Explainable CNN–Transformer Hybrids for Multiclass Lung Disease Classification in Chest X-Rays: A Survey and Future Research Directions

V Saranya[1], Deepika S.[2], Dharshini S.[3], Swetha R.[4], Thulasimani V.[5]
[1]*Assistant Professor of Computer Science Department, Sri Venkateshwaraa College of Engineering and Technology Ariyur, Puducherry*
[2,3,4,5]*Student, Department of Computer Science and Engineering, Sri Venkateshwaraa College of Engineering and Technology, Ariyur Puducherry, India*

***Abstract*—Chest X-ray (CXR) imaging is widely used to support the assessment of lung diseases, yet interpretation can be difficult because of overlapping anatomical structures, subtle abnormalities, and variation in image acquisition. Deep-learning methods have therefore been investigated for automated CXR classification. Convolutional neural networks (CNNs) learn local visual patterns, whereas vision transformers (ViTs) model relationships among image regions. Hybrid CNN–transformer architectures seek to combine these complementary representations. This narrative survey reviews CNN-based classifiers, segmentation-assisted and ensemble methods, hybrid architectures, image-enhancement techniques, and explainable artificial intelligence (XAI) approaches for CXR analysis. It also discusses commonly used datasets, evaluation metrics, reproducibility, class imbalance, data leakage, domain shift, and the limitations of visual explanations. Future research directions include efficient deployment, uncertainty estimation, multimodal learning, privacy-preserving multi-site training, and external clinical validation. Because datasets, class definitions, and evaluation protocols differ across studies, reported performance values should be interpreted in context rather than treated as directly comparable results.**

***Index Terms*—chest X-ray; lung disease classification; convolutional neural networks; vision transformers; feature fusion; explainable artificial intelligence; Grad-CAM; CLAHE; narrative survey.**

I. INTRODUCTION

Lung diseases, including pneumonia and COVID-19, contribute substantially to the global burden of disease [1]. Chest X-rays (CXRs) are relatively inexpensive and widely available, making them an important tool for assessing thoracic conditions. However, radiographic interpretation can be challenging because bones, blood vessels, the heart, and abnormal findings overlap in a two-dimensional image. Differences in reader experience and access to specialist radiologists have encouraged research into computer-aided diagnosis (CAD) systems [2].

Early deep-learning studies largely relied on convolutional neural networks (CNNs). CNNs are effective at learning local texture and shape patterns, but their standard operations may not capture long-range relationships as directly as attention-based methods. Vision transformers (ViTs) use self-attention to model relationships among image patches, while hybrid approaches combine convolutional and transformer representations. Explainability methods, including class-activation maps, can help users inspect regions that influence a prediction; however, a visually plausible heatmap is not, by itself, evidence of clinical validity.

The remainder of this paper is organised as follows. Section II introduces the background concepts. Section III surveys related work and commonly used datasets. Section IV presents future research directions, and Section V discusses practical recommendations and open challenges before the conclusion.

This survey synthesises recurring patterns in the literature and highlights evaluation practices needed to develop more reliable, transparent, and generalisable CXR classification systems.

II. BACKGROUND

A. Challenges in Chest X-Ray Interpretation

Three challenges recur in CXR analysis. First, low contrast can make subtle findings difficult to distinguish from normal lung texture. Second, different disease categories can have similar appearances; for example, pneumonia and other forms of lung opacity may overlap visually. Third, images vary across scanners, acquisition settings, patient populations, and institutions. Consequently, a model that performs well on one dataset may perform less reliably in another setting.

B. Contrast Enhancement with CLAHE

Contrast-limited adaptive histogram equalization (CLAHE) [4] improves local contrast by dividing an image into tiles and equalising the intensity histogram within each tile, subject to a contrast limit. It may make some structures more visible, but excessive enhancement can amplify noise or create artificial patterns.

CLAHE is a relatively inexpensive preprocessing option. Its value should be measured on the target dataset using a controlled comparison, with preprocessing applied consistently and without allowing information from evaluation data to influence model development.

