UMBC

DATA 606 Capstone in Data Science

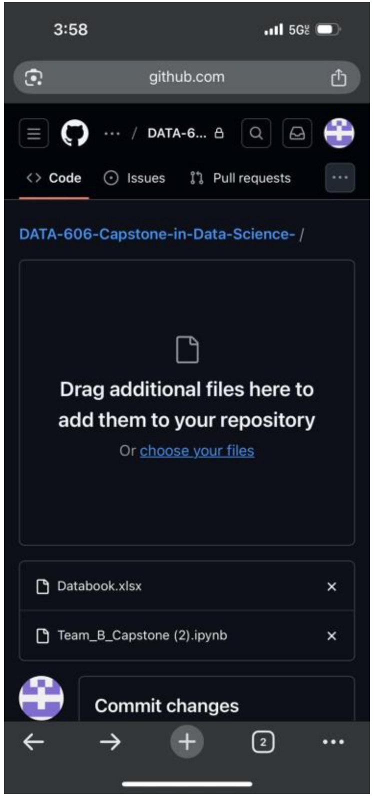
# Multimodal Patient Prognosis System

## Phase 2 – Exploratory Data Analysis & Model Construction

**Kaggle ID : capstoneprojectumbc**

**GitHub : https://github.com/shranya12/DATA-606-Capstone-in-Data-Science-**

**Team B:**
Tejaswini Veeramachaneni
Shranya Gandham
Dikshitha Tanneru


3:58
5G
github.com
DATA-6...
Code
Issues
Pull requests
DATA-606-Capstone-in-Data-Science- /
Drag additional files here to add them to your repository
Or choose your files
Databook.xlsx
Team_B_Capstone (2).ipynb
Commit changes
2

# Project Overview

**Problem Type : Multi-label Classification**
**Dataset Link  :** https://www.kaggle.com/datasets/nih-chest-xrays/data
**Size of subset: 3445 images(**Train 2 756 | Val 689**)**

## Features / Attributes:

- X-ray image (pixel arrays)
- Age, Gender, View Position (EHR)
- Text notes(csv file) / disease label

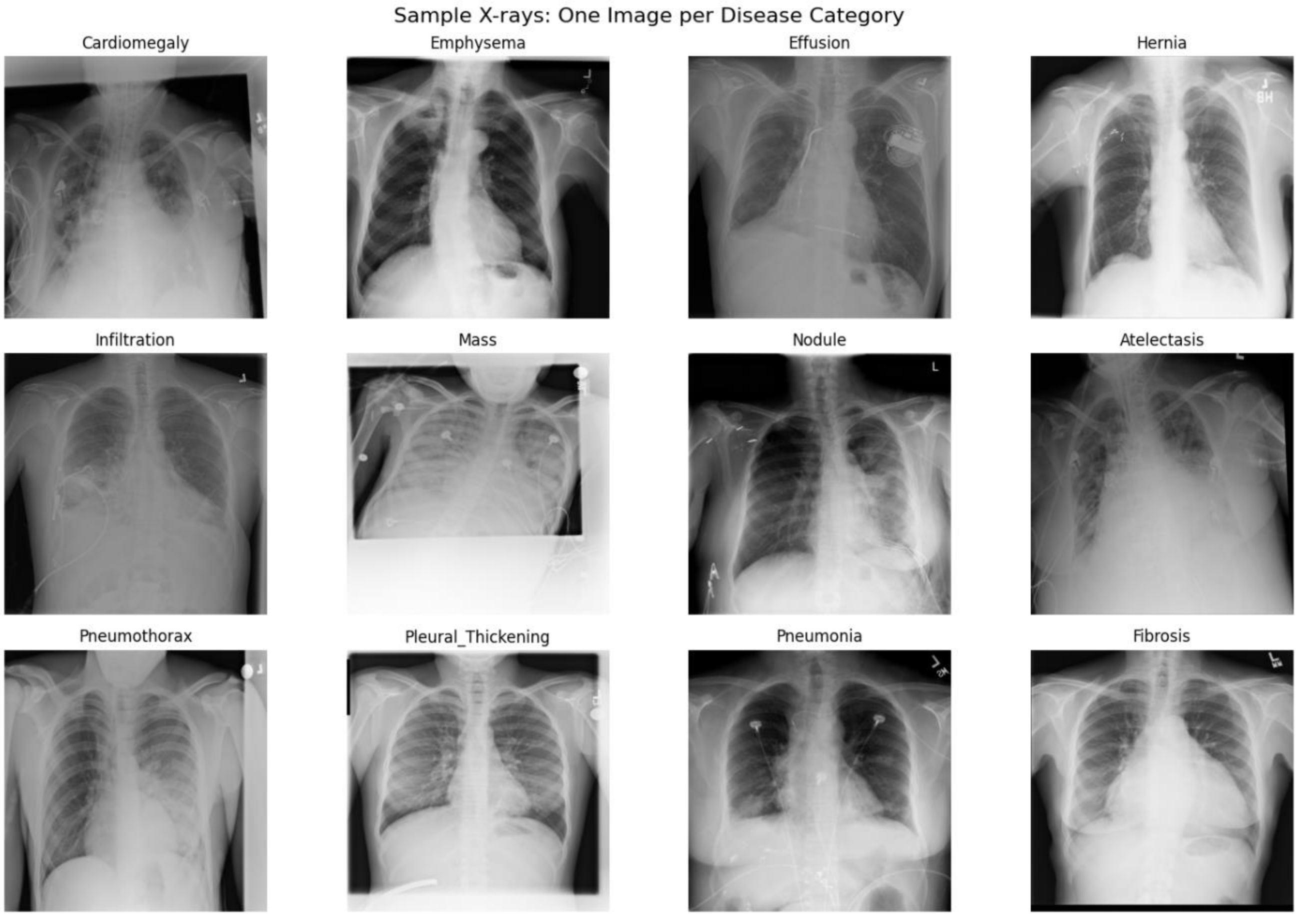
**Target Variable:** 15 binary disease columns
(*Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, Pneumothorax, Consolidation, Edema, Emphysema, Fibrosis, Pleural Thickening, Hernia, No Finding*)

