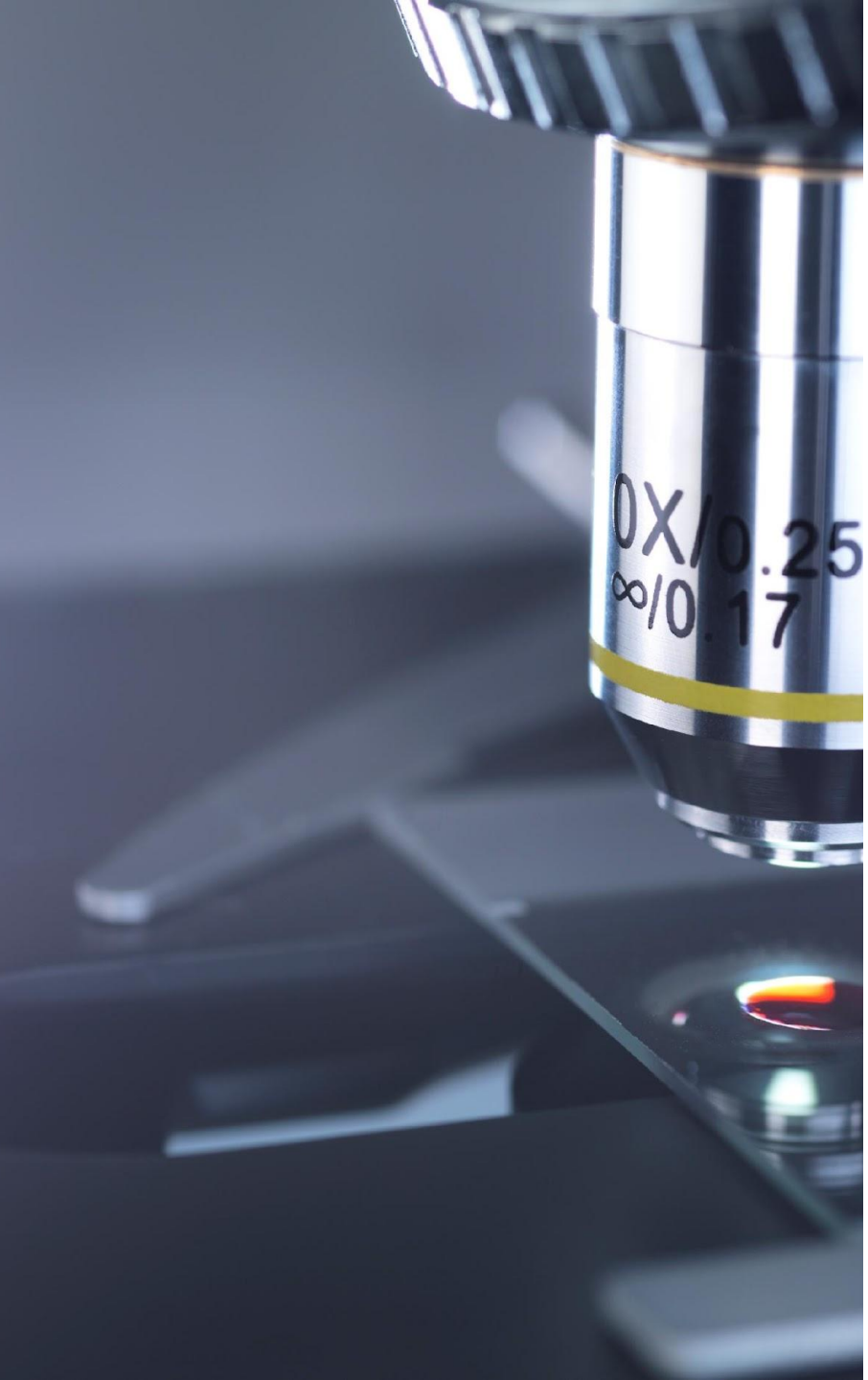
LEUKEMIA BLOOD SMEAR IMAGE CLASSIFICATION



RESEARCH PROBLEM, IMPORTANCE, AND IMPLICATIONS

- Acute Lymphoblastic Leukemia, or ALL, is a blood cancer that is notoriously difficult to identify and can be deadly within months if untreated.
- Current identification of diseased blood samples are completed by examination by an expert pathologist or oncologist.
- Our aim is to apply machine learning, specifically deep learning, to classify images of blood lymphocytes as normal or positive for ALL.
- Models such as ours can be used to aid in screening or diagnostic decisions for physicians.

“YOU SHOULD ASSUME THAT FOR MOST KINDS OF AN IMAGE THAT A HUMAN LOOKS AT, A COMPUTER COULD RECOGNIZE OBJECTS IN THAT IMAGE MORE QUICKLY AND MORE ACCURATELY THAN YOU CAN.”

