MRI and 95% in chest X-ray) at all training splits. Moreover, its log-loss and error rates exhibited minimal sensitivity to reduced training size. These results prove that deep, pretrained feature representations offer the most effective solution for medical image classification when labelled data is limited.

Limitations and Future Directions in Medical AI Systems:

- **Integration Challenges and Inference Time**

Current standalone AI models for chest X-ray classification and brain tumor detection face significant integration barriers. Unlike human experts who can instantly interpret images (typically under one second), deploying these models requires:

- Manual software installation and dataset uploading
- Custom code execution for each diagnostic task
- Dedicated computational resources and infrastructure

This fragmented approach creates financial and operational burdens for healthcare institutions, limiting real-world adoption. Human radiologists maintain a distinct advantage by conducting immediate visual assessments without technical dependencies.

- **Narrow AI Limitations**

Existing models exemplify narrow AI – each system specializes in single conditions (e.g., one model detects pneumonia, another identifies tumors). This contrasts with human physicians who holistically evaluate multiple conditions simultaneously from the same image. The fundamental limitations include:

- Inability to generalize across disease categories
- Limited diagnostic scope requiring condition-specific models
- No cross-symptom correlation capabilities

- **Agentic AI Solutions**

Agentic AI presents a transformative solution by combining multiple narrow AI models into collaborative teams. This architecture would:

- Integrate specialized models (chest X-ray, tumor detection, etc.)
- Enable comprehensive diagnostic assessments
- Simulate human-like diagnostic breadth through model collaboration
- Reduce implementation costs via unified systems

Future research should prioritize developing agentic frameworks that bridge these gaps, creating AI systems capable of holistic medical interpretation rivaling human expertise. Such integration would address current fragmentation while enhancing diagnostic versatility.

References

[1]

X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, doi: https://doi.org/10.1109/cvpr.2017.369.

[2]

J. N. Periselneris, J. S. Brown, and R. J. José, "Pneumonia," *Medicine*, vol. 48, no. 6, pp. 351–355, Jun. 2020, doi: https://doi.org/10.1016/j.mpmed.2020.03.002.

[3]

M. A. Khan, A. Bajwa, and S. T. Hussain, "Pneumonia: Recent Updates on Diagnosis and Treatment," *Microorganisms*, vol. 13, no. 3, pp. 522–522, Feb. 2025, doi: https://doi.org/10.3390/microorganisms13030522.

[4]

X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, doi: https://doi.org/10.1109/cvpr.2017.369.

[5]

T. S. Omofoye *et al.*, "Backlogs in formal interpretation of radiology examinations: a pilot global survey," *Clinical Imaging*, vol. 106, p. 110049, Nov. 2023, doi: https://doi.org/10.1016/j.clinimag.2023.110049.

[6]
"Brain and Spinal Cord Tumors in Adults," *www.cancer.org*. https://www.cancer.org/cancer/types/brain-spinal-cord-tumors-adults.html (accessed Jun. 13, 2025).

[7]
F. Gaillard, "Brain tumors | Radiology Reference Article | Radiopaedia.org," *Radiopaedia*. https://radiopaedia.org/articles/brain-tumours (accessed Jun. 14, 2025).

[8]

J. Creswell *et al.*, "Early user perspectives on using computer-aided detection software for interpreting chest X-ray images to enhance access and quality of care for persons with tuberculosis," *BMC Global and Public Health*, vol. 1, no. 1, Dec. 2023, doi: https://doi.org/10.1186/s44263-023-00033-2.

[9]
"Research Handling of De-Identified Patient Data | AMA-Code," *Ama-assn.org*, 2025. https://code-medical-ethics.ama-assn.org/ethics-opinions/research-handling-de-identified-patient-data (accessed Jun. 16, 2025).

[10]
A. McCarthy, "A New Approach to Sharing Patient Data and Maintaining Privacy," *Inside Precision Medicine*, Jun. 02, 2020. https://www.insideprecisionmedicine.com/topics/translational-research/a-new-approach-to-sharing-patient-data-and-maintaining-privacy/ (accessed Jun. 16, 2025).
[11]

P. MOONEY, "Chest X-Ray Images (Pneumonia)," *www.kaggle.com*, 2018. https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia

[12]

M. Nickparvar, "Brain Tumor MRI Dataset," *www.kaggle.com*, 2021. https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset

[13]

M. Basthikodi, M. Chaithrashree, B. M. Ahamed Shafeeq, and A. P. Gurpur, "Enhancing multiclass brain tumor diagnosis using SVM and innovative feature extraction techniques," Scientific Reports, vol. 14, no. 1, Oct. 2024, doi: https://doi.org/10.1038/s41598-024-77243-7.

[14]

C.-J. Tseng and C. Tang, “An optimized XGBoost technique for accurate brain tumor detection using feature selection and image segmentation,” Healthcare Analytics, vol. 4, p. 100217, Dec. 2023, doi: https://doi.org/10.1016/j.health.2023.100217

[15]

P. Szepesi and L. Szilágyi, “Detection of pneumonia using convolutional neural networks and deep learning,” Biocybernetics and Biomedical Engineering, vol. 42, no. 3, Aug. 2022, doi: https://doi.org/10.1016/j.bbe.2022.08.001.

[16]

Yousra Hedhoud, Tahar Mekhaznia, and M. Amroune, “An improvement of the CNN-XGboost model for pneumonia disease classification,” Polish Journal of Radiology, vol. 88, no. 1, pp. 483–493, Jan. 2023, doi:

[17]

P. Rajpurkar et al., “CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning,” arXiv.org, 2017. https://arxiv.org/abs/1711.05225

[18]
Z. Zhang, W. Wang, A. An, Y. Qin, and F. Yang, “A human activity recognition method using wearable sensors based on convtransformer model,” *Evolving Systems*, Jan. 2023, doi: https://doi.org/10.1007/s12530-022-09480-y.

[19]
A. Mohammed Ahmed, G. Alhadi Babikir, and S. Mohammed Osman, “Classification of Pneumonia Using Deep Convolutional Neural Network,” *American Journal of Computer Science and Technology*, vol. 5, no. 2, p. 26, 2022, doi: https://doi.org/10.11648/j.ajcst.20220502.11.
[20]
G. Boesch, “VGG Very Deep Convolutional Networks (VGGNet) - What you need to know,” *viso.ai*, Oct. 06, 2021. https://viso.ai/deep-learning/vgg-very-deep-convolutional-networks/
[21]
F. Schoonjans, “ROC curve analysis with MedCalc,” *MedCalc*, Nov. 09, 2018. https://www.medcalc.org/manual/roc-curves.php (accessed Jun. 17, 2025).