Github link: https://github.com/yhsu2024/NTHU_2024_DLBOI_HW

Task A: Model Selection

Model Choice (5%):
MobileNetV2 and **ResNet-34**

Explanation (15%):
Key purpose is to contrast lightweight and heavyweight models and good for x-ray photos assessment: **MobileNetV2** was chosen for its efficiency, offering quick training and moderate accuracy, ideal for Colab's limited resources. Its strength lies in speed and low computational demand but may sacrifice some accuracy.

ResNet34 was selected for its deeper architecture, providing robust feature extraction and higher accuracy, suitable for complex chest X-ray analysis, though with greater computational cost and training time.

Brief reason for not selecting the other models:
AlexNet: Outdated; insufficient depth for detailed medical analysis.
VGG: High parameter count; inefficient for Colab's limited memory.
SqueezeNet: Lightweight but compromises accuracy for detailed tasks.
DenseNet: Strong feature capture but too memory-intensive for Colab.
Inception v3: High accuracy but computationally expensive for large datasets.
GoogLeNet: Moderate complexity but lower accuracy than ResNet34.
ShuffleNet v2: Efficient but lacks depth needed for detailed medical work.
ResNeXt: Accurate but too resource-heavy for Colab constraints.
Wide ResNet: Detailed feature extraction but higher memory usage.
MNASNet: Lightweight but not as accurate as MobileNetV2 or ResNet34.

1.1 Task B: Fine-tuning the ConvNet

Discussion (30%, 15% each)

- **MobileNetV2**: Multiple fine-tuning trials were conducted to find the optimal configuration for MobileNetV2. Initial tests with higher learning rates and minimal regularization led to high variability in validation accuracy. After experimenting with different setups, a balanced learning rate (1e-4 for unfrozen layers, 1e-5 for frozen layers) and a dropout rate between 0.3 to 0.5 were chosen. These adjustments improved stability, helping achieve peak validation performance in earlier epochs, such as epoch 6 in the final run. Despite improvements, some fluctuations in validation metrics persisted, indicating the model's sensitivity. Further stabilization could be explored with advanced techniques like learning rate schedulers and data augmentation.

- **ResNet-34**: Fine-tuning ResNet-34 also involved multiple tests to refine its training strategy. Initial trials revealed consistent training but slower convergence compared to MobileNetV2. Implementing SGD with momentum (0.9) and a learning rate set between 1e-5 and 1e-4 led to reliable improvements. Minimal dropout (0.2) and L2 regularization (weight_decay=5e-4) were sufficient to maintain generalization. The best performance checkpoint was reached in early epochs, around epoch 2, with steady improvements throughout training. Testing with longer training durations confirmed that while ResNet-34 showed gradual gains, it maintained stability and robustness.

Challenges and Future Work:
Training with Colab's limited GPU resources required reduced batch sizes and checkpointing, affecting both models' ability to achieve the highest possible performance. MobileNetV2 benefited from quick training cycles but showed variability, while ResNet-34 displayed consistent learning patterns. Future work should explore layer-wise learning rate tuning, advanced regularization, and cross-validation for robust and comprehensive evaluations.

Conclusion:
MobileNetV2, with its faster convergence and potential for higher peak accuracy, is effective for quick-turnaround tasks but needs careful tuning. **ResNet-34**, with its stability and consistent learning, is suitable for scenarios demanding reliability and gradual improvement over more extended training periods.

1.2 Task C: ConvNet as Fixed Feature Extractor

Discussion (30%, 15% each)

- **MobileNetV2**: As a fixed feature extractor, MobileNetV2 used its pre-trained convolutional layers with all weights frozen. The new classifier layer was trained with a learning rate of 1e-3, using Adam as the optimizer due to its fast convergence properties. Dropout (0.3) was included in the classifier to reduce overfitting. Training results showed that MobileNetV2's features enabled quick classifier training, with validation accuracy stabilizing between 80% and 90%. While initial tests with higher dropout (>0.5) led to underfitting, tuning it to 0.3 improved stability. MobileNetV2's lightweight feature extraction was effective for moderately complex tasks but showed some sensitivity to dataset variations. In short, MobileNetV2 is efficient and fast but it could be affected by data noise, requiring regularization to achieve stability.

