Project Title: Medical Image Classification for Disease Diagnosis Using Deep Neural Networks with Hyperparameter Optimization

Group 3: Kirandeep Kaur, Ibrahim Balogun, Elijah Ojelabi, Maxwell Owusu-Edusei

1.1 Introduction

Medical image classification is a critical component of modern healthcare, particularly in the diagnosis of thoracic diseases. (Litjens et al., 2017). Chest X-rays are among the most common and vital radiological examinations used for diagnosing thoracic diseases such as pneumonia, tuberculosis, and lung cancer (Wang et al., 2017). Deep learning, especially convolutional neural networks (CNNs), has revolutionized medical image analysis by effectively learning discriminative features directly from raw images without manual feature engineering (Goodfellow et al., 2016).

Recent advances in deep learning, especially convolutional neural networks (CNNs), enable models to learn complex patterns directly from raw images without relying on hand-crafted features. (Kingma & Ba, 2015). Moreover, hyperparameter tuning techniques like random and grid search are essential for enhancing model performance and generalization (Chen et al.,2016).

1.2 Objectives

1. The primary objective of this study is to develop a deep learning model for classifying medical images using a pre-processed subset of the ChestX-ray14 dataset. Specifically, the study aims to:
2. **Model Development:**
   Implement a convolutional neural network **(CNN)**–based model using a pretrained DenseNet121 backbone, adapted for multi-label thoracic disease classification. The network uses sigmoid activation in the output layer and is trained with the **Adam optimizer** (learning rate = 1e-4 during initial training and 1e-5 during fine-tuning), along with early stopping and learning-rate scheduling to support stable convergence.
3. **Hyperparameter Tuning:**
   Improve model generalization through random search over key hyperparameters, including learning rate, batch size, and training configuration (e.g., freezing vs. unfreezing CNN layers). The configuration achieving the best validation macro AUC is selected as the optimal model.
4. **Model Validation and Evaluation:**
   Evaluate the CNN's performance on a separate test set using metrics such as AUC, precision, recall, and accuracy, and interpret the results in a clinical context. Performance of the baseline training setup is compared with the fine-tuned model to assess the impact of hyperparameter optimization.

2. Literature Review

Deep learning has become a foundational approach in modern medical image analysis. Litjens et al. (2017) provided one of the most comprehensive surveys in the field, demonstrating that convolutional neural networks (CNNs) consistently achieve state-of-the-art performance across various medical imaging tasks, including detection, classification, and segmentation. Their work established deep learning as a transformative methodology in healthcare diagnostics.

A major advancement enabling large-scale model training in medical imaging was the introduction of the ChestX-ray14 dataset by Wang et al. (2017). This publicly available collection of over 100,000 frontal chest radiographs, labeled for multiple thoracic diseases, accelerated research in automated disease detection and benchmarking of CNN-based classifiers.
The foundational principles behind these advancements stem from earlier work in deep learning theory and practice. Goodfellow et al. (2016) emphasized the strength of CNNs in hierarchical feature extraction, showing that these architectures outperform traditional handcrafted feature approaches in general image classification. Their contributions helped solidify CNNs as the dominant architecture for vision-based tasks.

Training deep learning models efficiently also relies on robust optimization techniques. Kingma and Ba (2015) introduced the Adam optimizer, which combines adaptive learning rates with momentum to provide stable and efficient convergence. Adam has since become one of the most widely used optimization algorithms in both research and applied machine learning.
Finally, Chen et al. (2016) explored strategies for hyperparameter optimization, highlighting how techniques such as random search, Bayesian optimization, and early-stopping mechanisms can significantly improve generalization performance. Their work underscores the importance of systematic configuration tuning in achieving high-performing deep learning models.

3. Methodology

3.1 Data Preprocessing

All images were resized to 128×128 pixels to ensure consistent input dimensions and efficient processing on the available hardware. Pixel intensities were normalized to the [0, 1] range by dividing by 255. Although the original ChestX-ray14 images are grayscale, they are loaded as three-channel images, which makes them compatible with CNN architectures such as DenseNet121 that expect RGB-shaped input. Standard Keras ImageDataGenerator utilities were used to perform on-the-fly data loading, normalization, augmentation, and batching during training.

