Malaria detection through RBC images using machine learning

Saurav, 2020243
Mohammad Osama Ataullah, MT21127
Sairam Mohan, 2019198

February 26, 2023

Abstract

Malaria is a disease caused by a plasmodium parasite and spread by the bite of an infected female Anopheles mosquito. The disease claims the lives of more than 400,000 people annually. The traditional approach for detecting Malaria is by preparing a blood smear and using a microscope to examine and find the parasite genus Plasmodium. This approach mainly relies on the knowledge of skilled specialists. In this research, using the red blood cell smears of the sampled cells, we will compare several feature extraction and classification algorithms. The National Institutes of Health provided RBC cell images dataset will be utilized in this study. To compare and choose the best performing architecture, evaluation criteria like accuracy, recall, precision, F1 score, and Area under curve (AUC) may be used

Keywords - Machine Learning, Deep Learning, Image classification, Feature Extraction, Malaria Detection

1 Introduction

1.1 Background

Malaria is caused by the protozoan parasites of the genus Plasmodium, that infects the red blood cells and causes this fatal disease. It spreads by mosquitoes bite of infected female Anopheles mosquitoes. In addition to flu-like symptoms, an infected person may also experience high fever, exhaustion, chills, septicemia, pneumonia, gastritis, enteritis, nausea, vomiting, headaches, and, in extreme circumstances, seizures and coma, which can be fatal. Although it cannot be passed from person to person, malaria can be transmitted from mother to foetus, through blood transfusions, or by sharing needles. The areas where malaria is most prevalent are those with warm, humid climates near freshwater sources, adjacent to Anopheles mosquito breeding grounds. Malaria is frequently identified by utilizing blood films and a microscope to examine blood cells. Every year, hundreds of millions of blood films are checked for malaria, which entails a qualified microscopist manually counting parasites and infected red blood cells. Not only are accurate parasite counts crucial for diagnosing malaria, but they

are also necessary for detecting treatment drug resistance, assessing therapeutic efficacy, and categorizing illness severity. False-negative cases result in needless antibiotic administration, a second appointment, missed workdays, and in some cases, progression into severe malaria. A misdiagnosis for false-positive instances requires the unnecessary use of anti-malaria medications and enduring their potential adverse effects, which may include nausea, abdominal pain, diarrhea, and occasionally serious complications. With the use of various machine learning approaches, we are attempting to automate the diagnostic process so that human professionals can make the most accurate diagnoses. We will examine different classifying algorithms such as Support Vector Machines, Decision Tree, and Logistic Regression after doing extensive preprocessing on the NIH malaria dataset. Also, we will try a number of preprocessing techniques on the dataset, including filters, feature extraction etc., to improve performance on the objective test set. We will also explore deep learning to improve our accuracy. Deep Learning is a machine learning technique designed to mimic the capability of processing information and making decisions in the human brain. Association, mindfulness, personality, and other aspects of the human mind go much beyond present profound learning capacities. We will began by employing an artificial neural network (ANN), which is a computational network based on biological neural networks that create the structure of the human brain, as our baseline model and then dive into deep learning architectures.

1.2 Literature survey/Research work

Currently, there are numerous assessment papers available that are related to our problem statement. [1] proposes using deep learning rather than traditional strategies that rely on tedious hand engineering feature extraction in an end-to-end arrangement that performs both feature extraction and classification directly from raw segmented patches of red blood smears. In [2] Shallow machine learning algorithms, such as AdaBoost, Random Forest, Decision Tree, and KNN, are used against the conventional strategy. Their proposed methodology detects malarial infection using captured images of patients without staining the blood or requiring the assistance of experts. [3] provide an overview and comparison of various techniques used in image analysis and machine learning for microscopic malaria diagnosis, such as imaging, image preprocessing, parasite detection and cell segmentation, feature computation, and automatic cell classification. [4] compared different types of features, feature extraction techniques, and why they are important. We also learned more about which features extraction techniques will be better and more valuable in which situations. In [5], global and local texture feature extraction is done using different algorithms. The significance of texture as a feature in an image is emphasised.The global texture features of an image are computed, such as homogeneity, correlation, contrast, dissimilarity, and maximum probability. Furthermore, [6] proposed CNN models (VGG16, VGG19) [7] with support vector machines (SVM) to determine the stages of parasite contamination and further developed the preparation time by using the pre-trained CNN models and transfer learning techniques. The

research contributed by fostering a CNN model to finetune the hyperparameter of the pre-trained model using transfer learning. [8] suggests employing cell segmentation approaches based on edge detection, watershed segmentation, and morphological segmentation to detect malaria parasites in blood smears. After segmenting cells,, the infections are identified using a threshold intensity pixel value for the infection, i.e. if the pixel value is in the range of the threshold the infections are recognized. Throughout the creation of this proposal, we referred to these papers and blog articles.

2 Materials and Methods- Dataset, Methodology, Novelty, Evaluation Metric(s)

2.1 Dataset

The dataset, titled NIH Malaria Dataset, was obtained from the National Institutes of Health (NIH). The dataset comprises of 27558 stained smears of the collected cells that were labeled as Uninfected or Infected with plasmodium virus. The 13779 parasitized images and 13779 uninfected images in the collection are spread equally. The segmented red blood cell patches have three channels (RGB), a channel depth of three, and a size range of 110–150 pixels. Plasmodium was present in positive samples, but plasmodium was absent in negative samples, however they might still contain other substances, such as staining impurities.