-JEREMY HOWARD

DATASET

- IEEE International Symposium on Biomedical Imaging (2019)
- Currently available on Kaggle
- Blood smear images of lymphocytes
- Training, validation, and test data is provided

DATA CLEANING AND EXPLORATION

"PATIENCE IS BITTER, BUT ITS FRUIT IS SWEET."
–JEAN-JACQUES ROSSEAU

DATA EXPLORATION

We explored the data provided by the challenge authors:

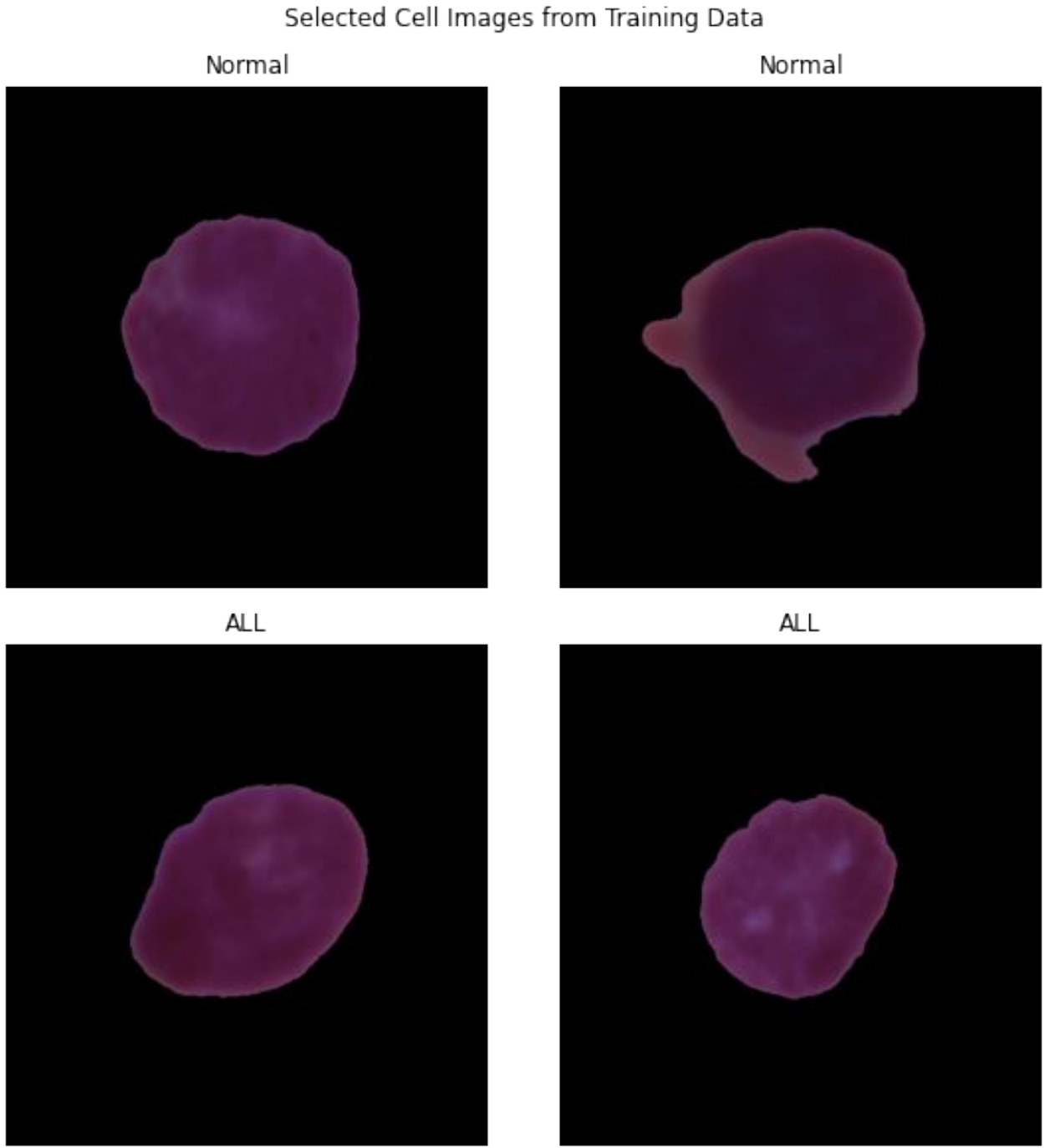
- Training Data
  - 10,661 images
- Validation Data
  - 1,867 images
- Test Data
  - 2,586 images

Other key findings

- Data structured in multiple folders, especially training data, where folds of data created
- No patient overlap between datasets or folds of data
- Ground truth label structure/location varies between training and validation folders
- No ground truth labels exist for test data, so we opted not to use it
- Between the remaining images in training and validation, there are 13,247 images of cells from 90 subjects

