The Impact of Clinical Features on the Performance of Machine Learning Models Detecting Pleural Effusion From Chest X-Ray Images

Gabriela Zhelyazkova (gazh@itu.dk)
Supervisors: Amelia Jiménez-Sánchez (amji@itu.dk), Veronika Cheplygina (vech@itu.dk)

Abstract—**Chest X-rays are the main tool when coming to diagnosing pleural effusion, a condition that can indicate a range of serious underlying health issues. While machine learning has increasingly been used to support this task, many models still depend only on the image itself and ignore important clinical details about the patient. This study was born out of the recognition of this limitation in current machine learning models in medical imaging. It investigates the impact of incorporating clinical features on the performance of machine learning models trained to detect pleural effusion. We use the CheXpert dataset for images and clinical labels and CheXmask for anatomical masks; we extracted multiple feature types — including lung asymmetry, fluid intensity, histogram distributions, and structural corners. We then combined these features with demographic and diagnostic labels to train XGBoost and ResNet-50-based models across various feature configurations. Results show that clinical features significantly improve both classification accuracy and calibration, with the XGBoost model trained exclusively on clinical variables achieving an AUC of 0.873 and near-perfect calibration, and ResNet-based model trained on clinical features only reached 0.860 AUC. Non-clinical features alone were insufficient, but dimensionality reduction through PCA did not led to degraded performance, but showed stable performance. These findings underscore the importance of clinical context in medical imaging tasks and highlight potential biases introduced by models trained on incomplete feature sets. The study thus emphasizes the need for multimodal approaches integrating visual and clinical data to improve generalizability and decision reliability in diagnostic applications. The relevant scripts for recreating this research and to continue it further could be found in our GitHub repository [1].**

Index Terms—**Pleural effusion, chest X-ray, clinical features, machine learning, deep learning, XGBoost, ResNet, feature extraction, medical imaging.**

I. Introduction

Chest X-ray examination continues to be essential in diagnosing pleural effusion—a condition characterized by fluid buildup in the pleural space, commonly associated with heart failure, infection, cancer, and other critical illnesses. Pleural effusion affects roughly 360 per 100,000 individuals globally each year, corresponding to around 1.5 million new cases annually in the United States alone, with these numbers continuing to rise [1]. The consequences of untreated or complicated effusions can be severe; hospitalized patients have a 30-day mortality rate ranging from 15% to 30%, which can increase to nearly 50% for ICU patients. Moreover, the one-year mortality rate approaches 50%, and malignant pleural effusions are especially dire, carrying a mortality rate exceeding 77% within one year [2].

Given these high stakes, developing reliable and clinically validated machine learning models to detect pleural effusion from chest radiographs has become increasingly important. However, current models often overlook the clinical context of patients, limiting their real-world effectiveness and generalizability [3].In this study we investigated how clinical features—such as lungs asymmetry, presence of fluids, presence of support devices — influence the performance of machine learning models. By conducting targeted experiments that mimic the comprehensive diagnostic approach radiologists use, this research aims to reveal hidden biases and enhance model reliability. Ultimately, this study aims to investigate how incorporating structured clinical features influences the performance and reliability of machine learning models for pleural effusion detection, highlighting the value of clinical context in medical imaging.

II. Related Work

Recent advances in machine learning have significantly impacted medical imaging, particularly in automating the diagnosis of chest pathologies from X-ray images. Rajpurkar et al. (2017) [4] demonstrated that deep learning models, specifically convolutional neural networks (CNNs) such as DenseNet [5], can achieve radiologist-level accuracy in classifying chest radiographs across various diseases, including pleural effusion. Their work emphasized the potential of automated diagnosis to alleviate radiologist workload and improve diagnostic throughput, though it highlighted the necessity of evaluating models on diverse patient populations and ensuring clinical applicability.

Similarly, another study conducted by Hassanpour et al. (2022) investigated the effectiveness of CNNs trained on chest X-rays for the detection of COVID-19-induced pneumonia. They combined radiographic data with clinical variables and found that integrating patient clinical characteristics significantly enhanced diagnostic accuracy and model generalizability. This research reinforced the importance of clinical metadata in radiological machine learning models, suggesting that patient demographics and clinical features can substantially influence model outcomes.

[1] https://github.com/gazhds/Bachelor-Project

Further illustrating the significance of clinical context in medical imaging, Budrys et al. (2018) pointed out that hidden biases, often described as ”artifacts,” such as inconsistent background features or unintended correlations with clinical variables, can affect model predictions. Their findings advocate careful consideration and explicit incorporation of clinical metadata during model training to improve robustness and reduce hidden biases. This aligns with the concept of hidden stratification, where subgroups with distinct clinical characteristics may experience degraded model performance despite strong overall metrics[6]. Recognizing and addressing such stratification is critical to ensuring model reliability across diverse patient populations.

Moreover, in a related investigation into detecting pleural effusion specifically, Golmohammadi et al. (2022) [7] demonstrated that models trained with comprehensive clinical data outperformed image-only models. Their results highlighted that clinical features such as age, sex, and presence of underlying diseases significantly influenced predictive performance, underscoring the value of multi-modal approaches combining imaging and clinical information.

These studies collectively underline the crucial role of clinical context in enhancing the diagnostic accuracy and generalization of machine learning models applied to medical imaging tasks. Building on this foundation, the present study narrows the focus to pleural effusion and investigates how a targeted set of expert-informed clinical features impact model performance. By systematically evaluating the diagnostic and calibration effects of integrating these features into machine learning pipelines, this work contributes a detailed, feature-level perspective on the role of clinical context in disease-specific model design.

