Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

# A Full-Stack, Explainable Chest X-ray Analysis Workspace Using DenseNet121: A Survey and Reference Design, with a Prototype Interface Case Study

**Hirva Gohel**
Department of Information Technology, Gyanmanjari Innovative University
Bhavnagar, Gujarat, India
ORCID: 0009-0005-6704-8441

***Abstract*—Chest X-ray is the most common radiology exam, yet is limited by the number and variability of radiologists who interpret each image. Deep convolutional neural networks, DenseNet121 in particular, have created a common baseline for Chest X-ray (CXR) interpretation, but accuracy alone cannot define a practically-useful system. We survey the literature for the elements required for a full-stack, explainable CXR interpretation workspace: the state of public data and label quality, DenseNet121 and alternatives, post-hoc explainability (mostly Grad-CAM), in-production architectures, role-based clinical review work-flow, and human-in-the-loop supervision. We describe how we combine these elements into a reference architecture with explainable inference, role-based access to information, logging, and audit trails. We review and compare the capabilities and failure modes of each of the components, review explanation methods per cost and reliability, and detail the role-based access control interface in a prototype workplace MediVision AI. Based on interface captures, the prototype, comprising a multi-label DenseNet121 classifier, Grad-CAM explanations, private workspaces for each user, and one-time expiring share links, is used as a case study to consider trade-offs that emerge in this setting. This demonstration is not a performance study and MediVision AI is not yet field-deployable. We conclude by highlighting open challenges (label noise, distribution shift, subgroup bias, calibration, and regulation) and directions for future research.**

**Index TermsChest X-ray, medical informatics, clinical decision aid, deep learning, DenseNet121, "open" AI, Grad-CAM, role-based access control.**

## I. INTRODUCTION

CXR can be performed at any point of time in the imaging department, as it is fast, low dose and low cost, which makes it one of the most performed diagnostic imaging studies performed [1]. It is able to identify various disease processes such as pneumonia, pleural effusion, pneumothorax, and cardiomegaly as well as many other thoracic pathologies. Reading chest x-ray images are challenging: the abnormalities can be very subtle, and anatomical structures can obscure pathology, often in facilities with high work rates well above the radiologist workforce capacity. This need has enabled research into artificial intelligence methods to triage studies and as a second reader and has contributed to a new humanAI collaborative way of working.

popularized thanks to these enablements. Tapping this upward trend from medical image deep learning [2], and following the publication of large annotated datasets [3], a 121-layer DenseNet trained on ChestX-ray14 (CheXNet) [4] outperformed radiologists on pneumonia detection, and later studied multi-label CNN vs. radiologists on 14 pathologies [5]. DenseNet121 [6] so became a common backbone in the CXR realm, and is by default a popular baseline. The prior CXR surveys [7] sort research papers as task/method, this work takes a workspace level perspective that and covers explanation, deployment, and clinician review.

But, a classifier AUROC> (good to great as a tool) may not be practical in an operational setting, even if it is ward-ready. It takes (i) a narrative rationale that justifies an image was scored how, (ii) the imaging review process sitting the responsibility of the final score with a human clinician, (iii) limiting each user's role-based capabilities and perspective, (iv) records of the reviewed images, deployed model version, and the score. An aspect of clinical AI review that relates to development, evaluation, and clinician's trust in it is that clinical impact is not guaranteed by the model's accuracy [8]. Critics have raised arguments against post-hoc explanations in clinical settings [9], and advocates make a case for inherently interpretable models in high-stakes circumstances [10]. These are a few reasons why workspace design and its evaluation matter.

This paper: contains a literature review we use to select a full-stack, explainable CXR analysis workspace, we built from DenseNet121 with role-based clinical review. Our contributions: (i) contains a review of available CXR datasets and their label quality issues; (ii) review of the state of the art in CXR classification architectures and a comparison of these; (iii) explanation method we used to develop an explanation evaluation scheme: including a review of explanation and explanation evaluation methods and another of explanation architectures on the literature; (iv) review of architectures and role-based review workflows we drew on from deployment and access-control literature; (v) case study of a CXR analysis prototype, MediVision AI, in its clinical task workflow interface explanations and operational response; (vi) discusses remaining issues and evaluation structure. This work is that means a survey of literature and work synthesis through design and a prototype case study; we do not produce new benchmark numbers for quantitative evaluation.

*Survey methodology:* This is a narrative survey and not a pre-registered systematic review. For the work published predominantly between 2017 and 2026 in November 2026, we searched publisher pages, PubMed/PubMed central records, arXiv, and academic search engines indexes (e.g. IEEE Xplore). PubMed/PubMed central records were searched using



(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

different combinations of several search phrases, including("chest X-ray", "CXR", "DenseNet121", "foundation model", "visionlanguage", "report generation", "Grad-CAM", "clinical deployment", "triage", "calibration", "role-based access control"). The search yielded 140 literature, from which we chose primary studies, benchmark articles, dataset articles, way articles, clinical trials, and topical articles, that dissected a component of a workflow, explanation, or governance on CXR, beyond method articles, standards, and standards articles. We discarded false-positive variants for which a peer reviewed or publisher was available, regardless of its credibility, unrelated to chest radiography, infor-mative, or viewpoint articles, commentary, or viewpoint articles without a CXR application and duplicate variants of a single non-preprint publisher sources. We addressed each workspace element with a citation so that the coverage is suggestive and not exhaustive and the coverage was not intended to allow meta-analyses.

II. Table of contents I. Executive summary II. Introduction and data sets III. Architectures IV. Explainability V. Full Stack architecture VI. Role-based review VII. MediVision AI case study VIII. Evaluation IX. Open challenges X. Conclusion and future directions.

