Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/200512

Agent-Based Intelligent Systems for Pneumonia Detection and Explainable Diagnostic Reporting: A Review

Esam Othman[1] *and Fadl Dahan*[2,*]

[1]Information Systems Department, College of Computer and Information Sciences, Prince Sultan University, Riyadh, Saudi Arabia
[2]Department of Management Information Systems, College of Business Administration - Hawtat Bani Tamim, Prince Sattam bin Abdulaziz University, Al-Kharj, 11942, Saudi Arabia

Received: 1 Jun. 2026, Revised: 12 Jul. 2026, Accepted: 24 Aug. 2026
Published online: 1 Sep. 2026

Abstract: Pneumonia remains one of the leading infectious causes of death worldwide, particularly among children under five and older adults; recent global burden estimates attribute approximately 2.5 million deaths in 2023 alone to the disease, including more than 600,000 children under five. Chest X-ray (CXR) imaging is the most widely available diagnostic modality, yet its interpretation remains subject to inter-reader variability and is constrained by the limited availability of specialist radiologists in many regions. Deep learning, and convolutional neural networks (CNNs) in particular, has driven a major shift in the automated detection of pneumonia from chest radiographs, beginning with seminal works such as CheXNet and the study of Kermany et al., and progressing through transfer-learning architectures including VGG, ResNet, DenseNet and EfficientNet. Nevertheless, most such models remain largely opaque "black boxes" that lack the transparency required to gain clinician trust, motivating explainable AI (XAI) techniques such as Grad-CAM, LIME and SHAP. More recently, multi-agent systems built on large language models (LLMs) that emulate the collaboration of multidisciplinary medical teams – exemplified by MedAgents, MDAgents, AgentClinic and KG4Diagnosis – have emerged as a promising direction for integrating visual perception, knowledge-based reasoning and report generation into a single explainable diagnostic pipeline. This paper presents a systematic literature review, guided by the PRISMA framework, of 32 studies covering (i) deep learning models for pneumonia detection, (ii) explainable AI techniques in medical imaging, and (iii) multi-agent systems for medical diagnosis and report generation, complemented by comparative tables of relevant studies and datasets and by a proposed conceptual architecture for an integrated agent-based system that combines diagnostic accuracy with clinical transparency. By supporting earlier and more accessible diagnosis of a leading cause of preventable death, this line of research also contributes to Sustainable Development Goal 3 (Good Health and Well-Being). The review concludes by discussing open challenges and future research directions in this emerging field.

Keywords: Chest X-ray, Deep learning, Explainable artificial intelligence, Multi-agent systems, Pneumonia detection

1 Introduction

Pneumonia is an acute infection that inflames the air sacs in one or both lungs, causing them to fill with fluid or pus and leading to breathing difficulty and reduced oxygenation. According to Global Burden of Disease (GBD 2023) estimates, pneumonia caused approximately 2.5 million deaths worldwide in 2023, including around 610,000 children under five and 79,000 children aged 5–14 [1]. UNICEF further reports that pneumonia remains the leading infectious cause of death among children under five, claiming the lives of more than 700,000 children annually – roughly 2,000 deaths per day [2]. Chest radiography remains the first-line and most widely accessible diagnostic tool for suspected pneumonia, but its interpretation requires specialized expertise and is affected by the visual similarity between pneumonia and other pulmonary conditions, as well as by the shortage of qualified radiologists in many resource-limited settings [3].

Deep learning has transformed medical image analysis over the past decade, enabling breakthroughs across classification, detection and segmentation tasks, as documented in the comprehensive survey by Litjens et al. covering more than 300 contributions to the field [4]. Within this broader trend, advances in deep learning have

* Corresponding author e-mail: f.naji@psau.edu.sa

enabled highly accurate image classification models that, in certain settings, rival physician performance – as demonstrated by CheXNet, whose performance exceeded the average performance of practicing radiologists on the F1 metric for pneumonia detection [5]. However, clinical adoption of such models is hindered by the "black-box" problem: a prediction must not only be accurate but also interpretable and justifiable to the treating physician and patient alike, which has driven the development of explainable AI (XAI) techniques.

In parallel, the emergence of large language models capable of operating as autonomous "agents" – able to plan, coordinate and invoke external tools – has opened the door to multi-agent systems that emulate multidisciplinary medical teams, where each agent specializes in a sub-task (visual perception, knowledge-based reasoning, quality control, report generation) and collaborates with the others through a coordinated architecture [6, 7]. Beyond their technical value, faster and more accessible diagnostic pipelines of this kind are also relevant to Sustainable Development Goal 3 (Good Health and Well-Being), which targets the reduction of preventable deaths from respiratory infections, particularly in settings where specialist care is scarce.

