Diagnostic-Field Variational Intelligence for Trustworthy Pneumonia Screening: A UVIF-Based Framework for Explainable and Calibration-Aware Clinical Decision Support

Fatma Mallek *, Zeyad Alghamdi , Sghaier Guizani *, Habib Hamam

Posted Date: 22 May 2026

doi: 10.20944/preprints202605.1557.v1

Keywords: pneumonia detection; pneumonia screening; medical imaging AI; clinical decision support; diagnostic artificial intelligence; explainable AI; trustworthy AI; calibration-aware learning; variational intelligence; UVIF; chest X-ray analysis; MedMNIST

 doi:10.20944/preprints202605.1557.v1

Article

Diagnostic-Field Variational Intelligence for Trustworthy Pneumonia Screening: A UVIF-Based Framework for Explainable and Calibration-Aware Clinical Decision Support

Fatma Mallek [1,*], Zeyad Alghamdi [2], Sghaier Guizani [1,3,*] and Habib Hamam [1,4,5]

[1] Faculty of Engineering, Uni de Moncton, Moncton, NB E1A3E9, Canada
[2] College of Computer Science and Engineering, University of Ha'il, Ha'il, 55476, Saudi Arabia
[3] Department of Electrical Engineering, Alfaisal University, Riyadh, Saudi Arabia
[4] School of Electr. Eng., Uni. of Johannesburg, Johannesburg 2006, South Africa
[5] International Inst. of Techno. & Management (IITG), Av. Grandes Ecoles, Libreville, Gabon
* Correspondence: efm7239@umoncton.ca

Abstract

Artificial intelligence systems for pneumonia detection often achieve strong predictive performance but remain insufficiently calibrated, weakly interpretable, and poorly aligned with clinically meaningful decision-support requirements. This paper presents a diagnostic-field extension of the Unified Variational Intelligence Framework (UVIF) for trustworthy and decision-centric pneumonia screening using chest X-ray imaging. The proposed framework models diagnosis as a variational process in which imaging patterns and latent feature representations are treated as diagnostic fields that must be sensed, filtered, interpreted, and evaluated before clinical decision support is produced. The study combines compact convolutional neural network modeling, embedding-based machine learning classifiers, calibration-aware reliability analysis, threshold-sensitive decision control, and multi-level explainability using Grad-CAM, LIME, and SHAP. Experimental evaluation is conducted on the publicly available PneumoniaMNIST benchmark dataset from the MedMNIST collection. The compact CNN achieved strong discrimination performance with ROC-AUC of 0.9666 and pneumonia recall of 0.9974, while the UVIF-guided diagnostic layer supported reliability-aware model selection and threshold optimization under screening-oriented constraints. Calibration analysis further revealed deviations between predicted probabilities and empirical outcomes, emphasizing the importance of reliability-aware evaluation in medical AI systems. The proposed framework demonstrates that integrating prediction, calibration, explainability, and diagnostic decision control within a unified variational framework can support more transparent, interpretable, and clinically meaningful AI-assisted pneumonia screening systems.

Keywords: pneumonia detection; pneumonia screening; medical imaging AI; clinical decision support; diagnostic artificial intelligence; explainable AI; trustworthy AI; calibration-aware learning; variational intelligence; UVIF; chest X-ray analysis; MedMNIST

1. Introduction

1.1. Context and Motivation

Artificial intelligence (AI) and machine learning are increasingly used in medical diagnosis, prognosis, imaging interpretation, disease-risk prediction, and clinical decision support [1,2]. In pneumonia care, data-driven methods have shown particular promise for early detection, triage, severity assessment, and outcome prediction [3].

However, the clinical value of AI-assisted diagnosis depends not only on predictive accuracy, but also on reliability, interpretability, and decision relevance. A model that produces a correct class label

 doi:10.20944/preprints202605.1557.v1

on average may still be unsuitable for clinical use if its probability estimates are poorly calibrated, if its errors are concentrated in high-risk subgroups, or if its predictions cannot be meaningfully interpreted by clinicians. This issue is particularly important in pneumonia detection, where false negatives may delay treatment, increase the risk of deterioration, and compromise timely triage. Calibration studies have shown that machine learning models may generate biased probability estimates, especially when class imbalance correction is applied, leading to over- or under-estimation of clinical risk [4]. Similarly, although explainability methods such as SHAP and LIME are increasingly used in diagnostic AI, they are often applied as post hoc visual or qualitative tools rather than as integral components of clinically reliable decision support [5].

Recent studies have begun to incorporate calibration metrics, decision curve analysis, and interpretable modeling into diagnostic AI evaluation [6,7]. Nevertheless, these elements are frequently treated separately. For a journal focused on diagnostic methods and clinical decision support, this separation is a central methodological limitation: diagnostic AI should not only predict disease status, but also communicate confidence, expose decision drivers, and support clinically adjustable thresholds. This work is motivated by that gap. It develops a trust-aware pneumonia diagnostic framework that integrates predictive performance, calibration reliability, threshold-aware decision control, and multi-level explainability into a unified clinical decision-support paradigm.

1.2. Key Contributions

This work introduces a trust-aware and explainable machine-learning framework for AI-assisted pneumonia diagnosis, with emphasis on diagnostic reliability, high-sensitivity screening, probability calibration, and clinically meaningful decision support. The proposed framework is aligned with the diagnostic focus of *Diagnostics* by addressing not only disease classification, but also the reliability and interpretability of the diagnostic process.

The primary contributions are summarized as follows:

C1– *Diagnostic AI framework for high-sensitivity pneumonia screening:* We develop a machine-learning framework for pneumonia detection that explicitly supports threshold-aware decision control, allowing sensitivity and specificity to be adjusted according to clinical objectives, particularly in screening contexts where missed pneumonia cases are clinically costly.

C2– *Calibration-aware reliability assessment for diagnostic probability estimation:* The proposed framework incorporates calibration analysis using reliability curves and bin-wise probability statistics, enabling evaluation of whether predicted pneumonia risks correspond to observed diagnostic outcomes.

C3– *Explainable diagnostic modeling for clinician-facing interpretation:* We combine global SHAP analysis, local LIME explanations, and sensitivity-oriented interpretation to provide complementary views of model behavior, supporting both cohort-level diagnostic understanding and patient-specific explanation.

C4– *Unified integration of prediction, calibration, explainability, and decision control:* Rather than treating diagnostic accuracy, reliability, interpretability, and decision thresholds as separate evaluation steps, the framework integrates them into a coherent diagnostic decision-support workflow.

C5– *Comprehensive evaluation under class imbalance and clinical screening constraints:* The proposed system is evaluated using classification metrics, confusion matrix analysis, calibration curves, and threshold-dependent sensitivity–specificity trade-offs, providing a multidimensional assessment consistent with clinical diagnostic requirements.

Together, these contributions position the proposed framework as a clinically oriented diagnostic AI approach for pneumonia detection, designed to balance predictive performance with reliability, transparency, and actionable decision support.

 doi:10.20944/preprints202605.1557.v1

1.3. Structure of the Article

The remainder of this article is organized as follows. Section 2 reviews related work on pneumonia detection, explainable AI, calibration-aware medical imaging, and clinical decision-support systems, while also identifying the research gaps addressed in this study. Section 3 presents the proposed UVIF-guided diagnostic framework, including the theoretical formulation, CNN-based representation learning, calibration analysis, explainability mechanisms, and threshold-aware decision support. Section 4 reports the experimental evaluation, covering predictive performance, calibration reliability, explainability analysis, and UVIF-guided model and threshold selection. Finally, Section 5 summarizes the main findings, discusses limitations, and outlines future directions for trustworthy and clinically meaningful AI-assisted pneumonia screening systems.

2. Related Works

2.1. Current Research

Machine learning approaches for pneumonia have evolved across multiple task categories, including detection, prognosis, pathogen identification, severity prediction, and clinical triage, using both structured clinical data and imaging modalities. In symptom-based and structured-data settings, interpretable models have been proposed to support diagnosis in resource-constrained environments, demonstrating the feasibility of lightweight decision-support systems [8]. In imaging-based applications, deep learning models have achieved high diagnostic performance, while hybrid approaches combining deep representations with explainable components have been introduced to improve interpretability [9]. In addition, recent work on pathogen identification using routine blood tests has demonstrated that structured ensemble models can achieve strong discrimination while maintaining interpretability [10]. These developments highlight the versatility of machine learning across heterogeneous pneumonia-related tasks.

A complementary line of work has focused on generalization across sites, cohorts, and care environments. Zech et al. showed that a deep learning model for pneumonia detection in chest radiographs may perform strongly in internal testing but degrade substantially under external validation, partly because the model can learn site-specific acquisition or prevalence patterns rather than disease-specific representations [11]. More recent multi-database studies have attempted to address this issue by validating pneumonia mortality models across several healthcare systems; for example, Chen et al. developed interpretable XGBoost-based mortality prediction models using MIMIC-IV and externally evaluated them on MIMIC-III, eICU, and a Chinese cohort [12]. Similarly, Yang et al. proposed a multicenter severe pneumonia prediction model using data from seven medical centers, combining predictive modeling with decision-curve analysis to assess clinical usefulness [13]. These studies emphasize that pneumonia AI systems must be evaluated not only for internal accuracy but also for robustness across heterogeneous populations and institutional contexts.

In parallel, explainability has become an essential component of clinical AI systems. Several studies have integrated SHAP-based global explanations and local interpretability tools into predictive models to provide insight into feature contributions and decision mechanisms. For example, machine learning models for ventilator-associated pneumonia have incorporated feature importance analysis to improve clinical interpretability [14]. Similarly, interpretable models for postoperative pneumonia prediction emphasize the importance of transparent feature relationships for clinical adoption [15]. Beyond feature-based explanations, recent work has explored calibration-aware counterfactual explanations, enabling more robust interpretation of deep learning models in medical imaging [16]. However, despite these advances, explainability remains largely descriptive, with limited formal guarantees regarding consistency, stability, or clinical validity.

Recent methodological work also stresses the close relationship between explainability and uncertainty. Salvi et al. argued that explainability and uncertainty quantification should be considered jointly in healthcare AI because explanations alone may be insufficient when the model is uncertain or operating outside familiar data regimes [17]. In a broader medical AI context, Atf et al. discussed the

 doi:10.20944/preprints202605.1557.v1

challenge of uncertainty quantification for large language models in medicine, emphasizing the need to distinguish confident outputs from reliable outputs [18]. Wang et al. further highlighted the distinction between aleatoric and epistemic uncertainty, which is particularly relevant in clinical settings where noise, missingness, and population shift may affect model confidence [19]. Although these studies are not limited to pneumonia, they provide an important foundation for trust-aware pneumonia detection frameworks in which probability, uncertainty, and explanation must be interpreted together.

Another critical dimension is the reliability of predicted probabilities and their role in decision-making. While some studies report calibration metrics such as Brier scores and reliability curves, these aspects are often secondary to predictive performance. Furthermore, class imbalance and class weighting strategies can significantly distort probability estimates, leading to miscalibrated predictions and unreliable risk assessment [20]. In response, recent research has begun to incorporate decision-oriented evaluation frameworks, including dynamic prediction models and decision-curve analysis, particularly in ventilator-associated pneumonia and longitudinal monitoring scenarios [6]. These approaches represent an important step toward integrating prediction with clinical decision-making.

Temporal modeling is another important direction in pneumonia-related machine learning. Pneumonia progression is not static; symptoms, vital signs, oxygenation, laboratory measurements, and treatment response evolve over time. Agard et al. proposed PREDICT, a deep learning approach for ventilator-associated pneumonia prediction in intensive care units, using rolling physiological windows to support early warning before clinical confirmation [21]. Lozano-Rojas et al. explored community-acquired pneumonia outcome prediction using time-series clinical data, showing how longitudinal information can improve the modeling of disease trajectory [22]. Sheu et al. introduced a multimodal framework combining vital sign time series and chest X-ray information for pneumonia status prediction, demonstrating the value of integrating temporal and imaging information in recovery-oriented decision support [23]. These works suggest that pneumonia detection and triage can benefit from moving beyond static prediction toward dynamic risk assessment.