## II. BACKGROUND AND DATASETS

### *A. Task Formulation*

Most tasks in chest X-ray are multi-label, because in a single image, an effusion, for example, is often connected to other findings (effusion and atelectasis). This results in using sigmoid output on each class, since they are non-mutually exclusive. The most common inputs are PA/AP (postero-anterior / antero-posterior), with occasional lateral views. While certain workflows include localization or segmentation, it's most common for multi-label image-level classification to be paired with a heatmap, first, and foremost for procedural guides. Pneumonia is not the only disease evaluated by CNN classifiers, but also tuberculosis [11], pneumothorax [12], malignant pulmonary nodules [13], and all major thoracic pathologies simultaneously [14].

### *B. Public Datasets*

Most of this work has focused on a small number of public datasets as in Table I. ChestX-ray8 [3], eventually expanded public as ChestX-ray14, a large label set extracted from radiology reports using NLP (scalable but noisy). CheXpert [15], which added explicit labels of uncertainty and a radiologist graded evaluation set. MIMIC-CXR [16], which provided links to radiology reports and included a wider range of clinical data. PadChest [17], a large Spanish language set with a hierarchical labelling scheme. A few other smaller or specialised sets have emerged, like for tuberculosis [18], the ZSHU-X-ray, with detailed annotations for pneumonia detection [19], the RSNA pneumonia set, with additional expert labels to ChestX-ray14 [20], and the VinDr-CXR set of bounding boxes as annotated by radiologists [21].

TABLE I
REPRESENTATIVE PUBLIC CHEST X-RAY DATASETS

| Dataset | Volume | Labels | Source and notes |
|---|---|---|---|
| ChestX-ray14 [3]† | 112,120 images | 14 labels | National Institutes of Health (NIH); report-mined labels |
| CheXpert [15] | 224,316 images | 14 labels | Stanford; uncertainty labels; expert-labeled test set |
| MIMIC-CXR [16] | 377,110 images | 14 labels | Beth Israel Deaconess Medical Center (BIDMC); free-text reports |
| PadChest [17] | >160,000 images | 174 findings | Hospital San Juan; Spanish reports; label hierarchy |
| VinDr-CXR [21] | 18,000 images | 6 global, 22 local | Vietnamese hospitals; radiologist bounding boxes |
| Montgomery, Shenzhen [18] | 138; 662 images | TB vs. normal | Small tuberculosis-screening sets |

*The original ChestX-ray8 publication [3] reported 108,948 images and 8 labels; the figures shown are for the subsequent expanded NIH Clinical Center release.

Once more, label quality is a concern. The quality of labels and dataset composition was scrutinised with several large public medical image datasets, and errors have been identified, even for ChestX-ray14 [22], and report-derived labels are simply a record of what the radiologists decided to note, rather than pixelwise ground truth. Results obtained from one dataset may not generalise to others.

### *C. Preprocessing and Training Practice*

Most pipelines rescales the radiographs to 224224 pixels, copies the single channel into three, and normalizes ImageNet statistics so ImageNet-pretrained weights [23] can be transferred. Transfer learning for general vision [24] is quite mature, but head-to-head comparison of retinal and CheXpert data showed ImageNet pretraining only a few points higher accuracy with most of the benefit coming from faster improvement and weight rescaling [25]. Contrast enhancement (e.g. contrast-limited adaptive histogram equalization (CLAHE) [26]) is sometimes applied, and moderate augmentation (small rotations, scale, and brightness/contrast variations) occurs; flip horizontal images along the vertical axis with caution as the perceived laterality of various objects (e.g. aortic knuckle) will change. The model handles class imbalance with weighted binary cross entropy or focal loss [27]. Make disjoint sets of patients between training and testing: if more than one study per patient is in both, there will be information leaks and the model will appear robust with an image-level random split but will perform poorly on an information-blind split.

## III. CNN ARCHITECTURES FOR CXR CLASSIFICATION

### *A. DenseNet121*

DenseNet [6] is a backbone architecture known for the concatenation of feature maps. Within a dense block, each layer receives feature maps from all preceding layers and passes its own feature map to every subsequent layer. This structure helps feature reuse, offers stable gradients, and results in a relatively compact model. DenseNet121 includes 4 dense blocks with 6 12 24, and 16 layers respectively, and has a growth rate of 32. These blocks are separated by transition layers that reduce spatial dimensions and the number of channels, with the model containing about 8 million parameters in total. CheXNet used this backbone [4], which is



(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

often selected for four practical reasons: parameter efficiency, easy, and stable fine-tuning from ImageNet weights, the final CNN feature map enabling CAM-style explanations, and a sufficiently small size suitable for inference on limited GPU or CPU resources.

### *B. Alternative Backbones*

The next most popular alternatives are residual networks [28], which have been considered for a long time as simple flexible yet optimised architectures. EfficientNet [29] is a scaleable network modelling the joint scale of depth, width, and resolution of the image for optimal performance with a given number of parameters. Vision Transformer (ViT) [30] models the complete context of the image with self-attention, they can be very powerful when scaled up with pretrained weights, but are more data hungry and computationally expensive to operate; hierarchical variants such as Swin [31] have addressed this while preserving the benefits of self-attention by reusing it over 'shifted' windows. More recently, self-supervised, and vision-language approaches have shown promise for detecting pathology in the absence of image-level labels: CheXzero [32] trains a CLIP [33] adaptation on radiographs and reports, and large scale self-supervised approaches have reached new and impressive heights in medical imaging [34]. Instruction-tuned visionlanguage foundation models like CheXagent [35] for CXR interpretation tasks that add report generation. These foundation models decrease reliance on annotations [36]. More recent work on CXR foundation-models have advanced beyond task-specific CNN baselines that pre-trained with self-supervision on 987,000 CXRs before transferred to in- and out-of-distribution tasks, [37], this paper categorizes CXR foundation models by pretraining model, architecture, and downstream use-case. This work on CXR foundation models strengthens the foundation for comparisons to DenseNet121, but will not eliminate the necessity for realworld validation.