This review aims to provide a comprehensive, up-to-date picture of the research intersection between three main strands: (a) automated deep-learning-based pneumonia detection, (b) explainability techniques in medical imaging, and (c) multi-agent systems for medical diagnosis and report generation. Relevant literature was gathered from academic databases and search engines (arXiv, IEEE Xplore, ScienceDirect, ACL Anthology, PubMed, SpringerLink) using keywords such as pneumonia detection, chest X-ray CNN, explainable AI medical imaging, multi-agent LLM diagnosis, and radiology report generation, with emphasis on both highly cited seminal works and recent studies published between 2012 and 2025. The identification and selection of sources followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) approach illustrated in Fig. 1. An initial pool of 425 records was retrieved through database searching and citation tracking; after removing 75 duplicate records, 350 records were screened by title and abstract, of which 90 underwent full-text assessment against the review's scope and quality criteria, yielding the 32 sources retained for qualitative synthesis in this review. Title/abstract screening and full-text eligibility assessment were each performed by a single reviewer against the criteria stated above; a second author cross-checked a sample of borderline decisions, with disagreements resolved by discussion. Of the 58 records excluded at the full-text stage, approximately 32 were excluded as off-topic (not addressing pneumonia detection, XAI, or medical multi-agent systems), 15 for insufficient methodological detail to extract comparable information, and 11 for being non-peer-reviewed without qualifying for the preprint exception described below. Preprints from established authors or research groups with substantial citation counts (e.g., CheXNet, VGG, AgentClinic, KG4Diagnosis) were retained as an explicit exception to the peer-review criterion, given their foundational or highly-cited status in the field; all other non-peer-reviewed records failing this exception were excluded at the full-text stage.

The present paper is organized as follows. In Section 2 we review deep learning models for pneumonia detection from chest radiographs. In Section 3 we discuss explainable AI techniques in medical imaging. In Section 4 we examine multi-agent systems in the medical domain. Section 5 presents the most commonly used benchmark datasets. In Section 6 we propose an integrated agent-based system architecture. Section 7 discusses open challenges, and Section 8 concludes the paper and outlines future research directions.

2 Deep Learning for Pneumonia Detection from Chest Radiographs

2.1 General pipeline of CNN-based classification systems

Most systems proposed in the literature follow a general pipeline that begins with chest X-ray acquisition, followed by pre-processing (resizing, normalization and data augmentation), then feature extraction via a pre-trained or custom-designed convolutional neural network, and finally a classification layer that outputs the probability of pneumonia, as illustrated in Fig. 2.

2.2 Seminal works

The modern era of CNN-based image classification is generally traced back to Krizhevsky, Sutskever and Hinton's AlexNet, which demonstrated the power of deep convolutional networks on large-scale image recognition and catalyzed the broader adoption of CNNs across computer vision and medical imaging alike [8]. Building on this foundation, the study by Kermany et al. is a landmark work in the pneumonia detection field: the authors applied transfer learning using the Inception V3 architecture on a pediatric chest X-ray dataset, achieving classification accuracy comparable to expert physicians, and subsequently released the dataset publicly, making it one of the most widely used benchmarks in follow-up research [3].

Around the same time, Rajpurkar et al. introduced CheXNet, a 121-layer DenseNet-121 model trained on the ChestX-ray14 dataset, comprising more than 100,000 frontal-view X-ray images spanning fourteen thoracic diseases. A comparison against four practicing academic radiologists showed that the model exceeded average radiologist performance on the F1 metric for pneumonia

