DIAGNOSING MALARIA IN CELL IMAGES

Thomas Friss

DATASET

- https://www.kaggle.com/iarunava/cell-images-for-detecting-malaria
- A collection of 27,558 images of cells, half infected with malaria(parasitized) and half are clean.
- The infected and clean cell images are in separate folders.
- Each folder has 13,780 files, all but 1 in the folders are images. The one non-image file is a .db file.
- Total memory size is 335 mb.

OBJECTIVE

- The goal was to create a network that can automatically diagnose images of cells to determine if they have been infected with malaria.
- Malaria diagnosis still relies on looking at images of blood cells. Any tool that can automate this can decrease the burden on medical staff, especially if someone is operating in an area with a significant presence of malaria.
- https://www.cdc.gov/malaria/diagnosis_treatment/diagnostic_tools.html

SOLUTION APPROACH

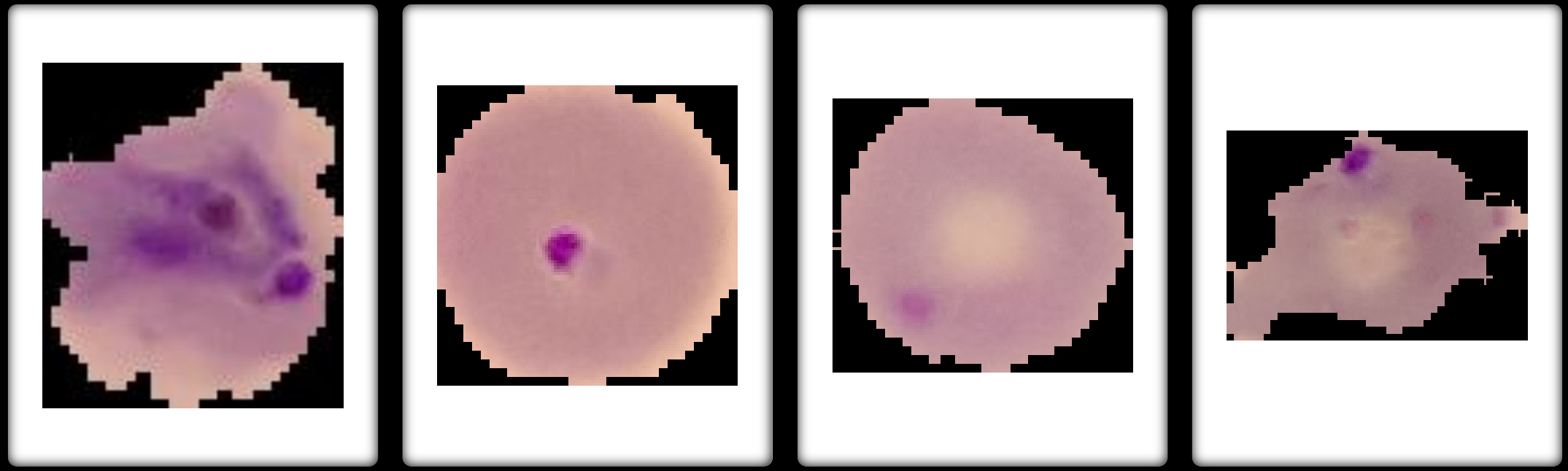
- Because the data consists of images a convolutional neural network was used.
- This is also a binary classification problem. The two categories being infected and uninfected.
- The sample consists of a 50/50 split between the two categories which simplifies model creation.

