Multimodal Analysis of Medical Data

TABLE OF CONTENTS

Introduction

Pneumonia is a prevalent and potentially life-threatening respiratory condition that remains a major contributor to morbidity and mortality in hospital and intensive care settings. Rapid and accurate diagnosis is critical to guiding timely treatment and improving patient outcomes. Clinicians commonly rely on structured electronic health record (EHR) data such as laboratory measurements and demographic information, alongside radiographic assessment of chest X-rays (CXRs) to inform diagnostic decisions. While both data modalities offer valuable clinical insights independently, integrating them through multimodal machine learning has the potential to enhance diagnostic performance beyond what unimodal models can achieve.

In this study, we leverage the Symile-MIMIC dataset, which pairs structured clinical data with CXR images, to systematically evaluate and compare predictive models across structured, unstructured, and multimodal frameworks. By assessing standard performance metrics including area under the receiver operating characteristic curve (AUROC), F1-score, precision, and recall, with particular emphasis on recall due to its clinical importance, we investigate whether multimodal fusion yields superior performance for pneumonia detection. Through this comprehensive analysis, we aim to clarify the value of multimodal learning in supporting automated, data-driven clinical decision making.

1. Business Understanding

1.1 Background

Pneumonia remains a leading cause of morbidity and mortality, particularly in acute and intensive care settings where timely clinical decision-making is essential. Early detection is critical, as delayed diagnosis can significantly worsen patient outcomes. In modern hospital environments, clinicians typically rely on a combination of structured electronic health record (EHR) data, such as laboratory results and demographic variables, and unstructured sources, most notably chest X-ray (CXR) imaging, to support diagnostic assessment. The increasing availability of multimodal clinical data presents an opportunity to evaluate how different data types contribute to predictive performance in automated pneumonia detection systems.

1.2 Problem Definition

This study investigates whether pneumonia can be accurately predicted using different modalities derived from the Symile MIMIC dataset. Specifically, the research examines three predictive settings: (1) models trained exclusively on structured EHR data, including laboratory measurements and demographic features; (2) models trained solely on unstructured CXR images; and (3) multimodal models that integrate both structured and image data. The core problem addressed is whether combining heterogeneous data sources yields superior diagnostic performance relative to unimodal approaches.

1.3 Project Objectives

The project pursues the following objectives:

1. Develop machine learning models using structured EHR data, comprising laboratory values and demographic information relevant to pneumonia risk.
2. Develop deep learning models using chest X-ray images to classify pneumonia from imaging alone.
3. Construct multimodal fusion models that jointly leverage structured and imaging data.
4. Compare the predictive performance of all model types to determine whether multimodal integration provides measurable improvements over unimodal methods.

1.4 Success Criteria

Project success is defined using standard diagnostic performance metrics, including Area Under the Receiver Operating Characteristic Curve (AUROC), F1-score, and precision/recall. A model will be considered successful if it achieves high discriminatory ability and clinically meaningful sensitivity and specificity. Ultimately, the multimodal fusion model is expected to outperform both structured-only and image-only models, demonstrating the added value of integrating complementary data modalities.

2. Data Understanding

2.1 Data Source

The data used in this study originate from Symile-MIMIC, a multimodal clinical dataset hosted on PhysioNet. Symile-MIMIC is constructed from the MIMIC-IV and MIMIC-CXR repositories and integrates structured EHR measurements with paired chest radiographs. All data are fully de-identified and made available only to credentialed researchers under PhysioNet’s data use agreement, ensuring compliance with ethical and regulatory standards for secondary use of clinical records. The dataset provides 11,622 hospital admissions with synchronized multimodal data components, partitioned into mutually exclusive training, validation, and test subsets.

2.2 Structured Data

The structured modality consists of demographic information and laboratory test results recorded within 24 hours of admission. Demographic variables include age, sex, and other admission-level descriptors that serve as baseline predictors for clinical risk estimation. The laboratory component comprises up to 50 routinely obtained blood tests, including hematologic indices (e.g., hemoglobin, hematocrit, white blood cell count), metabolic markers (e.g., sodium, potassium, creatinine, glucose), and additional clinically relevant analytes such as albumin and lactate. These values are standardized through percentile transformation relative to the training distribution, and each laboratory feature is paired with a binary indicator specifying whether the measurement was recorded.

Missingness is an inherent characteristic of the structured data: some admissions lack one or more laboratory results, and a dataset-level flag denotes cases with no laboratory tests available in the defined time window. The explicit encoding of missingness supports downstream modeling by allowing algorithms to differentiate between absent and physiologically normal values.

2.3 Unstructured Data

The unstructured component consists of frontal chest X-ray (CXR) images sourced from the MIMIC-CXR-JPG collection and aligned at the admission level within Symile-MIMIC. The dataset includes standard frontal projections, anteroposterior (AP) and posteroanterior (PA), which represent the typical imaging protocols for respiratory assessment in hospital settings.

Images undergo standardized preprocessing: the shorter image dimension is scaled to a fixed length, followed by square cropping (randomized for training and centered for evaluation). Pixel intensities are normalized using ImageNet-derived statistics to facilitate compatibility with convolutional neural networks initialized with pretrained weights. The dataset documentation notes the availability of pneumonia annotations but does not provide explicit label prevalence statistics; thus, the degree of class imbalance is acknowledged but not quantified here.

2.4 Target Variable

The target outcome for all predictive tasks is pneumonia presence or absence, defined directly from the CheXpert-generated labels included within the Symile-MIMIC dataset. CheXpert is an automated labeling system that extracts radiologic findings from free-text radiology reports using rule-based and machine-learning-based heuristics. In Symile-MIMIC, the pneumonia indicator derived from CheXpert serves as the authoritative supervision signal for both structured and unstructured modeling tasks. This ensures consistent labeling across modalities and enables direct comparison of unimodal and multimodal predictive frameworks.

2.5 Exploratory Data Analysis

Exploratory analyses were performed to characterise the target distribution, demographic profile, laboratory features, and comorbid radiographic conditions in the cohort. Pneumonia label distribution.

The initial distribution of pneumonia labels derived from CheXpert annotations showed substantial incompleteness and uncertainty. In the full dataset, the majority of entries were either missing (NaN) or marked uncertain (-1), with only a minority assigned definitive labels of 0.0 (no pneumonia) or 1.0 (pneumonia). Specifically, the unfiltered counts were: 15,717 missing, 2,723 non-pneumonia, 2,461 pneumonia, and 1,989 uncertain. This highlights the necessity of excluding ambiguous and missing labels before model development. (Fig 1)