C. Residual Networks and Vision Transformers

Residual networks (ResNets) use skip connections so that a block can learn a residual mapping, commonly expressed as $H(x) = F(x, W) + x$ [5]. ResNet-50 is often used as a CNN backbone to extract local visual features. A vision transformer divides an image into patches, embeds them as tokens, adds positional information, and applies self-attention to model relationships across the image [6]. These approaches provide complementary representations, but combining them does not automatically improve performance; the result depends on the data, fusion design, training strategy, and evaluation protocol.

D. Explainability

Gradient-weighted Class Activation Mapping (Grad-CAM) [7] uses gradients of a target-class score to weight convolutional feature maps and produce a coarse heatmap of influential regions. Grad-CAM++ modifies the weighting scheme, while Score-CAM uses changes in the model score associated with activation maps rather than gradients. Reviews of explainable AI (XAI) for diagnosis [8] emphasise that visual explanations can support model inspection, but an attractive heatmap is not necessarily faithful to the model or clinically meaningful.

E. Evaluation Metrics

Common evaluation metrics include accuracy, precision, recall, specificity, F1-score, and the area under the receiver operating characteristic curve (AUC). Accuracy measures the fraction of correctly classified images but can conceal poor performance on minority classes.

Recall measures the proportion of positive cases detected, while precision measures the proportion of predicted positives that are correct. Specificity measures the proportion of negatives correctly identified, and the F1-score balances precision and recall. For multiclass tasks, authors should state whether metrics are macro-, micro-, or weighted-averaged. Calibration should also be assessed because predicted confidence should correspond to observed correctness.

F. Transfer Learning

Because labelled medical images are often limited, many studies initialise models using weights pretrained on large natural-image collections and then fine-tune them on radiographs.

Transfer learning can accelerate convergence and reduce overfitting, but it introduces a domain gap: natural photographs differ from grayscale radiographs in texture, contrast, and semantic content. Self-supervised pretraining on large collections of chest X-rays is therefore a promising alternative that warrants further evaluation.

III. SURVEY OF RELATED WORK

The studies reviewed in this section use different datasets, class definitions, and evaluation splits. Their reported metrics should therefore be interpreted as study-specific findings rather than as entries in a directly comparable leaderboard.

Table I summarises selected approaches discussed in the source draft. Numerical performance values are withheld where they have not been independently verified against the original publications.

A. CNN-Based Classification

CNNs remain a widely used baseline for CXR classification. Sarkar et al. [9] presented a multi-scale CNN integrated with SHAP and Grad-CAM and reported strong benchmark performance; the exact numerical result should be checked against the original paper. Deepak and Bhat [10] investigated a multi-stage approach for classifying a broader set of lung conditions and reported accuracy and F1-score values in their study.

Mohan et al. [11] compared VGG16, Inception, ResNet, and a custom CNN for multiclass classification, while Mahamud et al. [12] studied fine-tuned transfer learning with explainability for multiple lung conditions. Although CNN-based methods can perform strongly on their development datasets, their generalisation may vary across institutions, and local convolutional features do not always represent global relationships explicitly.

B. Segmentation-Assisted Methods and Ensembles

Segmentation-assisted approaches attempt to focus classification on the lung region and reduce the influence of irrelevant image areas. Jangam et al. [14] investigated a multiclass framework for disease screening and diagnosis from chest radiographs. Ensembles can combine complementary predictions, but they may require more parameters, memory, and inference time, which can complicate deployment.

C. Hybrid CNN–Transformer Models

Ko et al. [15] investigated vision-transformer optimisation for lung-disease detection using chest X-rays. Chen et al. [16] evaluated a fine-tuned ViT-B/16 model on the COVID-19 Radiography Database and reported results for three- and four-class settings. Pham and Hoang [17] proposed CovEnViT, a CNN–ViT combination, while Deva and Dagur [3] described a ViT–ResNet fusion framework for multiclass CXR classification.