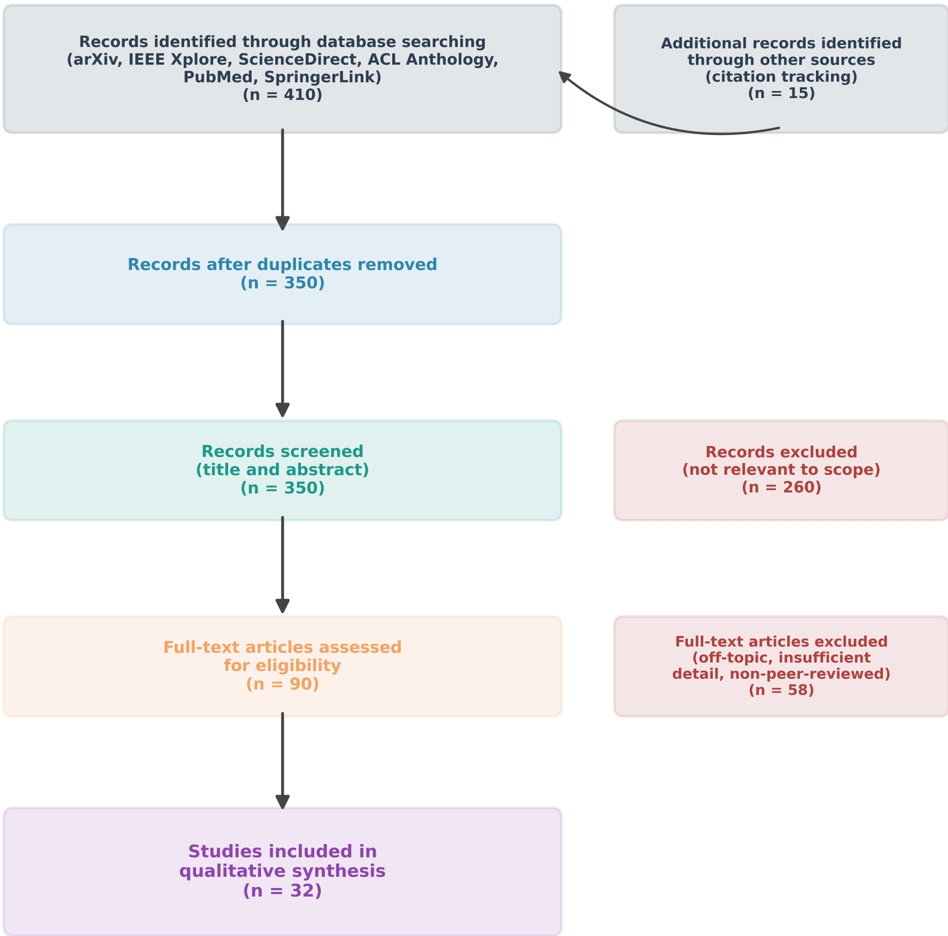

Fig. 1: PRISMA flow diagram of the literature search and selection process

detection [5]. The model also used class activation maps to localize probable disease regions, making it one of the earliest works to combine high diagnostic accuracy with an initial step toward visual interpretability.

2.3 Common convolutional architectures

Several standard architectures have been widely adopted as transfer-learning backbones for pneumonia classification, most notably VGG, known for its simple sequential design and moderate depth [9]; ResNet, which introduced residual connections to facilitate the training of very deep networks [10]; DenseNet, which connects each layer to all preceding layers to enhance feature reuse and reduce parameter count [11, 12]; and EfficientNet, which employs a compound scaling strategy to achieve a better trade-off between accuracy and computational efficiency [13]. Alongside these convolution-based architectures, recent years have seen efforts to combine Vision Transformers with CNNs within ensemble models that merge DenseNet, MobileNet and ViT, demonstrating improved generalization compared to single models [14].

2.4 Prior systematic reviews

Beyond individual empirical studies, comprehensive systematic reviews have surveyed research trajectories in this field. A notable example is the systematic literature review by Sharma and Guleria, which covered more than one hundred studies on deep-learning-based pneumonia detection from chest X-rays, discussing network architectures, datasets, evaluation metrics and common challenges [15]. The same authors had earlier proposed a dedicated VGG-16-based model augmented with additional dense layers, achieving strong accuracy on a limited training set [16]. Unlike this and other prior reviews, which focus primarily on CNN architectures and reported accuracy benchmarks, the present review is, to

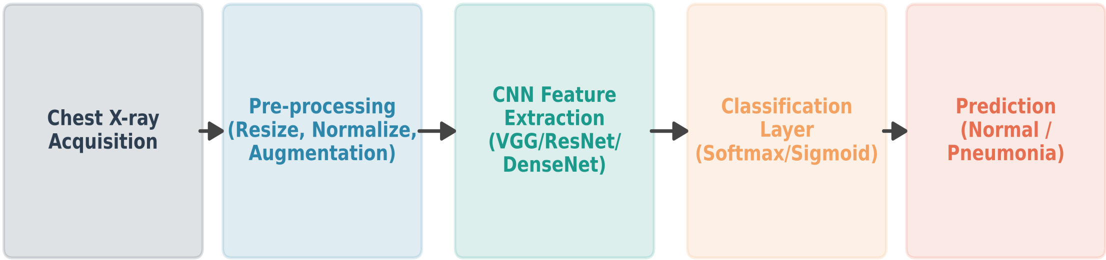

Fig. 2: General pipeline of CNN-based pneumonia detection from chest X-ray images

our knowledge, the first to jointly examine explainable AI techniques and LLM-based multi-agent frameworks as a single, integrated diagnostic pipeline for pneumonia detection.

2.5 Comparison of key studies

Table 1 summarizes several prominent deep-learning-based pneumonia detection studies in terms of architecture, dataset, reported accuracy and key contribution. Beyond the works listed, Rajasenbagam, Jeyanthi and Pandian proposed a deep CNN based on VGG19 combined with content-based image retrieval and Deep Convolutional Generative Adversarial Network (DCGAN) augmentation, achieving 99.34% accuracy on an unseen chest X-ray test set [17], while Bhatt and Shah demonstrated that a CNN ensemble model can further improve generalization over single-architecture classifiers [14].

3 Explainable Artificial Intelligence (XAI) in Medical Imaging

3.1 Why does medical diagnosis need explainability?

Tjoa and Guan, in their comprehensive survey, note that the adoption of highly complex models in high-stakes sectors such as medicine demands a high degree of accountability and transparency, and that the black-box nature of deep learning models remains a major obstacle to understanding and clinically justifying automated decisions [18]. This has motivated a class of techniques aimed at explaining a model's decisions after training (post-hoc explainability) without compromising its classification performance.