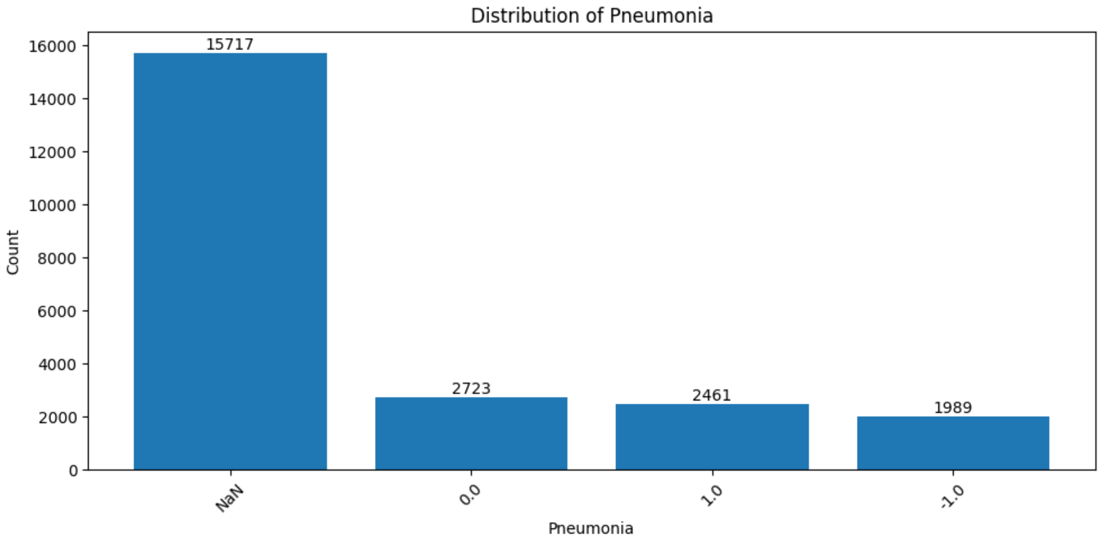


Fig 1. Distribution of pneumonia in the original Dataset

After filtering to retain only definitive labels (0 or 1) and deduplicating admissions to ensure unique patient encounters, the pneumonia cohort became substantially smaller but cleaner. The resulting analytical subset contained 1,316 non-pneumonia cases (53.1%) and 1,162 pneumonia cases (46.9%), producing a nearly balanced binary classification task. This deduplicated distribution reflects the ground-truth label set used for all supervised learning experiments in this study. (Fig 2)

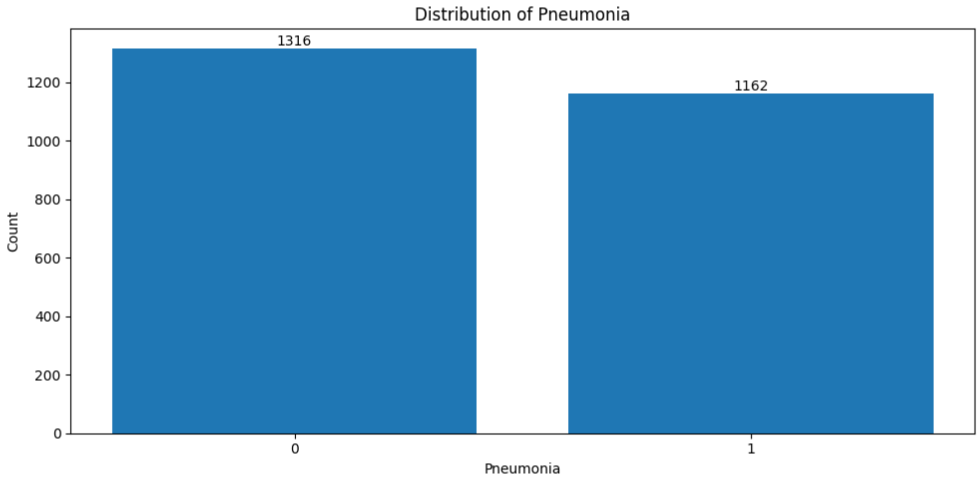


Fig 2. Distribution of pneumonia after filtering and deduplication

Demographic characteristics:

The age distribution, assessed using a kernel density estimate, showed a unimodal pattern with most patients in older adulthood; the density peaked in the late-middle to elderly age range, consistent with the known epidemiology of pneumonia in hospitalised populations.

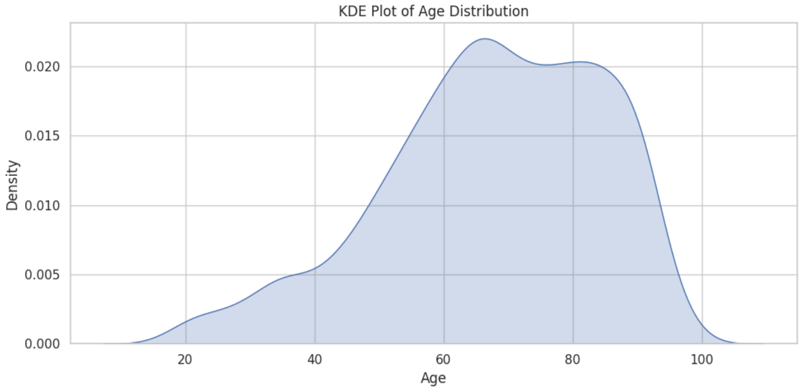


Fig 3.Distribution of Age

Gender distribution was approximately balanced with a slight male predominance (1,331 males vs. 1,147 females).

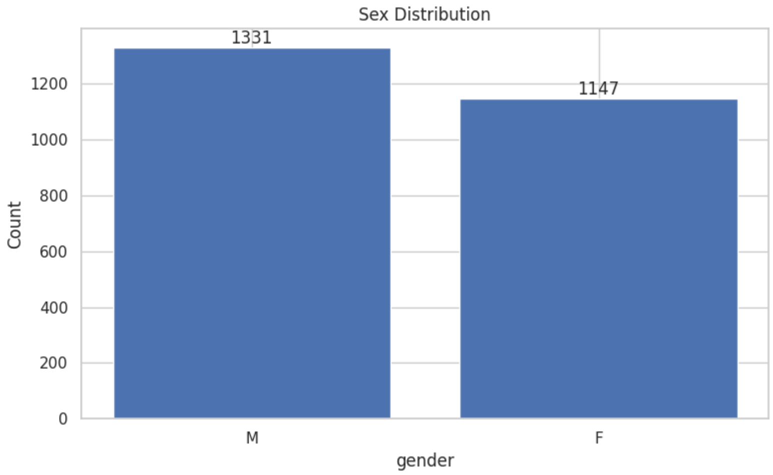


Fig 4. Distribution of Gender

Race was highly skewed: White patients constituted the majority, followed by Black/African American, with smaller counts in “Unknown,” “Other,” and a long tail of less frequent categories (e.g., specific European, Hispanic/Latino, and Asian subgroups). This skew should be considered when interpreting model performance across demographic strata.

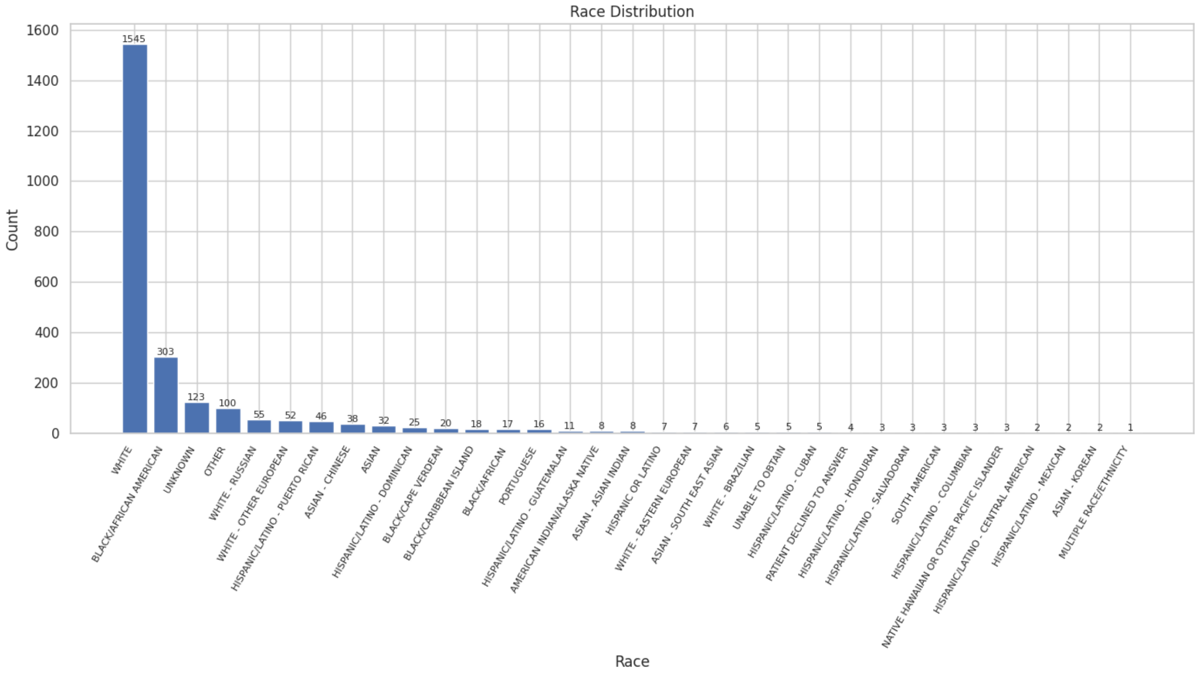


Fig 5. Distribution of Race

Laboratory variables and missingness:

The structured laboratory panel comprised 51 routinely collected tests, including: hematologic indices (e.g., Hematocrit, Hemoglobin, Platelet Count, Red and White Blood Cells, MCV, MCH, MCHC, RDW, RDW-SD), renal and electrolyte markers (Creatinine, Urea

Nitrogen, Sodium, Potassium, Chloride, Bicarbonate, Anion Gap, Magnesium, Calcium, Phosphate), liver and muscle enzymes (Alanine Aminotransferase, Aspartate Aminotransferase, Alkaline Phosphatase, Bilirubin, Total, Creatine Kinase), coagulation studies (INR(PT), PT, PTT), blood gas–related variables (pH, pO_2, pCO_2, Base Excess, Calculated Total CO_2), and differential counts (Neutrophils, Lymphocytes, Monocytes, Eosinophils, Basophils and their absolute counts, plus Immature Granulocytes), as well as flag variables (H, L, I).

Missingness was substantial and heterogeneous across tests. The flag variables H, L, and I had the highest missingness (about 90%), and several differential absolute counts and Immature Granulocytes were missing in roughly 70–76% of encounters. Important chemistry and gas parameters such as Creatine Kinase, pO_2, pCO_2, Calculated Total CO_2, Base Excess, and pH showed missingness around 60–67%. More routinely measured analytes, including Albumin, Alkaline Phosphatase, Bilirubin, and ALT, were somewhat better represented but still missing in approximately 40–48% of cases. These patterns underscore the need for models and preprocessing strategies that are robust to high and variable rates of missing laboratory data.

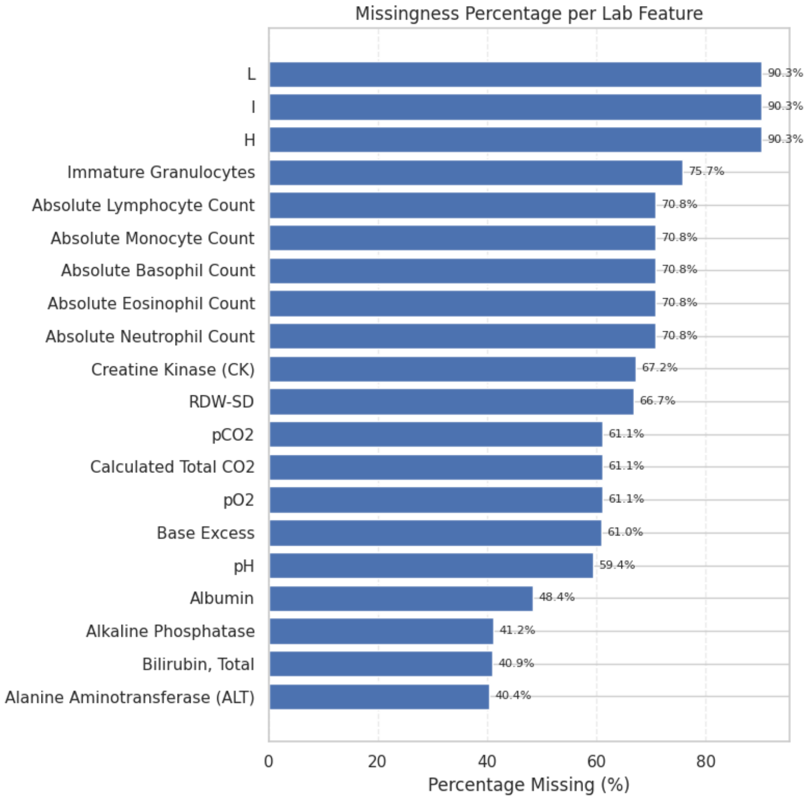


Fig 6. Missingness in Lab values

Co-occurring radiographic conditions:

To examine label complexity, an indicator of “any other CheXpert condition” was cross-tabulated with pneumonia status. Among non-pneumonia admissions, 367 (27.9%) had no other condition label while 949 (72.1%) had at least one additional finding. Among pneumonia admissions, only 142 (12.2%) had no other condition, whereas 1,020 (87.8%) had at least one co-occurring label. Thus, truly “clean” controls, patients with neither pneumonia nor any other CheXpert-positive finding, were relatively rare (367 encounters), and most pneumonia cases appeared in the context of additional radiographic abnormalities. This comorbidity structure is expected to increase the clinical realism of the task but may also complicate discrimination between pneumonia and other thoracic pathologies.

Table 1. Condition Co-Occurrence by Pneumonia Status

Pneumonia/ Has other Condition	FALSE	TRUE
0	367	949
1	142	1020

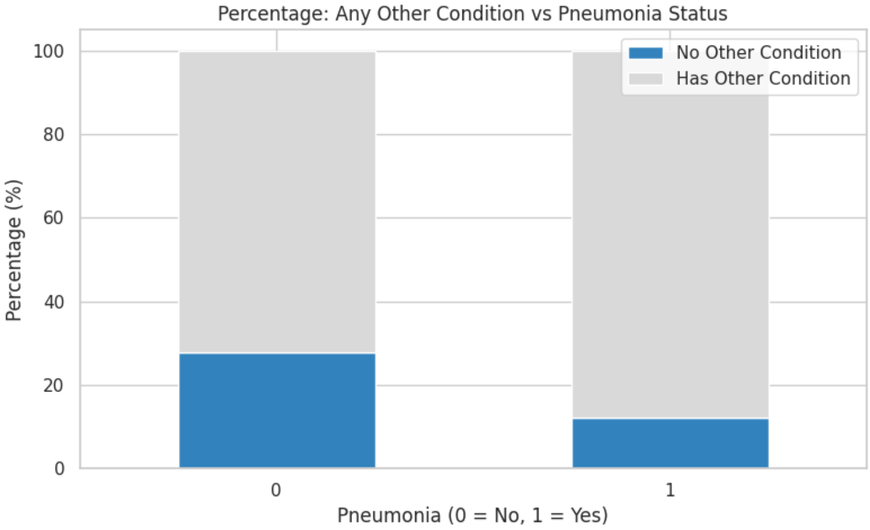


Fig 7. Other Condition vs Pneumonia Status in percentage

3. Data Preparation

This stage focused on constructing reproducible preprocessing pipelines for both structured and unstructured modalities and assembling the final datasets used for model training, validation, and evaluation. All transformations were applied systematically to ensure that no information from the validation or test sets leaked into model fitting.