# Dataset Overview

**Dataset:** NIH Chest X-ray 14 (Kaggle)

**Link:** https://www.kaggle.com/datasets/nih-chest-xrays/data

| **Property** | **Description** |
|---|---|
| Total Images(whole dataset)<br>Sample Images | 112,120<br>3445 |
| Patients | 30,805 |
| Labels | 14 thoracic diseases (multi-label) |
| Examples | Pneumonia, Edema, Effusion, Mass, Nodule, Atelectasis |
| Image Size | 1024×1024 px |
| Split | Train 80%, Val 20% |


Sample X-rays: One Image per Disease Category
Cardiomegaly
Emphysema
Effusion
Hernia
Infiltration
Mass
Nodule
Atelectasis
Pneumothorax
Pleural_Thickening
Pneumonia
Fibrosis

# Literature Review

| Study | Focus | Limitation | Our Extension |
| --- | --- | --- | --- |
| Wang et al. (2017) | Dataset + CNN baseline | Image-only learning | Added EHR + Text (multimodal) |
| Rajpurkar et al. (2017) | Single disease (pneumonia) | No ICU/clinical link | Multi-label across 15 diseases |
| Irvin et al. (2019) | Label uncertainty | No clinical context | Added BERT embeddings + ICU prognosis |

# Data Preprocessing

**Image Data:**

- Resized to 224×224 pixels
- Created **one-hot columns** for each disease (1/0)
- Augmentation (rotation, flip, zoom)

**EHR Data (future integration):**

- Imputation of missing values
- z-score scaling and encoding categoricals
- Built EHR features (age, gender, view) → scaled/encoded

**Text Data (future integration):**

- Clean notes (remove stopwords, symbols)
- Embeddings via BioBERT/ClinicalBERT
- Cleaned column names by standardizing Finding_Labels (| → ,strip spaces)