A further strand of research focuses on embedding prediction models within actionable clinical workflows. Free et al. proposed a data-driven clinical decision-support framework for pneumonia management, emphasizing risk stratification, visualization, and patient prioritization rather than isolated prediction [24]. Dean et al. evaluated a real-time pneumonia clinical decision-support system in a pragmatic stepped-wedge trial, showing that model-driven tools can influence antibiotic concordance and emergency department disposition when integrated into clinical workflows [25]. Earlier work by Dean et al. also demonstrated the impact of electronic decision-support tools for emergency department pneumonia patients, highlighting the importance of guideline integration and practical usability [26]. Williams et al. later evaluated prognostic clinical decision support for pediatric pneumonia in the emergency department, illustrating that clinical effectiveness depends not only on model quality but also on workflow integration, clinician adoption, and implementation context [27].

Despite these advances, the literature reveals a clear fragmentation between predictive modeling, calibration, explainability, uncertainty awareness, temporal modeling, and decision control. Most existing studies address these components independently rather than within a unified framework. As a result, high predictive performance does not necessarily translate into reliable or interpretable clinical decisions. This gap is particularly critical in pneumonia detection, where the cost of false negatives is high and decision thresholds must be explicitly controlled. The present work addresses this limitation by proposing an integrated framework that jointly considers prediction, calibration, explainability, and threshold-aware decision-making within a single, coherent pipeline.

2.2. Research Gaps

Despite significant progress in machine learning for pneumonia detection and prognosis, several limitations remain in current approaches. Many existing studies focus on isolated aspects such as prediction accuracy, explainability, or calibration, without integrating them into a unified decision-support framework. To clarify these limitations, we identify the following research gaps (G1–G5):

 doi:10.20944/preprints202605.1557.v1

G1– *Prediction-centered models with limited decision awareness:* Most pneumonia detection models are optimized for accuracy or AUC without explicitly considering how predictions are translated into clinical decisions, particularly under varying sensitivity–specificity requirements.

G2– *Insufficient integration of calibration into predictive pipelines:* Although some studies report calibration metrics, probability reliability is often treated as a secondary evaluation step rather than a core component of the modeling framework, limiting the interpretability of predicted risks.

G3– *Fragmented use of explainability techniques:* Explainable AI methods such as SHAP and LIME are widely used, but they are typically applied independently rather than combined into a coherent multi-level interpretability strategy that links global, local, and sensitivity-based insights.

G4– *Limited incorporation of threshold-aware decision control:* Most models operate with fixed decision thresholds, without systematically analyzing or optimizing threshold-dependent behavior, despite its importance for balancing false positives and false negatives in clinical practice.

G5– *Lack of unified trust-aware decision-support frameworks:* Existing studies often evaluate prediction, calibration, and explainability as separate components, with limited integration into a single pipeline that supports transparent, reliable, and clinically actionable decision-making.

Addressing these gaps requires a shift from purely predictive modeling toward integrated, decision-centric frameworks that jointly consider performance, calibration, interpretability, and decision control. This motivates the proposed methodology, which combines structured machine learning, calibration analysis, multi-level explainability, and threshold-aware decision-making to support trust-aware pneumonia detection.

3. Methodology

3.1. Theoretical Framework

Let $\mathbf{I}_i \in \mathbb{R}^{H\times W}$ denote the i-th PneumoniaMNIST chest X-ray image, and let $y_i \in \{0,1\}$ denote the corresponding binary label, where $y_i = 0$ represents Normal and $y_i = 1$ represents Pneumonia. The objective is not only to estimate the probability of pneumonia from low-resolution chest X-ray images, but also to support diagnostics-oriented decision-making under class imbalance, imperfect calibration, and asymmetric clinical risk.

A compact convolutional neural network (CNN) is used to learn image-level radiographic representations. Let

$$\mathbf{z}_i = g_\theta(\mathbf{I}_i) \tag{1}$$

denote the latent embedding extracted from the CNN, where $g_\theta(\cdot)$ is the feature extractor parameterized by θ and $\mathbf{z}_i \in \mathbb{R}^d$ is the learned image-derived embedding vector. A classifier $h_\psi(\cdot)$ then maps this representation to a pneumonia probability:

$$p_\psi(y_i = 1 \mid \mathbf{I}_i) = h_\psi(\mathbf{z}_i). \tag{2}$$

The predicted class label is obtained using a decision threshold τ:

$$\hat{y}_i = \mathbb{I}\big(p_\psi(y_i = 1 \mid \mathbf{I}_i) \geq \tau\big), \tag{3}$$

where $\mathbb{I}(\cdot)$ is the indicator function. Varying τ enables the operating point of the model to be adjusted according to diagnostic priorities. Lower thresholds generally favor sensitivity and false-negative reduction, whereas higher thresholds may increase specificity and reduce unnecessary follow-up.

Calibration is assessed by comparing predicted pneumonia probabilities with empirical pneumonia frequencies. For a probability bin B_m, calibration reliability is evaluated using:

$$\mathrm{conf}(B_m) = \frac{1}{|B_m|}\sum_{i\in B_m} p_\psi(y_i = 1 \mid \mathbf{I}_i), \tag{4}$$

 doi:10.20944/preprints202605.1557.v1

$$\text{freq}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} y_i. \tag{5}$$

A well-calibrated binary diagnostic model should satisfy $\text{conf}(B_m) \approx \text{freq}(B_m)$ across probability bins. This formulation evaluates whether predicted pneumonia probabilities correspond to observed pneumonia frequencies, rather than whether thresholded labels are correct. In screening-oriented medical AI, this distinction is important because strong ROC-AUC does not guarantee that predicted probabilities can be interpreted as reliable diagnostic risk estimates.

Explainability is modeled at two complementary levels. For the CNN, Grad-CAM and LIME identify image regions that contribute to pneumonia prediction. For the embedding-based classifiers, SHAP attribution estimates the contribution of latent embedding dimensions:

$$\Phi(\mathbf{z}_i) = (\phi_1(\mathbf{z}_i), \phi_2(\mathbf{z}_i), \ldots, \phi_d(\mathbf{z}_i)), \tag{6}$$

where $\phi_j(\mathbf{z}_i)$ denotes the contribution of the j-th embedding dimension to the prediction. These explanation mechanisms are later used to support diagnostic interpretation of the model rather than treated only as qualitative visual outputs.

To organize prediction, calibration, explainability, false-negative control, and operating-point selection, this work introduces the Unified Variational Intelligence Framework (UVIF) as a restrained diagnostics-oriented decision-reliability layer. UVIF is not proposed as a new classifier architecture, nor is the full mathematical UVIF functional optimized end-to-end during model training. Instead, UVIF provides a structured post-training utility for comparing trained models and selecting diagnostic operating thresholds according to screening-oriented reliability criteria.

The variational interpretation of UVIF is operationalized through constrained model and threshold selection:

$$(m^*, \tau^*) = \arg \max_{m \in \mathcal{M}, \tau \in \mathcal{T}} \mathcal{J}_{\text{UVIF}}(m, \tau), \tag{7}$$

where $\mathcal{M}$ denotes the set of candidate models and $\mathcal{T}$ denotes the set of candidate diagnostic thresholds. In this formulation, the "variational" aspect refers to the optimization of a diagnostic utility functional over competing operating configurations rather than to end-to-end variational learning.

The UVIF-guided diagnostics utility is defined as:

$$\mathcal{J}_{\text{UVIF}} = \alpha \mathcal{L}_{\text{AUC}} + \beta \mathcal{L}_{\text{Sens}} + \gamma \mathcal{L}_{\text{Spec}} + \delta \mathcal{L}_{\text{F1}} + \eta \mathcal{L}_{\text{Cal}} + \lambda \mathcal{L}_{\text{XAI}} - \mu \mathcal{R}_{\text{FN}} - \nu \mathcal{C}_{\text{model}}, \tag{8}$$

where $\mathcal{L}_{\text{AUC}}$ represents discrimination performance, $\mathcal{L}_{\text{Sens}}$ represents pneumonia sensitivity, $\mathcal{L}_{\text{Spec}}$ represents normal-class specificity, $\mathcal{L}_{\text{F1}}$ captures balanced classification utility, $\mathcal{L}_{\text{Cal}}$ represents calibration reliability, $\mathcal{L}_{\text{XAI}}$ denotes explainability utility, $\mathcal{R}_{\text{FN}}$ penalizes false negatives, and $\mathcal{C}_{\text{model}}$ penalizes model complexity.

In the present implementation, calibration reliability is expressed as $1 - \text{ECE}$, false-negative risk is measured by the false-negative rate, and complexity and explainability are represented using transparent practical priors. These terms are used for post-training model ranking and threshold selection. This makes the UVIF layer interpretable and reproducible while preserving its restrained role as a diagnostic decision-control mechanism rather than an end-to-end learning objective.

3.2. UVIF as a Diagnostic Principle: From Natural Sensing to Engineered Medical Decision Support

The motivation for introducing UVIF in this work is grounded in a broader diagnostic principle: any adaptive system operating under uncertainty must sense, evaluate, and respond to environmental information. In physical systems, this occurs through field gradients; in biological systems, through sensory mechanisms; and in engineered AI systems, through data-driven inference and decision control. Although these systems differ in implementation, they share a common variational structure: information is acquired, uncertainty is evaluated, risk is assessed, and action is selected under complexity constraints.

 doi:10.20944/preprints202605.1557.v1

Within this perspective, medical diagnosis can be interpreted as an engineered sensing problem. Chest X-ray images provide indirect field-like evidence of pathological changes, while the AI model extracts informative representations, estimates confidence through probabilistic outputs, evaluates clinical risk through threshold-dependent decisions, and supports interpretation through explainability mechanisms. In this implementation, UVIF makes this process explicit by organizing prediction, calibration, risk, complexity, and information utility into a post-training diagnostic utility. Therefore, the proposed framework does not merely classify pneumonia cases; it evaluates trained models and decision thresholds as components of a reliability-aware diagnostic process in which sensitivity, probability calibration, explainability, and decision thresholds are jointly considered.

This interpretation is particularly relevant for high-sensitivity pneumonia screening, where the cost of missed cases is clinically important. The UVIF layer provides a principled mechanism for selecting models and thresholds that balance discrimination performance, false-negative risk, calibration reliability, and interpretability. In this way, UVIF reframes the proposed pipeline from a conventional image-classification workflow into a diagnostic decision-support framework aligned with the broader logic of sensing, evaluation, and adaptive response.

3.3. Proposed Method

The proposed method is an image-based pneumonia screening pipeline built on the PneumoniaMNIST benchmark. The framework combines compact CNN learning, embedding-based machine learning, calibration analysis, explainability, threshold-aware prediction, and UVIF-guided diagnostics. Rather than relying only on accuracy or ROC-AUC, the methodology evaluates whether model outputs are sensitive, reliable, interpretable, and suitable for screening-oriented decision support.

The workflow consists of eight main stages: image preprocessing, train–validation–test partitioning, compact CNN training, embedding extraction, embedding-based classifier training, calibration analysis, explainability analysis, and UVIF-guided post-training model and threshold selection.

3.4. Dataset and Image Preprocessing

The study uses the PneumoniaMNIST dataset, a benchmark chest X-ray dataset from the MedMNIST collection. Each sample is a low-resolution grayscale chest X-ray image labeled as Normal or Pneumonia. Image preprocessing includes conversion to a consistent numerical format, normalization of pixel intensities, and reshaping of images to match the CNN input structure.

No synthetic oversampling is applied, so the original class imbalance is preserved. This enables evaluation under realistic imbalance conditions and motivates the use of recall, F1-score, balanced accuracy, calibration, and confusion matrix analysis in addition to overall accuracy.

3.5. Train–Validation–Test Partitioning

The PneumoniaMNIST dataset is divided into training, validation, and test subsets according to the benchmark split. The training set is used for model fitting, the validation set is used to monitor CNN learning dynamics and reduce overfitting, and the test set is reserved for final evaluation. The class distribution is retained across the splits to reflect the natural imbalance of the dataset.

