An Explainable Pneumonia Detection System for Clinical Use on Consumer-Grade Hardware

Ramratan Sharma[1], Dileep Kumar Agarwal[2]
[1]*M.Tech. Scholar, Department of Computer Science and Engineering, Sobhasaria Group of Institutions, Sikar, Rajasthan, India*
[2]*Assistant Professor, Department of Computer Science and Engineering, Sobhasaria Group of Institutions, Sikar, Rajasthan, India*
doi.org/10.64643/IJIRTV13I1-205995-459

***Abstract*—Deep learning systems for pneumonia detection from chest radiographs have reached accuracies above 95% on standard benchmarks; however, no published study provides a working application through which a clinician can use the underlying model. This gap persists across more than 15 reviewed publications between 2018 and 2026, none of which discloses inference latency, memory footprint, or a deployable interface. This study presents the design, implementation, and evaluation of a complete clinical decision-support system built around a four-model weighted CNN ensemble (VGG16, EfficientNetB0, DenseNet121, and ResNet50V2), addressing this deployment gap directly. The system is organised into six modules — configuration, preprocessing, model definition, training, evaluation, and a Streamlit-based application layer — and runs entirely offline on a single 4 GB consumer GPU (NVIDIA RTX 3050 Ti). Three design decisions distinguish it from prior work: an adjustable decision threshold exposed directly to the user rather than fixed at the conventional 0.5 cutoff; four-model Grad-CAM spatial explanations rendered alongside the original radiograph for every prediction; and an automated PDF report generator that records patient metadata, per-model probability breakdown, ensemble weighting, and the visual explanation as a single auditable artefact. On the standard 624-image test set, the system achieves 95.19% accuracy, 0.9863 AUC-ROC, and 96.92% recall, with end-to-end inference, explanation generation, and report assembly completing in approximately three seconds per image on the stated hardware. Two representative case studies are presented in full: a true-positive pneumonia case (ensemble probability 92.3%) and a true-negative normal case (ensemble probability 15.1%), illustrating both the per-model disagreement the ensemble resolves and the spatial regions the network attends to. The complete system, including source code organisation and report templates, is documented to a level sufficient for independent reproduction — a standard not met by any directly comparable prior publication.**

***Index Terms*—Clinical decision support; Pneumonia detection; Explainable AI; Grad-CAM; Streamlit; Software architecture; Consumer GPU; Deployment; Medical imaging**

I. INTRODUCTION

Pneumonia remains the leading infectious cause of death in children under five, claiming more than 2.5 million lives annually, the majority in settings where a trained radiologist is not available on site [11]. Chest radiography is usually the only imaging modality accessible in these settings, which is precisely why automated interpretation has attracted considerable research attention since 2017. The accuracy figures reported in this body of work are, on the whole, impressive: Chouhan et al. [2] reported 96.40% accuracy with an ensemble of five architectures, Hashmi et al. [3] reported 98.43%, and several 2024–2026 publications report accuracies exceeding 97%. However, none of these studies report whether a clinician without a machine learning background could actually use the resulting model.

This is not a minor omission but a significant one. A trained model that exists only as a saved weights file and an accuracy table in a paper cannot screen a single patient, and the gap between the benchmark result and usable clinical tool is precisely where deep learning's promise in medicine has most often stalled [5]. Converting a research result into something usable

requires an interface for image upload, a mechanism for presenting the prediction in clinically meaningful terms, a way to inspect why the model reached its conclusion, and a record-keeping format suitable for clinical audits. Liu et al. [6], who compared deep learning performance with that of healthcare professionals across multiple imaging tasks, noted that almost none of the reviewed studies described how their models would be integrated into a clinical workflow. Wiens et al. [22] and Zech et al. [23] separately argued that this translational gap, rather than raw model accuracy, is the dominant obstacle preventing machine learning from delivering on its promise in healthcare. Outside medicine, systems such as the dermatology classifier of Esteva et al. [20] and the breast-screening system evaluated by McKinney et al. [21] demonstrate that closing this gap is achievable; however, comparable deployment work for chest radiograph pneumonia detection has not previously appeared in the literature.