3.1 Structured Data Preprocessing

Preprocessing of the structured EHR subset involved standardising demographic variables, encoding categorical attributes, and preparing laboratory measurements for downstream modelling. Demographic variables were harmonised by mapping the numerous race categories into five consolidated groups, White, Black, Hispanic, Asian, and Other, to ensure statistical tractability while preserving clinically relevant distinctions. Age was discretised into four ordinal strata (18–40, 40–60, 60–80, and 80+) to capture broad demographic patterns and facilitate stratified analyses.

Laboratory values underwent a multi-step transformation pipeline. First, features with excessive sparsity, defined as more than 90% missingness, were removed to reduce noise and improve model stability. For the remaining laboratory tests, explicit missingness indicators were introduced to allow models to differentiate between absent and observed measurements. After splitting the dataset into training, validation, and test partitions, raw laboratory values were converted into percentile-based scores computed exclusively from the training distribution. This ensured that normalization reflected only information available during model development. Categorical variables, including demographic groupings, were one-hot encoded to produce model-compatible binary representations.

3.2 Image Data Preprocessing

Chest X-ray preprocessing followed a consistent procedure to standardise image inputs for deep learning models. Each radiograph was linked to its corresponding structured record, ensuring alignment across modalities prior to modelling. Images were resized and cropped to maintain aspect ratio and spatial consistency, and intensity values were normalised using fixed channel statistics to support stable optimisation. These steps produced a uniform tensor representation suitable for convolutional architectures and compatible with multimodal fusion pipelines.

3.3 Final Datasets

After modality-specific preprocessing, the structured features, chest X-ray images, and demographic attributes were merged into a unified multimodal dataset keyed by patient encounter. The integrated dataset was then partitioned into training (70%), validation (15%), and test (15%) subsets, ensuring mutually exclusive patient-level splits and preserving class balance across partitions.

The final dataset contained cleaned structured variables (with missingness indicators and percentile-normalised laboratory measurements), harmonised demographic categories, and preprocessed radiographs ready for ingestion by machine learning and deep learning models. This comprehensive preparation enabled reliable comparison between structured-only, image-only, and multimodal predictive approaches in subsequent modelling stages.

4. Structured Modelling

Structured models were developed using demographic variables, admission location, and percentile-normalized laboratory values as predictors of the binary pneumonia label. Because the clinical use case prioritises identifying true pneumonia cases, particular attention was paid to recall, in addition to AUROC and F1-score.

4.1 Model 1: Gradient-Boosted Trees (XGBoost)

Features and Preprocessing

The first structured model used all available laboratory features expressed as percentile scores, together with categorical demographics:

- Laboratory predictors: percentile-transformed versions of Hematocrit, Platelet Count, Creatinine, Potassium, Hemoglobin, White Blood Cells, MCHC, Red Blood Cells, MCV, MCH, RDW, Urea Nitrogen, Sodium, Chloride, Bicarbonate, Anion Gap, Glucose, Magnesium, Calcium (total), Phosphate, INR(PT), PT, PTT, differential counts (Basophils, Neutrophils, Monocytes, Eosinophils, Lymphocytes), RDW-SD, liver and muscle enzymes (ALT, AST, Alkaline Phosphatase, Creatine Kinase), Bilirubin (total), blood gas–related measures (pH, pO_2, pCO_2, Base Excess, Calculated Total CO_2), albumin, and absolute differential counts, plus Immature Granulocytes. Sparse lab flags H, L, and I were excluded due to high missingness.
- Categorical predictors: admission location, grouped race category, and age group.

Categorical variables were imputed with a constant “Unknown” category and one-hot encoded. Laboratory features were passed through as continuous predictors. The target label (Pneumonia) and the image identifier were excluded from the feature matrix.

The dataset was split into training (70%), validation (15%), and test (15%) subsets using stratified sampling to preserve the pneumonia prevalence.

Model and Hyperparameter Tuning

An XGBoost classifier with a logistic objective was trained within a scikit-learn pipeline that encapsulated preprocessing and modelling. Hyperparameters were tuned using

RandomizedSearchCV with a 5-fold stratified cross-validation scheme on the training set, optimising the ROC-AUC score. The search explored:

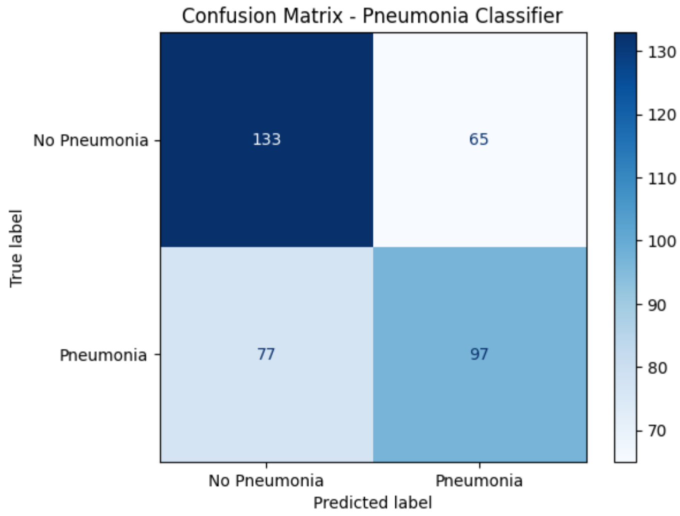
- Number of trees, learning rate, maximum depth, minimum child weight
- Subsample and column-subsampling rates
- Tree regularisation parameters (gamma, L1 reg_alpha, L2 reg_lambda)
- Class imbalance handling via scale_pos_weight (1 vs. ratio of negatives to positives)

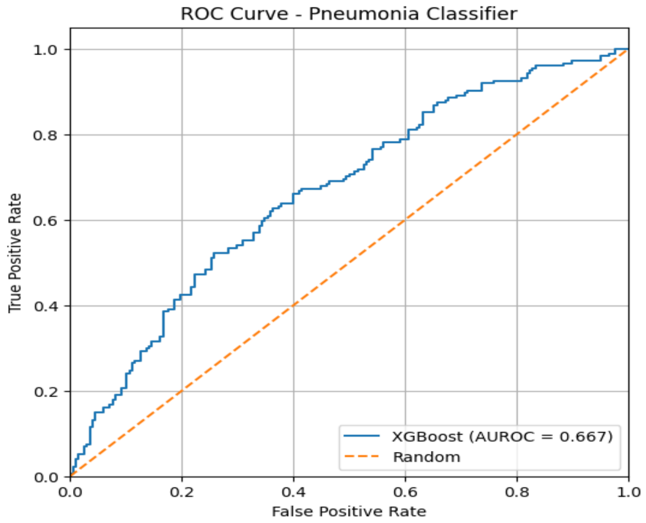
The best configuration achieved a cross-validated AUROC of 0.628 and selected: 200 trees, learning rate 0.01, max depth 6, min child weight 3, full subsampling (subsample 1.0), column subsampling 0.6, no additional gamma penalty, L1 regularisation of 1, and no L2 regularisation, with scale_pos_weight = 1.

Test Performance

On the held-out test set, the tuned XGBoost model achieved:

- ROC-AUC: 0.667
- Accuracy: 0.618
- Precision: 0.599
- Recall: 0.557
- F1-score: 0.577





The confusion matrix (threshold 0.5) showed 97 true positives and 77 false negatives among 174 pneumonia cases, indicating moderate sensitivity. The ROC curve demonstrates consistent improvement over the random baseline, but there remains room to improve recall if a higher-sensitivity operating point were chosen.