III. Data

A. CheXpert

For this study, the CheXpert dataset (Irvin et al., 2019) [8] was utilized as the primary source of chest radiograph images and corresponding clinical data. CheXpert is a publicly available, large-scale chest X-ray dataset compiled by Stanford University, containing 223,414 chest radiographs from 65,540 patients. The big difference in numbers is because most of the patients have more than one image made (between 1 and 5 images per patient). The dataset encompasses a wide spectrum of pathologies, including pleural effusion, and is designed specifically to facilitate machine learning research aimed at chest X-ray interpretation. Each image within CheXpert is accompanied by comprehensive clinical annotations, derived from radiology reports using automated natural language processing (NLP). These annotations provide labels for various pathologies, including pleural effusion, as positive, negative, uncertain, or not mentioned. For the purposes of this research, only radiographs explicitly labeled as positive, negative, or uncertain for pleural effusion were selected to ensure clarity and consistency in model training and evaluation.

The data consist of basic information of the patient, alongside the images, like sex and age, but also with data regarding the position the patient was in when the image was taken (AP (anterior-posterior) images are taken with the X-ray beam passing from front to back, typically used for bedridden patients, whereas PA (posterior-anterior) images are taken from back to front, commonly used in standard, upright imaging); most of the images, for which we have data, are taken from AP view (80%). Furthermore, the dataset provides labels for a range of findings and clinical conditions, including enlarged cardiomediastinum, cardiomegaly, lung opacity, lung lesion, edema, consolidation, pneumonia, atelectasis, pneumothorax, pleural effusion, other pleural abnormalities, fractures, and the presence of support devices.

The dataset includes 40% female and 60% male patients. The age distribution is unimodal and right-skewed, with the majority of patients falling between the ages of 50 and 80, peaking in the 60–70 age group (Fig.1)

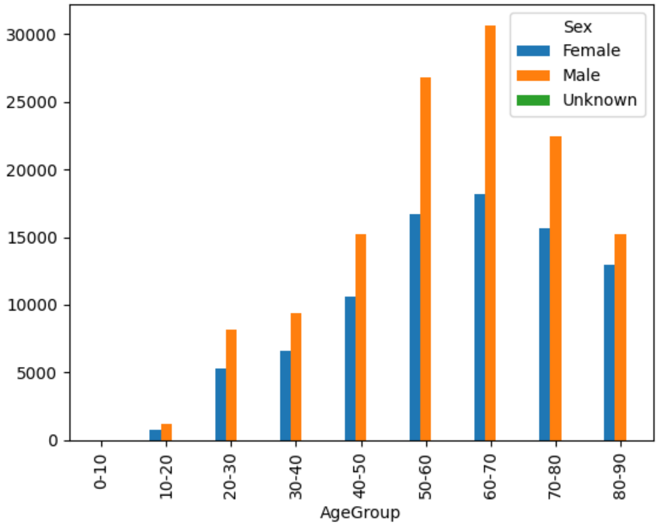

Fig. 1. Distribution of sex and age groups.

The data is not consistent through all the diseases, so for the sake of this project the disease with highest consistency rate was chosen (Pleural Effusion). Pleural Effusion has the following distribution: 86,187 Positives, 35,396 Negatives, and 11,628 Uncertain, which add up to 133,211 or more than half of the data we have. This means that in the end only those 133,211 images would be used for the model training and testing.

B. CheXmask

To incorporate anatomical context and enhance interpretability, segmentation masks from the publicly accessible CheXmask dataset (Gaggion et al, 2023) [9] were used. CheXmask complements CheXpert [8] by providing pixel-level annotations delineating specific anatomical structures and pathological regions on chest radiographs. The regions included left lung, right lung, and heart coordinates.

These masks were generated by expert radiologists and trained annotators, ensuring high annotation accuracy and quality.

The CheXmask dataset is compiled from six public databases: CANDID-PTX [10], ChestX-ray8 [11], Chexpert[8], MIMIC-CXR-JPG[12], Padchest[13], and VinDr-CXR[14], and consists of 676,803 masks. As for this project only CheXpert is used, accordingly only the masks for it are used for further work.

IV. Methodology

This study investigates the impact of clinically informed features on the performance of machine learning models for pleural effusion detection. The methodology comprises several key components: (1) preprocessing of chest X-ray images and associated metadata from the CheXpert dataset; (2) extraction of both image-based and clinically motivated features—including greyscale histograms, structural corners, lung asymmetry, and fluid intensity—from the original images and segmentation masks provided by CheXmask; and (3) classification using two model types: a deep learning model based on a dual-branch ResNet-50 architecture that integrates visual and structured tabular inputs, and an XGBoost classifier trained solely on tabular features. These models were evaluated across different combinations of clinical and non-clinical features to assess their relative contributions to diagnostic accuracy and calibration. A visual overview of the full pipeline is presented in Figure 2.