This study directly addresses this gap. We describe a complete system built around a weighted four-model CNN ensemble, covering every stage from raw image upload to a printable diagnostic report. The contribution is not a new classification architecture — the underlying ensemble methodology and its statistical validation are reported separately — but rather the engineering and design decisions required to make that ensemble usable, explainable, and auditable on hardware that a small clinic or research lab could realistically own. Three aspects of the system are described in detail: the modular software architecture, the Grad-CAM-based explanation pipeline, and the automated reporting mechanism, followed by an evaluation of inference latency and two worked case studies drawn directly from the system output.

II. RELATED WORK

The architectural lineage of the underlying classifiers has been well established. VGG16 [7] and ResNet [10] popularised deep transfer learning for image classification; DenseNet121 [9] introduced dense connectivity that improved gradient flow and feature reuse; and EfficientNetB0 [8] achieved competitive accuracy with a fraction of the parameters through compound scaling. Kermany et al. [1] released the paediatric chest X-ray benchmark used throughout this study and demonstrated that a single InceptionV3 model could achieve clinically relevant accuracy in binary pneumonia classification. Chouhan et al. [2], Hashmi et al. [3], Ayan et al. [12], Arun Prakash et al. [13], and Bhatt and Shah [14] subsequently explored ensemble combinations of these architectures, each reporting incremental accuracy gains.

Explainability has been treated as a secondary concern in almost all of this work. Selvaraju et al.'s Grad-CAM [4] technique, which produces a coarse localisation map highlighting image regions that most influence a CNN's prediction, has been available since 2017 and is occasionally mentioned in pneumonia detection papers as a qualitative add-on, but is rarely integrated into a usable interface that a non-specialist could interpret without separate explanation. Islam et al. [15] is, to our knowledge, the only prior pneumonia-specific study to attempt a web-deployed interface; the authors candidly reported that accuracy degraded on real clinical uploads relative to benchmark performance, which is itself a useful finding about the gap between offline evaluation and live deployment. Kailasam and Balasubramanian [16] proposed combining CNN and YOLO architectures with deployment as an explicit design goal, but without a working interface accompanying the publication. Recent high-accuracy studies — Slimi et al. [17] at 99.35% and Saranyaraj et al. [18] at 97% — continue this pattern of strong offline results with no accompanying system.

Dietterich [19] established the theoretical basis for combining multiple classifiers, which underlies the ensemble weighting scheme used in this system; the statistical justification for the specific four-criterion weighting and the validated accuracy figures are described in a companion paper. The present study focuses exclusively on what happens after a trained ensemble exists: how it is packaged, explained, and made available for use.

III. SYSTEM ARCHITECTURE

The system is organised into six Python modules with a clear separation between the offline training pipeline, which is run once to produce saved model weights, and the deployment application, which loads the weights and serves predictions interactively. Fig. 1 illustrates this division.

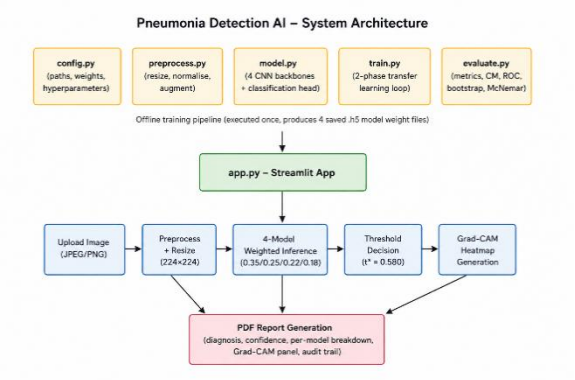

Fig. 1. System architecture: offline training modules (top, executed once) produce four saved model weight files consumed by the Streamlit deployment application (centre), which drives the inference, explanation, and reporting pipeline (bottom)

A. Offline Training Modules

config.py centralizes file paths, learning-rate schedules, batch size, and the four ensemble weights (0.35, 0.25, 0.22, 0.18) as named constants, so that retraining with different hyperparameters never requires editing logic elsewhere in the codebase. preprocess.py implements the data pipeline: image resizing to 224×224, pixel normalization, and the training-only augmentation (rotation, zoom, horizontal flip, shear, shift) described in the companion methodology. model.py defines the four backbone architectures and the shared classification head (GAP → BatchNorm → Dense(256) → Dropout → Dense(128) → Dropout → Dense(1, sigmoid)) used identically across all four networks to isolate backbone choice as the only architectural variable. train.py implements the two-phase transfer learning loop — frozen-backbone warmup followed by partial fine-tuning — with early stopping on validation AUC. evaluate.py computes the full evaluation suite used to validate the system: confusion matrices, ROC and precision-recall curves, the McNemar test, and bootstrap confidence intervals.