Fu et al. [18] introduced LungMaxViT, a hybrid architecture incorporating convolutional and multi-axis attention components, and reported results on COVID-19 and ChestX-ray14 datasets. Hadhoud et al. [19] investigated a two-stage ResNet-50 and ViT-B/16 pipeline for binary and multiclass tasks. The reported scores in these studies are task- and dataset-specific and should not be compared without accounting for differences in labels, splits, and evaluation procedures.

D. Preprocessing and Artifact Handling

A. Image enhancement and artifact handling are also active research directions. These methods aim to improve local contrast or reduce nuisance variation, but preprocessing can also introduce artifacts and must be evaluated using a controlled protocol.

Shamrat et al. [22] evaluated CLAHE and Gaussian filtering before MobileNetV2. Chehade et al. [23] investigated a CycleGAN-based preprocessing approach intended to reduce artifacts in ChestX-ray14. Nahiduzzaman et al. [24] proposed a parallel CNN and extreme-learning-machine framework for multiclass classification across numerous lung-disease categories.

Kumar et al. [21] described an enhanced multi-model deep-learning approach for pulmonary disease diagnosis. The exact performance figures and task definitions should be taken from the corresponding original publications rather than inferred from the titles. [21] combined histogram equalization with CNNs and report 98.88%. Shamrat et al. [22] applied CLAHE and Gaussian filtering ahead of MobileNetV2 and report 96.97%. Chehade et al. [23] used a CycleGAN to remove artifacts on ChestX-ray14 and improved the reported figure from 78.52% to 84.96% on a hard pneumonia-focused task. Nahiduzzaman et al. [24] paired a parallel CNN with an extreme learning machine for seventeen diseases.

TABLE I. Summary of representative studies. Reported metrics are task-specific and are not directly comparable.

Study	Approach	Reported result in source draft	Typical limitation
[9]	Multi-scale CNN with SHAP and Grad-CAM	Verify against the original publication before reporting a numerical result.	Generalisation requires external validation
[11]	Custom CNN compared with VGG16, Inception, and ResNet	Verify against the original publication before reporting a numerical result.	Dataset-specific result

[15]	Vision transformer optimised with RAdam	Verify against the original publication before reporting a numerical result.	External validation and explanation quality need assessment
[16]	Fine-tuned ViT-B/16 with Grad-CAM	Verify against the original publication before reporting a numerical result.	Results depend on task definition and split
[17]	CovEnViT: CNN–ViT combination	Verify against the original publication before reporting a numerical result.	Explanation evaluation requires further detail
[18]	LungMaxViT with convolutional and attention components	Verify against the original publication before reporting a numerical result.	Task-specific result
[19]	Two-stage ResNet-50 plus ViT-B/16	Verify against the original publication before reporting a numerical result.	Two-stage complexity
[22]	CLAHE and Gaussian filtering with MobileNetV2	Verify against the original publication before reporting a numerical result.	No transformer branch
[3]	CLAHE with ViT–ResNet fusion and explainability	Verify against the original publication before reporting a numerical result.	Verify evaluation protocol in original paper
[21]	Enhanced multi-model deep-learning approach	Verify against the original publication before reporting a numerical result.	Result and task definition need source-level confirmation

E. Gaps Motivating Hybrid and Explainable Designs

Three gaps recur across the reviewed literature. First, some high-performing models provide limited evidence that their explanations are faithful or clinically meaningful. Second, complex architectures can be costly to train and deploy. Third, relatively few studies evaluate image enhancement, hybrid feature extraction, and explanation methods together under a consistent multiclass protocol with external validation. These gaps motivate methods that balance predictive performance, computational cost, reproducibility, and clinically meaningful explanations.

F. Datasets Used in the Literature

Table II summarises datasets mentioned in the reviewed studies. Public collections may aggregate images from multiple hospitals or online sources, creating potential source-specific artifacts and label inconsistencies.

Class definitions and labelling criteria can also differ between studies, even when the same disease names are used. Dataset size and class composition should therefore be verified against the version actually used in each paper.

TABLE II. Datasets frequently used in the surveyed studies.