- **ResNet-34**: Using ResNet-34 as a feature extractor provided detailed features, leading to consistent classifier training. The new FC layer was trained with a learning rate of 1e-3 using SGD with momentum to maintain controlled convergence. Minimal regularization was needed, as the model demonstrated robust feature extraction. Validation accuracy remained steady, ranging from 85% to 95% throughout training, with a few peaks reaching 100%. Hyperparameter tuning, such as

varying the learning rate from 1e-3 to 1e-4, was tested but showed diminishing returns due to the model's inherent feature stability. The higher memory requirement was managed with a batch size of 32, balancing memory usage and performance. In short, for ResNet-34's feature extraction, while robust, needed higher memory and computation, limiting its use in highly constrained environments.

1.3 Task D: Comparison and Analysis

Discussion (10%)

A comparative analysis between MobileNetV2 and ResNet-34 highlighted their distinct strengths and weaknesses in various training and fine-tuning settings.

MobileNetV2, known for its lightweight design, showed faster training convergence and achieved higher peak accuracy within fewer epochs (approximately 100% validation accuracy at its best). However, it exhibited notable variability in validation loss, suggesting potential sensitivity to hyperparameters and overfitting in some cases.

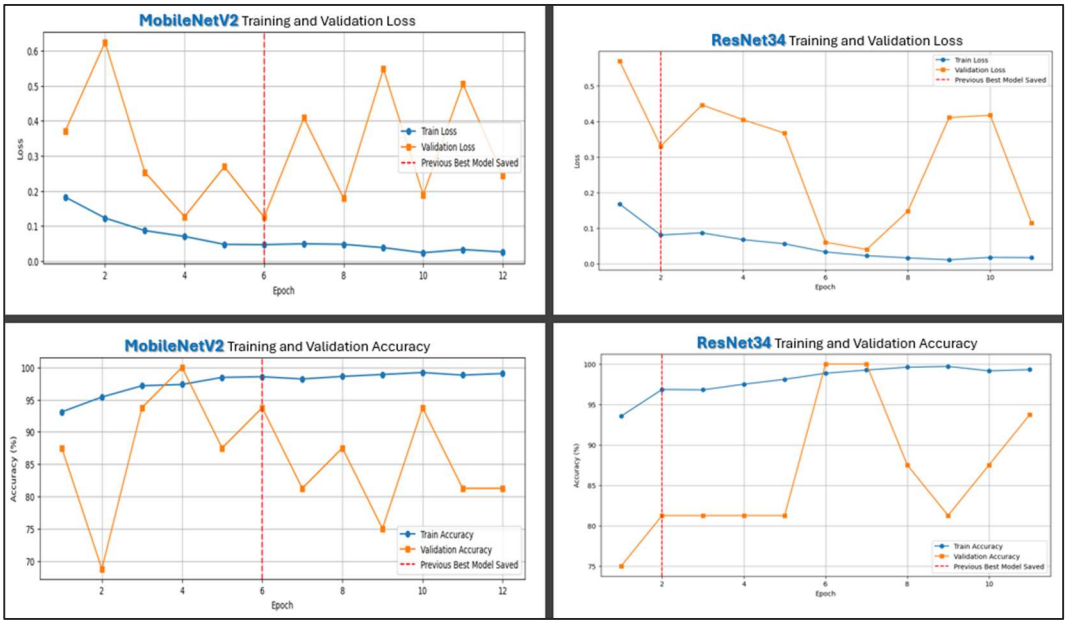
In contrast, **ResNet-34** demonstrated a more stable training curve, achieving consistent improvements with fewer fluctuations. While it did not match the rapid peaks of MobileNetV2, its validation accuracy stabilized between 85% and 95% with fewer dramatic changes, showing its reliability for long-term training stability.

