Malaria Detection

MIT-ADS Program
Capstone Project

Trush Majmudar 10/18/23

Problem Statement:

Malaria is detected by identifying parasitic cells in histopathology slides.

So far, this process has been manual, even with image processing softwares

This is tedious, time consuming, and costly process prone to misclassification.

Automating this process with high accuracy, low return times, and low costs is desirable.

We plan to solve this problem using Deep Learning methods.

Solution Approach: Convolutional Neural Networks

Data: Labeled RGB images of infected and uninfected cells.

Variety of shapes, colors noticeable.

Distinguishing features in parasitized cells are the dark blobs inside the cell body.

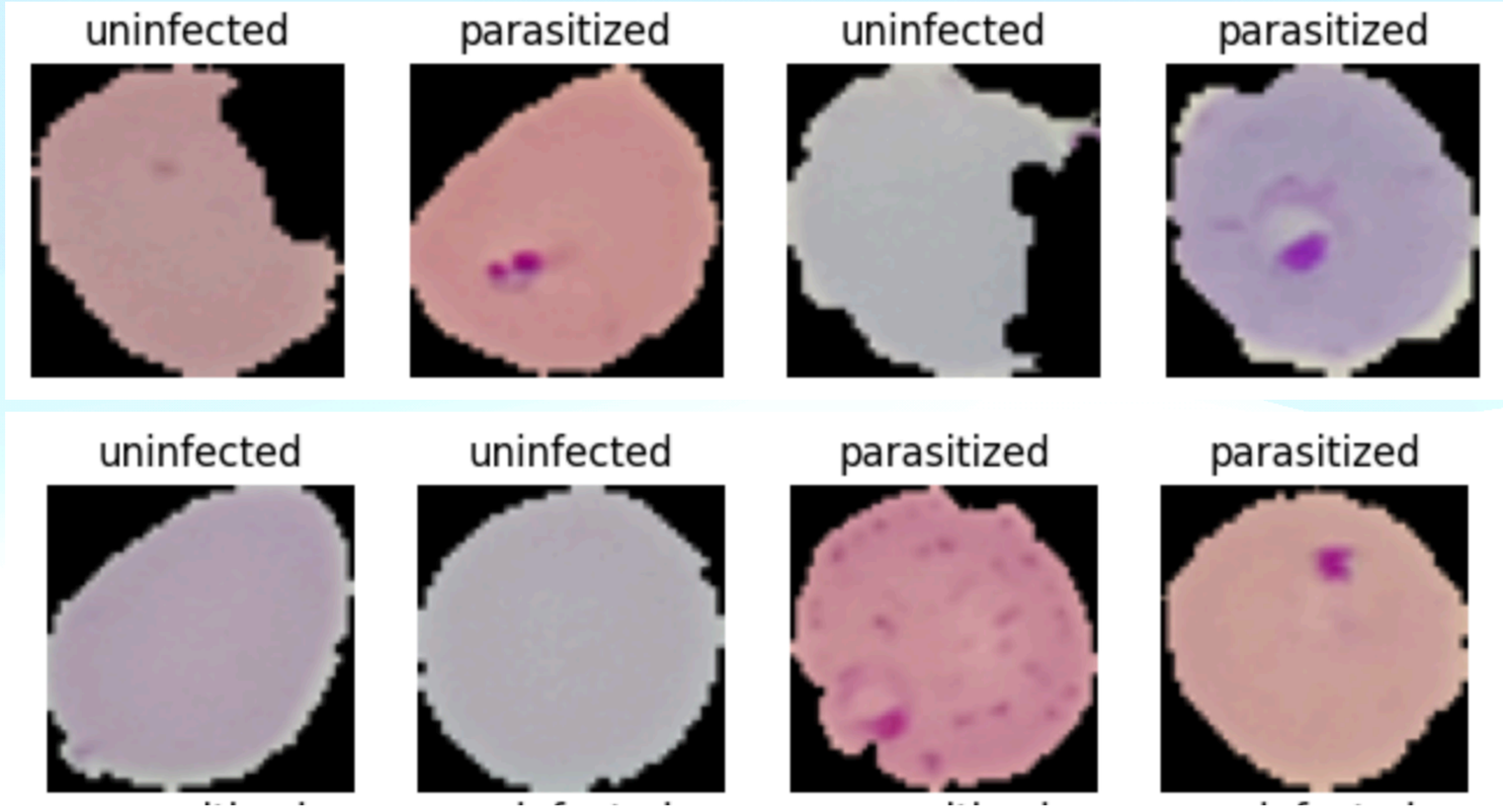
Convolutional Neural Networks have been known to work very well on classifying images and object recognition problems.

We will examine:

- Image Pre-Processing

- Data Augmentation

- Different architectures, regularization schemes, activation functions, and Transfer Learning.