TABLE II
BACKBONE COMPARISON FOR CXR WORKSPACES

| Architecture | Parameters | Key strengths | Practical limitations |
|---|---|---|---|
| DenseNet121 [6] | ≈8.0 M | Feature reuse; compact; CAM-friendly final block | Concatenations raise training memory use |
| ResNet-50 [28] | ≈25.6 M | Simple, widely supported strong baseline | More parameters than DenseNet121 |
| EfficientNet-B0 [29] | ≈5.3 M | High accuracy per parameter | Sensitive to training recipe; larger variants cost more |
| ViT-B/16 [30] | ≈86 M | Global context via self-attention | Data-hungry; heavier to serve |

Comparison across papers doesn't seem like a useful benchmark since the dataset version, the label set split resolution, and test protocol are all different. We because of this do not sort the backbones by how they rank in AUROC. We find DenseNet121 to be a reasonable baseline for deployable backends, since it is well-known and somewhat efficient, but newer backbones are generally better.

### *C. Fine-Tuning Strategy*

The classification head of an ImageNet-pretrained DenseNet121 is deleted and the model is trained to output one sigmoid per finding. (Usually (e.g. in pytorch [38]), all layers are finetuned with the Adam [39] optimizer at a small learning rate, learning rate schedule based on validation loss, the checkpoint is selected by mean validation AUROC, and confidence intervals are computed from bootstrapping over patients.) Decision thresholds for each finding should be chosen on validation data for the specified operating point (e.g. high sensitivity for triage) instead of default 0.5.

## IV. EXPLAINABILITY

### *A. Taxonomy*

There are different definitions of interpretability and the selection of method has to depend on the purpose of the explanation [40], [41]. Explanation methods for image classifiers can roughly be categorized into gradient/based on backpropagation methods (saliency maps [42], visualization of deconvolution-based methods [43], layer-wise relevance propagation [44], integrated gradients [45]), class activation mapping (CAM) [46], Grad-CAM [47], Grad-CAM++ [48]) based methods and perturbation or surrogate-model-based methods (local interpretable model-agnostic explanations (LIME) [49] and Shapley additive explanations (SHAP) [50]) due to visual noise may be a problem with saliency maps, the generalization is called SmoothGrad [51] which performs averaging of noisy versions of the maps. Table III compares explanation techniques.

TABLE III
EXPLANATION METHODS FOR CXR CLASSIFIERS

| Method | Category | Cost | Operational notes |
|---|---|---|---|
| Saliency [42] | Gradient | Low | Pixel-level but noisy and visually incoherent |
| CAM [46] | Activations | Low | Needs a global-pooling head; Grad-CAM removes this requirement |
| Integrated gradients [45] | Gradient path integral | Moderate | Tens of passes; result depends on the baseline image |
| Grad-CAM [47] | Gradient + activations | Low | Class-discriminative; coarse (7×7 for DenseNet121) |
| Grad-CAM++ [48] | Gradient + activations | Low | Same resolution limit; proposed for multiple instances |
| LIME [49], SHAP [50] | Perturbation / surrogate | High (many passes) | Slow; sensitive to segmentation and sampling choices |

### *B. Grad-CAM for DenseNet121*

The weight for each channel in a late convolutional feature map with Grad-CAM can be calculated by averaging the gradient of the output score for class c on each feature map. For class c and a feature map a withKchannels, the channel weights are

$$\alpha_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{i,j}^k} \tag{1}$$

where $Z$ is the number of spatial positions and $y^c$ is the pre-sigmoid score for class $c$; the class-specific heatmap is then

$$L_{\text{GradCAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right) \tag{2}$$

For DenseNet121, the last feature map is 77 at 224224, so the heatmap is too coarse on bilinear interpolation to get a visualisation, which is upsampled. This is fine for a coarse



(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

indication of where in the lung the basis of a prediction was, but not for contour boundaries, so very subtle nodules may be indistinct or completely invisible. Still, Grad-CAM++ [48] which changes the weightings was devised in particular for multiple instances of an object, which is just the case if a finding is in both lungs.

### *C. How Reliable Are the Heatmaps?*

There is conflicting evidence that heatmaps are grounded in medically-sound logic: sanity checks have demonstrated some saliency methods give similar maps when model weights are changed [52], so explanation methods should be validated and not assumed as truth. A benchmark on chest X-ray classification found some saliency methods were indistinguishable from a radiologist-drawn benchmark in localizing pathology [53]. Models can find "shortcuts, " a weakness of deep networks [54], for example, a pneumonia classifier was found to rely on references to hospital-system names [55], while COVID-19 detectors often rely on confounding signals rather than the virus [56]. Highlighting an area that should contain pathology does not mean it was found, while highlighting an area that is irrelevant can be counter-informative (Ghassemi et al recommend the latter as a method of model debugging over per-patient explanations [9]).

### *D. Evaluating Explanations*

Possible evaluation metrics are (i) localising to per expert label(s), e.g. intersection-over-union or pointing game; (ii) deletion/insertion curves where the scores should go down if the highlighted pixel(s) are removed; (iii) model and data randomisation sanity tests [52]; and (iv) reader studies whether explanations help clinicians' accuracy, trust, or reading time. Human-centred concept of causability [57] relates to how much an explanation offers the human with the level of causal understanding (E) identified will be the metric, not the visualisation.

## V. FULL-STACK WORKSPACE ARCHITECTURE

### *A. Reference Architecture*

The reference architecture is layered (Fig. 1), with a web client providing image upload, window/level and zoom adjustment, an overlay toggle for heatmaps, and a review interface. A rest API [58] offers authentication and role-based authorization, and transitions a case from the initial state through review states. An inference service performs image preprocessing, DenseNet121 scoring, and Grad-CAM generation. A relational database stores case statuses, predictions, and reviewer decisions. An object store stores images and heatmaps. An append-only audit log stores individual accesses and decisions. A model registry stores weight files, to ensure the correct file is used for all predictions, such that each prediction can be traced to the exact weights used to produce it.

