Model Descriptions

Brief Model(Model 1): This model has one Convolutional Layer, one fully connected layer and an output layer.

Deeper Model(Model 2): This model has three Convolutional layers, one fully connected layer and an output layer.

Deeper Model 2(Model 3): It has the same architecture as the deeper model, but has data augmentation (Zoom, Rotation, Width/Height Shift, Shear, Flip) performed on it.

Deeper Model 3(Model 4): It has the same architecture as the deeper model but has height shift and Elastic distortion applied to it.

ResNet Model(Model 5): Uses ResNet152 as base model

VGGNet Model(Model 6): Uses VGG16 as base model

Comparing Training and Test Times for Models

Model	Training time (s)	Test time (ms)
Model 1	129	279
Model 2	177	298
Model 3	1136	346
Model 4	**4743**	336
Model 5	1717	**5160**
Model 6	633	1770

Observations about Runtimes

- Model 4 has the longest training runtime
- Model 5 has the longest test runtime

Comparing Performance Measures for all models

Training Set

Model	Accuracy	Precision	Recall	F score
Model 1	**99.81**	**1.0**	**1.0**	**1.0**
Model 2	99.53	1.0	1.0	1.0
Model 3	89.49	0.9	0.89	0.89
Model 4	95.17	0.95	0.95	0.95

Model	Accuracy	Precision	Recall	F score
Model 5	82.23	0.82	0.82	0.82
Model 6	92.92	0.93	0.93	0.93

Validation Set

Model	Accuracy	Precision	Recall	F score
Model 1	89.37	0.89	0.89	0.89
Model 2	**91.95**	**0.92**	**0.92**	**0.92**
Model 3	86.57	0.87	0.87	0.87
Model 4	91.33	0.91	0.91	0.91
Model 5	80.28	0.8	0.8	0.8
Model 6	86.42	0.86	0.86	0.86

Test Set

Model	Accuracy	Precision	Recall	F score
Model 1	89.57	0.9	0.9	0.9
Model 2	**91.77**	**0.92**	**0.92**	**0.92**
Model 3	87.2	0.87	0.87	0.87
Model 4	**91.32**	**0.91**	**0.91**	**0.91**
Model 5	81.13	0.81	0.81	0.81
Model 6	85.42	0.86	0.86	0.86

Best Performance marked in Bold

Evaluating model performance with other metrics

The evaluation metrics used are Accuracy, Precision, Recall and F-score. These metrics were chosen for evaluation considering the nature of the problem (Medical classification).

- Precision decribes the proportion of positive cases correctly identified.
- Recall descripes the proportion of all positive cases that were actually detected.
- F-score is the harmonic mean of precision and recall.

Accuracy

- From the tables above, Model 1 had the highest training accuracy (99.81) followed by Model 2 (99.53) while Model 5 had the lowest training accuracy (82.23).
- Model 2 performed best on the validation set with an accuracy of 91.95 followed by Model 4 (91.33). Model 5 had the least accuracy on the validation set (80.23).
- On the test set, Model 2 performed best with an accuracy of 91.77 followed closely by Model 4 (91.33). Model 5 had the least accuracy compared to the other models on the test

set (81.13).

Precision

- Model 1 had the best precision(1.0) on the training set, Model 5 had the worst precision(0.82) on the training set.
- Model 1 had the best precision(0.92) on the training set, Model 5 had the worst precision(0.8) on the validation set.
- On the test set, Model 2 performed best with an precision of 0.92 followed closely by Model 4 (0.91). Model 5 had the least precision compared to the other models on the test set (0.81).

Recall

- Model 1 had the best recall(1.0) on the training set, Model 5 had the worst recall(0.82) on the training set.
- Model 1 had the best recall(0.92) on the training set, Model 5 had the worst recall(0.8) on the validation set.
- On the test set, Model 2 performed best with an recall of 0.92 followed closely by Model 4 (0.91). Model 5 had the least recall compared to the other models on the test set (0.81).

F-score

- Model 1 had the best F-score(1.0) on the training set, Model 5 had the worst F-score(0.82) on the training set.
- Model 1 had the best recall(0.92) on the training set, Model 5 had the worst F-score(0.8) on the validation set.
- On the test set, Model 2 performed best with an F-score of 0.92 followed closely by Model 4 (0.91). Model 5 had the least F-score compared to the other models on the test set (0.81).