MODEL EVALUATION

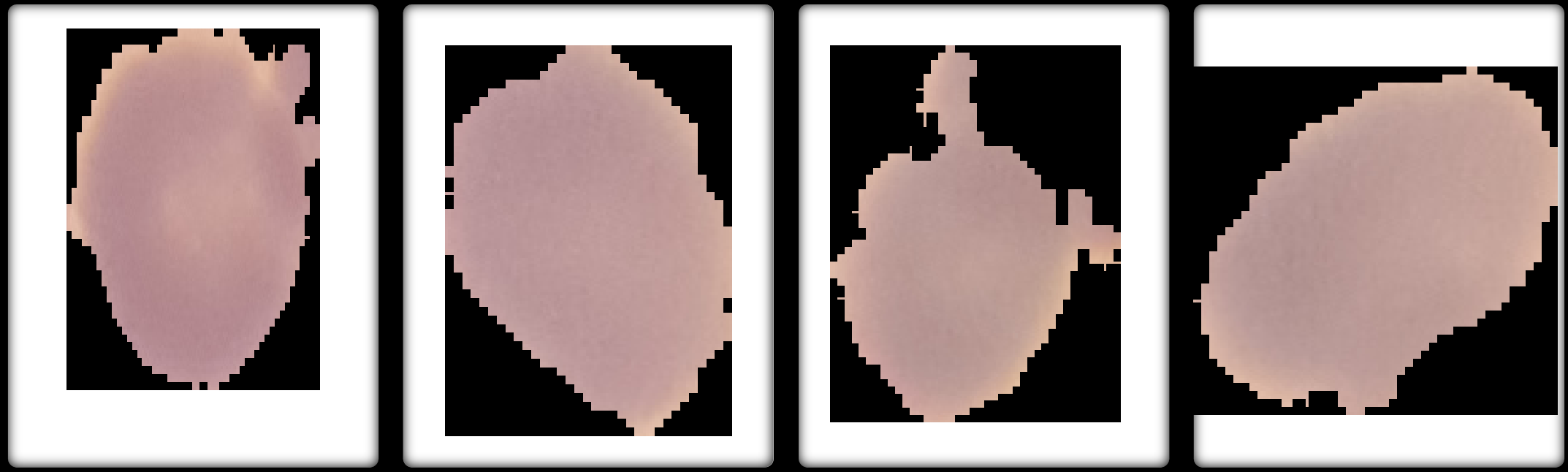
- There will be multiple metrics used but the primary metric for evaluation will be sensitivity, often referred to as recall
- Accuracy and precision will also be measured but not the focus of our evaluations.
- In diagnosing medical conditions our priority will be to avoid false negatives as much as possible.

PREPROCESSING AND EDA

- Due to the data consisting entirely of labeled images the initial exploratory data analysis consisted of visually examining some of the images to gain an intuitive understanding of the differences between clean and infected cell images.
- In these images the signs of infection often consisted of purple circles on the cell image.
- Many of the images have different pixel dimensions so all images were turned into 50,50 pixel images when imported into the model.