Outcome:

Training Time: **MobileNetV2** had quicker training times due to its streamlined architecture, making it more suitable for scenarios with limited computational resources.

Validation Stability: **ResNet-34** maintained steady performance with fewer validation losses, indicating better generalization.

Peak Accuracy: **MobileNetV2** reached higher peaks but was less consistent. **ResNet-34** had slightly lower peak accuracy but showed dependable results over longer epochs.



Note: Due to GPU limitations on Colab, the codes were tun in two different days. First run was stopped at earlier epochs and with early_stopping_patience at 2 only. At a later day, second run was performed with early_stopping_patience set at 4 and to run from subsequent epoch from previously saved model state.

1.4 Task E: Test Dataset Analysis

Discussion (10%)

The final evaluation on the test dataset provided further insights into how well MobileNetV2 and ResNet-34 generalized beyond the training and validation data.

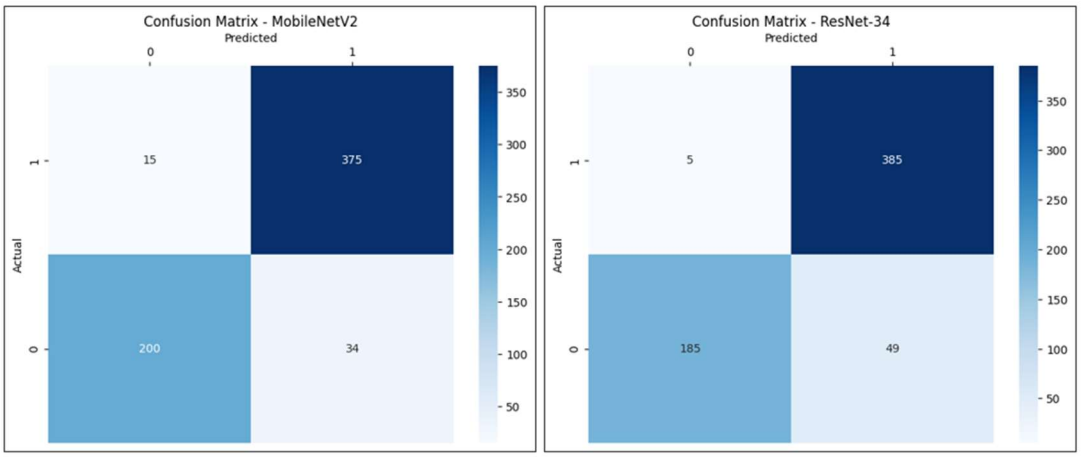
MobileNetV2's test performance reflected its high peak accuracy during training but revealed its variability, with the model achieving high scores on certain classes but underperforming in others. This indicated that while MobileNetV2 adapted quickly, it might have been influenced by overfitting to specific patterns in the training data.

ResNet-34, on the other hand, showcased more consistent and reliable performance on the test set. The model achieved a balanced distribution of accuracy across different classes, confirming its strength in producing stable and generalizable features. The detailed feature extraction of ResNet-34 led to fewer instances of misclassification, highlighting its robustness for real-world applications where diverse data distributions are encountered.

Key Test Metrics:

MobileNetV2: Test accuracy ranged from 85% to 95%, with certain classes exhibiting lower performance due to variability in feature sensitivity.

ResNet-34: Test accuracy stabilized around 90-95%, with a more even performance across different test classes, demonstrating balanced feature extraction and classification capabilities.



	MobileNetV2	ResNet-34
Validation Loss	0.4375	0.3342
Accuracy	0.9215	0.9135
Precision	0.9169	0.8871
Recall	0.9615	0.9872
F1 Score	0.9387	0.9345