3.2 Baseline CNN Architecture

The baseline model is a convolutional neural network (CNN) built using the pretrained DenseNet121 architecture as the feature extraction backbone. DenseNet121 consists of multiple densely connected convolutional blocks, enabling efficient feature. The pretrained convolutional layers were loaded with ImageNet weights, and the top classification layer was removed to allow adaptation to the multi-label thoracic disease task.
An additional GlobalAveragePooling2D layer was added after the DenseNet backbone to reduce the spatial feature maps into a compact representation suitable for classification. The final output layer consists of 14 sigmoid-activated neurons, one for each thoracic disease, supporting multi-label prediction.
During the baseline training stage, the DenseNet backbone was kept frozen, and only the classification head was trained. This approach reduces overfitting and speeds up training when working with a limited dataset.

3.3 Training Protocol

Training incorporated several optimization strategies to improve convergence and reduce overfitting. Early stopping was applied by monitoring the *validation AUC* with a patience of four epochs, restoring the model weights from the best-performing epoch. A learning-rate scheduling

mechanism (ReduceLROnPlateau) was used to reduce the learning rate by a factor of 0.2 when validation AUC failed to improve for two consecutive epochs, with a minimum learning rate of 1×10^{-7}.

The model was trained using the Adam optimizer, with a learning rate of 1×10^{-4} during the initial frozen-backbone stage and 1×10^{-5} during fine-tuning. Training was performed with a batch size of 64, and although the maximum epoch count was set higher, training effectively completed after approximately 12 epochs due to early stopping.

4. Experiments

4.1 Data Source

This study utilizes the NIH Chest X-ray Dataset (ChestX-ray14), a large-scale public dataset comprising 112,120 frontal chest radiographs annotated for 14 thoracic disease categories. The dataset includes images from 30,805 unique patients and provides accompanying demographic information such as age and gender, as well as image acquisition details including posteroanterior (PA) and anteroposterior (AP) view positions. The disease labels span 14 classes: Infiltration, Atelectasis, Pneumonia, Pneumothorax, Consolidation, Edema, Emphysema, Fibrosis, Effusion, Pleural Thickening, Cardiomegaly, Nodule, Mass, and Hernia. A notable characteristic of the dataset is its pronounced class imbalance, with the "No Finding" category representing 53.8% of all images and positive disease prevalence ranging from as low as 0.2% for Hernia to 17.7% for Infiltration.

4.3 Data Split

For practical experimentation within available computational resources, this study used a curated subset of the ChestX-ray14 dataset consisting of 53,822 training images and 11,221 validation images, derived from a larger pool of patient-level–filtered samples. An additional test set of 11,196 images was maintained for final model evaluation. All splits were generated using patient-level stratification, ensuring that no patient appeared in more than one subset. This approach prevents data leakage, maintains subject independence between splits, and preserves the validity of performance estimates.

4.4 Hyperparameter Tuning (Random Search)

A random search over 5 trials was conducted to explore key training hyperparameters and identify configurations that improved validation AUC. The search space included:

Hyperparameter	**Search Space**
Learning rate	[0.0001,0.00001]
Batch size	[16, 32, 64]
Backbone training mode	FrozenDenseNet121,PartiallyFine−TunedDenseNet121
Data augmentation strength	Randomized rotation, shift, zoom, and horizontal flip ranges

4.5 Decision Threshold Optimization

Due to the severe class imbalance in the ChestX-ray14 subset, using the default decision threshold of **0.50** caused the model to predict almost all labels as negative. As a result, precision and recall for most disease classes collapsed to near zero. To address this, lower decision

thresholds were evaluated to identify a more suitable operating point for multi-label classification.

- **Threshold = 0.20:**
  Provided the most practical balance between sensitivity and specificity. At this threshold, several diseases—especially Effusion, Edema, Pneumothorax, and Consolidation—showed meaningful recall and moderate precision, making it a reasonable clinical trade-off.
- **Threshold = 0.10:**
  Further increased recall but at the cost of more false positives. This setting is more sensitive for disease detection but produces reduced precision, making it less appropriate unless the goal is high-sensitivity triage or screening.

