

# Leukemia Classification with Tensorflow

Scott Askinosie  
Varun Reddy  
Cam Smugereski

# Acute lymphocytic/lymphoblastic leukemia (ALL)

- Cancer of the blood or bone marrow
- Progresses rapidly in the blood producing immature white blood cells (lymphocytes)
- Most common in children, 25% of all pediatric cancers, and has high chance for cure if caught early
- Caused by genetic mutation in bone marrow increasing number of immature, non-functional lymphocytes which build up and “crowd out” normal cells
- Identifying ALL can be costly and time consuming without high throughput approaches

# Executive Summary

- Early and accurate detection of cancer is crucial for best chances of successful recovery
- Developed a model that will use image processing tools like Tensorflow to distinguish between normal morphology of cells and cancerous morphology (ALL)
- Decrease time to diagnose – decrease human error
- Model will improve efficacy of cancer diagnoses and will detect cancerous morphology in cells that is inapparent to the human eye

# PROBLEM STATEMENT

This is a binary classification model to identify morphological characteristics of white blood cells to determine if they have normal morphology or morphology that is consistent with a leukemia blast. In general, the task of identifying immature leukemic blasts from normal cells under the microscope is challenging because morphologically the images of the two cells appear similar.

The ideal outcome is to have an accuracy score of above 80%. We also aim to minimize Type II errors to ensure that samples that contain cells with ALL morphology are not missed by our model.



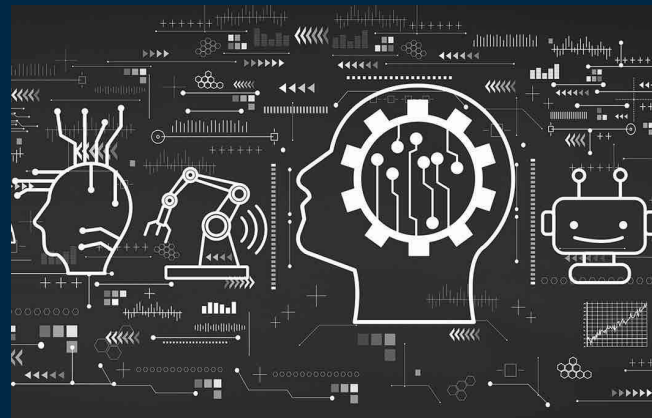
# DATA

- Cancer Imaging Archive
- Total subjects: 73
  - Cancer: 47, Normal: 26
- Images used: 3,537
  - Cancer: 2822, Normal: 705
- Null Model
  - Cancerous (ALL) 68%
  - Normal (HEM) 32%



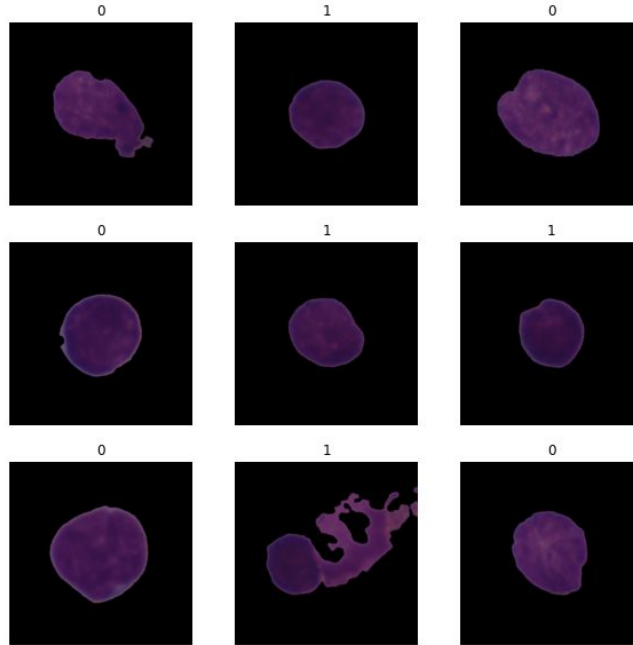
# OUR MODELS

- Neural Networks
  - Convolutional Neural Network (CNN)
  - Transfer Model (EfficientNetB0)
- Classification Models
  - k Nearest Neighbors (kNN)
  - Naive Bayes
  - Random Forest
  - Logistic Regression
  - ADA Boost (Random Forest)
  - Stacked



