WGU MSDA Capstone

Identifying Malaria in Patient Blood Samples Using Convolutional Neural Networks

Colton B. Horton

June 30, 2024

Personal Introduction

Colton B. Horton

- Master's student at Western Governors University studying Data Analytics.
- Living in Utah.
- Bachelor's degree in microbiology from Brigham Young University.
- My background in both data analytics and biology inspired this capstone project.

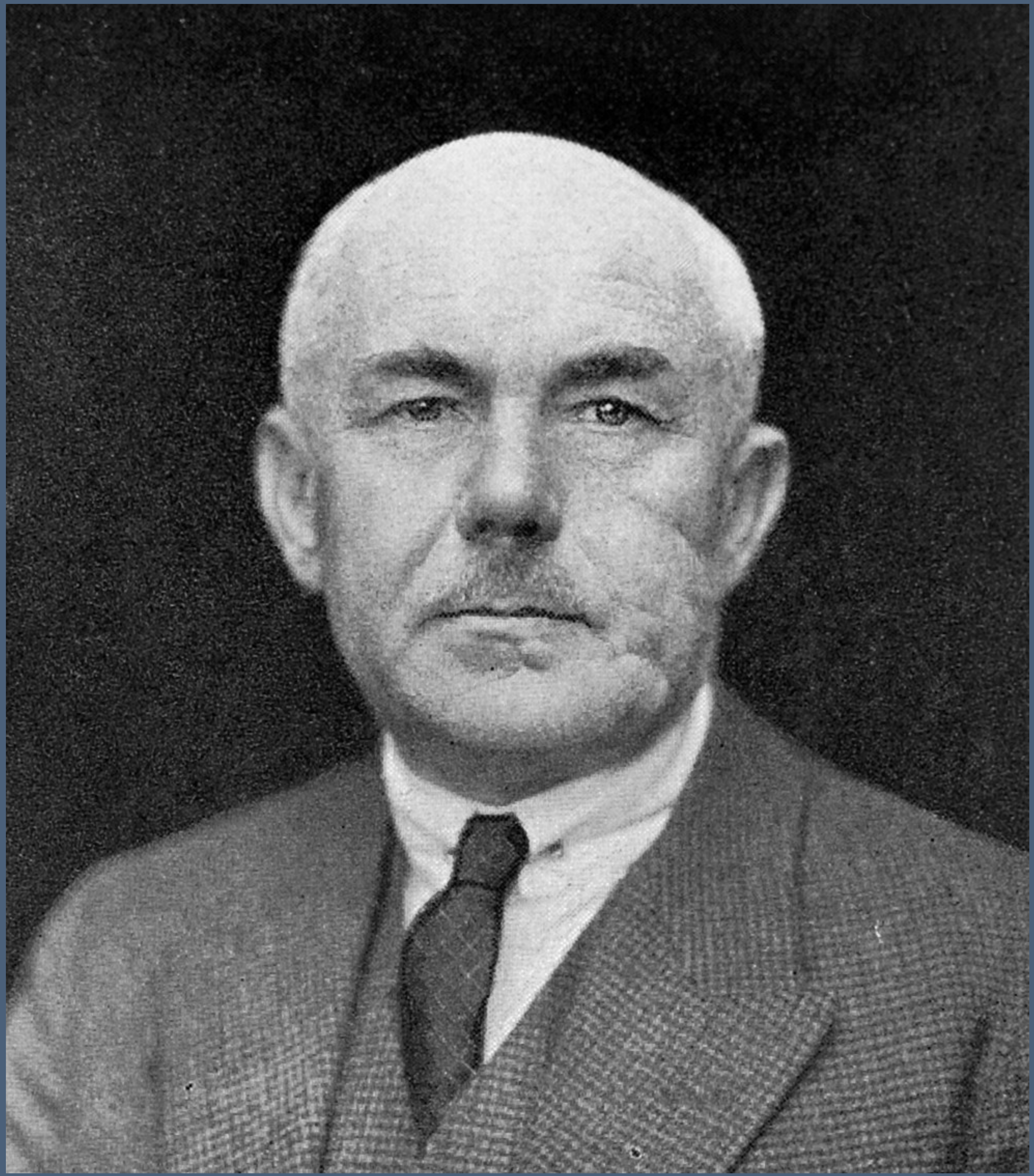
Research Problem

Malaria

A Global Health Challenge

- Malaria is a blood-borne infection caused by *Plasmodium* parasites and carried by mosquitoes. [1]
- It kills hundreds of thousands of people every year around the world. [2]
- It is a curable infection, but diagnosis remains a barrier to treatment, particularly in resource-limited areas. [1]

