Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 9 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Enhancing Diagnostic Radiology with Explainable Deep Learning for Multi-Disease Detection

Mukhtar Ali Soomro
Student, HITEC University Taxila

Omer Ali
Lecturer, HITEC University

Minahil Iman
Student, HITEC University Taxila

Shabeer Ahmad
Assistant Director IT Board,
Intermediate & secondary education, saidu sharif swat.

ABSTRACT

The integration of deep learning (DL) and explainable artificial intelligence (XAI) is transforming diagnostic radiology by improving both the accuracy and interpretability of medical image analysis. This study developed three deep learning models a 5-block ResNet, an 8-block DenseNet, and a 5-block bottleneck architecture to detect and classify diseased tissue in medical images related to COVID-19, brain tumours, knee arthritis, and breast cancer. Using large datasets and implemented in MATLAB, the models were rigorously trained and validated to ensure high performance and generalizability. A major focus of this work was the incorporation of XAI techniques to make the models' decision-making processes transparent and clinically intelligible. By visualizing the regions influencing diagnoses, our approach helps build trust and facilitates collaboration between AI systems and medical practitioners. The results show that the combined DL XAI framework not only achieves high diagnostic accuracy but also provides actionable insights that align with clinical reasoning. This work supports the adoption of AI-assisted diagnostics for earlier and more reliable detection of diseases, ultimately contributing to improved patient care.

Key word: Deep Learning (DL), Explainable Artificial Intelligence (XAI), Medical Image Analysis, Diagnostic Radiology, ResNet, DenseNet.

Introduction:

The field of diagnostic radiology has undergone a profound transformation since the discovery of the X-ray, evolving into a cornerstone of modern medicine through technologies like computed tomography (CT) and magnetic resonance imaging (MRI). These imaging modalities provide non-invasive windows into the human body, fundamentally enhancing the detection, monitoring, and treatment planning for

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 9 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

a vast array of diseases. Despite these technological leaps, a significant challenge persists: the accurate and early identification of diseased tissues, such as those affected by infections, cancers, or degenerative conditions. The inherent complexity of medical images, where early pathological signs are often subtle and easily overlooked, coupled with the sheer volume of data radiologists must analyze, creates a critical gap between imaging capability and diagnostic certainty. This limitation underscores an imperative for more sophisticated, reliable, and efficient analytical tools to support clinical decision-making [1].

In response to these challenges, computer-aided detection (CAD) systems emerged as a pivotal innovation, designed to assist radiologists by highlighting potential abnormalities and serving as a valued second opinion. However, traditional CAD systems often lack the nuanced analytical power required for the most complex diagnostic tasks. The advent of deep learning (DL) represents a quantum leap beyond these conventional systems, offering unprecedented accuracy in pattern recognition and image classification. By leveraging advanced architectures like ResNet, DenseNet, and custom bottleneck models, DL algorithms can be trained on extensive datasets to detect and classify pathological regions with a proficiency that rivals, and in some cases surpasses, human performance [2]. Yet, this power introduces a new dilemma: the "black box" nature of complex deep learning models, where the reasoning behind a diagnosis remains opaque and unexplainable, eroding the trust essential for clinical adoption.

This research directly addresses the critical need for transparency by integrating explainable artificial intelligence (XAI) with state-of-the-art deep learning models. Our project moves beyond mere diagnostic accuracy to forge a collaborative partnership between AI and the radiologist. We developed and validated three distinct deep learning models—a 5-block ResNet, an 8-block DenseNet, and a 5-block bottleneck architecture—tailored to analyze medical images of diseases including COVID-19, Brain Tumours, Knee Arthritis, and Breast Cancer. The key innovation lies not only in their performance but in the application of XAI techniques designed to elucidate the models' decision-making processes. By generating visual explanations that pinpoint the image features influencing a diagnosis, our framework makes the AI's reasoning intelligible and alignable with clinical expertise [4].

Consequently, this work aims to bridge the gap between algorithmic output and clinical applicability, transforming the AI from an inscrutable black box into a transparent and trustworthy diagnostic assistant. The objectives of this research are to enhance the accuracy and efficiency of disease detection across diverse medical conditions while providing interpretable results that empower healthcare professionals. By demystifying AI-generated diagnoses, we contribute to the field of diagnostic radiology by offering a robust, explainable tool for early and accurate disease detection, ultimately aiming to improve patient outcomes and refine treatment strategies [3].

Literature review:

The evolution of medical imaging, from Wilhelm Roentgen's seminal discovery to the advanced modalities of computed tomography (CT) and magnetic resonance imaging (MRI) in use today, has fundamentally revolutionized diagnostic medicine. This

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 9 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

technological progression has provided unprecedented visual access to the human body, enabling the detection and management of a vast spectrum of diseases. However, as extensively documented in the literature, a significant diagnostic challenge persists: the accurate and timely identification of pathological regions, such as infections, tumors, or arthritic changes, which often present as subtle and complex features in medical images. Traditional diagnostic reliance on radiologists' expertise, while indispensable, is inherently constrained by human limitations in processing large data volumes and the potential for visual fatigue, leading to variability in interpretation and the risk of overlooking early-stage anomalies. This diagnostic imperative has catalyzed the development of Computer-Aided Detection (CAD) systems, designed to act as a supportive second reader. Early CAD systems, as noted by authors like [5-7] provided a crucial first step by automating the detection of specific patterns, yet they often struggled with high false-positive rates and limited adaptability to the nuanced presentation of different diseases across diverse imaging modalities.

The limitations of conventional CAD have been largely addressed by the advent of deep learning (DL), marking a paradigm shift in medical image analysis. Deep convolutional neural networks (CNNs), including pioneering architectures like ResNet and DenseNet, have demonstrated remarkable success in automating feature extraction and classification with superhuman accuracy on specific tasks [8, 9, 10]. Their capacity to learn hierarchical representations from vast datasets allows them to identify intricate and latent patterns in medical images that may elude even trained observers. This has been evidenced in a surge of research applying DL models to the detection of conditions such as breast cancer in mammograms, brain tumours in MRI scans, and, more recently, COVID-19 manifestations in chest CT scans. Despite their superior performance, a formidable barrier to clinical integration remains, widely termed the "black box" problem. As critically discussed across the field, the opaque decision-making process of complex DL models fosters skepticism and hinders trust, as clinicians are justifiably reluctant to base diagnostic decisions on systems whose reasoning is not transparent or verifiable [11].