3.2 Key explainability techniques

Grad-CAM, proposed by Selvaraju et al. in 2017, relies on the gradients of the target class flowing into the last convolutional layer to produce a coarse localization map highlighting the regions most influential to the model's decision, and has since become the most widely used tool for visually explaining chest X-ray classification models [19]. LIME, introduced by Ribeiro, Singh and Guestrin in 2016, follows a different approach: it approximates the behavior of a complex model locally with a simple, interpretable linear model, by generating perturbations of the original input and observing their effect on the prediction [20]. SHAP, developed by Lundberg and Lee in 2017, is instead grounded in Shapley values from cooperative game theory, distributing the "contribution" of each feature or pixel to the final decision in a fair and mathematically consistent manner [21].

3.3 Applications of XAI in pneumonia detection

Several studies have combined these techniques to enhance the trustworthiness of classification systems. For instance, Bhandari et al. integrated XAI techniques within a classification system distinguishing COVID-19, pneumonia and tuberculosis cases from chest X-rays [22]. The recent IHRAS (Intelligent Humanized Radiology Analysis System) relies on Grad-CAM to visually localize pathological regions alongside anatomical segmentation and a large language model that generates a structured textual report using standardized medical terminology (SNOMED CT, Systematized Nomenclature of Medicine – Clinical Terms) [12], reflecting a clear shift from merely "explaining a decision" toward "generating a comprehensive, justified clinical report," as illustrated by the pipeline in Fig. 3.

3.4 Comparison of explainability techniques

Table 2 summarizes the four techniques discussed above in terms of their type, core principle, advantages and limitations.

Table 1: Comparison of key deep-learning-based pneumonia detection studies

Study / Year	Architecture	Dataset	Reported accuracy	Key contribution
Kermany et al., 2018 [3]	Transfer learning (Inception V3)	Pediatric CXR (Guangzhou)	92.8%	Landmark study achieving near-expert performance on pediatric chest X-ray classification
Rajpurkar et al. (CheXNet), 2017 [5]	DenseNet-121 (121 layers)	ChestX-ray14 (100,000+ images)	F1 exceeding average radiologist	Outperformed practicing radiologists on the F1 metric for pneumonia detection
Sharma & Guleria, 2023 [16]	VGG-16 + custom dense layers	Kaggle Chest X-ray dataset	∼90%	Fine-tuned VGG-16 with additional layers improves accuracy on limited data
Rajasenbagam et al., 2021 [17]	VGG19 + DCGAN augmentation	12,000 chest X-ray images	99.34%	Generative augmentation (DCGAN) substantially boosts classification accuracy
Bhatt & Shah, 2023 [14]	CNN ensemble model	Chest X-ray benchmark dataset	Higher than single models	Combining several CNN backbones in an ensemble improves generalization

Note: Reported figures are drawn directly from each cited study and are not directly comparable across rows, since they arise from different datasets, class balances, data splits and evaluation protocols; values should be interpreted as illustrative of the range of results reported in the literature rather than as a ranked comparison.

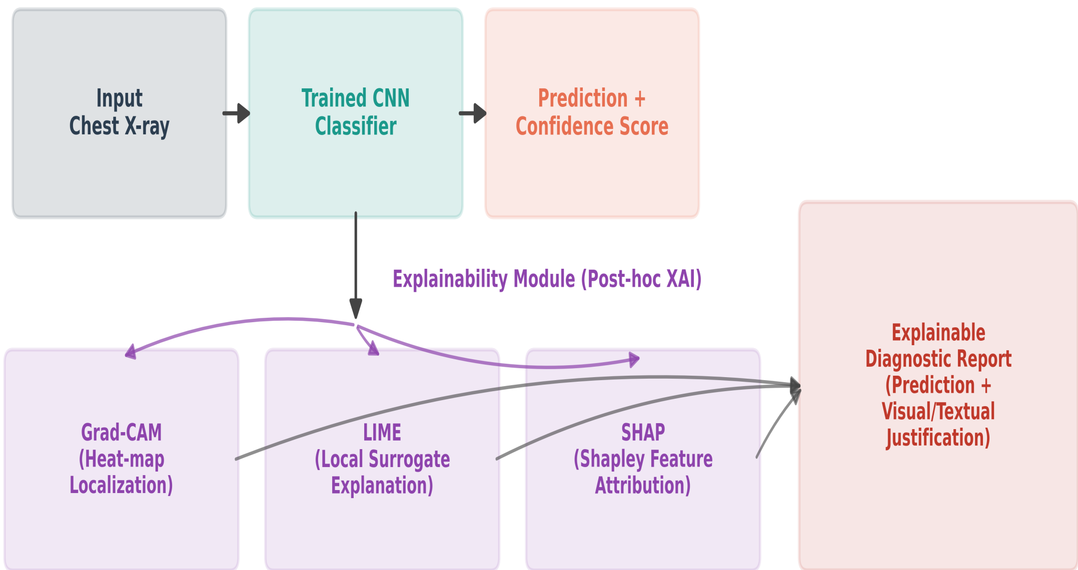

Fig. 3: Integration pipeline of explainability techniques (Grad-CAM, LIME, SHAP) with a CNN classifier