https://www.flaticon.com/free-icon/malaria_773852



https://en.wikipedia.org/wiki/Gustav_Giemsa

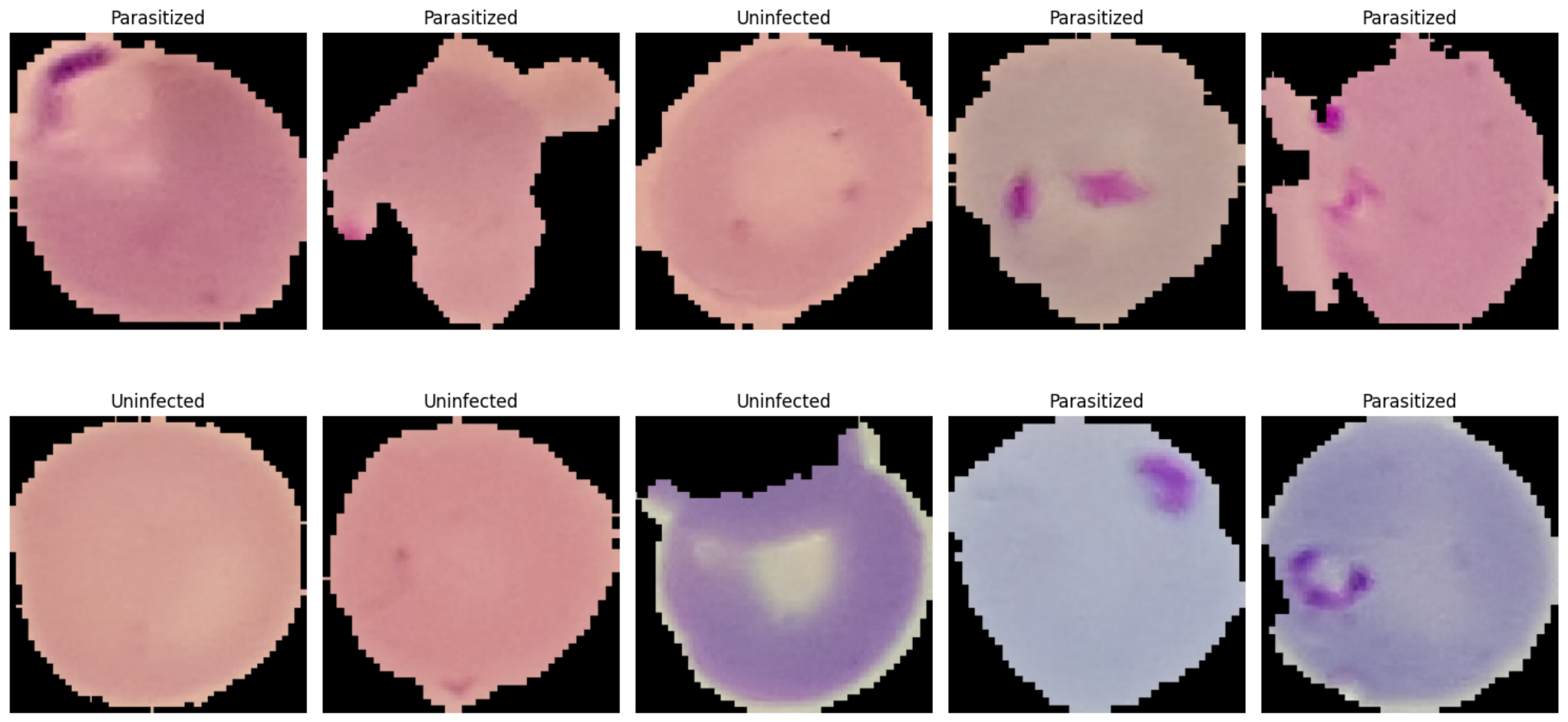
Malaria

Current Diagnostics

- In 1904, Gustav Giemsa created a staining protocol that analyzes human blood under a microscope for the presence of *Plasmodium* parasites. [3]
- It remains the gold standard for diagnosis today. [1]
- The diagnosis accuracy of Giemsa-stained blood samples, analyzed by trained microbiologists is ≈ 90%. [4]

Malaria

Giemsa Stain Example Images



https://www.kaggle.com/datasets/iarunava/cell-images-for-detecting-malaria

Malaria

Research Question

- Developing a quicker diagnostic tool that does not rely on trained microbiologists may: [5]