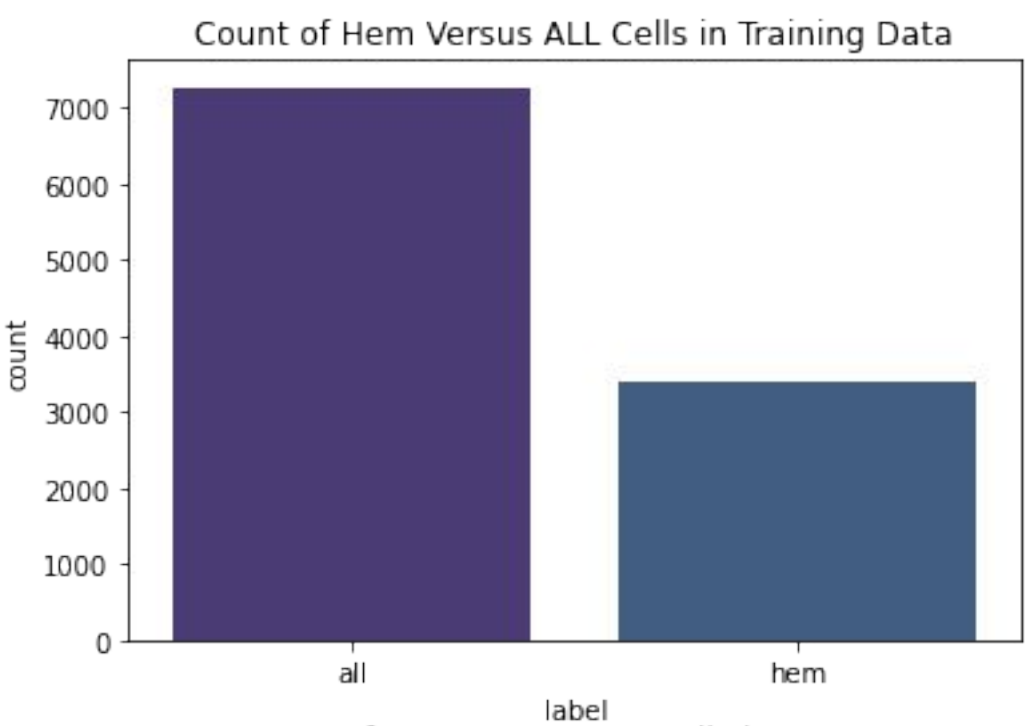
INITIAL DATA LOAD

- We loaded our data into a pandas DataFrame for easy manipulation. We have no other fields for features aside from images and labels.
- We visualized several randomly selected images and observed difficulty of differentiating normal and cancerous cells.
- Excess blank space was also observed in images and could benefit from cropping



FURTHER DATA REVIEW

- Imbalanced data in both training/validation and test datasets
- Roughly 2:1 ratio between cancerous:normal images
- Training/validation data:
  - ALL: 7,272
  - Hem (normal): 3,389
- Test data:
  - ALL: 1,219
  - Hem (normal): 648



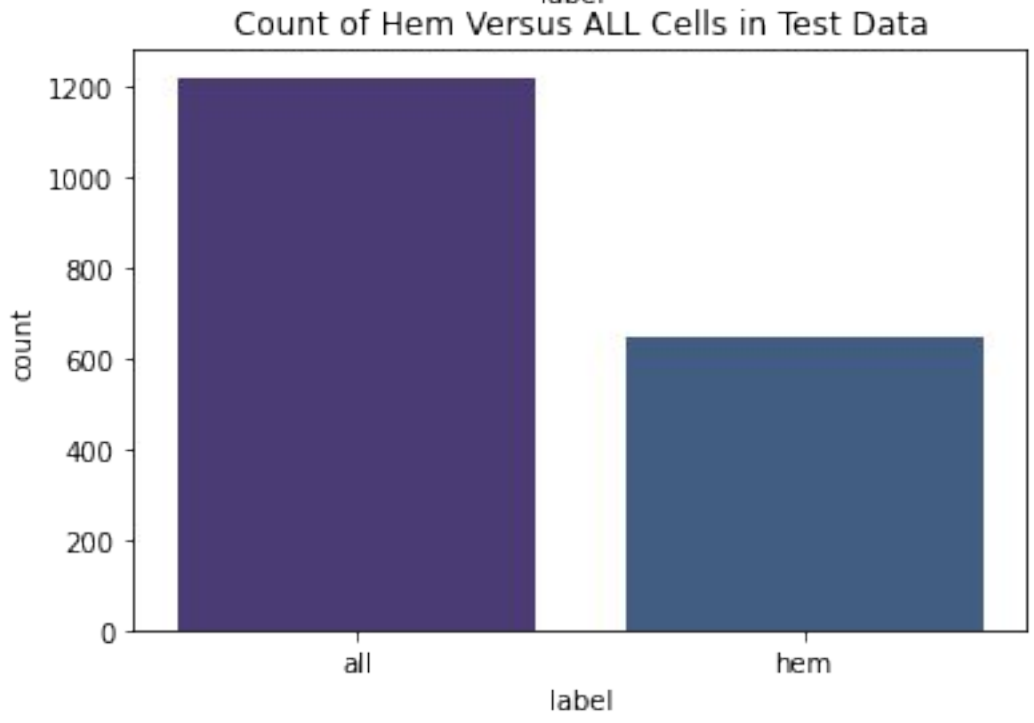
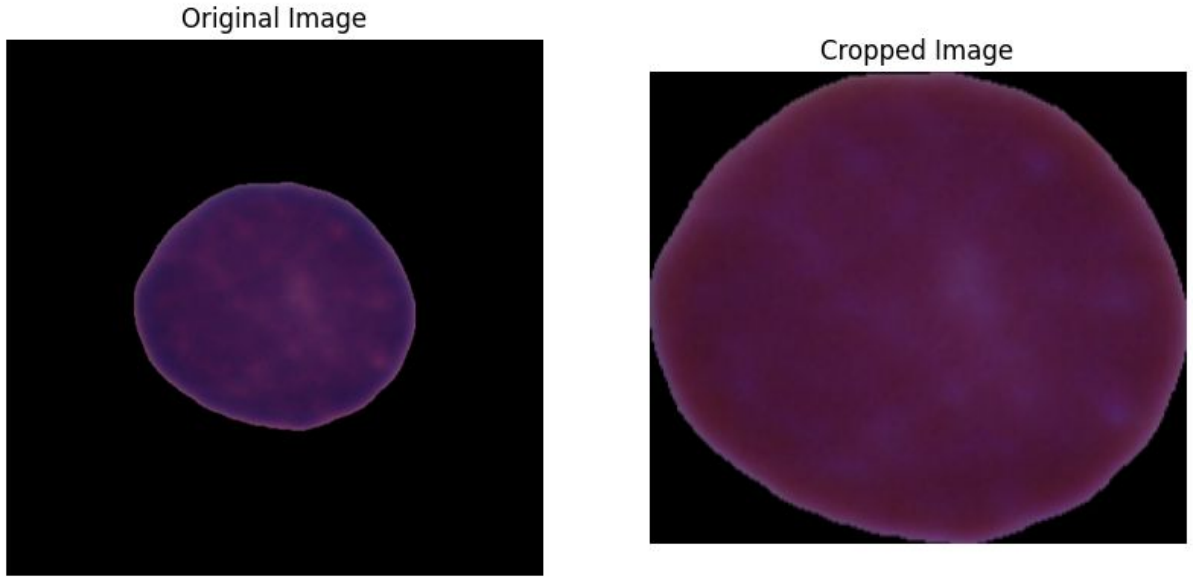