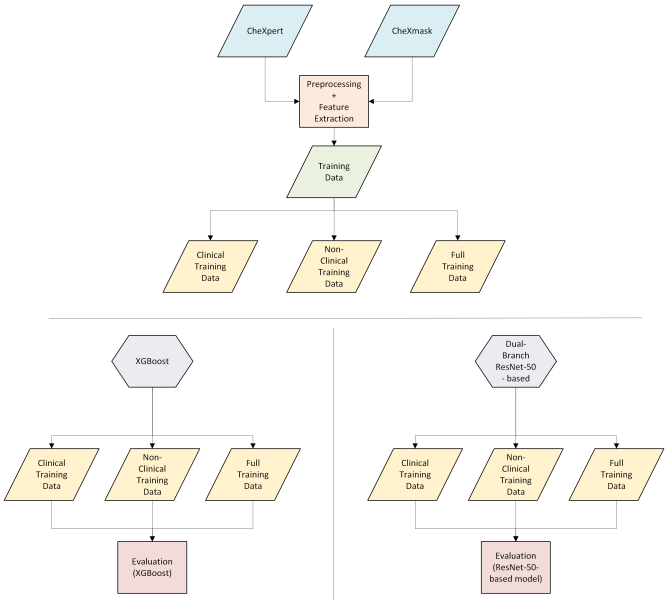

Fig. 2. Research Process Flowchart

A. Preprocessing

The CSV metadata from the CheXpert dataset was initially loaded and examined to ensure accurate representation and integrity. We selected columns containing patient demographics, clinical labels, and imaging details, and preliminary exploratory data analysis (EDA) was conducted. This involved inspecting dataset columns, checking distributions, and visualizing sample images to ensure data validity and quality. We filtered the dataset specifically to include entries explicitly labeled for the presence or absence of pleural effusion, removing entries with missing labels to maintain clarity in subsequent analyses.

The segmentation masks obtained from the CheXmask dataset were provided in Run-Length Encoding (RLE) format. Each mask annotation consisted of RLE strings stored in CSV files alongside corresponding image dimensions (height and width). To obtain usable pixel-level masks, a custom decoding function was implemented to convert RLE strings into binary mask arrays. These binary arrays explicitly represented anatomical structures, such as the right and left lungs, enabling accurate delineation of regions of interest (Fig. 2).

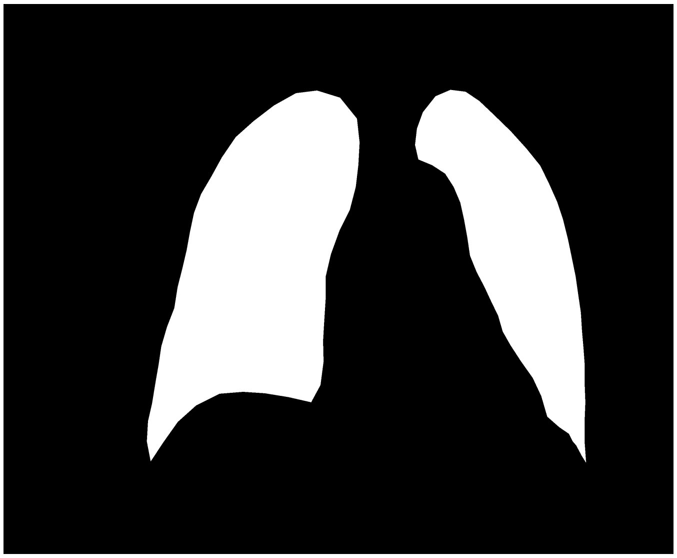

Fig. 3. Visualization of a binary mask

B. Features Extraction

The selection of features in this study was driven by established radiological reasoning and the clinical heuristics typically employed by radiologists when diagnosing pleural effusion. We categorized the features into two main groups: (1) clinical features, such as patient disease labels derived from the CheXpert metadata (e.g., presence of support devices and/or other diseases), as well as image-based clinical features designed to mimic diagnostic cues used by radiologists—namely, anatomical asymmetry and fluid intensity. (2) Non-clinical features included demographics (age, sex) and general image-derived features, such as greyscale histograms and Harris corner detection[15].

In clinical radiology, symmetry between the lungs is a key visual cue assessed when diagnosing pleural effusion. Radiologists often look for differences in size, shape, or brightness between the left and right lungs—changes that can indicate fluid buildup, collapsed lung regions, or mass effect. For instance, pleural effusion typically causes one side of the chest to appear whiter (more opaque) due to the presence of fluid. These asymmetries are well-established diagnostic indicators of conditions like pleural effusion, atelectasis, or lung compression[16]. Another common clinical sign is the accumulation of fluid at the bottom of the lungs, which appears as blunting of the normally sharp costophrenic angle (formed by the points at which the chest wall and diaphragm meet) or a curved "meniscus" line[17]. These intuitive visual features guide radiologists in practice and form the rationale for including asymmetry and fluid intensity as key input features in this study's machine learning models.

Base level features. To quantify the overall brightness distribution of the chest X-ray images, a greyscale histogram was computed directly on the original, unmasked radiographs. Each image, represented in 8-bit format, contained pixel intensity values ranging from 0 (black) to 255 (white). These values were binned into a fixed number of intervals, and the number of pixels falling into each bin was counted to form the histogram. The result was a feature vector capturing the overall intensity profile of the image, reflecting not only the anatomical structures but also potential artifacts, background shading, and device presence. While this approach introduced non-lung regions into the histogram, it provided a global view of radiographic density that could still reflect pleural effusion and other abnormalities. To ensure consistency across images of varying sizes and exposure levels, the histograms were normalized by the total pixel count.

Corner-based features were extracted from the full, unmasked chest X-ray images using the Harris Corner Detection algorithm [15]. Each image was first loaded and converted to grayscale to ensure consistent processing across the dataset. To enhance structural details and improve the detection of relevant features, histogram equalization was applied, thereby increasing the contrast between anatomical structures and surrounding tissues. The Harris corner response was then computed, generating a map that identified regions with strong local intensity variation. Prominent corners were extracted using a peak detection method with a minimum distance threshold to avoid clustering. The resulting corner coordinates were flattened into one-dimensional feature vectors. To standardize feature input across all samples, each feature vector was either padded with zeros or truncated to a fixed length of 200 elements. For missing or unreadable images, a zero-filled feature vector was used. This approach captured meaningful structural complexity across the radiographs, providing a robust descriptor of variations potentially associated with pleural effusion, lung abnormalities, or medical devices.