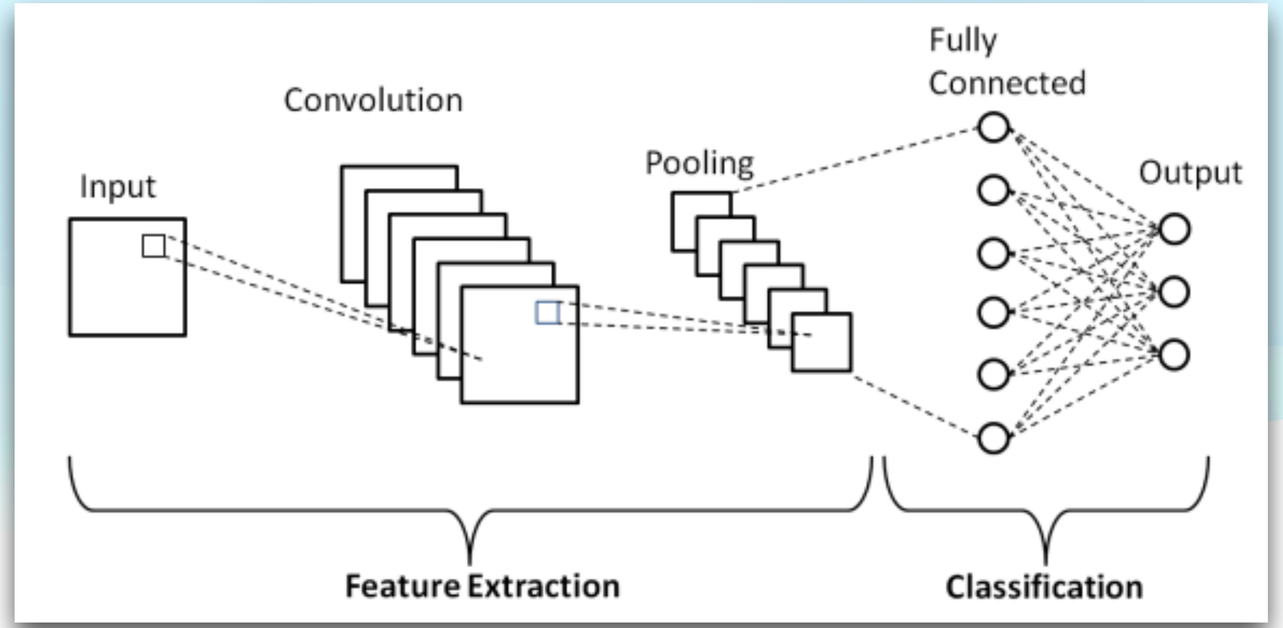


Training Images: 24,958 split 80/20 between training and validation

Testing Images: 2600

Basic Convolutional Neural Network Architecture

Typical Scheme:
Binary classification will have two output neurons at the end—
Uninfected and Parasitized



Convolutional Layers pick out features

Pooling Layers reduce the number of parameters

Dropout layers switch off some neurons to reduce overfitting

Batch Normalization is a regularization method to ensure “in-distribution” normalized outputs

- Fully connected dense layers do the classification. Final output layer has as many neurons as there are classes.

Figure from: Van Hiep Phung and Eun Joo Rhee, MDPI Applied Sciences, 2019

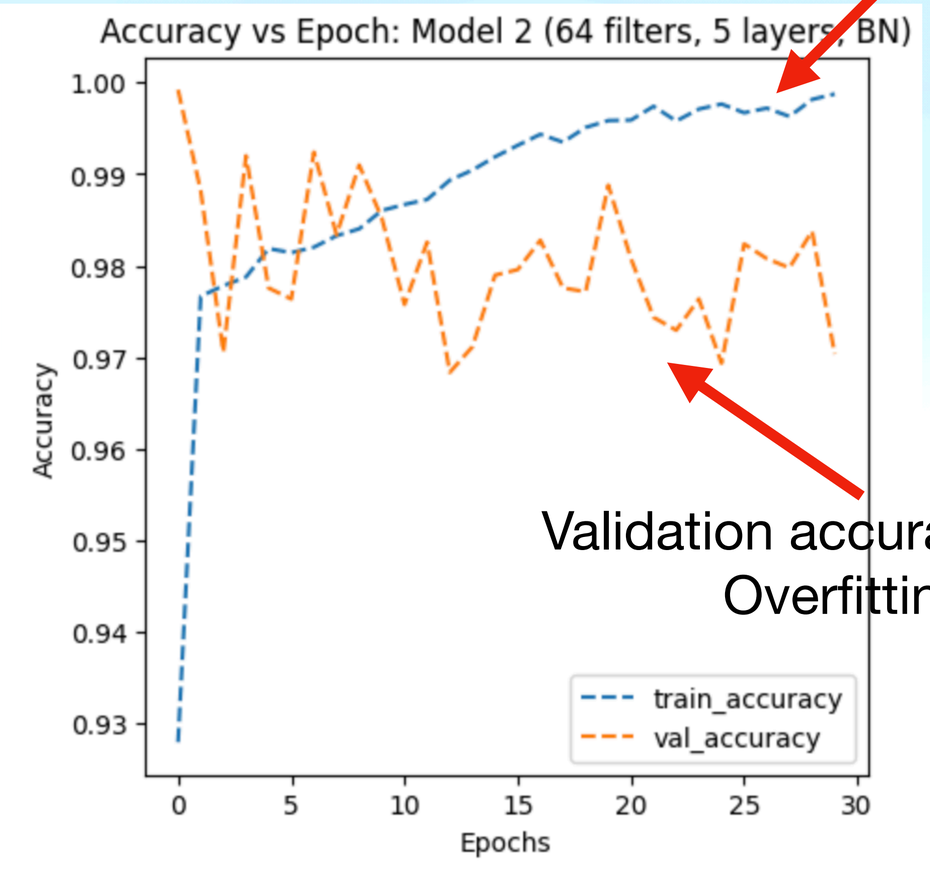
Models Comparison:

Model Architecture	Training Accuracy	Validation Accuracy	Testing Accuracy	Average Precision	Average Recall
Basic CNN: 3 layers: 32 filters each, Max pooling, dropout 512 filter Dense layer	0.981	0.983	0.984	0.98	0.98
Model 1: 5 layers: 32 filters each, Max pooling, dropout 512 filter Dense layer	0.975	0.977	0.977	0.98	0.98
Model 2: 5 layers: 64 filters each, Batch Normalization 512 filter Dense layer, Dropout at the end	0.99	0.97	0.978	0.98	0.98
Model 3: Data Augmentation, (32) 5 layers, Leaky ReLU Batch Normalization, 512 filter Dense layer, Dropout	0.977	0.982	0.978	0.98	0.98
Model 4: Transfer Learning VGG16 + 3 Dense Layers: 256, 128, 64	0.935	0.955	0.957	0.96	0.96

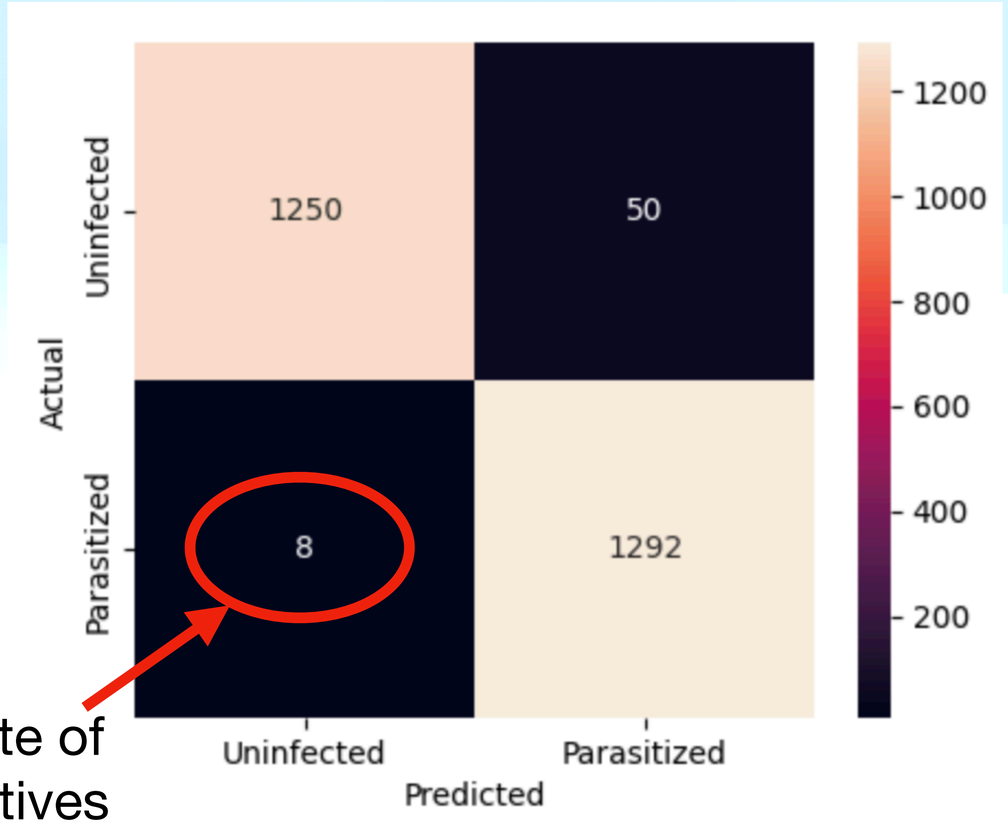
It appears as if most models are performing more or less the same, but…
we have to see the Accuracy vs. Epochs curves and Confusion Matrix

Example: Model 2 (not chosen)

still increasing…
still trainable…..



Validation accuracy dropping…
Overfitting issue

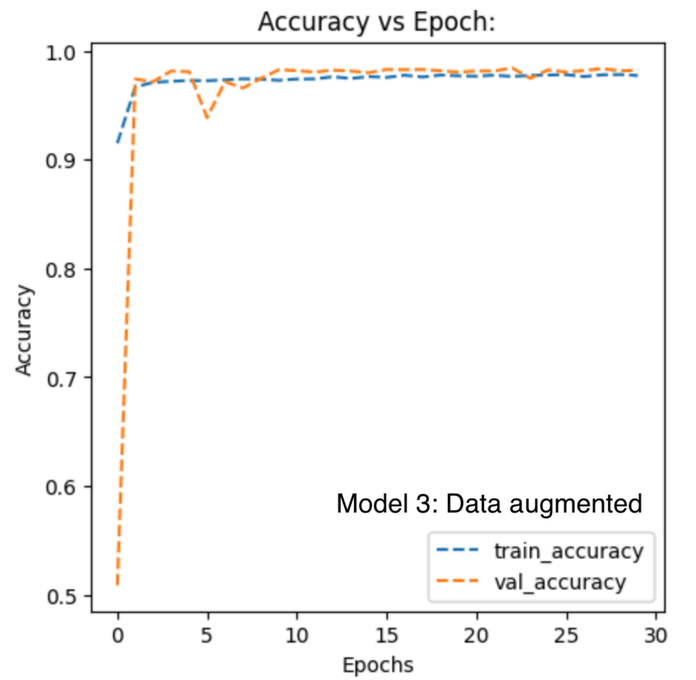


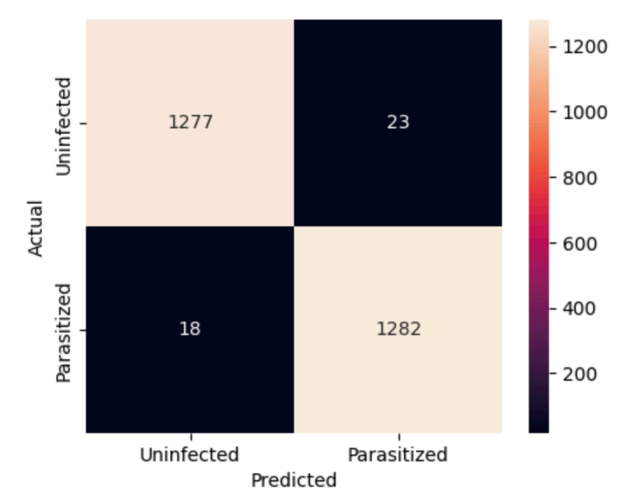
Very low rate of
False Negatives

Chosen Model Architecture
Model 3 (with Data Augmentation and Batch Normalization)