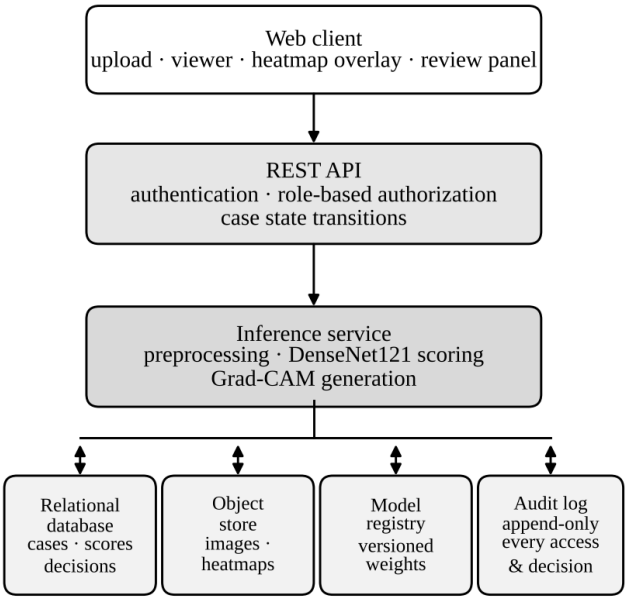



Fig. 1. Layered reference architecture for an explainable CXR workspace. Arrowed lines indicate request flow from client to inference service and read/write accesses by each service to the four storage building blocks.

### *B. Model Serving*

Inference can be performed in-process with the API, with a separate model server, or exported to a high-performance runtime (e.g. using the Open Neural Network Exchange [ONNX] Runtime). Running inference on a second server decouples compute-intensive inference from other parts of the pipeline, scales independently, and enables batching. Running inference as a separate service also brings increased separation of compute time from other pipeline operations. Note that Grad-CAM takes a backwards pass on the model, so it is costlier than scoring and needs modifications to export and serve. For a 224224 input and a lightweight model like DenseNet121, on-device inference will probably be fast enough to work interactively, but this should be measured on the target hardware rather than inferred. Operational topics to consider include model registration and versioning, input validation, monitoring for input drift, and what should happen when the service is unreachable.

### *C. Interoperability, Privacy, and Security*

An interface between the clinic and a workspace (like DICOM [59]), or between a workspace and the rest of the hospital (for example using the HL7 FHIR interface [60]) is also required. These images and data are protected health information, so a workspace should discard or not attempt to catalog unnecessary DICOM header fields, and encrypt data at rest and in transit, not include patient labels in application logs and adhere to legislation such as the Health Insurance Portability and Accountability Act [61] or the General Data Protection Regulation [62]. Many research-based systems, as well as some image processing approaches, upload PNG or JPEG images rather than directly connect to a picture archiving and communication system (PACS); such solutions may be suitable for tutorials and demonstration but are not appropriate for clinical use without substantial engineering and validation. An alternative is federated learning where model



(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

updates are sent rather than images; this avoids the transfer of protected health information when data should not leave a site [63]. The role of federated learning in digital health has been reviewed [64].

## VI. ROLE-BASED CLINICAL REVIEW

### *A. Role-Based Access Control*

Role-based access control (RBAC) brings access based on the permissions (privileges) set for roles and the assignment of users to roles [65]. A standards effort aims to standardize the core, hierarchial, and constrained variants of RBAC [66]. It mirrors clinical work, where roles are based on function, and reduces, least, privileges, and separation of duties. Example of a role set are in Table IV.

TABLE IV
ILLUSTRATIVE ROLES AND PERMISSIONS

| Role | Permitted actions | Workflow restrictions |
|---|---|---|
| Technician | Upload studies; track status of own submissions | Cannot edit findings or finalize reports |
| Radiologist | View image, scores, and heatmap; confirm, amend, or reject AI findings; sign off report | Cannot manage users or model configuration |
| Referring clinician | View finalized reports and images for own patients | Sees reviewed results only; cannot edit |
| Admin | Manage users and roles; inspect audit logs; manage model versions | Cannot alter clinical decisions |

### *B. Human-in-the-Loop Workflow*

ai output should be decision support, not a diagnosis Figure 2 a simple task state flow that we can evolve through a case Uploaded, analysed by the ai, Pending review, Reviewed (amended confirmed rejected) and Finalized (finalised). Only a radiologist role user can progress the case to reviewed or finalised so that a trained clinician keeps rapport accountability for the report, then be presented with calibrated confidence, not raw scores, the heatmap should be optional so it does not anchor the reader, and viewers should be logged if the reviewer does not agree with the model.

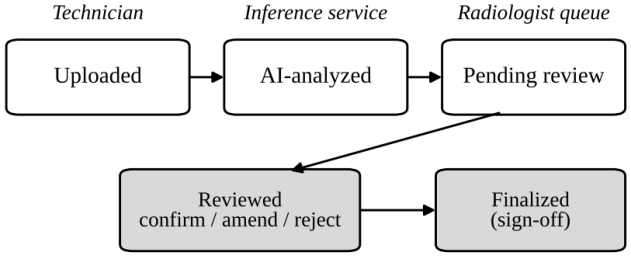



Fig. 2. Case-state workflow with role-restricted transitions.

A noted risk of blind trust in automation (automation bias) in decision support [67] and other studies suggests that there should be no issue in letting the reader formulate a decision before revealing the output of the AI; a reasonable hypothesis for future empirical testing. While promising evidence of benefit is patchy. Multireader studies of chest radiographs found increased radiologist accuracy [68], [69] and efficiency [70] with support from AI, whereas a dermatology study found benefit was dependent on both reader experience and the decision support technology [71].

### *C. Audit and Feedback*

Each action should have an immutable audit log with user role date and time case id model version and decision. Override Reviewer overrides are helpful: aggregate override rates by find can reveal model problems or drift; curated disagreements can work as case data for retraining following governance. Audit logs can support investigations and accountability.