  - Increase speed to diagnosis.
  - Improve access to treatment.
  - Potentially decrease malaria mortality.
- Can a deep learning model be used as an alternative diagnostic tool to potentially achieve these outcomes?
  - The deep learning model of choice will be a "convolutional neural network."

Hypotheses

Null hypothesis:

- Convolutional neural networks **can not** predict cells infected with malaria with statistical significance, achieving at least 90% accuracy.

Alternative hypothesis:

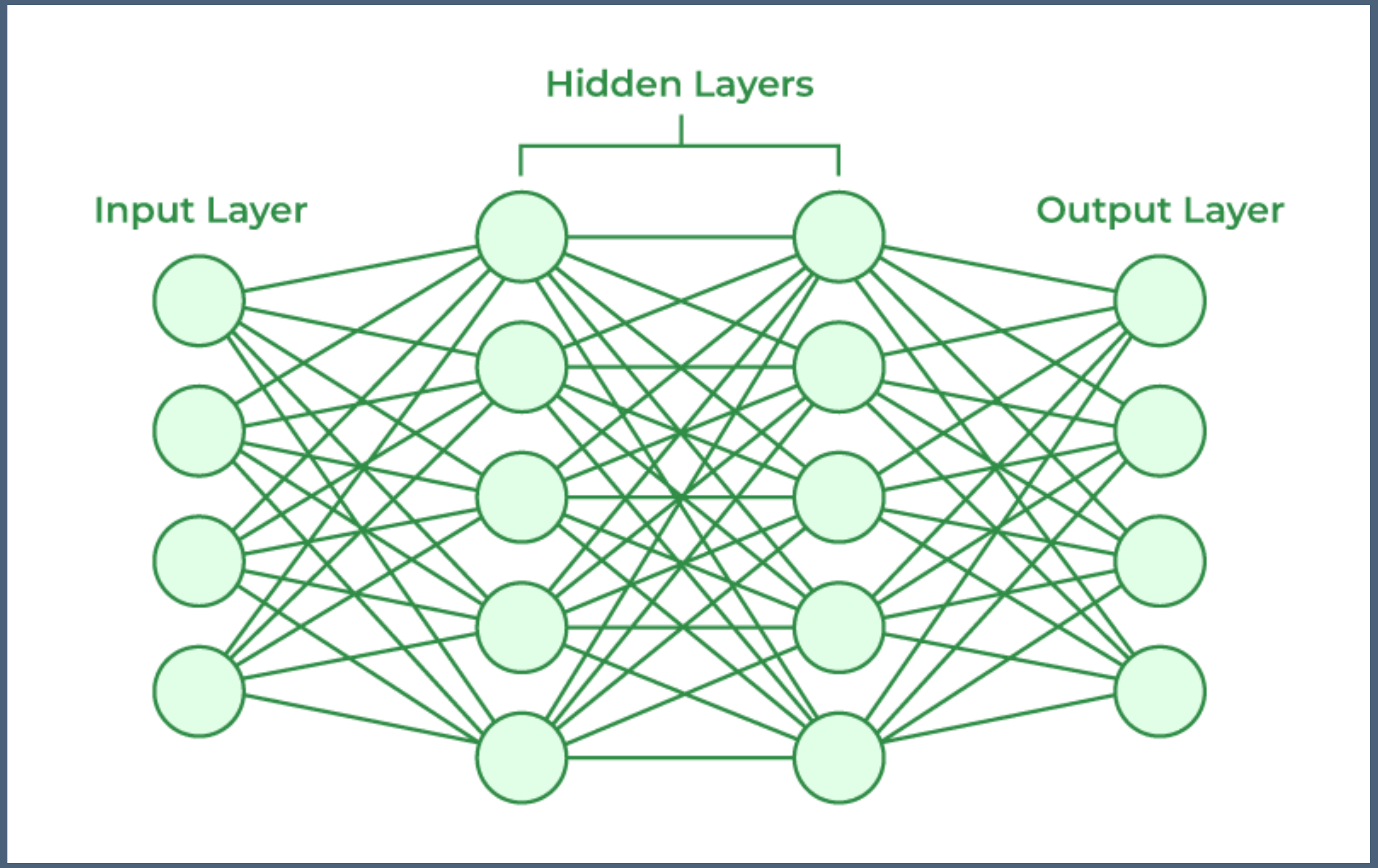
- Convolutional neural networks **can** predict cells infected with malaria with statistical significance, achieving at least 90% accuracy