# Visualizing Cancerous Cells (ALL) Vs Healthy Cells (HEM)

ALL (0) Vs HEM (1) Cells

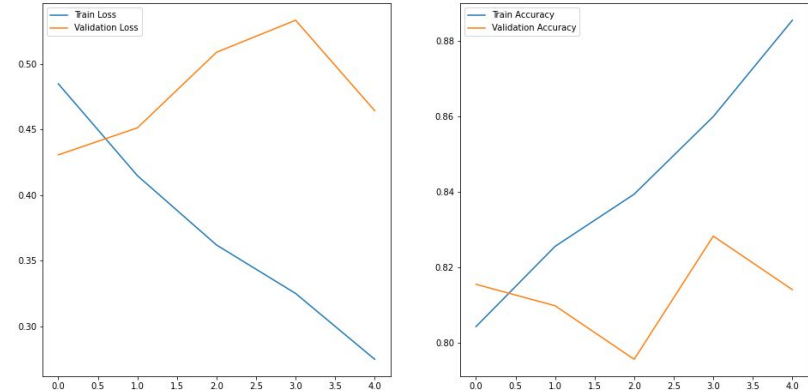


# CNN Base Model on Original Cell Images

## No Augmentation

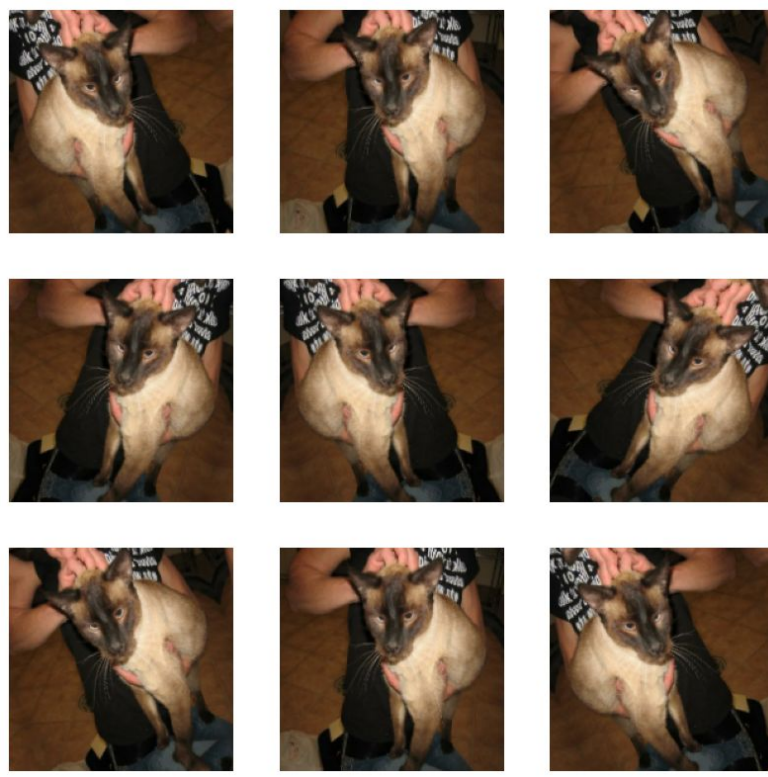
	Train	Test
Loss	0.26	0.54
Accuracy	89%	80%