Asymmetry. For the calculation of the asymmetry first a rectangular was put around each lung masks. Then the right lung was flipped horizontally so that to be in the same direction as the left one. The mirrored mask was put on top of the left lung mask and then the fact that the masks are binary was used. The binary mask is actually a matrix consisting of 0s and 1s where each entry is actually a pixel. When putting the right mask on top of the left one, we are actually performing matrix addition. That means pixels could have a value of 0 (background), 1 (non-overlapping area from one lung), or 2 (overlapping area where the flipped right lung and left lung matched) (Fig.3). The asymmetry ratio was then calculated as the proportion of non-overlapping lung pixels (value = 1) relative to the total number of pixels belonging to either lung (sum of pixels with value 1 and value 2). Mathematically, this was computed as the sum of pixels with value 1 divided by the sum of all pixels with values 1 and 2 (as shown in equation 1). This method provided a robust, image-based metric of asymmetry, accounting for both shape and position, and allowed for more nuanced analysis of anatomical abnormalities such as those caused by pleural effusions.

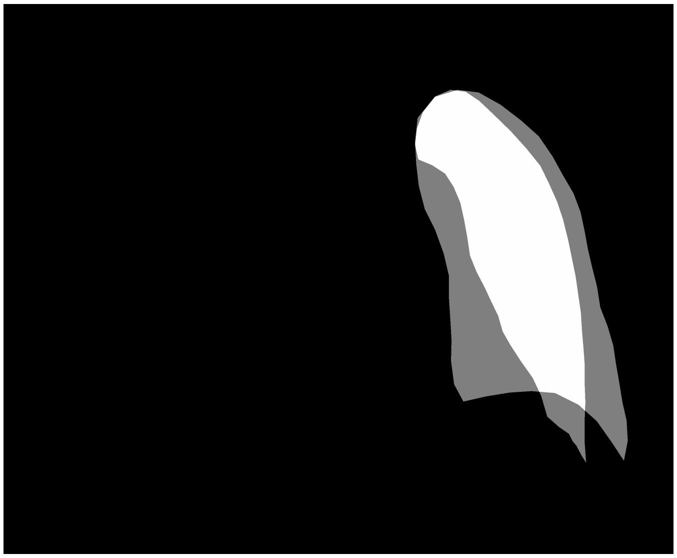

Fig. 4. Visualization of overlapping lungs

$$\text{Ratio} = \frac{\text{Non-overlapping lung pixels}}{\text{Total lung pixels}} = \frac{Sum(1s)}{Sum(1s) + Sum(2s)} \tag{1}$$

Fluid Levels. For the calculation of the fluid levels feature, the pleural effusion mask from the CheXmask dataset was first extracted and overlaid onto the corresponding chest X-ray image. The region corresponding to the detected fluid accumulation was isolated by applying the binary mask directly to the image pixels. In the resulting masked area, the distribution of pixel intensity values was analyzed. Since fluid appears more radiopaque (whiter) compared to aerated lung tissue on X-ray images, pixel brightness was used as a proxy to quantify fluid presence. The mean pixel intensity within the masked region was calculated, representing the average "whiteness" level associated with the effusion. Additionally, higher-order statistics such as median intensity and intensity variance were computed to better capture the spread and distribution of fluid-related opacity (Fig.4). This approach provided a robust, quantitative metric for characterizing the extent and density of pleural fluid, allowing for a more nuanced interpretation of disease severity based on radiographic features.

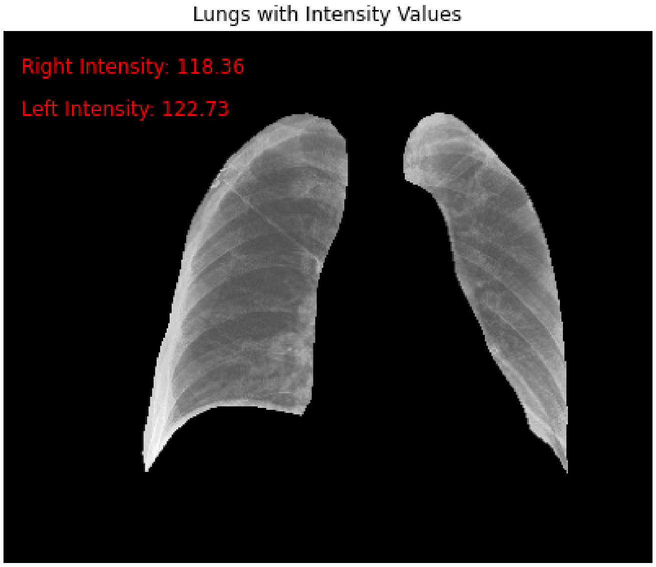

Fig. 5. Visualization of fluid levels

V. Models

To investigate the impact of different types of features on pleural effusion detection, and to compare performance across model types, two parallel modeling strategies were adopted: a deep neural network incorporating chest X-ray images and structured data, and a gradient boosting decision tree model (XGBoost) trained on structured features alone.