Dataset	Classes / content as described in the source draft	Studies cited in the source draft
COVID-19 Radiography Database (approximately 21,000 images)	COVID-19, lung opacity, viral pneumonia, and normal	[16], [3]
ChestX-ray14 (112,120 images)	14 thoracic findings	[24], [23]
COVID-QU-Ex	COVID-19, non-COVID / infection, and normal	[18]
Curated CXR set (9,209 images)	Normal, COVID-19, bacterial pneumonia, and viral pneumonia	[17]
NIH and SCH collections	Normal, pneumonia, pneumothorax, and tuberculosis	[25]
COVID-Pneumonia-Normal (Kaggle)	COVID-19, pneumonia, and normal; external test set as described in draft	[26]

G. Comparison of Explanation Techniques

Table III compares commonly used XAI techniques. None should be treated as a universal gold standard. Gradient-based class-activation methods are relatively

fast, while perturbation-based methods can be computationally expensive.

Model-agnostic tools such as SHAP and LIME can also be expensive or unstable for image inputs. Where feasible, explanations should be evaluated using multiple methods and compared with relevant annotations or clinical expertise. The evaluation should include failure cases, not only visually persuasive examples.

TABLE III. Common XAI techniques for chest X-ray models.

Method	Uses gradients?	Strength	Limitation
Grad-CAM	Yes	Fast and widely used	Coarse spatial resolution
Grad-CAM++	Yes	Can better represent multiple relevant regions	Still spatially coarse
Score-CAM	No	Uses changes in model score to weight activation maps	Computationally slower
SHAP	No	Principled feature-attribution framework	Can be expensive for image inputs
LIME	No	Provides local explanations	May be unstable
Attention rollout	No	Uses transformer attention information	Not necessarily class-specific or faithful

H. Trends Across the Literature

- Many studies report high accuracy on selected benchmarks, but small differences between headline scores are difficult to interpret when datasets, splits, and class definitions differ. Such differences may reflect evaluation choices rather than meaningful improvements.
- The research focus is increasingly broadening from accuracy alone to trustworthiness, including explanation quality, confidence calibration, and external validation.
- Hybrid designs combine CNNs and transformers sequentially, in parallel, or in multistage pipelines.
- Preprocessing remains an active area of investigation. CLAHE and histogram equalisation are relatively inexpensive, but their benefit should be measured on the intended dataset.
- Task definitions vary considerably in class count, label criteria, and dataset composition. Direct comparison of headline metrics is therefore risky unless the evaluation settings are aligned.

I. How Hybrid Models Combine CNN and Transformer Features

Hybrid architectures differ in how their components are connected, and this choice affects computational cost and model behaviour.

Four common patterns are described below.

- Sequential CNN-to-transformer design: a CNN produces feature maps that are converted into tokens for a transformer. This preserves a local-feature inductive bias in early layers and adds attention-based context later, as in compact approaches such as CovEnViT [17].
- Parallel late fusion: independent CNN and transformer backbones process the same image, and their output vectors are concatenated before classification. This design is straightforward to implement and evaluate through ablation studies, but the branches exchange information only at the fusion stage.
- Two-stage cascades: an initial model separates broad categories, such as normal and abnormal, and a second model refines the prediction, as in the ResNet-50 plus ViT-B/16 pipeline discussed in [19]. Cascades can be effective but add complexity and may compound errors across stages.
- Attention-augmented blocks: squeeze-and-excitation or window-attention components are incorporated into a CNN or transformer, as in LungMaxViT [18]. Such designs may improve feature selection, although their behaviour and explanations can be more difficult to interpret.

Cross-attention between branches allows CNN feature maps and transformer tokens to exchange information directly. Its potential value in CXR classification warrants careful study through controlled comparisons and ablation experiments.

IV. FUTURE ENHANCEMENTS

A. Proposed Dataset and Planned Implementation

For the planned implementation, the COVID-19 Radiography Database will be considered as the

primary dataset for multiclass chest X-ray classification. The dataset is described in the literature as containing four categories: COVID-19, lung opacity, viral pneumonia, and normal. Before use, the downloaded version, class labels, image counts, usage conditions, and possible duplicate or corrupted images will be checked. The images will then be prepared using a consistent preprocessing pipeline, and the data will be divided into training, validation, and test sets with attention to preventing data leakage. A hybrid Vision Transformer (ViT) and ResNet-50 approach, with optional CLAHE preprocessing, will be investigated as a candidate model. Model performance will be assessed after implementation using accuracy, precision, recall, F1-score, and a confusion matrix. These steps describe planned work only; the dataset has not been evaluated and no experimental performance results are claimed in this survey draft.