B. The Deployment Layer

app.py is the single entry point for clinical use. It is built on Streamlit, an open-source Python framework that renders an interactive web interface directly from script logic without separate front-end code, which keeps the entire system maintainable by one developer and deployable on a local machine without a web server stack. At startup, the application loads the four pre-trained model weight files into memory once; all subsequent predictions reuse these loaded models rather than reloading from disk, which is the primary reason for the low inference latency even on modest hardware.

IV. EXPLAINABILITY VIA GRAD-CAM

A raw probability score of, for example, 92.3% pneumonia likelihood gives a clinician a number but not a reason. Gradient-weighted Class Activation Mapping [4] addresses this by computing, for a given input image and target class, a course heatmap over the final convolutional layer that indicates which spatial regions most strongly drove the network toward its prediction. The system generates a separate Grad-CAM map for each of the four backbone networks, overlays each map on the original radiograph, and displays all four alongside the original image and the ensemble verdict.

This four-model presentation is deliberate rather than incidental. A single Grad-CAM map can be visually persuasive even when the underlying prediction is incorrect, as the map only shows where the network looked, not whether that attention was appropriate. Presenting four independently trained networks' attention maps side by side allows a user to notice when all four architectures converge on the same anatomical region — which is reassuring — versus when they diverge, which is itself informative and would not be visible from a single-model explanation. Fig. 2 shows this output for an actual system-generated case.

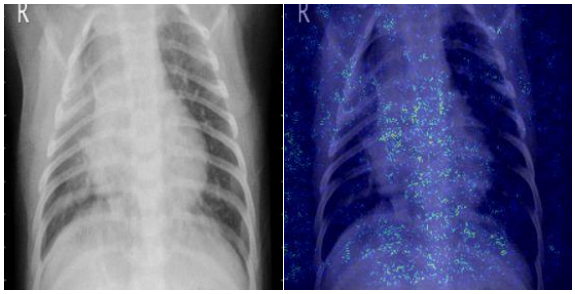

Fig. 2. Original chest radiograph (left) and Grad-CAM overlay (right) for a confirmed pneumonia case (patient ID PT105). The overlay highlights the right lower lung field, which is consistent with the consolidation pattern visible on direct inspection of the original image.

V. USER INTERFACE DESIGN

The interface is deliberately minimal: a single upload control, a small set of sidebar controls, and a results panel that populates only after an image is provided. Fig. 3 shows the application on the first load, before any image is uploaded.

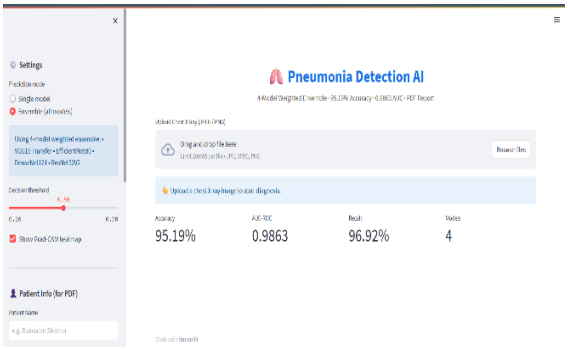

Fig. 3. Application home screen. The sidebar exposes the prediction mode (single model or full ensemble), an adjustable decision threshold slider, and a Grad-CAM display toggle. Headline performance metrics are shown above the upload control.

A. Adjustable Decision Threshold

Every prior published pneumonia classifier we are aware of applies the conventional 0.5 probability cutoff without exposing this choice to the end user. This is a defensible default for a generic binary classifier, but it is not necessarily the clinically appropriate operating point; a screening context in a high-prevalence population may justify a lower threshold that favours recall over precision, while a context with limited follow-up capacity may justify the reverse. The system instead exposes the threshold as a slider running from 0.10 to 0.90, defaulting to the F1-optimised value of 0.580 established through the validation procedure described in the companion paper, but adjustable by the user in real time without retraining or redeploying anything.