4.2 Model 2: Embedded Multilayer Perceptron (MLP)

Features and Representation

The second structured model used the same underlying clinical information but represented laboratory and categorical variables differently:

- Laboratory percentile features were transformed into ordinal bins (≤10th percentile, 10–25th, 25–75th, 75–90th, >90th, and an implicit bin for missing values). Sparse lab flags (H, L, I) were again dropped.
- Categorical variables (admission location, race group, age group) were integer-encoded with an explicit UNKNOWN category for unseen or missing levels.

A custom PyTorch dataset assembled per-encounter vectors of lab bins and categorical codes for use in the neural network.

Model Architecture and Hyperparameter Tuning

The PneumoniaMLP model embedded each binned lab variable and each categorical feature into low-dimensional vectors and concatenated them before passing them through a feed-forward network:

- One embedding per lab feature (6 bins) with a learnable lab embedding dimension.
- One embedding per categorical variable with a learnable categorical embedding dimension.
- A stack of fully connected layers with batch normalisation, ReLU activations, and dropout, followed by a single sigmoid output neuron.

The model was trained with BCEWithLogitsLoss and the AdamW optimiser, using early stopping based on validation AUROC.

Hyperparameters were tuned with Optuna over 20 trials, maximising validation ROC-AUC. The search varied learning rate, batch size, lab and categorical embedding dimensions, hidden size, number of layers, dropout, and weight decay. The best trial achieved a validation AUROC of 0.608, with the following configuration:

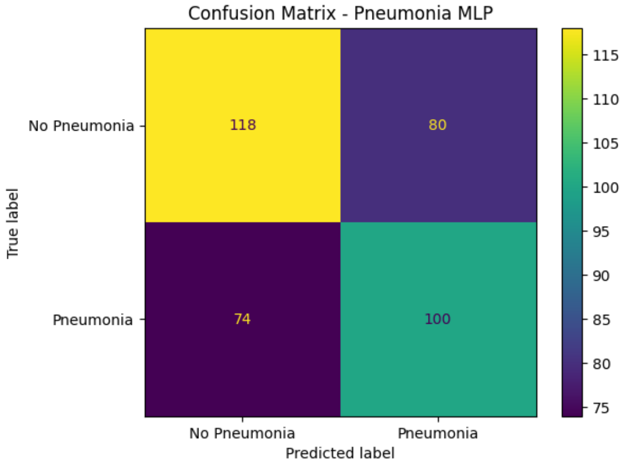
- Learning rate = 0.002
- Batch size 32
- Lab embedding dimension 8
- Categorical embedding dimension 16
- Hidden layer width 128, 3 layers
- Dropout 0.5
- No weight decay

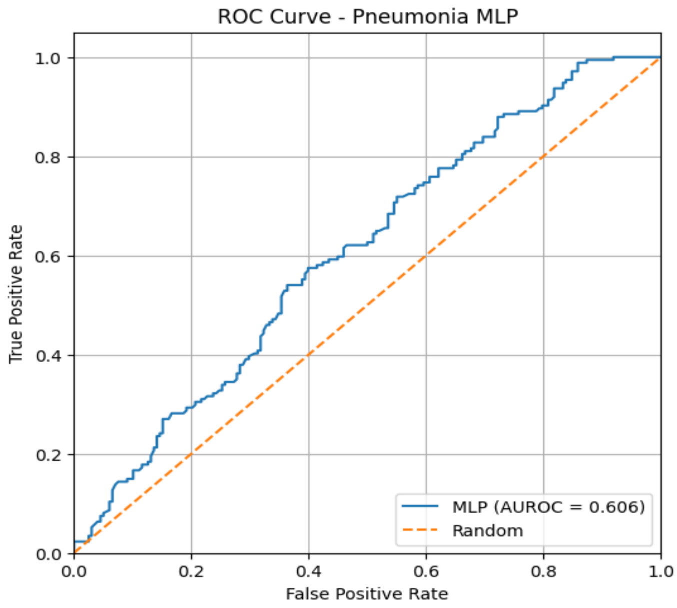
A final model was trained on the full training set using these hyperparameters and evaluated on the held-out test set.

Test Performance

On the test set, the MLP achieved:

- ROC-AUC: 0.606
- Accuracy: 0.586
- Precision: 0.556
- Recall: 0.575
- F1-score: 0.565





The model's sensitivity (recall 0.575) was slightly higher than that of XGBoost at the default threshold, although at the cost of lower AUROC and similar overall accuracy. The ROC curve for the MLP indicates modest discriminative ability above chance but below that of the tree-based model.

4.3 Comparison and Role of Recall

Both structured models demonstrated moderate discriminative performance, with AUROC in the range of 0.61–0.67 and F1-scores around 0.57. From a recall-focused perspective, the MLP achieved marginally higher sensitivity (0.575 vs. 0.557), whereas XGBoost delivered superior overall ranking performance (AUROC 0.667 vs. 0.606). These results suggest that structured EHR data alone provide some signal for pneumonia detection, but their ability to reliably capture all true pneumonia cases is limited, motivating the subsequent exploration of image-only and multimodal fusion models.

5. Unstructured Modelling (Chest X-ray Only)

Unstructured models were trained using only the chest X-ray images and their associated view position, with the binary CheXpert pneumonia label as the target. As in the structured setting, the clinical priority is to maximise recall for pneumonia, while still monitoring AUROC and F1-score.

All image models share the same data handling pipeline. The original dataframe contains an image field (already normalised tensors or NumPy arrays) and a ViewPosition field (e.g. *PA*, *AP*). Missing view positions are mapped to "Unknown", and ViewPosition is encoded as an integer view_position_id. The dataset is split into train (70%), validation (15%), and test (15%) subsets using stratified sampling on the pneumonia label.

A custom XrayImageDataset converts each image into a FloatTensor of shape [3, 320, 320], handling grayscale or differently shaped inputs by channel replication and bilinear resizing when necessary. During training, tensor-level augmentation applies random horizontal flips and random rotations by multiples of 90 degrees; validation and test images are passed without augmentation.

5.1 Model 1: Custom CNN with View-Position Embedding

Image Representation and Inputs

Each training example consists of:

- A preprocessed image tensor [3×320×320]
- A binary pneumonia label (0 = normal, 1 = pneumonia)
- An integer view_position_id indicating the CXR projection

The view position is treated as a categorical covariate and modelled through a learnable embedding.

Network Architecture

The CustomCNN is a convolutional network designed specifically for this task:

1. Convolutional feature extractor
    - The network comprises num_blocks (2–4) repeated blocks.
    - Block *b* takes curr_in channels and outputs out_ch = base_filters × 2^(b) feature maps.
    - Each block applies:
        - Conv2d(curr_in, out_ch, kernel_size=3, padding=1)
        - BatchNorm2d(out_ch)
        - ReLU activation
        - Optional Dropout2d with probability dropout_conv
        - MaxPool2d(2,2) to halve the spatial resolution
    - Across blocks, the number of channels increases geometrically while spatial resolution is progressively downsampled, yielding a high-level representation of the image.
2. Global pooling
    - After the final block, an AdaptiveAvgPool2d(1) compresses each feature map to a single value, producing a [B, C] representation (where *C* is the final number of channels).

    - This provides a translation-invariant summary of the learned features.

3. View-position embedding
    - An nn.Embedding layer maps each view_position_id to a low-dimensional vector of size 8.
    - The pooled image feature vector and the view embedding are concatenated, explicitly conditioning the classifier on projection type (e.g. AP vs PA).
4. Fully connected head
    - The concatenated vector passes through a small MLP:
        - Either a single Linear(fc_in_dim, 1) layer, or
        - Linear(fc_in_dim, fc_hidden_dim) -> ReLU -> Dropout(dropout_fc) -> Linear(fc_hidden_dim, 1)
    - The final output is a single logit per image; a sigmoid is applied at inference to obtain pneumonia probabilities.