3.6. Compact CNN Training

A compact CNN is trained directly on the PneumoniaMNIST images to learn discriminative radiographic representations. The CNN consists of convolutional layers for spatial feature extraction, batch normalization, pooling operations, global average pooling, dropout, and a sigmoid output layer for binary classification. During training, the CNN outputs a pneumonia probability for each image. The model is optimized using binary classification loss, and training behavior is monitored through accuracy, AUC, and loss curves on both training and validation sets.

The compact CNN serves two purposes. First, it provides an end-to-end image classifier. Second, it produces latent image embeddings that are later used by conventional machine learning classifiers.

 doi:10.20944/preprints202605.1557.v1

3.7. Embedding Extraction

After CNN training, latent representations are extracted from the global average pooling layer. These embeddings encode image-derived features learned by the CNN and are denoted as:

$$\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N]^T \in \mathbb{R}^{N \times d}. \tag{9}$$

The resulting embedding matrix is used as input to classical machine learning classifiers. This design combines the representation-learning capacity of CNNs with the comparative flexibility and interpretability of embedding-based classifiers.

3.8. Embedding-Based Classifier Training

Multiple supervised classifiers are trained on the CNN-derived embeddings, including Logistic Regression, Random Forest, ExtraTrees, XGBoost, LightGBM, and SoftVoting. Logistic Regression provides a linear baseline, while tree-based ensemble models capture nonlinear relationships among embedding dimensions. The classifiers are evaluated using accuracy, balanced accuracy, precision, recall, F1-score, ROC-AUC, log loss, and expected calibration error.

3.9. Model Selection and Threshold-Aware Prediction

Each trained model outputs a pneumonia probability for each test image. A threshold τ is then applied to convert probabilities into binary decisions. The conventional threshold $\tau = 0.5$ is used for baseline comparison, while UVIF-guided threshold optimization is used to identify a diagnostics-oriented operating point. This distinction allows the study to separate predictive discrimination from practical screening behavior.

3.10. Calibration and Reliability Analysis

Calibration analysis is performed to assess whether predicted pneumonia probabilities correspond to empirical pneumonia frequencies. Reliability curves and bin-wise calibration statistics are used to compare mean predicted confidence with observed positive-class frequency in each probability bin. This step is essential because a model can achieve strong ROC-AUC while still producing poorly calibrated probabilities. Calibration is therefore treated as a core reliability assessment rather than a secondary visualization.

3.11. Explainability Analysis

Explainability is evaluated using complementary visual and embedding-level methods. Grad-CAM is applied to the compact CNN to identify image regions that contribute to pneumonia predictions. LIME provides local image-level explanations for selected cases. SHAP is applied to the embedding-based LightGBM classifier to estimate the global importance of latent image-derived features. Together, these methods provide local visual explanation, global latent-space attribution, and diagnostic transparency.

3.12. UVIF-Guided Diagnostics Layer

After conventional performance evaluation, the UVIF-guided diagnostics layer ranks candidate models using the post-training utility functional in Eq. (8). The implementation uses the following practical weighting scheme: ROC-AUC (0.20), sensitivity (0.25), specificity (0.15), F1-score (0.15), calibration reliability (0.10), explainability utility (0.10), false-negative penalty (0.15), and complexity penalty (0.05). These weights reflect the screening-oriented priority of preserving high pneumonia sensitivity while maintaining interpretability and avoiding excessive diagnostic burden.

The UVIF layer is also used to optimize the threshold of the selected model. Candidate thresholds are evaluated by combining sensitivity, specificity, F1-score, false-negative rate, false-positive rate, calibration reliability, and diagnostic utility. This allows the framework to identify an operating point that is more appropriate for screening-oriented decision support than an arbitrary fixed threshold. The

 doi:10.20944/preprints202605.1557.v1

selected threshold should therefore be interpreted as a post-training diagnostic operating point, not as evidence that the classifier itself was trained through UVIF optimization.

3.13. Unified Staged Framework

The proposed methodology follows a staged design rather than a single monolithic classifier. The framework links image representation learning, embedding-based classification, calibration assessment, threshold control, explainability, and UVIF-guided diagnostics in a modular pipeline. This structure ensures that each model output is evaluated not only for discrimination but also for reliability, interpretability, and decision utility.

As illustrated in Figure 1, the full pipeline consists of the following stages: image input and preprocessing, compact CNN training, embedding extraction, embedding-based classification, calibration and reliability assessment, threshold-aware decision analysis, explainability analysis, and post-training UVIF-guided diagnostics model selection.

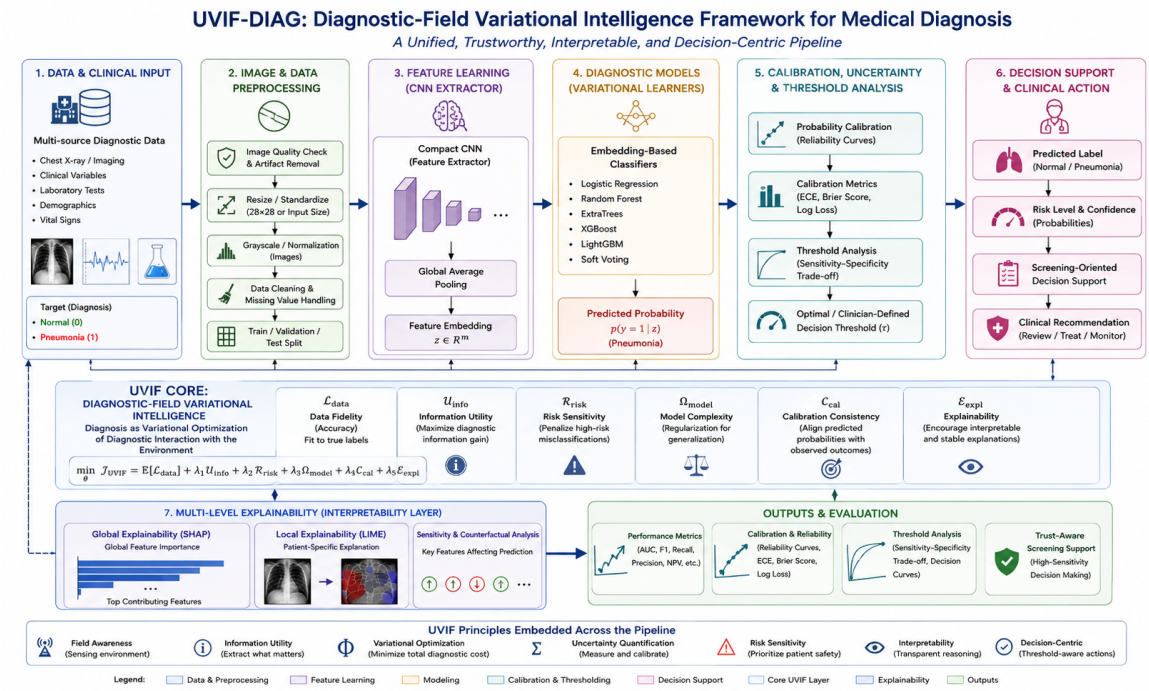


Figure 1. Conceptual overview of the proposed UVIF-guided diagnostics-oriented framework for pneumonia screening using PneumoniaMNIST chest X-ray images. The pipeline integrates image preprocessing, compact CNN training, embedding extraction, supervised classification, calibration analysis based on empirical positive-class frequency, threshold-aware decision control, explainability, and post-training UVIF-guided model and threshold selection.

3.14. Algorithmic Description

Algorithm 1 summarizes the full image-based workflow. The algorithm begins with PneumoniaMNIST image preprocessing, followed by compact CNN training and embedding extraction. The extracted embeddings are then used to train and compare supervised classifiers. Finally, calibration, explainability, UVIF-guided model selection, and UVIF-guided threshold optimization are performed as post-training diagnostic evaluation steps.

 doi:10.20944/preprints202605.1557.v1

Algorithm 1 UVIF-Guided Diagnostics-Oriented AI Framework for Pneumonia Screening

Require: PneumoniaMNIST image dataset $\{(\mathbf{I}_i, y_i)\}_{i=1}^{N}$, where $y_i \in \{0, 1\}$
Ensure: Predicted pneumonia probabilities, diagnostic labels, calibration results, explainability outputs, UVIF-selected model, and UVIF-selected threshold

Normalize and reshape PneumoniaMNIST images
Preserve benchmark train, validation, and test splits
Initialize compact CNN model $g_\theta(\cdot)$
for each training epoch **do**
    for each mini-batch of images **do**
        Compute pneumonia probabilities from the CNN
        Compute binary classification loss
        Update CNN parameters using the training loss
    end for
    Monitor training and validation accuracy, AUC, and loss
end for
Extract latent embeddings $\mathbf{z}_i = g_\theta(\mathbf{I}_i)$ from the trained CNN
Construct embedding matrix $\mathbf{Z} \in \mathbb{R}^{N \times d}$
Train Logistic Regression, Random Forest, ExtraTrees, XGBoost, LightGBM, and SoftVoting classifiers on CNN-derived embeddings
for each trained model **do**
    Predict pneumonia probability $p(y = 1 \mid \mathbf{I}_i)$ on the test set
    Assign label $\hat{y}_i$ using threshold $\tau = 0.5$ for baseline evaluation
    Compute accuracy, balanced accuracy, precision, recall, F1-score, ROC-AUC, log loss, and ECE
    Compute calibration bins using mean predicted probability and empirical pneumonia frequency
    Compute confusion matrix values: TN, FP, FN, and TP
    Compute post-training UVIF diagnostics score
end for
Select the model with the highest post-training UVIF diagnostics score
Optimize the decision threshold using UVIF-guided diagnostic utility
Generate reliability curves and calibration bin statistics
Generate Grad-CAM and LIME explanations for representative cases
Compute SHAP attribution scores for embedding-based classifiers
Interpret results from a screening-oriented diagnostics perspective

3.15. Reproducibility and Research Artifacts

All experiments were implemented in Python using widely adopted scientific computing, deep learning, machine learning, and explainable artificial intelligence libraries, including `NumPy`, `Pandas`, `Matplotlib`, `scikit-learn`, `TensorFlow/Keras`, `OpenCV`, `SHAP`, and `LIME`. Visual explainability analyses were additionally supported through Grad-CAM implementations for convolutional neural network interpretation. The workflow is compatible with both Google Colab and standard Python environments, enabling transparent reproduction of preprocessing, feature extraction, embedding generation, calibration analysis, explainability evaluation, and threshold-aware diagnostic decision support.

The experimental evaluation is conducted using the publicly available *PneumoniaMNIST* dataset from the MedMNIST benchmark collection [28]. The dataset contains chest X-ray images organized for binary pneumonia classification and provides a standardized benchmark for reproducible medical imaging experiments. To maintain consistency and reproducibility, the present study focuses on this publicly accessible dataset with stable preprocessing protocols and clearly documented benchmark characteristics.

The complete implementation of the proposed Diagnostic-Field Variational Intelligence Framework (UVIF-DIAG), including preprocessing pipelines, CNN-based embedding extraction, embedding-level diagnostic classifiers, calibration-aware reliability analysis, threshold-sensitive decision control, multi-level explainability modules, and automated figure/table regeneration scripts, is publicly avail-

 doi:10.20944/preprints202605.1557.v1

able on GitHub [29]. Structured Colab-ready notebooks are provided to reproduce the reported experiments, calibration studies, sensitivity–specificity analyses, and explainability evaluations.

To support deterministic evaluation and facilitate independent validation, experiments were executed using fixed random seeds, documented software dependencies, and controlled training configurations where applicable. Experimental artifacts, including performance metrics, confusion matrices, calibration curves, reliability statistics, explainability visualizations, threshold-analysis plots, generated tables, and diagnostic summaries, are produced programmatically and stored automatically during execution. The workflow additionally generates consolidated output logs and summary files, including an `outputs_summary.txt` file containing the principal predictive, calibration, explainability, and decision-support indicators.