Overall, a threshold of **0.20** was selected for reporting results, as it offered a more stable balance between identifying true positives and controlling false detections.

Test Set Composition

The held-out test set contained 11,196 chest radiographs, each annotated with 14 possible thoracic disease labels. All images were successfully loaded with no missing files. The distribution of positive cases across disease classes was highly imbalanced, with common conditions such as Infiltration (1,985 cases), Effusion (1,488 cases), and Atelectasis (1,173 cases) appearing far more frequently than rare conditions such as Hernia (20 cases), Pneumonia (134 cases), and Fibrosis (139 cases). This imbalance is consistent with the characteristics of the original NIH ChestX-ray14 dataset and significantly influences model performance, particularly recall on rare diseases.

Clinical Interpretation

The final DenseNet121 model achieved a macro AUC of 0.691, showing moderate ability to rank diseased vs. non-diseased cases. Common conditions such as Effusion, Edema, Pneumothorax, and Consolidation were detected more reliably, while rare labels like Hernia, Fibrosis, and Pneumonia showed very low recall due to limited examples.

A threshold of 0.50 produced almost all-negative predictions, so a lower threshold of 0.20 was used to balance sensitivity and precision. At this level, the model achieved micro precision = 0.303 and micro recall = 0.263, offering usable—but not diagnostic—performance.

Overall, the model is suitable as a computer-aided detection (CAD) support tool for highlighting likely abnormalities, with clinicians making final decisions. Class imbalance remains the main limitation and suggests the need for rebalancing or larger datasets in future work.

5.0 Conclusion and Recommendations

5.1 Conclusion

The fine-tuned DenseNet121 model achieved a macro AUC of 0.691, showing moderate ability to rank disease likelihood across thoracic conditions. Lowering the decision threshold to 0.20 improved sensitivity, highlighting the importance of threshold tuning in clinical settings. Although performance was strongest for common conditions such as Effusion and Edema, rare diseases remained difficult to detect due to imbalance. Overall, the model shows potential as a computer-aided detection (CAD) tool for prioritizing likely abnormalities, with clinicians making final decisions.5.2

5.2 Limitations

This study used a **curated subset** of the ChestX-ray14 dataset rather than the full 112,120 images, which may limit generalizability. Although DenseNet121 provides strong feature extraction, the model still struggled with rare classes because of pronounced class imbalance. The evaluation relied on a single patient-level split, meaning results may vary under different partitioning strategies.

5.3 Future Work

Future work should explore training on a larger portion of the full dataset and adopting additional imbalance-handling strategies such as class weighting, focal loss, or oversampling. More extensive fine-tuning of pretrained backbones and testing alternative architectures may also improve performance. Incorporating **cross-validation** and more advanced threshold-optimization methods would increase robustness and support development of a clinically aligned CAD tool.

References:

Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. 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). 2017. Available at: https://arxiv.org/abs/1705.02315. Accessed November 25, 2025.

Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. International Conference on Learning Representations (ICLR). 2015. Available at: https://arxiv.org/abs/1412.6980. Accessed November 25, 2025.

Goodfellow I, Bengio Y, Courville A. Deep Learning. MIT Press; 2016. Chapters 6 and 8. Available at: https://www.deeplearningbook.org/. Accessed November 25, 2025.

Chen R, Stewart WF, Sun J, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA. 2016;304(6):649-656. doi:10.1001/jama.2016.13416. Accessed November 25, 2025.

GeeksforGeeks. ML - Stochastic Gradient Descent (SGD). Available at: https://www.geeksforgeeks.org/machine-learning/ml-stochastic-gradient-descent-sgd/. Accessed November 25, 2025.

Wikipedia. Area under the ROC curve. Available at: https://en.wikipedia.org/wiki/Receiver_operating_characteristic. Accessed November 25, 2025.

Kaggle Dataset. NIH Chest X-ray Dataset. Available at: https://www.kaggle.com/datasets/nih-chest-xrays/data. Accessed November 25, 2025.