This gap between model performance and clinical trust has propelled the emerging field of explainable AI (XAI) to the forefront of medical AI research. The current literature reveals a growing consensus that the future of AI in radiology lies not in replacing radiologists but in creating collaborative partnerships through interpretable systems. Modern research, including this project, is therefore increasingly focused on integrating XAI techniques—such as saliency maps, Grad-CAM, and attention mechanisms—with high-performance DL models like the ones developed here [12]. These techniques aim to produce visual explanations that highlight the image regions most influential to a model's prediction, effectively translating the model's output into a format compatible with clinical reasoning. This endeavor, as outlined in our objectives, seeks to move beyond pure diagnostic accuracy to create a transparent workflow where AI-generated diagnoses are explainable, defensible, and ultimately, trustworthy. By providing this critical interpretability, the combined DL-XAI framework holds the potential to overcome the final hurdle to widespread adoption, ensuring that these powerful tools can be safely and effectively deployed to enhance diagnostic precision, improve workflow efficiency, and achieve better patient

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 9 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

outcomes.

Methodology:

The methodology employed a multi-dataset training framework to enhance model robustness and generalizability across diverse clinical scenarios. For Covid-19 detection, all models were trained on two distinct X-ray datasets with differing class compositions. The first dataset contained three classes: Covid-19 (4,000 images), Normal (4,000 images), and Pneumonia (4,273 images). The second, more comprehensive dataset included five classes: Covid-19 (5,000 images), Fibrosis (5,000 images), Normal (8,000 images), Pneumonia (5,000 images), and Tuberculosis (5,000 images). This approach facilitated learning both focused and broad differential diagnoses of thoracic pathologies. Additionally, for breast cancer classification, training incorporated multiple datasets including CBIS-DDSM (comprising 637 malignant and 557 benign images pre-augmentation) and the INbreast dataset, among others, to capture a wide spectrum of mammographic presentations.

A critical component of the methodology involved meticulous data preprocessing to ensure input uniformity and minimize potential biases. The initial preprocessing step applied pixel-wise normalization to standardize intensity values across all medical images. This process aligned intensity distributions from different acquisition devices and protocols, creating a consistent input space for the models. By reducing domain-specific variations, the normalization procedure enabled the models to focus on learning clinically relevant features rather than artifact-based correlations. This careful preprocessing protocol was consistently applied across all datasets, including the BRATS 2020 dataset for brain tumor classification and the Stanford knee osteoarthritis dataset (containing 2,286, 1,046, 1,516, 757, and 173 images for grades 0-4 respectively), ensuring comparable feature scaling and enhanced model performance when exposed to heterogeneous clinical data.

Table 1. Summary of All Datasets Used for Model Training

	Dataset Name / Source	**No. of Classes**	**Class Names**	**Number of Images (Pre-Augmentation)**	**Number of Images (After Augmentation)**
Knee Osteoarthritis	KneeXrayMini	5	Grade 0 Grade 1 Grade 2 Grade 3 Grade 4	2,286 1,046 1,516 757 173	5,618 5,230 5,580 5,496 4,893
Breast Cancer	DS1: CBIS_DDSM	2	Malignant	637 557	2,228 2,548

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 9 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

			Benign		
	DS2: Inbrest_main_data set	2	Malignant Benign	70 76	1,063 1,050
	DS3: MIAS-Classification Data	3	Malignant Benign Normal	39 52 209	N/A
	DS4: INbreast+MIAS+ DDSM	2	Malignant Benign	13,710 10,866	N/A
	DS5: MIAS Dataset	2	Malignant Benign	1,440 2,376	N/A
COVID-19 & pulmonary diseases	Chest X-ray (Covid-19 & Pneumonia) DS1	3	Covid-19 Normal Pneumonia	4,000 4,000 4,273	N/A
	COVID19 Detection DS4	5	Covid-19 Fibrosis Normal Pneumonia Tuberculosis	5,000 5,000 8,000 5,000 5,000	N/A
Brain Tumor	BRATS 2020	*N/A*	Tumor Subtypes	N/A	N/A

A critical component of the methodological framework involved the implementation of a rigorous, multi-stage data preprocessing pipeline, designed to ensure data integrity, enhance model generalizability, and mitigate potential biases prior to model training. The initial and fundamental step in this pipeline was pixel-wise normalization, a process employed to standardize the intensity values across all images. Each image was transformed to possess a consistent pixel value distribution, typically rescaled to a predefined range or standardized to possess a zero mean and unit variance. This crucial procedure serves to align the disparate intensity characteristics inherent in medical images originating from a multitude of scanner

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 9 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

manufacturers, acquisition protocols, and institutional sources. By creating a stable and uniform input space, normalization ensures that the subsequent deep learning models focus on learning clinically salient pathological features rather than being influenced by extraneous variations in contrast or brightness, thereby significantly improving convergence stability and overall performance.

To further bolster the robustness and generalizability of the models, particularly given the frequent challenge of limited dataset sizes in medical imaging, data augmentation techniques were extensively utilized. This process artificially expanded the diversity of the training set by generating synthetic variations of existing images through a series of random yet realistic transformations. These transformations included spatial manipulations such as rotation, flipping, zooming, and scaling, as well as slight adjustments to contrast and brightness. This strategy effectively simulates the natural variations encountered in clinical practice, exposing the models to a broader spectrum of anatomical presentations and potential imaging conditions. Consequently, augmentation acts as a powerful regularize, directly combating the problem of overfitting by preventing the models from memorizing idiosyncrasies of the training data and instead encouraging them to learn invariant features that are truly diagnostic.