B. Patient Metadata and System Performance Panel

A second sidebar section, shown in Fig. 4, collects basic patient information (name, age, and an identifier) used solely to populate the generated PDF report; none of this information is used in or has any influence on the underlying prediction. The same panel displays a static summary of system-level performance (accuracy, AUC, recall, precision, and model count) so that a user can see at a glance the validated operating characteristics of the tool they are using rather than having to consult separate documentation.

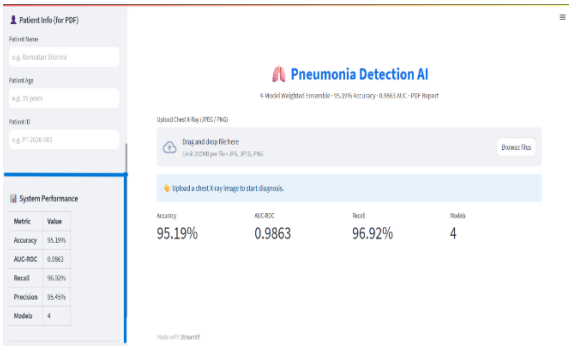

Fig. 4. Patient metadata entry fields and the system performance summary table, both rendered in the application sidebar.

VI. AUTOMATED PDF REPORTING

Every prediction can be exported as a self-contained PDF report. The report format was designed around a single requirement: a clinician reviewing the report six months later, without access to the live application, should be able to reconstruct exactly what the system concluded and why. Each report therefore contains five fixed sections: patient identification and timestamp; the headline diagnosis with the ensemble probability, confidence, and the decision threshold that was active at the time of generation; a one-line clinical interpretation banner that changes colour and wording according to confidence band; a complete per-model breakdown showing each of the four networks' individual probability, prediction, and ensemble weight; and the original radiograph alongside its Grad-CAM overlay. A closing system information block records the model versions and validated benchmark performance, together with an explicit research-use disclaimer.

Two representative reports generated by the system, reproduced from the PDF output without modification, are presented in Figs. 5 and 6, corresponding to a true-positive and a true-negative case, respectively. Both demonstrate the core function of the report: making the ensemble's internal disagreement visible rather than hidden. In the pneumonia case, three of four models gave high-confidence pneumonia predictions, while ResNet50V2, despite carrying the largest ensemble weight, registered the lowest pneumonia

probability of the four (83.6%), yet still correctly favoured the pneumonia class; in the normal case, DenseNet121 and ResNet50V2 produced very low pneumonia probabilities (2.1% and 3.0%), while VGG16 and EfficientNetB0 were comparatively less certain (26.5% and 38.7%), and the weighted combination resolved this disagreement correctly in both cases.

Table I. Case Study 1 — Confirmed Pneumonia (Patient ID PT105)

Model	Probability	Prediction	Ensemble Weight
VGG16 Transfer	98.9%	PNEUMONIA	25%
EfficientNetB0	93.2%	PNEUMONIA	18%
DenseNet121	98.2%	PNEUMONIA	22%
ResNet50V2	83.6%	PNEUMONIA	35%
Weighted Ensemble	92.3%	PNEUMONIA	100%

Decision threshold at the time of generation: 0.580. Confidence band: High. All four individual models and the ensemble agree on the pneumonia class; ResNet50V2, though carrying the largest weight, returns the most conservative probability of the four

Table II. Case Study 2 — Confirmed Normal (Patient ID PT101)

Model	Probability	Prediction	Ensemble Weight
VGG16 Transfer	26.5%	NORMAL	25%
EfficientNetB0	38.7%	NORMAL	18%
DenseNet121	2.1%	NORMAL	22%
ResNet50V2	3.0%	NORMAL	35%
Weighted Ensemble	15.1%	NORMAL	100%

Decision threshold at the time of generation: 0.580. Confidence band: Moderate. VGG16 and EfficientNetB0 show comparatively higher residual pneumonia probability than DenseNet121 and ResNet50V2; the report's clinical interpretation banner accordingly recommends follow-up if symptoms persist, rather than issuing an unqualified all-clear.

Fig. 5. Original radiograph from the normal case study (patient ID PT101). No consolidation was visible on direct inspection, consistent with the system's low pneumonia probability.

