Multi-class Classification of Malaria Parasite Life cycle using Single-cell Transcriptomes

Thesis submitted in partial fulfillment
of the requirements for the degree of

*Master of Science in **Computational Natural Sciences** by Research*

by

Swarnim Shukla
20161113
swarnim.shukla@research.iiit.ac.in

International Institute of Information Technology, Hyderabad
(Deemed to be University)
Hyderabad - 500 032, INDIA
June 2023

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "Multi-class classification of Malaria Parasite Life cycle using Single-cell Transcriptomes" by Swarnim Shukla, has been carried out under my supervision and is not submitted elsewhere for a degree.

Date

Adviser: Dr. Bhaswar Ghosh

To my dear family members and friends!

Acknowledgments

My sincere and heartfelt gratitude to everyone who assisted me in finishing my research thesis and thereby helped me to develop a researcher's perspective. First and foremost, I would like to express my sincerest gratitude to my adviser, Dr. Bhaswar Ghosh, for his overwhelming support ever since I got associated with him. He honed my research skills so as to bring me to my research work submission status. This thesis was only made possible by his unwavering commitment to high-quality work.

I would like to particularly thank Arghya who helped me gain very useful technical insights and the necessary core domain knowledge pertinent to my research. He was always willing to help me with any queries I had and brainstorm ideas. I would also like to thank Soham and Gayathri for always helping and guiding me and making this research possible.

I would also like to express my gratitude to my mother Ruchi, father Mukul, and brother Adi for their continued affection, inspiration, and encouragement throughout my research. It would have been much more challenging to complete my research without their unwavering love and support. I would also like to thank all my friends Mayank, Ankit, Vikrant, Swapnil, Suhan, Ritvik, and the rest of the bois who made my college life @ IIITH memorable and were an absolute source of motivation, brotherhood, and fun in life. Thank you bois!

In the end, I would like to thank one and all those who contributed in some way towards the successful completion of my thesis work!

Abstract

Malaria, which is spread by the female anopheles mosquito, is a highly fatal disease that affects many parts of the world, with up to 0.4 million deaths reported worldwide. The detection of malaria infection levels is based on vital gene expressions. Experts quantify malaria parasite-infected RBCs and classify their life cycle stages at the macroscopic level in order to make informed decisions. Several computational approaches have recently been proposed to avoid the dimensionality problem and produce accurately predicted results. Our study presents a theoretical framework to select diagnosis markers and drug targets by implementing ML techniques on sc-RNA-seq data. The main objective is to select the top-ranked genes from the scRNA-seq profiles at different stages of the Plasmodium falciparum (Pf) life cycle inside infected RBC. We employ a supervised learning algorithm coupled with feature selection algorithms to extract the most relevant genes to predict the life cycle stages of Pf inside RBC.

The first stage of modeling is to optimize the quality of data from the dataset (5066 features) by removing the irrelevant features. Genetic Algorithm (GA) based search technique is popularly used for feature selection and dealing with high dimensionality datasets. This reduced subset (378) is further utilized in the second stage of high accuracy multi-class classification.

In this work, a GA-based dimensionality reduction technique is used on single-cell transcriptomics to obtain an optimised subset of features from a larger data set. To separately transform the selected elements into a lower dimension, features are chosen based on their class variants, taking into account increased efficiency and accuracy. We constructed the protein-protein interaction network (PPIN) of these genes and performed topological analysis using the Search Tool for the Retrieval of Interacting Genes/ Proteins database (STRING 11.0 b) and Gephi software to provide hierarchies according to the importance of the genes in the network. Various topological measures are estimated to evaluate the node characteristics in the PPINs, including degree, between centrality, eccentricity, closeness centrality, eigenvector centrality, and clustering coefficient. Proteins having a high degree and betweenness centrality tend to assert more control over the network function. We also performed gene ontology analysis to determine the role of proteins in the parasite's life cycle progression.

For the multi-class classification of the life cycle of malaria parasite based on oriented gradients and local binary pattern features, a three-pronged approach employing the multi-class Support Vector Machine (SVM), Logistic Regression (LR), and Random Forest (RF) techniques are used. On using these 378 features, RF performed best with a classification accuracy of 92% while SVM had a 91%

accuracy and LR gave 88% accuracy. By merely using the 378 features, we achieved similar or better performance scores for all four classes, across all three models. Further, randomly chosen features from our dataset of 378 were also evaluated using the SVM, LR, and RF models. We achieved an accuracy of 81%, 79%, and 80% for the three respective models. This proves the robustness of the features selected using the GA-based approach. The proposed research methodology can be likely used for improved malaria diagnosis and drug targets.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Malaria

Malaria is a deadly disease caused by the Plasmodium parasite and is transmitted through the bite of a female Anopheles mosquito. The four common Plasmodium species are Plasmodium (falciparum, vivax, malariae and ovale) [7] and two of these species falciparum and vivax are of the greatest threat. *Plasmodium falciparum (Pf)* is the most prevalent on the African continent. *P. vivax* is dominant in most countries outside of sub-saharan Africa. This plasmodium attacks the red blood cells (RBCs) and the degree of malaria can be estimated by the quantity of infected RBCs [8]. If not treated, Pf malaria can progress to death within a span of 24 hours. In 2020, 241 million cases of malaria were estimated worldwide, malaria deaths stood at 627000 in 2020 [9]. Plasmodium species cause human malaria, with majority of the estimated 0.4 million annual deaths accounted for the deadliest unicellular, protozoan malaria causing parasite Pf. Of all the Plasmodium species, Pf has the quickest time for the development of infection symptoms. The incubation phase may last nine to thirty days. Malaria symptoms include high fever, tiredness, vomiting and headache and even seizures and death in some severe cases. If not properly treated, the disease may recur even months later. The most common symptom of malaria is paroxysm. Every two days, there is a cyclical occurrence of sudden coldness followed by shivering, fever, and sweating. Malaria can lead to several serious complications including development of respiratory disorder. Non-cardiogenic pulmonary oedema, pneumonia, and severe anaemia are a few potential reasons. Infection with Pf can cause severe cerebral malaria, which is distinguished from other types of fever by retinal whitening.It causes spleen and/or liver enlargement, as well as low blood sugar, severe headache, and haemoglobin in the urine with kidney failure. [10, 11].

1.2 Malaria Parasite Life Cycle

The growth and survival of plasmodium through the reproduction and development process in different hosts during its entire life cycle occurs by more than 5,000 genes and associated proteins. These

genes also evade the different responses of the host that otherwise may lead to their destruction.

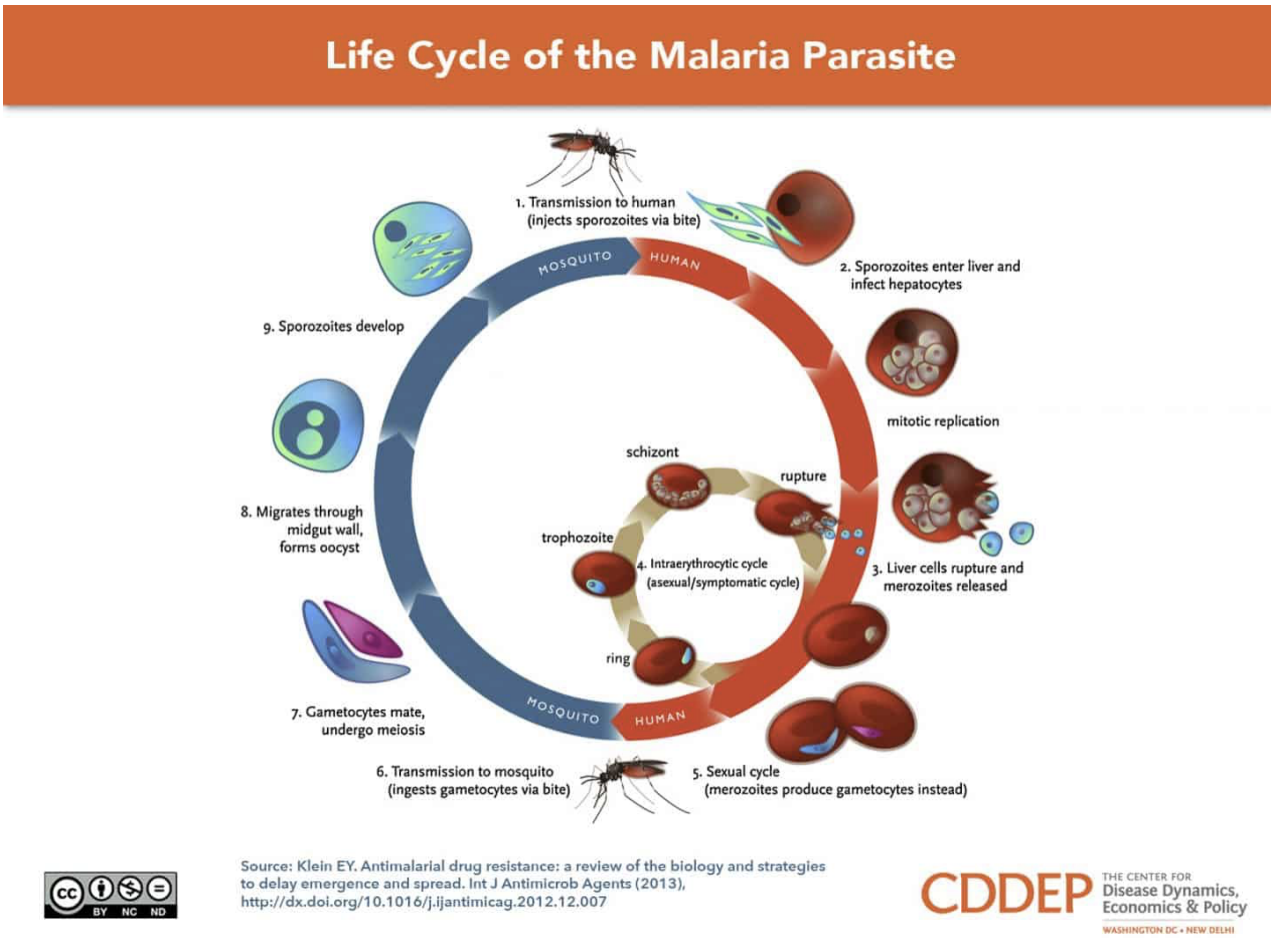


Figure 1.1: The life cycle of a malaria parasite.

The complex life cycle (Figure 1.1) of malaria parasites adapts diverse developmental strategies, each of which is adapted to thrive in the particular host environment. Malaria transmission occurs through female anopheles mosquito (vector). When it feeds on an infected human it ingests gametocytes - the sexual form of the parasite [12]. Male and female gametocytes mate in the mosquito's gut and after meiosis they move through the midgut wall forming an oocyst and leading to the development of thousands of sporozoites which are then injected into a human during the next bite, rapidly reaching the liver and infecting the hepatocytes. Later they start replicating asexually. Around 15 days later the liver schizonts are ruptured, and thousands of merozoites are released into the blood further invading the RBCs. Over the next 48 hours, the parasite replicates, passing through different stages (ring, trophozoite and schizont), and producing nearly 16 new merozoites per schizont. The schizonts further populate asynchronously with other parasites, producing the dominant fever cycle. After each replication, some

of these merozoites also develop into gametocytes, infecting susceptible mosquitoes, bringing full circle to the transmission cycle [13].

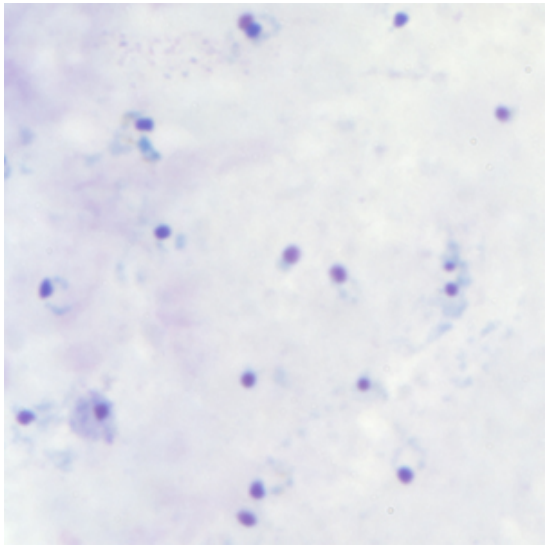
During the blood stage growth, Plasmodium being intra-cellular, provides protection to the parasite due to the immune response of the host and is vulnerable in the extra-cellular stage [14].The development of blood stage begins when a freshly released, extracellular parasite (a merozoite) attacks an erythrocyte, forming the ring stage of infection, progressing into the trophozoite stage. During this stage the infected erythrocyte is largely modified enabling the proliferation of parasite. Thereafter, the parasite sub-divides to form a connected network of daughter cells, called schizont, which finally loses the host erythrocyte, and releases the new born merozoites to attack the new erythrocytes. These steps are in together termed the intra erythrocytic developmental cycle (IDC) [15]. The parasite then invades RBCs and develops into a ring stage within 48 hrs and soon followed by stages trophozoite and schizont. Depending on differentiation intra erythrocytically to gametocytes inside a human host, the transmission of malaria parasites to mosquitoes is done [16].

On attachment of the parasite to RBC, its cell membrane deforms forming a junction allowing the penetration of the parasite in to the cell using various protein structures. The parasite which begins to form a ringed shape in the cell, creates a parasitophorous vacuole which separates it from the RBC's intra-cellular environment. For biosynthesis, haemoglobin the primary nutrition source and amino acids are used. The breakdown of molecule is followed by the continued proliferation and increase in numbers and size of the RBCs. The parasite continuously divides within the cell going through different stages and producing trophozoites and finally schizonts. As their number increases by asexual reproduction, the cell bursts producing new merozoites which infect other RBCs while some of the parasites turn into gametocytes leaving the RBCs uninvaded. These non-pathogens infect the mosquitoes when it feeds on infected individual persons. In the host of the mosquito, the gametocytes keep on fusing and continue sexual reproduction allowing the overall cycle to go on and on [17, 18].

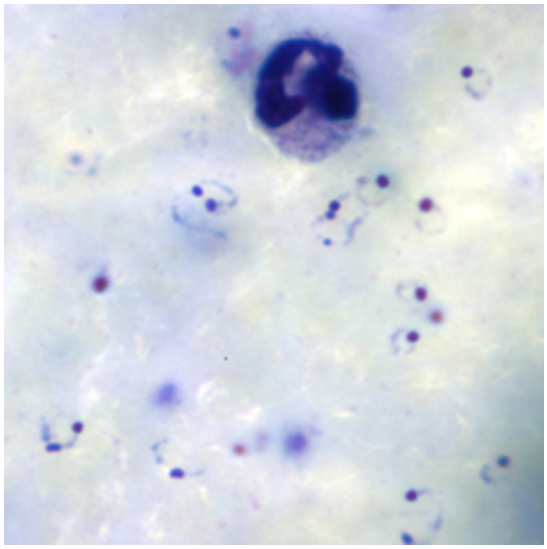
1.2.1 Various stages of the IDC

In Pf infections, the RBCs are sized normally. Generally, we can only see gametocytes and rings unless the blood settled before the smear preparation. The different stages of the IDC are depicted next [19].

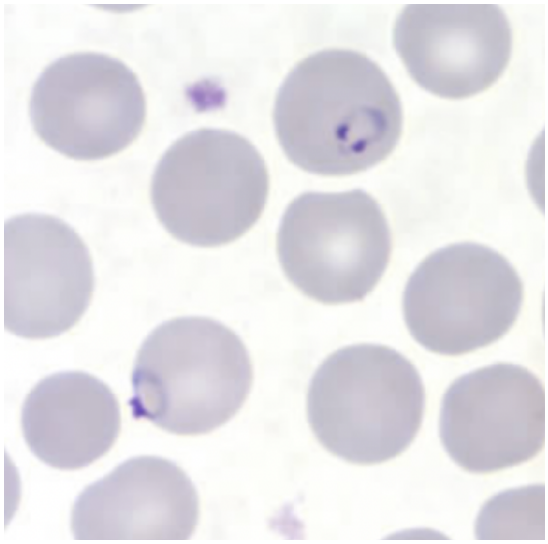
1. **Rings:** The cytoplasm of Pf rings is delicate and has a few small chromatin dots and the infected RBCs do not enlarge. A few times appliqué forms (rings appearing on the RBC periphery)[20] can also be found, as seen in Figure 1.2.



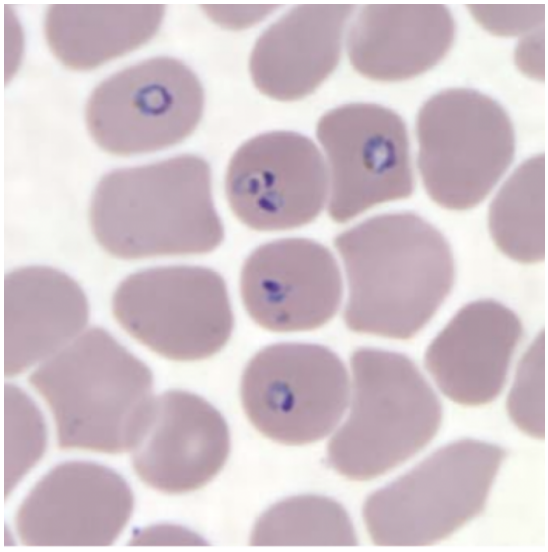
(a) Rings in a thick blood smear.



(b) Images from a thick blood smear. Note the classic ”headphones” appearance of many rings.



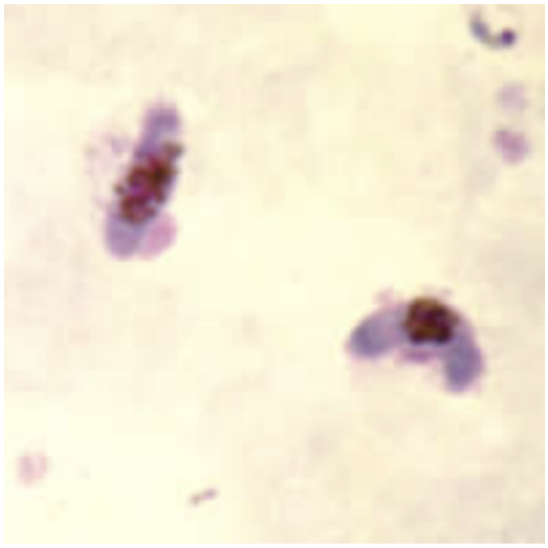
(c) Thin, delicate rings in a thin blood smear. Note the double chromatin dot in the infected RBC at the top, and the appliqué form in the infected RBC at the bottom [21].



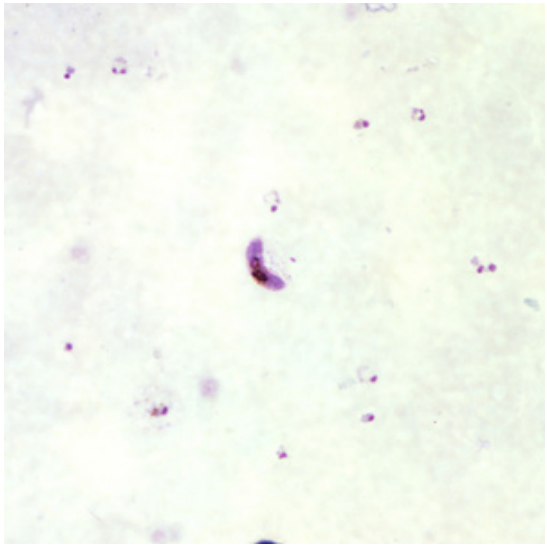
(d) Rings in a blood smear. Note the multiply-infected RBCs.

Figure 1.2: Ring stage of malaria parasite life cycle.

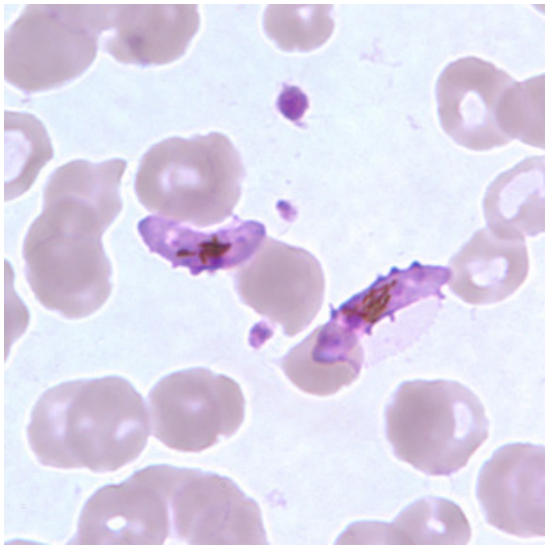
2. **Gametocytes:** Pf gametocytes are shaped like a sausage or crescent while the chromatin is in the form of a single mass (macrogamete) or diffuse (microgamete) (Figure 1.3).



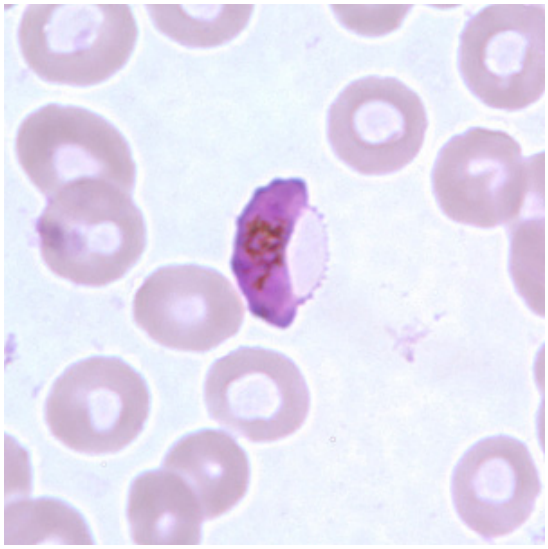
(a) Gametocytes in a thick blood smear.