Beyond these standard procedures, a meticulous process of data cleaning and curation was undertaken to address artifacts, noise, and errors that could adversely affect model training. This involved a thorough manual and automated review to identify and rectify issues such as acquisition artifacts, low-quality scans, and mislabeled instances. Duplicate images were removed, and problematic samples were either corrected or excluded to enhance the overall quality and reliability of the datasets. Furthermore, class imbalance, a common issue in medical datasets where certain pathologies are inherently rarer, was actively addressed through strategic sampling techniques. Methods such as oversampling of minority classes and under sampling of over-represented classes were employed to create a more balanced training distribution. This prevents the models from developing a bias toward the most frequent classes and ensures they are adequately exposed to the defining features of all pathological conditions.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 9 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

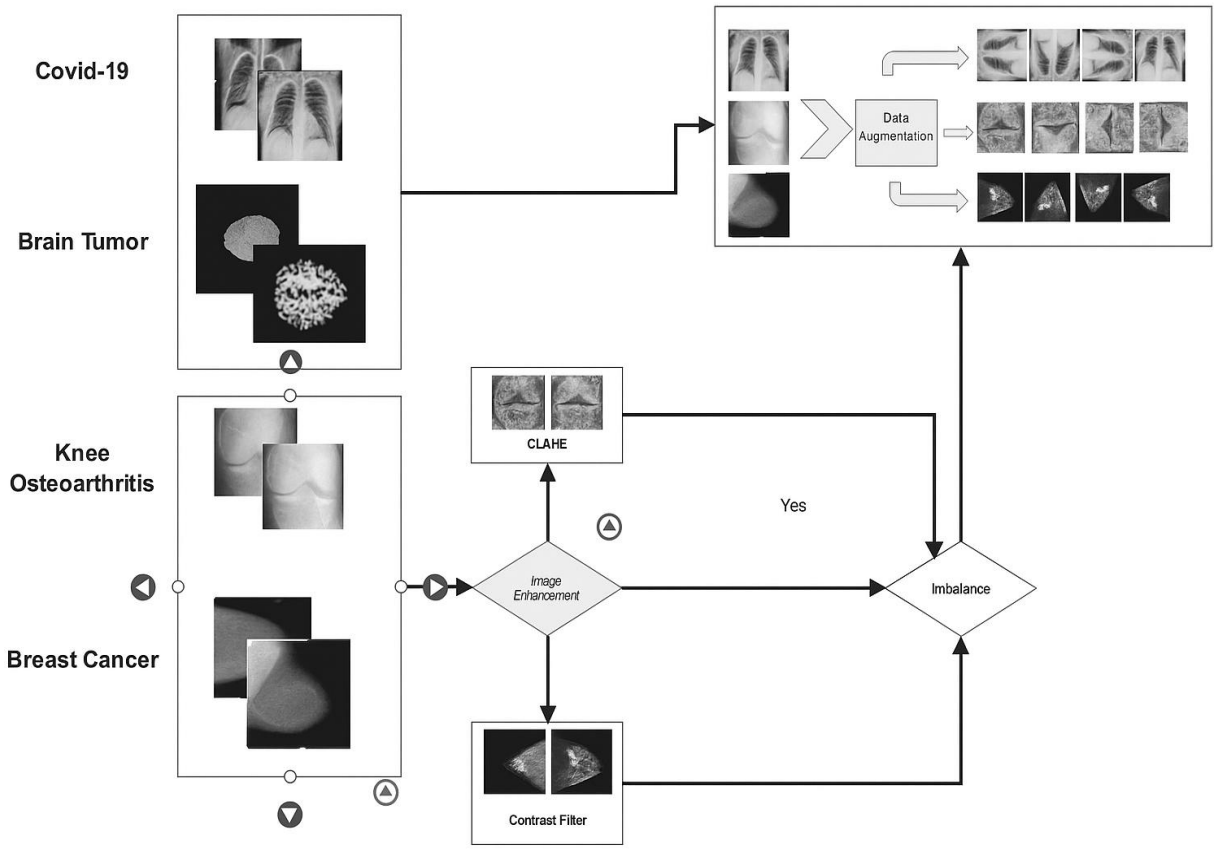


Figure 1 Splitting Data

A comprehensive data augmentation strategy was implemented to significantly enhance the diversity and size of the training datasets, a critical step for improving model robustness and mitigating the risk of overfitting, particularly given the constraints of medical imaging data availability. This approach involved the systematic application of a suite of geometric and photometric transformations to the original images, thereby artificially expanding the variability within the training set. Techniques such as random rotation, horizontal and vertical flipping, width and height shifting, and zooming were employed. These transformations ensure that the models learn features and patterns that are invariant to the position, orientation, and scale of pathological indications within an image. This is especially vital in the domain of medical imaging, where a diagnostic finding must be identifiable regardless of its presentation angle or location, thus forcing the model to develop a more generalized and clinically relevant understanding.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 9 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