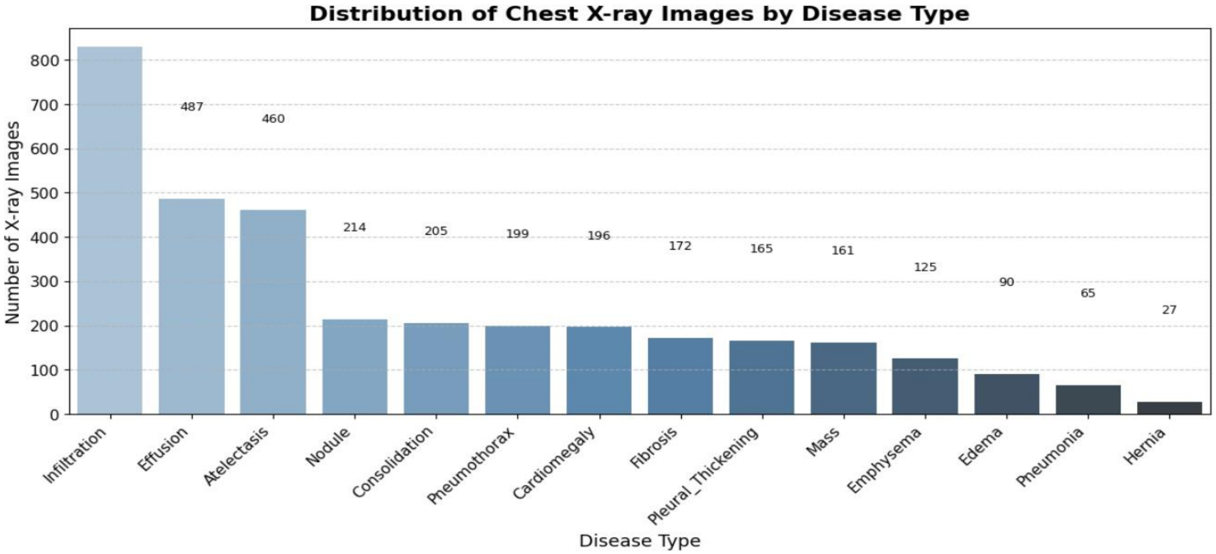
# Exploratory Data Analysis (EDA)

**Main Findings:**

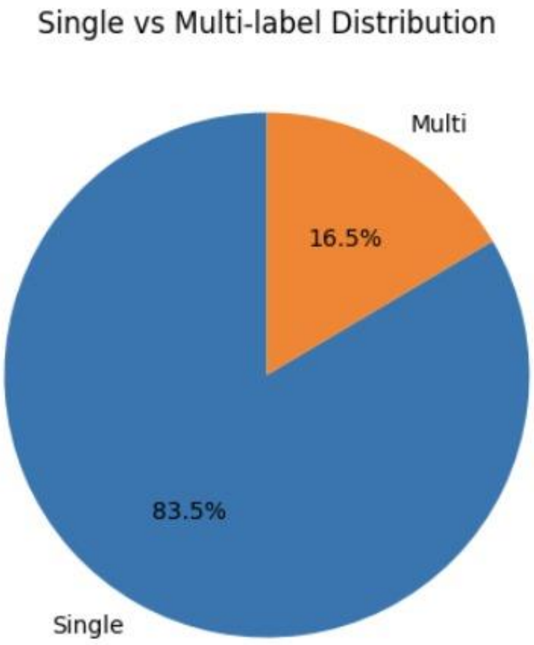
- “ No Finding ” class dominates → imbalance issue.
- Rare diseases (Hernia, Fibrosis)
- Common co-occurrence: Effusion + Atelectasis.
- Slight male majority in gender split.
- Patients range from infants to elderly.

# Visualization

Bar chart – Disease Frequency(excluding ‘No findings’)