(b) Gametocytes in a thick smear. Note also the presence of several rings.



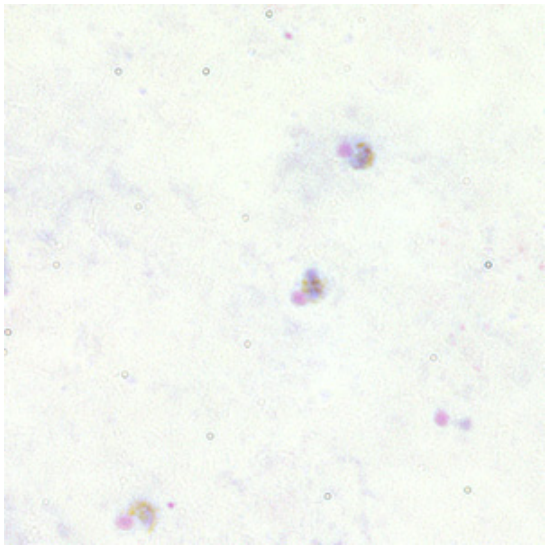
(c) Two Gametocytes in a thin smear.



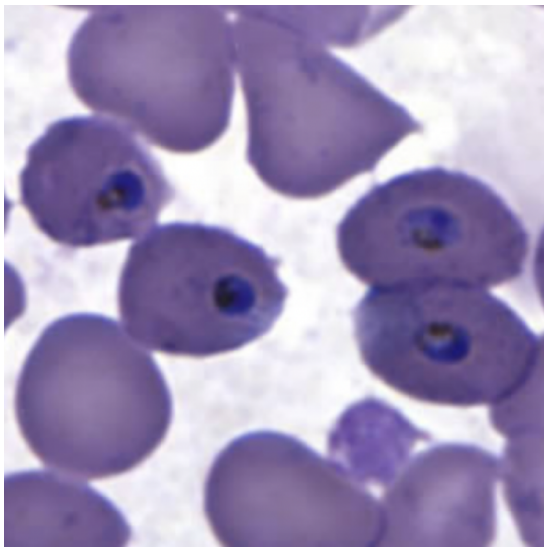
(d) Gametocytes in a thin smear showing the membrane of RBC.

Figure 1.3: Gametocyte stage of malaria parasite life cycle.

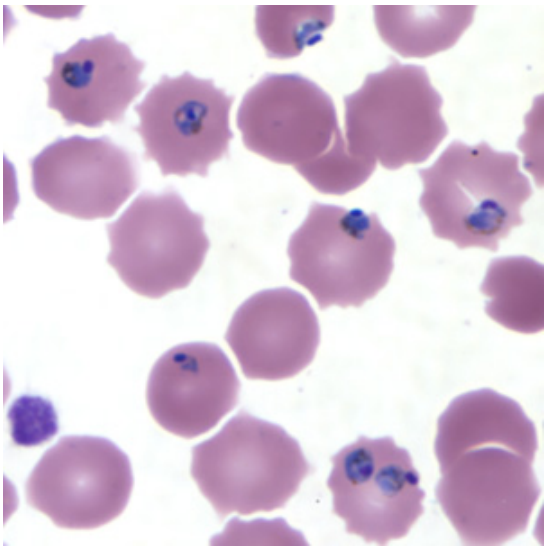
3. **Trophozoites:** Older, ring stage parasites or Pf trophozoites can rarely be seen in peripheral blood smears. The cytoplasm of younger rings tends to be less dense than in mature trophozoites. As Pf trophozoites grow, they continue to retain their ring-like amoeboid shape (Figure 1.4). At times traces of yellow pigment are also visible within the cytoplasm.



(a) Trophozoites in a thick blood smear.



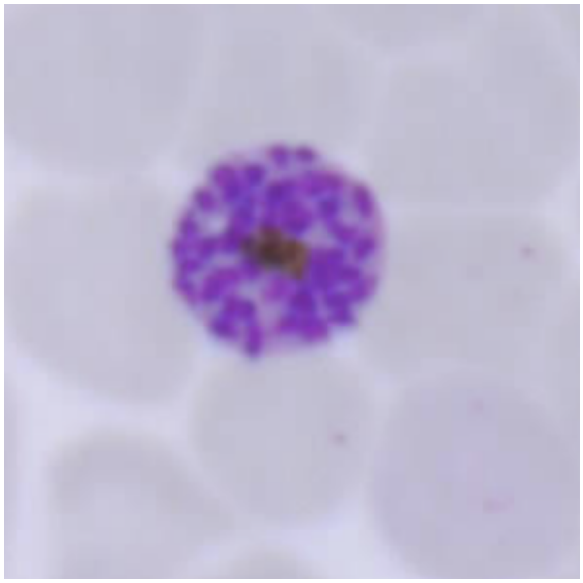
(b) Mature, compact trophozoites in a thin blood smear.



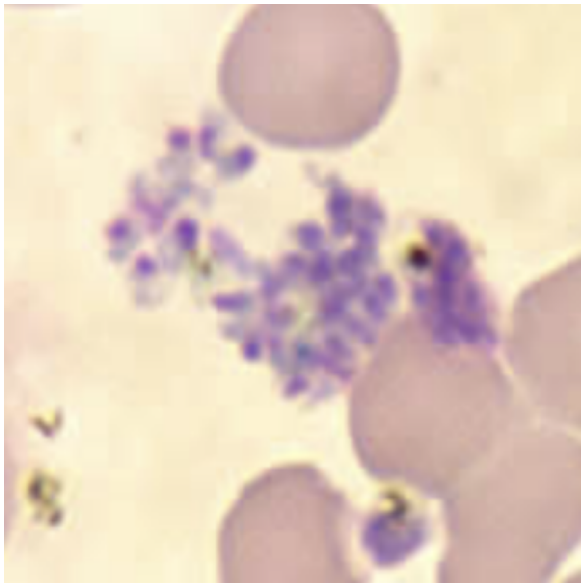
(c) Compact trophozoites in a thin blood smear.

Figure 1.4: Trophozoites stage of malaria parasite life cycle.

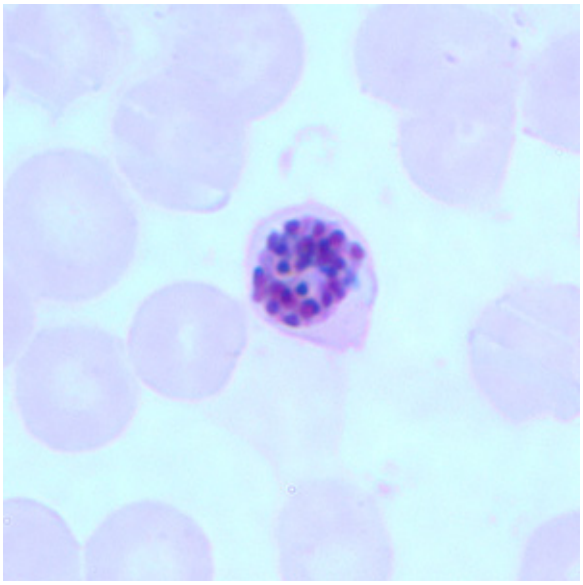
4. **Schizonts:** Pf schizonts are hardly visible in peripheral blood., dark pigment, grouped in one mass can be seen in mature schizonts (Figure 1.5).



(a) Mature schizont in a thin blood smear.



(b) Ruptured schizont in a thin blood smear.



(c) Another schizont in a thin blood smear.

Figure 1.5: Schizont stage of the malaria parasite life cycle.

1.3 Plasmodium invasion proteins: Structural aspect

An insight into the mechanism of interaction, invasion, and inhibition at the interface of the host–parasite is provided by the structural details of Plasmodium.

The structure of Pf is not fixed and keeps changing continuously during the entire life cycle. The spindle-shaped sporozoite of 10–15 μm length grow into 30–70 μm diameter ovoid schizont. Each schizont then continues to produce merozoites of 1 μm diameter and 1.5 μm length. Merozoites of the erythrocyte form a ring structure to turn in to a trophozoite feeding on the haemoglobin and forming the granular pigment haemozoin [22, 23]. Contrary to the other species of Plasmodium, the Pf gametocytes are crescent-shaped and elongated (3–6 μm wide and 8–12 μm long), which sometimes helps in their identification. The ookinete is further longer in size (18–24 μm nearly) while an oocyst is round in shape and grows in diameter up to 80 μm [7].
When a blood film is examined under a microscope, only early stage ring-shaped trophozoites and gametocytes are found. While mature trophozoites or schizonts are typically hidden in tissues in blood smears, occasionally, lighter comma-shaped red spots can be seen on the surface of erythrocytes. These ”Maurer’s cleft” dots are secretory organelles and proteins that produce enzymes and are necessary for the processes of immune evasion and nutrition uptake (Figure 1.6) [24].

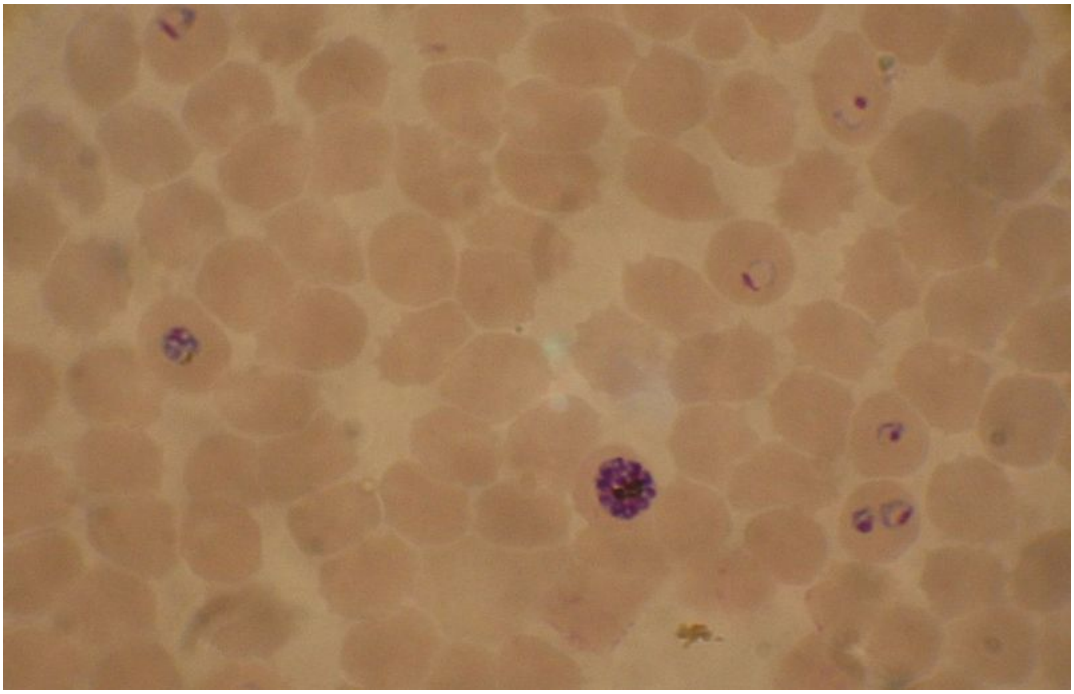


Figure 1.6: Blood smear from a Pf culture.

1.4 Literature Review and Treatment

Malaria continues to be a major concern for humans in many tropical and subtropical regions. Just in Africa itself, millions of children die annually from this disease attributed to the Pf parasite species. The disease becomes severe when Pf modifies the surface of infected RBCs by inserting parasite proteins. By the process of cytoadherence, these parasitized erythrocytes bind to the host endothelial cells in the brain, leading to the occurrence of cerebral malaria [25]. The current day malaria control and treatment techniques include bed nets impregnated with insecticide and chemotherapy. Although large scale efforts have been made towards developing a vaccine, none of the immunization approaches existing so far are effective. Moreover, with the growing resistance against existing antimalarial drugs, reliable prophylaxis is impossible, making the cure of malaria even more difficult. As a collaborative effort, the Malaria Genome Sequencing Consortium was setup to alleviate these problems, by sequencing the entire genome of Pf. "The consortium aims to generate almost all of the Pf genome sequence in unfinished form and making its entire gene complement accessible for malaria researchers" [26].
Oral medicines like Artemisinin may be used to treat simple malaria. The combination of ACT artemisinin and other anti-malarial drugs remains the most effective treatment for Pf infection. This decreases single drug component resistance. Artemisinin-naphthoquine combination therapy is also used to treat falciparum malaria. However, more research is needed for a reliable treatment. In low transmission settings, the performance of Artesunate+Mefloquine is better than that of Mefloquine. Atovaquone-Proguanil is effective against uncomplicated Pf with 5% to 10% possible failure rate while Amodiaquine+Sulfadoxine-Pyrimethamine is said to display lesser treatment failures in uncomplicated Pf malaria. Studies on treating uncomplicated malaria with Chlorproguanil-Dapsone are scarce. The combination of primaquine and artemisinin-based combination therapy for falciparum malaria reduces transmission on days 3-4 and 8 after infection [27, 28]. Chloroquine remains the mainstay of treatment for Plasmodium vivax malaria despite increasing reports of treatment failure [29]. For quick and successful patient recovery, it is crucial to diagnose and quickly treat malarial infection. If the malaria life cycle stages are somehow ascertained then the treatment of disease becomes easier. Various techniques are being used to examine malaria, including:

1. rapid diagnosis tests (RDTs)

2. quantitative buffy coat

3. peripheral blood smear microscopic examination, etc.

In majority of developing countries, the stained blood smears microscopic examination is yet considered to be the standard diagnostic method [30, 31]. Experienced medical professionals frequently examine a large number of blood films to detect malaria infection. Microscopists normally visualize the thin and thick blood smears to identify a disease or its cause. However, the accuracy depends upon the quality of smear and expertise to classify and count the parasite and non-parasite cells. It is fairly challenging

to number the parasites and infected RBCs manually and needs an expert microscopist for a quality diagnosis [32].

1.4.1 Machine Learning based Diagnostics

In Pf the classification is inefficient as the changes in geometric features are not seen. Colour pixel classification with feature characterization based on k-NN was reported for infected erythrocyte identification [33]. A semi-automatic method was presented by Diaz et al. [34] using SVM classifier to quantify and classify malarial erythrocytes. Hu set of relative shape measurement, invariant moment, intensity histogram, Laplacian features, gradient features, co-occurrence matrix, flat texture, run length matrix was used for the characterization of erythrocytes by Spring et al. [35]. Khan et al. [36] used first order feature and Hu moment based on intensity histogram with FF-BPNN classifier while Soni et al. [37] published an automatic classification of erythrocytes and Pf.
A variety of image processing techniques are used for the diagnosis and stage detection of the malaria parasite. This diagnosis is carried out using textural and statistical features in blood images of the malaria parasite [24]. Nowadays, digital image processing and machine learning techniques are contributing to a higher diagnostic accuracy in ultrasound imaging, CT, microscopy, etc. Computerized diagnosis of malaria based on analysis of microscopic images of thin blood smear exists is reported in the literature [15, 38, 39]. Geometric and texture features are used for classifying the malaria infected erythrocyte as the morphological features of the infected erythrocyte change. Detection based on eccentricity features and relative size using feed forward back propagation neural network (FF-BPNN) was reported in [38].
Convolutional Neural Networks (CNNs) have been successfully used for automatic classification of malaria parasites from blood smear images, enabling quicker diagnoses. However, it was limited to binary mode as infected and non-infected. A review of CNN techniques used for malaria diagnosis, focusing on the data preprocessing, preparation and classification alongwith NN architectures, performance and properties is presented. The authors in [40] used CNN based deep learning models for attribute extraction and categorization. For achieving higher categorization accuracy, they selected certain dominating features including size, color, shape, and cell count, from the images. Similarly, a more effective two-stage approach based on CNN on a larger dataset was also proposed by [41]. It remains an especially challenging task to distinguish the multiple growth stages of parasites. Seng et al. [42] developed a deep learning approach for the recognition of multi-stage malaria parasites in blood smeared images using a novel deep transfer graph convolutional network (DTGCN). They reported higher accuracy and effectiveness compared to a wide range of state-of-the-art approaches.

CNNs when used for feature extraction exhibit better performance than learning from scratch and transfer learning approaches. However, research employing private datasets for training and testing the CNN models cannot be easily compared with other methods. Future research would employ available public datasets to allow comparison of proposed CNN methods. Multi-class CNN models for classifi-

cation of species and life stages of malaria-causing plasmodium are needed [43].

1.4.2 ML based Feature Classification

Various machine learning (ML) approaches have been proposed for accurate gene features classification. Karthik and Sudha [44], reviewed ML methods for classifying gene expression model or computational analytical structure for complicated diseases, by identifying several differentially expressed gene techniques. Numerous ML approaches have been proposed in the literature to enhance gene expression data classification such as clustering, classification, dimensional reduction, among others [45]. Training of ML models using initial high-dimensional features performs unsatisfactorily in practice and may result in network over-fitting and increased redundant information. This problem was addressed using the random forests classifier in [46, 47]. Hossain et al. [30] designed an effective variational quantum circuit (VQC-based) approach to recognize the existence of malaria in RBC images through the classification of optimized feature set extracted from them. Murad et al. [48] used algorithms based on multifilter and hybrid approaches to feature selection, leading to accuracy in excess of 90%. Mei et al. [49] suggested a dimensionality reduction method to classify tumor gene expression data. Arowolo et al. [50] and Li et al. [51] proposed a dimensionality reduction approach for classifying gene expression.
To overcome the dimensionality problem, Rokach et al. devised a genetic algorithm-based feature selection method. They evaluated the fitness function of several obvious tree classifiers using a new encoding approach [52]. Zhang et al. proposed a classifier ensemble with feature selection based on GA. The authors of this work created a new hybrid method that combines a multi-objective genetic algorithm with an ensemble of classifiers. The GA-ensemble approach was tested on a variety of datasets and its performance was compared using a variety of classifiers [53]. Cheng-Lung Huang [54] suggested a feature selection method based on GA and SVM optimization. The ultimate goal was to improve the SVM classification accuracy while optimising the feature subset and parameters. Chaung et al. [55] employed a hybrid technique that began with a genetic algorithm with a dynamic variable to pick a sample of genes, which were then ranked using Chi Square analysis, and the level of accuracy of the selection was assessed using SVM. Shutao et al [56] used Particle Swarm optimisation and GA in order to perform highly accurate classification. The authors in [57] achieved a classification accuracy of 90.32 % using GA for feature selection and a SVM Classifier.

1.4.3 Single-cell RNA sequencing

The Pf genome is sized at 23.3 Mb and is encoded by over 5400 genes [15]. Transcriptional regulation of a majority of parasite genes occurs during the IDC. The same is seen at multiple points of time but has a maximally abundant single peak per gene. cDNA microarray technology was employed in the initial analysis of Pf IDC. RNA-seq when applied initially to the Pf IDC led to gene model alterations in more than 10% of the 5400 genes, as also 121 new coding sequences were identified [32]. By this

research 84 alternate splicing cases and 75% of predicted splice sites were also confirmedly detected. However, due to the limitations of the then RNA-seq technology, the extremely AT-rich UTRs went undetected on a genome-wide scale. The probable reason for this was attributed to a multiple difficulties generating PCR bias and AT-rich cDNA against the AT-rich sequences [58].

Ribonucleic acid sequencing (RNA-Seq) is used to express multiple levels of transcripts simultaneously. Several tools are used to develop and classify in to several pre-defined classes, after studying the RNA-Seq data of pathogens or viruses, depending on their attributes. Various machine learning approaches are providing powerful toolboxes for classifying RNA-Seq data [38]. Single-cell transcriptomics is used to map the genes during the entire transmission cycle of the etiological agent of malaria, Pf. Single-cell RNA sequencing (scRNA-seq) has allowed us to solve the cell-level heterogeneity issue in complex cell populations [59, 60, 61]. In fact, recent studies have elucidated the role of heterogeneity in enabling a small fraction of Plasmodium population inside the human host to remain ready to enter into the mosquito host by making transition to gametogenesis stage [62]. Similarly, heterogeneity plays a crucial role in the Plasmodium stress response inside the RBC [16]. Off late scRNA-seq views of the entire life cycle of Pf, have captured at high resolution the fine-scale developmental transitions driving their progression [58, 40, 41, 42, 63]. The Malaria Cell Atlas [58, 41] data resource and website provides the scientists the option to investigate gene expression patterns in individual parasites across the different developmental stages and for a variety of parasite species.

The differences between individual cells of unicellular and multi-cellular organisms can have far-reaching functional consequences. The single-cell mRNA-sequencing methods developed recently allow the high-throughput, unbiased, high-resolution transcriptomic analysis of individual cells. This has opened newer vistas of biology by way of transcription dynamics, tissue composition, and regulatory relationships between genes. This goes on to provide additional transcriptomic information in comparison to traditional bulk cell populations profiling methods. Fast paced technological advances in the field of cell capture, phenotyping, bioinformatics and molecular biology seem to project a bright future in multiple medical and biological applications [64].