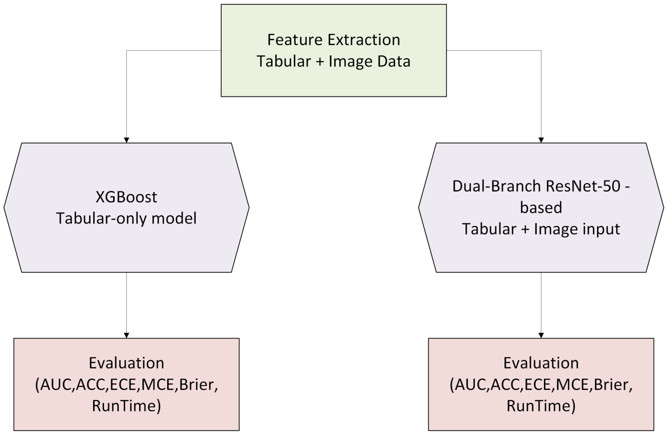

Fig. 6. Comparison of the two models

A. *XGBoost*

Initially, we considered simpler machine learning models such as logistic regression and k-nearest neighbors (KNN) for classifying pleural effusion based on tabular features [18]. However, the high proportion of missing values across various clinical attributes, such as undocumented disease labels, posed a significant challenge. Traditional models failed to train effectively due to the prevalence of NaN values. After researching potential strategies for handling missing data, it became clear that common approaches, such as imputing missing values with the mean, standard deviation, or most frequent label, would artificially alter the dataset and potentially obscure meaningful patterns. Given these considerations, we selected XGBoost[19] for its inherent ability to natively handle missing values during training without the need for imputation. Alternative gradient boosting methods like CatBoost[20] and LightGBM[21] were also considered; however, XGBoost was ultimately chosen due to its proven reliability, tree-based structure, and familiarity. This choice ensured that the model could fully leverage the available data while preserving the integrity of the original feature distribution.

B. *ResNet-50-based dual-branch classifier for image and structured data*

The primary objective of this study is to investigate whether the integration of structured, clinically validated features—rooted in radiological expertise and anatomical segmentation—can enhance the performance of deep learning models in classifying pleural effusion from chest radiographs. To this end, we extended the standard ResNet-50 architecture into a multi-input, dual-branch neural network that combines visual and structured feature modalities.

At the core of the architecture lies a modified ResNet-50 backbone[22], adapted to process grayscale chest X-ray images from the CheXpert dataset. The initial convolutional layer was reconfigured to accept single-channel input, and the classification head was replaced with an identity mapping. This allows the CNN to function purely as a visual feature extractor, producing a 2048-dimensional embedding per image.

In parallel, a second branch processes structured tabular data, composed of expert-informed clinical non-clinical features. These include asymmetry scores between lung fields, fluid intensity distributions, histogram-based brightness metrics, and corner-based structural cues.

To address missing values in the tabular data, we applied a k-nearest neighbors (k-NN) imputation strategy (k=5) based on Euclidean distances. This preserves local structure in the data while mitigating biases from global statistical assumptions[23]. The cleaned features are processed by a two-layer MLP with 128 and 64 ReLU-activated neurons, respectively, resulting in a 64-dimensional structured embedding.

This embedding is concatenated with the 2048-dimensional image vector, forming a unified 2112-dimensional representation. A shared classification head—comprising a dense layer with 64 ReLU-activated units and a softmax output—produces predictions across three classes: positive, negative, and uncertain.

This late fusion design allows each modality to be processed independently and then merged at the decision level, providing modularity and interpretability. Compared to early or intermediate fusion, this strategy better mirrors clinical workflows, where structured observations complement visual assessments. Our architecture prioritizes clinical alignment and explainability over state-of-the-art complexity, serving as a testbed for evaluating the diagnostic value of expert-informed features.

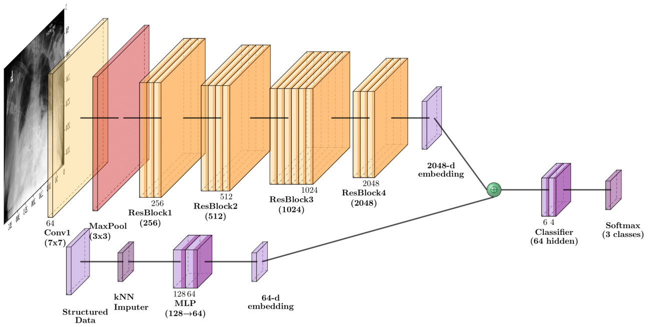

Fig. 7. Architecture Diagram

C. Configuration

All models were trained using an 80/20 stratified train-test split to maintain class balance. The neural network models were trained for 10 epochs with a batch size of 64, using the Adam optimizer and sparse categorical cross-entropy loss. XGBoost models were trained with default regularization and early stopping based on validation performance.

VI. Experiments

To evaluate the impact of different types of features on pleural effusion classification, both models—XGBoost and the Dual-Branch ResNet-50 neural network—were trained using three distinct experimental configurations. These configurations are referred to throughout the paper as:

- **Non-Clinical:** Includes only demographic variables (age, sex) and engineered non-clinical features derived from images, such as greyscale histograms and Harris corners. Structured clinical variables were excluded.
- **Clinical Only:** Includes only structured clinical features, such as diagnostic labels and relevant pathology indicators. Demographic and non-clinical image-derived features were excluded.
- **Full:** Combines all available structured features—demographic, clinical, and non-clinical. For the ResNet-based model, these were paired with the chest X-ray images; for XGBoost, only the structured features were used.

These configurations were designed to isolate the contribution of each feature group and assess their individual and combined effects on classification performance and calibration.

A. Evaluation metrics

We evaluated all models on the test set using a combination of classification and calibration metrics to comprehensively assess performance. Specifically, we measured area under the receiver operating characteristic curve (AUC) and accuracy to evaluate discriminative power. However, recent work has highlighted the limitations of relying solely on classification metrics in clinical decision support contexts[24].