Input: Data Augmented images (rotation, horizontal+vertical flip, shear, zoom)
32 Conv+Leaky ReLU+Max Pooling+ Batch Normalization
32 Conv+Leaky ReLU+Max Pooling+ Batch Normalization
32 Conv+Leaky ReLU+Max Pooling+ Batch Normalization
32 Conv+Leaky ReLU+Max Pooling+ Batch Normalization
32 Conv+Leaky ReLU+Max Pooling+ Batch Normalization
Flatten+Dense (512)+Leaky ReLU +Dropout
Dense (2) +Softmax Activation
Trainable Parameters: 105282

Chosen Model Architecture:





Salient Features and Advantages:

★ Both curves have stabilized to flat curves.

★ Perfect matching between training and validation accuracy.

★ Optimal fitting— no overfitting or under-fitting visible.

★ Low and similar numbers for False Negatives and False Positives.

This model is trained on a more realistic data that captures the natural variation and noise in data collection.

It is better suited to be applicable in real situations.

Proposed Business Solution

Mid-to-Large Enterprise, Clinical Labs, Pharma, Hospitals

Two-tiered Approach:

A well-benchmarked and trained model like Model 3 for quick assessment and…

Fine-Tuned Transfer Learning (TL) Model for benchmarked solution.

Drawback: TL models are trained on very different kinds of images. Some general features may be detected but not medically relevant details.

Field Medics, Small Enterprise

- Need low costs, low data and connectivity requirements
- Need quick but accurate prediction
- On site or in-house assessment
- Mobile app or Web-Browser based Analysis

Solution:

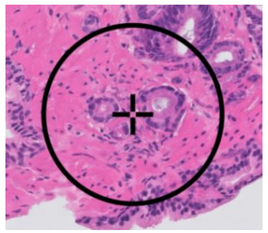
Optimized version of Model 3

Trained on Data Augmented images to reflect realistic situations.

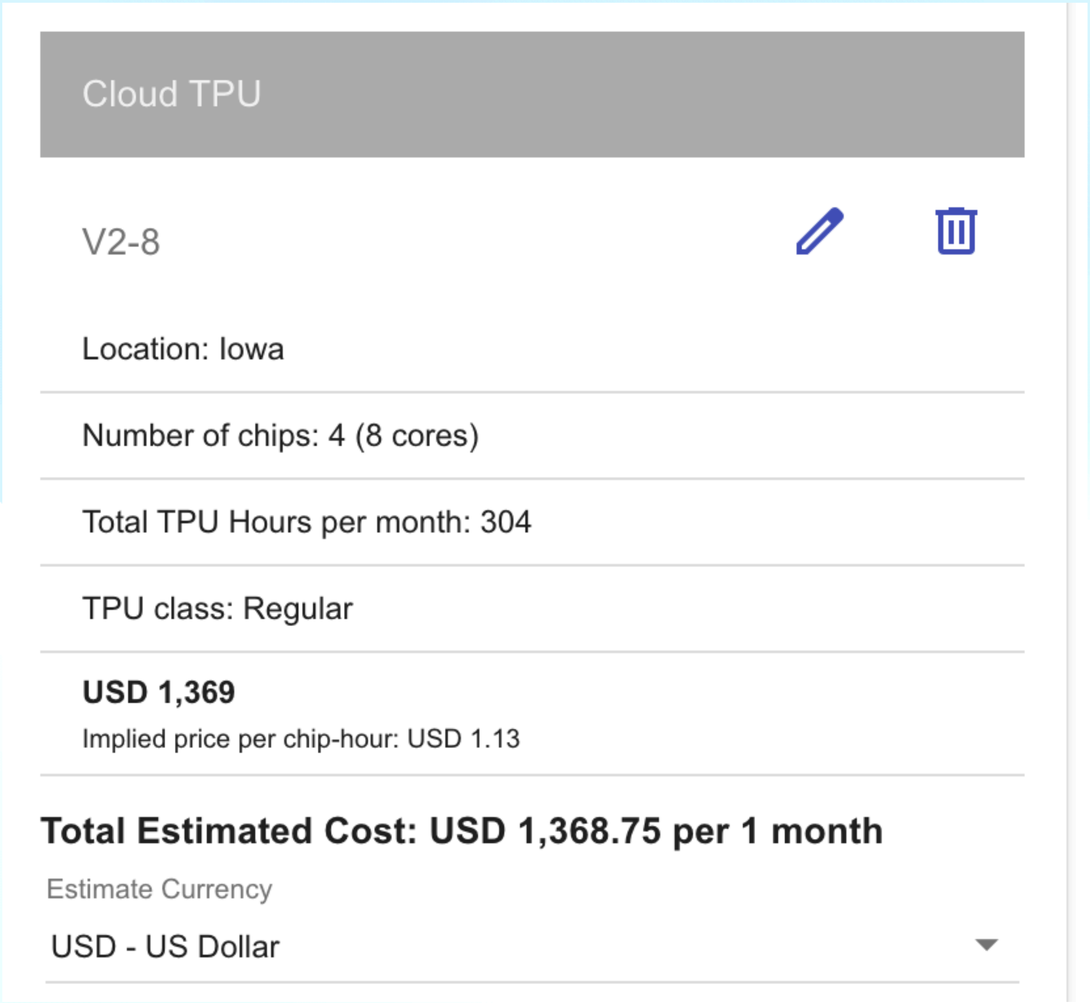
Wait for…

Microsoft and Paige Announce Collaboration on Image-based AI Model in Fight Against Cancer

Paige, a provider of end-to-end digital pathology solutions and clinical AI, has announced a collaboration with Microsoft to build what they refer to as the world's largest image-based AI models for digital pathology and oncology.