Single-cell RNA sequencing technology could be possibly used to further reveal how the parasite resists drugs or bypasses the immune system. This may also help in the future in developing drugs that disturb the synchronisation of these genes, disrupting the parasite's development and further stopping the cause of disease [60, 65]. Recent studies in next generation sequencing (NGS) via RNA-Seq have enabled the simultaneous measurement of expression levels of multiple transcripts [38]. RNA-Seq is a powerful transcriptome profiling NGS technique which provides in-depth details of RNA transcripts. The RNA-Seq analysis methodology in general involves the following steps:
1) raw sequence disease data expression analysis, after its normalization;
2) individual network modules construction for gene identification; and
3) examination of their aggregate network properties [40].
This approach assists in the identification of multiple disease related genes which can then assist as a target in the further process of drug discovery. RNA-Seq readings can be exon, gene or other regions-

of-interest. These expression levels have led to the development of various classification algorithms. However, the analysis of complex datasets using computational tools is still inadequate, and there is no single technique for classifying and analysing the RNA-Seq data. Additionally, the high dimensional NGS data, including RNA-Seq, makes the problem even more challenging. The development of classification models for molecular level disease monitoring, diagnosis, and classification of diseases and research into potential disease biomarkers is highly desired.

1.4.4 Conclusions from Literature Review

Parasites of malaria can be easily identified by analysing the digitized microscopic blood smears. The same is, however, error prone, time consuming, and tedious. As a result, automation of the diagnostic process is critical to reducing the time-consuming manual review and diagnosis process. The classification of healthy and unhealthy cells into four types according to the respective life stages was done using as many as 11 CNN-based deep learning models. The robustness of the models was evaluated by experimenting on two cross-datasets of different type.

Whereas ResNet-18's accuracy in binary classification was 97.68%, DenseNet-201's accuracy in multi-class classification was 99.40%. The cross-dataset experiments highlight the lack of robustness in the deep learning approach as a weakness. Mobile-oriented architectures seem to be promising and performed satisfactorily in the recognition and classification of type and stage of malaria parasites [66].

1.5 Motivation

The complex life cycle of malaria parasites features diverse developmental strategies for unique adaptation in specific host environments. The main objective of our study is to select the top-ranked genes from the scRNA-seq profiles at different stages of the Pf life cycle inside the infected RBC. We employ a supervised learning algorithm coupled with feature selection algorithms to extract the most relevant genes to predict the life cycle stages of plasmodium inside RBC. The first stage of the proposed model is to optimize the quality data from the dataset by removing the redundant, noisy and irrelevant genes (features). From the systematic literature review it can be concluded that:
1) GA based search technique is popularly used for feature selection.
2) GA is well suited for dealing with high dimensionality datasets.
3) GAs showed better performance than the other selection algorithms.

Hence, GA is employed in our proposed approach as the search algorithm in the feature selection process. This subset can be further utilized in the second stage of the process, Classification, to produce high classification accuracy. We tested the subsets using three classifiers: SVM, LR, and RF to ensure that the investigation is carried out rigorously. The combination of the first and second stages of the proposed model will achieve a better identification of the different Malaria Life Cycle stages. Additionally,

the feature selection method is able to identify genes that significantly change expression across the life cycle stages. UMAP projection of the cells based on these features supports the distinction of stages using these features. We constructed the protein interaction network of these genes and performed a set of topological analysis to provide hierarchies according to the importance of the genes in the network. These genes can be used for disease diagnosis and drug targets.
Feature extraction, selection and classification of malarial erythrocytes are the major issues faced. The features comprises of a combination of features such as the channel difference histogram, prediction error R-G color and the co-occurrence of binary linear pattern. Feature selection and classification of erythrocyte based on optimal feature set and a hybrid (k-NN, SVM and Naive Bayes) classifier was attempted here. The hybrid classifier is hoped to provide improved results and sensitivity compared to the three individual classifiers [7].
Our study presents a theoretical framework to select diagnostic markers and drug targets by implementing ML techniques on sc-RNA-seq data.

1.6 Dimensionality Reduction

The sc-RNA expression dataset is highly dimensional. The dataset contains redundant characteristics that behave as noise during model training. As a result, classification performance is degraded, and computing time is increased. Dimensionality Reduction (DR) techniques are required to eliminate redundancy and to retrieve irrelevant details that hinder performance. When the data contains a large number of features, a model can become more complex. Complex models tend to overfit the data. By lowering the number of features (dimensionality) in the data, DR reduces model complexity. The details of the two methods for reducing the dimensionality of data is shown in Figure 1.7:

1.6.1 Feature Extraction

Feature extraction is a way to find the most important features, traits, or attributes of a dataset. Data with several dimensions require the use of feature extraction to provide a more concise summary of its categorization. New variables are created as combinations of the original variables to reduce the dimensionality. I t is classified into two types: linear and nonlinear.

Linear feature extraction adopts data on a low-dimensional subspace. Matrix factorization is used to project them on this subspace. Principal Component Analysis (PCA) is a linear feature extraction technique that is mostly utilized for dimensionality reduction. It finds the principal components of the data using the covariance matrix, eigenvalues and eigenvectors. These components represent a portion of the variance in the data.

Several methods of nonlinear dimensionality reduction exist. For instance, mapping a low-dimensional surface to a high-dimensional can reveal a nonlinear nature between the features. To map the features onto higher-dimensional space, a lifting function is used. The relationship between the features in a

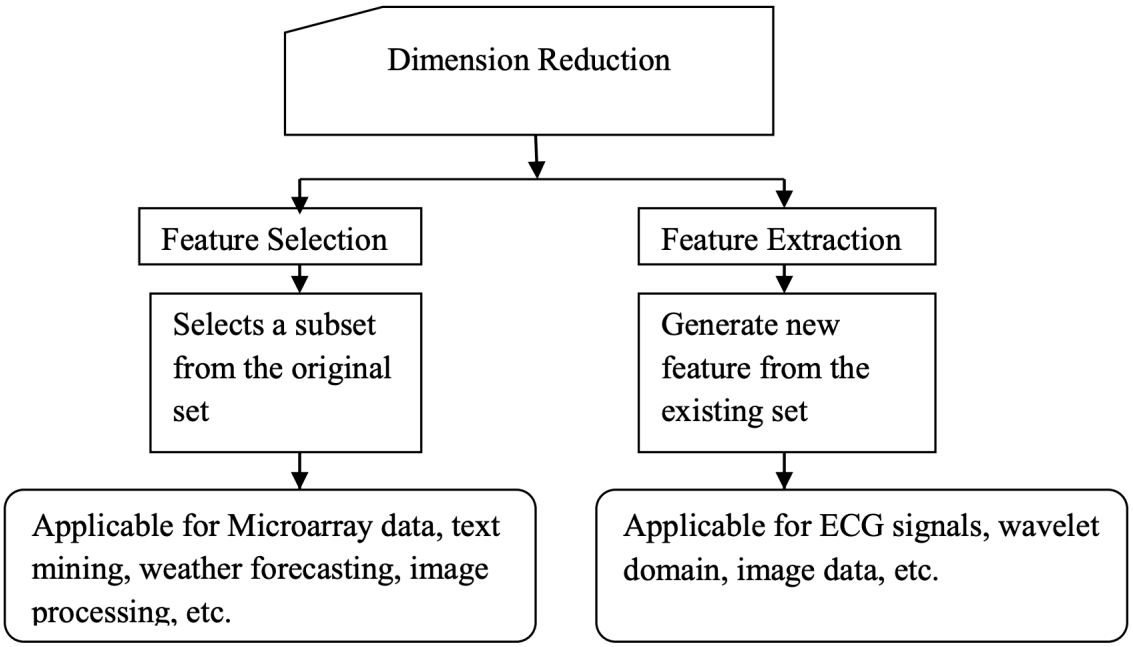


Figure 1.7: Dimensionality Reduction [1].

higher space is linear, and thus it is easily detected. This is converted back into lower-dimensional space, revealing the nonlinear nature. The difference between linear and nonlinear problems is shown in Figure 1.8.

1.6.2 Feature Selection

A feature in a dataset is a quantifiable property of the observed process. The feature (gene) selection method facilitates data comprehension, reduces processing requirements, alleviates the curse of dimensionality, and enhances the performance of the classifier. Feature selection is the process of removing noisy, and redundant features and identifying the relevant feature subset upon which the learning algorithm improves its performance [67]. The feature selection process is usually used on datasets that contain thousands of features with small sample size. Some of the main objectives of the feature selection process are to reduce overfitting, to provide fast and cost-effective models, and lastly, to gain a better insight into the underlying process that generated the data.
In addition to the standard method of feature selection, soft computing techniques (fuzzy logic, neural networks, and evolutionary algorithms are some examples) are used to choose features from high-dimensional data. Numerous evolutionary techniques, such as the genetic algorithm, the ant colony optimization algorithm, and the particle optimization algorithm, are commonly employed to achieve effective feature selection in high-dimensional datasets [68]. The filter, wrapper, and embedded methods are the three main types of feature selection methods [69].

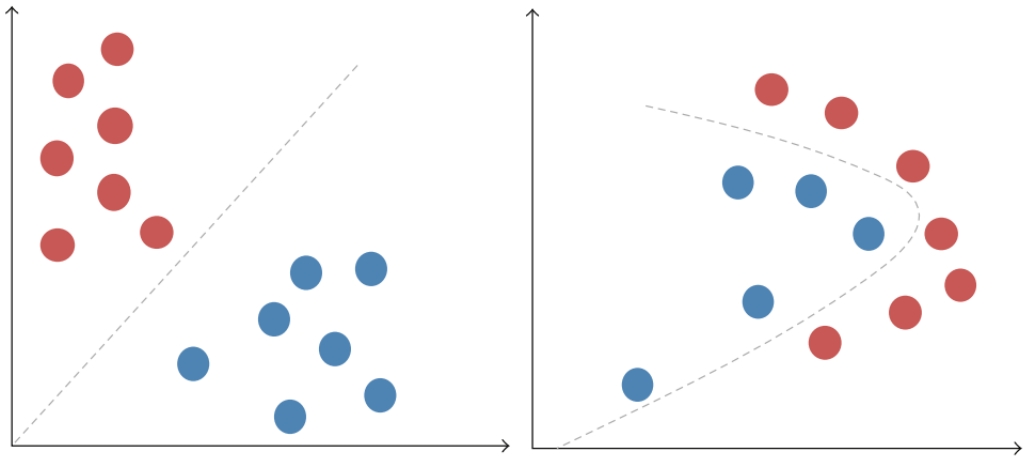


Figure 1.8: Linear versus nonlinear classification problems [2].

1.6.2.1 Filter Method

In this method, the features are filtered based on the intrinsic properties of the data. The low-scoring features are removed, and the optimal subset of features is input into the classification algorithm. This method does not consider a predictive model in the evaluation. It is fast and simple and scales well with high-dimensional data. It avoids overfitting, but sometimes does not pick the best features. Some of the ranking methods are as follows [70]:

- **Correlation criteria:** This is used to determine the linear relationship between two features. The Pearson's correlation coefficient is used to determine it:

$$R(i) = \frac{cov(x_i, y)}{\sqrt{var(x_i) * var(y)}} \tag{1.1}$$

  where cov is the covariance between the variables x and y and var is the variance of each of the two variables.

- **Mutual information:** This metric is used to quantify the interdependence of features. A value of 0 indicates that two qualities are unrelated.

1.6.2.2 Wrapper Method

This is a feedback method that uses a machine learning algorithm to help choose the best features. They use the performance of the classifier to figure out which features are good. They look through the space of feature subsets and figure out how well one learning algorithm will do for each feature from a feature set. To find the best subset of features, the wrapper method uses a blind search. There is no way to be sure that this is the best subset without getting all the possible subsets. It is hard to choose features in this way because it is NP-hard. It has a lower error rate. It selects a nearly perfect subset but

the method's primary disadvantage is that it is computationally costly. Because of the classifier, there is a risk of overfitting than there is with filter approaches.

1.6.2.3 Embedded Method

Despite the filter method's shorter computation time, a major disadvantage is that it is independent of the classifier, typically resulting in inferior performance to the wrappers. The wrapper model, on the other hand, has a significant computational cost, which is exacerbated further by the high dimensionality. An intermediate solution is to employ hybrid or embedded algorithms that leverage the classifier to provide criteria for ranking features. In comparison to the wrapper technique, embedded approaches are more tractable and efficient and has a lower risk of overfitting.

1.7 Thesis Outline

In this research, we use machine learning to predict the various stages of the malaria life cycle. In Chapter 2 we have briefly discussed the methodology, including the dataset, feature selection, and prediction models that have been used in our study. Further, we also discussed the pipeline used in our study. Chapter 3 presents a detailed account of the experiments performed and includes the results and discussions. In Chapter 4, we have summarised the major inferences and contributions of our research, as well as a discussion of the potential future directions.

Chapter 2

Methodology

2.1 RNA-seq Dataset details

The single cell RNA-seq dataset utilized here is derived from the Malarial Cell Atlas, an open source database of single cell transcriptomic data spanning the complete life cycle of malarial parasites. It is freely accessible through a dynamic, user-friendly web interface (www.sanger.ac.uk/science/tools/mca/mca/) [33]. For the current study, we considered the 10X scRNA-seq data of the intra-erythrocytic stages of Pf in the human host. The dataset has 5066 rows and 6737 columns. Each row corresponds to scRNA-seq read counts of a gene and each column corresponds to the same for a single cell. There are 5066 features in this dataset, which correspond to all the genes in each cell of the parasite. Additionally, each cell is assigned one label among the four blood cycle stages (i.e. ring, early trophozoite, late trophozoite, and schizont). Thus, we set out to utilize classification ML algorithms (see Section 2.3) which would allow us to predict the life cycle stage of a cell based on the gene expression pattern.

2.2 Feature Selection using Genetic Algorithm

The idea behind evolutionary computation was that it could be utilised as a tool for optimization and that solutions to problems could be evolved through the application of natural selection operators [71]. Earlier methods required describing jobs as finite-state machines and conducting mutations by changing the state diagrams randomly. Genetic Algorithms (GA) were devised by John Holland as a population-based algorithm. GAs work with a population of individuals that represent potential solutions to a given problem. Each individual or chromosome is evaluated using a fitness function that measures how well-suited this chromosome is as a solution. The best-fitting parents that survived crossover to produce offspring. It is necessary to perform mutations in order to prevent GA from being caught at a good but not ideal solution. The chromosomes that have survived in the population are the most optimal solutions after many generations of evolution. [72]. As shown in Figure 2.1, the basic steps followed in GA are as follows [73]:

1. Randomly initialize a population of n chromosomes.
2. Evaluate fitness of each chromosome.
3. The best fitted individuals create new chromosomes by crossover and mutation.
4. Evaluate the new chromosomes.
5. If the termination condition is reached then return the optimal solutions, otherwise go to Step 3.

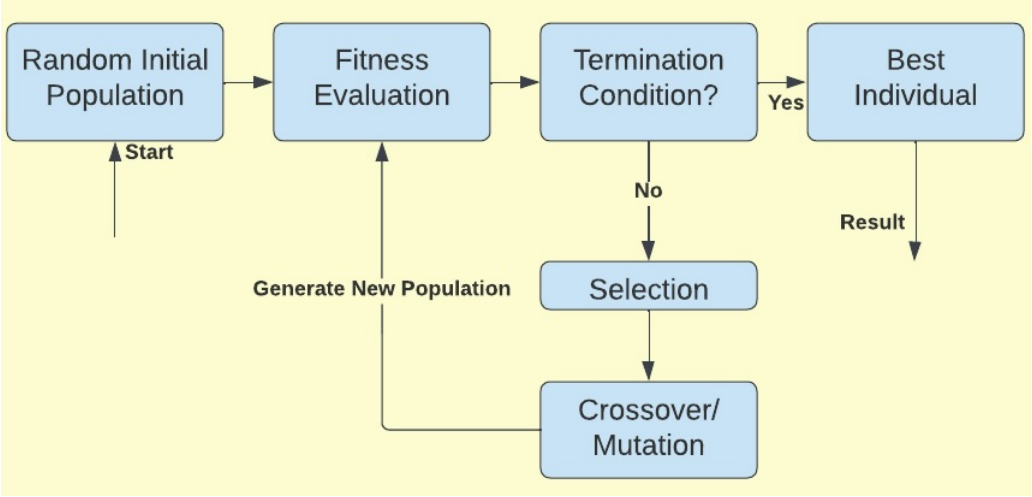


Figure 2.1: Basic steps of Genetic Algorithm.

2.2.1 GA Parameters

In this section, we will discuss the basic GA parameters [74].

2.2.1.1 Representation

GAs can employ any representation of individual genomes. The binary-coded GA is the most commonly used type of GA and is used for most of the development effort. Each chromosome in a binary coding system is a vector comprising of 0s and 1s. A value of one indicates that the feature is currently selected, whereas a value of zero indicates that it is currently not selected. As shown in Table 2.1, Features 1, 3, 4 and 6 are selected, while Features 2 and 5 are not selected for that chromosome. These features are also called genes of the chromosome.

	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6
Chromosome	1	0	1	1	0	1

Table 2.1: Representation of the chromosomes for a dataset containing 6 features.

2.2.1.2 Initialization

It is important to start off the population with a wide variety of individuals or chromosomes to avoid premature convergence. Usually, the first population is chosen at random, and the subsets are made from the search space with a uniform distribution.

2.2.1.3 Selection

The Selection operator selects which individuals will be copied into the mating pool, from which the next generation will be readied. The selection is performed to yield a mating pool with the same size as the original population. The mating pool then serves as the parents on which genetic operators (crossover and mutation) are used, interchanging and modifying the gene sets to generate a new generation. The probability of an individual being replicated into the mating pool is determined by the fitness function of the individual. Individuals with a higher fitness level have a higher chance of being selected.

Some of the commonly used selection methods are tournament selection, rank selection and roulette wheel selection. In the K-Way tournament selection procedure, we randomly select K individuals from the population and choose the finest among them to become parents. The same procedure is followed to choose the next parent. The individual with the maximum fitness level will win (Figure 2.2). In rank

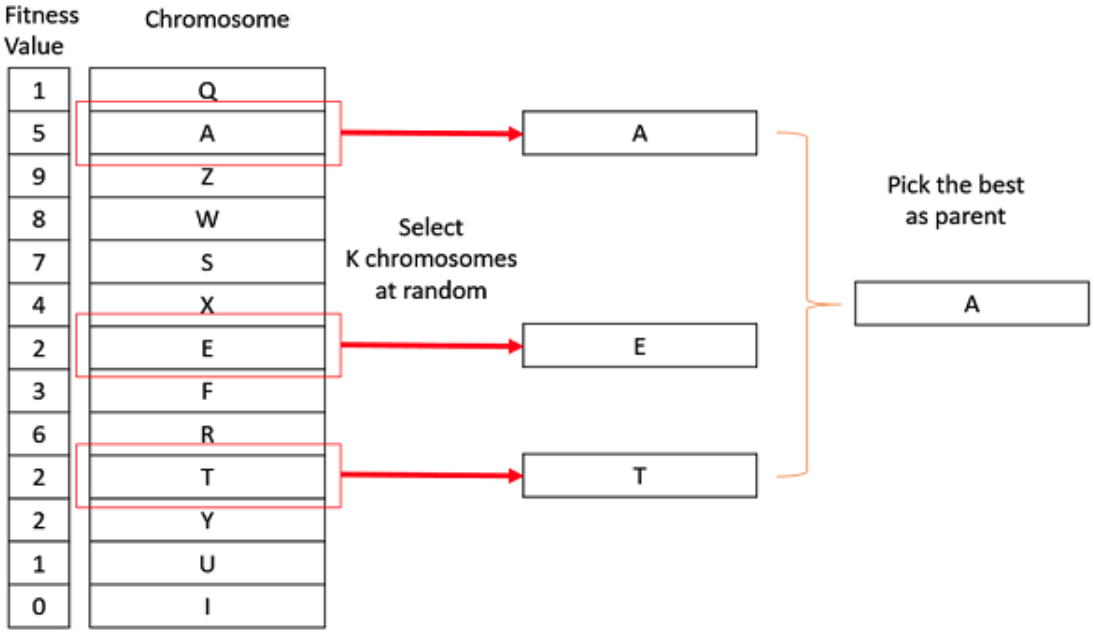


Figure 2.2: Tournament selection [3].

selection, every individual in the population is ranked according to its fitness. The selection probability depends upon their rank and not their fitness. Those who are ranked higher are favoured over those who are ranked lower. In a roulette wheel selection, the roulette wheel is divided into n pies, where n equals

the population's total number of individuals. Each member receives a proportional share of the circle based on its fitness score. The wheel is turned around a fixed point on its circumference. The portion of the wheel immediately ahead of the fixed point is designated as the parent. An identical procedure is repeated for the second parent. Clearly, a fitter individual has a larger pie on the wheel, which means they have a better chance of landing ahead of the fixed point. Because of this, an individual's chances of being selected are directly related to their fitness.

2.2.1.4 Fitness Function

The fitness function is a function that takes a candidate solution to the problem as an input and tells how "fit" or "good" the solution is for the problem at hand. The fitness function should have a smaller computation time. It must be able to measure how fit a given solution is or how fit an offspring can be made from that solution.