IMAGE PROCESSING

We performed image processing in batches in preparation for our models

- Cropped images
- Added filepaths for cropped images to DataFrames
- Split train/validation
  - Train: 7,035 images
  - Validation: 1,759 images
- Test Data: 1,867

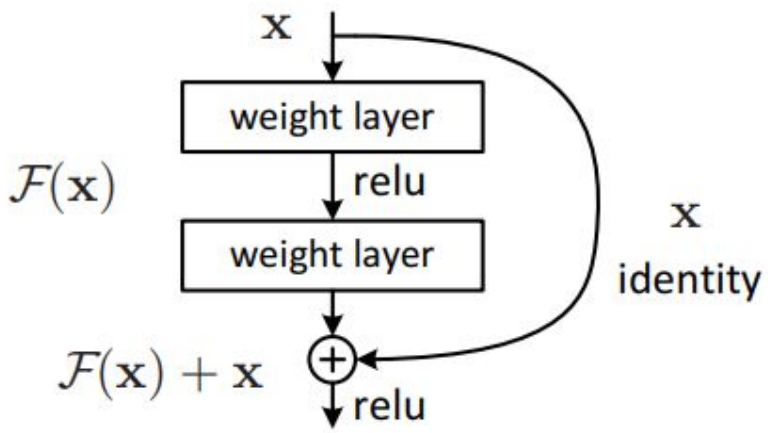


Original versus cropped image of randomly selected image

MODEL TRAINING AND ANALYSIS

"PATIENCE IS BITTER, BUT ITS FRUIT IS SWEET."
–JEAN-JACQUES ROSSEAU

PRE-TRAINED MODEL – RESNET50



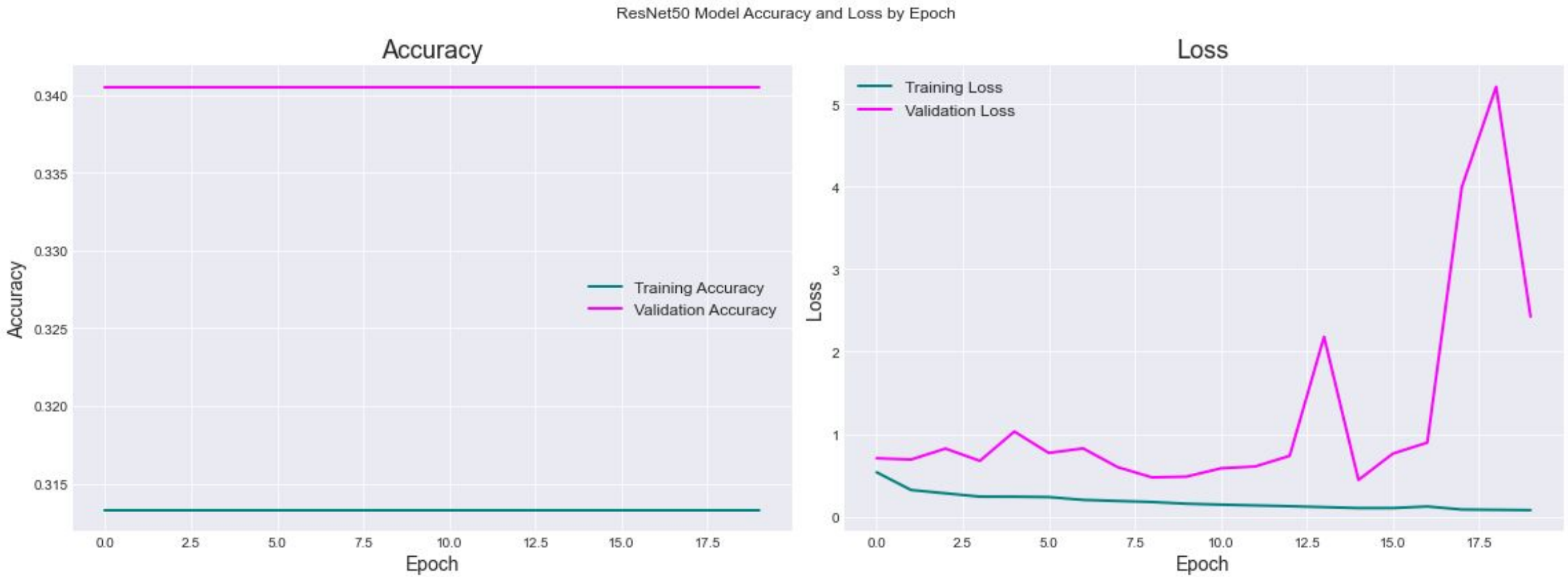
Residual learning block, from the original paper