Overall, this architecture combines local convolutional feature extraction with global summarisation and an explicit embedding of the view position.

Training and Hyperparameter Optimisation

The CustomCNN is trained using BCEWithLogitsLoss and the Adam optimiser on GPU when available. Hyperparameters are tuned with Optuna, maximising validation ROC-AUC. The search space includes:

- Learning rate in [10−5,10−3] (log scale)
- Batch size {8, 16, 32}
- Number of convolutional blocks (2–4) and base filter width {16, 32, 64}
- Convolutional dropout ∈ {0.0, 0.2, 0.5} and fully connected dropout {0.0, 0.3, 0.5}
- Hidden dimension of the fully connected head {0, 64, 128, 256}
- Weight decay in [10−6,10−3]

For each trial, the model is trained for up to 15 epochs with early stopping triggered after three epochs without improvement in validation AUROC. The best-performing configuration is then retrained on the combined train+validation set (with augmentation) and evaluated on the held-out test set.

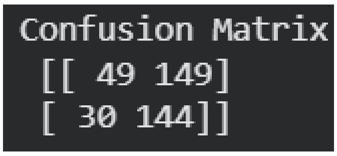
Test Performance

On the test set, the final CustomCNN yields:

- ROC-AUC: 0.566
- Accuracy: 0.519
- Precision: 0.492

- Recall: 0.828
- F1-score: 0.617

The confusion matrix



```
Confusion Matrix
 [[ 49 149]
 [ 30 144]]
```

It shows that the model correctly identifies the majority of pneumonia cases (high recall) but misclassifies many normal radiographs as pneumonia, leading to low specificity and moderate precision. From a recall-oriented clinical perspective, this behaviour is acceptable for screening but would impose a substantial burden of false positives.

5.2 Model 2: ResNet-18 with View-Position Embedding

Image Representation and Inputs

The second unstructured model uses the same preprocessed images, labels, and view-position IDs as Model 1, including the same augmentation strategy during training.

Network Architecture

The ResNet18Custom model replaces the handcrafted convolutional stack with a standard ResNet-18 backbone:

1. ResNet-18 backbone
   - A torchvision resnet18 model is instantiated without pretrained weights (weights=None), so all parameters are learned from the pneumonia task.
   - The default fully connected classification layer is replaced by nn.Identity, exposing a 512-dimensional feature vector from the global average pooling layer for each image.
2. View-position embedding
   - As in Model 1, an nn.Embedding layer maps each view_position_id to an 8-dimensional vector.
   - This embedding is concatenated with the 512-dimensional ResNet feature vector, allowing the network to modulate its predictions based on view type.
3. Fully connected head
   - The concatenated vector of size 512+8 feeds into:
     - Either a single Linear(fc_in_dim, 1) layer, or
     - Linear(fc_in_dim, fc_hidden_dim) -> ReLU -> Dropout(dropout_fc) -> Linear(fc_hidden_dim, 1)

    - As before, the output is a single logit and a sigmoid is used to obtain probabilities.

Thus, this model combines a deep residual network trained end-to-end with an auxiliary view-position embedding, offering greater representational capacity than the custom CNN.

Training and Hyperparameter Optimisation

Training again uses BCEWithLogitsLoss and the Adam optimiser, with early stopping on validation ROC-AUC. Optuna is used to tune:

- Learning rate in [10-5,10-3]
- Batch size {8, 16, 32}
- Fully connected dropout {0.0, 0.3, 0.5}
- Hidden dimension of the head {0, 64, 128, 256}
- Weight decay in [10-6,10-3]

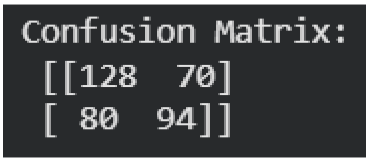
Each trial trains for up to 15 epochs with early stopping patience of three epochs. The best trial's configuration is used to train a final model on the combined train+validation set, which is then evaluated on the test set.

Test Performance

The final ResNet-18–based model achieves:

- ROC-AUC: 0.636
- Accuracy: 0.597
- Precision: 0.573
- Recall: 0.540
- F1-score: 0.556

The confusion matrix



```
Confusion Matrix:
 [[128  70]
 [ 80  94]]
```

It indicates a more balanced trade-off between sensitivity and specificity than the CustomCNN. Compared with Model 1, this model reduces false positives substantially and improves overall discrimination (higher AUROC), but at the cost of a lower pneumonia recall.

5.3 Comparison and Role of Recall

Both unstructured models exploit only the CXR images and view position, yet they exhibit different operating characteristics:

- The CustomCNN attains very high recall (0.83) but relatively low AUROC (0.57) and accuracy, functioning as an aggressive screen that flags most pneumonia cases along with many false positives.
- The ResNet-18 model provides better overall discrimination (AUROC 0.64) and accuracy (~0.60), with a more balanced precision–recall profile, but misses more pneumonia cases (recall 0.54).

Given the project's emphasis on recall, the CustomCNN demonstrates that image-only models can be tuned toward high sensitivity, whereas the ResNet-18 architecture offers superior ranking and calibration. These unstructured results set a benchmark for subsequent multimodal fusion models, which aim to leverage both imaging and structured EHR data to improve recall without sacrificing overall discriminative performance.

6. Multimodal Modelling

Multimodal learning aims to combine structured EHR features with unstructured chest X-ray data in a unified predictive framework. The hypothesis is that laboratory findings and demographic context provide complementary information to imaging, enabling models to detect pneumonia more accurately than either modality alone. In this study, a two-stage multimodal design was implemented: (i) a gradient-boosted decision tree model trained exclusively on structured data, and (ii) a convolutional neural network trained to integrate image features with the structured model's output. This approach treats the structured model's prediction as a high-level clinical signal that the CNN can refine using visual evidence.

6.1 Structured Component: XGBoost for Tabular Representation

The structured branch of the multimodal pipeline is an XGBoost classifier trained on binned laboratory values, categorical demographic variables, admission location, and a scaled age feature. Laboratory percentiles were discretised into ordered bins, sparse lab indicators were dropped, and all categorical attributes (e.g., race, age group, admission location, and radiographic view position) were encoded through integer vocabularies derived from the training set.

To address label imbalance, XGBoost's scale_pos_weight parameter was tuned according to the ratio of negative to positive cases. The model was trained with 300 trees, depth 5, learning rate 0.05, and subsampling/column sampling rates of 0.8. After training, XGBoost produced a probability score for pneumonia for every encounter, which served two purposes:

1. Standalone performance evaluation, representing the structured-only baseline.

2. Use as an input feature for the multimodal fusion model, where it acts as a distilled clinical representation summarising the tabular data.

These probability outputs (train/validation/test) were saved and later injected into the CNN fusion architecture.

6.2 Unstructured Component: CNN-based Image Encoder

The multimodal model incorporates a custom convolutional neural network to extract features from chest radiographs. The network includes:

- Four convolutional blocks with increasing channel depths (32 - 64 - 128 - 256),
- ReLU activations and max-pooling for downsampling large spatial resolutions,
- A flattening layer followed by a fully connected layer that embeds the resulting feature map into a 128-dimensional image representation.

This CNN is trained end-to-end jointly with the fusion head, enabling it to adapt image features specifically to the multimodal decision task. Unlike the structured model, which captures laboratory and demographic signals, the CNN learns visual biomarkers of pneumonia such as consolidations, opacities, and asymmetrical lung patterns.