## VII. CASE STUDY: MEDIVISION AI PROTOTYPE

### *A. System Overview*

In support of the survey, here's a sample interface evaluation of the MediVision AI research prototype, which was extracted from interface screenshots taken during its use. It is limited to what can be seen in the screenshots. It isn't a specification at the code level, and if a feature isn't present in the screenshots it doesn't mean that the feature isn't present in the software. The public landing page offers a research-quality imaging workspace based on DenseNet121 transfer learning, multi-label CXR analysis, and Grad-CAM visual explanations, and allows account creation and sign-in. This prototype is included here to contextualize the interface review against the workspace ideas.

After signing in, the user is taken to a private workspace that includes four sections: a general overview, an upload area, a report library, and a profile and security section. Uploads are restricted to JPG, JPEG, and PNG file types with a maximum size of 10 MB; once uploaded, files undergo validity checks and are then displayed on the private history page where their quality status can be reviewed. If an upload is removed, it remains listed and is marked with a "Removed" status. The report library allows searching by name, finding, and model, filtering by status, sorting, and sharing individual results through expiring secure links. The chest X-ray image shown in this paper was obtained from the NIH ChestX-ray14 dataset, which is available for research [3] and contains deidentified medical images. The correspondence between the findings and the survey sections is provided in Table V.

TABLE V
OBSERVED PROTOTYPE FEATURES AND RELATED SURVEY SECTIONS

| Feature | Observed implementation | Section |
|---|---|---|
| Input validation | JPG, JPEG, or PNG up to 10 MB; per-upload quality status | V-B |
| Private archive | Per-user upload history; removal shown as a status | V-C |
| Inference | Four probability outputs: Normal, Pneumonia, Cardiomegaly, Pleural_Effusion | II-A |
| Explanation | Original, processed, heatmap, and overlay views; class selector; on-demand generation | IV-B |
| Traceability | Timestamped analysis record with model version | V-A |
| Sharing | Per-report print and expiring secure links | V-C |
| Monitoring | Dashboard counts of scans, successes, failures, and mean processing time | V-B |
| Safeguard | Per-report notice that outputs are not diagnoses | VI-B |



(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

### *B. Illustrative Report*

The report-view interface includes four panels displaying the original radiograph, the processed image, a Grad-CAM heatmap, and an overlay. This layout illustrates the explanation workflow and demonstrates the approximate spatial resolution found in a 77 DenseNet121 feature map.

Signs of 2% Normal, 53% Pneumonia, 30% Cardiomegaly, 31% Pleural_Effusion are in the results panel. >100% adds up as this is independent sigmoid outputs and not mutually exclusive combinations of classes. The collapsed view does not show what the operating threshold is or any amount of calibration so these values are not calibrated clinical probabilities.

### *C. Operational Observations*

The dashboard snapshot includes 7 uploaded studies (3 done, none processing, 4 failed, averaging 7 s processing time for done reports). This is not a benchmark. The views maintain the failed items with "No findings yet" with a no available model, and the interface do not document reasons for failure, so they are not in the record. The views maintain the completed items with the version label "v2.0-multimodal", again the interface do not document the meaning of this label, and we adopt DenseNet121 as the classification backbone because the landing-page description suggests so.

### *D. Gap Analysis Against the Reference Design*

Table VI compares the observed interface with the reference design of Sections V and VI.

TABLE VI
PROTOTYPE CAPABILITIES VERSUS REFERENCE DESIGN

| Requirement / concept | Status | Observation |
|---|---|---|
| Authentication and private archive | Observed | Account sign-in; private upload history |
| Upload validation | Observed | Format, size limit, quality status checks |
| Multi-label scores and Grad-CAM | Observed | Four outputs; heatmap and overlay display |
| Per-analysis model versioning | Observed | Timestamped analysis record included |
| Expiring share links | Observed | Per-report sharing capabilities |
| Role separation (RBAC) | Not observed | Single user type visible in captured views |
| Reviewer sign-off (confirm/amend/reject) | Not observed | Reports display model output only |
| Audit-log view | Not observed | May exist outside captured interface views |
| Threshold and calibration display | Not visible | Technical parameter details collapsed |

This illustrative interface analysis example examines only one report and one snapshot of a dashboard. It does not evaluate diagnostic accuracy, localization accuracy calibration reliability, or clinical utility. There are no explanations for these four failed analyses, the deployed model is not quantitatively evaluated in this paper, and the prototype code is not released publicly. As a consequence of these limitations, this case study is presented as a design illustration, not a classifier validation.

## VIII. EVALUATION FRAMEWORK

We propose that a workspace of this kind be evaluated at four levels.

1) *Model level:* AUROC for each finding with bootstrap CIs over patients; area under the precision-recall curve under class imbalance; sensitivity and specificity at the chosen operating point; and calibration as measured by expected calibration error [72]. Measures of uncertainty (e.g. Monte Carlo dropout [73] or deep ensembles [74]) can be used to test whether high-uncertainty cases are linked to errors (disentangling data from model uncertainty [75]).
2) *Generalization level:* Internal test set splits on the patient level and external validation with data from a different hospital, device, or population, as performance usually generalize poorly [55], [76]. Metrics should be reported separately for sex, age, and other covariates available [77].
3) *Explanation level:* Localization agreement, sanity checks, and reader-rated usefulness, as in Section IV-D.
4) *Workflow level:* Turnaround time, reviewer override percentage, turnaround time per case, agreement with and without AI, standardised usability questionnaires and so on. Security testing as well: there need to be tests of role-permission: no role should ever be allowed to undertake a task not listed in table IV, and all audit-logging must be complete.

## IX. CHALLENGES AND OPEN ISSUES

### *A. Label Noise and Dataset Bias*

Labels derived from reports encode reporting habits and NLP errors [3], [22]. Radiologist-adjudicated test sets, as used in CheXpert [15], reduce but do not remove this problem.