Loss and Accuracy for Train Vs Validation data



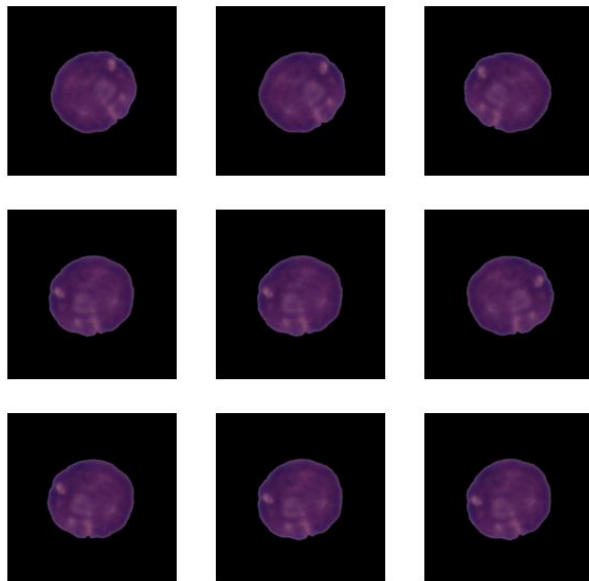


# Image Augmentation



# Image Augmentation on Cell Images

Leukocyte With Augmentation Applied



# CNN Base Model on Augmented Images

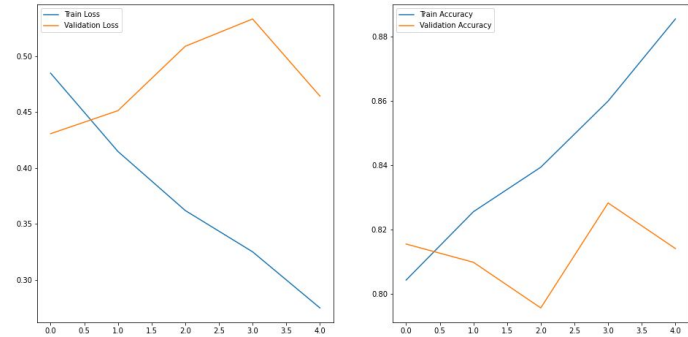
## No Augmentation

	Train	Test
Loss	0.26	0.54
Accuracy	89%	80%

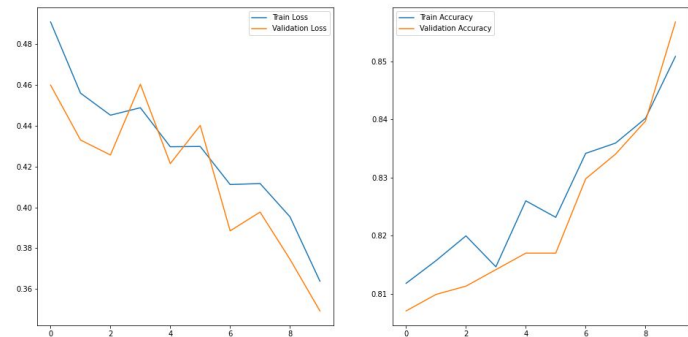
## With Augmentation

	Train	Test
Loss	0.30	0.28
Accuracy	88%	89%

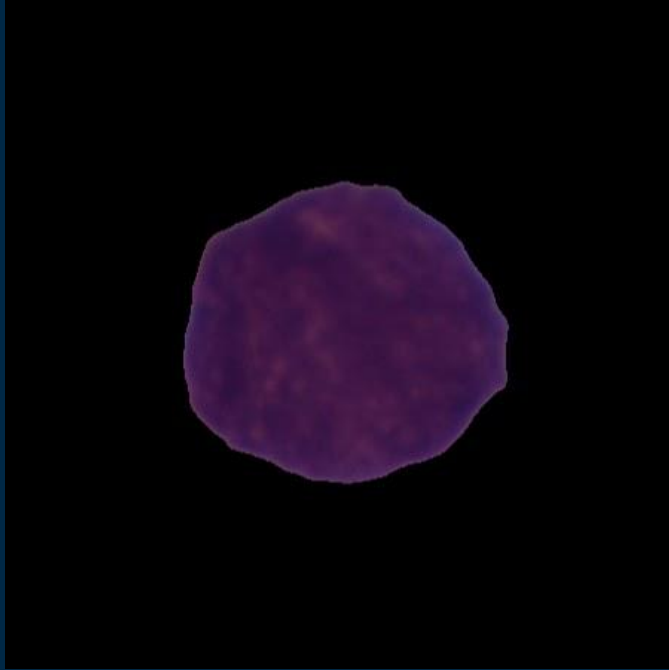
Loss and Accuracy for Train Vs Validation data