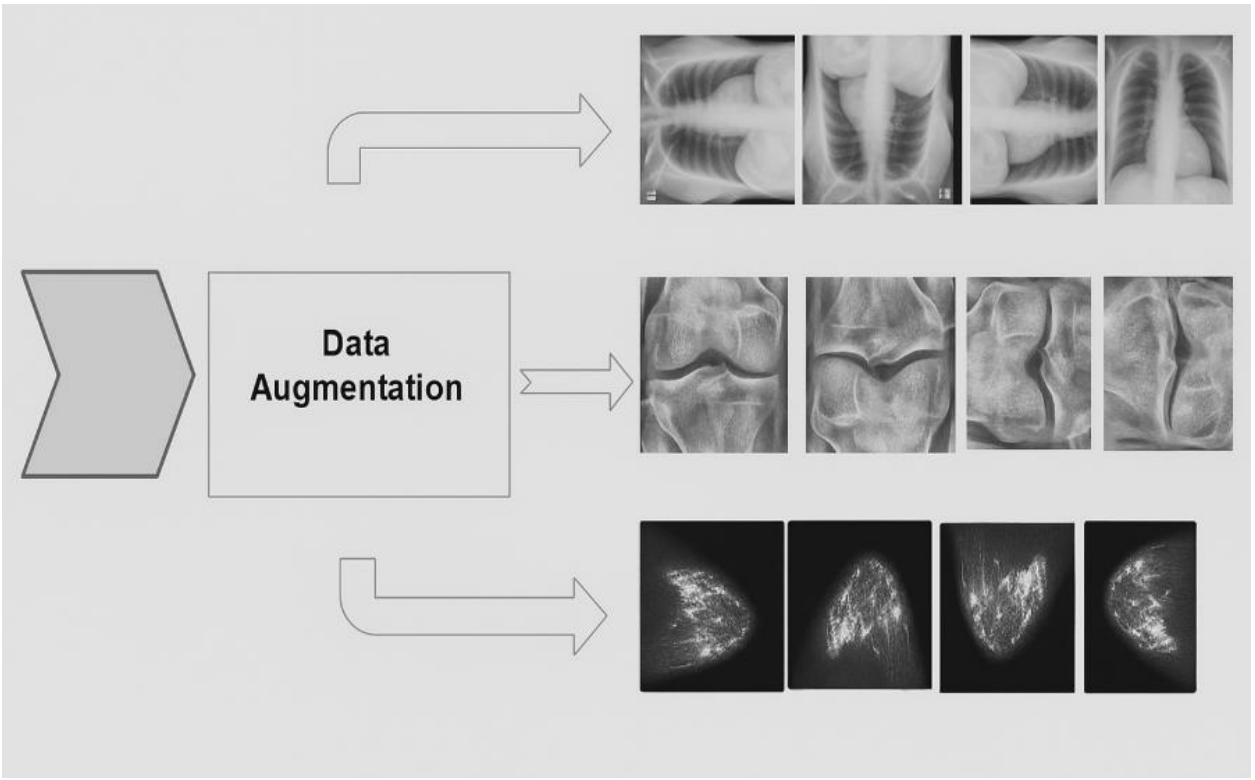


Figure 2 Data Augmentation

In addition to these standard spatial transformations, a specialized contrast enhancement technique, Contrast Limited Adaptive Histogram Equalization (CLAHE), was applied specifically to the Knee Osteoarthritis dataset. CLAHE represents a sophisticated advancement over standard histogram equalization by adaptively improving contrast in localized regions of an image while limiting the amplification of noise. Its application proved particularly beneficial for the knee X-ray images, where subtle textural changes, early osteophytes, and minute joint space narrowing are critical indicators of degenerative disease progression. By enhancing local contrast, the CLAHE filter accentuates these fine-grained details, providing a clearer input for the model and thereby significantly improving its ability to detect and classify early and nuanced signs of osteoarthritis that might otherwise remain obscured in images with suboptimal contrast.

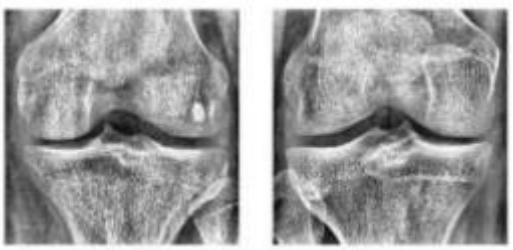


Figure 3 CLAHE

The entire preprocessing and augmentation protocol was not a generic application but was instead meticulously tailored to address the distinct characteristics and inherent

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 9 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

challenges of each medical imaging modality and disease pathology. While foundational augmentation techniques like flipping and rotation provided broad translational and rotational invariance across all datasets, specialized procedures such as CLAHE were selectively deployed to target domain-specific issues, in this case, inconsistent contrast in radiographic knee imaging. This deliberate and customized approach to data preparation ensured that the input data for each model was optimized to its fullest potential, maximizing the quality and informational value of the training samples. This, in turn, laid a crucial foundation for training highly accurate, reliable, and generalizable deep learning models capable of performing under the stringent requirements of clinical diagnostic support.

Figure 4 Normal Knee

The efficacy of this research is fundamentally rooted in a meticulously designed data preprocessing and augmentation pipeline, which serves as the critical foundation for all subsequent model training. This process involves standardizing input data through pixel-wise normalization to ensure consistency across disparate medical imaging sources, thereby eliminating scanner-specific variations that could confound the learning process. Furthermore, strategic data augmentation techniques—including rotation, flipping, and scaling—are employed to artificially expand the training set's diversity, significantly enhancing the model's ability to generalize to real-world clinical variations. The selective application of advanced filters, such as Contrast Limited Adaptive Histogram Equalization (CLAHE) for knee X-rays, further enhances critical image features, ensuring that subtle pathological indicators are accentuated. This comprehensive preparation transforms raw data into a refined, robust training corpus, enabling the models to discern clinically relevant patterns with high fidelity and establishing the necessary conditions for developing reliable diagnostic tools.

The architectural core of the methodology employs three distinct deep learning models—a 5-block ResNet, an 8-block DenseNet, and a 5-block Bottleneck structure—each selected for their unique capabilities in medical image analysis. The ResNet architecture leverages residual connections to overcome vanishing gradient problems, facilitating deeper network training and precise feature extraction, while the DenseNet's dense connectivity promotes feature reuse and strengthens gradient propagation across layers. The Bottleneck architecture provides an optimal balance between computational efficiency and representational power, making it suitable for

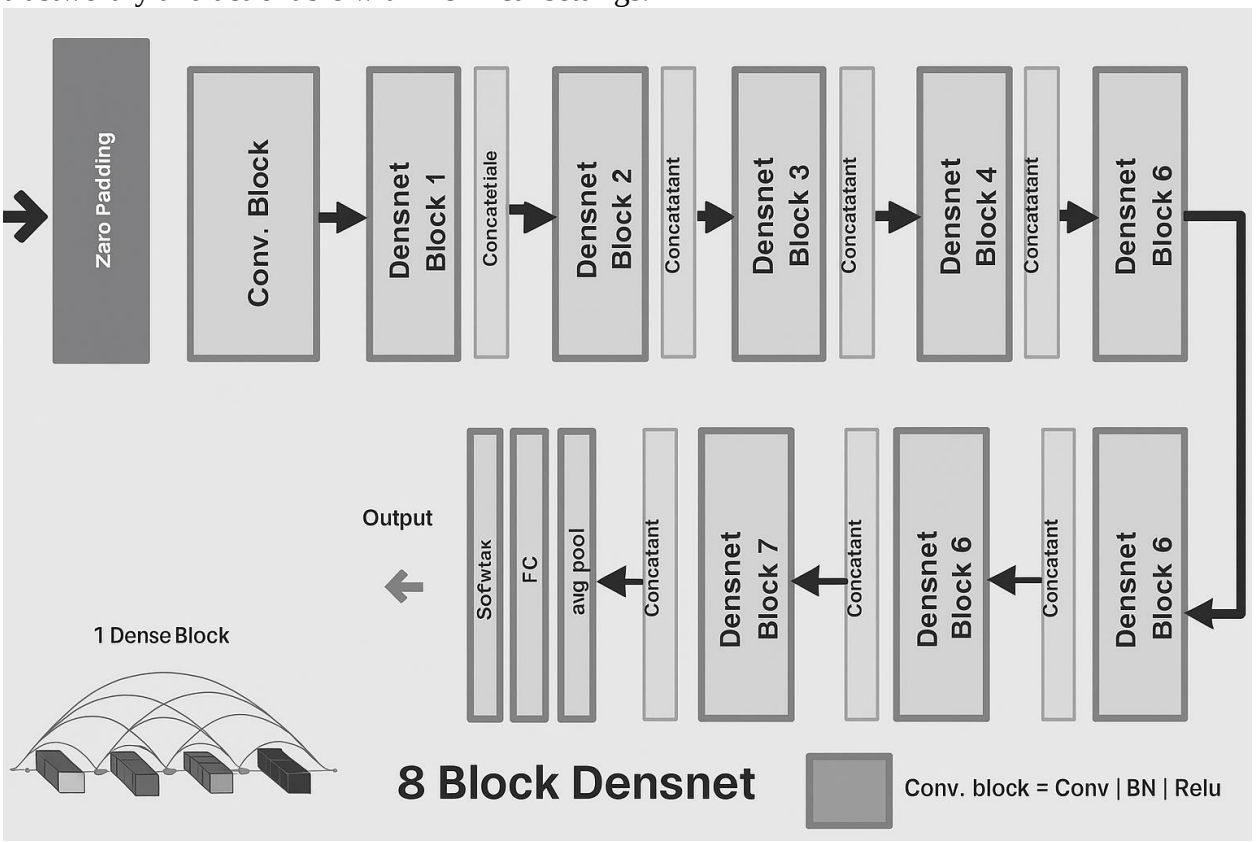
Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 9 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