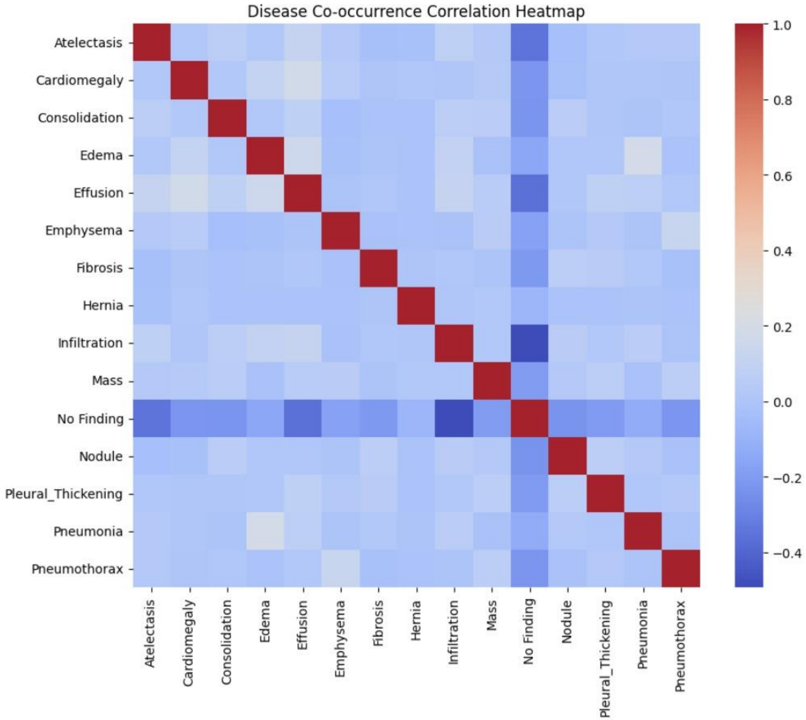


# Pie Chart- Proportion of single vs Multi-labeled X-ray




Single-labeled images: 4175
Multi-labeled images: 824

# Heatmap- Co-occurrence Correlation



# Model Construction

**Model Type:** Multimodal Deep Learning Network

- □ **Image Branch:** DenseNet121 CNN extracts visual features from X-rays.
- 📊 **EHR + Text Branch:** MLP processes patient data and clinical notes.
- 🔗 **Fusion Layer**: Combines both feature sets for final prediction.
- **Text branch**: ClinicalBERT embeddings (768-D).

# Training & Validation

**Splitting Strategy:**

Train 80% | Validation 20%

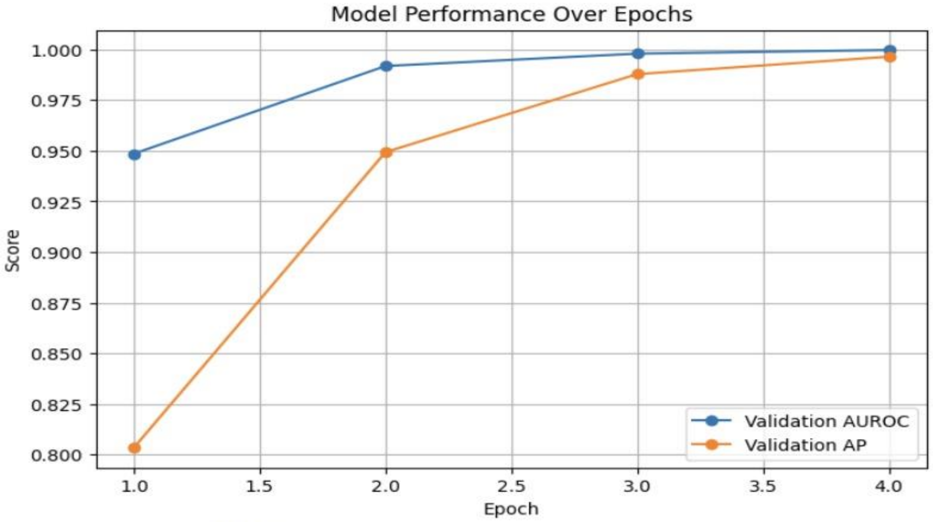
- Loss = BCEWithLogitsLoss for multi-label tasks.
- Optimizer = AdamW (lr = 1e-4).
- Batch = 16, Epochs = 4 (for subset).
- Evaluation = AUROC, Average Precision.

**Cross-Validation:** 10-Fold Stratified K-Fold

**Results (Sample Subset):**

- Epoch 1: Train loss = 0.28, Val AUROC = 0.71, Val AP = 0.23.
- Best Validation AUROC ≈ 0.91 (on final subset).
- ICU Outcome demo (confusion matrix): TN = 139, TP = 25, FN = 26, FP = 10 → 82 % accuracy.