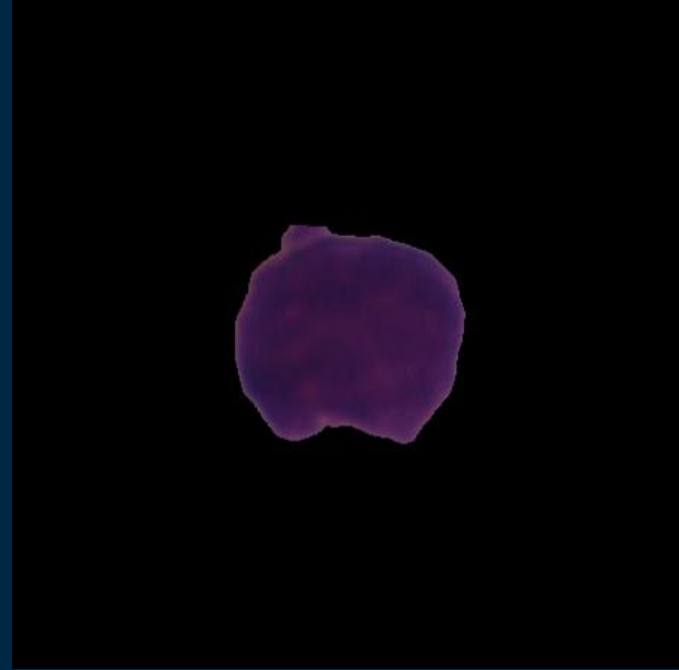
Loss and Accuracy for Train Vs Validation data