### *B. Shortcut Learning and Distribution Shift*

Models could pick up acquisition artifacts, device specificities, and hospital-dependent signals [55], [56]. Solutions are multi-centre training, image harmonisation, invariance objectives, and explanation-guided auditing, but none is assured.

### *C. Fairness*

Seyyed-Kalantari et al. observed that classifiers trained on open CXR collections exhibited higher underdiagnosis among disenfranchised patient groups [77]. Subgroup performance should be monitored during deployment and all health-care workspaces should display it. Further examples showed the presence of proxy outcome bias in health-care risk scores [78] and that race can be recovered from medical images by deep models, thwarting efforts to reduce bias [79].

### *D. Faithfulness of Explanations*

Heatmaps are coarse and of uncertain benefit to clinicians [9], [53]. Prospective reader studies, better explanation metrics, and alternatives such as inherently interpretable models [10] should be evaluated.

### *E. Calibration and Uncertainty*

Recent deep networks are frequently miscalibrated [72]. Here, the uncertainty estimates [73] must be calibrated so the reviewer can know when to trust a given score and when to escalate.



(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

*F. Regulation and Governance*

The CXR classifier being used is most likely a medical device software [80], which are in many jurisdictions subject to clinical validation, change control, post-market surveillance and quality requirements, and both in the EU and many jurisdictions subject to regulation as a medical device or a safety component of a medical device, [81], which may also be classified as a high-risk AI system, with requirements being phased in over several years. Regulatory requirements may differ across jurisdictions and change over time; teams should check current regulators' guidance; no jurisdiction-specific claims of compliance are made in this survey.

*G. Foundation Models*

While self-supervised and vision-language models [32], [35], [36], [37] may ease annotation costs, recent benchmark results have revealed the unreliable nature of automated abnormal CXR report generation [82] and variability across hospitals. Despite the computational expense and unpredictable generation artifacts of foundation model s simple predictable classifiers like DenseNet121 are still desirable in production settings. Clinical validation of AI-based CXR triage in a clinical trial has found a shorter turnaround time in the real world, underscoring the importance of evaluating operational impact independently of model accuracy [83]. Text-supervision is not always a prerequisite for CXR encoding [84] like with image-only pretraining like RAD-DINO yet this introduces further concerns about verification, erroneous generated text, and the correct approach to attribution.

## X. CONCLUSION AND FUTURE DIRECTIONS

DenseNet121 and Grad-CAM still represent a simple, studied starting point for an explainable CXR workspace. The literature found model accuracy, explanation coherence and workspace design to be separate issues and leaving readers with separate set of evidence for each. Role-based review with calibrated outputs and audit trails decouple a classifier to a traceable, accountable decision-support device. The MediVision AI case study shows explanations, versioned records and explicit failure states were achievable with a light-weight prototype, but role-based review and quantitative validation were not.

Finally, future work could consider (i) multi-institution external validation with subgroup-level performance, (ii) prospective reader studies evaluating heatmap interpretability and automation bias, (iii) uncertainty-aware triage, (iv) ground-truth calibrated decision thresholds, (v) federated or privacy-preserving training [63], [64], (vi) continuous monitoring with focused feedback from reviewer overrates to identify failures, and (vii) addressing the lack of preexisting standards. Addressing these could help bridge the gap between laboratory prototype and clinically-reliable AI system.

## ACKNOWLEDGMENT

This work received no external funding. The author declares no competing interests.

Funding: no sources of funding were used to prepare this article.

Competing interests: The author has no competing interests to declare.

Author Contributions and AI-Artificial Intelligence Assistance AThe author had written the full manuscript independently, without the use of AI-powered tools for generating content, language editing, or formatting. The author is the sole contributor to the content of the manuscript.

Data availability: we used the openly available NIH ChestX-ray14 dataset, published by NIH Clinical Center (Wang et al (Wang et al., 2017). 2017; URL: https://nihcc.box.com/v/ChestXray-NIHCC) for this study, and no data were created in this work.

*Code availability:* The prototype implementation developed for this study is not currently publicly available. No standalone software package is being released with this paper.

## REFERENCES

[1] E. J. Topol, "High-performance medicine: The convergence of human and artificial intelligence," *Nat. Med.*, vol. 25, no. 1, pp. 44–56, Jan. 2019, doi: 10.1038/s41591-018-0300-7.

[2] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017, doi: 10.1016/j.media.2017.07.005.

[3] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 3462–3471, doi: 10.1109/CVPR.2017.369.

[4] P. Rajpurkar *et al.*, "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," 2017, *arXiv:1711.05225*.

[5] P. Rajpurkar *et al.*, "Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists," *PLoS Med.*, vol. 15, no. 11, Art. no. e1002686, 2018, doi: 10.1371/journal.pmed.1002686.

[6] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2261–2269, doi: 10.1109/CVPR.2017.243.

[7] E. Çallı, E. Sogancioglu, B. van Ginneken, K. G. van Leeuwen, and K. Murphy, "Deep learning for chest X-ray analysis: A survey," *Med. Image Anal.*, vol. 72, Art. no. 102125, 2021, doi: 10.1016/j.media.2021.102125.

[8] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," *BMC Med.*, vol. 17, Art. no. 195, 2019, doi: 10.1186/s12916-019-1426-2.

[9] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," *Lancet Digit. Health*, vol. 3, no. 11, pp. e745–e750, 2021, doi: 10.1016/S2589-7500(21)00208-9.

[10] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nat. Mach. Intell.*, vol. 1, pp. 206–215, 2019, doi: 10.1038/s42256-019-0048-x.

[11] P. Lakhani and B. Sundaram, "Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks," *Radiology*, vol. 284, no. 2, pp. 574–582, 2017, doi: 10.1148/radiol.2017162326.

[12] A. G. Taylor, C. Mielke, and J. Mongan, "Automated detection of moderate and large pneumothorax on frontal chest X-rays using deep convolutional neural networks: A retrospective study," *PLoS Med.*, vol. 15, no. 11, Art. no. e1002697, 2018, doi: 10.1371/journal.pmed.1002697.