To capture the quality of the predicted probabilities, we also included Expected Calibration Error (ECE), which quantifies the average difference between predicted confidence and actual accuracy across bins, and Maximum Calibration Error (MCE), which reflects the worst-case deviation. Additionally, we reported the Brier score as a proper scoring rule that penalizes both miscalibrated and overconfident predictions. We also recorded training time for each configuration to compare computational efficiency.

VII. Results

This section presents the performance outcomes of both the XGBoost and ResNet-based models across different feature configurations, highlighting how clinical, non-clinical, and combined inputs impact classification accuracy and calibration.

A. Model Performance by Feature Type

XGBoost. The strongest performing configuration for XGBoost was the model trained solely on **clinical features**. This model achieved the highest AUC of **0.869**, accuracy of **0.849**, and the lowest calibration error (ECE: **0.006**, Brier Score: **0.079**). It also had the shortest runtime (3.3 seconds), highlighting its computational efficiency. These results suggest that clinical metadata alone captures most of the discriminative signal for pleural effusion detection.

The **full dataset configuration**, which included all clinical, non-clinical, and engineered image features (e.g., histogram and corner features), yielded a slightly lower AUC of 0.865 and accuracy of 0.847, but achieved the best maximum calibration error (MCE: **0.052**) and required more computation (41.4 seconds).

By contrast, using **only non-clinical features** led to the weakest performance among these configurations (AUC: 0.653, Accuracy: 0.696). These features—such as demographic variables and image-derived structural metrics—were not sufficient to capture the complexity of pleural effusion patterns, and calibration was also poorer

(ECE: 0.014, Brier: 0.144).

TABLE I
XGBOOST RESULTS – MODEL PERFORMANCE BY FEATURE TYPE

Metric	Non-Clinical	Clinical	Full
AUC	0.653	**0.869**	0.865
Accuracy	0.696	**0.849**	0.847
ECE	0.014	**0.006**	0.008
MCE	0.103	0.091	**0.052**
Brier	0.144	**0.079**	0.080
Runtime	51.2s	**3.3s**	41.4s

ResNet. The ResNet-based model showed a similar trend in which the **clinical configuration** again outperformed others, achieving the highest AUC of **0.860**, accuracy of 0.837, and the lowest Brier score (0.084). The model trained on the **full feature set** performed slightly worse in terms of discrimination (AUC: 0.822, Accuracy: 0.814) and required the least training time (69 minutes), suggesting that feature complexity does not necessarily improve neural network performance in this setting.

The **non-clinical configuration**, while performing better than its XGBoost counterpart, still lagged behind in calibration and predictive performance (AUC: 0.766, Brier: 0.100, ECE: 0.033), indicating that image-derived and demographic features alone are not sufficient for robust clinical classification. These findings mirror the XGBoost results and further underscore the central role of clinically meaningful features in driving diagnostic performance across both model types.

For the ResNet-50-based models, it is important to note that training was conducted on a High-Performance Computing (HPC) system equipped with GPUs. Due to the shared nature of the cluster and the scheduling policies in place, the model trained on the full dataset was executed on a different compute node than the other two experiments. This variation in hardware and scheduling conditions may partially account for the observed differences in runtime for the full-dataset model.

TABLE II
RESNET RESULTS – MODEL PERFORMANCE BY FEATURE TYPE

Metric	Non-Clinical	Clinical	Full
AUC	0.766	**0.860**	0.822
Accuracy	0.811	**0.837**	0.814
ECE	0.033	**0.014**	0.031
MCE	**0.047**	0.063	0.048
Brier	0.100	**0.084**	0.096
Runtime	94.68 m	91.34m	**69.02 m**

B. PCA-Based Feature Configurations

To reduce dimensionality and improve computational efficiency, Principal Component Analysis (PCA) was applied to the histogram and corner-based engineered features. These subsets introduced a high number of variables—particularly the flattened corner coordinates—which were potentially redundant. PCA was used to compress these into a more compact representation before training XGBoost models. The transformation was performed using an explained variance ratio (EVR) threshold of 80%, which resulted in 105 retained components.

In the context of XGBoost, PCA proved to be a reasonable design choice. The **PCA + Full** configuration, which combined clinical, non-clinical, and PCA-compressed engineered features, achieved an AUC of 0.867 and accuracy of 0.847—nearly identical to the non-PCA full-feature setup (AUC: 0.865, Accuracy: 0.847). Calibration metrics remained strong (ECE: 0.008, Brier: 0.079), and the runtime was significantly reduced from 41.4 seconds to just 29.65 seconds—a nearly twofold improvement in computational speed.

The **PCA + Non-Clinical** setup, while still underperforming compared to configurations with clinical features, achieved similar accuracy (0.687) to its non-PCA counterpart (0.696) and even showed a lower MCE (0.024 vs. 0.103), suggesting slightly improved confidence alignment. However, the AUC remained low (0.644), and the Brier score was the worst among all setups (0.147), indicating limited predictive power when clinical features were absent.

These results suggest that when applied selectively, PCA can serve as a useful preprocessing step for XGBoost—maintaining strong predictive and calibration performance while substantially reducing computational cost. Unlike in the deep learning pipeline, where feature representation is learned end-to-end, tree-based models like XGBoost can benefit from well-structured dimensionality reduction, particularly when input feature sets are large and partially redundant.

In summary, PCA did not significantly degrade model quality in the full-feature setting and yielded measurable improvements in efficiency. This balance of performance and speed makes it a viable option for structured tabular pipelines, especially when working with engineered high-dimensional inputs.