A hint on the costs…



- A detailed cost/benefit analysis is essential before any implementation…
- A glimpse at just one of the associated costs reveals that cloud computing solutions are also expensive.
- Recall that malaria is prevalent in developing nations with minimal on-field resources.
- In such situations, an in-house executable model using free open source platforms is ideal.

Executive Summary

❖ Convolutional Neural Networks and Deep Learning methods can accurately predict infected and uninfected malaria cells from histology images.

❖ Automation of this process is feasible and gives very good accuracy at a relatively low turn-around time and low cost.

Solution to be adapted will depend on the scale of the enterprise and other constraints.

For larger organizations with computing infrastructure, data storage and handling pipelines, and good cloud computing platforms, a two-tiered approach is more suitable.
- a combined Model 3 type assessment and a future Transfer Learning model, when available.

For smaller setups and on-field situations, a fast, low cost, low resource requirement solution like mobile or web-based platform that implements optimized Model 3 is the preferred solution.

Thank you!

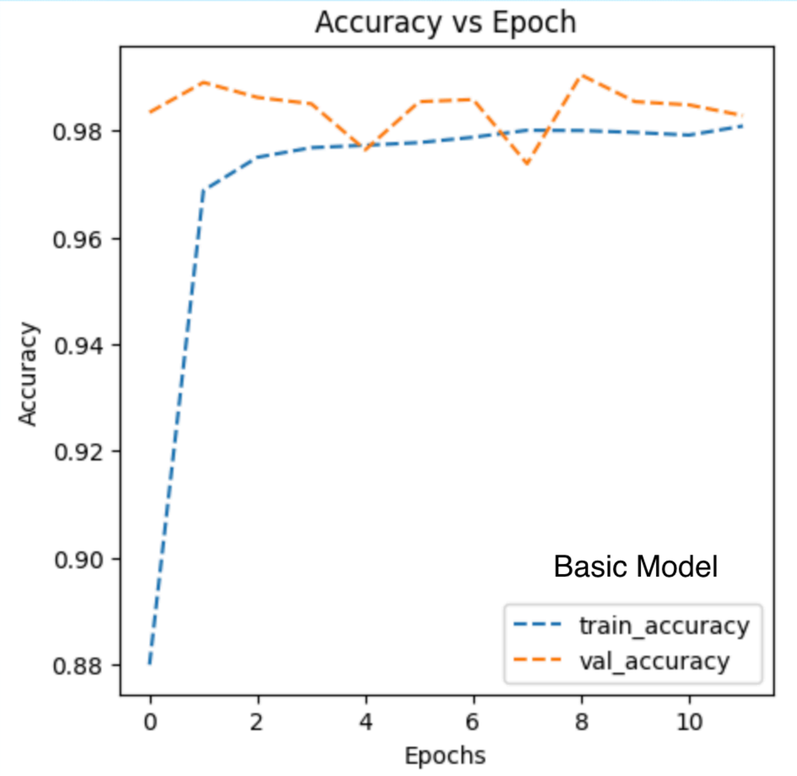
I would like to thank

- The Great Learning Program Staff
- The Instructors
- Industry Mentors and their Mentoring Sessions
- My Cohort for this course

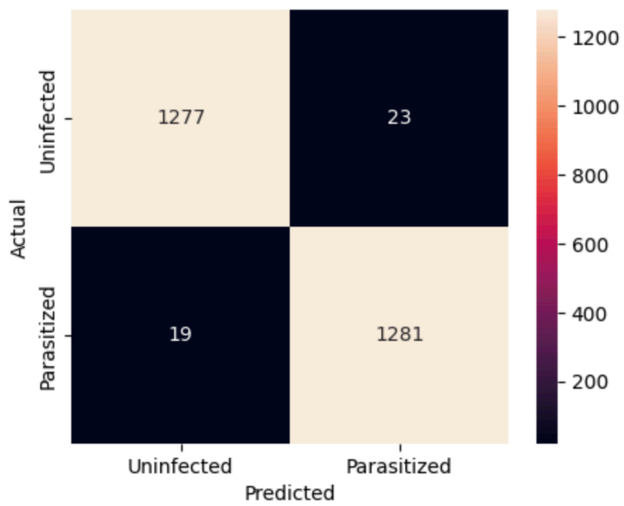
Appendix

Basic Model:

Three 32 neuron CNN with dropout and max pooling



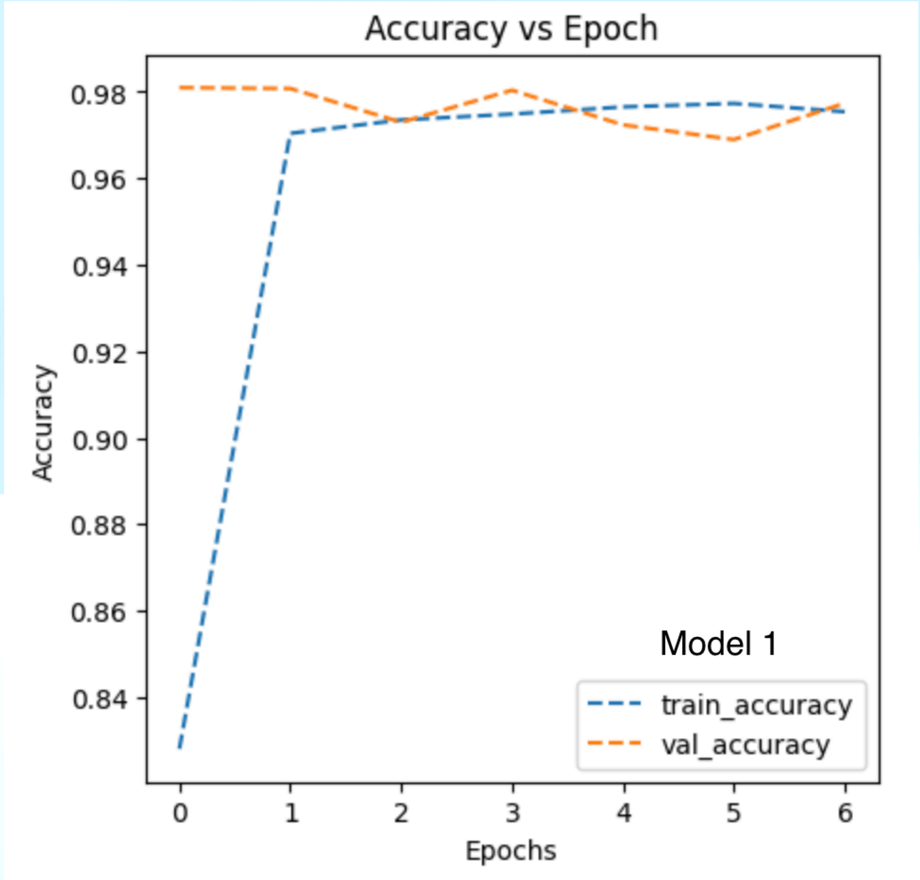
	precision	recall	f1-score	support
0	0.99	0.98	0.98	1300
1	0.98	0.99	0.98	1300
accuracy			0.98	2600
macro avg	0.98	0.98	0.98	2600
weighted avg	0.98	0.98	0.98	2600



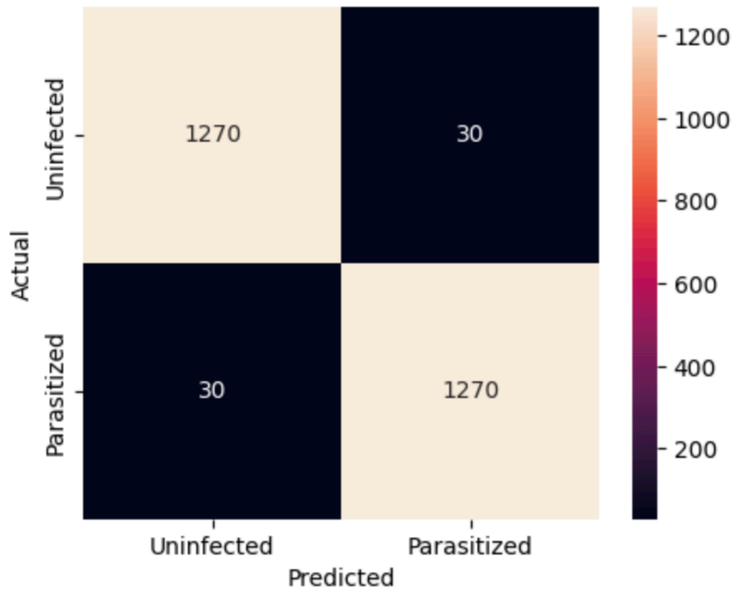
Numbers are good but validation accuracy consistently higher than training accuracy => Undercutting

Model 1:

5 layers, 32 neurons each, max pooling, dropout



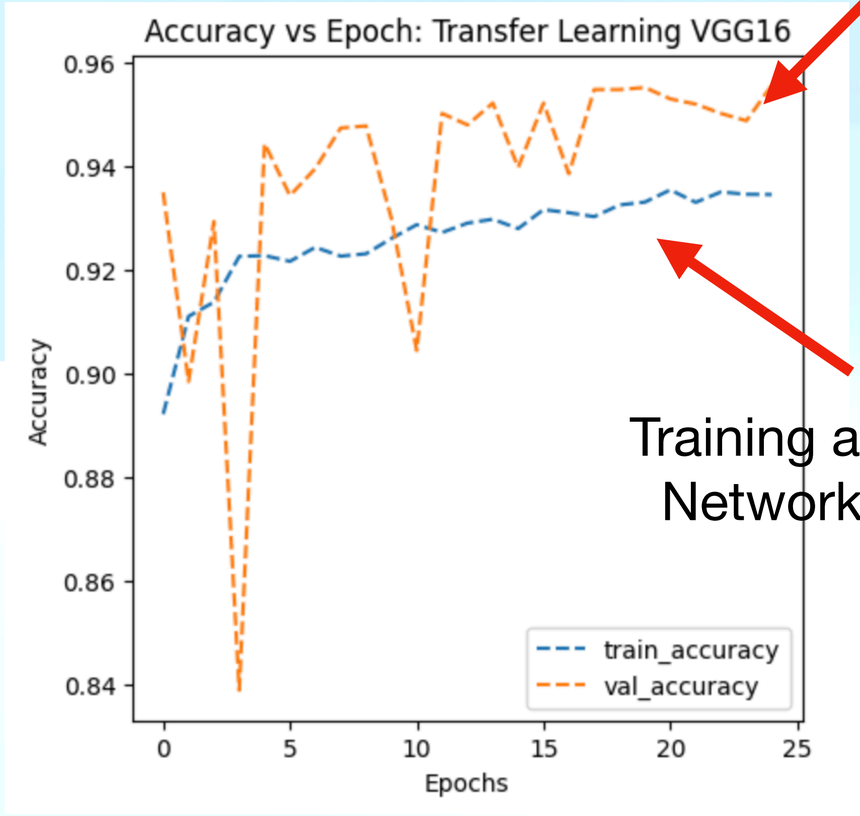
0	0.98	0.98	0.98	1300
1	0.98	0.98	0.98	1300
accuracy			0.98	2600
macro avg	0.98	0.98	0.98	2600
weighted avg	0.98	0.98	0.98	2600