4 Multi-Agent Systems for Medical Diagnosis and Report Generation

4.1 From single models to collaborative agent teams

The theoretical foundations of software agents and multi-agent coordination were laid out well before the LLM era, notably in Wooldridge's classic treatment of agent architectures, communication protocols and coordination mechanisms [23]. The limitations of single-model approaches in handling complex diagnostic cases have more recently motivated the design of LLM-based multi-agent frameworks that emulate multidisciplinary medical teams, where each agent specializes in a specific role (general practitioner, radiology specialist, quality reviewer) and interacts with the other agents through a defined coordination protocol [6, 7].

Table 2: Comparison of key explainable AI techniques used in medical imaging

Method	Type	Core principle	Advantages	Limitations
Grad-CAM [19]	Model-specific (CNN)	Uses gradients of the target class flowing into the last convolutional layer to produce a heat map	Fast, requires no retraining, visually intuitive	Coarse spatial resolution; architecture-dependent
LIME [20]	Model-agnostic	Fits a simple local linear surrogate model around one prediction by perturbing the input	Applicable to any classifier	Can be unstable and sensitive to the sampling strategy
SHAP [21]	Model-agnostic	Distributes Shapley values from cooperative game theory to attribute feature contributions	Strong theoretical foundation with fair attribution properties	Computationally expensive for high-dimensional images
Grad-CAM++ / Score-CAM	Extension of Grad-CAM	Refines per-pixel weighting or replaces gradients with confidence scores	Sharper localization of multiple pathological regions	Added computational overhead versus vanilla Grad-CAM

4.2 Representative multi-agent frameworks

The MedAgents framework has a number of LLM agents adopt specialized medical roles (e.g., cardiology or radiology) and engage in a multi-round discussion to simulate real-world clinical reasoning without requiring additional training [6]. The MDAgents framework instead introduces an adaptive strategy that dynamically determines whether a case requires solo or group processing based on its medical complexity, and has shown a notable accuracy improvement of up to 11.8% when combining moderator review with external medical knowledge [7].

AgentClinic provides an interactive simulation environment in which one agent plays the role of the doctor and another the role of the patient, requiring the "doctor agent" to elicit a diagnosis through dialogue and active data collection rather than relying on static multiple-choice questions; the evaluation spans nine medical specialties and seven languages [24]. KG4Diagnosis, in turn, adopts a two-tier hierarchical architecture: a "general practitioner" agent performs initial assessment and triage, followed by specialist agents for in-depth diagnosis, with reasoning enhanced through a knowledge graph automatically built from medical text [25].

4.3 Toward integrating visual perception with multi-agent reasoning

Most closely related to the topic of this review, the IHRAS system combines four integrated components: deep convolutional networks for classifying fourteen thoracic conditions, Grad-CAM for visual localization, a SAR-Net network for anatomical segmentation, and a large language model (DeepSeek-R1) guided by the CRISPE (Capacity and Role, Insight, Statement, Personality, Experiment) prompt-engineering framework to generate structured diagnostic reports using standardized SNOMED CT terminology [12]. This system represents a practical example of transforming classification and visual-explanation outputs into a fully integrated clinical text report. Earlier work on weakly-supervised, end-to-end radiological report generation similarly demonstrated the feasibility of mapping chest X-ray features directly onto free-text clinical narratives [26].

4.4 Comparison of related multi-agent frameworks

Table 3 compares the agent architecture, application domain and key characteristics of the frameworks reviewed in this section.

5 Benchmark Datasets

The development and evaluation of pneumonia detection systems rely on a limited number of publicly available datasets, which vary in size, annotation granularity and level of detail, as summarized in Table 4. Beyond the pediatric CXR, ChestX-ray14 and CheXpert datasets discussed earlier, two additional large-scale resources are widely used for related tasks: MIMIC-CXR, which pairs over 377,000 chest radiographs with free-text radiology reports from the Beth Israel Deaconess Medical Center [27], and PadChest, a Spanish-language dataset of more than 160,000 images annotated with 174 radiographic findings mapped to standardized medical terminology [28]. For pneumonia localization

Table 3: Comparison of key multi-agent frameworks in the medical domain

Framework	Agent architecture	Application domain	Key characteristics
MedAgents [6]	Multiple role-specialized agents (collaborative discussion)	Zero-shot medical reasoning	Simulates multidisciplinary discussion among LLM agents to improve diagnostic accuracy
MDAgents [7]	Dynamically assigned solo/group collaboration structure	Complex medical decision-making	Adapts the number of agents and collaboration mode to case complexity
AgentClinic [24]	Doctor agent and patient agent in a simulated environment	Interactive multimodal clinical diagnosis evaluation	Dialogue-driven environment simulating clinical data collection under incomplete information
KG4Diagnosis [25]	Hierarchical: general-practitioner agent + specialist agents	Multi-domain medical diagnosis	Integrates a knowledge graph to enhance diagnostic reasoning
IHRAS [12]	Integrated modules: classification + segmentation + Grad-CAM + LLM	Explainable chest X-ray report generation	Links deep classification, visual explanation and standardized-terminology report generation

specifically, the RSNA Pneumonia Detection Challenge dataset augments a subset of the NIH ChestX-ray8 collection with expert bounding-box annotations of probable pneumonia opacities [29].