Given the efficiency gains and minimal performance loss observed in the XGBoost pipeline, it would be worthwhile to explore the integration of PCA into the dual-branch ResNet-50-based architecture. Future research may investigate the impact PCA would have on the performance of the deep learning model.

TABLE III
XGBOOST RESULTS – PCA-BASED CONFIGURATIONS

Metric	PCA + Full	PCA + Non-Clinical
AUC	0.867	0.644
Accuracy	0.847	0.687
ECE	0.008	0.015
MCE	0.037	0.024
Brier	0.079	0.147
Runtime	29.65s	29.85s

VIII. DISCUSSION

In this section, we interpret the comparative results of XGBoost and ResNet, assess the role of different feature groups, and reflect on the broader implications of calibration, model complexity, and feature selection for clinical machine learning.

The results from the XGBoost experiments clearly indicate the critical role of clinical features in the reliable detection of pleural effusion from chest radiographs. Among all feature configurations, models trained exclusively on clinical data consistently outperformed those using only non-clinical or engineered features, achieving the highest AUC and best calibration performance while maintaining extremely low computational cost. This reinforces findings from prior work that underscore the diagnostic relevance of structured clinical variables in medical imaging models.

The performance gap between clinical and non-clinical configurations further illustrates the limitations of relying on demographic or image-derived features—such as histogram distributions and corner-based texture metrics—when used in isolation. While these non-clinical features may capture general anatomical information or image structure, they appear insufficient for characterizing disease-specific signals without accompanying clinical context. This suggests that such features may serve a supplementary rather than primary role in decision-making frameworks for pathology detection.

The PCA-based experiments offer valuable insight into managing high-dimensional engineered features. In the full-feature setting, PCA reduced runtime by nearly 30% without any meaningful drop in AUC, accuracy, or calibration performance. This suggests that dimensionality reduction can be a viable strategy for improving computational efficiency when applied to subsets of features with potential redundancy, such as histograms and corner coordinates. The decreased components (from 425 to 105), but the stable performance showed that the use of features, extracted from the images without medical background and reasoning, is not the right strategy behid optimizing machine learning models used for medical scenarios.

Importantly, calibration metrics such as ECE, MCE, and Brier score remained nearly identical across PCA and non-PCA configurations in the full-feature setup. This indicates that PCA did not compromise the model's ability to generate reliable probability estimates. These results reinforce that, in structured medical data pipelines, PCA can simplify input space and reduce computational cost without degrading either classification accuracy or probabilistic reliability—provided that strong features, such as clinical variables, remain intact.

In high-stakes applications such as diagnostic support for pleural effusion, where decisions depend on predicted confidence levels, these distinctions are not only statistically relevant but clinically essential [25].

These observations from XGBoost are echoed in the performance of the ResNet-based model. The ResNet-based experiments mirrored many of the findings observed with XGBoost. The model trained solely on clinical features consistently outperformed the other configurations, achieving the highest AUC, accuracy, and best calibration metrics among the ResNet runs. Interestingly, while the full feature configuration slightly reduced training time—likely due to more compact gradient updates from redundant inputs—it did not lead to improved performance. This suggests that the addition of engineered image-derived features (histograms and corners) did not contribute meaningful complementary information beyond what was already captured in the convolutional layers processing the X-ray images.

Compared to XGBoost, the neural network model benefited more from the inclusion of visual inputs but still depended heavily on structured clinical data for optimal performance. This reinforces the idea that clinical features are not just informative—they are essential for robust, calibrated classification of pleural effusion, even in more complex architectures that incorporate raw imaging data.

Taken together, the findings across both models suggest that while non-clinical features may offer marginal performance gains when combined with clinical inputs, they are insufficient on their own and should be viewed primarily as augmentative. Clinical features remain the core drivers of reliable classification. Furthermore, attempts to reduce complexity through dimensionality reduction must be further researched and applied carefully.

Ultimately, the results support the broader conclusion that multimodal models combining clinical and imaging information hold promise for pleural effusion detection, but only when the clinical component is robustly represented. Moreover, achieving trustworthy performance requires not only strong classification scores but also well-calibrated probability estimates—especially in high-stakes clinical settings where overconfidence can have serious implications.

IX. FUTURE WORK

There are several directions this work could be extended in the future. One of the most immediate steps would be to test the current models on external datasets (e.g., MIMIC-CXR[12], PadChest[13], VinDr-CXR[14], and ChestX-ray14[11]) to better understand how well they generalize across different hospitals, imaging equipment, and patient populations. Since this study focused solely on CheXpert, it's unclear how models trained on this data would perform in a completely new setting.

It would also be valuable to move beyond simple binary classification. Predicting the severity or progression of pleural effusion, or identifying co-existing conditions, would bring the models closer to how radiologists work in real life. Lastly, in situations where clinical data isn't readily available, future research could look into whether this information can be estimated from the image itself, for example through multi-task learning or auxiliary models trained to detect support devices or other diagnoses visually.

X. Limitations

The limitations of this study can be broadly divided into technical and data-related aspects. First, all experiments were conducted using the CheXpert dataset, which provides structured clinical metadata alongside the chest radiographs. This setup may not reflect the conditions of other datasets where such metadata is missing, inconsistently labeled, or embedded in unstructured formats. In those cases, relying on tabular clinical features—as this study shows to be key for XGBoost performance—may not be practical or even possible.

Additionally, it is valuable to mention that the data provided alongside with the images is crucial for the high performance of the model. In case where another dataset is used, the researchers may have to extract a lot of the features themselves, e.g. the presence of support devices like tubes, etc.