Data Analysis

Model Building

General Neural Networks

- Neural networks were inspired by the connections in the human brain. [6]
- The goal of a neural network is to learn complex patterns in data.
- Consists of interconnected layers of nodes (or neurons) that identify these relationships.
- There are 3 types of layers: Input, hidden, and output.
  - The hidden layers are the heavy lifters.



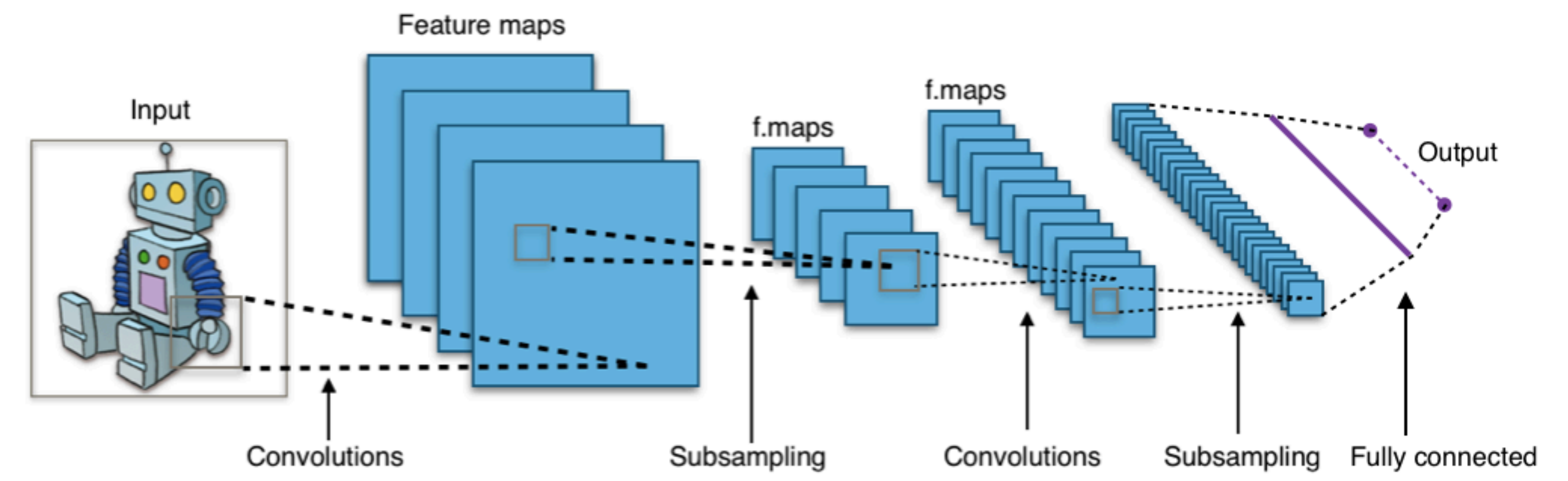
https://www.geeksforgeeks.org/artificial-neural-networks-and-its-applications/