The literature points to several practical research priorities, summarised in Table IV.

A. Lightweight Models for Deployment

Running a ViT and ResNet-50 together can be computationally expensive. Knowledge distillation can train a smaller student model, while post-training quantisation or structured pruning may reduce resource requirements. Mobile-friendly backbones are another option. Evaluations should report inference latency, memory use, model size, and, where relevant, energy consumption alongside predictive metrics.

B. Uncertainty and Calibration

A diagnostic-support system should indicate when a prediction is uncertain. Monte Carlo dropout, deep ensembles, temperature scaling, and conformal prediction can help estimate uncertainty or produce prediction sets. Low-confidence cases can then be referred for expert review. These methods require careful validation and do not replace clinical judgement.

C. Improved Feature Fusion

Feature concatenation is simple, but the branches do not interact before the final fusion stage. Cross-attention between CNN feature maps and ViT tokens, gated fusion, or learnable branch weighting may allow the model to adapt the contribution of each branch to an individual image. Self-supervised pretraining on unlabelled CXR collections, for example through masked-image modelling, may also reduce dependence on natural-image pretraining.

D. Multi-Site and Privacy-Preserving Training

Training with data from multiple hospitals may improve coverage of population and acquisition variability, but sharing patient images can be restricted by privacy, legal, or institutional requirements. Federated learning enables institutions to train a shared model without directly centralising raw images, although model updates may still require privacy safeguards. Domain adaptation and intensity harmonisation may help address differences across scanners. Patient-level splits and held-out-site evaluation should be prioritized.

E. Multimodal Inputs

A radiograph is rarely interpreted in isolation. Patient age, symptoms, oxygen saturation, laboratory findings, prior images, and other clinically relevant information may help contextualise ambiguous findings. Radiology reports can also provide useful supervision, although their labels may be incomplete or noisy. Multimodal systems should be evaluated carefully to ensure that additional inputs improve performance without introducing leakage or unfair access assumptions.

F. Quantitatively Evaluated Explanations

Future studies should treat explanation quality as an evaluation target. Transformer-aware methods, such as attention rollout or relevance propagation, can complement class-activation maps. Faithfulness can be examined with deletion and insertion tests, while localisation can be measured against lung masks or expert annotations where suitable reference data exist. Lung-region masking may reduce reliance on image borders or text markers, but its effect should be measured. Explanation interfaces should also be assessed with intended users, including radiologists.

G. Clinical Validation, Fairness, and Governance

A computer-aided diagnosis tool must demonstrate value beyond benchmark accuracy. Multicentre validation and reader studies can assess whether it improves diagnostic accuracy, efficiency, or workflow. Subgroup analyses by relevant demographic and acquisition characteristics can identify performance disparities. Post-deployment

monitoring is needed to detect drift, and privacy, regulatory, and accountability requirements must be addressed. Extending the label set to other common findings may improve utility, provided each label is supported by reliable data and clear definitions.

H. A Possible Next-Generation Architecture

Figure 3 presents a proposed framework for future investigation rather than a validated system. Images from participating sites would first undergo lung-region alignment, CLAHE, and intensity harmonisation. A lightweight CNN would extract local textures and shapes, while a distilled or self-supervised ViT would model longer-range relationships. Cross-attention or gated fusion could combine the two feature streams before classification. If longitudinal images are available and appropriately linked, a separate component could examine changes over time. Explanation methods would generate visual summaries of regions influencing the output. Each component would need to be evaluated independently and as part of the complete pipeline.

A calibrated classification head could produce both a predicted class and an associated confidence estimate. Explanation maps should be evaluated for faithfulness and localisation, and uncertain cases should be flagged for expert review. For a federated implementation, institutions would retain their raw images locally, while the privacy implications of shared model updates would still need to be assessed.