[13] J. G. Nam *et al.*, "Development and validation of deep learning–based automatic detection algorithm for malignant pulmonary nodules on chest radiographs," *Radiology*, vol. 290, no. 1, pp. 218–228, 2019, doi: 10.1148/radiol.2018180237.

[14] E. J. Hwang *et al.*, "Development and validation of a deep learning–based automated detection algorithm for major thoracic diseases on



(This work is licensed under a Creative Commons Attribution 4.0 International License.)

chest radiographs," *JAMA Netw. Open*, vol. 2, no. 3, Art. no. e191095, 2019, doi: 10.1001/jamanetworkopen.2019.1095.

[15] J. Irvin *et al.*, "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 590–597, doi: 10.1609/aaai.v33i01.3301590.

[16] A. E. W. Johnson *et al.*, "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," *Sci. Data*, vol. 6, Art. no. 317, 2019, doi: 10.1038/s41597-019-0322-0.

[17] A. Bustos, A. Pertusa, J.-M. Salinas, and M. de la Iglesia-Vayá, "PadChest: A large chest x-ray image dataset with multi-label annotated reports," *Med. Image Anal.*, vol. 66, Art. no. 101797, 2020, doi: 10.1016/j.media.2020.101797.

[18] S. Jaeger *et al.*, "Two public chest X-ray datasets for computer-aided screening of pulmonary diseases," *Quant. Imaging Med. Surg.*, vol. 4, no. 6, pp. 475–477, 2014, doi: 10.3978/j.issn.2223-4292.2014.11.20.

[19] D. S. Kermany *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018, doi: 10.1016/j.cell.2018.02.010.

[20] G. Shih *et al.*, "Augmenting the National Institutes of Health chest radiograph dataset with expert annotations of possible pneumonia," *Radiol., Artif. Intell.*, vol. 1, no. 1, Art. no. e180041, 2019, doi: 10.1148/ryai.2019180041.

[21] H. Q. Nguyen *et al.*, "VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations," *Sci. Data*, vol. 9, Art. no. 429, 2022, doi: 10.1038/s41597-022-01498-w.

[22] L. Oakden-Rayner, "Exploring large-scale public medical image datasets," *Acad. Radiol.*, vol. 27, no. 1, pp. 106–112, 2020, doi: 10.1016/j.acra.2019.10.006.

[23] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2009, pp. 248–255, doi: 10.1109/CVPR.2009.5206848.

[24] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2014, pp. 3320–3328.

[25] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning for medical imaging," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2019, pp. 3347–3357.

[26] S. M. Pizer *et al.*, "Adaptive histogram equalization and its variations," *Comput. Vis., Graph., Image Process.*, vol. 39, no. 3, pp. 355–368, 1987, doi: 10.1016/S0734-189X(87)80156-1.

[27] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2999–3007, doi: 10.1109/ICCV.2017.324.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[29] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 6105–6114.

[30] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.

[31] Z. Liu *et al.*, "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 10012–10022, doi: 10.1109/ICCV48922.2021.00986.

[32] E. Tiu *et al.*, "Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning," *Nat. Biomed. Eng.*, vol. 6, pp. 1399–1406, 2022, doi: 10.1038/s41551-022-00936-9.

[33] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 8748–8763.

[34] S. Azizi *et al.*, "Big self-supervised models advance medical image classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 3478–3488, doi: 10.1109/ICCV48922.2021.00346.

[35] Z. Chen *et al.*, "A vision-language foundation model to enhance efficiency of chest X-ray interpretation," 2024, *arXiv:2401.12208*, doi: 10.48550/arXiv.2401.12208.

[36] Z. Yang, X. Xu, J. Zhang, G. Wang, M. K. Kalra, and P. Yan, "Chest X-ray foundation model with global and local representations integration," *IEEE Trans. Med. Imag.*, vol. 44, no. 12, pp. 4787–4799, Dec. 2025, doi: 10.1109/TMI.2025.3581907.

[37] A. Wang, Z. Yang, S. Rais-Bahrami, and P. Yan, "Chest X-ray foundation models: A survey and future directions," *Meta-Radiology*, vol. 4, no. 1, Art. no. 100203, Mar. 2026, doi: 10.1016/j.metrad.2025.100203.

[38] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2019, pp. 8024–8035.

[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015.

[40] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," 2017, *arXiv:1702.08608*.

[41] Z. C. Lipton, "The mythos of model interpretability," *Commun. ACM*, vol. 61, no. 10, pp. 36–43, 2018, doi: 10.1145/3233231.

[42] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *Proc. Int. Conf. Learn. Represent. (ICLR) Workshop*, 2014.

[43] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 818–833.

[44] S. Bach *et al.*, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS ONE*, vol. 10, no. 7, Art. no. e0130140, 2015, doi: 10.1371/journal.pone.0130140.

[45] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2017, pp. 3319–3328.

[46] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 2921–2929, doi: 10.1109/CVPR.2016.319.

[47] R. R. Selvaraju *et al.*, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.

[48] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, 2018, pp. 839–847, doi: 10.1109/WACV.2018.00097.

[49] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.

[50] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 4765–4774.

[51] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "SmoothGrad: Removing noise by adding noise," 2017, *arXiv:1706.03825*.

[52] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2018, pp. 9505–9515.

[53] A. Saporta *et al.*, "Benchmarking saliency methods for chest X-ray interpretation," *Nat. Mach. Intell.*, vol. 4, pp. 867–878, 2022, doi: 10.1038/s42256-022-00536-x.

[54] R. Geirhos *et al.*, "Shortcut learning in deep neural networks," *Nat. Mach. Intell.*, vol. 2, pp. 665–673, 2020, doi: 10.1038/s42256-020-00257-z.