2.2.1.5 Crossover

Selecting the best chromosomes from the existing pool and putting them in the mating pool is the purpose of the selection stage. Crossover which is sexual reproduction occurs first, followed by mutation, in the mating pool. To provide the best possible offspring, two strings are randomly selected from the mating pool and crossed. Some of the different methods of crossover are:

- **Single Point Crossover**: On the parent strings, a crossing point is randomly chosen. After that point, the string is swapped between the two parents. As shown in Table 2.2, the parents P_1 and P_2 mate to produce children or chromosomes C_1 and C_2 using the single point crossover method. There are eight features in this example, and the crossover happens at the fourth index.

P_1	1	0	1	0	0	1	1	1
P_2	1	1	1	1	0	1	1	0
C_1	1	0	1	0	0	1	1	0
C_2	1	1	1	1	0	1	1	1

Table 2.2: Single point crossover operation.

- **Uniform Crossover**: A random selection of one of the parent chromosomes' genes is used for each gene. The crossover probability determines the parent at each gene position in this method. As shown in Table 2.3, the parents P_1 and P_2 mate to produce chromosomes C_1 and C_2 using the uniform crossover method.

P_1	0	0	0	0	0	0	0	0
P_2	1	1	1	1	1	1	1	1
C_1	1	0	0	0	1	1	0	1
C_2	0	1	1	1	0	0	1	0

Table 2.3: Uniform crossover operation.

- **OR Crossover**: In the OR method, we do the OR operation between the two parents, P_1 and P_2 to produce the offspring C. Table 2.4 shows the OR crossover operation.

P_1	1	0	1	0	0	1	1	1
P_2	1	1	1	1	0	1	1	0
C	1	1	1	1	0	1	1	1

Table 2.4: OR crossover operation.

- **AND Crossover**: In the AND method, we do the AND operation between the two parents, P_1 and P_2 to produce the offspring C. Table 2.5 shows the AND crossover operation.

P_1	1	0	1	0	0	1	1	1
P_2	1	1	1	1	0	1	1	0
C	1	0	1	0	0	0	1	0

Table 2.5: AND crossover operation.

- **XOR Crossover**: In the XOR method, we do the XOR operation between the two parents, P_1 and P_2 to produce the offspring C. Table 2.6 shows the XOR crossover operation.

P_1	1	0	1	0	0	1	1	1
P_2	1	1	1	1	0	1	1	0
C	0	1	0	1	0	1	0	1

Table 2.6: XOR crossover operation.

2.2.1.6 Mutation

It is possible that individuals in the population will undergo mutation as a result of the crossover operation that was used to produce the new offspring. A mutation is a small, random change to a value that happens with a very small probability, which is called the mutation probability. Mutation is critical because it keeps the population exploring the search space. It enhances population diversity by preventing the population from becoming saturated with identical chromosomes. Some of the different methods of mutation are:

- **Bit Flip Mutation**: We flip one or more random bits in bit flip mutation (Figure 2.3). This is used for GAs that are binary encoded.

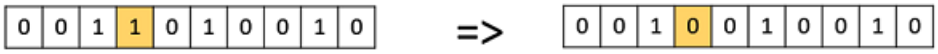


Figure 2.3: Bit Flip Mutation Operation

- **Swap Mutation**: In swap mutation, we randomly choose two locations on the chromosome and switch their values (Figure 2.4).

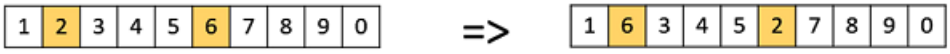


Figure 2.4: Swap Mutation Operation

2.2.1.7 Generate New Population

In GA there are numerous approaches of managing populations from one generation to the next. Given the limitation of maintaining a fixed population size, some individuals must be removed. Some of the methods for generation replacement are:

- **Elitist Replacement:** In this method, the fittest person in the population is always passed down to the next generation. This will guarantee that the best-fitted individuals are not destroyed.
- **Total Replacement:** In this method, only the new offspring of the previous generation enters the next generation, and the parents of the previous generation are completely discarded. So, in each generation, you get a new set of individuals.
- **Steady State Replacement:** At any given time, only a single population of individuals is maintained using this strategy. Two individuals are chosen from a population based on their fitness, and their characteristics are subsequently modified through mutation and crossover. The replacement operator then selects the chromosomes to be removed so that the newly created individuals can join this single population. You have the option of specifying how much of the population should

be replaced in each generation, as well as the replacement criterion that will be utilised in each generation.

2.2.1.8 Termination Condition

In GA the termination condition is critical in defining when the GA run will terminate. It has been noted that while the GA initially moves rapidly with improved solutions appearing every few iterations, this tendency tends to saturate in the later phases with very small improvements. Some of the termination conditions commonly applied include:

- When the population has remained unchanged for x iterations.
- When an absolute number of generations is reached.

2.3 Classification Algorithms

Classification is the process of identification of which of a set of categories or sub-populations an observation belongs to. Usually, the individual observations are grouped into a set of quantifiable properties called features. These features may either be categorical or ordinal, integer or real-valued. In the field of machine learning, observations are called instances, the variables termed as features are grouped to form a feature vector, and the to be predicted categories are called classes. We have used three classification algorithms in our study viz. Support Vector Machine, Logistic Regression and Random Forest [75, 76].

2.3.1 Support Vector Machine

Support Vector Machine (SVM) is a popular supervised learning technique that can be used for classification and regression tasks. To classify the data points, SVM finds a hyperplane in an N-dimensional space (N is the number of features). It sorts the data into two or more categories with the help of a boundary to distinguish similar categories. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as SVM.

As shown in Figure 2.5, the data can be classified into two categories, positive and negative. Let us assume that the blue sample (positive) is female and the green sample (negative) is male. Our goal is to differentiate between the males and females based on first studying the characteristics of both genders and then accurately labeling the unseen data. Our next idea is to find a line that separates the points. Let the equation of the line be:

$$mx + c = 0 \tag{2.1}$$

The hyperplane equation can now easily be written as:

$$w^T(x) + b = 0 \tag{2.2}$$

where b is the intercept and bias term of the hyperplane equation. If a blue point is substituted in this hyperplane equation, we will get a positive value mathematically:

$$w^T(x) + b > 0 \tag{2.3}$$

And the predictions from the negative group in the hyperplane equation would give a negative value of that number, i.e.

$$w^T(x) + b < 0 \tag{2.4}$$

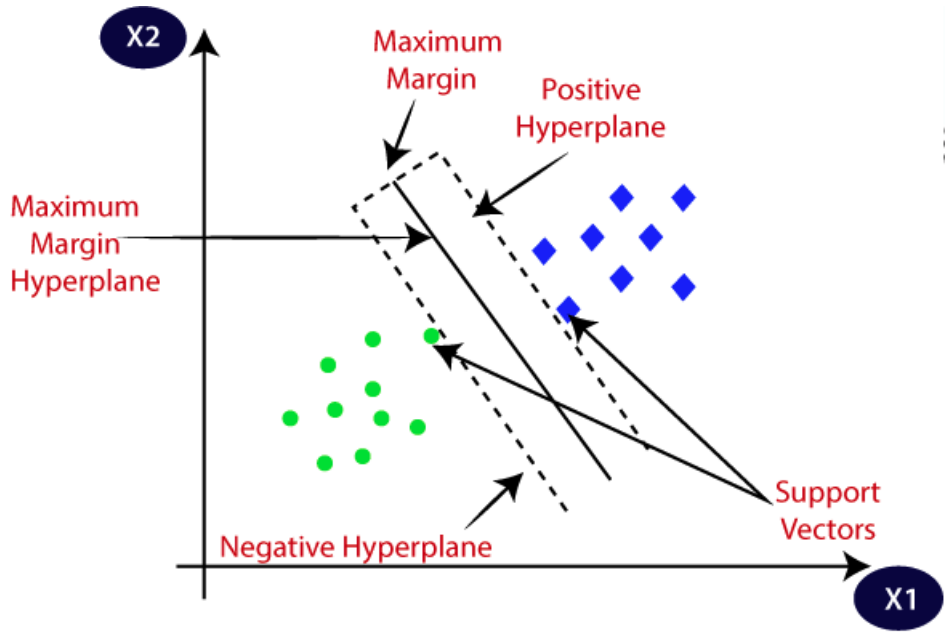


Figure 2.5: SVM classifier generating a two-dimensional line separating the two classes (green and blue) into two separate groups [4]

There is no situation when everything is perfect and sometimes the classes are not linearly separable. This means that we cannot expect the model to give us a hyperplane equation that is perfect for both genders. There will always be one or more points that do not fall into their category even when the best hyperplane equation is found. To solve this we can use kernels to convert the linear classifier to a non-linear classifier to help us solve the problem.

2.3.2 Logistic regression

Logistic regression (LR) is a classification procedure that uses a discrete set of classes to assign observations to them. Classification issues include determining if an email is spam or not, whether an

online transaction is fraudulent or not, and whether a tumour is malignant or benign, etc. [77]. Logistic regression translates its result into a probability value using the logistic sigmoid function (Figure 2.6). The sigmoid function is used to convert expected values into probabilities. The function converts any real value to a value in the range of 0 to 1.

$$\sigma(t) = \frac{1}{1 + e^{-t}} \tag{2.5}$$

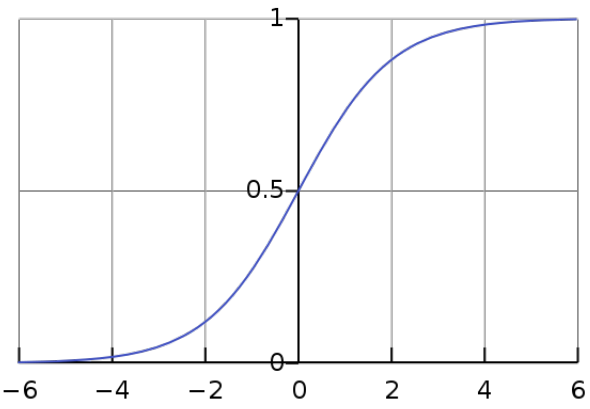


Figure 2.6: Sigmoid function [5]

where t is given by:

$$t = \theta \mathbf{X} \tag{2.6}$$

and X are the input features. In the case of multiple features it becomes:

$$t = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \theta_3 X_3 + ... \tag{2.7}$$

where X_1, X_2, and X_3 are input features, and each input feature will have a randomly initialised theta, θ_0 being the initial bias term. The objective of this algorithm is to continuously update the theta value in order to establish a link between the input data and the output label. We can use this to define our LR model f with some threshold value:

$$f(t) = \begin{cases} 1 & \sigma(t) \geq threshold \\ 0 & otherwise \end{cases}$$

2.3.3 Random forest

Random Forest (RF) is a supervised ML algorithm that is used for classification and regression tasks. It is a collection of decision trees constructed from a randomly chosen subset of training data. The steps followed in the RF algorithm are as follows:

1. Select random samples from the dataset.
2. Construct decision tree for every sample. The decision tree predicts the results.
3. Voting will take place for every predicted result.

Figure 2.7 explains the working of the RF algorithm.

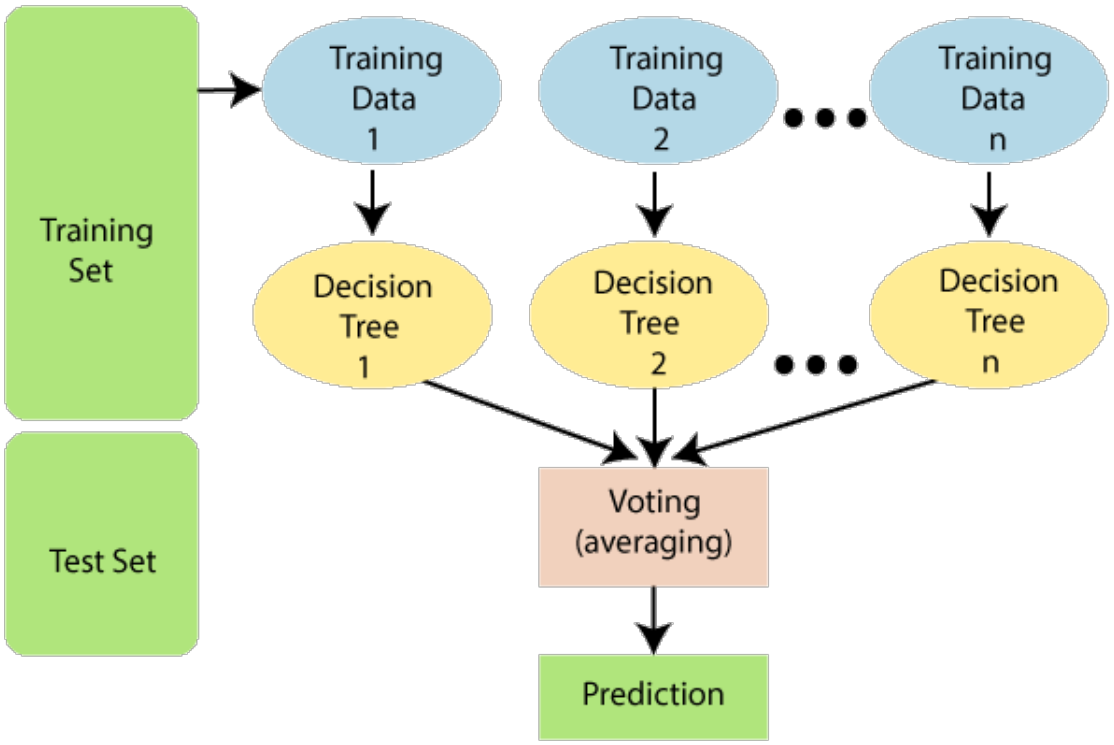


Figure 2.7: Working of the Random Forest algorithm [6]

2.4 System Design

The flowchart of the entire pipeline comprises the following stages (Figure 2.8):

1. Dimensionality Reduction with Feature Selection.

2. Classification Pipeline.

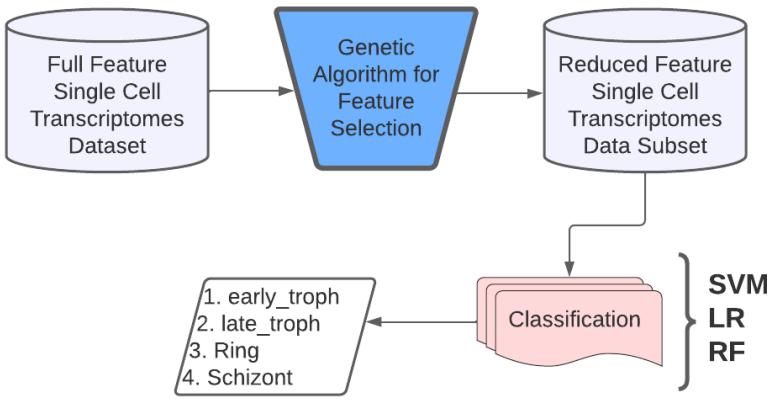


Figure 2.8: The system design of the entire pipeline.

The dimensionality of the gene expression dataset is high. The dataset has redundant features which act as noise while training a model. This results in poor classification performance and long computational time. Hence, the first stage is the feature selection stage using GA. This stage outputs Solution #1 as shown in the pseudo code in Figure 2.9. Solution #1 is then passed to stage 2 of the pipeline. These optimized features are then used to train different models. We have used the SVM, LR and RF models in our study. This yields the different evaluation metrics of the four different classes viz early_troph, late_troph, schizont and ring.

input : Training Set
Testing Set
output : Selected Features
Classification Accuracy

Begin: General Steps for feature selection using GA

Initialization:
$t \leftarrow 0$
Initialize population P_t randomly from training set.
EVALUATE_FITNESS(P_t)
while *termination condition not met* **do**
    Select individuals from P_t (fitness proportionate)
    Crossover of individuals
    Mutate individuals
    EVALUATE_FITNESS(modified_individuals)
    $P_{t+1} \leftarrow$newly created individuals
    $t \leftarrow t + 1$
end
Solution#1 $\leftarrow$Individuals with maximum fitness

Function `EVALUATE_FITNESS(`P`):`
    $clf \leftarrow Random\_Forest\_Classifier()$
    for *each individual* $i \epsilon P$ **do**
        $fitness(i) \leftarrow accuracy(clf(i))$
    end

End Feature Selection Process

Begin: Classification Process

Receive the best fitted individuals(Solution#1)
Compute the classification accuracy of the selected features using different classifiers(SVM, RF and LR)
Return the classification accuracy and evaluation results.

End Classification Process

Figure 2.9: Pseudo code of the proposed method.

2.5 Performance Evaluation Metrics

After selecting features and implementing a model and obtaining some outputs in the form of a probability or a class, the next step is to determine the model's effectiveness using test datasets. Various performance metrics are used to compare various ML algorithms [78]. In our study, we have used the following metrics:

2.5.1 Confusion Matrix

The confusion matrix is a table that compares the actual labels to the predictions made by the model. The confusion matrix is divided into rows and columns, with each row representing instances of the actual class and each column representing instances of the predicted class. In a sense, the Confusion Matrix is not a performance metric per se, but rather a foundation upon which other performance metrics are evaluated.

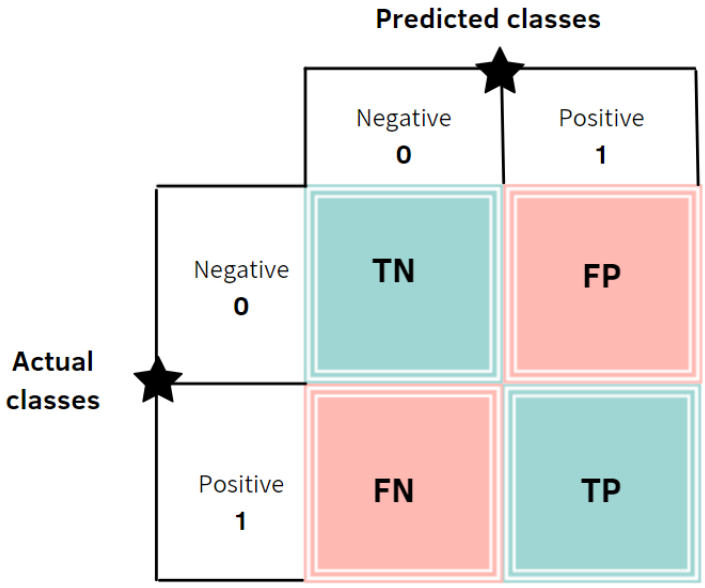


Figure 2.10: Confusion Matrix

As shown in Figure 2.10, 1 is a positive outcome and 0 is a negative outcome of some imaginary model. Some of the terms associated with Confusion Matrix are:

- **True Positive (TP):** This is the scenario when both the actual and the predicted class of the data point are 1.
- **True Negative (TN):** This is the scenario when both the actual and the predicted class of the data point are 0.

- **False Positive (FP):** This is the scenario when the actual class is 0 and the predicted class of the data point is 1.

- **False Negative (FN):** This is the scenario when the actual class is 1 and the predicted class of the data point is 0.

2.5.2 Mutual Information

Mutual Information (MI) is a measure of how much one variable's uncertainty is reduced when the other variable's value is known. It is given by the formula [79]:

$$I(X_1; X_2) = \sum_{X_1} \sum_{X_2} P(X_1, X_2) \log \frac{P(X1, X2)}{P(X_1)P(X_2)} \tag{2.8}$$

where $P(X_1, X_2)$ is the joint distribution of the two variables. $P(X_1)$ and $P(X_2)$ are the marginal distribution of the two variables. It is a dimensionless quantity that is measured in bits. Each element of the confusion matrix represents the conditional probability of predicting a class y' given the true category y - $p(y'|y)$. The joint probability of $P(y, y')$ is equal to the multiplication of the probability of the true label $P(y)$ and the conditional probability $P(y'|y)$. $P(y')$ is given by the sum of joint probability over true label y. We have used this to find the $I(y; y')$.

2.5.3 Accuracy

It is the most common way to measure the performance of classification algorithms and is defined by the ratio of true predictions to the sum of true and false predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.9}$$

2.5.4 Precision

Precision is given by the ratio between the True Positives and all the Positives.

$$Precision = \frac{TP}{TP + FP} \tag{2.10}$$

2.5.5 Recall

Recall attempts to calculate which portions of actual positives were correct. It can be defined mathematically as:

$$Recall = \frac{TP}{TP + FN} \tag{2.11}$$

2.5.6 F1 score

F1 score is the harmonic mean of precision and recall. F1 score is calculated mathematically as the weighted average of precision and recall. F1 has a maximum value of 1 and a minimum value of 0. It can be calculated using:

$$F1 = \frac{2 * precision * recall}{precision + recall} \tag{2.12}$$

2.5.7 Matthews' Correlation Coefficient (MCC) score

MCC returns scores between -1 and 1, with 1 indicating perfect classification performance and -1 indicating 100 percent inaccuracy. Random prediction is represented by an MCC value of zero. The coefficient can be calculated using:

$$F1 = \frac{TP \times TN - FP \times FN}{\sqrt{((TP + FN)(TP + FP)(TN + FN)(TN + FP))}} \tag{2.13}$$

The next chapter presents the details of the experiment carried out and the results obtained followed by their discussions.

Chapter 3

Experimentation and Results

The results in terms of the number of features selected and the output of the classification pipeline with and without feature selection is presented in this section.