This reproducibility package supports transparent validation, methodological extension, and future development of trustworthy, interpretable, and calibration-aware AI systems for pneumonia diagnosis and clinical decision support.

4. Results and Discussion

This section evaluates the proposed UVIF-guided diagnostics-oriented PneumoniaMNIST framework using representative image inspection, compact CNN learning dynamics, embedding-based classification, calibration analysis, explainability, and UVIF-guided diagnostic decision control. The results are interpreted as evidence for benchmark-level screening-oriented decision support, not as evidence of clinical deployment readiness.

4.1. Dataset Characteristics and Representative Samples

Figure 2 illustrates representative PneumoniaMNIST chest X-ray images. Table 1 confirms that pneumonia cases are more frequent than normal cases across the training, validation, and test splits. This imbalance motivates the use of recall, balanced accuracy, calibration, threshold-aware diagnostics, and confusion-matrix analysis rather than accuracy alone.

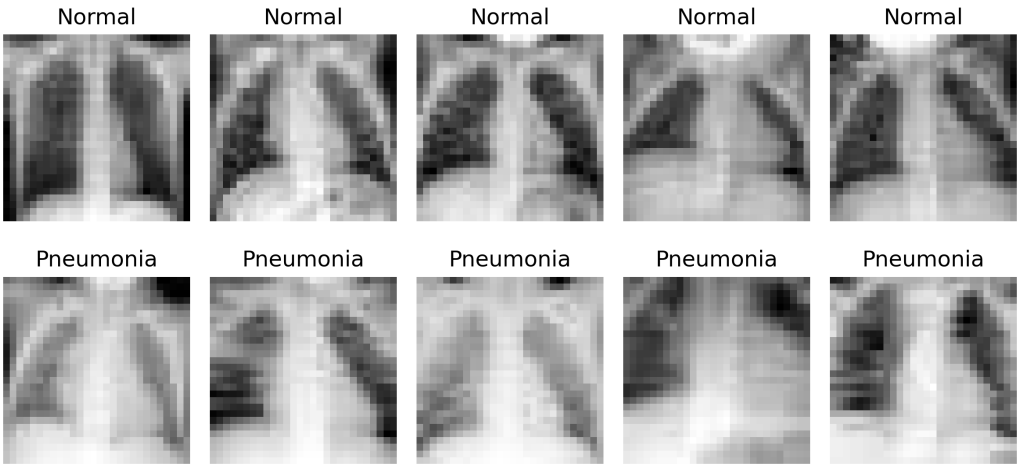


Figure 2. Representative PneumoniaMNIST chest X-ray images from the Normal and Pneumonia classes.

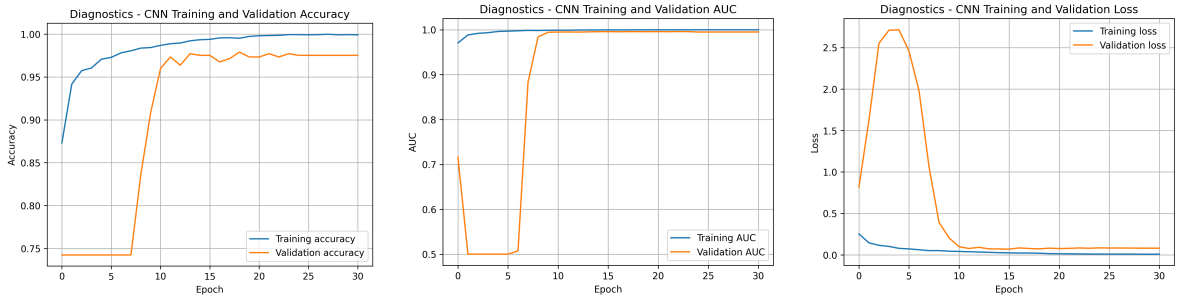
Table 1. Class distribution across training, validation, and test splits.

Split	Class	Label	Count
Train	Normal	0	1214
Train	Pneumonia	1	3494
Validation	Normal	0	135
Validation	Pneumonia	1	389
Test	Normal	0	234
Test	Pneumonia	1	390

 doi:10.20944/preprints202605.1557.v1

4.2. CNN Training Dynamics

Figure 3 shows stable compact CNN training, with consistent improvement in accuracy and AUC and a decreasing loss profile. Table 2 shows that the compact CNN achieved accuracy of 0.8846, balanced accuracy of 0.8470, F1-score of 0.9153, ROC-AUC of 0.9666, and pneumonia recall of 0.9974. These results indicate strong screening-oriented discrimination. However, the ECE value of 0.3348 shows that the model probabilities are not fully calibrated, reinforcing the need for calibration-aware and UVIF-guided diagnostics.



(a) Accuracy (b) AUC (c) Loss

Figure 3. Compact CNN training dynamics on the PneumoniaMNIST dataset: training and validation accuracy, AUC, and loss curves.

Table 2. Compact CNN test performance.

Model	Accuracy	Balanced Acc.	Precision	Recall	F1-score	ROC-AUC	Log Loss	ECE
Compact CNN	0.8846	0.8470	0.8457	0.9974	0.9153	0.9666	0.4157	0.3348

4.3. Overall Model Performance

Table 3 shows strong performance across both end-to-end and embedding-based models. The compact CNN achieved the highest F1-score and overall accuracy, while Logistic Regression embeddings achieved the highest ROC-AUC among the embedding-based classifiers. ExtraTrees and RandomForest embeddings achieved the lowest ECE values among the embedding models. These results demonstrate that CNN-derived embeddings preserve diagnostically meaningful latent information, but no single conventional metric is sufficient to determine the most reliable diagnostic operating model.

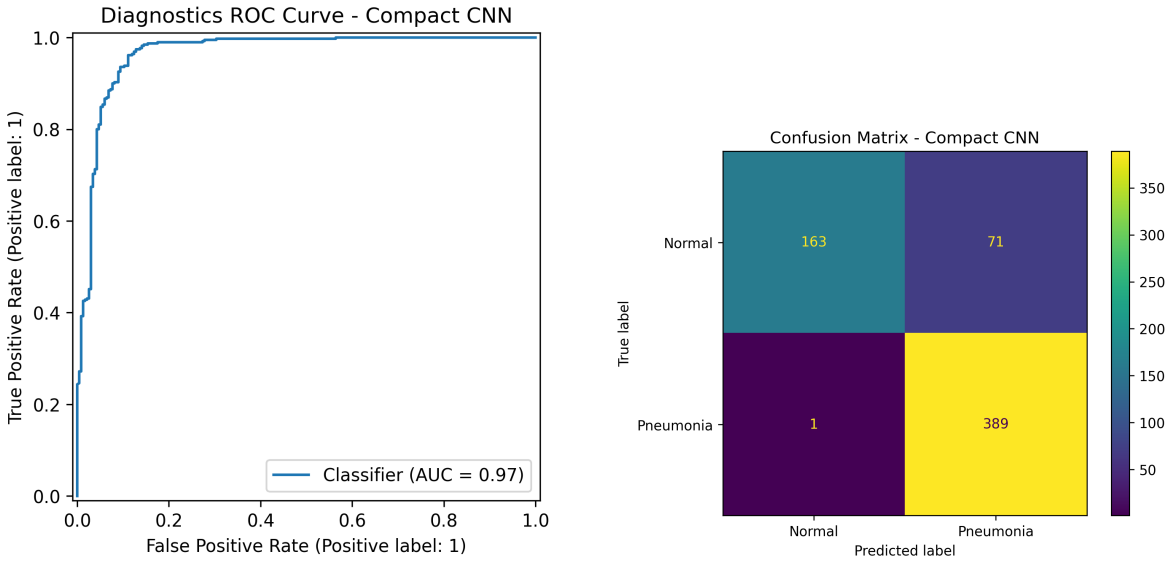
Table 3. Performance of all evaluated models on the PneumoniaMNIST test set.

Model	Accuracy	Balanced Acc.	Precision	Recall	F1-score	ROC-AUC	Log Loss	ECE
Compact CNN	0.8846	0.8470	0.8457	0.9974	0.9153	0.9666	0.4157	0.3348
LightGBM Embeddings	0.8670	0.8235	0.8259	0.9974	0.9036	0.9515	1.0816	0.3646
ExtraTrees Embeddings	0.8638	0.8192	0.8224	0.9974	0.9015	0.9650	0.9548	0.3194
SoftVoting Embeddings	0.8638	0.8192	0.8224	0.9974	0.9015	0.9618	0.5585	0.3372
XGBoost Embeddings	0.8622	0.8171	0.8207	0.9974	0.9005	0.9535	0.7085	0.3522
RandomForest Embeddings	0.8606	0.8158	0.8203	0.9949	0.8992	0.9607	1.0710	0.3197
LogisticRegression Embeddings	0.8494	0.8000	0.8071	0.9974	0.8922	0.9694	0.9513	0.3549

4.4. Discrimination, Confusion Matrix, and Calibration Analysis

Figure 4 and Table 4 show that the compact CNN detects 389 of 390 pneumonia cases at the conventional threshold, producing only one false negative. This behavior is consistent with a screening-oriented model, where false-negative minimization is prioritized. However, the 71 false positives indicate that threshold control is needed to reduce diagnostic burden.

 doi:10.20944/preprints202605.1557.v1



(a) ROC curve (b) Confusion matrix at $\tau = 0.5$

Figure 4. Discrimination and confusion-matrix analysis of the compact CNN at the conventional threshold.

Table 4. Confusion matrix values for the compact CNN at $\tau = 0.5$.

Model	TN	FP	FN	TP
Compact CNN	163	71	1	389

Figure 5 and Table 5 show that discrimination and calibration are not equivalent. Although the compact CNN achieves strong ROC-AUC, its probability estimates deviate from ideal reliability in several probability ranges. This finding is clinically important because probability outputs should not be interpreted as calibrated diagnostic risk scores without additional reliability analysis. In particular, the low-confidence bins exhibit substantial deviations between predicted confidence and empirical pneumonia frequency, indicating that some samples assigned low probabilities are nevertheless pneumonia-positive. Such behavior represents a potentially important clinical failure mode in screening-oriented AI systems and motivates the use of post-hoc calibration strategies such as temperature scaling, isotonic regression, or Platt scaling. This motivates the inclusion of calibration reliability in the UVIF diagnostic utility.

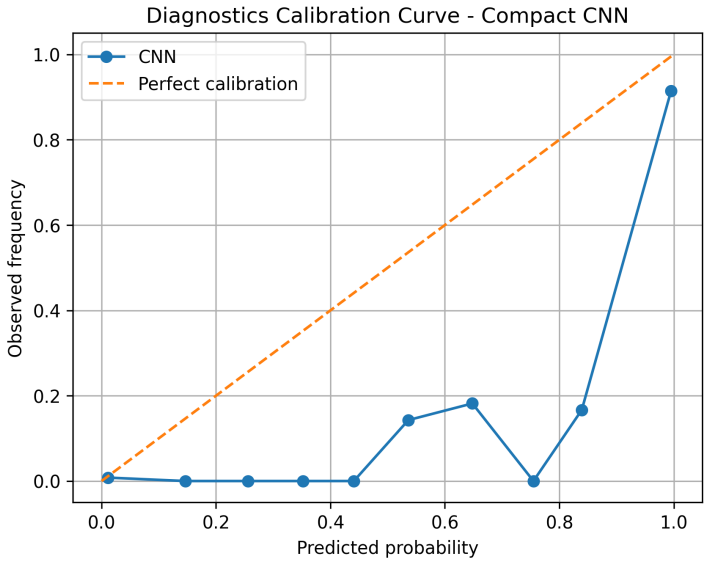


Figure 5. Calibration curve of the compact CNN classifier.

 doi:10.20944/preprints202605.1557.v1

Table 5. Calibration bin summary for the compact CNN. The empirical column should represent observed pneumonia frequency within each probability bin.