6.3 Fusion Mechanism: Integrating Image and Tabular Signals

The multimodal fusion architecture (CNNXGBFusion) merges the two data streams as follows:

1. The CNN produces a 128-dimensional image feature vector

$$z_{\text{img}} \in \mathbb{R}^{128}$$

2. The XGBoost model provides a single probability score

$$s_{\text{xgb}} \in [0, 1]$$

3. These two representations are concatenated:

$$h = [z_{\text{img}} \parallel s_{\text{xgb}}]$$

4. The fused vector passes through a two-layer MLP:
   - ○ Linear -> ReLU -> Dropout -> Linear -> Logit
   - ○ Producing the pneumonia prediction.

This design treats the XGBoost output as a compact summary of all structured information, thereby simplifying the fusion process while still preserving all tabular signals.

Importantly, the CNN learns conditional visual representations influenced by the structured prior: if XGBoost signals a high likelihood of pneumonia, the CNN is encouraged to look for confirming visual evidence; if XGBoost predicts low risk, the CNN must extract stronger or more specific visual cues to overturn that estimate. This leads to a cooperative modelling strategy in which tabular data guides, but does not override, image interpretation.

6.4 Training Strategy

The multimodal model is trained end-to-end using:

- BCEWithLogitsLoss with a positive-class weight proportional to the class imbalance,
- AdamW optimisation with learning rate 1e-4 and weight decay 1e-4,
- A batch size of 16 and early stopping on validation AUC.

During training:

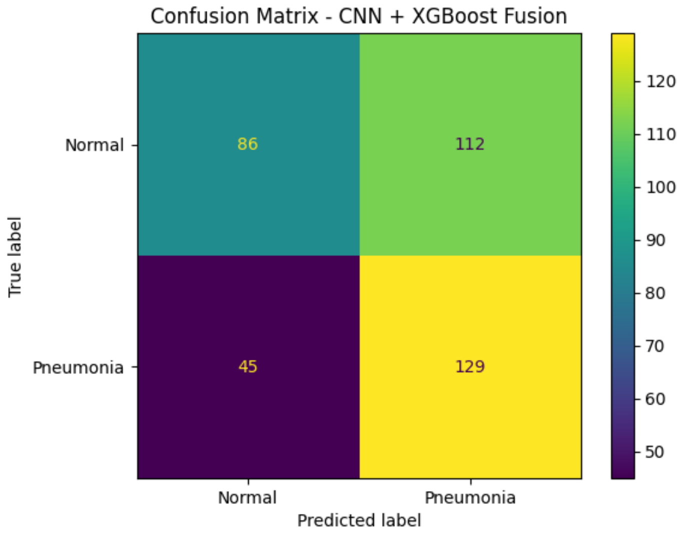
- The CNN receives augmented images (random flips and rotations).
- The XGBoost branch remains frozen; only its probability outputs participate in training.
- Fusion layers and the CNN backbone jointly learn a consistent multimodal representation.

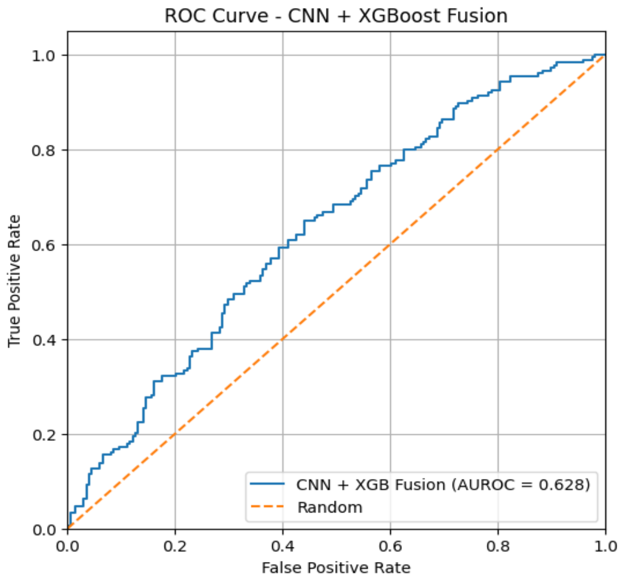
6.5 Test Performance of the Multimodal Model

The final CNN+XGB fusion model achieved the following metrics on the held-out test set:

- ROC-AUC: 0.628
- Accuracy: 0.578
- Precision: 0.535
- Recall: 0.741
- F1-score: 0.622

Confusion matrix:





The multimodal model captures substantially more pneumonia cases than XGBoost alone (recall 0.74 vs. 0.56) and significantly more than the ResNet-18 model (0.54). It also achieves a more favourable precision–recall balance than the custom CNN (0.49 precision, 0.83 recall). The result suggests that structured predictions enrich the image encoder with clinically grounded prior information, allowing the model to detect pneumonia more consistently than using either modality in isolation.

7. Is Multimodal Actually Required?

Single-Stage CNN with Full Clinical Data

To test whether an explicitly multimodal design (separate structured and image models with a fusion layer) is truly necessary, a single-stage deep model was trained that ingests all available information at once: chest X-ray images, binned laboratory values, demographic variables, view position, admission location, and age.

7.1 Model Construction and Inputs

The experiment reuses the same stratified 70/15/15 train–validation–test split on the pneumonia label. Pre processing mirrors the earlier multimodal pipeline:

- Laboratory percentile features are discretised into ordinal bins using fixed quantile edges; sparse flags (H, L, I) are removed.
- Categorical variables (admission_location, ViewPosition, race_grouped, age_group) are integer-encoded via training-set vocabularies with an explicit unkonwn category.
- These binned labs, categorical codes, and a scaled age feature form a tabular vector of length num_tab_features for each encounter.

The CXRFullDataset returns three components per example:

1. A preprocessed image tensor [3, H, W], with grayscale images expanded to three channels.
2. A tabular feature tensor containing all binned labs and encoded categorical variables as float32.
3. The binary pneumonia label.

7.2 Network Architecture: CNNFullFusion

The CNNFullFusion model extends the earlier image-only CNN by appending the full tabular vector directly at the fusion layer:

1. Image encoder
   - Four convolutional blocks with increasing channel counts (32 -> 64 -> 128 -> 256).
   - Each block applies Conv2d -> ReLU -> MaxPool2d, progressively downsampling the image and extracting higher-level visual features.
   - The final feature map is flattened and passed through a fully connected layer to produce a 128-dimensional image embedding zimg.
2. Tabular branch
   - The tabular input is not embedded or processed by a separate network; the full feature vector xtab (binned labs, categorical codes, age) is concatenated directly with the image embedding.
3. Fusion head
   - The fused representation [zimg ‖ xtab] is passed through:
     - Linear(d_img + num_tab_features, fusion_width) -> ReLU -> Dropout
     - Linear(fusion_width, 1)
   - The output is a single logit, which is converted to a probability with a sigmoid during evaluation.

It consumes both image and structured data, but does so in a single monolithic network, without a separate specialised structured model.