3.1 Data visualisation and dimensionality reduction of the scRNA-seq data

We used Seurat [80], an R-based Bioconductor package, to visualise and then apply dimensionality reduction on the single cell RNA-seq data. We integrated the raw expression counts and metadata generated by Howick et al. [33], for downstream analysis to visualize the cells on a suitable manifold. Since the published data had already undergone quality control, the cell counts were subjected to normalisation using the 'LogNormalize' method of the Seurat package. This involves a global normalisation of cell counts with respect to the total expression, followed by log transformation. For further analysis, it is useful to focus on genes that exhibit high variation over all the cells in the dataset. Hence, we selected 1000 highly variable features (genes) from the data using the FindVariableFeatures() function. The data was then subjected to scaling before applying the standard dimension reduction techniques like PCA and UMAP. Next, PCA was performed on the data and the clusters produced from this linear dimensional reduction were annotated based on the blood cycle stages. Using the first 10 PCs, we also performed a non-linear UMAP-based dimension reduction on the cells for a better projection and annotated the clusters based on blood cycle stage. RunUMAP() function was used with dims = 1:10. Figure 3.1 represents the UMAP of scRNA-seq counts of all the 5066 features. Non-linear dimensional reduction of the expression values of all genes in the dataset is carried out. UMAP projection of the four clusters for ring, early trophozoite, late trophozoite and schizont can be noticed distinctly. The cell clusters are coloured based on the blood cycle stages of Pf.

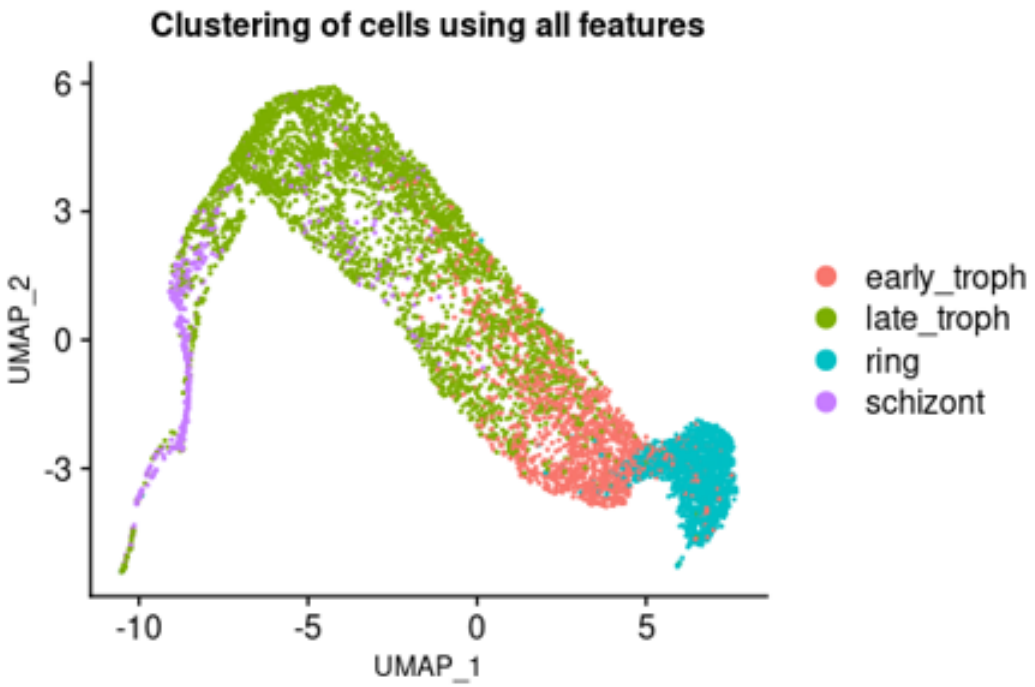


Figure 3.1: **Three dimensional visualization of the cells from the scRNA dataset shows distinct cluster of life cycle stages**. UMAP of cells based on scRNA-seq counts of all variable features. The cell clusters are colored based on the blood cycle stages of *P.falciparum.*

3.2 Classification without Feature Selection

This section presents the details of the classification results without using the feature selection stage. Using all the features, the datset was trained on SVM, LR and RF. Figure 3.2 shows the accuracy of SVM, LR, and RF. This is the baseline for our experiment. Without feature selection, SVM and RF performed best with a measured classification accuracy of 89%. The least accurate algorithm was LR (86%).

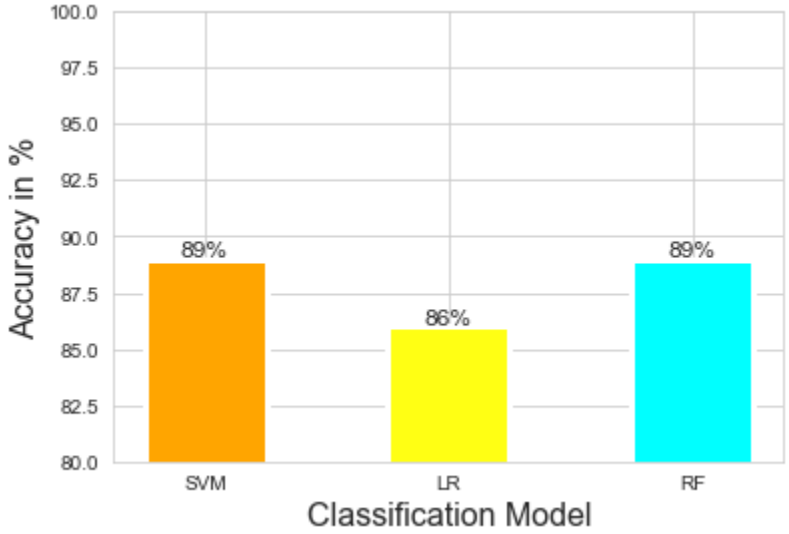


Figure 3.2: Classification accuracy of different models without feature selection. The classification accuracy are shown for different machine learning protocols namely SVM, LR and RF.

Table 3.1 presents the precision, recall, and F1 scores of the SVM model for the four classes. For the late_troph and ring classes, we achieved an F1 score of 0.91 and 0.95 each, while we got an average F1 score for early_troph as 0.83. Schizont's F1 score was the lowest at 0.74.

Metric (%) / Malaria Life Cycle Stage	precision	recall	F1-score
early_troph	0.89	0.77	0.83
late_troph	0.87	0.95	0.91
ring	0.93	0.97	0.95
schizont	0.86	0.64	0.74

Table 3.1: Test results of SVM model without feature selection.

Table 3.2 presents the precision, recall, and F1 scores of the LR model for the four classes. For the late_troph and ring classes we achieved an F1 score of 0.88 and 0.94, respectively, while we got an average F1 score for early_troph as 0.78. Schizont was again the lowest at 0.68.

Metric (%) / Malaria Life Cycle Stage	precision	recall	F1-score
early_troph	0.80	0.76	0.78
late_troph	0.86	0.91	0.88
ring	0.93	0.95	0.94
schizont	0.78	0.61	0.68

Table 3.2: Test results of LR model without feature selection.

Table 3.3 presents the precision, recall, and F1 scores of the RF model for the four different classes. For late_troph and ring we have achieved an F1 score of 0.91 and 0.95, respectively. We got an average F1 score for early_troph of 0.82. Schizont was found to be the lowest at 0.72.

Metric (%) / Malaria Life Cycle Stage	precision	recall	F1-score
early_troph	0.89	0.76	0.82
late_troph	0.87	0.94	0.91
ring	0.93	0.98	0.95
schizont	0.86	0.62	0.72

Table 3.3: Test results of RF model without feature selection.

3.3 Experiment

This section introduces each of the followed steps in detail. A summary of the implementation of the entire pipeline is depicted in Figure 3.3. In order to create an independent test set and improve the classification validity and accuracy, the input data was divided into the training and testing sets in a ratio of 80% and 20% respectively.

The training set was created to validate the feature selection while the test set served a similar validation role in the classification process. The training set is then processed through the GA pipeline [30]. GA is a stochastic evolutionary optimization technique. It starts with an initial randomized set of population of features and then creates another population using subsets of the available features whose individuals are evaluated using a predictive model for the target task. The selection technique is used to pick the higher fitness subsets to be carried forward into the next generation for applying the cross-over (updating the winning feature sets with features from the other winners) and mutation (probabilistically introducing or removing some features) genetic operators. This process is iterated to yield the optimum features for the set termination criteria. It is important to assess the performance of the GA pipeline.

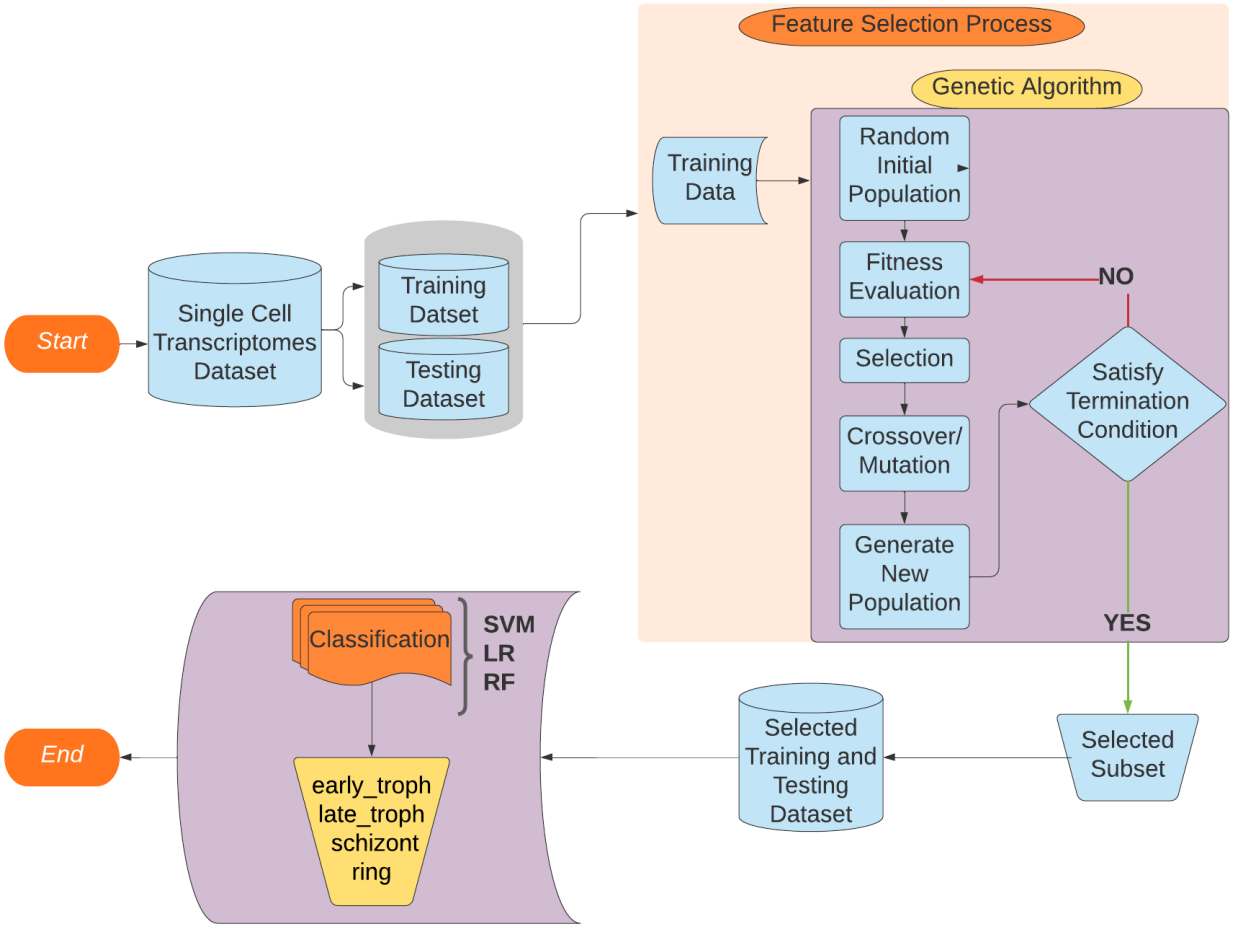


Figure 3.3: Flowchart of the entire pipeline.

How does the GA perform with different predictive models? Are the solutions obtained from the GA pipeline optimal? How many generations does it take to reach the optimal solution? How much time does it take to reach the optimal solution? We intend to address these questions in the coming subsection. Our study's objective is to identify the optimal set of features within a reasonable time. We performed two different experiments for that.

After the GA has selected the optimal features, these features are then subjected to different classification algorithms (SVM, RF, LR) to measure the classification accuracy of the selected feature set. This yields us the classification accuracy of the four classes viz early_troph, late_troph, schizont and ring.

3.3.1 Results and Discussion

In Experiment 1 we have used the RF as the predictive model where as in Experiment 2 we have used the LR as the predictive model. The common parameters used for both the experiments include:

- Initial Random Population: 50

- Selection Operator: Tournament Selection
- Crossover Operator: Uniform Crossover with crossover probability 0.5
- Mutation: Bit Flip Mutation with mutation probability 0.2
- Generate New Population: Elitist Replacement
- Termination Condition: Number of generations 50 and if no change in the best-fitted individual for 5 generations.

We ran both the experiments using a different number of features every time. The number of features used in the searches is 50, 100, 150, 200, 250 and 300. The results presented in Table 3.4 compares the accuracy rates achieved by both the experiments for 50, 100, 150, 200, 250 and 300 features. Comparing the numbers for both the experiments, we can infer that both the experiments have achieved a similar accuracy rate. Experiment 1 gave better accuracy for 150 and 300 features in comparison to Experiment 2. Experiment 2 gave better accuracy with 50 and 250 features when compared to Experiment 1. For 100 and 200 features, both the experiments were comparable. Next, we measured the time taken by the experiments to select the optimal features.

Number of Features	Best accuracy rate by Experiment 1	Best accuracy rate by Experiment 2
50	0.79	0.85
100	0.81	0.82
150	0.87	0.83
200	0.85	0.87
250	0.84	0.87
300	0.89	0.84

Table 3.4: Classification accuracy rates for Experiment 1 and Experiment 2.

The results presented in Table 3.5 compare the time taken by both the experiments for 50, 100, 150, 200, 250 and 300 features. The time taken by Experiment 1 is in the range of 309 - 709 seconds whereas the time taken by Experiment 2 is in the range of 3109 - 12203 seconds. As shown in Figure 3.4, Experiment 2 has taken significantly more runtime compared to Experiment 1 under the same environment.

Number of Features	Time taken by Experiment 1 (sec)	Time taken by Experiment 2 (sec)
50	709	2274
100	550	3109
150	431	3205
200	309	8155
250	309	12203
300	465	4274

Table 3.5: Time taken to select optimal features for Experiment 1 and Experiment 2.

We achieved similar accuracy rates, so we decided to prioritise the time taken for running the experiments. Hence, for all further experiments, we eventually decided to use the Random Forest model for the fitness function of the GA pipeline.

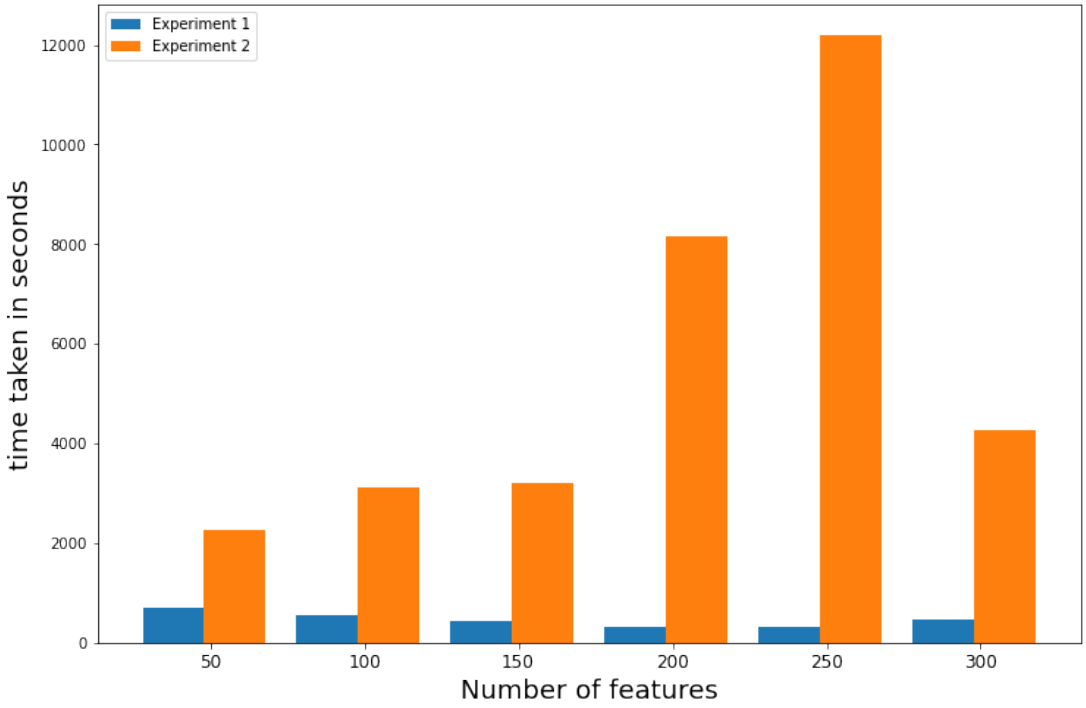


Figure 3.4: Time taken by Experiment 1 and Experiment 2 for different number of features.

3.4 Classification with Feature Selection

The main aim of our study was to remove the redundant noise in our dataset so that the overall classification accuracy can be improved. We used the GA pipeline with the following parameters for this experiment:

- Initial Random Population: 500
- Selection Operator: Tournament Selection
- Crossover Operator: Uniform Crossover with crossover probability 0.5
- Fitness Operator: MCC score of Random Forest model
- Mutation: Bit Flip Mutation with mutation probability 0.2
- Generate New Population: Elitist Replacement
- Termination Condition: Number of generations 100 and if no change in the best-fitted individual for 20 generations.

The number of features that we considered for this experiment were 500, 1000, 1500, 2000, 2500 and 3000. The results in Table 3.6 show the time taken and the accuracy rates for different number of features. We have taken the best accuracy results amongst all the three models viz. SVM, LR, and RF.

Number of Features	Time taken by experiment (sec)	Best accuracy rate by experiment
500	356	0.87
1000	449	0.89
1500	380	0.90
2000	368	0.90
2500	578	0.90
3000	682	0.90

Table 3.6: Time taken and classification accuracy rates for Classification with Feature Selection.

The total number of features in the dataset was 5066. The best classification accuracy achieved without feature selection was 89% (Section 3.2). We can see from Table 3.6 that as we reduce the number of features, the best accuracy achieved by our model is very similar. This proves that there is a lot of redundant data that can be removed from the dataset. Hence, we set the maximum number of features at 500 for our further experiments. This means that we will want to reduce the dataset from 5066 to at least 500. This leads to a reduction of the dataset by 90.1%. In the experiment in this section, we changed some of the parameters so that we could get better accuracy results for a maximum of 500 features. We changed the initial random population to 500 and we set the termination condition to 100 generations or 20 generations if there is no change in the generations. In the coming subsections, we will discuss the results of the GA pipeline with the maximum number of features set at 500.

3.4.1 GA convergence

We used the MCC scores of the Random Forest classifier as a fitness function for our experiment. Table 3.7 shows the maximum MCC scores achieved in each generation by GA. Generation 0 is the

Number of generations	Maximum MCC scores
0	0.81628465
1	0.82654331
2	0.83440726
3	0.83892729
4	0.84507526
5	0.84507526
6	0.84511502
7	0.84511502
8	0.84695042
9 - 28	0.84695042

Table 3.7: Maximum MCC scores achieved by the GA pipeline.

initial random individual selection with a maximum MCC score of 0.81628465. We can see that from generation 1 to generation 8, the maximum value of MCC scores of the population has increased from 0.82654331 to 0.84695042, indicating that the GA was able to reach a better optimal solution from generation to generation. After that, from generation 9 till generation 28, the maximum value of MCC scores remains the same (0.84695042) for 20 generations, indicating that the GA has converged to an optimal solution. This was the termination criteria set by us to end the GA pipeline.

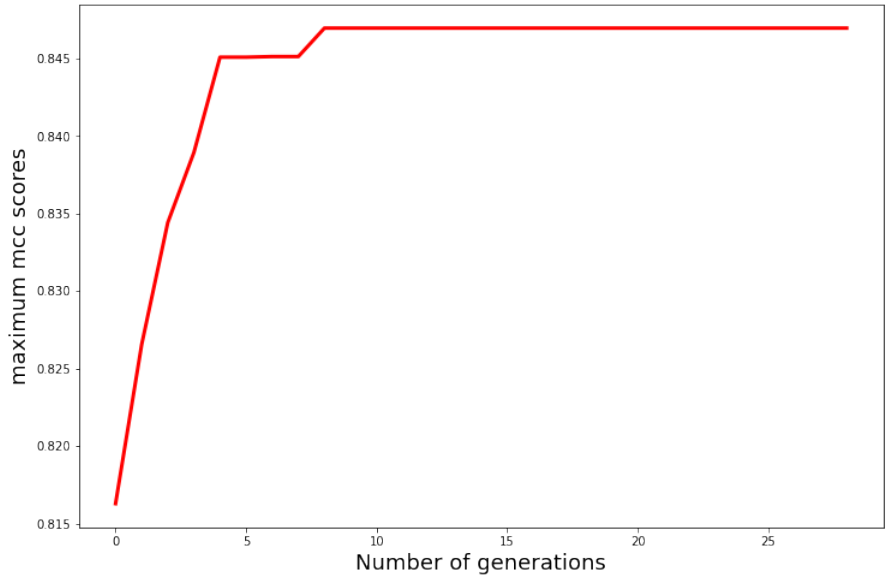


Figure 3.5: Maximum MCC scores Vs Number of GA generations.