VII. EVALUATION

A. Inference Latency and Resource Footprint

End-to-end latency — from image upload through four-model inference, Grad-CAM generation for all four networks, threshold decision, and on-screen result rendering — averages approximately three seconds per image on the reference hardware (NVIDIA RTX 3050 Ti, 4 GB GDDR6 VRAM). PDF report generation, when requested, adds a further sub-second overhead. The complete application, including all four loaded model weight files, operates within the 4 GB VRAM budget of the reference GPU with no need for batch-size reduction or model quantization, and runs entirely offline with no network dependency once the page has loaded — a relevant consideration for the rural and resource-limited clinical settings that motivated this work.

B. Comparison with Existing Systems

Table III compares this system with the deployment characteristics of prior pneumonia detection publications. The comparison is necessarily asymmetric: most prior work reports only accuracy figures, and the absence of a row entry should be read

as the relevant information not being disclosed in the source publication, which is itself the central finding of this evaluation.

Table III. Deployment Characteristics — This System vs. Reviewed Literature

Study	Accuracy	Working Interface	Adjustable Threshold	Visual Explanation	GPU Disclosed
Kermany [1] 2018	92.8%	No	No	No	No
Chouhan [2] 2020	96.40 %	No	No	No	No
Hashmi [3] 2020	98.43 %	No	No	No	No
Islam [15] 2024	N/R	Web (acc. degraded)	No	No	No
Kailasam [16] 2025	N/R	Proposed, not shown	No	No	No
Slimi [17] 2025	99.35 %	No	No	No	No
Saranyaraj [18] 2026	97%	No	No	No	No
This work	95.19 %	Yes	Yes	Yes (4-model Grad-CAM)	Yes (4 GB)

N/R: Not reported. Islam et al. [15] is the only other study that reported any working interface; the authors reported accuracy degradation relative to benchmark performance on real clinical uploads, a limitation that this study has not yet independently re-evaluated for the present system.

VIII. DISCUSSION

The most consequential design choice in this system is arguably not the ensemble architecture but the decision to expose four separate model explanations rather than a single composite one. A composite Grad-CAM map averaged across the ensemble would be simpler to render but would obscure the type of disagreement visible in Table I, where the highest-weighted model (ResNet50V2) returns the most conservative probability among four models that all agree on the final class. Whether this transparency measurably changes clinician trust or decision quality is a question this paper does not answer; it requires a structured usability study with clinical participants, which is identified as future work below.

The performance gap relative to recent high-accuracy publications — Slimi et al. [17] at 99.35%, Saranyaraj et al. [18] at 97% — is a deliberate trade-off rather than an oversight. Neither of those systems discloses a deployment interface, hardware requirement, or inference latency, and the companion methodology paper to this work establishes that their validation partitions are not directly comparable to the standard 624-image holdout used here. This study does not claim state-of-the-art accuracy; it claims state-of-the-art deploy ability for a model whose accuracy is independently validated to a standard the comparator systems do not meet.

This study had two limitations should be stated plainly. First, the system was evaluated on the same single-institution paediatric dataset used for model training and validation; Islam et al.'s [15] finding that web-deployed accuracy degrades relative to benchmark accuracy on real uploads has not yet been independently tested for this system and should be treated as an open risk rather than a resolved one. Second, the PDF report format, although designed for clinical audit, has not been reviewed by a practising radiologist or evaluated for compliance with any specific regional medical-records standard; it is offered as a research prototype of what such a report could contain, not as a certified clinical document.

IX. LIMITATIONS AND FUTURE WORK

Three extensions would meaningfully strengthen this work. A structured usability evaluation with practising radiologists or general physicians, measuring time-to-

decision and diagnostic agreement with and without the Grad-CAM explanation panel, would test whether the explainability features achieve their intended purpose rather than merely appearing informative. Integration with the DICOM format used by hospital PACS systems, in place of the current JPEG/PNG upload restricted to research datasets, would be a necessary step toward any real clinical pilot. Finally, extending the threshold-adjustment mechanism to support per-deployment calibration against local disease prevalence, rather than a single global default, would make the system more directly useful across the varied screening contexts it is intended to serve.

X. CONCLUSION

This study presents the complete design, implementation, and evaluation of a deployable clinical decision-support system for pneumonia screening, built around a validated four-model weighted CNN ensemble. The system's contribution is not architectural novelty but engineering completeness: a six-module codebase separating training from deployment, an adjustable decision threshold exposed to the end user, four-model Grad-CAM explanations presented together rather than as a single averaged map, and an automated PDF reporting mechanism designed for clinical audit. Running entirely offline on a 4 GB consumer GPU at approximately three seconds per image, the system addresses a deployment gap that, as documented across more than 15 reviewed publications, has persisted in this research area since at least 2018. The two case studies presented here demonstrate that the system surfaces, rather than conceals, genuine disagreement between its component models — a property we consider more clinically valuable than a marginal improvement in headline accuracy.