Table 4: Key chest X-ray datasets used in pneumonia detection research

Dataset	Size	Annotation type
Pediatric CXR (Guangzhou / Kermany) [3]	5,863 images	Binary (Normal / Pneumonia)
ChestX-ray14 (NIH) [30]	112,120 images, 30,000+ patients	14 thoracic diseases (multi-label)
CheXpert [31]	224,316 images	14 findings, uncertainty labels
MIMIC-CXR [27]	377,110 images / 227,835 studies	Free-text radiology reports
PadChest [28]	160,000+ images	174 findings (multi-label, Spanish)
RSNA Pneumonia Challenge [29]	~30,000 images	Bounding boxes for opacities

6 Proposed Agent-Based System Architecture

Building on the preceding review, this paper proposes a conceptual architecture for an integrated intelligent system that combines the diagnostic accuracy of CNN models, the transparency offered by XAI techniques, and the coordination and reasoning capabilities of LLM-based multi-agent systems. The proposed system consists of five core agents coordinated through a central orchestrator agent, as illustrated in Fig. 4:

– **Perception Agent**: extracts features and classifies the chest X-ray using a pre-trained convolutional architecture (e.g., DenseNet or ResNet), outputting a diagnosis probability with a confidence score.
– **Explainability Agent**: applies Grad-CAM, LIME and SHAP to the Perception Agent's outputs to produce heat maps and local explanations that clarify the visual basis of the decision.
– **Knowledge/Reasoning Agent**: a large language model augmented with structured medical knowledge (a medical ontology or knowledge graph) that links the visual findings to the patient's clinical context.
– **Validation & Quality Control Agent**: reviews the consistency of the preceding agents' outputs, flags potential contradictions, and recommends re-evaluation when confidence is low.
– **Report Generation Agent**: integrates all outputs into a structured, explainable diagnostic report presented to the treating physician under a human-in-the-loop model for the final decision.

The core benefit of this architecture lies in the separation of concerns among perception, explanation, reasoning and validation tasks, allowing each agent to be developed and improved independently while the orchestrator agent maintains overall system coherence, consistent with classical multi-agent coordination principles [23]. The dedicated validation and quality-control agent also provides an additional safety layer, reducing the risk of error propagation or potential hallucination by the large language model in this sensitive clinical context.

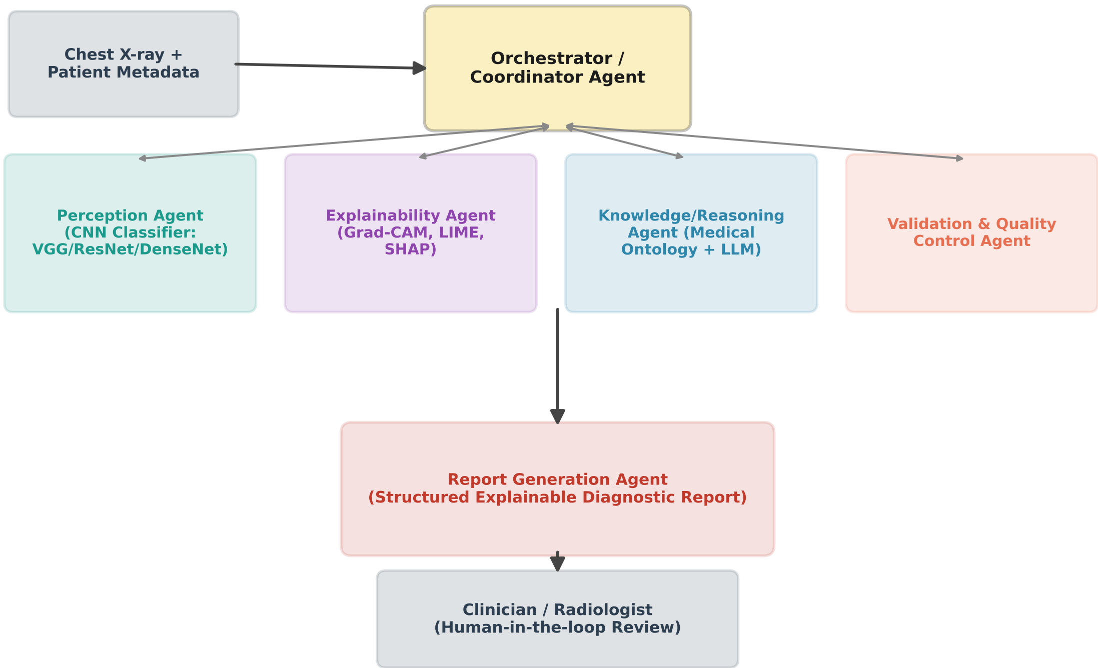

Fig. 4: Proposed agent-based architecture for pneumonia detection and explainable diagnostic reporting