We kept the maximum number of generations at 100, and the GA converged within this number of generations to reach the optimal solution. Figure 3.5 displays the relationship between the number of generations needed to reach the optimal value using GA. We can see that from generation 0 till generation 8, the MCC value increased, and then for the next 20 generations the value remained constant, indicating the convergence.

3.4.2 Number of features selected

The GA pipeline outputs the most optimal features and has removed the redundant features. Table 3.8 shows the details of the number of features. The initial number of features was 5066 out of which a subset of 378 features was selected using the GA. Thus the dataset was reduced by 92.5%.

	Number of Features
Full Dataset	5066
Features Selected after GA pipeline	378

Table 3.8: Numbers of Features selected

3.4.3 Classification Results

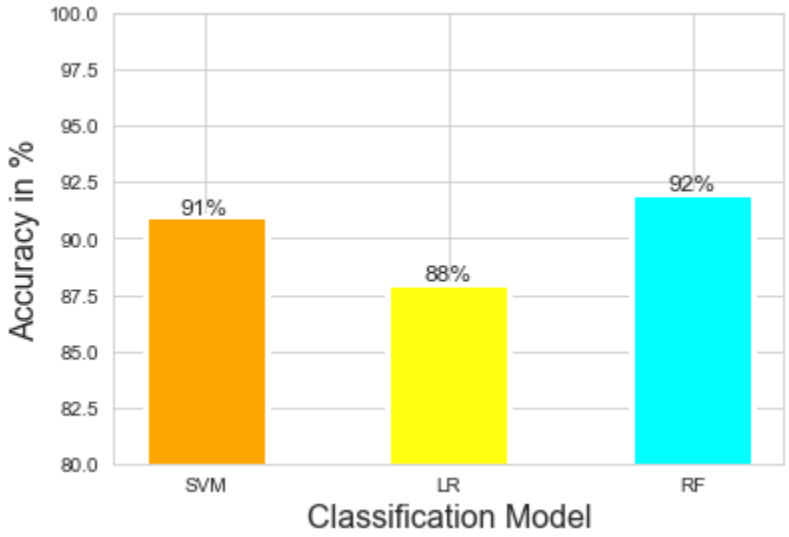


Figure 3.6: Classification accuracy of different models with feature selection. The classification accuracy are shown for different machine learning protocols namely SVM, LR and RF after selection of the 378 feature following genetic algorithm.

.

The next step was to use the GA based optimally selected 378 features in the classification pipeline. We trained our SVM, LR, and RF models using the optimal features. Figure 3.6 shows the classification accuracy of SVM, LR, and RF models. RF performed best with the classification accuracy measured as 92%. SVM and LR gave 91% and 88% accuracy respectively. The followup subsections present a detailed review of the test results for the SVM, LR, and RF models, respectively.

3.4.3.1 Using multi-class Support Vector Machine

Table 3.9 presents the precision, recall, and F1 scores of the SVM model for the four classes. For late_troph and ring, we have achieved an F1 score of 0.93 and 0.96, respectively. We got a F1 score for early_troph at 0.85. Schizont was the worst, at 0.79.

Metric (%) / Malaria Life Cycle Stage	precision	recall	F1-score
early_troph	0.91	0.79	0.85
late_troph	0.89	0.97	0.93
ring	0.94	0.97	0.96
schizont	0.91	0.70	0.79

Table 3.9: Test results of SVM model with Feature Selection.

3.4.3.2 Using Logistic Regression

Table 3.10 presents the precision, recall, and F1 scores of the LR model for the four different classes. For late_troph and ring, we have achieved an F1 score of 0.90 and 0.95, respectively. We got a F1 score for early_troph at 0.83. Schizont was the worst, at 0.68.

Metric (%) / Malaria Life Cycle Stage	precision	recall	F1-score
early_troph	0.86	0.79	0.83
late_troph	0.88	0.92	0.90
ring	0.94	0.96	0.95
schizont	0.74	0.63	0.68

Table 3.10: Test results of LR model with Feature Selection.

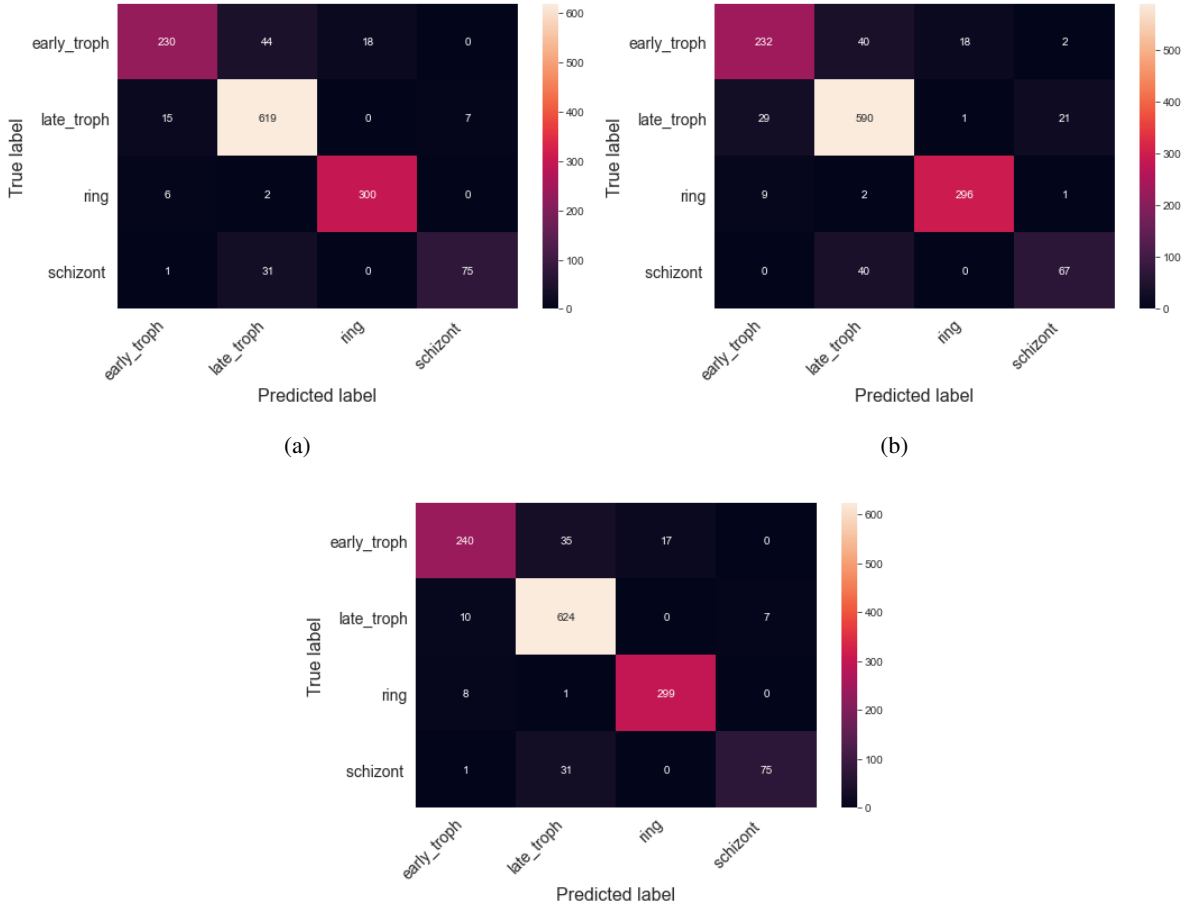
3.4.3.3 Using Random Forest

Table 3.11 presents the precision, recall, and F1 scores of the RF model for the four different classes. For late_troph and ring, we have achieved an F1 score of 0.94 and 0.96, respectively. We got a F1 score for early_troph at 0.87. Schizont was the worst at, 0.79.

Malaria Life Cycle Stage \ Metric (%)	precision	recall	F1-score
early_troph	0.93	0.82	0.87
late_troph	0.90	0.97	0.94
ring	0.95	0.97	0.96
schizont	0.91	0.70	0.79

Table 3.11: Test results of RF model with Feature Selection.

3.4.4 Confusion matrix for the three models



(a) (b)

(c)

Figure 3.7: Confusion matrix of different models show the prediction accuracy for different stages. The heatmaps display the confusion matrix in predicting the four different stages as indicated after feature selection for three different models (A) SVM (B) LR (C) RF models.

Figure 3.7 shows the confusion matrix along with the heatmaps in predicting the four stages after feature selection for the three models. The confusion matrix for the SVM model shows that 44 samples were predicted as late troph which should have been labeled as early troph. Similarly, 31 samples were predicted as late troph which were otherwise labeled as schizont. For late troph class, 15 samples were misclassified as early troph. For ring class, 6 samples were misclassified as early troph. The confusion matrix for the LR model shows that 40 samples were predicted as late troph which should have been labeled as early troph. Similarly, 40 samples were predicted as late troph which were otherwise labeled as schizont. For late troph class, 29 samples were misclassified as early troph. For ring class, 9 samples were misclassified as early troph. The confusion matrix for the RF model shows that 35 samples were predicted as late troph which should have been labeled as early troph. Similarly, 31 samples were predicted as late troph which were otherwise labeled as schizont. For late troph class, 10 samples were misclassified as early troph. For ring class, 8 samples were misclassified as early troph. These were some of the common misclassifications in all three models.

3.5 Comparison of classification with feature selection and without feature selection

Figure 3.8 shows a comparison between the classification accuracy of without feature selection vs. with feature selection for the three models. It demonstrates the legitimacy of the selected features. We have reduced our feature set from 5066 to 378, using which we achieved an improved accuracy of 91% in the SVM model, 88% in the LR model, and 92% in the RF model. For the SVM model, without feature selection, we got a F1 score of 0.83, 0.91, 0.95, and 0.74, whereas, with feature selection, we got a F1 score of 0.85, 0.93, 0.96, and 0.79 for early troph, late troph, ring, and schizont, respectively. For the LR model, without feature selection, we got a F1 score of 0.78, 0.88, 0.94, and 0.68, whereas, with feature selection, we got a F1 score of 0.83, 0.90, 0.95, and 0.68 for early troph, late troph, ring, and schizont, respectively. For the RF model, without feature selection, we got a F1 score of 0.82, 0.91, 0.95, and 0.72, whereas, with feature selection, we got a F1 score of 0.87, 0.94, 0.96, and 0.79 for early troph, late troph, ring, and schizont, respectively. Using the selected features, we achieved similar or better F1 scores across all four classes, in all three models. This proves the robustness of the features selected from the GA pipeline. For the early troph class, we achieved the best F1 score of 0.87 from the RF model. For the late troph class, we achieved the best F1 score of 0.94 from the RF model. For the ring class, we have achieved the best F1 score of 0.96 from both the SVM and RF models. For schizont class, we have achieved the best F1 score of 0.79 from the SVM and RF model. The schizont class has seen lesser F1 scores than the others, this could be because of the lesser number of schizont cells in the dataset.

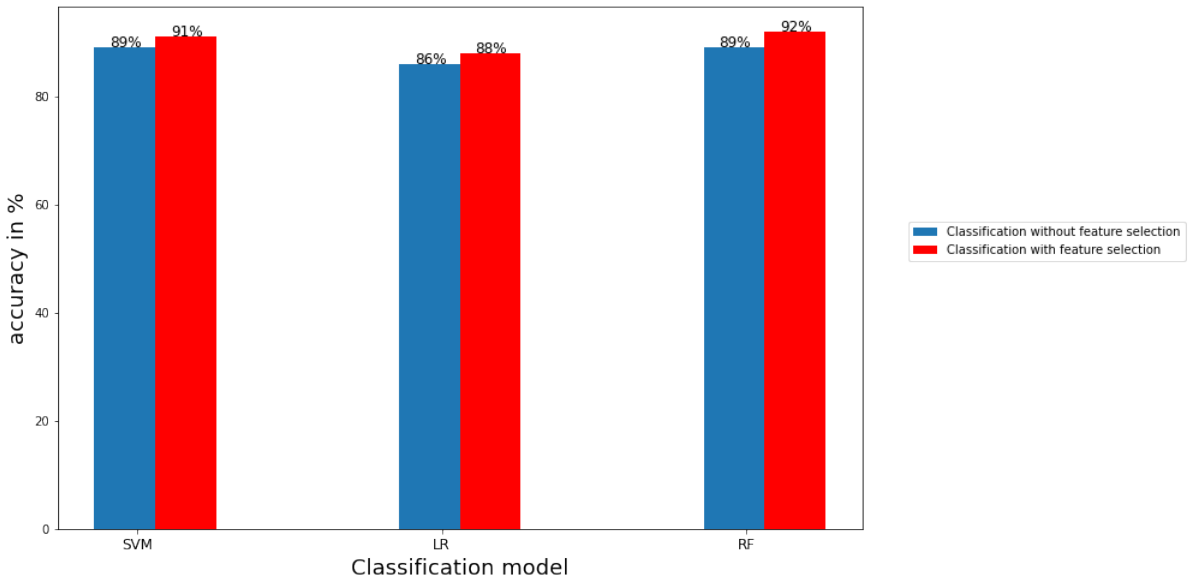


Figure 3.8: Classification accuracy with feature selection vs without feature selection demonstrate the legitimacy of the selected features. The bar graphs display a comparison between the values of accuracy for the three models and for classification with feature selection and without feature selection as indicated.

We also calculated the mutual information (MI) between the predicted labels and the true labels of the three models using the joint probabilities from the confusion matrix (see Subsection 2.5.2). For instance, C(1,1) of the confusion matrix represents the joint probability P(X, Y) where X= true label of early troph and Y corresponds to correctly predicted early troph. Similarly C(1,2) would reflect the joint probability P(X,Y) where X = true label of early_troph while Y = incorrectly predicted to be late_troph. Figure 3.9 shows the comparison of MI with and without feature selection. One of the advantages of displaying accuracy using mutual information is that the upper limit of the mutual information is exactly known. So, the accuracy of the model can be compared with the ideal case. In our case, since the number of labels is four, the maximum possible mutual information for an error-free case is 2 bits, however, maximum information acquired by the models is 1.28 bits here.

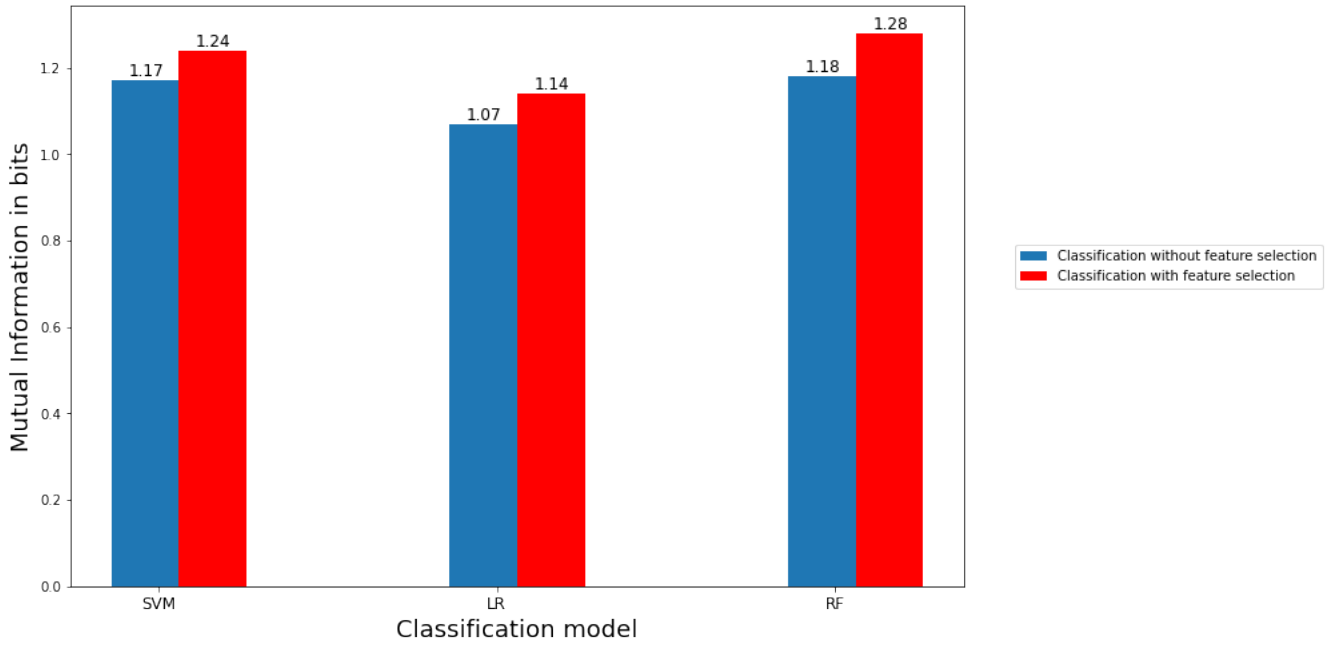


Figure 3.9: Mutual information with and without feature selection. The bar graphs display a comparison between the values of mutual information in bits between predicted and actual labels for the three models and for classification with feature selection and without feature selection as indicated.

3.5.1 Classification with randomly selected 378 features

In order to test whether the GA-based feature selection algorithm is able to select the features appropriately, we randomly chose 378 features from our dataset and evaluated the prediction accuracy using the SVM, LR, and RF models. We achieved an accuracy of 0.81, 0.79, and 0.80 for the models. Table 3.12 shows the F1 scores of the different classes for the three models.

F1 scores / Malaria Life Cycle Stage	SVM	LR	RF
early_troph	0.63	0.61	0.60
late_troph	0.86	0.85	0.86
ring	0.87	0.84	0.86
schizont	0.72	0.69	0.71

Table 3.12: F1 scores of different models of different classes with randomly selected 378 features.

The accuracy and the F1 scores of this experiment were lower when compared to the classification results with feature selection using the GA pipeline (see subsection 3.4.3). This results demonstrate the legitimacy and supremacy of the feature selection method.

3.6 Construction and Analysis of Protein-Protein Interaction Network

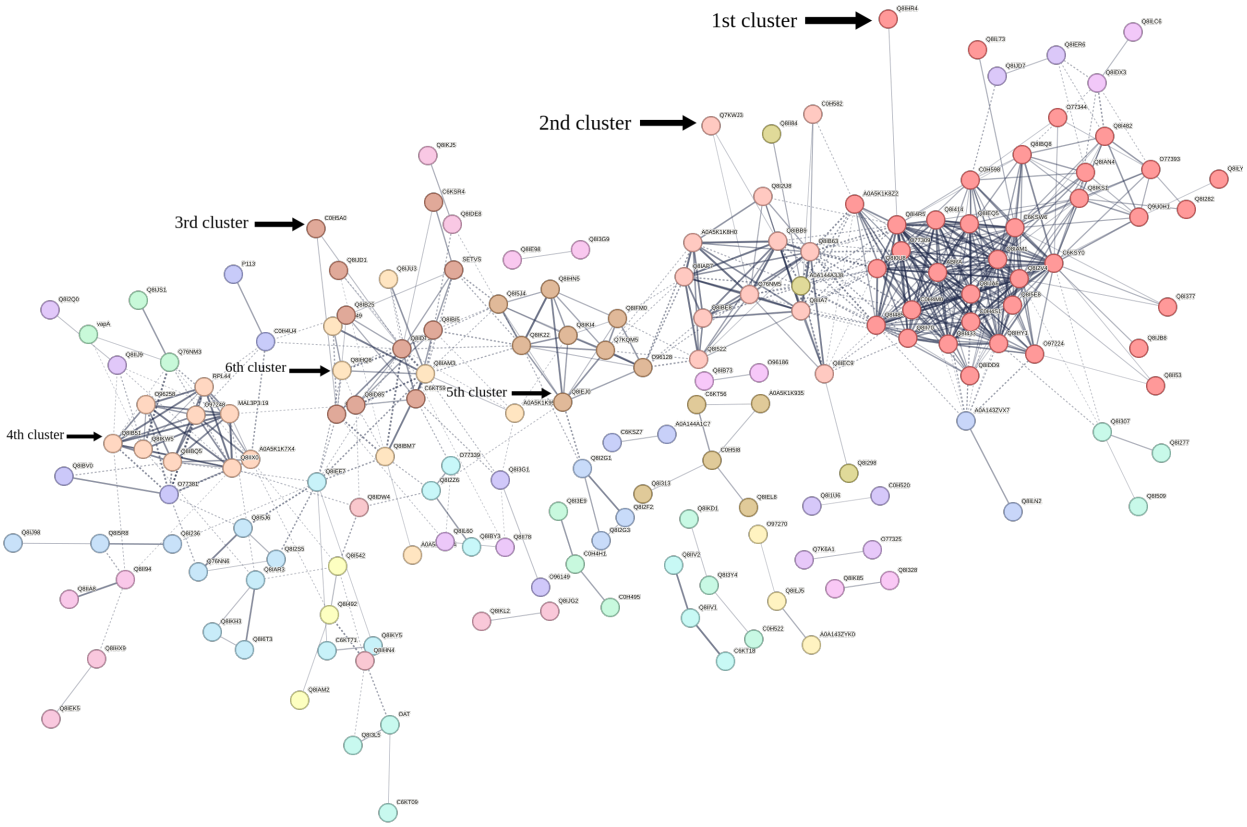


Figure 3.10: Protein-protein interaction network exhibits different clusters.
The graph shows the protein protein interaction network of the 378 proteins selected by the feature selection method. The different colors indicate different identified clusters.