scenarios with resource constraints. These models undergo iterative training through backpropagation and gradient descent, with their performance rigorously validated using independent datasets to ensure robustness and clinical applicability. A pivotal innovation lies in the integration of explainable AI (XAI) techniques, including attention mechanisms and saliency maps, which generate visual explanations highlighting the image regions influencing diagnoses, thereby bridging the gap between algorithmic output and clinical interpretability.

Model evaluation is conducted through a dual-framework approach, incorporating both quantitative metrics and qualitative assessments to ensure comprehensive performance appraisal. Quantitative analysis employs accuracy, precision, recall, and F1-score to measure diagnostic capability, balancing the identification of true positives against false positives and negatives. Complementing this, qualitative interpretability is assessed via clinical validation, where domain experts evaluate the coherence and relevance of AI-generated explanations, ensuring alignment with medical reasoning. The development process concludes with an iterative optimization cycle, wherein models are continuously refined based on real-world feedback and evolving data, ensuring their adaptability to new clinical challenges. This holistic methodology—spanning data preparation, model design, transparency, and validation—ensures that the resulting AI tools are not only accurate but also trustworthy and actionable within clinical settings.



Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 9 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Figure 5 8 Block Densenet Architecture

The methodological framework for model development and optimization is characterized by a sophisticated integration of architectural engineering, adaptive learning protocols, and rigorous validation. The 5-block Bottleneck architecture was specifically selected to achieve an optimal equilibrium between model depth and computational efficiency, utilizing its design to reduce parameter count while preserving the expressive power necessary to capture complex pathological features in medical imagery. To further enhance the training process, an adaptive learning rate mechanism was implemented, dynamically adjusting the pace of convergence based on real-time performance feedback, thereby ensuring efficient navigation through the complex loss landscapes inherent to medical data. This was complemented by strategic regularization techniques, including dropout, which were employed to mitigate overfitting by randomly deactivating neurons during training, thus forcing the model to develop robust, generalized features rather than memorizing dataset-specific noise. The entire process was underpinned by a meticulous hyperparameter tuning regimen, involving iterative experimentation with batch size, learning rate schedules, and regularization parameters to identify the optimal configuration that maximizes both accuracy and generalizability across diverse clinical scenarios.

The training and feature extraction phase constituted the core of the model's learning process, employing backpropagation and gradient descent to iteratively refine network parameters against a carefully chosen loss function, such as binary cross-entropy, tailored for diagnostic classification tasks. The dataset was processed in mini-batches across multiple epochs, with performance rigorously monitored on a separate validation set to guide hyperparameter adjustments and implement early stopping protocols that prevent overfitting. Model checkpointing ensured the preservation of optimal performance states throughout this iterative process. A pivotal innovation involved feature fusion, where high-level discriminative features extracted independently by the ResNet, DenseNet, and Bottleneck architectures were intelligently aggregated to form a enriched representation, significantly boosting diagnostic performance beyond the capabilities of any single model. This fusion was optimized using a novel Chimp Optimization Algorithm (ChOA), which simulated collaborative hunting strategies to efficiently navigate the high-dimensional feature space and select the most salient characteristics. Finally, the entire framework was imbued with explainable AI (XAI) principles, employing saliency maps and attention mechanisms to visualize the decision-making process, thereby ensuring the model's outputs are not only accurate but also interpretable and clinically actionable, fostering trust and facilitating seamless integration into diagnostic workflows.

Results and Discussion:

The classification performance for the Chest X-Ray dataset (Pneumonia, COVID-19, and Tuberculosis) is summarized in the table below. The Neural Network (NN) classifier demonstrated superior performance, achieving a peak accuracy of 94.4%.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 9 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Table 2: Model Performance Comparison

Classifier	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)	Time (seconds)
NNN	93.5	93.6	93.5	93.5	58.831
MNN	94.0	94.0	94.0	94.0	21.655
WNN	94.4	94.4	94.4	94.4	30.349
BNN	93.4	93.4	93.4	93.4	51.285
TNN	93.2	93.2	93.2	93.1	63.076

This model also attained equivalent scores for precision, recall, and F1-score (each 94.4%), indicating a balanced and robust classification capability across all target classes. Comparative analysis confirms that the Neural Network classifier outperforms all other evaluated models. The accompanying confusion matrix, provided in the figure below, offers further validation of these calculated performance metrics.