7.3 Training Procedure

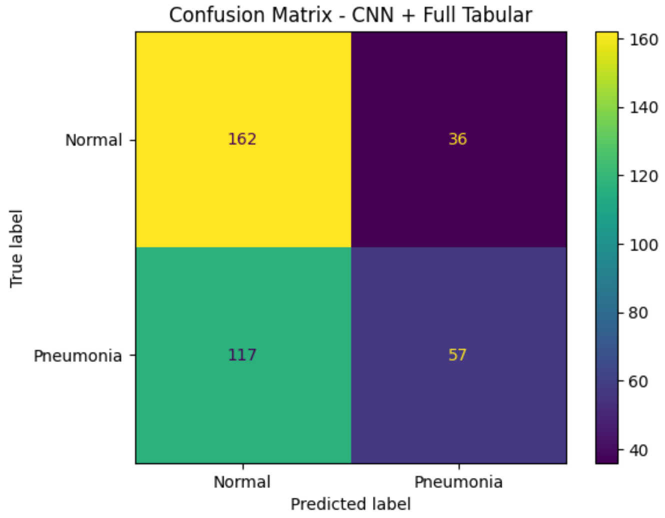
Training uses BCEWithLogitsLoss with a positive-class weight proportional to the ratio of negative to positive cases to mitigate class imbalance. Optimisation is performed with AdamW, mini-batch size 16, and early stopping based on validation ROC–AUC with a patience of five epochs. The same data augmentations as in earlier CNN experiments (random flips and rotations) are applied to the images during training.

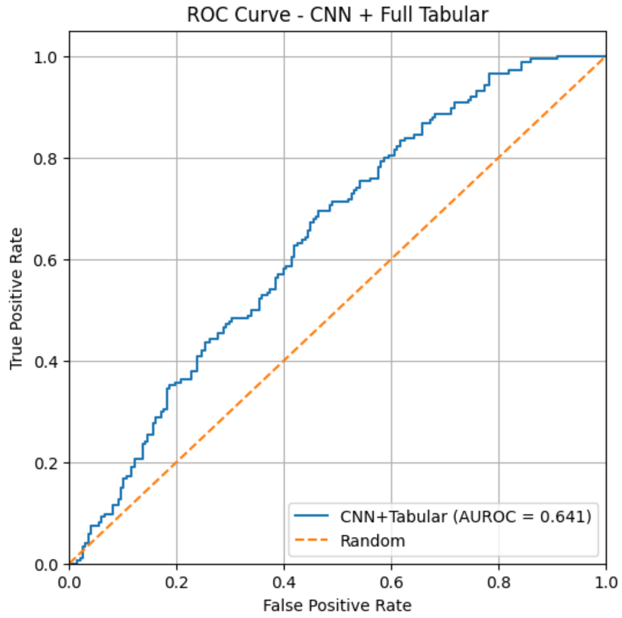
7.4 Test Performance

On the held-out test set, the CNN+full-tabular model achieves:

- ROC-AUC: 0.641
- Accuracy: 0.589
- Precision: 0.613
- Recall: 0.328
- F1-score: 0.427

The confusion matrix





It shows that the model is highly specific but poorly sensitive: it correctly classifies most normal cases but misses more than two-thirds of pneumonia cases. This behaviour contrasts sharply with the high-recall image-only CNN and the CNN+XGB fusion model.

7.5 Is Explicit Multimodality Necessary?

Comparing across all models:

- Structured-only (XGBoost / MLP): AUROC ≈ 0.61–0.67, recall ≈ 0.56–0.58.
- Image-only:
  - Custom CNN: very high recall (0.83) but low AUROC (0.57).
  - ResNet-18: better AUROC (0.64) but moderate recall (0.54).
- Two-stage multimodal (CNN + XGB score): AUROC 0.63, recall 0.74, F1 ≈ 0.62.
- Single-stage CNN + full tabular (this section): AUROC 0.64, recall 0.33, F1 ≈ 0.43.

Although the CNN + full-tabular model attains a competitive AUROC (slightly higher than the CNN+XGB fusion), it fails the clinical requirement of high recall, performing worse than both the structured-only and image-only models in terms of sensitivity. In contrast, the explicit multimodal fusion model (CNN+XGB) achieves substantially higher recall while maintaining a balanced precision–recall profile.

Interpretation: simply concatenating raw tabular features to a CNN embedding does not automatically yield a superior multimodal model. The two-stage design, where a dedicated structured model first produces a calibrated risk score that is then fused with image features, appears better suited to exploiting complementary information from EHR data and CXRs under a recall-focused objective.

In this sense, multimodality is indeed beneficial, but it must be implemented carefully. The results suggest that:

- A naive “all-in-one” CNN over images and tabular features is not sufficient to replace more principled multimodal fusion.
- Explicitly modelling the structured modality (e.g., with XGBoost) and then integrating its output into an image network provides clear advantages in recall and overall clinical usefulness.

8. Current Research on Multimodal Integration in Medical Data

In clinical practice and medical AI research, multimodal analysis, where diverse data types such as medical imaging, laboratory results, clinical notes, and EHR records are combined, has increasingly demonstrated real utility and impact. Unlike traditional models that operate on a single modality (e.g., imaging alone), multimodal AI mirrors how clinicians reason: by synthesising visual, numerical, and contextual information to inform diagnosis and prognosis. Such approaches have been applied across a range of clinical applications including disease diagnosis, risk prediction, and prognostic modelling. For example, systematic reviews have documented dozens of studies where models that fuse medical imaging with structured EHR data outperform unimodal counterparts in tasks spanning neurological disorders, cancer diagnosis, cardiovascular disease prediction, and COVID-19 detection, underscoring the practical value of multimodal strategies in healthcare analytics.

Moreover, curated surveys highlight that multimodal AI is not only a research trend but also a conceptual framework gaining traction in translational and clinical research communities. These studies show that fusion of heterogeneous modalities enhances diagnostic accuracy, robustness, and the ability to capture complementary signals across data sources, an essential attribute for complex conditions like pneumonia where imaging and clinical context jointly inform decision-making.

Representative publications include:

- *Artificial Intelligence-Based Methods for Fusion of Electronic Health Records and Imaging Data*, a review summarising multiple clinical tasks where multimodal fusion improved performance over single-modality models.
- *Multimodal Machine Learning in Image-Based and Clinical Biomedicine,* which outlines how fused modalities are being integrated into clinical decision support systems.
- Narrative and scoping reviews that document application of multimodal AI across diverse disease domains and provide implementation guidelines for combining imaging with structured data in clinical models.

These examples signal that multimodal models are already being actively developed and evaluated in the biomedical domain, reinforcing their relevance to your pneumonia prediction problem.

Conclusion

In this study, we investigated the relative effectiveness of structured EHR data, unstructured chest X-ray imaging, and multimodal fusion techniques for automated pneumonia detection using the Symile-MIMIC dataset. Structured machine learning models demonstrated moderate discriminatory ability, while deep learning models trained on imaging alone achieved high sensitivity but with substantial false positives. Importantly, multimodal fusion, particularly a two-stage architecture combining tabular risk scores with image representations, yielded a balanced performance profile with improved recall relative to most unimodal baselines.

Despite the complexity of clinical data and the inherent noise in automated labelling, our systematic comparison reveals that explicit multimodal integration can harness complementary information from diverse data sources more effectively than naïve single-stage architectures. This underscores the value of principled multimodal design in clinical predictive modelling, especially when recall is a critical objective. Future work aimed at improving label fidelity, representation learning, and external validation will further enhance the clinical applicability of these models.

Reference

[1] Symile-MIMIC: a multimodal clinical dataset of chest X-rays, electrocardiograms, and blood labs from MIMIC-IV. https://physionet.org/content/symile-mimic/1.0.0/

[2] Artificial Intelligence-Based Methods for Fusion of Electronic Health Records and Imaging Data.https://arxiv.org/abs/2210.13462?utm_source=chatgpt.com

[3] Multimodal Machine Learning in Image-Based and Clinical Biomedicine: Survey and Prospects. https://link.springer.com/article/10.1007/s11263-024-02032-8?utm_source=chatgpt.com

[4] Artificial intelligence-based methods for fusion of electronic health records and imaging data. https://www.nature.com/articles/s41598-022-22514-4.pdf?utm_source=chatgpt.com