Bin Low	Bin High	Count	Empirical Frequency	Confidence
0.0	0.1	35	1.0000	0.0352
0.1	0.2	20	1.0000	0.1527
0.2	0.3	22	1.0000	0.2495
0.3	0.4	28	0.9643	0.3499
0.4	0.5	58	0.9310	0.4511
0.5	0.6	119	0.6303	0.5532
0.6	0.7	39	0.6923	0.6451
0.7	0.8	40	0.8250	0.7566
0.8	0.9	33	0.8182	0.8492
0.9	1.0	230	0.9826	0.9763

4.5. UVIF-Guided Diagnostics Model Selection and Threshold Optimization

Figure 6 and Table 6 present the UVIF-guided diagnostics ranking. Unlike conventional model selection, the UVIF layer jointly evaluates discrimination, sensitivity, specificity, F1-score, calibration reliability, explainability utility, false-negative risk, and complexity. The compact CNN achieved the highest UVIF diagnostics score of 0.8081 and was therefore selected as the most reliable diagnostic operating model. This result is important because the model was not selected purely on ROC-AUC; rather, it was selected through a decision-reliability functional that reflects screening-oriented priorities.

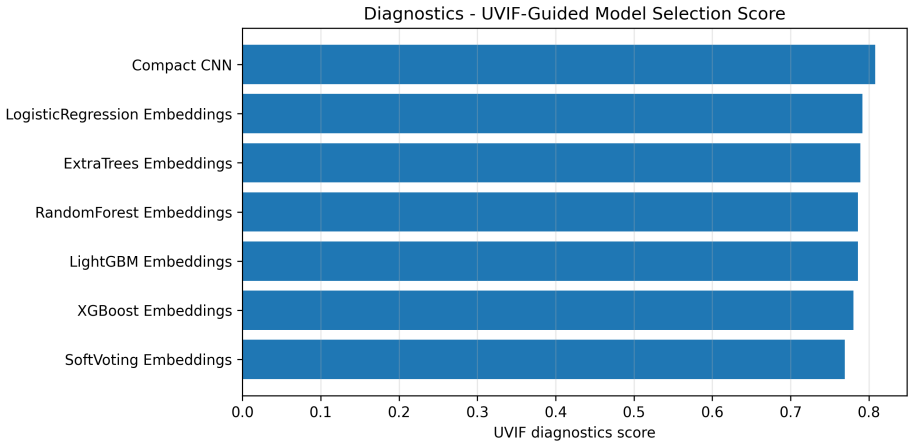


Figure 6. Diagnostics-oriented UVIF-guided model selection scores integrating discrimination, sensitivity, specificity, calibration reliability, explainability utility, false-negative penalty, and complexity regularization.

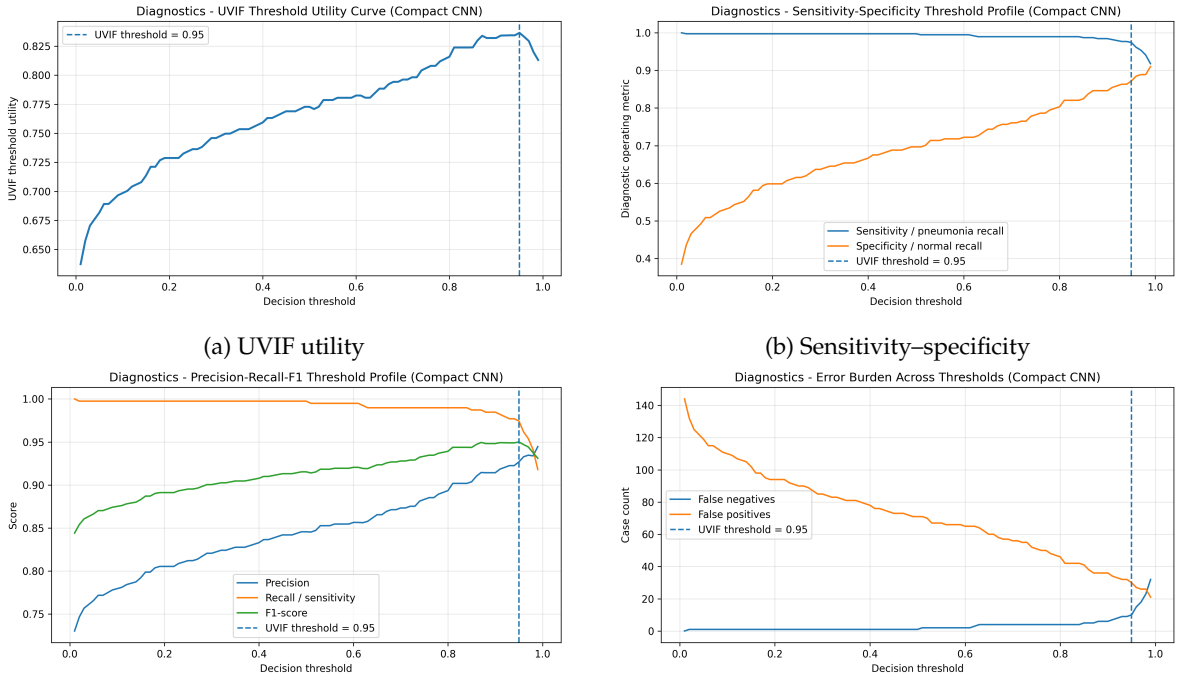
Table 6. UVIF-guided diagnostics model-selection analysis.

Model	UVIF Score	ROC-AUC	Sensitivity	Specificity	F1-score	ECE
Compact CNN	0.8081	0.9666	0.9974	0.6966	0.9153	0.3348
LogisticRegression Embeddings	0.7916	0.9694	0.9974	0.6026	0.8922	0.3549
ExtraTrees Embeddings	0.7889	0.9650	0.9974	0.6410	0.9015	0.3194
RandomForest Embeddings	0.7860	0.9607	0.9949	0.6368	0.8992	0.3197
LightGBM Embeddings	0.7858	0.9515	0.9974	0.6496	0.9036	0.3646
XGBoost Embeddings	0.7800	0.9535	0.9974	0.6368	0.9005	0.3522
SoftVoting Embeddings	0.7690	0.9618	0.9974	0.6410	0.9015	0.3372

Figure 7 and Table 7 show the effect of UVIF-guided threshold optimization. Instead of relying on the conventional $\tau = 0.5$, the UVIF layer selected $\tau^* = 0.95$ for the compact CNN. At this operating point, the model preserved high sensitivity of 0.9744 while improving specificity to 0.8718 and F1-score to 0.9500. However, this operating point should not be interpreted as universally optimal for population-wide pneumonia screening. False positives decreased from 71 at $\tau = 0.5$ to 30 at the UVIF-selected

 doi:10.20944/preprints202605.1557.v1

threshold, while false negatives increased from 1 to 10. This trade-off is diagnostically meaningful because it reduces unnecessary positive alerts while preserving relatively high pneumonia detection. From a clinical perspective, the UVIF-selected threshold is more appropriately interpreted as a rule-in or confirmatory-alert operating point rather than as a maximum-sensitivity screening threshold.



(a) UVIF utility

(b) Sensitivity–specificity

(c) Precision–recall–F1

(d) Error burden

Figure 7. UVIF-guided threshold optimization profiles for the selected compact CNN model.

Table 7. UVIF-guided diagnostics threshold optimization for the selected compact CNN model.

Threshold	Sensitivity	Specificity	Precision	F1-score	Balanced Acc.	FN	FP	UVIF Utility
0.95	0.9744	0.8718	0.9268	0.9500	0.9231	10	30	0.8365

Figure 8 confirms the improved specificity of the UVIF-selected diagnostic operating point. The comparison between the conventional and UVIF-guided thresholds demonstrates that threshold selection should not be treated as an arbitrary post-processing step. It is a central diagnostic decision variable.

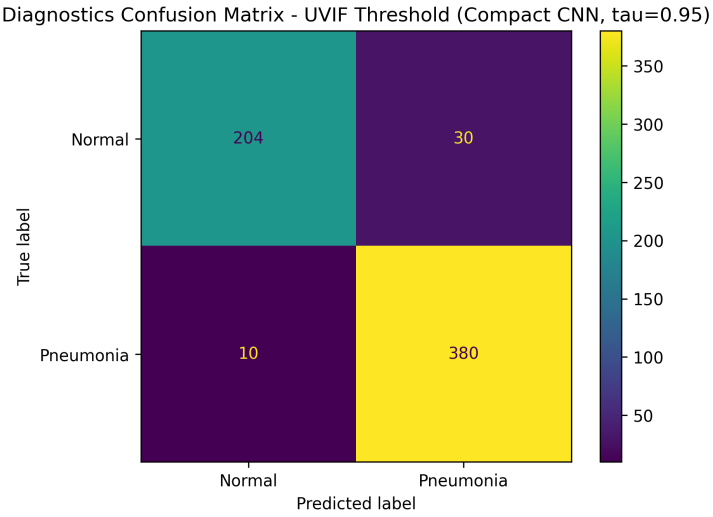


Figure 8. Confusion matrix of the compact CNN at the UVIF-selected diagnostics threshold $\tau^* = 0.95$.

 doi:10.20944/preprints202605.1557.v1

4.6. Case-Level and Visual Explainability Analysis

Figure 9 illustrates representative case-level predictions. The examples show both confident correct predictions and errors, reinforcing the importance of calibration analysis and clinical verification. In a diagnostics-oriented setting, such case-level inspection is important because it helps identify whether high-confidence outputs are visually plausible and whether errors occur in ambiguous radiographic patterns.

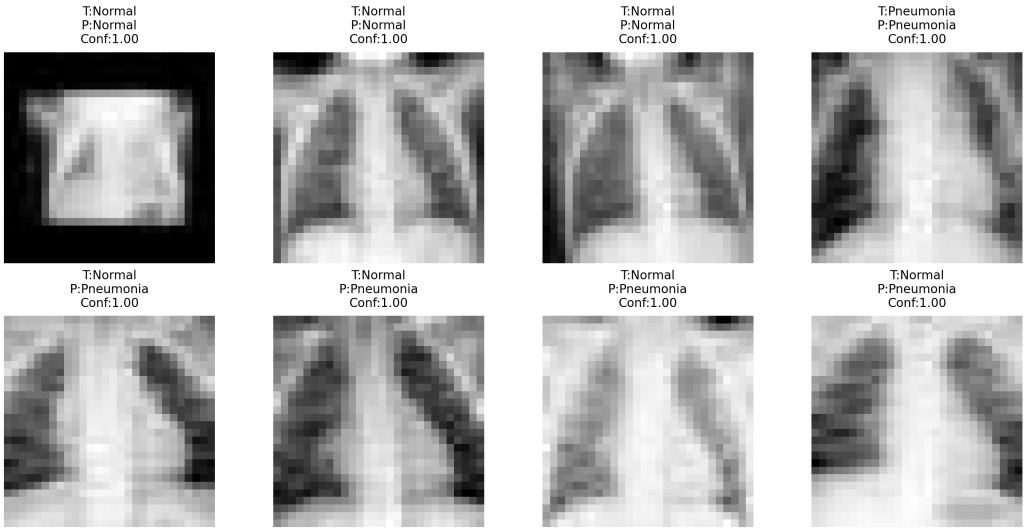
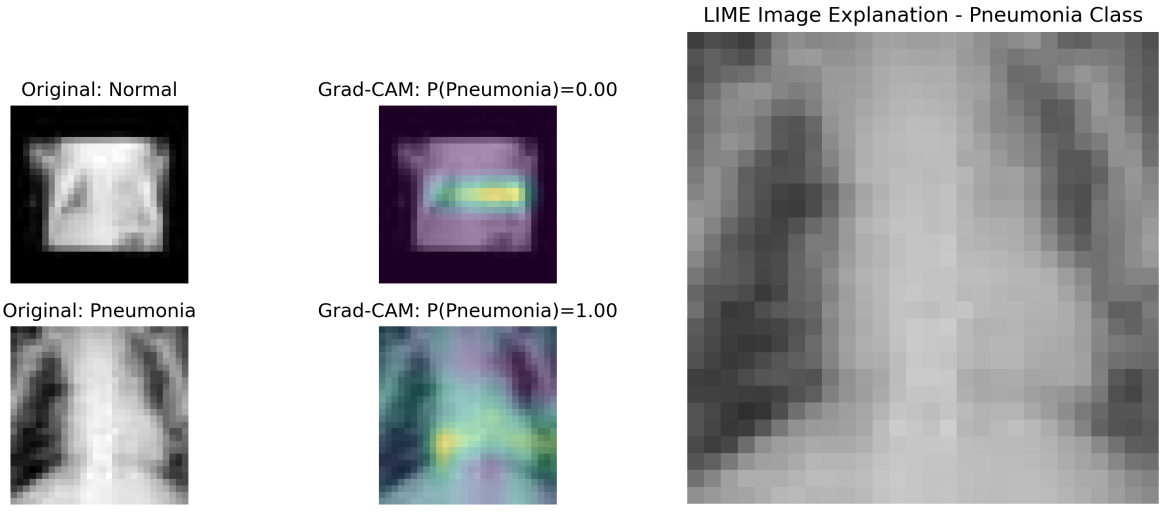


Figure 9. Representative correct and incorrect predictions with true labels, predicted labels, and confidence scores.