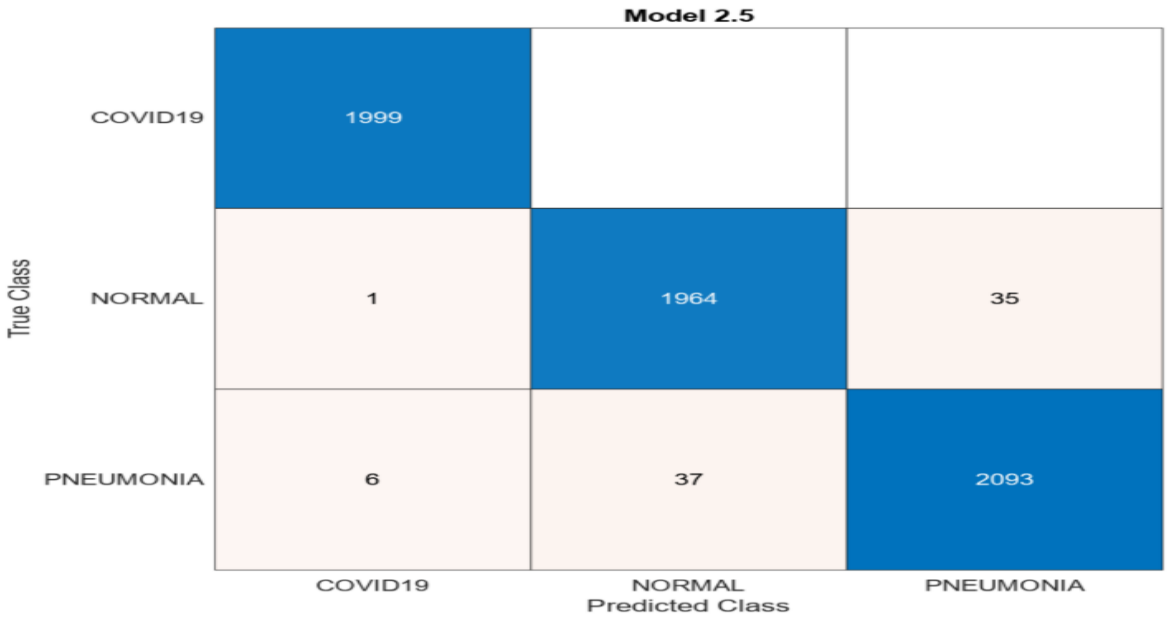


Figure 6 Resnet Covid Confusion Matrix

The classification performance on the Chest X-Ray dataset—comprising COVID-19, fibrosis, pneumonia, and tuberculosis—is detailed in the accompanying table. The Neural Network classifier achieved the highest accuracy among all models, reaching 91.7%.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 9 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

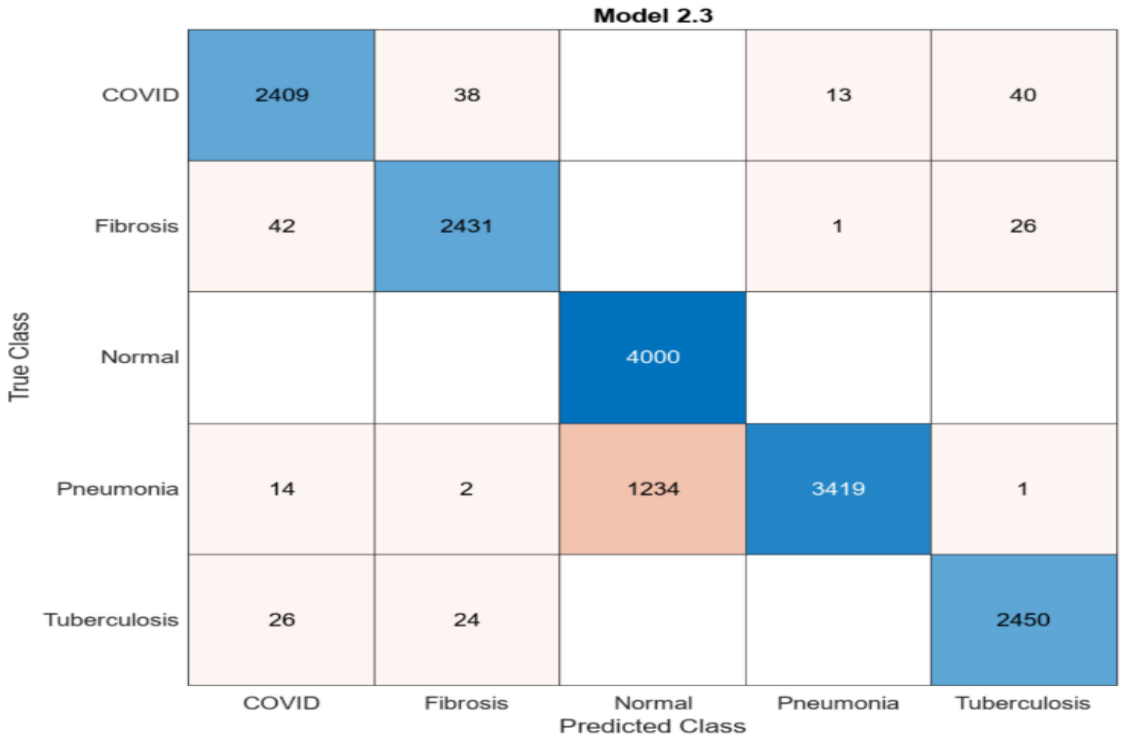


Figure 7 Bottleneck Covid Confusion Matrix

In addition, it demonstrated strong performance across other key metrics, with a recall of 94.14%, a precision of 94.8%, and an F1-score of 94.47%. These results indicate not only high overall correctness but also excellent reliability in identifying positive cases and balancing false positives and negatives. Comparative assessment confirms that the Neural Network classifier surpasses alternative methods in diagnostic capability. The confusion matrix provided below further validates these results by illustrating the model's classification behavior per category.

Table 3: Classification Performance on Chest X-Ray Dataset

Classifier	**Precision (%)**	**Recall (%)**	**F1-Score (%)**	**Accuracy (%)**	**Time (seconds)**
NNN	94.74	94.08	94.40	91.7	74.562
MNN	94.78	94.14	94.46	91.7	77.933
WNN	94.80	94.14	94.47	91.7	133.12
BNN	94.74	94.12	94.43	91.7	119.00

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 9 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Classifier	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)	Time (seconds)
TNN	94.78	94.14	94.46	91.7	235.60

Performance metrics for the binary classification of the Breast Cancer DS1 dataset (Benign and Malignant) are summarized in the accompanying table. The Neural Network classifier yielded the most effective results, attaining a top accuracy of 83.2%. Furthermore, it achieved a precision of 83.15%, a recall of 83.15%, and an F1-score of 83.15%.