[55] J. R. Zech *et al.*, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study," *PLoS Med.*, vol. 15, no. 11, Art. no. e1002683, 2018, doi: 10.1371/journal.pmed.1002683.

[56] A. J. DeGrave, J. D. Janizek, and S.-I. Lee, "AI for radiographic COVID-19 detection selects shortcuts over signal," *Nat. Mach. Intell.*, vol. 3, pp. 610–619, 2021, doi: 10.1038/s42256-021-00338-7.

[57] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," *WIREs Data Min. Knowl. Discov.*, vol. 9, no. 4, Art. no. e1312, 2019, doi: 10.1002/widm.1312.

[58] R. T. Fielding, "Architectural styles and the design of network-based software architectures," Ph.D. dissertation, Dept. Inf. Comput. Sci., Univ. California, Irvine, CA, USA, 2000.



(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

[59] *Digital Imaging and Communications in Medicine (DICOM) Standard, PS3*, National Electrical Manufacturers Association, Rosslyn, VA, USA. [Online]. Available: https://www.dicomstandard.org

[60] D. Bender and K. Sartipi, "HL7 FHIR: An agile and RESTful approach to healthcare information exchange," in *Proc. IEEE Int. Symp. Comput.-Based Med. Syst. (CBMS)*, 2013, pp. 326–331.

[61] *Health Insurance Portability and Accountability Act of 1996*, Pub. L. No. 104-191, 110 Stat. 1936, 1996.

[62] *Regulation (EU) 2016/679 of the European Parliament and of the Council (General Data Protection Regulation)*, *Off. J. Eur. Union*, vol. L119, pp. 1–88, May 2016.

[63] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Int. Conf. Artif. Intell. Stat. (AISTATS)*, 2017, pp. 1273–1282.

[64] N. Rieke *et al.*, "The future of digital health with federated learning," *npj Digit. Med.*, vol. 3, Art. no. 119, 2020, doi: 10.1038/s41746-020-00323-1.

[65] R. S. Sandhu, E. J. Coyne, H. L. Feinstein, and C. E. Youman, "Role-based access control models," *Computer*, vol. 29, no. 2, pp. 38–47, Feb. 1996, doi: 10.1109/2.485845.

[66] D. F. Ferraiolo, R. Sandhu, S. Gavrila, D. R. Kuhn, and R. Chandramouli, "Proposed NIST standard for role-based access control," *ACM Trans. Inf. Syst. Secur.*, vol. 4, no. 3, pp. 224–274, 2001, doi: 10.1145/501978.501980.

[67] K. Goddard, A. Roudsari, and J. C. Wyatt, "Automation bias: A systematic review of frequency, effect mediators, and mitigators," *J. Amer. Med. Inform. Assoc.*, vol. 19, no. 1, pp. 121–127, 2012, doi: 10.1136/amiajnl-2011-000089.

[68] J. C. Y. Seah *et al.*, "Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: A retrospective, multireader multicase study," *Lancet Digit. Health*, vol. 3, no. 8, pp. e496–e506, 2021, doi: 10.1016/S2589-7500(21)00106-0.

[69] Y. Sim *et al.*, "Deep convolutional neural network–based software improves radiologist detection of malignant lung nodules on chest radiographs," *Radiology*, vol. 294, no. 1, pp. 199–209, 2020, doi: 10.1148/radiol.2019182465.

[70] J. S. Ahn *et al.*, "Association of artificial intelligence–aided chest radiograph interpretation with reader performance and efficiency," *JAMA Netw. Open*, vol. 5, no. 8, Art. no. e2229289, 2022, doi: 10.1001/jamanetworkopen.2022.29289.

[71] P. Tschandl *et al.*, "Human–computer collaboration for skin cancer recognition," *Nat. Med.*, vol. 26, pp. 1229–1234, 2020, doi: 10.1038/s41591-020-0942-0.

[72] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2017, pp. 1321–1330.

[73] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2016, pp. 1050–1059.

[74] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 6402–6413.

[75] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 5574–5584.

[76] J. P. Cohen, M. Hashir, R. Brooks, and H. Bertrand, "On the limits of cross-domain generalization in automated X-ray prediction," in *Proc. Mach. Learn. Res. (MIDL)*, vol. 121, 2020, pp. 136–155.

[77] L. Seyyed-Kalantari, H. Zhang, M. B. A. McDermott, I. Y. Chen, and M. Ghassemi, "Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations," *Nat. Med.*, vol. 27, pp. 2176–2182, 2021, doi: 10.1038/s41591-021-01595-0.

[78] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019, doi: 10.1126/science.aax2342.

[79] J. W. Gichoya *et al.*, "AI recognition of patient race in medical imaging: A modelling study," *Lancet Digit. Health*, vol. 4, no. 6, pp. e406–e414, 2022, doi: 10.1016/S2589-7500(22)00063-2.

[80] *Software as a Medical Device (SaMD): Key Definitions*, IMDRF SaMD Working Group, Tech. Rep. IMDRF/SaMD WG/N10FINAL:2013, Dec. 2013.

[81] *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)*, *Off. J. Eur. Union*, OJ L, 2024/1689, Jul. 12, 2024.

[82] X. Zhang *et al.*, "Automated chest X-ray report generation remains unsolved," in *Biocomputing 2026: Proc. Pacific Symp.*, vol. 31, 2026, pp. 236–250, doi: 10.1142/9789819824755_0017.

[83] S. Sridharan *et al.*, "Real-world evaluation of an AI triaging system for chest X-rays: A prospective clinical study," *Eur. J. Radiol.*, vol. 181, Art. no. 111783, Dec. 2024, doi: 10.1016/j.ejrad.2024.111783.

[84] F. Pérez-García *et al.*, "Exploring scalable medical image encoders beyond text supervision," *Nat. Mach. Intell.*, vol. 7, pp. 119–130, 2025, doi: 10.1038/s42256-024-00965-w.



(This work is licensed under a Creative Commons Attribution 4.0 International License.)