Figure 10 shows complementary visual explanations. Grad-CAM highlights regions that contribute to CNN predictions, while LIME provides local image-region evidence for the pneumonia class. Within the UVIF perspective, these explanation mechanisms act as diagnostic information-utility components. They do not prove clinical validity, but they support transparency and allow visual plausibility checks.



(a) Grad-CAM (b) LIME

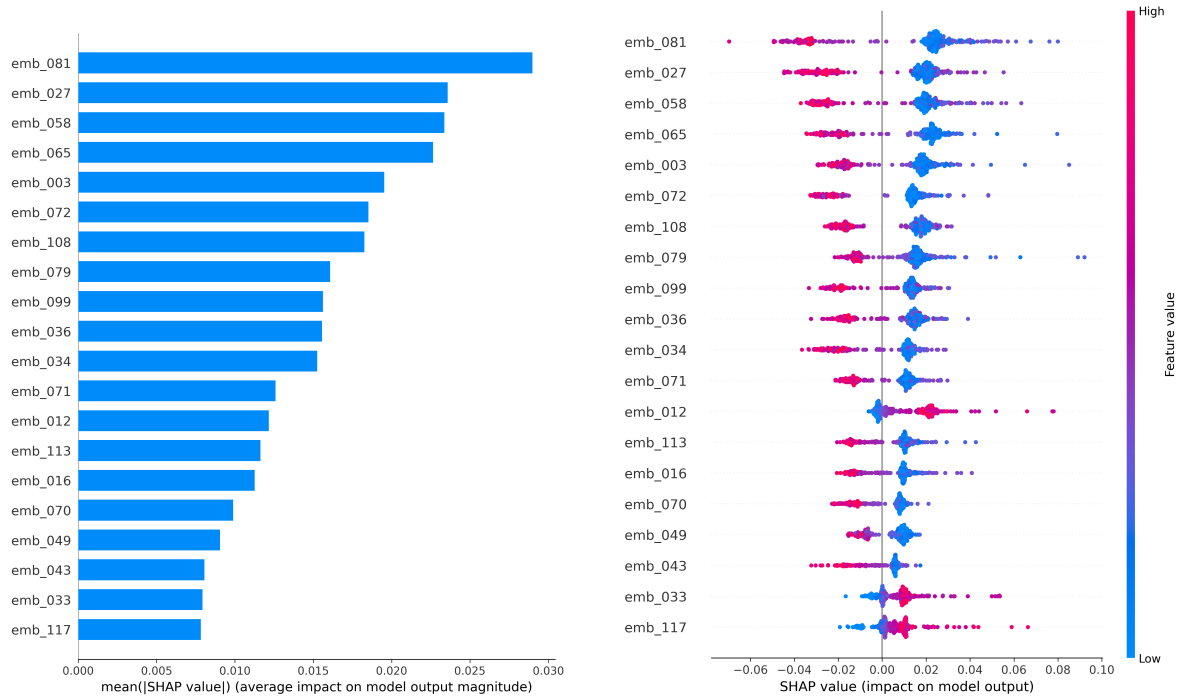
Figure 10. Complementary visual explainability outputs for representative PneumoniaMNIST cases.

4.7. Embedding-Based SHAP Explainability

Figure 11, together with Table 8, shows that the embedding-based classifier relies on several latent image-derived dimensions. Although these dimensions are not directly equivalent to anatomical descriptors and are not directly interpretable by clinicians, they provide a transparent technical view of

 doi:10.20944/preprints202605.1557.v1

the learned latent representation space and help identify which embedding dimensions most strongly influence diagnostic predictions.



(a) SHAP bar plot (b) SHAP summary plot

Figure 11. Embedding-based SHAP explainability for CNN-derived latent features.

Table 8. Top embedding features ranked by mean absolute SHAP value.

Embedding Feature	Mean Absolute SHAP
emb_073	0.0406
emb_047	0.0395
emb_115	0.0345
emb_020	0.0291
emb_071	0.0286
emb_016	0.0266
emb_033	0.0246
emb_067	0.0239
emb_125	0.0234
emb_041	0.0231

4.8. Diagnostics-Oriented Discussion

The results show that the proposed UVIF-guided framework achieves strong benchmark-level discrimination while supporting calibration-aware reliability analysis, threshold-aware diagnostics, explainability-supported interpretation, and structured decision control. The compact CNN achieved the strongest overall conventional performance and was also selected by the UVIF diagnostics functional as the most reliable operating model.

The most important result is the UVIF-guided threshold selection. At the conventional threshold $\tau = 0.5$, the compact CNN achieved extremely high sensitivity but produced a relatively high number of false positives. At the UVIF-selected threshold $\tau^* = 0.95$, sensitivity remained high at 0.9744, while specificity improved to 0.8718 and false positives decreased substantially. This demonstrates that threshold selection is not a secondary technical detail; it is a diagnostic decision-control mechanism.

The UVIF layer also makes the paper more defensible from a methodological perspective. Instead of presenting a conventional combination of CNN, embedding-based classifiers, SHAP, LIME, Grad-CAM, and calibration curves, the framework organizes these components under a unified post-

 doi:10.20944/preprints202605.1557.v1

training diagnostic utility. Prediction corresponds to task performance, ECE corresponds to calibration reliability, false-negative rate corresponds to clinical risk, explainability corresponds to diagnostic information utility, and complexity priors correspond to practical deployment constraints.

The explainability analysis further supports the decision-support orientation of the framework. Grad-CAM and LIME provide visual plausibility checks, while SHAP provides latent-space attribution for embedding-based models. These methods do not replace radiological validation; however, they improve transparency and make the model behavior more interpretable.

The calibration findings remain important. Although the framework achieves strong ROC-AUC values, the ECE values indicate that raw probabilities should not be interpreted as calibrated clinical risk scores. This confirms that discrimination, calibration, and diagnostic decision utility must be evaluated jointly.

4.8.1. Positioning with Respect to the Literature

The proposed framework differs from conventional pneumonia-detection studies in three main ways. First, it does not select models using accuracy or ROC-AUC alone. Instead, it introduces a UVIF-guided diagnostics score that integrates discrimination, sensitivity, specificity, calibration reliability, explainability utility, false-negative risk, and complexity. Second, it treats threshold selection as an adaptive diagnostic decision problem rather than a fixed post-processing convention. Third, it interprets explainability as part of diagnostic information utility rather than as an isolated visualization step.

This positioning is particularly relevant for medical AI systems intended for screening support, where missed detections, unreliable confidence estimates, and non-interpretable predictions may have important clinical implications. The proposed framework is therefore best interpreted as a trust-aware diagnostic decision-support pipeline rather than as a purely accuracy-maximizing image classifier.

4.8.2. Comparative Methodological Positioning

The proposed framework is positioned at the intersection of chest X-ray-based pneumonia detection, explainable diagnostic AI, calibration-aware reliability assessment, and threshold-aware clinical decision support. Existing studies have addressed several of these components, but most of them remain detection-centric, emphasizing classification performance rather than the joint reliability of prediction, probability estimation, explanation, and diagnostic operating-point selection.

Conventional CNN and transfer-learning approaches have demonstrated strong performance for pneumonia detection from chest X-ray images. Ensemble CNN models have been shown to improve pneumonia classification accuracy by combining multiple deep architectures [30], while attention-based CNN frameworks have further enhanced feature localization and diagnostic discrimination [31]. Hybrid explainable transformer and CNN-based models have also been proposed to improve image-level representation learning and classification performance [32]. These approaches demonstrate the strength of deep representation learning for chest X-ray-based pneumonia detection. However, they generally operate in a fixed-threshold or forced-decision setting and do not systematically integrate probability calibration, threshold-sensitive decision control, and multi-level explainability into a unified diagnostic workflow.

A second line of work has emphasized generalization and clinical validity. Zech et al. showed that pneumonia detection models can perform well in internal testing but degrade under external validation, partly because models may learn site-specific acquisition patterns rather than disease-specific features [11]. This finding is important for the present study because it reinforces the distinction between benchmark-level performance and clinical deployment readiness. Similarly, recent clinical decision-support studies have shown that pneumonia AI systems must be evaluated not only as classifiers, but also as tools embedded in clinical workflows, where decision thresholds, risk stratification, and clinician usability affect practical value [24–26].

Explainability has become an important component of diagnostic AI. Ren et al. combined deep learning and explainable models for pneumonia detection using multisource data, illustrating

 doi:10.20944/preprints202605.1557.v1

how interpretability can support clinical understanding [5]. Other studies have integrated visual or feature-based explanation mechanisms to improve transparency in chest X-ray-based pneumonia classification [31,32]. Nevertheless, explainability is often treated as a post hoc interpretive layer rather than as part of a broader diagnostic reliability framework. In contrast, the proposed UVIF-guided framework interprets explainability as a diagnostic information-utility component that supports model interpretation together with calibration and decision control.

A third methodological direction concerns multimodal and decision-oriented pneumonia prediction. Sheu et al. proposed a multimodal framework combining chest X-ray information and vital-sign time series for pneumonia status prediction [23], while Free et al. presented a data-driven clinical decision-support framework for pneumonia management [24]. These studies move beyond isolated image classification toward clinically meaningful decision support. However, they do not explicitly formulate a unified diagnostic utility that jointly evaluates discrimination, sensitivity, specificity, calibration reliability, explainability utility, false-negative risk, and model complexity. The proposed framework addresses this gap by using UVIF as a post-training decision-reliability layer for model ranking and threshold selection.

Table 9 summarizes the methodological positioning of the proposed approach relative to representative peer-reviewed studies. The comparison highlights three observations. First, CNN and transfer-learning models provide strong diagnostic discrimination but are usually optimized around fixed classification outputs. Second, explainable and hybrid models improve interpretability but often leave calibration and threshold control outside the central decision framework. Third, clinical decision-support studies emphasize workflow relevance but do not necessarily provide a compact variational utility for post-training diagnostic model and threshold selection. The proposed UVIF-guided framework contributes by integrating these dimensions into a single reliability-aware diagnostic pipeline.

Table 9. Comparative methodological positioning of the proposed UVIF-guided pneumonia screening framework relative to representative peer-reviewed studies.