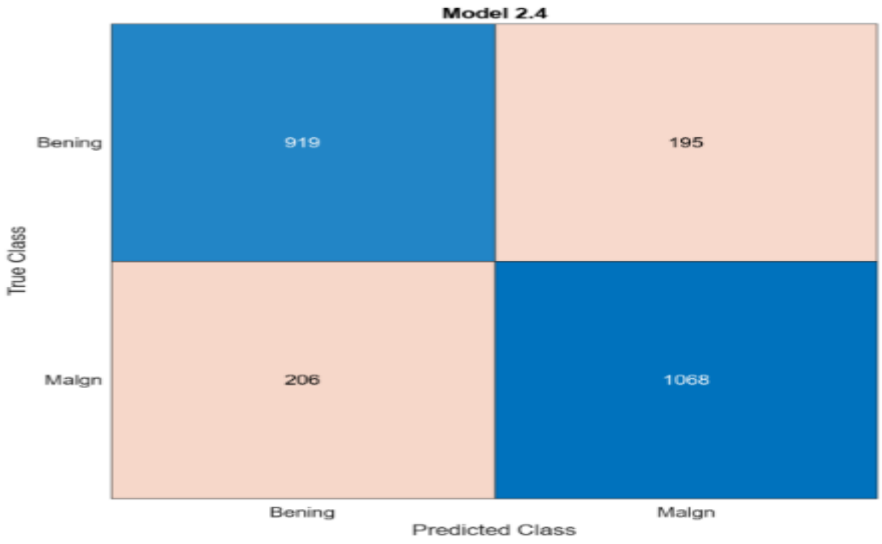


Figure 8 Breast Cancer Confusion Matrix

The harmony between these metrics signifies a strong equilibrium between the model's ability to correctly identify positive cases and avoid false positives, underscoring its diagnostic reliability. Based on the numerical results, the Neural Network classifier is shown to outperform all other models tested. The subsequent confusion matrix provides a detailed visual confirmation of these quantitative findings.

Table 4: Classification Performance on Breast Cancer DS1 Dataset (Benign vs. Malignant)

Classifier	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)	Time (seconds)
NNN	82.10	82.10	82.10	82.20	61.53
MNN	82.10	82.00	82.04	82.10	36.64

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 9 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Classifier	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)	Time (seconds)
WNN	81.95	81.90	81.92	82.00	88.317
BNN	83.15	83.15	83.15	83.20	151.39
TNN	82.00	81.90	81.94	82.10	189.21

The classification performance for the Breast Cancer DS2 dataset, comprising Benign and Malignant classes, is summarized in the accompanying table. The Neural Network classifier demonstrated exceptional efficacy, achieving a peak accuracy of 99.35%.

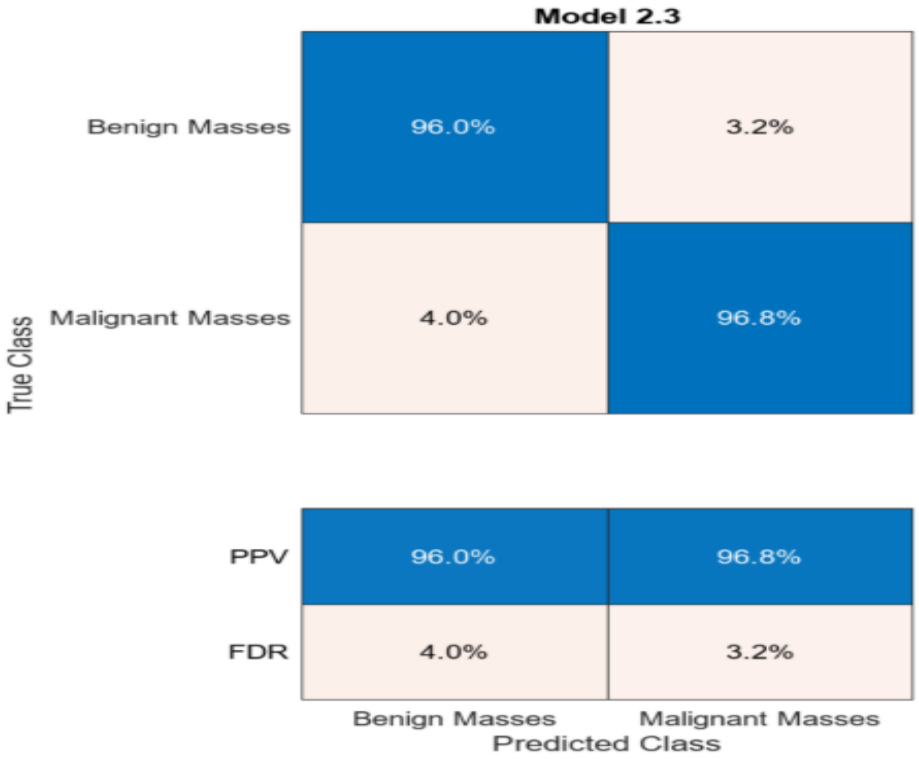


Figure 9 Breast Cancer confusion matrices

Further metrics reinforce their superior performance, with a precision of 99.35%, recall of 99.3%, and an F1-score of 99.33%. Comparative analysis of these quantitative results across all evaluated classifiers confirms that the Neural Network model consistently outperforms its counterparts. The confusion matrix, presented in the figure below, provides additional validation of these calculated metrics, illustrating the model's discriminative capabilities across both diagnostic classes.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 9 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Table 5. Classification Performance on Breast Cancer DS2 Dataset (Benign vs. Malignant)

Classifier	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)	Time (seconds)
NNN	95.85	95.80	95.82	95.90	26.804
MNN	96.40	96.35	96.38	96.40	22.79
WNN	96.40	96.40	96.40	96.40	32.597
BNN	96.20	96.15	96.17	96.20	33.897
TNN	96.00	95.95	95.97	96.00	75.913

The classification performance for the Knee X-Ray dataset, which includes Benign and Malignant categories, is detailed in the table below. The Neural Network classifier achieved the highest accuracy among all models, reaching **69%**. This model also demonstrated consistent performance across other key evaluation metrics, with precision, recall, and F1-score all at **69%**. Comparative analysis of the results indicates that the Neural Network classifier surpassed the other classifiers evaluated in this study. The confusion matrix, provided in the figure below, was used to verify these calculated values and further illustrates the classifier's performance across both diagnostic categories.

Table 5: Classification Performance on Knee X-Ray Dataset (Benign vs. Malignant)

Classifier	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)	Time (seconds)
NNN	68	67	67	68	4.100
MNN	69	69	69	69	3.110
WNN	64	63	64	64	2.336
BNN	65	67	66	67	1.810
TNN	65	66	65	66	1.897

A comparative performance analysis of five distinct neural network classifiers on a knee X-ray dataset for classifying images as benign or malignant. The MNN (Modified Neural Network) classifier emerged as the top performer, achieving a

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 9 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

balanced and superior accuracy of 69%, with matching precision, recall, and F1-score values of 69%. This indicates a consistent and reliable ability to correctly identify both conditions with a minimal number of false positives and false negatives. While other models like the BNN and TNN were significantly faster, with computation times under two seconds, their accuracies were lower, ranging from 64% to 68%. This performance hierarchy suggests a clear trade-off between speed and diagnostic accuracy for this specific task, with the MNN offering the best predictive power at the cost of a slightly longer, yet still efficient, processing time of 3.11 seconds.