Model Building

Convolutional Neural Networks (CNNs)

- CNNs are great for working with images, perfect for this project! 7
- A "convolution" is a mathematical process that transforms an image into another one called a "feature map."
  - The image in the upper right shows an edge-detection convolution / feature map.
- CNNs processes images through the use of many feature maps.

https://en.wikipedia.org/wiki/Kernel_(image_processing)



https://en.wikipedia.org/wiki/Convolutional_neural_network

Model Building

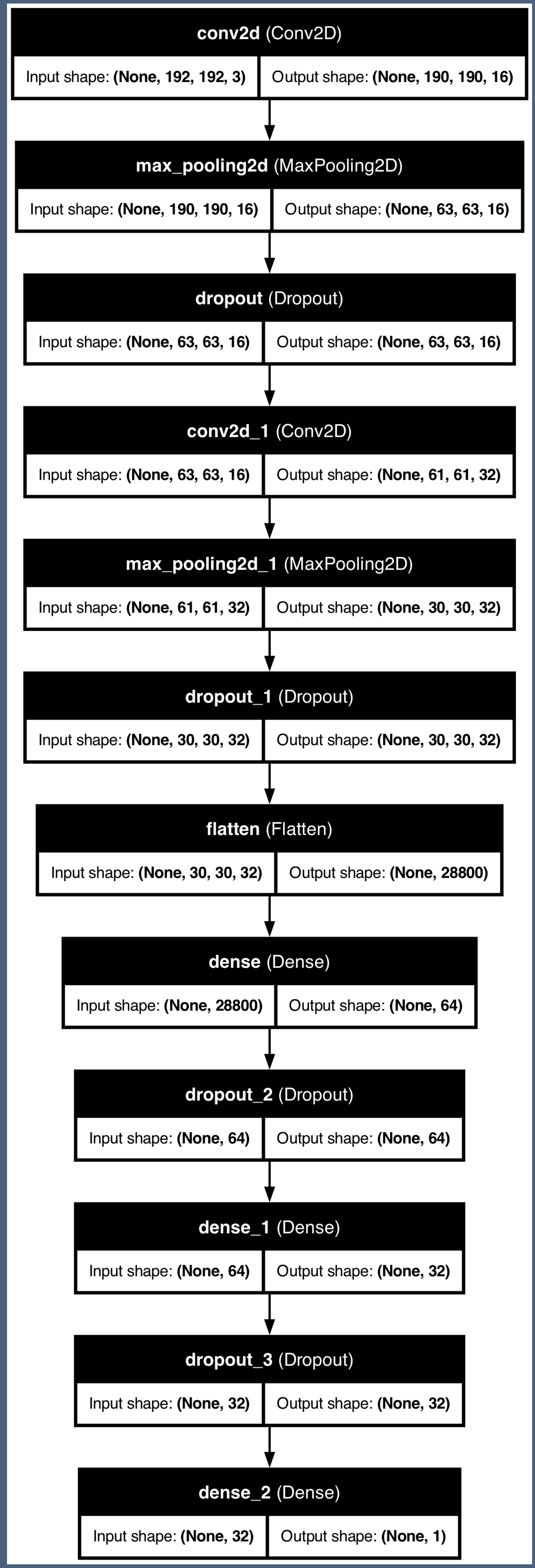
Sequential Model Building

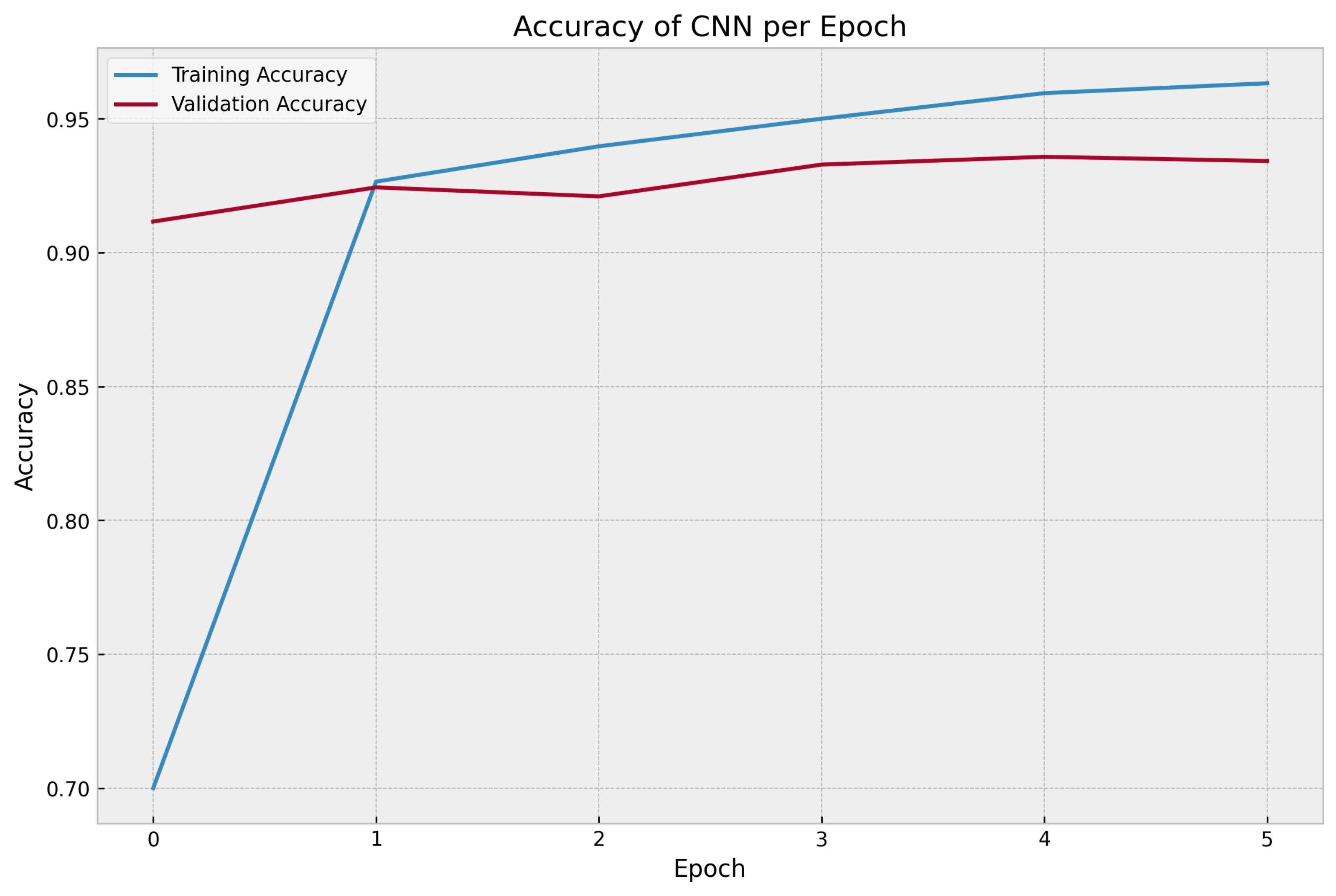
- Neural networks are very complex and require careful consideration when building them. [8]

- For this project, I will be building the model one layer at a time and analyzing how it impacts predictive performance.

  - Once the predictions are no longer improving, the model building process can stop.