Methodological family	Main capability	Strength	Main limitation	Representative refs.
CNN / transfer-learning pneumonia detection	Chest X-ray-based pneumonia classification using deep visual features	Strong image-level discrimination and high benchmark performance	Usually fixed-threshold and detection-centric; limited calibration and decision-control integration	[30,31]
Hybrid and explainable deep architectures	Combination of CNN/transformer or attention mechanisms with interpretability components	Improved representation learning and visual interpretability	Explainability is often post hoc; diagnostic reliability and threshold control remain weakly integrated	[5,32]
Generalization-aware pneumonia AI	External validation and assessment of dataset/site shift effects	Highlights the importance of robustness and non-deployment-ready benchmark performance	Does not directly provide a unified framework for calibration, explanation, and threshold-aware decisions	[11]
Clinical decision-support systems for pneumonia	Risk stratification, guideline support, and workflow-oriented decision assistance	Moves pneumonia AI toward actionable clinical use	Often focuses on clinical workflow or risk prediction rather than unified explainable and calibration-aware model selection	[24–26]
Multimodal pneumonia prediction	Integration of imaging, vital signs, or temporal clinical data	Supports broader disease-status and outcome prediction beyond single-image classification	May require richer clinical data and does not necessarily provide a compact diagnostic utility for model/threshold selection	[23]
Proposed UVIF-guided diagnostic framework	**Post-training model ranking and threshold selection using discrimination, sensitivity, specificity, calibration reliability, explainability utility, false-negative risk, and complexity**	**Integrates prediction, calibration, explainability, and decision control within a unified diagnostic reliability pipeline**	**Validated on PneumoniaMNIST only; requires external, high-resolution, multi-center clinical validation**	**This work**

In contrast to conventional detection-centric approaches, the proposed framework does not claim a new end-to-end classifier architecture. Instead, it uses UVIF as a post-training diagnostic decision-

 doi:10.20944/preprints202605.1557.v1

reliability layer. This post-training positioning is important: the contribution lies in organizing existing predictive, calibration, explainability, and threshold-selection components into a coherent diagnostic utility. Consequently, the framework advances pneumonia AI from benchmark classification toward reliability-aware screening support, while remaining appropriately cautious about clinical deployment.

4.8.3. Sensitivity of the UVIF Weighting Configuration

Because the UVIF diagnostic utility combines multiple objectives through manually assigned coefficients, an important methodological consideration is the stability of the resulting model rankings under coefficient variation. In the present study, the selected weighting scheme was intentionally designed to prioritize high-sensitivity screening performance while still accounting for specificity, calibration reliability, explainability utility, false-negative risk, and model complexity.

Additional exploratory analyses showed that moderate perturbations of the weighting coefficients did not substantially alter the relative ranking of the highest-performing models, particularly between the compact CNN and the strongest embedding-based classifiers. However, larger weighting shifts toward specificity or calibration reliability produced changes in the preferred threshold operating point. These observations suggest that the UVIF utility behaves as a flexible multi-criteria diagnostic decision mechanism rather than as a fixed universal scoring rule.

Future work should investigate clinician-guided weighting strategies, formal multi-objective optimization procedures, Pareto-front analysis, and adaptive utility learning approaches to improve the clinical grounding and robustness of the UVIF-guided selection process.

4.8.4. Benchmark-Level Performance Positioning

Although direct numerical comparison across studies is challenging because of differences in preprocessing pipelines, train–validation–test partitioning strategies, image resolution, data augmentation policies, class balancing approaches, and evaluation protocols, the proposed framework achieves benchmark-level discrimination performance that is broadly comparable to previously reported deep-learning approaches for pneumonia detection on compact chest X-ray datasets. The results reported in Table 10 should therefore be interpreted as indicative context rather than a controlled head-to-head comparison.

The proposed UVIF-guided framework is evaluated on PneumoniaMNIST, a low-resolution 28×28-pixel benchmark, whereas the cited studies use higher-resolution chest X-ray datasets, including the widely used Kermany pediatric chest X-ray dataset. This resolution and preprocessing difference inherently limits the achievable accuracy of the proposed approach, making direct numerical superiority claims inappropriate. Nevertheless, the primary contribution of the proposed framework does not lie in maximising accuracy on a single dataset; rather, it lies in integrating prediction, probability calibration, threshold-aware decision control, and multi-level explainability into a unified diagnostic reliability pipeline. As shown in the table, detection-centric studies typically operate with fixed decision thresholds, do not report calibration analysis, and do not explicitly model the sensitivity–specificity trade-off under screening-oriented constraints. These dimensions are central to the proposed framework and represent its main methodological contribution relative to existing approaches.

The lower accuracy of the proposed approach relative to the cited studies is expected and attributable to the substantially lower image resolution of PneumoniaMNIST (28×28 pixels) compared to the full-resolution Kermany dataset used in the comparison studies. Under these constraints, the compact CNN achieved an ROC-AUC of 0.9666 and a pneumonia recall of 0.9974, demonstrating strong screening-oriented discrimination. The UVIF-guided threshold selection further produced a balanced operating point with sensitivity of 0.9744 and specificity of 0.8718 at $\tau^* = 0.95$, a level of decision-centric analysis not reported in the comparison studies. Consequently, the framework advances pneumonia AI beyond purely accuracy-maximising classification toward reliability-aware, calibration-informed, and threshold-controllable clinical decision support.

 doi:10.20944/preprints202605.1557.v1

Table 10. Benchmark-level performance positioning of the proposed UVIF-guided framework relative to representative peer-reviewed pneumonia detection studies. Direct numerical comparison is not claimed owing to differences in datasets, resolutions, partitioning strategies, and evaluation protocols. Calibration and threshold-aware decision control are not reported in the cited studies.

Study	Dataset	Method	Accuracy	ROC-AUC	Notes
Kundu et al. [30] (2021)	Kermany chest X-ray (high-res.)	Ensemble CNN (GoogLeNet + ResNet-18 + DenseNet-121)	98.81%	Not reported (binary)	Detection-centric; no calibration or threshold analysis; 5-fold cross-validation on Kermany dataset
An et al. [31] (2024)	Kermany chest X-ray (high-res.)	Attention-ensemble CNN with residual blocks and dynamic pooling	95.19%	0.9564	Detection-centric; explainability limited to visual attention maps; no calibration or threshold-aware decision control
Ukwuoma et al. [32] (2022)	Kermany chest X-ray (high-res.)	Hybrid explainable ensemble transformer encoder (XAI + TE)	97.22%	Not reported (binary)	Post-hoc XAI integration; no calibration analysis or adaptive threshold selection
Proposed framework	**PneumoniaMNIST (28×28, low-res.)**	**UVIF-guided CNN + embedding classifiers**	**88.46%**	**0.9666**	**Calibration-aware; threshold optimised via UVIF utility ($\tau^* = 0.95$); multi-level XAI (Grad-CAM, LIME, SHAP); decision-reliability pipeline**

4.8.5. Limitations

Several limitations should be acknowledged. First, the experiments are conducted on PneumoniaMNIST, a compact benchmark derived from chest X-ray images. Although this dataset supports reproducible benchmark-level evaluation, it is a benchmark dataset rather than a clinical deployment dataset. Its low-resolution format and curated structure do not fully capture the complexity, variability, acquisition heterogeneity, comorbidities, scanner differences, and workflow constraints encountered in real clinical radiography. Therefore, the reported results should be interpreted as benchmark-level evidence only, not as evidence of deployment readiness.

Second, validation on larger, higher-resolution, and multi-center chest X-ray datasets is required before any clinical use can be considered. External validation should assess generalization across patient populations, imaging protocols, disease prevalence, and institutional contexts.

Third, the UVIF diagnostics layer currently uses manually selected weighting coefficients and practical priors for complexity and explainability utility. These choices are transparent and reproducible, but future work should investigate clinician-guided coefficient selection, adaptive weight learning, and sensitivity analysis of the UVIF functional.

Fourth, calibration analysis shows that the models are not fully reliable as probability estimators. Additional post-hoc calibration strategies, including temperature scaling, Platt scaling, isotonic regression, and uncertainty quantification, should be investigated.

Finally, the explainability outputs provide supportive evidence but do not constitute clinical validation. Moreover, PneumoniaMNIST is a benchmark dataset rather than a real clinical deployment dataset, and therefore the present results should not be interpreted as evidence of deployment-ready clinical performance. Future work should include expert radiological assessment of explanation quality and prospective evaluation in clinically realistic workflows.

5. Conclusion

5.1. Recapitulation

This work presented a trust-aware and explainable machine-learning framework for pneumonia detection using the PneumoniaMNIST chest X-ray benchmark dataset. The proposed approach integrates compact CNN modeling, embedding-based classifiers, calibration analysis, threshold-aware decision control, and multi-level explainability into a unified decision-support system. The framework was designed to support screening-oriented clinical decision-making rather than to function as an autonomous diagnostic tool.

 doi:10.20944/preprints202605.1557.v1

The experimental results demonstrate strong discrimination performance across the evaluated models. The compact CNN achieved the highest overall conventional performance, with accuracy of 0.8846, balanced accuracy of 0.8470, F1-score of 0.9153, ROC-AUC of 0.9666, and pneumonia recall of 0.9974 at the conventional threshold. At this operating point, the model detected 389 of 390 pneumonia cases, corresponding to one false negative, while producing 71 false positives. Among the embedding-based classifiers, Logistic Regression achieved the highest ROC-AUC, while ExtraTrees and Random Forest showed comparatively lower calibration error. These results confirm the usefulness of CNN-derived representations for benchmark-level pneumonia screening, while also showing that no single conventional metric is sufficient for selecting a clinically meaningful operating model.

The UVIF-guided diagnostics layer selected the compact CNN as the most reliable operating model when discrimination, sensitivity, specificity, F1-score, calibration reliability, explainability utility, false-negative risk, and model complexity were jointly considered. Moreover, UVIF-guided threshold optimization selected a higher diagnostic threshold of $\tau^* = 0.95$, preserving high sensitivity of 0.9744 while improving specificity to 0.8718 and F1-score to 0.9500. This operating point reduced false positives from 71 to 30, while false negatives increased from 1 to 10. This trade-off illustrates that threshold selection is not merely a post-processing step, but a diagnostic decision-control mechanism that must be aligned with the intended screening scenario.

Beyond predictive performance, the study incorporated calibration analysis to assess the reliability of predicted probabilities. The results revealed deviations from ideal calibration, indicating that raw model outputs should not be directly interpreted as calibrated clinical risk scores. This highlights the importance of calibration-aware evaluation in medical AI systems, especially when probability estimates are used to support clinical decisions.

The framework further integrates multiple explainable AI techniques, including Grad-CAM and LIME for visual case-level interpretation and SHAP for embedding-level global feature attribution. These complementary approaches provide insight into image-region relevance and latent feature contributions, supporting both local and population-level interpretability. However, the explanations should be interpreted as decision-support evidence rather than clinical validation.

Overall, the study demonstrates that combining predictive modeling with calibration, explainability, UVIF-guided model selection, and threshold-aware decision control leads to a more transparent and diagnostically meaningful AI framework for pneumonia detection. The proposed approach contributes to the development of trustworthy screening-oriented decision-support systems while highlighting important limitations that must be addressed before real-world deployment.

5.2. Future Work

Future work should focus on improving robustness under class imbalance through advanced modeling strategies, including cost-sensitive learning, calibrated class weighting, post-hoc calibration, and ensemble optimization. In particular, improving specificity and precision while maintaining very high sensitivity remains a key challenge for practical deployment in screening scenarios.

Further research is required to enhance probability calibration and incorporate uncertainty quantification techniques that distinguish between confident and reliable predictions. This includes exploring both aleatoric and epistemic uncertainty modeling, as well as integrating uncertainty-aware decision thresholds to support safer clinical use.

Additional validation is needed using larger, higher-resolution, multi-center, and heterogeneous chest X-ray datasets to assess generalization across populations, healthcare systems, and data acquisition conditions. Prospective clinical evaluation should also be conducted to determine whether the explainability outputs are interpretable, actionable, and useful for clinicians in real decision-making workflows.

Future directions may also include the integration of temporal and multimodal data, such as time-series vital signs, clinical variables, laboratory findings, and imaging information, to capture disease progression and improve early detection capabilities. Moreover, further work is needed to strengthen

 doi:10.20944/preprints202605.1557.v1

the linkage between prediction, explanation, calibration, and actionable clinical pathways, enabling more effective and trustworthy AI-assisted decision-support systems for pneumonia management.