Higher rate of
False Negatives

Transfer Learning: VGG16

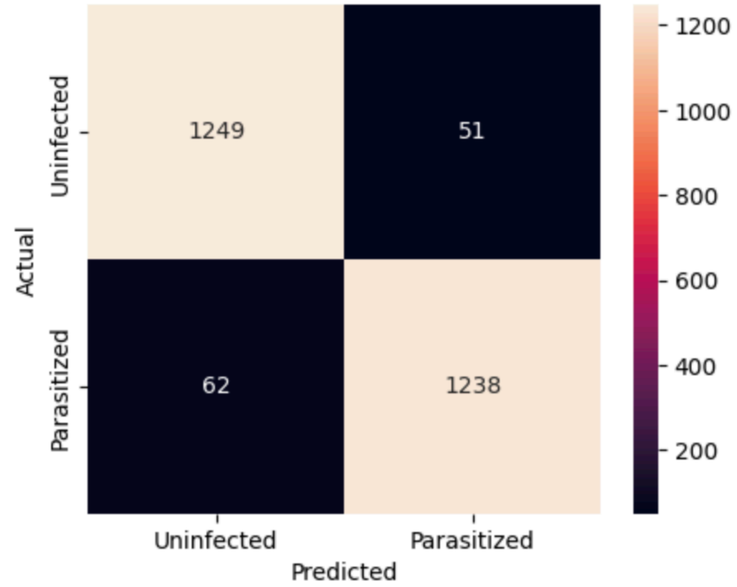
3 Dense layers, 256, 128, 64 neurons each, batch norm., dropout



Validation accuracy fluctuating and higher than training… Undercutting

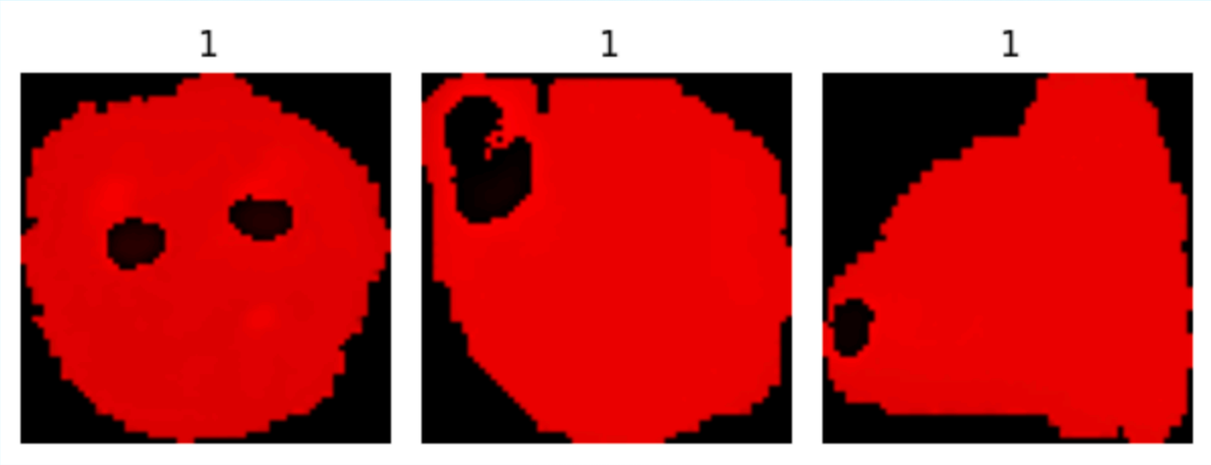
Training accuracy rising… Network is still learning

	precision	recall	f1-score	support
0	0.95	0.96	0.96	1300
1	0.96	0.95	0.96	1300
accuracy			0.96	2600
macro avg	0.96	0.96	0.96	2600
weighted avg	0.96	0.96	0.96	2600