Final Model Architecture





Final training accuracy: 96.4%
Final validation accuracy: 93.4%

Model Building

Final Model

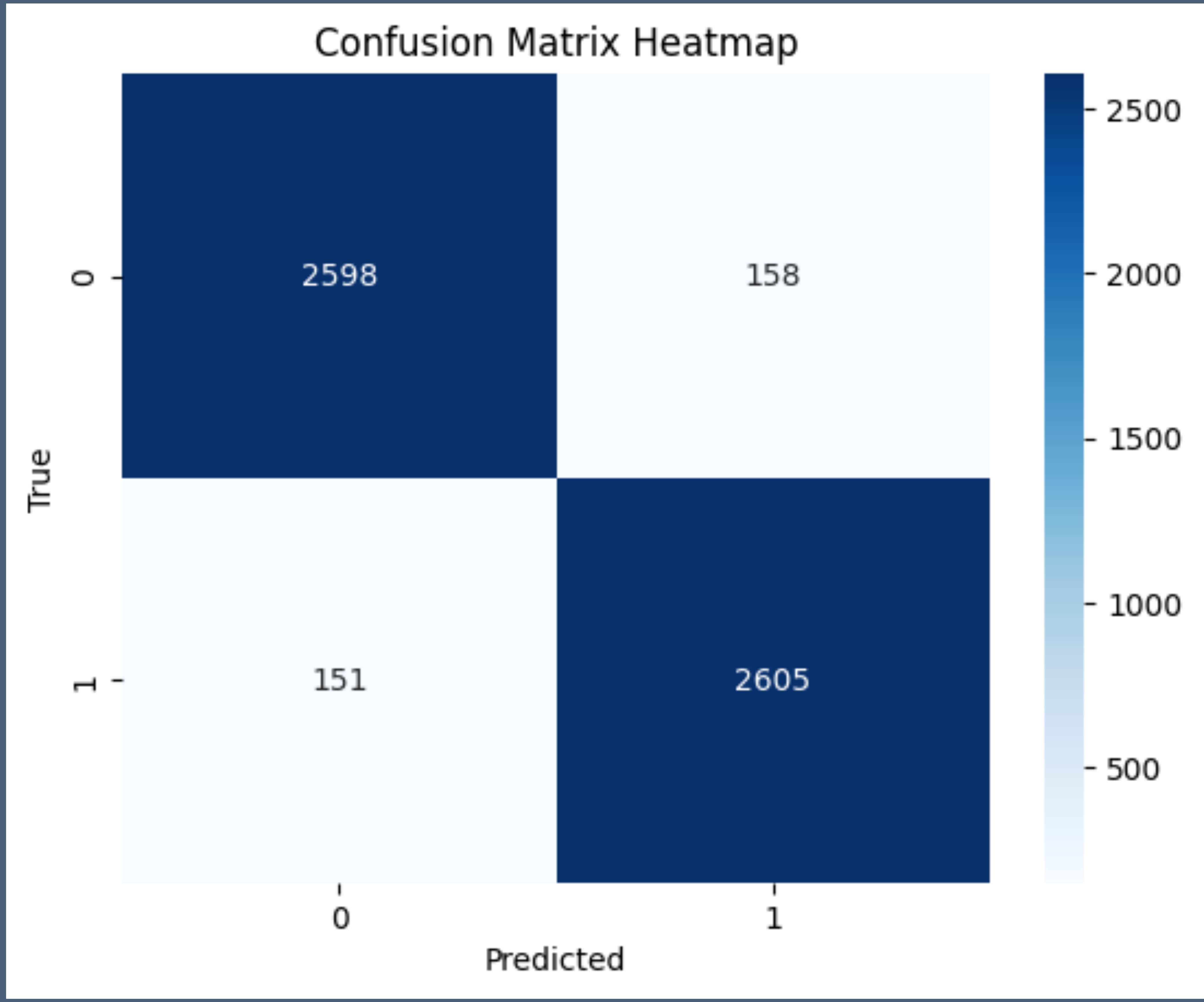
- Before fitting CNNs, data are split into training, validation, and testing subsets. 9
  - The training and validation subsets are used for fitting the neural network, the testing subset is used for a final examination of the model.
- CNNs are trained in batches called "epochs" 10
  - The accuracy can be examined for each epoch of the training process.
- The final model architecture can be seen on the left as well as the accuracy progression of the training process.