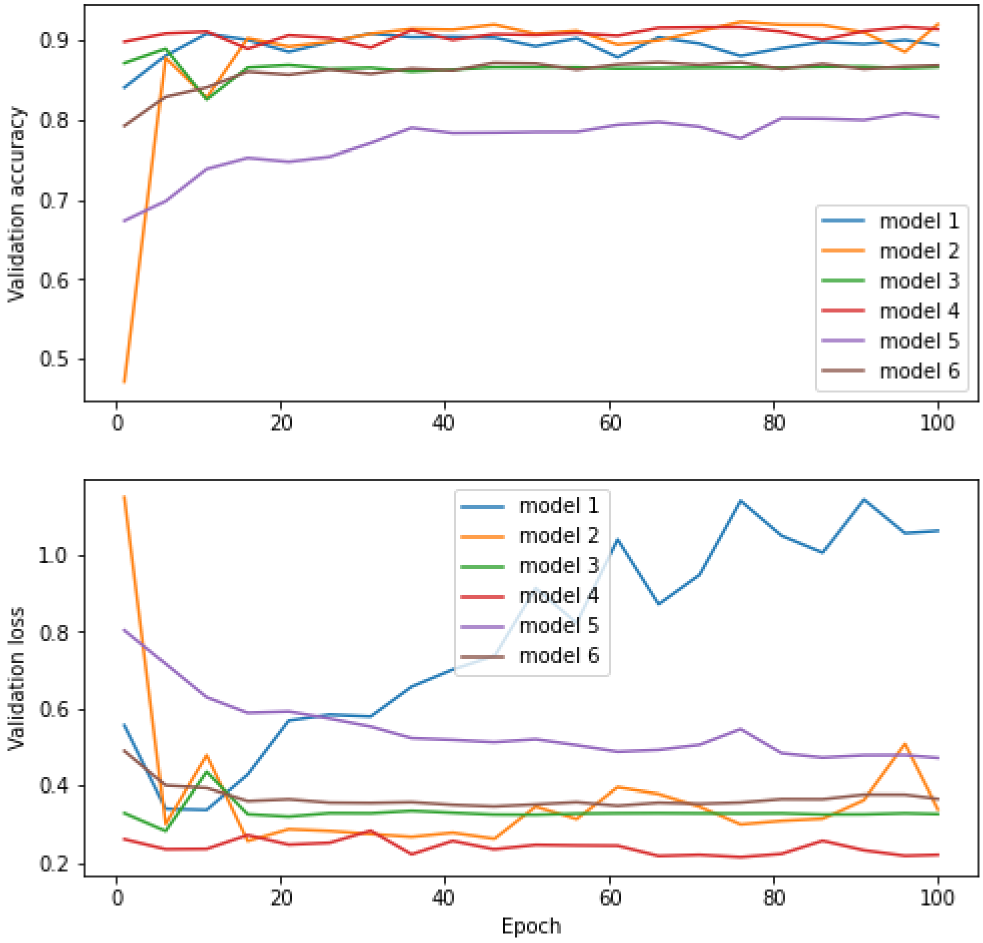
Comparison of Different Models

In [ ]:

```
fig, (ax1, ax2) = plt.subplots(2, figsize=(8, 8))
count = 1
epochs = list(range(1,100, 5))
epochs.append(100)
for history in history_list:
    label = 'model ' + str(count)
    val_accuracy = [history.history['val_accuracy'][i] for i in range(len(histor
    val_loss = [history.history['val_loss'][i] for i in range(len(history.histor
    val_accuracy.append(history.history['val_accuracy'][-1])
    val_loss.append(history.history['val_loss'][-1])
    ax1.plot(epochs, val_accuracy, label= label)
    ax2.plot(epochs, val_loss, label=label)
    count += 1

ax1.set_ylabel('Validation accuracy')
ax2.set_ylabel('Validation loss')
ax2.set_xlabel('Epoch')
ax1.legend()
ax2.legend()
```