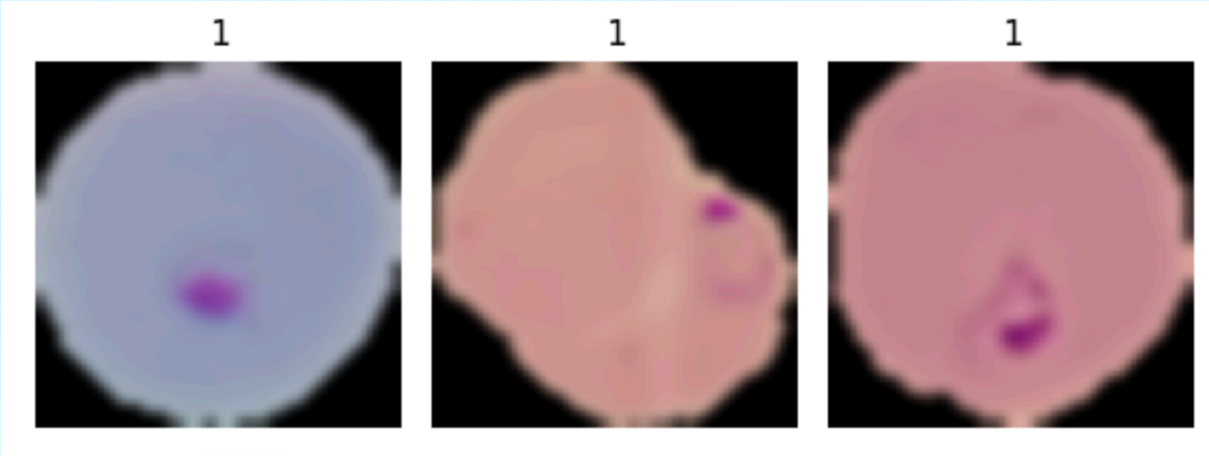


Highest rate of False Negatives and lowest accuracy

Image Pre-Processing

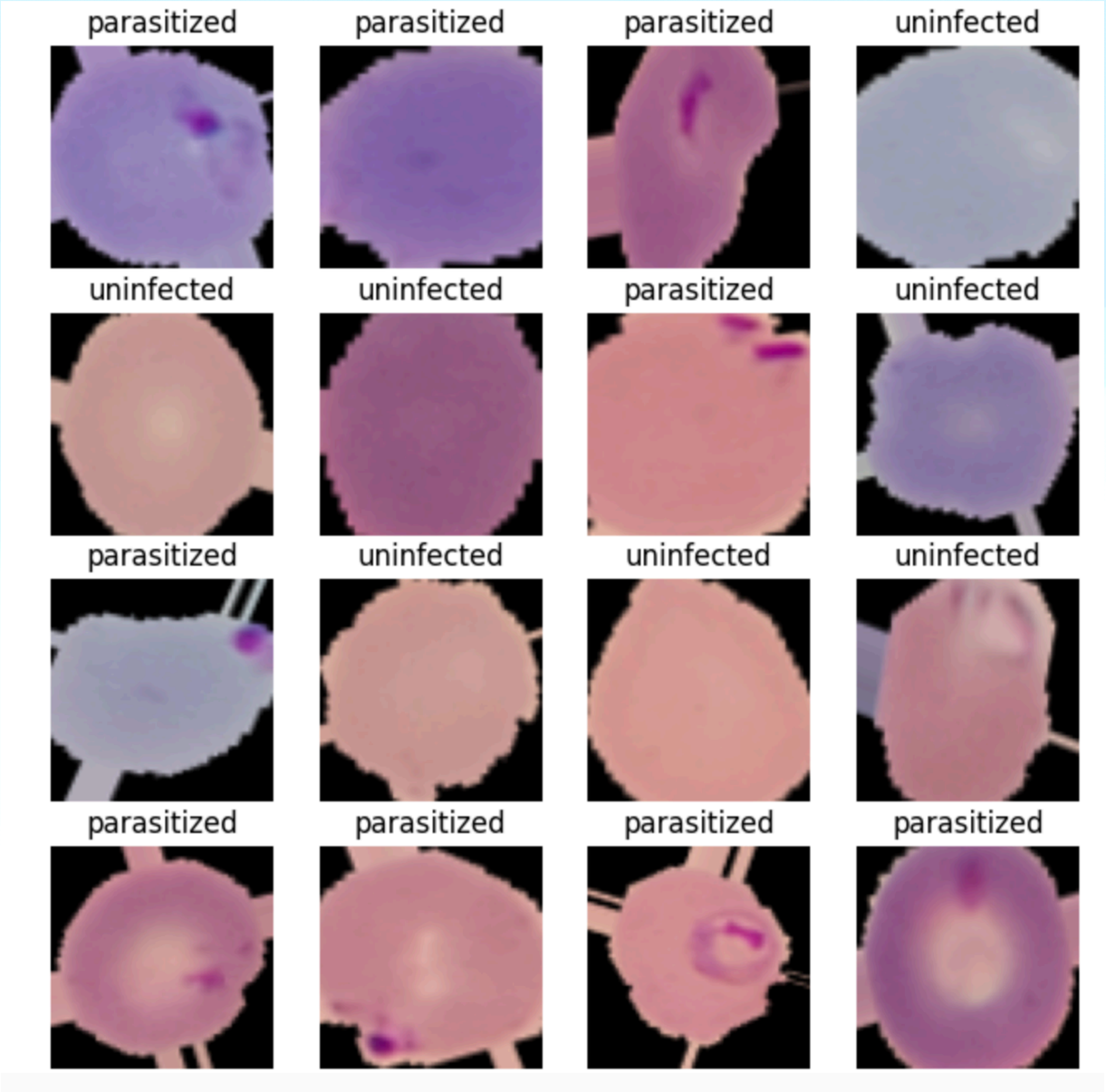


HSV Conversion:
Not sure if it helps.
I know that HSV produces smoother pallet,
which helps with gradient calculations but
I have to investigate how touse HSV effectively.



Gaussian Blurring:
it helps to remove salt-pepper noise
better gradients
diminishes distinctive features of parasitized cells.

Data Augmentation



Horizontal flips
Vertical flips
Rotations
Shearing
Zoom