EXAMPLES OF INFECTED CELLS

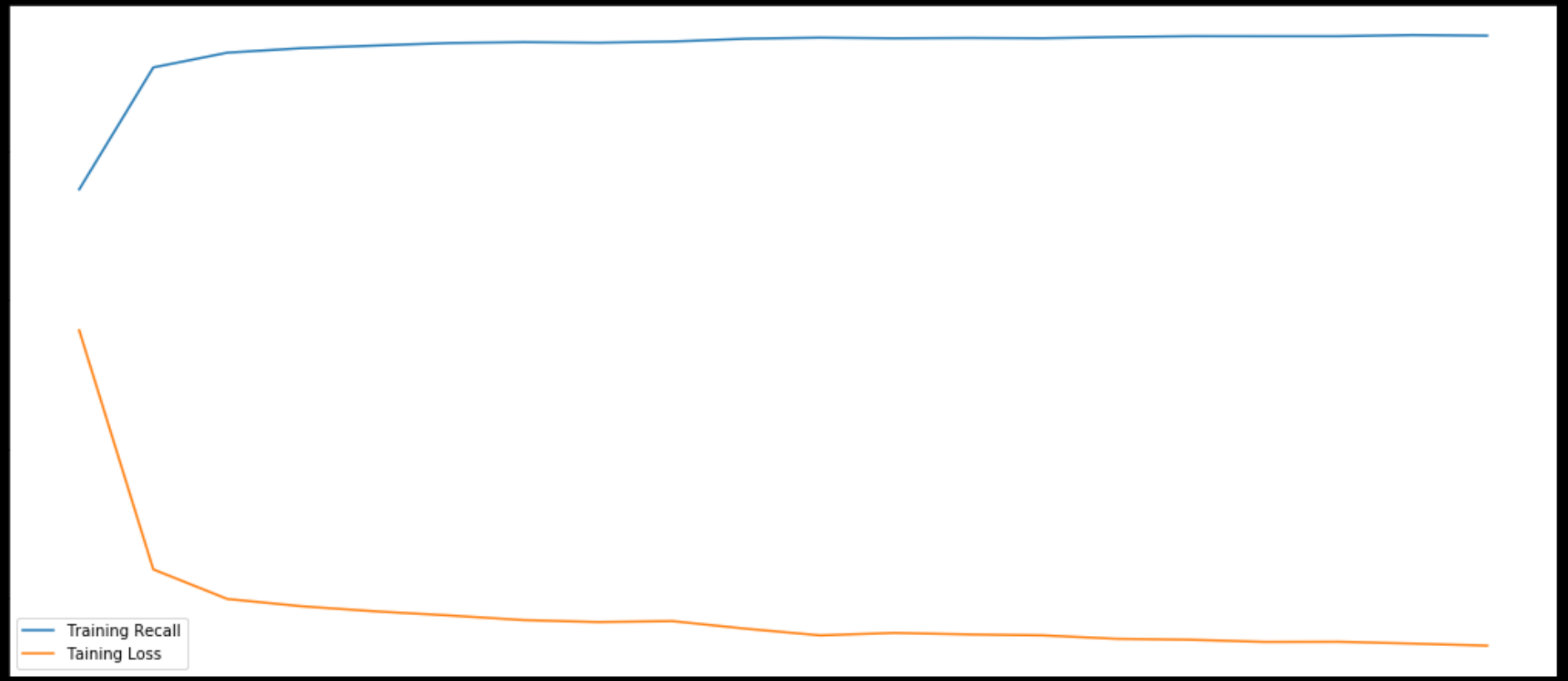


EXAMPLES OF CLEAN CELLS

MODELING

- Multiple convolutional neural networks were tested and all networks were also tested with data augmentation to determine if the augmentation would have a positive effect on sensitivity.
- A rectified linear unit(ReLU) activation function was used for most of the networks but a sigmoid activation function was also tested.
-

TRAINING RECALL AND TRAINING LOSS



CHALLENGES

- The .db file in the folders means a simple for loop to import images will not work but a try/except loop will do the job and if it is working properly there will be one use of the except line when reading each folder.
- Due to the amount of images being used training models without GPU support is a time consuming process. With augmented data the issue becomes even worse with 20 epochs of training often taking more then an hour of time.

RESULTS TABLE

Model	Accuracy	Recall
Model 1	96.44%	95.57%
Model 1 Augmented	96.25%	95.86%
Model 2	95.41%	92.78%
Model 2 Augmented	94.89%	94.21%
Model 3	95.05%	91.93%
Model 3 Augmented	94.89%	95.75%
Model 4 Augmented	95.59%	94.21%
Model 5	95.97%	93.71%
Model 5 Augmented	94.53%	94.96%

FINDINGS AND TAKEAWAYS

- Some models created promising results that if deployed could potentially assist medical staff in diagnosing malaria, especially in high infection areas.
- Using a GPU had drastic effects on processing speed. Augmented epochs could take 300 seconds or more without GPU support. After the GPU was used the epoch time was cut down to 15-20 seconds.
- Model 1 Augmented was the most successful model with a recall of 96.25%

FURTHER WORK

- Better optimization of the models is possible and could be something worth doing in time.
- Taking this network and placing in a full stack application that allows medical staff to easily upload images and easily get results back.
- There are other diseases where these techniques could be applied as well.
- Designing an algorithm that could diagnose several different kinds of illnesses would be extremely useful with the end goal of being able to feed blood sample images into an application and automatically receive diagnoses.