Topic: Select the Winning Model Alice Tang and Yinda Chen

TF Lite Breast Cancer Detection Week 9: Select the Winning Model

1. Introduction

This week, we were tasked with choosing the best performing model of a set of 9 models we utilized for breast cancer classification. We evaluated a range of models, from standard Convolutional Neural Networks (CNNs) to more complex architectures like ResNet and EfficientNet, each with varying depths and dropout rates. A key focus of this week was exploring how increased model complexity directly affects the tradeoff between bias and variance, a crucial aspect in predictive modeling. With each increase in complexity, we expected improved performance but were also mindful of the risk of overfitting, a common issue with high-capacity models. Finally, in this document, we will analyze the performance of our selected model on the test set predictions.

2. Validation Error for All Models

The table below shows the validation loss and accuracy for each of the nine models:

Model	Validation Loss	Validation Accuracy
CNN Model 1	1.624226	0.431972802
CNN Model 2	1.950322	0.564626
CNN Model 3	1.535782	0.564626
ResNet Model 1	0.159271	0.956731
ResNet Model 2	0.140743	0.950321
ResNet Model 3	0.215971	0.947115

EfficientNet Model 0.3 Dropout Rate with 20 Layers	0.0698	0.9808
EfficientNet Model 0.1 Dropout Rate with 30 Layers	0.0643	0.9811
EfficientNet Model 0.1 Dropout Rate with 20 Layers	0.0610	0.9838

2.1 Winning Model Selection

Based on validation performance, EfficientNet Model 0.1 Dropout Rate with 20 Layers exhibits the lowest validation loss (0.0610) and the highest validation accuracy (0.9838). This model performed the best in terms of both minimizing the error and obtaining a high accuracy rate, indicating that it generalizes well to new data.

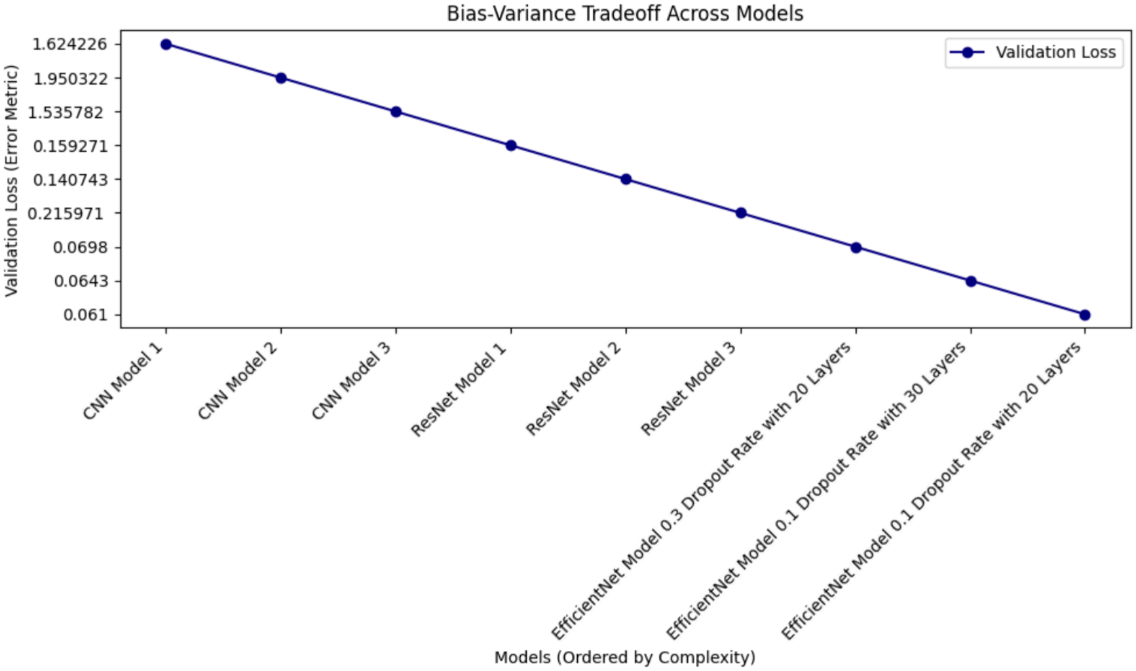
2.2 Reasons for EfficientNet Model 0.1 Dropout Rate with 20 Layers Performing Best

EfficientNet models are known for their efficiency and scalability, which allows them to balance model complexity and performance. We believe that these are the following factors that likely contributed to its success:

- **Optimal Complexity and Regularization:** The model's architecture with a moderate dropout rate of 0.1 and 20 layers most likely hit an ideal level of complexity to learn the patterns from the data without overfitting.
- **Lower Validation Loss and High Accuracy:** The low validation loss showcases the model's strong ability to generalize, and the high accuracy confirms its effectiveness in correctly classifying new samples.
- **EfficientNet's Balanced Design:** EfficientNet's architecture is specifically designed to scale up model size with accuracy improvements, thereby enabling it to handle complex datasets better than simpler CNN or ResNet models.