Findings

Results

Predictions From Testing Images

- Chi-square p-value of the predictions confusion matrix: approximately zero.
- Overall model accuracy: 94.4%.
- False positive and false negative rate are both approximately 5.6%.
- Reject the null hypothesis:
  - CNNs **can** predict cells infected with malaria with statistical significance, achieving at least 90% accuracy.



Limitations

Technological Requirements

- While CNNs boast strong predictive performance, they require the proper technology to host them at testing sites.
- Local infrastructure may not support this requirement.
- Each testing location will have unique challenges and addressing them will require careful consideration.
  - Creating portable, self-sufficient diagnostic devices independent of infrastructure.

Proposed Actions

CNN implementation

- Partner with humanitarian and medical aid programs to assess how and where the CNN can address needs and relieve strain on the medical system.

  - Ensure the CNN is integrated into existing healthcare frameworks for effective and sustainable implementation in resource-limited settings.

Expected Benefits

Enhanced Diagnostics and Global Health Impacts

- Implementing this model could benefit resource-limited areas by providing individuals with a fast and effective malaria diagnostic tool.
- By increasing the ease of receiving a diagnosis, it could improve the outcomes of malaria patients all around the world by empowering them to get treatment sooner (or at all).
- This project also underscores the benefit that deep learning models can have in the medical field, both in resource-limited areas as well as resource-abundant areas.
  - Successful implementation could help support more medical deep learning research.

References

1. Centers for Disease Control and Prevention. (2024, April 24). Malaria diagnostic tests. Centers for Disease Control and Prevention. https://www.cdc.gov/malaria/hcp/diagnosis-testing/malaria-diagnostic-tests.html
2. World Health Organization. (2023, December 4). Fact sheet about malaria. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/malaria
3. Wikimedia Foundation. (2023, December 8). Gustav Giemsa. Wikipedia. https://en.wikipedia.org/wiki/Gustav_Giemsa
4. Stauffer, W. M., Cartwright, C. P., Olson, D. A., Juni, B. A., Taylor, C. M., Bowers, S. H., Hanson, K. L., Rosenblatt, J. E., & Boulware, D. R. (2009). Diagnostic performance of rapid diagnostic tests versus blood smears for malaria in US clinical practice. Clinical Infectious Diseases, 49(6), 908–913. https://doi.org/10.1086/605436
5. Jaegar, S., & Yang, F. (2021, October 27). Malaria datasheet. U.S. National Library of Medicine. https://lhncbc.nlm.nih.gov/LHC-research/LHC-projects/image-processing/malaria-datasheet.html
6. Mijwel, M. M., Esen, A., & Shamil, A. (2023). Overview of Neural Networks. Babylonian Journal of Machine Learning, 2023, 42–45. https://doi.org/10.58496/bjml/2023/008
7. Yamashita, R., Nishio, M., Do, R. K., & Togashi, K. (2018). Convolutional Neural Networks: An overview and application in Radiology. Insights into Imaging, 9(4), 611–629. https://doi.org/10.1007/s13244-018-0639-9
8. Gupta, V., Li, Y., Peltekian, A., Kilic, M. N., Liao, W., Choudhary, A., & Agrawal, A. (2024). Simultaneously improving accuracy and computational cost under parametric constraints in materials property prediction tasks. Journal of Cheminformatics, 16(1). https://doi.org/10.1186/s13321-024-00811-6
9. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., … Zheng, X. (2015). TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/
10. Li, H., Rajbahadur, G. K., Lin, D., Bezemer, C.-P., & Jiang, Z. M. (2024, January 18). Keeping Deep Learning Models in Check: A History-Based Approach to Mitigate Overfitting. arXiv. https://arxiv.org/html/2401.10359v1

Thank You!