2.2 Preprocessing

Image patches' raw structured pixel data won't be particularly useful for the categorization task. Instead, we employ a representation that is unaffected by translation, rotation, and intensity offsets. The Plasmodium detection issue is mostly concerned with the form of the objects in the input patches. We need to scale the various-sized photographs we've collected, and we need a representation that isn't affected by changes in translation, intensity, or rotation. We propose the use of edge detection filters for edge detection and preprocessing in our dataset as the Parasitized Cell Images contains stained purple patches. Hence we decided to use edge detection to get the enhanced image of the infected cell, which would further help us in feature extraction

2.3 Proposed Methodology

We will first examine a number of edge detection techniques, including Canny, Sobel, Laplacian of Gaussian (LoG), and Scharr, as well as Adaptive Histogram Equalization, or CLAHE. Moreover, we might employ local and global feature extraction methods.The former looks for features using a colour histogram, shape, and texture. We will investigate some of the most popular methods for local feature extraction, such as SIFT, MSER, KAZE, and others. We will then begin modelling by applying various baseline machine learning models, such

as Decision Tree, Logistic Regression, Support Vector Machine, and ANN, to our processed data. Then, utilising various bagging and boosting approaches like Random Forest and Adaboost for hyperparameter tweaking of our models, we will execute Ensembling on the best-performing models. We will further explore Deep Learning, starting with ANN (Artificial Neural Network). Moving on to more advanced architectures, like CNN (Convolution Neural Network). We will put the most effective action function to and test when evaluating the performance of well-known pre-trained CNNmodels like VGG, ResNet, etc. on our preprocessed dataset.

3 Plan and tentative timeline

3.1 Planning and problem definition(1-2 weeks) [6 Feb - 18 Feb 2023]:

Defined the problem to be solved and established the project goals and objectives Determined the scope of the project and identified the required resources, including dataset source. Conducted a preliminary analysis of the data to identify any data quality issues or data preprocessing requirements Selected a possible set of appropriate machine learning algorithms and techniques to be applied on our problem.

3.2 Data Preprocessing (1-2 weeks)[26 Feb - 4 March 2023]:

Would preprocess the data appropriately to extract relevant features, with local and global feature extraction algorithms, data cleaning, normalization, and feature engineering. Would Perform exploratory data analysis to identify any patterns or relationships in the data that could inform the machine learning model selection or feature engineering. Use appropriate ratio for train test split. Perform any necessary data augmentation to increase the size or diversity of the dataset

3.3 Model Development and Training (2-3 weeks)[6 march - 20 march 2023]:

Train various baselines and our proposed machine learning model using the training data and validate the model using the validation set. Optimize the hyperparameters of each model using techniques such as grid search or random search Evaluate the performance of the model using appropriate metrics such as accuracy, precision, recall, or F1 score Tune the model using techniques such as early stopping, regularization, or ensembling to improve performance

3.4 Model Evaluation (1-2 weeks)[21 March - 26 March 2023]:

Evaluate every final model including the baselines and proposed models using the testing set. Compare and report the performance metrics to identify the best performing model.

3.5 Documentation (1 week)[27 March - 31 March 2023]:

Document the entire process, including data preparation, model development, evaluation, and deployment, to facilitate future replication or extension of the project Create a final report that communicates the project objectives, methodology, results, and conclusions.

4 Distribution of work among group members

Mohammad Osama Ataullah:Literature Survey, Preprocessing, Local Feature Extraction, Deep Learning
Saurav: Literature survey, Global Feature Extraction, Hyper Parameter tuning, Deep Learning
Sairam Mohan: Literature survey, Preprocessing, Machine learning, baseline implementation

References

1. Aimon Rahman, Hasib Zunair, M Sohel Rahman, Jesia Quader Yuki, Sabyasachi Biswas, Md Ashraful Alam, Nabila Binte Alam, and M. R. C. Mahdy. 2019. Improving malaria parasite detection from red blood cell using deep convolutional neural networks.
2. G.B. Saiprasath, N. Babu, J. ArunPriyan, R. Vinayakumar, V. Sowmya, and Dr Soman K. P. Performance comparison of machine learning algorithms for malaria detection using microscopic images". IJRAR19RP014 International Journal of Research and Analytical Reviews, 6(1).
3. Gaurav Kumar and Pradeep Kumar Bhatia. 2014. A detailed review of feature extraction in image processing systems. In 2014 Fourth International Conference on Advanced Computing Communication Technologies, pages 5–12.
4.Mahdieh Poostchi, Kamolrat Silamut, Richard J. Maude, Stefan Jaeger, and George Thoma. 2018. Image analysis and machine learning for detecting malaria. Translational Research, 194:36–55. InDepth Review: Diagnostic Medical Imaging.
5.Nagarajan Deivanayagampillai, Suruliandi A, and Kavitha Jc. 2017. Melanoma detection in dermoscopic images using global and local feature extraction. International Journal of Multimedia and Ubiquitous Engineering,scopus, 12:19–27
6.Vijayalakshmi and Rajesh Kanna. 2019. Deep learning approach to detect malaria from microscopic images. Multimedia Tools and Applications, 79(21-

22):15297–15317
7.Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition.
8.Kirti Motwani, Abhishek Kanojiya, Cynara Gomes, ,Abhishek Yadav (2021). "Malaria Detection using Image Processing and Machine Learning" International Journal of Engineering Research & Technology (IJERT)