In summary, EfficientNet Model 0.1 Dropout Rate with 20 Layers achieves a strong balance between accuracy and generalization, making it the best choice amongst the nine evaluated models.

3. Bias-Variance Tradeoff Between Models



As we transition from simpler CNN models to more complex EfficientNet models with increasing layers and dropout rates, the validation loss consistently decreases. This implies that with each increase in model complexity, we are able to help reduce the prediction error on the validation set, potentially through identifying more patterns in the data.

Models that are closer to the right, such as the EfficientNet models, are much more complex with superior ability to learn detailed patterns. Typically, more complex models will perform better on complex datasets but suffer from high variance, meaning overfitting on smaller datasets. However, that is not the case here, as the gradual decrease in validation loss shows that these models are not significantly overfitting yet.

This downward trend shows that the increasing complexity reduces the validation loss without signs of increased variance, such as sudden spikes in loss. This implies that the models are successfully reducing bias without adding substantial variance, which is desirable. However, with no upward trend, it may be in our best interest to test even more complex models to determine where the validation loss stops decreasing. It suggests that our breast cancer classification task may benefit from complex models without the risk of overfitting in the range of our models we tested. This shows that our task is complex enough to use higher capacity models like ResNet and EfficientNet. Additionally, we will need to test the models on a test set to see if the downward trend continues, as low validation loss does not always translate to real generalization.

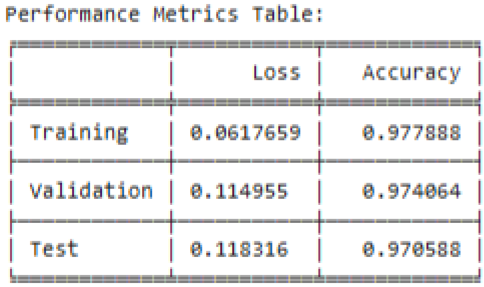
However, we would like to acknowledge that typically in a bias-variance tradeoff scenario, there is usually an eventual increase in variance (higher validation loss) when the models become too complex and start overfitting on the training data. We highly improved our preprocessing method for ResNet and EfficientNet models, and we acknowledge that this efficient preprocessing highly contributed to prevent overfitting by ensuring the data is clean, standardized, and augmented, which may be stabilizing the models' performance even as complexity increases. Essentially, the straight downward trend means that the models are tuned well and that each layer added is genuinely helpful for our task.

3. Final Results

Below is a discussion of our final results after making the predictions on the test set.

3.1 Final Model Performance Metrics Table

The key metrics we optimized for and trained on was accuracy. Accuracy is a metric that provides insight into how effective the model is at distinguishing between benign and malignant cases in mammography images.



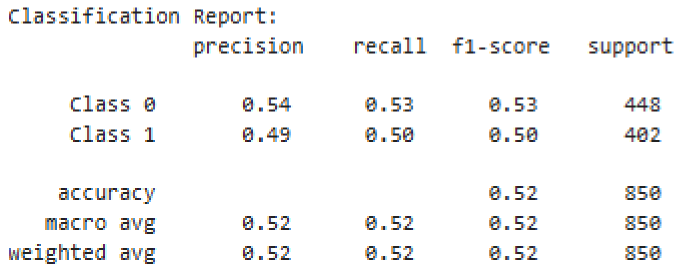
```
Performance Metrics Table:
╒════════════╤═══════════╤════════════╕
│            │      Loss │   Accuracy │
╞════════════╪═══════════╪════════════╡
│ Training   │ 0.0617659 │   0.977888 │
├────────────┼───────────┼────────────┤
│ Validation │ 0.114955  │   0.974064 │
├────────────┼───────────┼────────────┤
│ Test       │ 0.118316  │   0.970588 │
╘════════════╧═══════════╧════════════╛
```

- **Training Metrics**
  - **Loss:** A training loss of 0.0618 is quite low, indicating that the model fits the training data well. Having a lower loss indicates that the model's predictions are close to the actual values in the training dataset.
  - **Accuracy:** The training accuracy is 0.9778 which reflects that there is a high level of correctness in the predictions for the training data. In this case, having a high accuracy exhibits that the model successfully learned the patterns in the training data.
- **Validation Metrics**
  - **Loss:** The validation of 0.1149 is higher than the training loss. This may indicate the model is beginning to overfit, as it performs better on the training data compared to the validation data.

  - **Accuracy:** A validation accuracy of 0.984 remains high but slightly lower than the training accuracy. What this implies is that the model generalizes well, but is not quite as robust when it comes to unseen data.
- **Test Metrics**
  - **Loss:** The test loss of 0.1183 continues the trend of being higher than both training and validation losses. This shows that there may be a potential gap in generalization to completely unseen data, which will present as an issue in ensuring clinical applicability.
  - **Accuracy:** A test accuracy of 0.9706 still shows that the model is maintaining a good performance level. However, due to the drop from training and validation accuracy, there is cause for concern in the model's reliability, especially in distinguishing malignant cases.