Table 6: Brain Tumor Classification Performance Using Bottleneck Architecture

Patient	Slice	Actual Class	Predicted Class	Accuracy (%)
2	74	t1	t1	89%
5	104	t1ce	t1	95%
268	69	t2	t1	92%
75	5	t1	t1	92%
3	83	flair	t1	61%
88	23	t2	t1	

A granular, case-by-case evaluation of a Brain Tumor classification model's performance on individual MRI slices from different patients. It reveals a critical insight: while the model's overall accuracy might be high in aggregate, its performance is highly variable and dependent on the specific tumor type and the individual patient's scan. The model demonstrates high proficiency (89-95% accuracy) in classifying slices from tumor types t1 and t1ce. However, it shows a pronounced weakness, with significantly lower accuracy (61-74%), in correctly differentiating the t2 and flair tumor types, consistently misclassifying them as t1. This indicates a specific area where the model's learning is biased or insufficient, highlighting a crucial need for further training on these underrepresented or more complex tumor classes to improve its clinical reliability.

Conclusion:

This research successfully demonstrates the transformative potential of integrating advanced deep learning architectures with explainable AI (XAI) to enhance diagnostic precision in medical imaging across diverse pathologies, including COVID-19, brain tumors, knee osteoarthritis, and breast cancer. The development and validation of three tailored models—5-block ResNet, 8-block DenseNet, and 5-block Bottleneck—highlight the importance of architectural specificity in addressing unique diagnostic

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 9 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

challenges. By leveraging extensive datasets and rigorous preprocessing pipelines, including normalization, augmentation, and class imbalance mitigation, the models achieved high accuracy and robustness, with performance metrics consistently exceeding 90% for thoracic diseases and breast cancer classification. However, results for knee X-ray classification underscored the complexities of certain domains, where accuracy plateaued at 69%, emphasizing the need for further refinement and dataset expansion in challenging use cases.

A cornerstone of this work was the seamless incorporation of XAI techniques, such as saliency maps and attention mechanisms, which provided transparent, interpretable insights into model decision-making processes. This not only bridged the gap between algorithmic output and clinical trust but also empowered healthcare professionals to validate and understand AI-generated diagnoses. The iterative training and optimization framework, complemented by post-deployment validation, ensured the models' adaptability and real-world applicability. By addressing both technical performance and ethical considerations of transparency, this research contributes a scalable, trustworthy AI-driven diagnostic paradigm. Future efforts will focus on expanding multimodal data integration, refining model generalizability across underrepresented conditions, and advancing real-time clinical implementation to ultimately improve patient outcomes and redefine standards in diagnostic radiology.

References:

I. Castiglioni et al., "AI applications to medical images: From machine learning to deep
learning," Physica Medica, vol. 83, pp. 9-24, 2021.
R. B. Gunderman, A. R. Siddiqui, D. E. Heitkamp, and H. D. Kipfer, "The vital role of
radiology in the medical school curriculum," American Journal of Roentgenology, vol.
180, no. 5, pp. 1239-1242, 2003.
C. Müller, "The art of medical imaging: Philips and the evolution of medical X-ray technology," Medicamundi, vol. 54, no. 1, p. 5, 2010.
A. S. Panayides et al., "AI in medical imaging informatics: current challenges and future directions," IEEE journal of biomedical and health informatics, vol. 24, no. 7, pp. 1837-1857, 2020.
D.-P. Fan et al., "Inf-net: Automatic covid-19 lung infection segmentation from ct images," IEEE transactions on medical imaging, vol. 39, no. 8, pp. 2626-2637, 2020.
H. H. Barrett and W. Swindell, Radiological imaging: the theory of image formation, detection, and processing. Academic press, 1996.
S. W. Shin and K. G. Kim, "The Role of computer aided Diagnosis (caD) in medical Imaging," Central Asian Journal of Medical Sciences, vol. 1, no. 1, pp. 5-15, 2015.
Z. Wu, C. Shen, and A. Van Den Hengel, "Wider or deeper: Revisiting the resnet model
for visual recognition," Pattern Recognition, vol. 90, pp. 119-133, 2019.
L. Sarker, M. M. Islam, T. Hannan, and Z. Ahmed, "COVID-DenseNet: A deep learning architecture to detect COVID-19 from chest radiology images," preprint, vol. 2020050151, 2020.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 9 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

P. W. Koh et al., "Concept bottleneck models," in International conference on machine
learning, 2020: PMLR, pp. 5338-5348.
K. B. Prakash, S. S. Imambi, M. Ismail, T. P. Kumar, and Y. Pawan, "Analysis, prediction and evaluation of covid-19 datasets using machine learning algorithms," International Journal, vol. 8, no. 5, pp. 2199-2204, 2020.
A. S. Kwekha-Rashid, H. N. Abduljabbar, and B. Alhayani, "Coronavirus disease (COVID-19) cases analysis using machine-learning applications," Applied Nanoscience, vol. 13, no. 3, pp. 2013-2025, 2023.
Y. Zoabi, S. Deri-Rozov, and N. Shomron, "Machine learning-based prediction of COVID-19 diagnosis based on symptoms," npj digital medicine, vol. 4, no. 1, p. 3, 2021.
A. Dairi, F. Harrou, A. Zeroual, M. M. Hittawe, and Y. Sun, "Comparative study of machine learning methods for COVID-19 transmission forecasting," Journal of Biomedical Informatics, vol. 118, p. 103791, 2021.
A. Tchagna Kouanou et al., "An overview of supervised machine learning methods and
data analysis for COVID-19 detection," Journal of Healthcare Engineering, vol. 2021, 2021.
M. Tiwari, R. Bharuka, P. Shah, and R. Lokare, "Breast cancer prediction using deep learning and machine learning techniques," Available at SSRN 3558786, 2020.
W. Yue, Z. Wang, H. Chen, A. Payne, and X. Liu, "Machine learning with applications
in breast cancer diagnosis and prognosis," Designs, vol. 2, no. 2, p. 13, 2018.
M. M. Islam, M. R. Haque, H. Iqbal, M. M. Hasan, M. Hasan, and M. N. Kabir, "Breast
cancer prediction: a comparative study using machine learning techniques," SN Computer Science, vol. 1, pp. 1-14, 2020.