# Model Evaluation

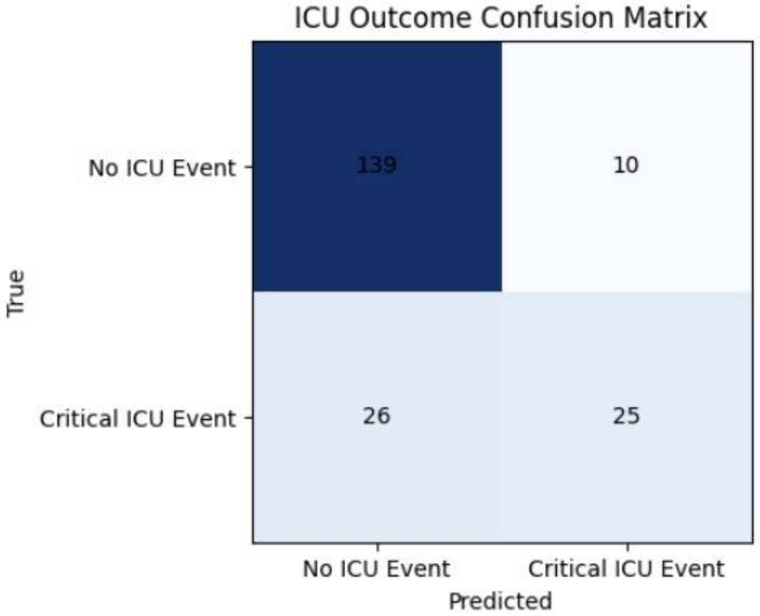



Results saved successfully!


```
{
  "best_epoch": 4,
  "best_AUROC": 0.9997,
  "best_AP": 0.9964,
  "train_loss_at_best": 0.0173,
  "history_csv": "/content/drive/MyDrive/606_Capstone/training_history.csv",
  "plot_path": "/content/drive/MyDrive/606_Capstone/performance_plot.png"
}
```

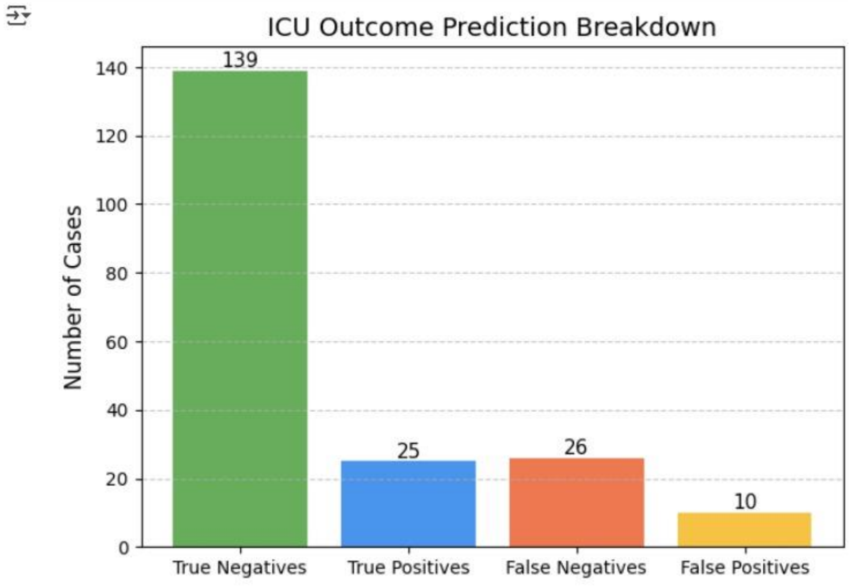
```
Simulated positive ICU outcome rate: 0.25

 ICU Outcome Model Results
Accuracy: 0.820
AUROC: 0.910
Confusion Matrix:
 [[139  10]
 [ 26  25]]
```



# Results: ICU Outcome Demo

- Overall correct ≈ **82%**; AUROC ≈ **0.91** (pilot)




TP/TN = correct predictions

FP/FN = incorrect predictions (FP = false alarm, FN = missed case).

# Expected Outcomes & Next Steps

- Train on **much larger sample** (tens of thousands)
- Better thresholding/calibration; per-class AUROC/AP
- Add **Grad-CAM** heatmaps for explainability
- Build a **Gradio UI** for quick clinician-style testing
- Document everything in GitHub & add reproducible scripts

# References


1. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr.2017.369
2. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., & Ng, A. Y. (2017). CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. ArXiv.org. https://arxiv.org/abs/1711.05225
1. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D. A., Halabi, S. S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C. P., Patel, B. N., Lungren, M. P., & Ng, A. Y. (2019). CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. Proceedings of the AAAI Conference on Artificial Intelligence, 33, 590–597. https://doi.org/10.1609/aaai.v33i01.3301590.

# Thank You