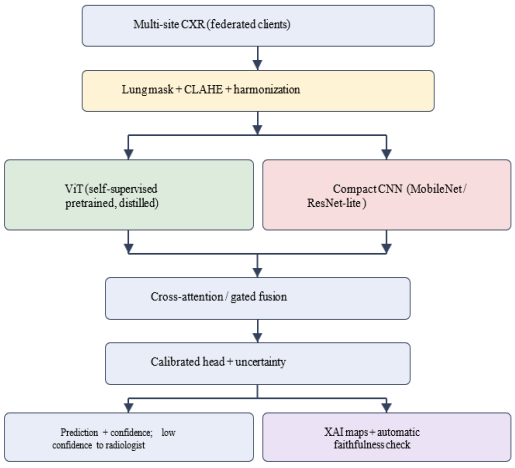


Fig. 3. Proposed next-generation pipeline combining the future enhancements discussed in this section.

Finally, a calibrated classification head processes the result. It produces both a prediction and a confidence value. Explanations are. Automatically checked for faithfulness. Cases with confidence are passed to a radiologist, for review. Training is carried out i manner that images never leave their hospital.

I. A Stricter Evaluation Protocol

Future studies should adopt a transparent evaluation checklist. Split data by patient before augmentation, and keep the test set untouched during model selection. Report variation across multiple random seeds and, where appropriate, confidence intervals obtained by bootstrapping the test predictions. Evaluate on at least one external dataset or held-out institution and describe any performance decline. Report class-wise sensitivity, specificity, calibration, and a confusion matrix, along with inference time and memory use. Quantitative evidence should support explanation figures, and code, configuration files, and trained weights should be shared when licensing, privacy, and institutional policies permit.

J. Ethics, Safety, and Human Oversight

A tool that informs diagnosis carries responsibilities beyond accuracy. Training data may under-represent some populations, so performance should be audited across relevant subgroups. A convincing heatmap may encourage over-trust, known as automation bias; outputs should therefore be presented as decision support rather than a definitive diagnosis. Patient privacy must be protected in stored data and, for federated systems, in shared model updates. Deployment may also require regulatory review, post-deployment monitoring, and clear accountability for errors.

K. Integration into Clinical Workflows

A model must fit existing hospital workflows. Within a picture archiving and communication system (PACS), a tool might analyse an image after storage and provide a prioritisation score or explanatory heatmap alongside the original radiograph. Uncertain predictions should be flagged for review rather than silently assigned a definitive label. Prediction logs and radiologist feedback can support audits and monitoring for performance drift. Workflow design is therefore part of safe deployment.

L. A Staged Research Roadmap

A staged roadmap can make development more

manageable. In the near term, a team could implement patient-level data splits, calibration, quantitative explanation checks, and model compression. Over approximately one year, it could investigate cross-attention fusion, self-supervised pretraining, an expanded label set, and external validation with a partner institution. Longer-term work could include federated multicentre training, multimodal inputs, and prospective reader studies. The timing depends on data access, ethics approvals, resources, and collaboration agreements.

TABLE IV. Proposed future enhancements and indicative effort.

Research direction	Potential benefit	Indicative effort
Distillation / quantisation	Smaller models and lower resource use	Medium
Uncertainty estimation and calibration	More transparent confidence and referral of uncertain cases	Low–medium
Cross-attention fusion	Richer interaction between CNN and transformer features	Medium
Federated multi-site training	Collaboration without centralising raw images	High
Multimodal inputs	Additional context for ambiguous cases	High
Quantitative XAI and lung masks	More rigorous explanation evaluation	Medium
Prospective clinical study	Evidence of clinical utility	High

V. DISCUSSION AND PRACTICAL RECOMMENDATIONS

The reviewed literature supports several practical recommendations for researchers developing or evaluating CXR classifiers.