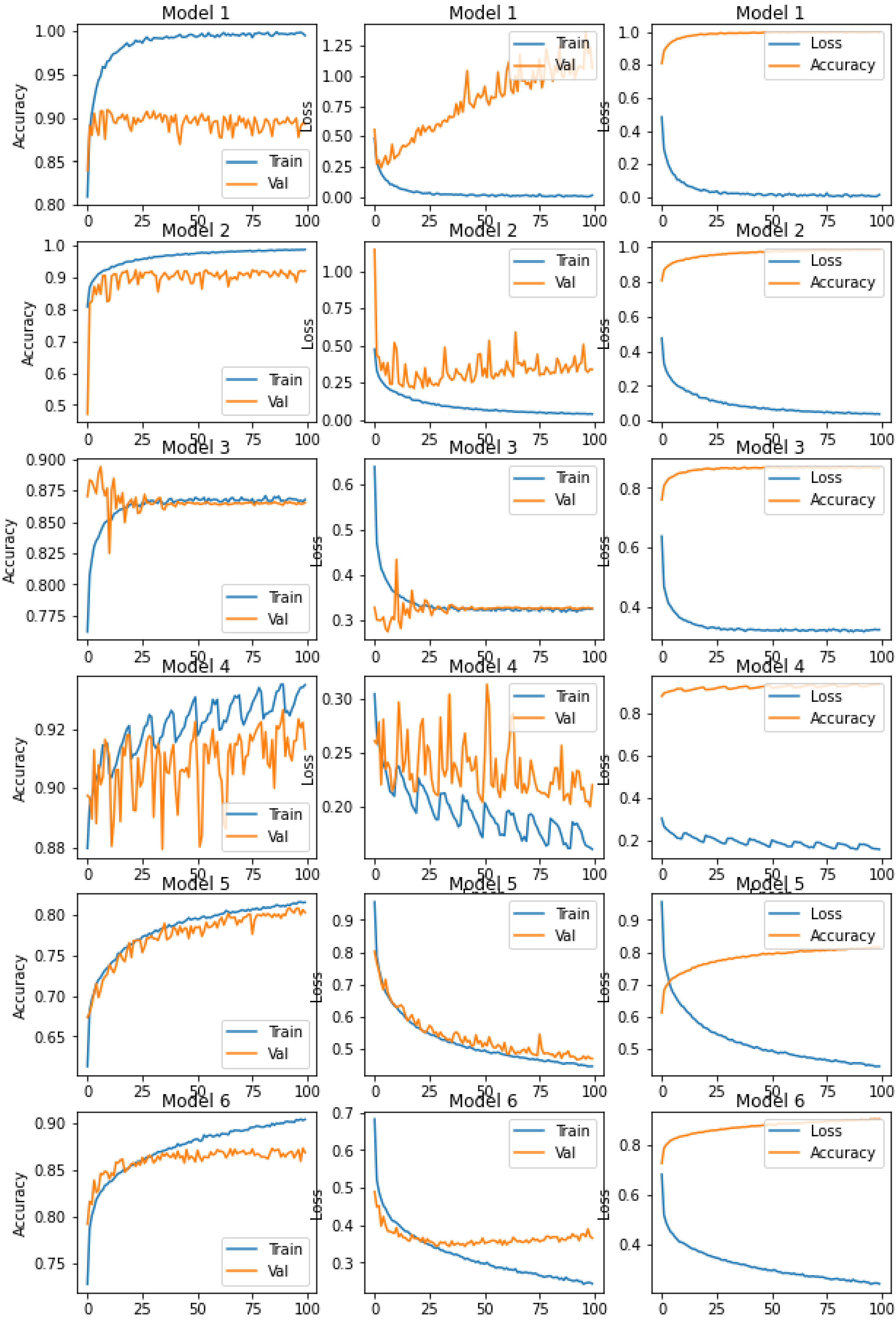
```
Out[ ]: <matplotlib.legend.Legend at 0x7f4ee84bfd10>
```



Observations

- Model 4 (ResNet-152) has the lowest Validation Accuracy
- Model 1 (Brief Model) has the highest Validation Loss

```
In [ ]: compare_accuracy_loss(history_list)
```



Observations about Accuracy and Loss plots

- Models 2, 4, 5, 6 have relatively lower overfitting
- Models 1 and 2 have significant overfitting