Finally, no external validation was performed using independent datasets. The model's calibration and discriminative performance were assessed using internal train-validation splits, which may not reflect generalization to new populations, imaging equipment, or clinical workflows.

XI. Conclusion

This study set out to explore how incorporating clinical features can improve the performance and reliability of machine learning models detecting pleural effusion from chest X-rays. By combining image-based inputs with structured clinical data, the experiments showed that models benefit greatly from the added context, achieving both higher accuracy and better calibration. Clinical metadata played a central role in driving performance, especially in models relying on structured tabular data.

These findings suggest that models trained without access to clinical information may be limited in their effectiveness, particularly for complex or ambiguous cases. At the same time, the results highlight that when clinical features are available in structured form—as with the CheXpert dataset—they can significantly enhance model performance even without deep image analysis. In scenarios where such metadata is unavailable, alternative strategies may be needed, such as training models to infer these signals directly from the image.

Finally, this study demonstrates that prioritizing clinically grounded features can yield both accurate and well-calibrated diagnostic models. While ResNet-50 was used as a representative deep learning model, its architecture was not tuned for peak performance, as the study's aim was not model optimization but understanding feature contribution. Future research could extend these findings by validating models on external datasets, exploring advanced fusion strategies, and investigating methods to embed clinical insights directly into models when structured metadata is unavailable.

References

[1] S. Walker *et al.*, "Epidemiology and management of pleural effusion," *BMJ*, vol. 377, e069196, 2022.

[2] D. Feller-Kopman and R. Light, "Pleural effusion," *New England Journal of Medicine*, vol. 389, no. 12, pp. 1129–1142, 2023.

[3] L. Seyyed-Kalantari, H. Zhang, M. McDermott, I. Chen, and M. Ghassemi, "Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations," *Nature Medicine*, vol. 27, no. 12, pp. 2176–2182, 2021. DOI: 10.1038/s41591-021-01595-0. [Online]. Available: https://www.nature.com/articles/s41591-021-01595-0.

[4] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, *et al.*, "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *arXiv preprint arXiv:1711.05225*, 2017.

[5] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4700–4708.

[6] L. Oakden-Rayner, A. L. Beam, and L. J. Palmer, "Hidden stratification causes clinically meaningful failures in machine learning for medical imaging," *Proceedings of the National Academy of Sciences*, vol. 117, no. 48, pp. 30 678–30 686, 2020. DOI: 10.1073/pnas.1919012117. [Online]. Available: https://arxiv.org/abs/1909.12475.

[7] A. Golmohammadi *et al.*, *Enhancing pleural effusion detection using clinical data in deep learning models*, CS229 Final Report, Stanford University, 2022.

[8] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, C. Ciurea-Ilcus, C. Chute, *et al.*, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 590–597.

[9] N. Gaggion, A. Iannessi, P. Ballester, *et al.*, *Chexmask: Chest x-ray segmentation masks across multiple datasets*, https://physionet.org/content/chexmask-cxr-segmentation-data/1.0.0/, Accessed: 2025-05-13, 2023.

[10] A. Smit, W. Halsey, V. Harish, *et al.*, "Candid-ptx: A dataset for pneumonia localization and severity grading on chest radiographs," *arXiv preprint arXiv:2211.05158*, 2022. [Online]. Available: https://arxiv.org/abs/2211.05158.

[11] 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," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2097–2106, 2017.

[12] A. E. Johnson, T. J. Pollard, S. J. Berkowitz, *et al.*, "Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs," *arXiv preprint arXiv:1901.07042*, 2019. [Online]. Available: https://arxiv.org/abs/1901.07042.

[13] 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," *Medical Image Analysis*, vol. 66, p. 101 797, 2020. DOI: 10.1016/j.media.2020.101797.

[14] H. H. Nguyen, Q. B. Tran, H. H. Le, *et al.*, "Vindr-cxr: An open dataset of chest x-rays with radiologist annotations," *Scientific Data*, vol. 9, no. 1, pp. 1–7, 2022. DOI: 10.1038/s41597-022-01181-4.

[15] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proceedings of the Alvey Vision Conference*, vol. 15, Manchester, UK, 1988, pp. 10–5244.

[16] R. W. Light, *Pleural Diseases*, 6th. Philadelphia, PA: Lippincott Williams & Wilkins, 2013.

[17] T. Collins and S. Sahn, "Pleural fluid analysis: Diagnostic utility and practical aspects," *Clinics in Chest Medicine*, vol. 19, no. 2, pp. 219–240, 1998.

[18] E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, and B. Van Calster, "A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models," *Journal of Clinical Epidemiology*, vol. 110, pp. 12–22, 2019.

[19] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2016, pp. 785–794. DOI: 10.1145/2939672.2939785.

[20] A. V. Dorogush, V. Ershov, and A. Gulin, "Catboost: Gradient boosting with categorical features support," in *Proceedings of the Workshop on ML Systems at NIPS 2017*, 2017. [Online]. Available: http://learningsys.org/nips17/assets/papers/paper_11.pdf.

[21] G. Ke, Q. Meng, T. Finley, *et al.*, "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[23] G. E. Batista and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 20–29, 2002.

[24] L. Maier-Hein, A. Reinke, P. Godau, *et al.*, "Metrics reloaded: Recommendations for image analysis validation," *arXiv preprint arXiv:2206.01653*, 2022.

[25] R. Roelofs, C. Mitchell, A. Berke, S. Bhadra, M. J. Sheller, B. Do, *et al.*, "Common pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using chest radiographs and ct scans," *arXiv preprint arXiv:2008.06388*, 2020.