Understanding protein-protein interactions (PPIs) is critical for cell physiology in normal and pathological states because they are required for practically every process in a cell [81]. Protein-protein interaction networks (PPIN) are graphs of the interactions between proteins in a cell. Protein-protein interaction happens in specified binding areas and serves a specific function. The feature selection method provided us with 378 proteins in Plasmodium falciparum. We used the Search Tool for the Retrieval of Interacting Genes/Proteins database (STRING 11.0b) [82] to construct the PPI network associated with these proteins. STRING can then construct a PPI network containing all of these proteins and their connections. Their interactions were generated with high confidence from high-throughput lab experiments and prior information in curated databases (sources: experiments, databases; $Scores \geq 0.90$). The network construction shows a set of highly connected modules (Figure 3.10).

3.6.1 Topological Analysis of the PPIN

Various topological measures are generally used to evaluate both the global and node characteristics in the PPINs, including degree (k), between centrality (BC), eccentricity, closeness centrality (CC), eigenvector centrality (EC), and clustering coefficient [83]. These are the definitions of these measures:

- Network size: The total number of nodes N is called the size of the network
- Degree: The number of links a node has (i.e., the number of its direct neighbors) is called its degree k.
- Network paths: A network path refers to a sequence of links that connect two nodes A and B
- Network diameter: The diameter dmax of a network is the longest of all shortest paths between any two nodes.
- Between Centrality: Between Centrality counts the number of shortest paths that pass through a node.
- Eigenvector Centrality: measure of the degree of the node as well as the degree of its neighbors.
- Closeness Centrality: measure of how close is a vertex to the other vertices [sum of the shortest path distances].
- eccentricity : The eccentricity of a node in a graph is defined as the length of a longest shortest path starting at that node.
- Clustering coefficient: clustering coefficient is a measure of the degree to which nodes in a graph tend to cluster together.

Here, highest degree nodes are identified using the degree distribution. Additionally, we have used Markov Clustering (MCL) Algorithm to find the clusters in the network. Among these clusters, we identified clusters which also contain the node with the highest degree and high BC.

This PPIN is composed of 378 nodes with the number of edges: 600, average node degree: 3.17, average local clustering coefficient: 0.309, expected number of edges: 621, PPI enrichment p-value: 0.00015. We can see that proteins in the red cluster (designated as 1st cluster) have the highest degree and high betweenness centrality. So, we can consider the red cluster as disease module. We analysed other topological properties like degree, BC, eccentricity, CC, EC, clustering coefficient, etc of this Red cluster using Gephi [84].

The proteins in Table 3.13 from red cluster have high degree and betweenness centrality (BC). In this cluster, the number of nodes: 36, number of edges: 252, average node degree: 14, average local clustering coefficient: 0.83, expected number of edges: 127, PPI enrichment p-value $< 10^{-6}$. We can see that this cluster has lesser nodes with high clustering coefficient. So, this is a small world network. We can see from the above table that C6KSW6 and C6KSY0 have the highest degree with high BC. We

considered these two proteins as the hubs or bottlenecks as these nodes have high degree (k) and BC. We have chosen 3 more proteins that have high degree and BC to consider as the backbone of the PPIN. These proteins are Q8I2V4, Q8IAM1, and Q8I4R5. These 5 proteins are highly connected in PPIN and have control over the network.

Proteins name	Degree	betweeness centrality	Description
C6KSW6	29	116.68	Leucine-rich repeat protein
C6KSY0	29	83.89	AP2 domain transcription factor, putative
Q8I2V4	25	31.35	Regulator of chromosome condensation-PP1-interacting protein
Q8IAM1	25	26.76	AP2 domain transcription factor, putative
Q8I4R5	23	41.62	Rhoptry neck protein 3

Table 3.13: Topological analysis of the PPI network of the selected proteins.

In order to delineate the role of the PPN clusters, gene ontology (GO) enrichment analysis were performed separately for different proteins belonging to the 6 clusters as designated in the Figure 3.10 and GO terms having enrichment p-values less than 0.05 are selected. The 1st cluster is found to be enriched for the Rhoptry protein family which is known to play crucial role in the virulence of the parasite inside the host [85]. The 2nd cluster proteins predominately belong to the apical complex family which mediate host penetration and invasion [86]. The 4th cluster is enriched with ribosomal protein plausibly to regulate translation during the IE life cycle stages [87, 88]. The fifth cluster is composed of proteins belonging to symbiont containing vacuole membrane which is likely central to nutrient acquisition, host cell remodeling, waste disposal, environmental sensing, and protection from innate defense etc [89]. One of the components of 6th cluster is found to be the proteins in the nucleolus which are important for regulation of ribosomal biogenesis [90]. Out of the 5 proteins with high degree and betweeness in the PPI network, Q8I4R5 (from the red cluster) showed p-value less than 0.05 in the GO enrichment analysis. We see that Q8I4R5 is the UniProt ID (RON3 - rhoptry protein) [91]. It could be a potential target for drug design as RON3 affects functional translocation of exported proteins and glucose uptake.

The function of a membrane protein complex called the Plasmodium translocon of exported proteins (PTEX), which exports specific parasite proteins across the parasitophorous vacuolar membrane (PVM) that encases the parasite in the host RBC cytoplasm, is essential for Plasmodium spp. survival within the host red blood cell (RBC). The core of PTEX has three proteins: EXP2, PTEX150, and the HSP101 ATPase. Only EXP2 is a membrane protein out of these three proteins. Studying the PTEX-dependent transport of members of the exportome, we found that when the parasite rhoptry protein RON3 was conditionally disrupted, exported proteins such as the ring infected erythrocyte surface antigen (RESA) were unable to move in parasites. Additionally, RON3-deficient parasites did not progress through the ring stage, and their intake of glucose was drastically reduced. The results show that RON3 affects two translocation processes, including the movement of the parasite exportome through PTEX and the

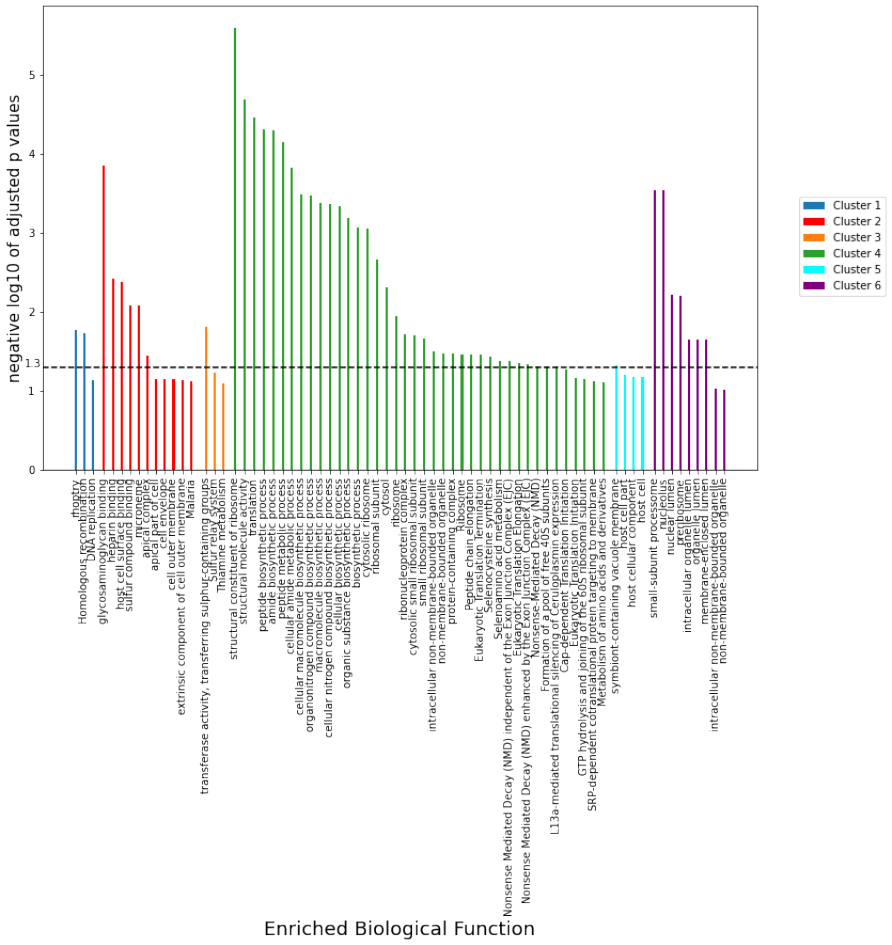


Figure 3.11: **Different enriched biological function for first six protein-protein interaction clusters**. The p-values of the enrichment of different gene ontologies for the six clusters of PPI network as indicated by the color code. The horizontal dashed line represents a threshold of 0.05.

movement of glucose from the RBC cytoplasm to the parasitophorous vacuolar (PV) space, where it can enter the parasite via the hexose transporter (HT) in the parasite plasma membrane [91]. (see Figure 3.11)

3.7 Expression Profile of the Selected Features

The analysis above provides us a set of proteins which are associated with the progression of the malaria pathogen through different stages of the life cycle. Thus, the expression pattern of these proteins would elicit the identity of the stages. In order to investigate the overall expression pattern of the genes across the different stages, we extracted the selected 378 features from the dataset. For each feature, we find the average RNA-seq read counts for all the four classes (early_troph, late_troph, schizont and ring). The average values are then transformed into log scale. We observed that genes fall into different

clusters according to the expression patterns (Figure 3.12) and also the expression patterns vary among the stages. For instance, the genes at the bottom have a very low expression in ring phase. Similarly, genes at the top cluster are displaying low expression for all stages. These expression patterns may be harnessed to look for specific markers for different stages. Additionally, we visualised the clustering behaviour of cells after feature selection by GA, using the 378 features via the Seurat package. As done previously,the normalized counts were subjected to linear and non-linear dimensionality reduction using PCA and UMAP respectively. Figure 3.13 shows clear clusters of all the four blood cycle stages - ring, early troph, late troph and schizont, which supports that the selected features can serve as markers for the respective stages.

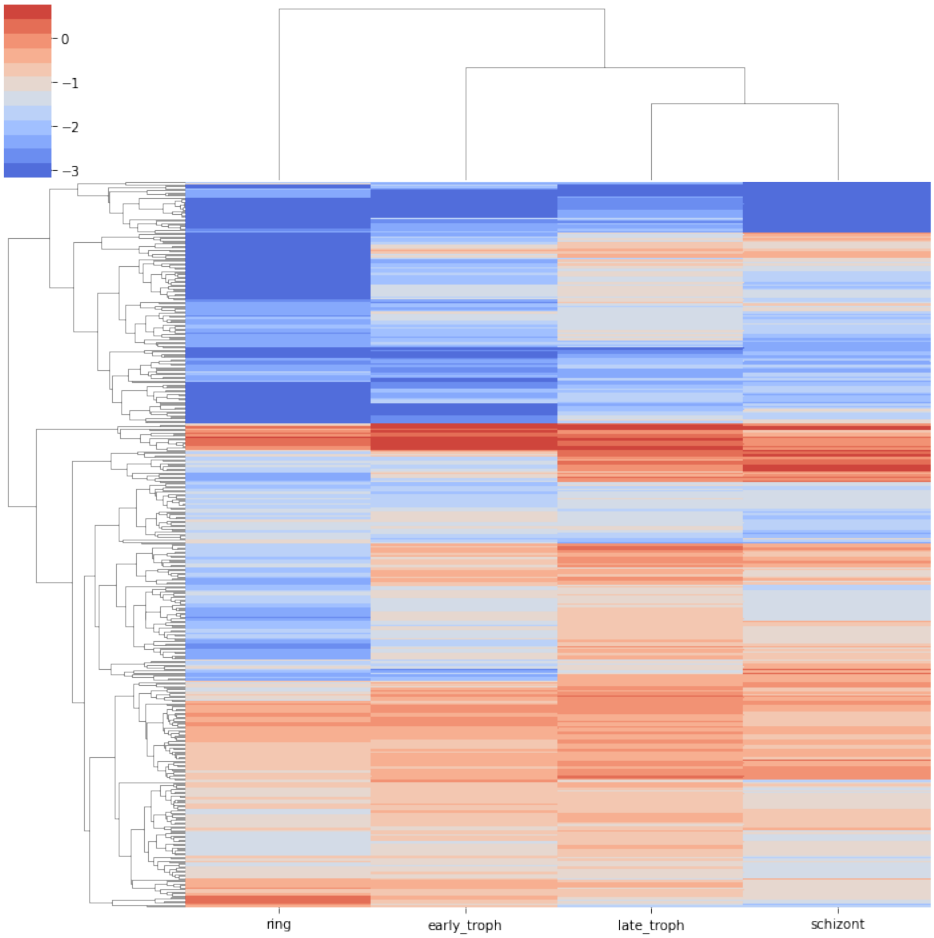


Figure 3.12: **The expression profiles are distinct among the stages**. Expression Profile of the selected genes across the different stages. The heatmap shows the average RNA-count of the selected 378 genes across the different stages as indicated. A hierarchical clustering is performed on the expression levels in order group genes with similar expression patterns indicated by the dendrogram.

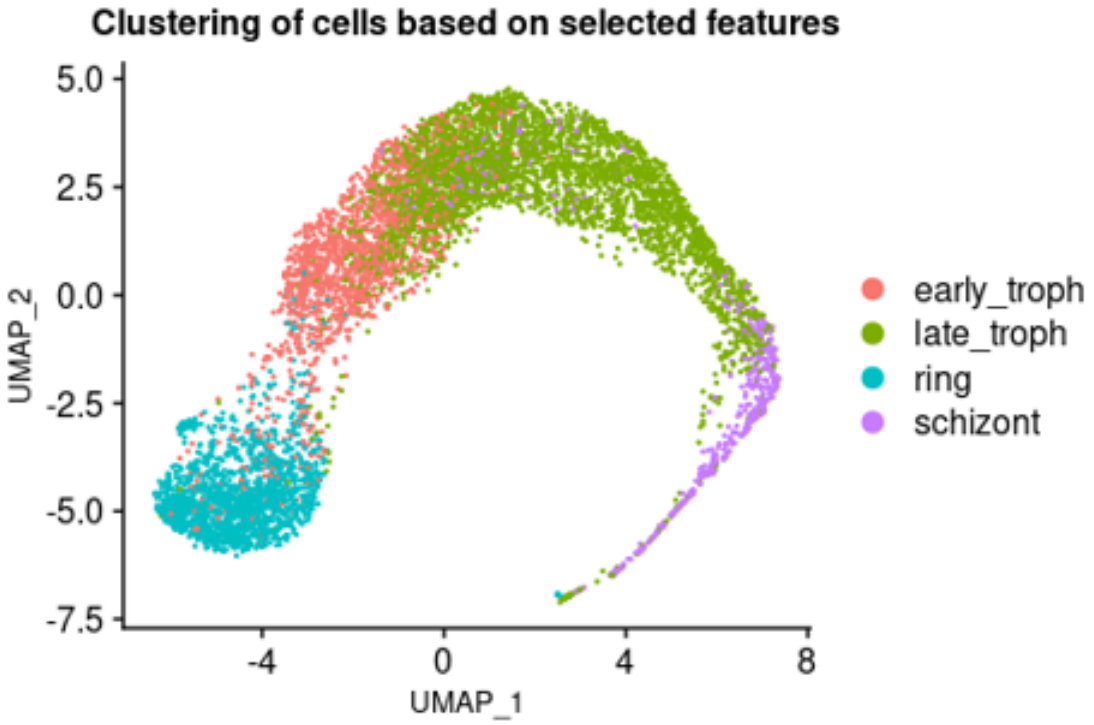


Figure 3.13: **Three dimensional visualization of the cells based on selected features**.UMAP of cells using 378 features. The cell clusters are colored based on the blood cycle stages of *P.falciparum.*

3.8 Tools Utilized

The data we used is freely accessible as a processed dataset through a user-friendly web interface (www.sanger.ac.uk/science/tools/mca/mca/) [33]. Our dataset has 5066 rows and 6737 columns. Each row corresponds to single cell and each column corresponds to a gene. We have 5066 features in our dataset and have four malaria life cycle stages (early_troph, late_troph, ring, and schizont).

Ada, the High Performance Computing Data Center of International Institute of Information Technology Hyderabad, India was utilized for the computation. It consists of 92 nodes, each equipped with dual Intel Xeon E5-2640 v4 processor, 128 GB RAM, two scratch disks (2 TB SATA and 960 GB SSD SATA) and four Nvidia GTX 1080 Ti / RTX 2080 Ti GPUs. The cluster has a total of 1472512 GPU cores, 3680 CPU cores and 11776 GB RAM. For our experiment we have used 40 cores with maximum memory per CPU as 2 GB on a Linux Ubuntu operating system. The proposed model is implemented using Python with the genetic_selection library for the GA implementation and the sklearn library for the classification algorithms. The relevant data and python scripts can be found in this github **code** link.

We used the R-based Seurat (v4.1.0) package developed by Satija lab [80] for visualisation and dimensionality reduction of single cell RNA-seq data. This was implemented in R (v4.1.3), run on RStudio environment (v1.3.1093). We followed the standard pre-processing workflow, normalisation, linear and

non-linear dimensionality reduction recommended by Seurat developers with default parameters, unless otherwise mentioned in the results section.

The feature selection method provided us with 378 proteins in Pf. We used the Search Tool for the Retrieval of Interacting Genes/Proteins database (STRING 11.0b) [82] to construct the PPIN associated with these proteins. STRING software `https://string-db.org/` can then construct a PPIN containing all of these proteins and their connections. Their interactions were generated with high confidence from high-throughput lab experiments and prior information in curated databases (sources: experiments, databases; $Scores \geq 0.90$). We can see that proteins in red cluster have highest degree and high BC. So, we can consider red cluster as disease module. We have also analysed other topological properties like degree, BC, eccentricity, CC, EC, clustering coefficient, etc of this Red cluster using Gephi [84] software.

In the next chapter we present the major results and discussion of our research along with the potential future research directions.

Chapter 4

Conclusions

In this research we employ supervised learning algorithm coupled with feature selection algorithms to extract the most relevant genes in predicting the life cycle stages of Plasmodium falciparum inside RBCs. The present study presents a two-stage model for feature selection and classification leading to improved classification of the different stages of the Malaria Life Cycle. This was achieved in the first stage of modeling by extracting the relevant features from the total data set (5066 features) considered for analysis. The popular GA based search optimization technique for feature selection was applied on single-cell transcriptomics for dealing with high dimensionality datasets and dimensionality reduction. Features are chosen based on their class variants for higher efficiency and accuracy, transforming the selected elements into a lower dimension. The study's main finding is that using a feature selection procedure before applying a classification algorithm results in more accurate predictions. The use of GA as a feature selection process significantly reduced the number of features included in the dataset.

Next, we constructed PPINs between our proteins obtained after employing feature selection algorithm and conducted network analysis on this PPIN using the STRING 11.0b and Gephi software. A set of topological analysis was performed using various topological measures (including degree, between centrality, eccentricity, closeness centrality, eigenvector centrality, and clustering coefficient to estimate and evaluate the node characteristics in the PPINs [83]. We found degree and betweeness centrality of each protein though this calculation to provide hierarchies according to importance of the genes in the network. Proteins having high degree and betweeness centrality tend to assert more control over the network function and can thus be considered as drug targets for future studies.

In the second stage the reduced subset of 378 features is further utilized for high accuracy multi-class classification. For the four-class classification of the life cycle of malaria parasite based on oriented gradients and local binary pattern features, a three-pronged approach employing SVM, LR and RF techniques is used. On using the reduced 378 features, RF performed best with a classification accuracy of 92% while SVM had a 91% accuracy and LR gave 88% accuracy. Even for the reduced features dataset we achieved similar performance for all the four classes, across all the three models. Further, randomly chosen features from our dataset of 378 were also evaluated using the SVM, LR, and RF models. We achieved an accuracy of 81%, 79%, and 80% for the three respective models, hence, proving

the robustness of the features selected using the GA approach. The proposed research methodology can be likely used for improved malaria diagnosis and drug targets. For further research, the hybrid methods for feature selection, the impact of parameter fine tuning on various algorithms' levels and the use of other methods including Ensemble Learning may be attempted.

Related Publications

1. Swarnim Shukla, Soham Choudhuri, Gayathri Priya Iragavarapu, Bhaswar Ghosh, Machine learning approaches for classification of Plasmodium falciparum life cycle stages using single-cell transcriptomes, bioRxiv doi: https://doi.org/10.1101/2022.06.22.497155; July 24, 2022. [92]

2. Swarnim Shukla, Soham Choudhuri, Gayathri Priya Iragavarapu, and Bhaswar Ghosh. Supervised learning of Plasmodium falciparum life cycle stages using single-cell transcriptomes identifies crucial proteins. Journal of Bioinformatics and Systems Biology. 6 (2023): 31-46[93]

Bibliography

[1] Barnali Sahu, Satchidananda Dehuri, and Alok Jagadev. A study on the relevance of feature selection methods in microarray data. *The Open Bioinformatics Journal*, 11(1), 2018.