- Begin with a clear baseline, such as a fine-tuned ResNet or ViT, and assess preprocessing choices such as CLAHE through controlled ablation studies.
- Split data at the patient level before applying augmentation, and apply augmentation only to training data. Report a confusion matrix and class-wise metrics rather than relying on accuracy alone. Check that values are consistent across the text, tables, and figures.
- Treat explanations as hypotheses to be tested. Where possible, compare them with lung masks or expert annotations and report failure cases as well as illustrative examples.
- Consider deployment constraints early, including model size, inference latency, memory use, and the handling of uncertain predictions.
- Evaluate models on data from other institutions before making claims about generalizability.

A. Reproducibility Checklist

- Document training, validation, and test partitions while protecting patient privacy and maintaining patient-level separation.
- Report framework and library versions, random seeds, pretrained weights, and the target layer used by each explanation method.
- Describe the training configuration consistently and ensure that the text, tables, and figures report matching values.
- Share code and trained weights where licensing, privacy, and institutional policies permit, so that results can be reproduced.

B. Open Challenges

Several challenges recur across the literature. Public datasets may contain noisy labels, which can affect measured performance. Class imbalance can make aggregate metrics misleading, and augmentation or resampling must be applied only after patient-level splitting to avoid leakage. Models may learn shortcuts, such as text markers, borders, or scanner-specific patterns, instead of disease-related features. Domain shift can reduce performance on images from a different institution or population. Further evidence is also needed to establish whether these tools improve clinical workflows or patient outcomes.

C. Limitations of This Survey

This paper is a focused narrative survey rather than a systematic review. It concentrates on deep-learning methods for lung-disease classification from chest radiographs, particularly CNN-based, transformer-based, hybrid, and explainability-oriented approaches. Work on CT, segmentation-only methods, and foundation models is covered only briefly. The discussion synthesises findings reported in the cited literature; no new experiments were conducted. Because datasets, class definitions, and evaluation protocols vary, reported performance values should not be treated as a direct leaderboard.

D. Glossary of Key Terms

Table V defines selected terms in plain language for readers from clinical and engineering backgrounds.

TABLE V. Glossary of terms used in this paper.

Term	Plain-language definition
AUC	A measure of how well a model ranks positive cases above negative cases; interpretation depends on the task and averaging method.
Calibration	The degree to which predicted confidence agrees with observed correctness.
Federated learning	A training approach in which institutions collaborate without directly transferring raw patient data to a central server.
Knowledge distillation	Training a smaller model to reproduce useful behaviour of a larger model.

VI. CONCLUSION

Research on CXR classification has progressed from conventional CNNs toward approaches that combine local visual features with broader image context. Explainability methods can make model behaviour easier to inspect, but they do not by themselves establish reliability. A trustworthy evaluation also requires patient-level data separation, transparent reporting, calibration, external validation, and quantitative assessment of explanations. Future work should investigate efficient deployment, uncertainty estimation, multimodal context, privacy-preserving multi-site learning, and prospective clinical evaluation. Since studies use different datasets and protocols, reported metrics must be interpreted in context rather than compared directly.

REFERENCES

[1] G. S. Gould et al., "Recognising the importance of chronic lung disease: A consensus statement from the Global Alliance for Chronic Diseases (lung diseases group)," *Respiratory Research*, vol. 24, no. 1, Art. no. 15, 2023.

[2] M. V. Sanida, T. Sanida, A. Sideris, and M. Dasygenis, "An advanced deep learning framework for multi-class diagnosis from chest X-ray images," *J*, vol. 7, no. 1, pp. 48–71, 2024, doi: 10.3390/j7010003.

[3] R. Deva and A. Dagur, "ViT-ResNet Fusion: An explainable hybrid framework for high-accuracy multiclass lung disease classification in chest X-rays," *IEEE Access*, vol. 14, pp. 1358–1372, 2026, doi: 10.1109/ACCESS.2025.3649109.

[4] K. Zuiderveld, "Contrast limited adaptive histogram equalization," in *Graphics Gems IV*. Academic Press, 1994, pp. 474–485.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[6] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learn. Represent. (ICLR), 2021.

[7] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 618–626.

[8] A. A. Biswas, "A comprehensive review of explainable AI for disease diagnosis," Array, vol. 22, Art. no. 100345, 2024.