# Testing CNN Model on Additional Images

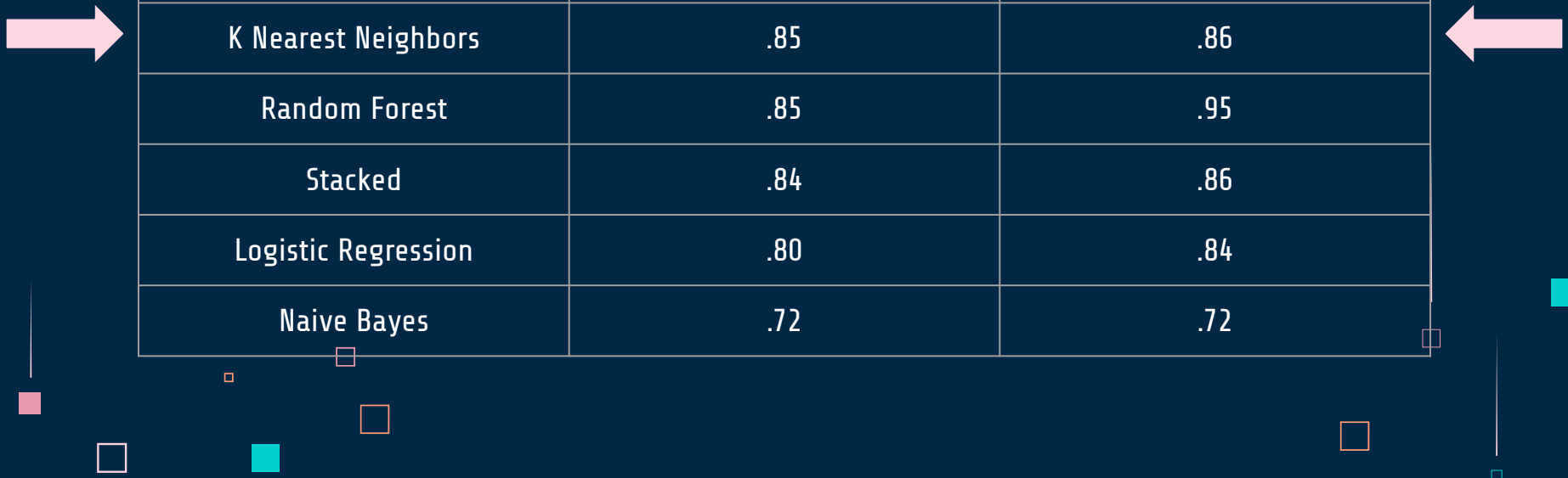


The CNN model is over 95%  
certain this is a cancerous cell



The CNN model is over 99%  
certain this is a normal cell

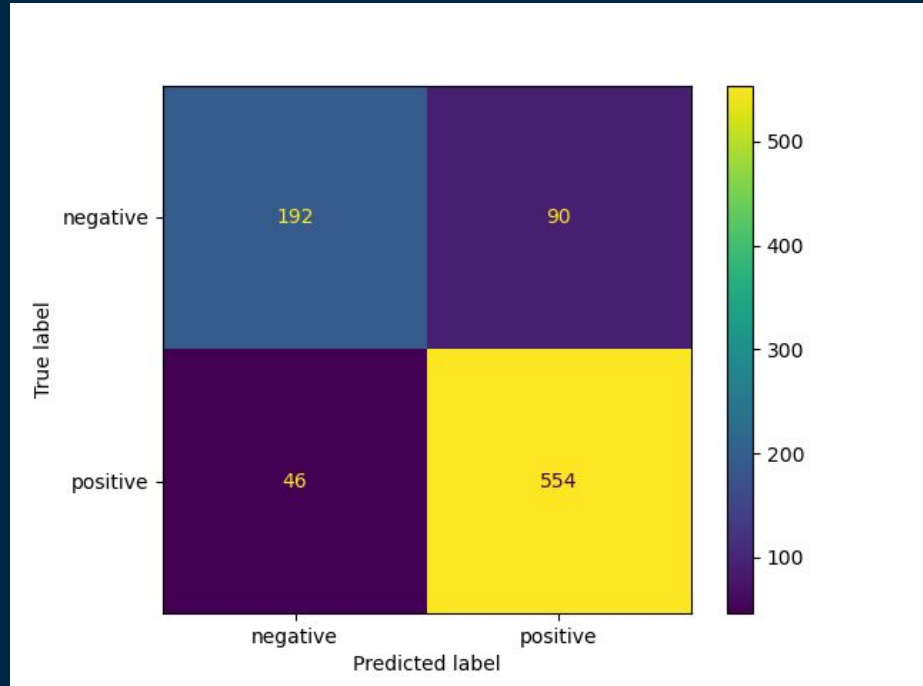
# Classification Models



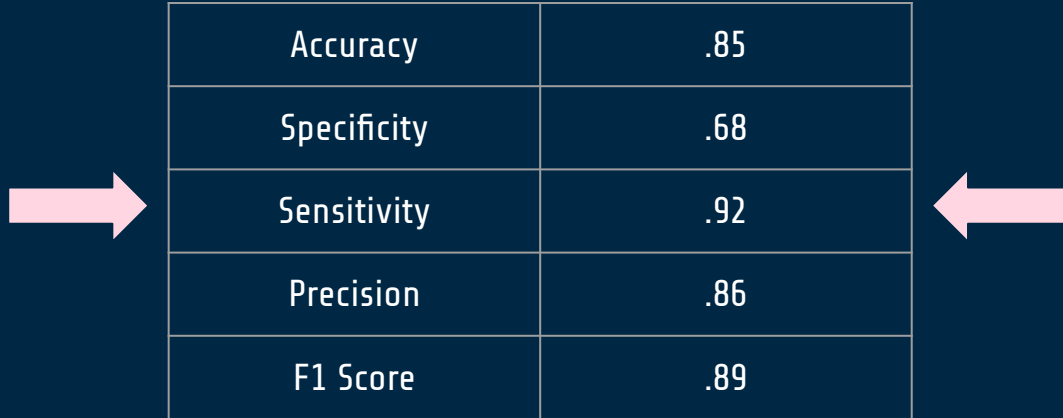
Model	Test Accuracy	Train Accuracy
Ada Boost (RF)	.86	.98
K Nearest Neighbors	.85	.86
Random Forest	.85	.95
Stacked	.84	.86
Logistic Regression	.80	.84
Naive Bayes	.72	.72

# kNN Confusion Matrix

Type II (false negatives) errors : 5.3%



# Model Analysis Values – kNN



Accuracy	.85
Specificity	.68
Sensitivity	.92
Precision	.86
F1 Score	.89

# Conclusions and Next Steps

Exceeded our goal of 80% accuracy.

The highest performing model is the CNN model with image augmentation which produced a 89% test accuracy score and 88% train accuracy score.

Next steps:

- Optimizing Type 2 errors with the CNN model