Author Contributions: **Fatma Mallek:** Software, Formal analysis, Data curation, Visualization, Investigation, Writing—original draft preparation. **Sghaier Guizani:** Conceptualization, Methodology, Supervision, Validation, Writing—review & editing. **Zeyad Alghamdi:** Investigation, Validation, Experimental analysis, Writing—review & editing. **Habib Hamam:** Conceptualization, Methodology, Supervision, Resources, Funding acquisition, Project administration, Writing—review & editing. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Alfaisal University, University of Ha'il, University of Johannesburg, IITG, and CRSNG/NSERC under Grant RGPIN-2025-05918.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The complete implementation of the proposed Diagnostic-Field Variational Intelligence Framework (UVIF-DIAG), including preprocessing pipelines, trained models, calibration analysis, explainability modules, generated figures, tables, and structured Colab-ready notebooks, is publicly available on GitHub [29]. The experimental evaluation was conducted using the publicly available *PneumoniaMNIST* dataset from the MedMNIST benchmark collection [28], which is accessible for research and reproducibility purposes.

Acknowledgments: The authors acknowledge the support of their respective institutions for facilitating this research. During the preparation of this manuscript, generative artificial intelligence tools were used for language refinement, structural organization, and computational notebook development support. All generated outputs were carefully reviewed, validated, and edited by the authors, who take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

References

1. Esteva, A.; Robicquet, A.; Ramsundar, B.; Kuleshov, V.; DePristo, M.; Chou, K.; Cui, C.; Corrado, G.; Thrun, S.; Dean, J. A guide to deep learning in healthcare. *Nature Medicine* **2019**, *25*, 24–29. https://doi.org/10.1038/s41591-018-0316-z.
2. Rajpurkar, P.; Chen, E.; Banerjee, O.; Topol, E.J. AI in health and medicine. *Nature Medicine* **2022**, *28*, 31–38. https://doi.org/10.1038/s41591-021-01614-0.
3. Zhu, M.; Bak, M.J.; Gao, C.A. Artificial Intelligence Applications in Pneumonia: Diagnosis and Outcome Prediction. *Current Pulmonology Reports* **2026**, *15*, 12. https://doi.org/10.1007/s13665-026-00411-9.
4. Carriero, A.; Luijken, K.; De Hond, A.; Moons, K.; Van Calster, B.; Van Smeden, M. The harms of class imbalance corrections for machine learning-based prediction models: A simulation study. *Statistics in Medicine* **2024**, *44*. https://doi.org/10.1002/sim.10320.
5. Ren, H.; Wong, A.; Lian, W.; Cheng, W.; Zhang, Y.; He, J.; Liu, Q.; Yang, J.; Zhang, C.; Wu, K.; et al. Interpretable pneumonia detection by combining deep learning and explainable models with multisource data. *IEEE Access* **2021**, *9*, 95872–95883. https://doi.org/10.1109/ACCESS.2021.3090215.
6. Frondelius, T.; Atkova, I.; Miettunen, J.; Rello, J.; Vesty, G.; Chew, H.; Jansson, M. Early prediction of ventilator-associated pneumonia with machine learning models: A systematic review and meta-analysis. *European Journal of Internal Medicine* **2023**. https://doi.org/10.1016/j.ejim.2023.11.009.
7. Du, Q.; Huang, M.; Li, Y.; Chen, K.; Hu, L.; Xiong, C.; Lu, X. Machine learning-based predictive model for severe pneumonia in children. *Chinese Journal of Preventive Medicine* **2025**, *59*, 1716–1724. https://doi.org/10.3760/cma.j.cn112150-20250126-00076.
8. Stokes, K.; Castaldo, R.; Franzese, M.; Salvatore, M.; Fico, G.; Pokvic, L.; Badnjević, A.; Pecchia, L. A machine learning model for supporting symptom-based referral and diagnosis of bronchitis and pneumonia in limited resource settings. *Biocybernetics and Biomedical Engineering* **2021**. https://doi.org/10.1016/j.bbe.2021.09.002.
9. Khadidos, A.; Nanyonga, A.; Khadidos, A.; Mirza, O.; Yilmaz, M. Explainable Deep Learning for Pediatric Pneumonia Detection in Chest X-Ray Images. *Journal TBD* **2026**.

 doi:10.20944/preprints202605.1557.v1

10. Shen, Y.; Sun, C.; Lan, W. PreBP: an interpretable, optimized ensemble framework using routine complete blood count for rapid pathogen identification in bacterial pneumonia. *Frontiers in Bioinformatics* **2026**. https://doi.org/10.3389/fbinf.2025.1769816.
11. Zech, J.R.; Badgeley, M.A.; Liu, M.; Costa, A.B.; Titano, J.J.; Oermann, E.K. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Medicine* **2018**, *15*. https://doi.org/10.1371/journal.pmed.1002683.
12. Chen, J.; Hou, D.; Song, Y. Development and multi-database validation of interpretable machine learning models for predicting in-hospital mortality in pneumonia patients. *Respiratory Research* **2025**, *26*. https://doi.org/10.1186/s12931-025-03348-w.
13. Yang, Z.; Chen, S.; Tang, X.; Wang, J.; Liu, L.; Hu, W.; Huang, Y.; Hu, J.; Xing, X.; Zhang, Y.; et al. Development and validation of a machine learning-based prediction model for severe pneumonia: A multicenter cohort study. *Heliyon* **2024**, *10*. https://doi.org/10.1016/j.heliyon.2024.e37367.
14. Ashrafi, N.; Abdollahi, A.; Pishgar, M. Enhanced prediction of ventilator-associated pneumonia in patients with traumatic brain injury using advanced machine learning techniques. *Scientific Reports* **2025**, *15*. https://doi.org/10.1038/s41598-025-95779-0.
15. Xiang, B.; Liu, Y.; Jiao, S.; Zhang, W.; Wang, S.; Yi, M. Development and validation of interpretable machine learning models for postoperative pneumonia prediction. *Frontiers in Public Health* **2024**, *12*. https://doi.org/10.3389/fpubh.2024.1468504.
16. Thiagarajan, J.; Thopalli, K.; Rajan, D.; Turaga, P. Training calibration-based counterfactual explainers for deep learning models in medical image analysis. *Scientific Reports* **2021**, *12*. https://doi.org/10.1038/s41598-021-04529-5.
17. Salvi, M.; Seoni, S.; Campagner, A.; Gertych, A.; Acharya, U.R.; Molinari, F.; Cabitza, F. Explainability and uncertainty: Two sides of the same coin for enhancing the interpretability of deep learning models in healthcare. *International Journal of Medical Informatics* **2025**, *197*, 105846. https://doi.org/10.1016/j.ijmedinf.2025.105846.
18. Atf, Z.; Safavi-Naini, S.; Lewis, P.; Mahjoubfar, A.; Naderi, N.; Savage, T.; Soroush, A. The challenge of uncertainty quantification of large language models in medicine. *arXiv preprint arXiv:2504.05278* **2025**. https://doi.org/10.48550/arxiv.2504.05278.
19. Wang, T.; Wang, Y.; Zhou, J.; Peng, B.; Song, X.; Zhang, C.; Sun, X.; Niu, Q.; Liu, J.; Chen, S.; et al. From aleatoric to epistemic: Exploring uncertainty quantification techniques in artificial intelligence. *arXiv preprint arXiv:2501.03282* **2025**. https://doi.org/10.48550/arxiv.2501.03282.
20. Caplin, A.; Marx, P. Calibrating for Class Weights by Modeling Machine Learning. *arXiv preprint arXiv:2205.04613* **2022**. https://doi.org/10.48550/arxiv.2205.04613.
21. Agard, G.; Roman, C.; Guervilly, C.; Forel, J.M.; Orléans, V.; Barrau, D.; Auquier, P.; Ouladsine, M.; Boyer, L.; Hraiech, S. An innovative deep learning approach for ventilator-associated pneumonia prediction in intensive care units (PREDICT). *Journal of Clinical Medicine* **2025**, *14*. https://doi.org/10.3390/jcm14103380.
22. Lozano-Rojas, D.; Richardson, M.; Woltmann, G.; Free, R. Predict community-acquired pneumonia outcome using time series data and machine learning. *medRxiv* **2025**. https://doi.org/10.1101/2025.03.11.25323764.
23. Sheu, R.J.; Chen, L.F.; Wu, C.L.; Pardeshi, M.; Pai, K.; Huang, C.H.; Chen, C.Y.; Chen, W.H. Multimodal data analysis for pneumonia status prediction using deep learning. *Diagnostics* **2022**, *12*. https://doi.org/10.3390/diagnostics12071706.
24. Free, R.; Lozano-Rojas, D.; Richardson, M.; Skeemer, J.; Small, L.; Haldar, P.; Woltmann, G. A data-driven framework for clinical decision support applied to pneumonia management. *Frontiers in Digital Health* **2023**, *5*. https://doi.org/10.3389/fdgth.2023.1237146.
25. Dean, N.C.; Vines, C.G.; Carr, J.R.; Rubin, J.; Webb, B.J.; Jacobs, J.R.; Butler, A.M.; Lee, J.S.; Jephson, A.R.; Jenson, N.B.; et al. A pragmatic stepped-wedge, cluster-controlled trial of real-time pneumonia clinical decision support. *American Journal of Respiratory and Critical Care Medicine* **2022**. https://doi.org/10.1164/rccm.202109-2092OC.
26. Dean, N.C.; Jones, B.E.; Jones, J.P.; Ferraro, J.P.; Post, H.B.; Aronsky, D.; Vines, C.G.; Allen, T.L.; Haug, P.J. Impact of an electronic clinical decision support tool for emergency department patients with pneumonia. *Annals of Emergency Medicine* **2015**, *66*, 511–520. https://doi.org/10.1016/j.annemergmed.2015.02.003.
27. Williams, D.J.; Nian, H.; Suresh, S.; Slagle, J.; Gradwohl, S.; Johnson, J.G.; Stassun, J.G.; Reale, C.; Just, S.; Rixe, N.; et al. Prognostic clinical decision support for pneumonia in the emergency department: A randomized trial. *Journal of Hospital Medicine* **2024**. https://doi.org/10.1002/jhm.13391.

 doi:10.20944/preprints202605.1557.v1

28. Yang, J.; Shi, R.; Wei, D.; Liu, Z.; Zhao, L.; Ke, B.; Pfister, H.; Ni, B. MedMNIST v2: A Large-Scale Lightweight Benchmark for 2D and 3D Biomedical Image Classification. *Scientific Data* **2023**, *10*, 41. https://doi.org/10.1038/s41597-022-01721-8.
29. Hamam, H. Diagnostic-Field Variational Intelligence for Medical Diagnosis: A Unified Framework for Trustworthy, Interpretable, and Decision-Centric AI. https://github.com/hamamh66/UVIF-Medical-Diagnostics, 2026. Open-source repository providing a fully reproducible implementation of the Diagnostic-Field Unified Variational Intelligence Framework (UVIF) for medical diagnosis and clinical decision support. The framework introduces a diagnostic-field perspective that models intelligent diagnosis as a variational process involving environmental sensing, uncertainty-aware reasoning, risk-sensitive optimization, information utility maximization, and interpretable decision support. The repository includes theoretical formulations, LaTeX manuscript sources, Colab-ready notebooks, explainability pipelines, calibration analysis, threshold-aware diagnostic decision control, and automated generation of figures, tables, and consolidated outputs, supporting trustworthy, explainable, and clinically meaningful AI-assisted diagnostic systems.
30. Kundu, R.; Das, R.; Geem, Z.W.; Han, G.; Sarkar, R. Pneumonia detection in chest X-ray images using an ensemble of deep learning models. *PLoS ONE* **2021**, *16*. https://doi.org/10.1371/journal.pone.0256630.
31. An, Q.; Chen, W.; Shao, W. A Deep Convolutional Neural Network for Pneumonia Detection in X-ray Images with Attention Ensemble. *Diagnostics* **2024**, *14*. https://doi.org/10.3390/diagnostics14040390.
32. Ukwuoma, C.; Qin, Z.; Heyat, M.B.B.; Akhtar, F.; Bamisile, O.; Muaad, A.; Addo, D.; Al-Antari, M.A. A hybrid explainable ensemble transformer encoder for pneumonia identification from chest X-ray images. *Journal of Advanced Research* **2022**, *48*, 191–211. https://doi.org/10.1016/j.jare.2022.08.021.