[9] O. Sarkar et al., "Multi-scale CNN: An explainable AI-integrated unique deep learning framework for lung-affected disease classification," Technologies, vol. 11, no. 5, Art. no. 134, 2023.

[10] G. D. Deepak and S. K. Bhat, "A multi-stage deep learning approach for comprehensive lung disease classification from X-ray images," Expert Systems with Applications, vol. 277, Art. no. 127220, 2025, doi: 10.1016/j.eswa.2025.127220.

[11] G. Mohan, M. M. Subashini, S. Balan, and S. Singh, "A multiclass deep learning algorithm for healthy lung, COVID-19 and pneumonia disease detection from chest X-ray images," Discover Artificial Intelligence, vol. 4, no. 1, Art. no. 20, 2024.

[12] E. Mahamud et al., "An explainable artificial intelligence model for multiple lung diseases classification from chest X-ray images using fine-tuned transfer learning," Decision Analytics Journal, vol. 12, Art. no. 100499, 2024.

[13] E. Jangam, C. S. R. Annavarapu, and A. A. D. Barreto, "A multi-class classification framework

for disease screening and disease diagnosis of COVID-19 from chest X-ray images," Multimedia Tools and Applications, vol. 82, no. 10, pp. 14367–14401, 2023.

[14] J. Ko, S. Park, and H. G. Woo, "Optimization of vision transformer-based detection of lung diseases from chest X-ray images," BMC Medical Informatics and Decision Making, vol. 24, Art. no. 191, 2024.

[15] T. Chen et al., "A vision transformer machine learning model for COVID-19 diagnosis using chest X-ray images," Healthcare Analytics, vol. 5, Art. no. 100332, 2024.

[16] T.-A. Pham and V.-D. Hoang, "Chest X-ray image classification using transfer learning and hyperparameter customization for lung disease diagnosis," Journal of Information and Telecommunication, vol. 8, no. 4, pp. 587–601, 2024.

[17] X. Fu, R. Lin, W. Du, A. Tavares, and Y. Liang, "Explainable hybrid transformer for multi-classification of lung disease using chest X-rays," Scientific Reports, vol. 15, Art. no. 6650, 2025, doi: 10.1038/s41598-025-90607-x.

[18] Y. Hadhoud et al., "From binary to multi-class classification: A two-step hybrid CNN-ViT model for chest disease classification based on X-ray images," Diagnostics, vol. 14, no. 23, Art. no. 2754, 2024.

[19] R. Kumar et al., "Enhanced multi-model deep learning for rapid and precise diagnosis of pulmonary diseases using chest X-ray imaging," Diagnostics, vol. 15, no. 3, Art. no. 248, 2025.

[20] F. J. M. Shamrat et al., "High-precision multiclass classification of lung disease through customized MobileNetV2 from chest X-ray images," Computers in Biology and Medicine, vol. 155, Art. no. 106646, 2023.

[21] A. Hage Chehade et al., "Advancing chest X-ray diagnostics: A novel CycleGAN-based preprocessing approach for enhanced lung disease classification in ChestX-Ray14," Computer Methods and Programs in Biomedicine, vol. 259, Art. no. 108518, 2025.

[22] M. Nahiduzzaman et al., "Parallel CNN-ELM: A multiclass classification of chest X-ray images to identify seventeen lung diseases including COVID-19," Expert Systems with Applications, vol. 229, Art. no. 120528, 2023.

[23] S. Kim et al., "Deep learning in multi-class lung diseases' classification on chest X-ray images," Diagnostics, vol. 12, no. 4, Art. no. 915, 2022.

[24] S. Kumar, "COVID-Pneumonia-Normal Chest X-Ray Images," Kaggle dataset. [Online]. Available: COVID-Pneumonia-Normal Chest X-Ray Images. Accessed: Oct. 25, 2026.

[25] T. Rahman et al., "COVID-19 Radiography Database," Kaggle dataset. [Online]. Available: COVID-19 Radiography Database. Accessed: Oct. 25, 2026.