[2] Zena M Hira and Duncan F Gillies. A review of feature selection and feature extraction methods applied on microarray data. *Advances in bioinformatics*, 2015, 2015.

[3] Genetic algorithms - parent selection[https://www.tutorialspoint.com/genetic_algorithms/genetic_algorithms_parent_selection.htm], Jul 2022.

[4] Support vector machine (svm) algorithm - javatpoint[https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm, Mar 2022.

[5] Jun Han and Claudio Moraga. The influence of the sigmoid function parameters on the speed of backpropagation learning. In *International workshop on artificial neural networks*, pages 195–201. Springer, 1995.

[6] Random forest (rf) algorithm - javatpoint[https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_classification_algorithms_random_forest.htm, Mar 2022.

[7] Salam Shuleenda Devi, Amarjit Roy, Joyeeta Singha, Shah Alam Sheikh, and Rabul Hussain Laskar. Malaria infected erythrocyte classification based on a hybrid classifier using microscopic images of thin blood smear. *Multimedia Tools and Applications*, 77(1):631–660, 2018.

[8] Naveed Abbas, Tanzila Saba, Amjad Rehman, Zahid Mehmood, Hoshang Kolivand, Mueen Uddin, and Adeel Anjum. Plasmodium life cycle stage classification based quantification of malaria parasitaemia in thin blood smears. *Microscopy research and technique*, 82(3):283–295, 2019.

[9] Daniel J Weiss, Tim CD Lucas, Michele Nguyen, Anita K Nandi, Donal Bisanzio, Katherine E Battle, Ewan Cameron, Katherine A Twohig, Daniel A Pfeffer, Jennifer A Rozier, et al. Mapping the global prevalence, incidence, and mortality of plasmodium falciparum, 2000–17: a spatial and temporal modelling study. *The Lancet*, 394(10195):322–331, 2019.

[10] Cdc - about malaria disease [https://www.cdc.gov/malaria/about/disease.html], Mar 2022.

[11] Richard Idro, Kevin Marsh, Chandy C John, and Charles R J Newton. Cerebral malaria: Mechanisms of brain injury and strategies for improved neurocognitive outcome, Oct 2010.

[12] Giulia Siciliano and Pietro Alano. Enlightening the malaria parasite life cycle: bioluminescent plasmodium in fundamental and applied research. *Frontiers in microbiology*, 6:391, 2015.

[13] EY Klein. Antimalarial drug resistance: a review of the biology and strategies to delay emergence and spread. *International journal of antimicrobial agents*, 41(4):311–317, 2013.

[14] Hirdesh Kumar and Niraj H Tolia. Getting in: The structural biology of malaria invasion. *PLoS Pathogens*, 15(9):e1007943, 2019.

[15] Lia Chappell, Philipp Ross, Lindsey Orchard, Timothy J Russell, Thomas D Otto, Matthew Berriman, Julian C Rayner, and Manuel Llinás. Refining the transcriptome of the human malaria parasite plasmodium falciparum using amplification-free rna-seq. *BMC genomics*, 21(1):1–19, 2020.

[16] Mukul Rawat, Ashish Srivastava, Shreya Johri, Ishaan Gupta, and Krishanpal Karmodiya. Single-cell rna sequencing reveals cellular heterogeneity and stage transition under temperature stress in synchronized plasmodium falciparum cells. *Microbiology spectrum*, 9(1):e00008–21, 2021.

[17] Niall D Geoghegan, Cindy Evelyn, Lachlan W Whitehead, Michal Pasternak, Phoebe McDonald, Tony Triglia, Danushka S Marapana, Daryan Kempe, Jennifer K Thompson, Michael J Mlodzianoski, et al. 4d analysis of malaria parasite invasion offers insights into erythrocyte membrane remodeling and parasitophorous vacuole formation. *Nature communications*, 12(1):1–16, 2021.

[18] Plasmodium https://www.microscopemaster.com/plasmodium.html, Jul 2022.

[19] Plasmodium ovale [https://www.cdc.gov/dpdx/resources/pdf/benchaids/malaria/povale_benchaidv2.pdf], july 23, 2022.

[20] Laboratory diagnosis of Plasmodium falciparum - MOAM.INFO — moam.info. `https://moam.info/laboratory-diagnosis-of-plasmodium-falciparum_59caae2a1723dd056c5a1be4.html`. [Accessed 27-Feb-2023].

[21] Malaria seminar — slideshare.net. `https://www.slideshare.net/sayanmanta/malaria-seminar`. [Accessed 27-Feb-2023].

[22] Richard Lucius, Brigitte Loos-Frank, Richard P Lane, Robert Poulin, Craig Roberts, and Richard K Grencis. *The biology of parasites*. John Wiley & Sons, 2017.

[23] Michael Lanzer, Hannes Wickert, Georg Krohne, Laetitia Vincensini, and Catherine Braun Breton. Maurer's clefts: a novel multi-functional organelle in the cytoplasm of plasmodium falciparum-infected erythrocytes. *International journal for parasitology*, 36(1):23–36, 2006.

[24] Vinayak K Bairagi and Kshipra C Charpe. Comparison of texture features used for classification of life stages of malaria parasite. *International Journal of Biomedical Imaging*, 2016, 2016.

[25] Janet Storm, Jakob S Jespersen, Karl B Seydel, Tadge Szestak, Maurice Mbewe, Ngawina V Chisala, Patricia Phula, Christian W Wang, Terrie E Taylor, Christopher A Moxon, et al. Cerebral malaria is associated with differential cytoadherence to brain endothelial cells. *EMBO Molecular Medicine*, 11(2):e9164, 2019.

[26] Malcolm J Gardner, Neil Hall, Eula Fung, Owen White, Matthew Berriman, Richard W Hyman, Jane M Carlton, Arnab Pain, Karen E Nelson, Sharen Bowman, et al. Genome sequence of the human malaria parasite plasmodium falciparum. *Nature*, 419(6906):498–511, 2002.

[27] François Nosten and Nicholas J White. Artemisinin-based combination treatment of falciparum malaria. *Defining and Defeating the Intolerable Burden of Malaria III: Progress and Perspectives: Supplement to Volume 77 (6) of American Journal of Tropical Medicine and Hygiene*, 2007.

[28] Nicholas M Douglas, Nicholas M Anstey, Brian J Angus, Francois Nosten, and Ric N Price. Artemisinin combination therapy for vivax malaria. *The Lancet infectious diseases*, 10(6):405–416, 2010.

[29] Robert J Commons, Julie A Simpson, Kamala Thriemer, Georgina S Humphreys, Tesfay Abreha, Sisay G Alemu, Arletta Añez, Nicholas M Anstey, Ghulam R Awab, J Kevin Baird, et al. The effect of chloroquine dose and primaquine on plasmodium vivax recurrence: a worldwide antimalarial resistance network systematic review and individual patient pooled meta-analysis. *The Lancet Infectious Diseases*, 18(9):1025–1034, 2018.

[30] Muhammad Minoar Hossain, Md Abdur Rahim, Ali Newaz Bahar, and Mohammad Motiur Rahman. Automatic malaria disease detection from blood cell images using the variational quantum circuit. *Informatics in Medicine Unlocked*, 26:100743, 2021.

[31] Eliana Real, Virginia M Howick, Farah A Dahalan, Kathrin Witmer, Juliana Cudini, Clare Andradi-Brown, Joshua Blight, Mira S Davidson, Sunil Kumar Dogga, Adam J Reid, et al. A single-cell atlas of plasmodium falciparum transmission through the mosquito. *Nature communications*, 12(1):1–13, 2021.

[32] Mahdieh Poostchi, Kamolrat Silamut, Richard J Maude, Stefan Jaeger, and George Thoma. Image analysis and machine learning for detecting malaria. *Translational Research*, 194:36–55, 2018.

[33] Virginia M Howick, Andrew JC Russell, Tallulah Andrews, Haynes Heaton, Adam J Reid, Kedar Natarajan, Hellen Butungi, Tom Metcalf, Lisa H Verzier, Julian C Rayner, et al. The malaria cell atlas: Single parasite transcriptomes across the complete plasmodium life cycle. *Science*, 365(6455):eaaw2619, 2019.

[34] Gloria Díaz, Fabio A. González, and Eduardo Romero. A semi-automatic method for quantification and classification of erythrocytes infected with malaria parasites in microscopic images. *Journal of Biomedical Informatics*, 42(2):296–307, 2009.

[35] Vít Špringl et al. Automatic malaria diagnosis through microscopy imaging. *Higher Diploma, Faculty of Electrical Engineering*, 2009.

[36] Mohammad Imroze Khan, Bhibhudendra Acharya, Bikesh Kumar Singh, and Jigyasa Soni. Content based image retrieval approaches for detection of malarial parasite in blood images. *International Journal of Biometrics and Bioinformatics (IJBB)*, 5(2):97, 2011.

[37] Jigyasha Soni, Nipun K. Mishra, and Chandrashekhar Kamargaonkar. Automatic differentiation between rbc and malarial parasites based on morphology with first order features using image processing. 2011.

[38] Almas Jabeen, Nadeem Ahmad, and Khalid Raza. Machine learning-based state-of-the-art methods for the classification of rna-seq data. In *Classification in BioApps*, pages 133–172. Springer, 2018.

[39] G. Karthik, S. Muttan, M. P. Saravanan, R. Seetharaman, and V. Vignesh. Automated malaria diagnosis using microscopic images. In *2019 Third International Conference on Inventive Systems and Control (ICISC)*, pages 514–517, 2019.

[40] Jaspreet Singh Chima, Abhishek Shah, Karan Shah, and Rekha Ramesh. Malaria cell image classification using deep learning. *International Journal of Recent Technology and Engineering*, 8(6):5553–59, 2020.

[41] Qazi Ammar Arshad, Mohsen Ali, Saeed-ul Hassan, Chen Chen, Ayisha Imran, Ghulam Rasul, and Waqas Sultani. A dataset and benchmark for malaria life-cycle classification in thin blood smear images. *Neural Computing and Applications*, 34(6):4473–4485, 2022.

[42] Sen Li, Zeyu Du, Xiangjie Meng, and Yang Zhang. Multi-stage malaria parasite recognition by deep learning. *GigaScience*, 10(6):giab040, 2021.

[43] Doni Setyawan, Retantyo Wardoyo, Moh Edi Wibowo, E Elsa Herdiana Murhandarwati, and Joharotul Jamilah. Malaria classification using convolutional neural network: A review. In *2021 Sixth International Conference on Informatics and Computing (ICIC)*, pages 1–9. IEEE, 2021.

[44] S Karthik and M Sudha. A survey on machine learning approaches in gene expression classification in modelling computational diagnostic system for complex diseases. *International Journal of Engineering and Advanced Technology*, 8(2):182–191, 2018.

[45] Micheal Olaolu AROWOLO, Marion Olubunmi ADEBIYI, Chiebuka Timothy NNODIM, Sulaiman Olaniyi ABDULSALAM, and Ayodele Ariyo ADEBIYI. An adaptive genetic algorithm

with recursive feature elimination approach for predicting malaria vector gene expression data classification using support vector machine kernels. *Walailak Journal of Science and Technology (WJST)*, 18(17):9849–11, 2021.

[46] Qingwen Li, Benzhi Dong, Donghua Wang, and Sui Wang. Identification of secreted proteins from malaria protozoa with few features. *IEEE Access*, 8:89793–89801, 2020.

[47] Sushil Kumar Mishra. Human malaria detection and stage classification using random forest classifier.

[48] Murad Al-Rajab, Joan Lu, and Qiang Xu. A framework model using multifilter feature selection to enhance colon cancer classification. *Plos one*, 16(4):e0249094, 2021.

[49] Qinglin Mei, Huaxiang Zhang, and Cheng Liang. A discriminative feature extraction approach for tumor classification using gene expression data. *Current Bioinformatics*, 11(5):561–570, 2016.

[50] Micheal O Arowolo, Marion Olubunmi Adebiyi, Ayodele Ariyo Adebiyi, and Olatunji Julius Okesola. A hybrid heuristic dimensionality reduction methods for classifying malaria vector gene expression data. *IEEE Access*, 8:182422–182430, 2020.

[51] Jie Li, Zhun Zhao, Li Zhou, and Yadong Wang. Y-spcr: A new dimensionality reduction method for gene expression data classification. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 401–408. IEEE, 2019.

[52] Lior Rokach. Genetic algorithm-based feature set partitioning for classification problems. *Pattern Recognition*, 41(5):1676–1700, 2008.

[53] Zili Zhang and Pengyi Yang. An ensemble of classifiers with genetic algorithmbased feature selection. *The IEEE intelligent informatics bulletin*, 9(1):18–24, 2008.

[54] Cheng-Lung Huang and Chieh-Jen Wang. A ga-based feature selection and parameters optimizationfor support vector machines. *Expert Systems with applications*, 31(2):231–240, 2006.

[55] Li-Yeh Chuang, Chao-Hsuan Ke, Hsueh-Wei Chang, and Cheng-Hong Yang. A two-stage feature selection method for gene expression data. *OMICS A journal of Integrative Biology*, 13(2):127–137, 2009.

[56] Shutao Li, Xixian Wu, and Mingkui Tan. Gene selection using hybrid particle swarm optimization and genetic algorithm. *Soft Computing*, 12(11):1039–1048, 2008.

[57] Mohd Saberi Mohamad, Safaai Deris, and Rosli Md Illias. A hybrid of genetic algorithm and support vector machine for features selection and classification of gene expression microarray. *International Journal of Computational Intelligence and Applications*, 5(01):91–107, 2005.

[58] Manaswini Pradhan. Evolutionary computational algorithm by blending of ppca and ep-enhanced supervised classifier for microarray gene expression data. *IAES International Journal of Artificial Intelligence*, 7(2):95, 2018.

[59] Adam J Reid, Arthur M Talman, Hayley M Bennett, Ana R Gomes, Mandy J Sanders, Christopher JR Illingworth, Oliver Billker, Matthew Berriman, and Mara KN Lawniczak. Single-cell rna-seq reveals hidden transcriptional variation in malaria parasites. *elife*, 7:e33105, 2018.

[60] Juliana M Sa, Matthew V Cannon, Ramoncito L Caleon, Thomas E Wellems, and David Serre. Single-cell transcription analysis of plasmodium vivax blood-stage parasites identifies stage-and species-specific profiles of expression. *PLoS biology*, 18(5):e3000711, 2020.

[61] Katelyn A Walzer, Hélène Fradin, Liane Y Emerson, David L Corcoran, and Jen-Tsan Chi. Latent transcriptional variations of individual plasmodium falciparum uncovered by single-cell rna-seq and fluorescence imaging. *PLoS genetics*, 15(12):e1008506, 2019.

[62] Asaf Poran, Christopher Nötzel, Omar Aly, Nuria Mencia-Trinchant, Chantal T Harris, Monica L Guzman, Duane C Hassane, Olivier Elemento, and Björn FC Kafsack. Single-cell rna sequencing reveals a signature of sexual commitment in malaria parasites. *Nature*, 551(7678):95–99, 2017.

[63] Divya Jain and Vijendra Singh. An efficient hybrid feature selection model for dimensionality reduction. *Procedia Computer Science*, 132:333–341, 2018.

[64] Aleksandra A Kolodziejczyk, Jong Kyoung Kim, Valentine Svensson, John C Marioni, and Sarah A Teichmann. The technology and biology of single-cell rna sequencing. *Molecular cell*, 58(4):610–620, 2015.

[65] Mtakai Ngara, Mia Palmkvist, Sven Sagasser, Daisy Hjelmqvist, Åsa K Björklund, Mats Wahlgren, Johan Ankarklev, and Rickard Sandberg. Exploring parasite heterogeneity using single-cell rna-seq reveals a gene signature among sexual stage plasmodium falciparum parasites. *Experimental cell research*, 371(1):130–138, 2018.

[66] Andrea Loddo, Corrado Fadda, and Cecilia Di Ruberto. An empirical evaluation of convolutional networks for malaria diagnosis. *Journal of Imaging*, 8(3):66, 2022.

[67] C Gunavathi, K Premalatha, and K Sivasubramanian. A survey on feature selection methods in microarray gene expression data for cancer classification. *Res. J. Pharm. Technol*, 10:1395–1401, 2017.

[68] D Asir Antony Gnana Singh, E Jebamalar Leavline, R Priyanka, and P Padma Priya. Dimensionality reduction using genetic algorithm for improving accuracy in medical diagnosis. *International Journal of Intelligent Systems and Applications*, 8(1):67, 2016.

[69] Matthew Shardlow. An analysis of feature selection techniques. *The University of Manchester*, 1(2016):1–7, 2016.

[70] Madeline McCombe. Intro to feature selection methods for data science [https://towardsdatascience.com/intro-to-feature-selection-methods-for-data-science-4cae2178a00a], Jun 2019.

[71] Adam Slowik and Halina Kwasnicka. Evolutionary algorithms and their applications to engineering problems. *Neural Computing and Applications*, 32(16):12363–12379, 2020.

[72] Marion O Adebiyi, Micheal O Arowolo, and Oludayo Olugbara. A genetic algorithm for prediction of rna-seq malaria vector gene expression data classification using svm kernels. *Bulletin of Electrical Engineering and Informatics*, 10(2):1071–1079, 2021.

[73] D.E. Goldberg. *Genetic Algorithms*. Pearson Education, 2013.

[74] Melanie Mitchell. *An introduction to genetic algorithms*. MIT press, 1998.

[75] Alexander Statnikov, Lily Wang, and Constantin F Aliferis. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC bioinformatics*, 9(1):1–10, 2008.

[76] Mohammadmehdi Saberioon, Petr Císař, Laurent Labbé, Pavel Souček, Pablo Pelissier, and Thierry Kerneis. Comparative performance analysis of support vector machine, random forest, logistic regression and k-nearest neighbours in rainbow trout (oncorhynchus mykiss) classification using image-based features. *Sensors*, 18(4):1027, 2018.

[77] Logistic regression [https://datax.berkeley.edu/wp-content/uploads/2020/09/slides-m140-logistic-reg-sklearn.pdf], Jul 2022.

[78] Aayush Bajaj. Performance metrics in machine learning [complete guide], Jul 2022.

[79] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[80] Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36:411–420, 2018.

[81] Roded Silverbush, Dana Sharan. A systematic approach to orient the human protein–protein interaction network. *Nature Communications*, 2019.

[82] Nastou KC Lyon D Kirsch R Pyysalo S Doncheva NT Legeay M Fang T Bork P Jensen LJ von Mering C. Szklarczyk D, Gable AL. The string database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.*, 49(605-612):91–107, 2021.

[83] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

[84] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: An open source software for exploring and manipulating networks, 2009.

[85] Natalie A Counihan, Ming Kalanon, Ross L Coppel, and Tania F de Koning-Ward. Plasmodium rhoptry proteins: why order is important. *Trends in parasitology*, 29(5):228–236, 2013.

[86] Noriko Okamoto and Patrick J Keeling. The 3d structure of the apical complex and association with the flagellar apparatus revealed by serial tem tomography in psammosa pacifica, a distant relative of the apicomplexa. *PloS one*, 9(1):e84653, 2014.

[87] Heather J Painter, Neo Christopher Chung, Aswathy Sebastian, Istvan Albert, John D Storey, and Manuel Llinás. Genome-wide real-time in vivo transcriptional dynamics during plasmodium falciparum blood-stage development. *Nature communications*, 9(1):1–12, 2018.

[88] Jessey Erath, Sergej Djuranovic, and Slavica Pavlovic Djuranovic. Adaptation of translational machinery in malaria parasites to accommodate translation of poly-adenosine stretches throughout its life cycle. *Frontiers in Microbiology*, 10:2823, 2019.

[89] Tobias Spielmann, Georgina N Montagna, Leonie Hecht, and Kai Matuschewski. Molecular make-up of the plasmodium parasitophorous vacuolar membrane. *International Journal of Medical Microbiology*, 302(4-5):179–186, 2012.

[90] Sylvie Briquet, Asma Ourimi, Cédric Pionneau, Juliana Bernardes, Alessandra Carbone, Solenne Chardonnet, and Catherine Vaquero. Identification of plasmodium falciparum nuclear proteins by mass spectrometry and proposed protein annotation. *PLoS One*, 13(10):e0205596, 2018.

[91] Leanne M Low, Yvonne Azasi, Emma S Sherling, Matthias Garten, Joshua Zimmerberg, Takafumi Tsuboi, Joseph Brzostowski, Jianbing Mu, Michael J Blackman, and Louis H Miller. Deletion of plasmodium falciparum protein ron3 affects the functional translocation of exported proteins and glucose uptake. *MBio*, 10(4):e01460–19, 2019.

[92] Swarnim Shukla, Soham Choudhuri, Gayathri Priya Iragavarapu, and Bhaswar Ghosh. Machine learning approaches for classification of plasmodium falciparum life cycle stages using single-cell transcriptomes. *bioRxiv*, 2022.

[93] Swarnim Shukla, Soham Choudhuri, Gayathri Priya Iragavarapu, and Bhaswar Ghosh. Supervised learning of plasmodium falciparum life cycle stages using single-cell transcriptomes identifies crucial proteins, 2023.