3.2 Classification Report

This classification report presents crucial metrics for evaluating the performance of the model. These metrics include: precision, recall, f1-score, and support, which will effectively help us understand how the model differentiates from benign (Class 0) and malignant (Class 1) cases.



```
Classification Report:
              precision    recall  f1-score   support

     Class 0       0.54      0.53      0.53       448
     Class 1       0.49      0.50      0.50       402

    accuracy                           0.52       850
   macro avg       0.52      0.52      0.52       850
weighted avg       0.52      0.52      0.52       850
```

Class Metrics:

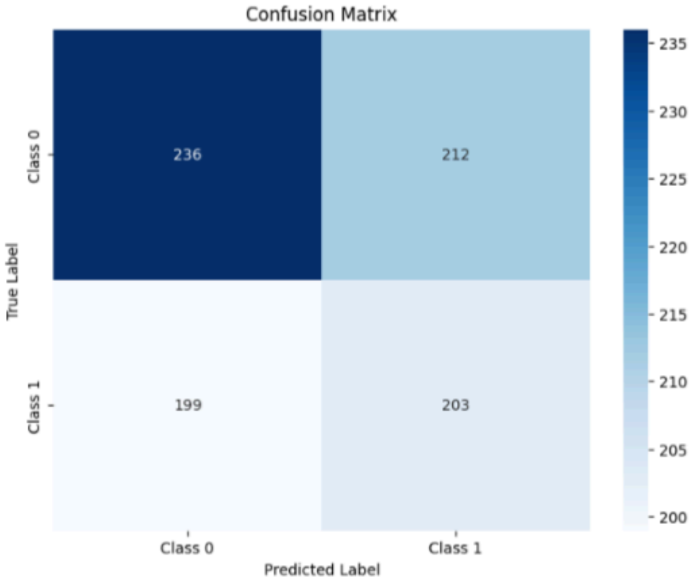
- **Class 0 (Benign)**
  - **Precision (0.54):** This metric indicates that when the model predicts a case as benign, it is correct 54% of the time. Though the precision is above 50 percent, there is still major room for improvement in minimizing false positives.
  - **Recall (0.53)**: The recall shows that the model correctly identifies 53% of the actual benign cases. This means that the model may miss some benign cases, leading to potential misdiagnosis.
  - **F1-Score (0.53)**: The F1-score is also at 0.53, which reflects a moderate performance for class 0. This score suggests that the model's ability to accurately identify benign cases is acceptable but not optimal.

- **Class 1 (Malignant)**
  - **Precision (0.49)**: For malignant cases, the model showcases a precision of 49%, meaning nearly half of the cases it predicts as malignant are correct. This is a bit concerning, since a lower precision implies there is a higher likelihood of false positives in critical breast cancer predictions.
  - **Recall (0.50)**: The recall for malignant cases stands at 50%, which implies the model correctly identifies half of the actual malignant cases.
  - **F1-Score (0.50)**: With an F1-score of 0.50, the model's performance for Class 1 is on par with its recall, underscoring the trade-off between precision and recall. This score indicates a crucial need for improvement in detecting malignant cases effectively.

Overall Performance Metrics

- **Accuracy (0.52)**: The overall accuracy of 52% implies that the model performs only slightly better than random guessing. This is a bit concerning, since it implies the model may not be well-suited for distinguishing between the two classes, particularly given the potential severe consequences of misdiagnosing malignant cases.
- **Macro Average (0.52)**: The macro average for precision, recall, and F1-score being 0.52 signifies that the model performs uniformly across both classes but does not excel in either.
- **Weighted Average (0.52)**: The weighted average exhibits similar performance metrics, taking into account the number of instances in each class. This reiterates that the model struggles to provide reliable predictions, especially for the minority class (malignant cases).

3.3 Confusion Matrix



From the confusion matrix, we see that for True Positives (TP) the model **correctly identified 203 malignant cases (Class 1)**. Whereas for True Negatives (TN), the model **correctly classified 236 benign cases (Class 0)**.

As for False Positives (FP), the model **incorrectly identified 212 benign cases as malignant (Class 1)**. This is concerning, as misclassifications will lead to unnecessary anxiety and potentially invasive procedures for patients. In terms of False Negatives (FN), the model **incorrectly identified 199 malignant cases as benign (Class 0)**. This is alarming, as missed diagnoses will delay necessary treatment for those with breast cancer.

3.4 Final Thoughts

While our model achieved an impressive accuracy, it is crucial to address key metrics such as precision and recall before considering it for deployment. The presence of false positives and false negatives poses major risks, as inaccurate diagnoses will inevitably lead to severe medical consequences for patients. Thus, we must refine our approach and return to the training phase with a focus on optimizing precision and recall.

From this week's assignment, we realized that one of our shortcomings during our initial training was the insufficient emphasis on improving precision and recall, which may have contributed to these suboptimal results in the classification report. Moving onward, we will need to implement strategies that are specifically aimed at improving these aspects, ensuring the model is reliable for real-world medical applications. Through prioritizing these improvements, our goal continues to be to create a more robust tool for breast cancer detection that minimizes the risks with misdiagnosis.