7 Open Challenges

7.1 Technical challenges

- Limited spatial resolution of techniques such as Grad-CAM, which produce coarse heat maps that may not precisely align with fine anatomical lesion boundaries.
- High computational cost of SHAP when applied to high-dimensional images, limiting its use in time-critical settings.
- Instability of some LIME explanations due to sensitivity to the random sampling strategy around the prediction point.
- Risk of hallucination in large language models when generating medical reports, requiring strict verification mechanisms before any diagnostic text is adopted.

7.2 Data and organizational challenges

- Limited size and diversity of publicly available datasets compared to actual clinical data volumes in hospitals, with associated risk of demographic bias.
- Privacy and health-data-protection constraints, which have motivated solutions such as federated learning – in which model updates rather than raw patient data are exchanged across institutions [32] – to train models without transferring original patient data outside hospitals.
- Absence of standardized metrics for evaluating explanation quality itself, as opposed to conventional classification accuracy evaluation.

7.3 Clinical and ethical challenges

- The need for large-scale clinical validation studies before adopting multi-agent systems in real-world care settings.
- The necessity of preserving the human-in-the-loop principle and avoiding full replacement of the final medical decision by automated system outputs.
- Legal and ethical accountability issues in the event of a diagnostic error arising from a complex interaction among multiple AI agents.

8 Conclusion and Future Directions

Based on a PRISMA-guided synthesis of 32 studies, this review has traced the research trajectory from early convolutional classification models for pneumonia detection, through explainable AI techniques that addressed the “black-box” problem, to multi-agent systems built on large language models that open new avenues for integrating visual perception, knowledge-based reasoning and report generation into a single, transparent diagnostic pipeline.

The reviewed literature suggests that the most likely future direction lies in hybrid systems combining the strength of convolutional networks in extracting fine-grained visual features with the contextual reasoning and natural communication capabilities of large language models, within a well-organized and verifiable multi-agent architecture. Among the most promising future research directions are:

– developing standardized quantitative metrics to evaluate medical explanation quality rather than classification accuracy alone;
– integrating federated learning with multi-agent systems to preserve patient data privacy across multiple institutions [32];
– conducting controlled clinical trials to assess the impact of such systems on diagnostic decision quality and clinician/patient trust; and
– developing clear governance and regulatory frameworks for legal accountability when deploying multi-agent systems in critical diagnostic contexts.

More broadly, advancing such systems supports Sustainable Development Goal 3 by making reliable pneumonia diagnosis more accessible in settings where specialist care is scarce.

Declarations

Funding. The authors extend their appreciation to Prince Sattam bin Abdulaziz University for funding this research work through the project number (PSAU/2026/01/33486).
Conflicts of interest/Competing interests. The authors declare no conflict of interest.
Availability of data and material. Not applicable.
Code availability. Not applicable.

Acknowledgement

The authors extend their appreciation to Prince Sattam bin Abdulaziz University for funding this research work through the project number (PSAU/2026/01/33486). They also acknowledge the valuable and insightful contributions of researchers at Prince Sultan University.