"Deep Residual Learning for Image Recognition"

ResNet50 is a convolutional neural network that uses residual learning

- 50 layer deep model
- Residual building blocks, or shortcut layers, that are trained later when the model is re-trained
- Allows better evaluation of feature space of image by the model
- Addresses vanishing gradient issue seen in deeper networks
- Available in Keras library
- Added additional layers on top of model to customize for our images

RESNET50 MODEL PERFORMANCE



ResNet model training and validation accuracy and loss over 20 epochs

CUSTOM-BUILT CNN MODEL

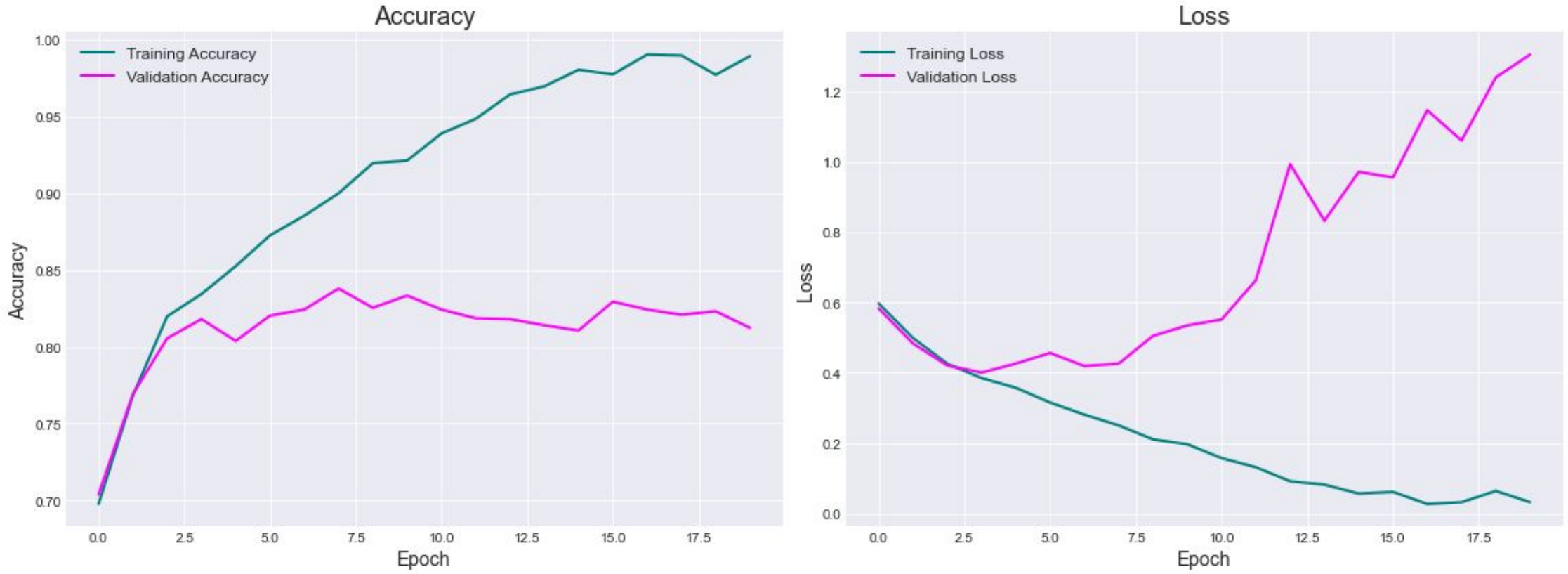
```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv2d (Conv2D)             (None, 222, 222, 32)      896

 max_pooling2d (MaxPooling2  (None, 111, 111, 32)      0
 D)

 conv2d_1 (Conv2D)           (None, 109, 109, 64)      18496

 max_pooling2d_1 (MaxPoolin  (None, 54, 54, 64)        0
 g2D)

 conv2d_2 (Conv2D)           (None, 52, 52, 64)        36928

 max_pooling2d_2 (MaxPoolin  (None, 26, 26, 64)        0
 g2D)

 flatten (Flatten)           (None, 43264)             0

 dense (Dense)               (None, 64)                2768960

 dense_1 (Dense)             (None, 1)                 65

...
Total params: 2825345 (10.78 MB)
Trainable params: 2825345 (10.78 MB)
Non-trainable params: 0 (0.00 Byte)
```

Initial model summary from Keras

Created a Sequential model in Keras with the following architecture:

- 3 alternating layers each of convolutional layers and max pooling layers
- Flatten layer to convert 2D to 1D feature vectors
- Fully connected layer with ReLU activation
- Output layer with sigmoid for binary classification