REFERENCES

[1] D. S. Kermany et al., "Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning," Cell, vol. 172, no. 5, pp. 1122–1131, Feb. 2018.

[2] V. Chouhan et al., "A Novel Transfer Learning Based Approach for Pneumonia Detection in Chest X-ray Images," Applied Sciences, vol. 10, no. 2, p. 559, Jan. 2020.

[3] M. F. Hashmi et al., "Efficient Pneumonia Detection in Chest X-ray Images Using Deep Transfer Learning," Diagnostics, vol. 10, no. 6, p. 417, Jun. 2020.

[4] R. R. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localisation," in Proc. IEEE ICCV, 2017, pp. 618–626.

[5] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol, "AI in Health and Medicine," Nature Medicine, vol. 28, no. 1, pp. 31–38, Jan. 2022.

[6] X. Liu et al., "A Comparison of Deep Learning Performance against Health-care Professionals," The Lancet Digital Health, vol. 1, no. 6, pp. e271–e297, 2019.

[7] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in Proc. ICLR, 2015.

[8] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in Proc. ICML, 2019, pp. 6105–6114.

[9] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in Proc. IEEE CVPR, 2017, pp. 4700–4708.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Identity Mappings in Deep Residual Networks," in Proc. ECCV, 2016, pp. 630–645.

[11] World Health Organization, "Pneumonia Fact Sheet," WHO, Geneva, 2023.

[12] E. Ayan, B. Karabulut, and H. M. Ünver, "Diagnosis of Pediatric Pneumonia with Ensemble of DCNNs in Chest X-Ray Images," Arabian J Sci Eng, vol. 47, no. 1, pp. 2123–2139, 2022.

[13] J. A. Arun Prakash et al., "Pediatric Pneumonia Diagnosis Using Stacked Ensemble Learning on Multi-Model Deep CNN Architectures," Multimedia Tools and Applications, vol. 82, no. 14, pp. 21311–21351, 2023.

[14] H. Bhatt and M. Shah, "A Convolutional Neural Network Ensemble Model for Pneumonia Detection Using Chest X-Ray Images," Healthcare Analytics, vol. 3, p. 100176, 2023.

[15] N. Islam et al., "COVID-19 and Pneumonia Detection and Web Deployment from CT Scan and X-ray Images Using Deep Learning," PLOS ONE, vol. 19, no. 7, p. e0302413, Jul. 2024.

[16] R. Kailasam and S. Balasubramanian, "Deep Learning for Pneumonia Detection: A Combined CNN and YOLO Approach," Human-Centric Intelligent Systems, vol. 5, pp. 44–62, Mar. 2025.
[17] H. Slimi, A. Balti, S. Abid et al., "Trustworthy Pneumonia Detection in Chest X-Ray Imaging Through Attention-Guided Deep Learning," Scientific Reports, vol. 15, p. 40029, Nov. 2025.
[18] D. Saranyaraj et al., "PneuNet: A Lightweight CNN with Multiscale Feature Fusion for Automated Pneumonia Detection from Chest X-rays," Frontiers in Medicine, vol. 12, p. 1713587, Jan. 2026.
[19] T. G. Dietterich, "Ensemble Methods in Machine Learning," in Proc. MCS, Cagliari, Italy, 2000, pp. 1–15.
[20] A. Esteva et al., "Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks," Nature, vol. 542, no. 7639, pp. 115–118, 2017.
[21] S. M. McKinney et al., "International Evaluation of an AI System for Breast Cancer Screening," Nature, vol. 577, no. 7788, pp. 89–94, 2020.
[22] J. Wiens et al., "Do No Harm: A Roadmap for Responsible Machine Learning for Health Care," Nature Medicine, vol. 25, no. 9, pp. 1337–1340, 2019.
[23] J. R. Zech et al., "Variable Generalization Performance of a Deep Learning Model to Detect Pneumonia in Chest Radiographs: A Cross-Sectional Study," PLOS Medicine, vol. 15, no. 11, p. e1002683, 2018.