References

[1] UNICEF DATA, Pneumonia in Children Statistics, UNICEF Global Databases (updated 2025), https://data.unicef.org/topic/child-health/pneumonia/ (2025).
[2] Forum of International Respiratory Societies (FIRS), Pneumonia Remains Leading Cause of Child Mortality, World Pneumonia Day Press Release (2025).
[3] D. S. Kermany, M. Goldbaum, W. Cai, C. C. S. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan et al., Identifying medical diagnoses and treatable diseases by image-based deep learning, *Cell*, **172(5)**, 1122–1131 (2018).
[4] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken and C. I. Sánchez, A survey on deep learning in medical image analysis, *Med. Image Anal.*, **42**, 60–88 (2017).
[5] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren and A. Y. Ng, CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning, *arXiv:1711.05225* (2017).
[6] X. Tang, A. Zou, Z. Zhang, Z. Li, Y. Zhao, X. Zhang, A. Cohan and M. Gerstein, MedAgents: Large language models as collaborators for zero-shot medical reasoning, in *Findings of the Assoc. Comput. Linguistics (ACL)*, 599–621 (2024).
[7] Y. Kim, C. Park, H. Jeong, Y. S. Chan, X. Xu, D. McDuff, H. Lee, M. Ghassemi, C. Breazeal and H. W. Park, MDAgents: An adaptive collaboration of LLMs for medical decision-making, in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, **37** (2024).
[8] A. Krizhevsky, I. Sutskever and G. E. Hinton, ImageNet classification with deep convolutional neural networks, in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, **25**, 1097–1105 (2012).
[9] K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition, *arXiv:1409.1556*; presented at the *Int. Conf. Learn. Represent. (ICLR)* (2015).
[10] K. He, X. Zhang, S. Ren and J. Sun, Deep residual learning for image recognition, in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 770–778 (2016).
[11] G. Huang, Z. Liu, L. Van Der Maaten and K. Q. Weinberger, Densely connected convolutional networks, in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 4700–4708 (2017).
[12] G. A. P. Rodrigues, A. L. M. Serrano, G. D. Bispo, G. P. R. Filho, V. P. Gonçalves and R. I. Meneguette, IHRAS: Automated medical report generation from chest X-rays via classification, segmentation, and LLMs, *Bioengineering (MDPI)*, **12(8)**, Art. 795 (2025).
[13] M. Tan and Q. Le, EfficientNet: Rethinking model scaling for convolutional neural networks, in *Proc. Int. Conf. Mach. Learn. (ICML), PMLR*, 6105–6114 (2019).
[14] H. Bhatt and M. Shah, A convolutional neural network ensemble model for pneumonia detection using chest X-ray images, *Healthcare Anal.*, **3**, Art. 100176 (2023).
[15] S. Sharma and K. Guleria, A systematic literature review on deep learning approaches for pneumonia detection using chest X-ray images, *Multimed. Tools Appl.*, **83(8)**, 24101–24151 (2024).
[16] S. Sharma and K. Guleria, A deep learning based model for the detection of pneumonia from chest X-ray images using VGG-16 and neural networks, *Procedia Comput. Sci.*, **218**, 357–366 (2023).

[17] T. Rajasenbagam, S. Jeyanthi and J. A. Pandian, Detection of pneumonia infection in lungs from chest X-ray images using deep convolutional neural network and content-based image retrieval techniques, *J. Ambient Intell. Humaniz. Comput.*, 1–8 (2021).
[18] E. Tjoa and C. Guan, A survey on explainable artificial intelligence (XAI): Toward medical XAI, *IEEE Trans. Neural Netw. Learn. Syst.*, **32(11)**, 4793–4813 (2020).
[19] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 618–626 (2017).
[20] M. T. Ribeiro, S. Singh and C. Guestrin, 'Why should I trust you?': Explaining the predictions of any classifier, in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. (KDD)*, 1135–1144 (2016).
[21] S. M. Lundberg and S.-I. Lee, A unified approach to interpreting model predictions, in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, **30**, 4765–4774 (2017).
[22] M. Bhandari, T. B. Shahi, B. Siku and A. Neupane, Explanatory classification of CXR images into COVID-19, pneumonia and tuberculosis using deep learning and XAI, *Comput. Biol. Med.*, **150**, Art. 106156 (2022).
[23] M. Wooldridge, *An Introduction to MultiAgent Systems*, 2nd ed., John Wiley & Sons, Chichester, UK (2009).
[24] S. Schmidgall, R. Ziaei, C. Harris, E. P. Reis, J. Jopling and M. Moor, AgentClinic: A multimodal agent benchmark to evaluate AI in simulated clinical environments, *arXiv:2405.07960* (2024).
[25] K. Zuo, Y. Jiang, F. Mo and P. Liò, KG4Diagnosis: A hierarchical multi-agent LLM framework with knowledge graph enhancement for medical diagnosis, *arXiv:2412.16833* (2024).
[26] S. Zhang, X. Xin, Y. Wang, Y. Guo, Q. Hao, X. Yang, J. Wang, J. Zhang, B. Zhang and W. Wang, Automated radiological report generation for chest X-rays with weakly-supervised end-to-end deep learning, *arXiv:2006.10347* (2020).
[27] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-Y. Deng, R. G. Mark and S. Horng, MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports, *Sci. Data*, **6**, Art. 317 (2019).
[28] A. Bustos, A. Pertusa, J.-M. Salinas and M. de la Iglesia-Vayá, PadChest: A large chest X-ray image dataset with multi-label annotated reports, *Med. Image Anal.*, **66**, Art. 101797 (2020).
[29] G. Shih, C. C. Wu, S. S. Halabi, M. D. Kohli, L. M. Prevedello, T. S. Cook, A. Sharma, J. K. Amorosa, V. Arteaga, M. Galperin-Aizenberg et al., Augmenting the National Institutes of Health chest radiograph dataset with expert annotations of possible pneumonia, *Radiol. Artif. Intell.*, **1(1)**, Art. e180041 (2019).
[30] 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, in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2097–2106 (2017).
[31] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya et al., CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison, in *Proc. AAAI Conf. Artif. Intell.*, **33**, 590–597 (2019).
[32] B. McMahan, E. Moore, D. Ramage, S. Hampson and B. Agüera y Arcas, Communication-efficient learning of deep networks from decentralized data, in *Proc. 20th Int. Conf. Artif. Intell. Stat. (AISTATS), PMLR*, 1273–1282 (2017).