INITIAL CNN MODEL PERFORMANCE

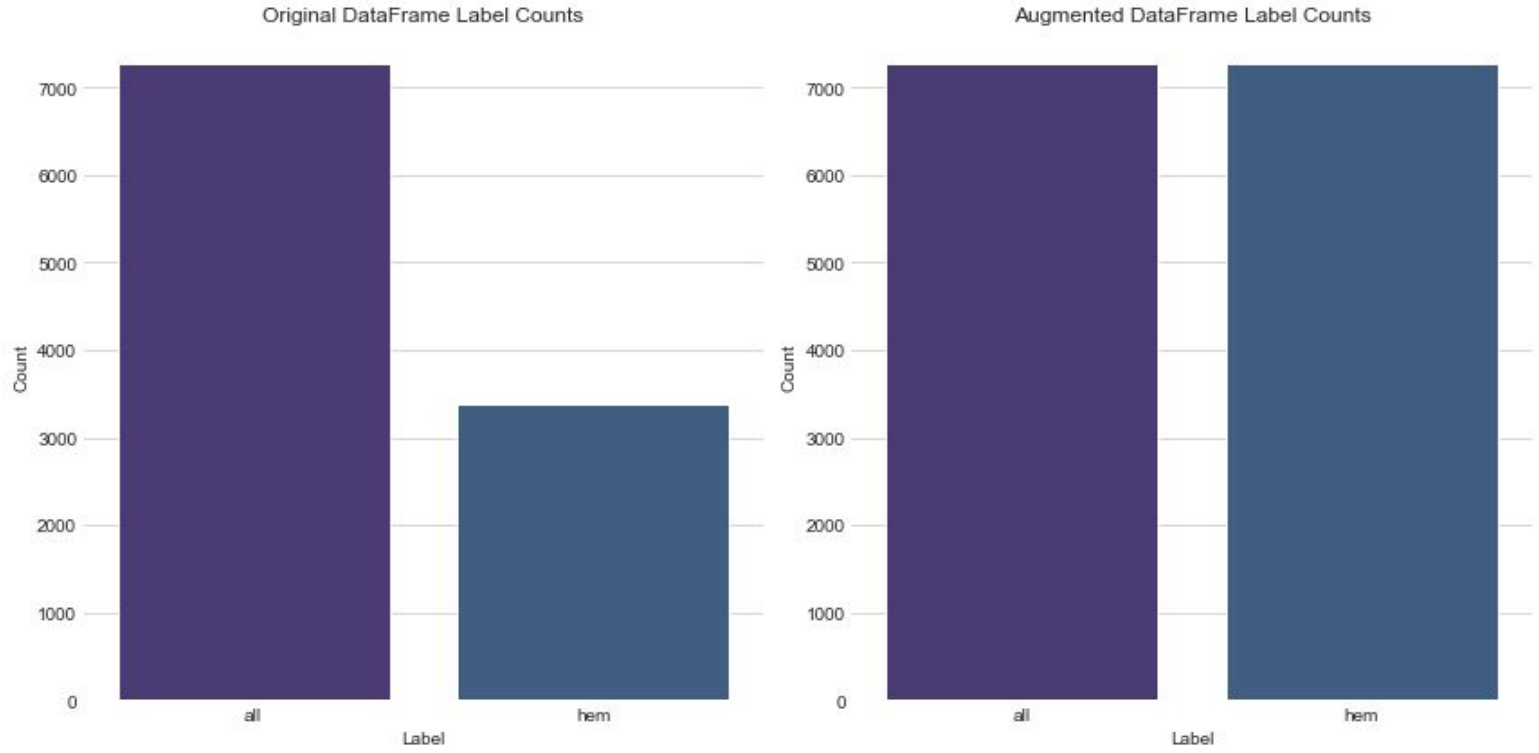


Initial CNN model training and validation accuracy and loss over 20 epochs

INITIAL MODEL REFINEMENT

Data augmentation process:

- New DataFrame featuring normal images randomly rotated, shifted, or flipped.
- Utilized train_test_split from sklearn.model_selection to re-divide the dataset
  - Train: 10,141
  - Validate: 2,536
  - Test: 1,867
- Utilized ImageDataGenerator Settings to augment the train set further with randomly flipping images horizontally and vertically.



DATA AUGMENTATION

AUGMENTED INITIAL CNN MODEL PERFORMANCE

Initial CNN model training and validation accuracy and loss over 20 epochs on augmented dataset

INITIAL MODEL REFINEMENT

Motivation for Hyperparameter Tuning:

- Enhanced Performance: Despite augmented data yielding promising results, recognized potential for further optimization.
- Fine-tuning Objective: Aimed to explore epochs, learning rates, and dropout rates to optimize model performance.

Adjustment Rationale:

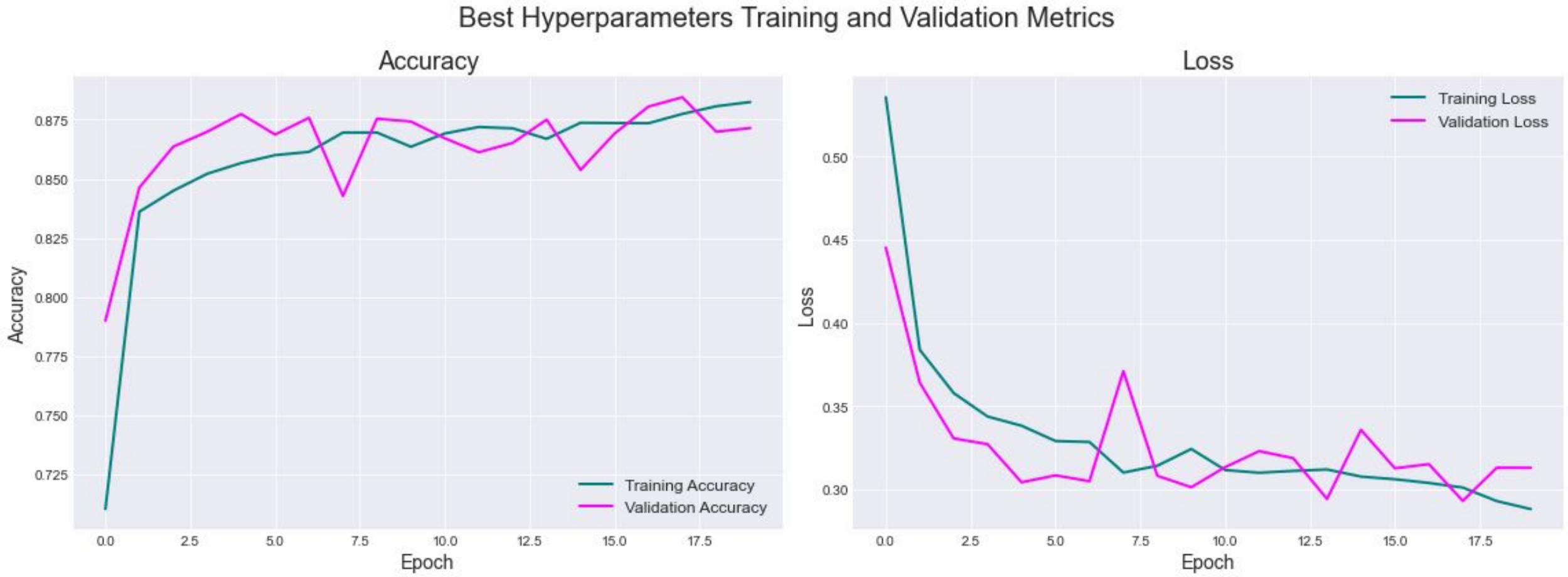
- Epochs: Balance between model complexity and training time to prevent overfitting (epochs tested: [10, 15, 20])
- Learning Rate: Efficient navigation of optimization landscape to converge to optimal solution (learning rates tested: [0.001, 0.01, 0.1])
- Dropout Rate: Regularization to enhance model generalization by preventing over-reliance on specific features (dropout rates tested: [0.0, 0.25, 0.5])

Implementation of Hyperparameter Tuning:

- Grid Search Approach: Utilized grid search for comprehensive exploration of hyperparameter space.
- The best Hyperparameters were found to be: {'dropout_rate': 0.25, 'epochs': 20, 'learning_rate': 0.001}

HYPERPARAMETER GRIDSEARCH

INITIAL MODEL REFINEMENT



Best hyperparameters CNN model training and validation accuracy and loss

HYPERPARAMETER GRIDSEARCH

INITIAL MODEL REFINEMENT

Early Stopping CNN model training and validation accuracy and loss (patience = 3)

EARLY STOPPING

INITIAL MODEL REFINEMENT

Cross-Validation Procedure:

- Utilized K-Fold cross-validation with 5 folds to evaluate model performance comprehensively.
- Cloned the base model for each fold to ensure unbiased evaluation and prevent information leakage.
- Trained and evaluated the model on each fold, storing validation metrics for subsequent analysis.

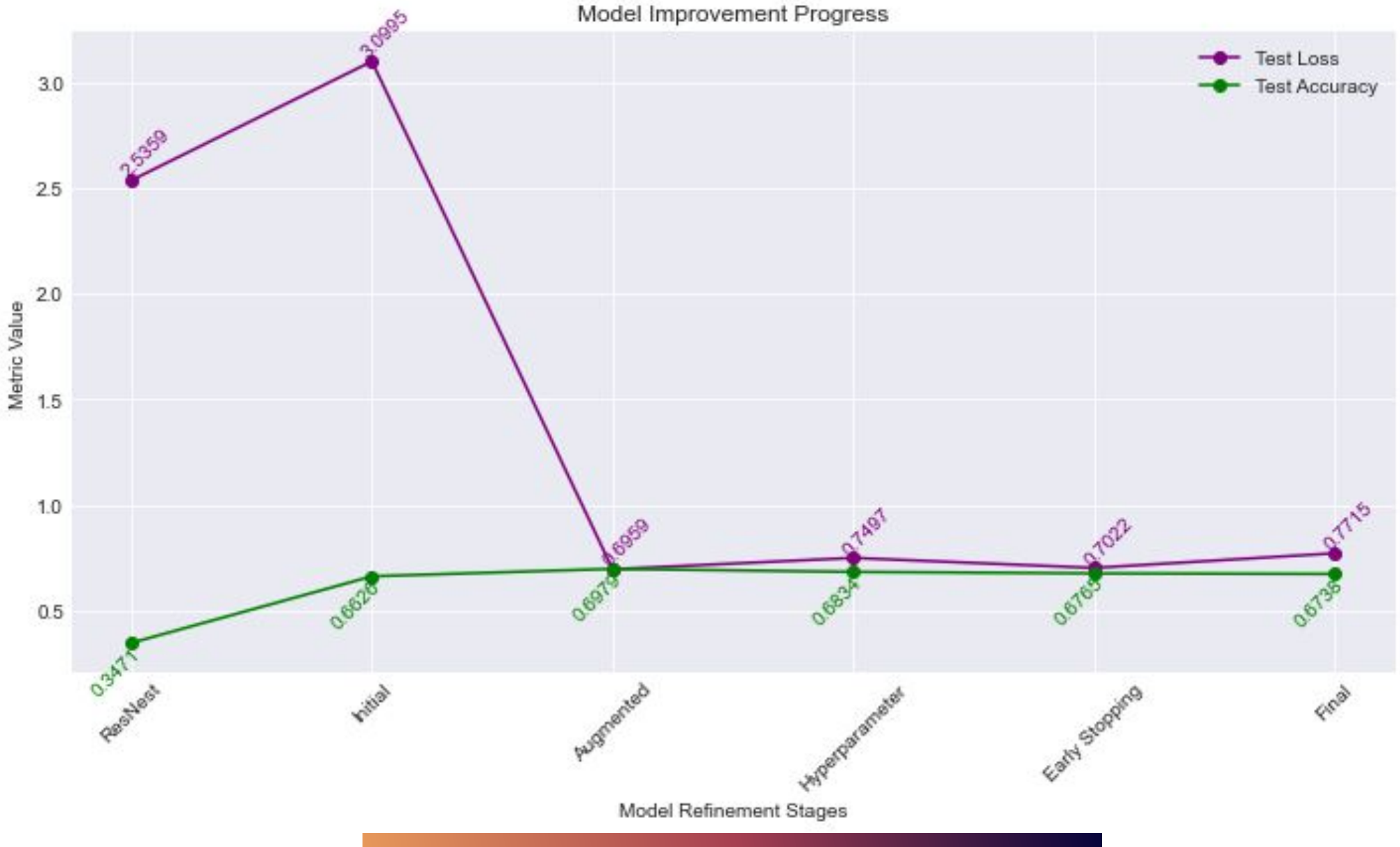
Iterative Model Evaluation:

- Repeated the cross-validation process for all folds to obtain robust and reliable performance estimates.
- Facilitated the identification of stable hyperparameters and validation of model stability.
- Contributed to the iterative refinement of the model, ensuring optimal performance on unseen data.

Average Validation Loss (CV): 0.3080
Average Validation Accuracy (CV): 0.8709

K-FOLDS CROSS VALIDATION

MODEL IMPROVEMENT



FINAL CONCLUSION

"PREDICTING THE FUTURE ISN'T MAGIC, IT'S ARTIFICIAL INTELLIGENCE."
-DAVE WATERS

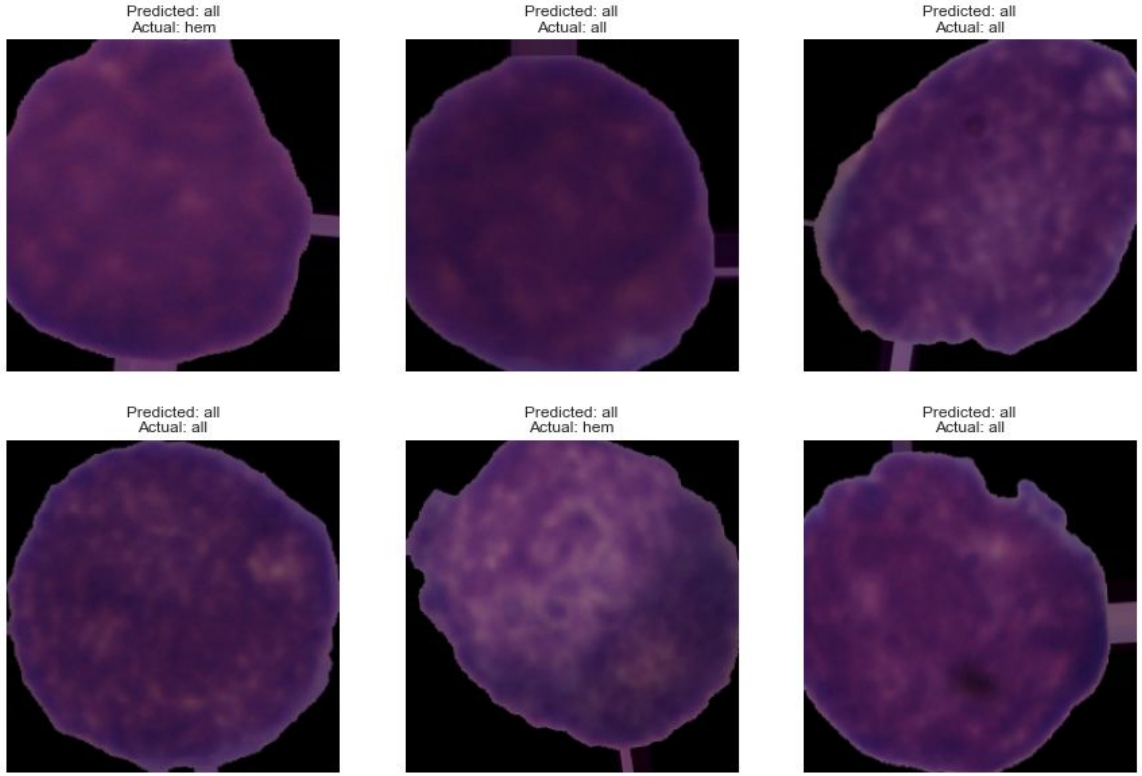
FINAL CONCLUSIONS

Performance Enhancement:

- Our model achieves an accuracy of 0.6738 and a loss of 0.7715, surpassing the ResNet pretrained model significantly.
- Doubled accuracy and reduced loss by 1.95 compared to ResNet's initial metrics.

Significance and Impact:

- Accurately discerns cancerous cells, vital for diagnostics.
- Demonstrates reliability in clinical applications, showcasing machine learning's transformative power in improving patient outcomes.



Generating predictions with final CNN model

THANK YOU

Amy Schneider
https://www.linkedin.com/in/amyschneider/

Alonna Guerrero
https://www.linkedin.com/in/alonna-g