Machine learning approaches for classification of Plasmodium falciparum life cycle stages using single-cell transcriptomes

Swarnim Shukla, Soham Choudhuri, Gayathri Priya Iragavarapu, and Bhaswar Ghosh*

International Institute of Information Technology, Gachibowli, Hyderabad, 500 032 Telangana, INDIA

E-mail: bhaswar.ghosh@iiit.ac.in

Phone: 040-6653 1000 Ext: 1512

Abstract

Malaria, spread by the female *Anopheles* mosquito, is a highly fatal disease widespread in many parts of the world, causing 0.4 million deaths globally. Vital gene expressions form the basis in the detection of malaria infection levels. Quantification of malaria parasite infected RBCs and classification of its life cycle stages are done at macroscopic level by experts, for making informed decisions. Off late multiple computational approaches have been proposed to circumvent the problem of dimensionality leading to accurate predicted results. In this work a dimensionality reduction technique based on Genetic Algorithm (GA) is applied on *P. falciparum* single-cell transcriptomics to arrive at an optimized subset of features from the larger dataset. Features are chosen based on their class variants considering increased efficiency and accuracy, to separately transform the selected elements into a lower dimension. For the classification

of the life cycle of malaria parasite based on single cell transcriptome data, a three-pronged approach employing the multiclass Support Vector Machine (SVM), Logistic Regression (LR) and Random Forest (RF) techniques is used. Distribution of cells was visualised and mapped using the R-based Seurat package. Further, we constructed protein interaction networks of the genes identified by the feature selection method and elucidated the role of the proteins in progression of the parasite through it's life cycle. Our approach presents a novel protocol to implement ML techniques on scRNA seq datasets and subsequently harnessing the extracted information for biomarker/drug target detection.

Introduction

Malaria is a deadly disease caused by the *Plasmodium* parasite and is transmitted through the bite of a female Anopheles mosquito. This *Plasmodium* attacks the red blood cells (RBCs) and the degree of malaria can be estimated by the quantity of infected RBCs.[1] The complex life cycle of malaria parasites features diverse developmental strategies, each of which is uniquely adapted to thrive in the particular host environment. Six *Plasmodium* species cause human malaria, with majority of the estimated 0.4 million annual deaths caused by *Plasmodium falciparum* (*P. falciparum*). Blood stage development begins when a newly released, extracellular parasite (a merozoite) invades an erythrocyte, establishing the ring stage of infection, progressing to the trophozoite stage. During this stage the infected erythrocyte is extensively modified to enable parasite proliferation. Thereafter the parasite divides to form a connected group of daughter cells, called schizont, which eventually lyses the host erythrocyte, releasing the newly formed merozoites to invade new erythrocytes. These steps are collectively known as the intraerythrocytic developmental cycle (IDC).[2] Malaria symptoms include high fever and headache and in some severe cases even seizures and death are caused. Malaria mostly affects the economically weak communities of the world, where medical treatment is not readily available. For a quick and successful recovery of any patient,

it is vital to diagnose and treat the malarial infection early. If the malaria life cycle stages are somehow ascertained then the treatment of disease becomes easier. Experienced medical professionals frequently examine a large number of blood films to detect malaria infection. Microscopists normally visualize the thick and thin blood smears to identify a disease or its cause. However, the accuracy depends upon smear quality and the expertise in classifying and counting the parasite and non-parasite cells. It is fairly challenging to number the parasites and infected RBCs manually and needs an expert microscopist for quality diagnosis.[3] Recent advances in single cell RNA-sequencing (sc-RNA)techniques paved new ways to characterize gene expression changes during the development stages of the plasmodium life cycle .[4,5] Analysis of the gene expression regulation may allow us to identity new diagnostics marker as well as potential target for a new drug. Indeed, many studies have already been conducted in the last few years by utilizing sc-RNA experiments for *Plasmodium falciparum*. One of the central advantages of employing sc-RNA methods is the scope of exploring cell to cell heterogeneity in the population by uncovering hidden variability in gene expression among single cells [5–7] . 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 .[8] Similarly, heterogeneity plays crucial role in *Plasmodium* stress response inside the RBC .[9] However, the diagnostics of scRNA-Seq is challenging as its outcome suffers from lack of fit due to high dimensional gene expression data. Advanced computational skills are needed to study and process the massive volume of proteomic and genomic data obtainable freely from several repositories [10,11] and harness them to reveal new biological insights.

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. There are two methods for reducing the dimensionality of data: Feature Extraction and Feature Selection. Feature selection is further

divided into filter, wrapper, and embedded methods.[12] In the filter method, mathematical measures are used to select the optimal features. Wrapper is a feedback method that uses a machine-learning algorithm to help choose the best features. The embedded approach is a hybrid of the filter and wrapper methods. This paper proposes a wrapper-based feature selection technique using the Genetic Algorithm to select the optimal features and remove the redundant noise in the dataset. In order to evaluate the performance of these features, SVM, LR, and RF classification models are used.

The main objective of our study is to select top-ranked genes from the scRNA-seq profiles at different stages of the plasmodium falciparum life cycle inside infected RBC. We employ supervised learning algorithm coupled with feature selection algorithms to extract most relevant genes in predicting the life cycle stages of plasmodium inside RBC. The first stage of the proposed model is to optimize the quality data from the data set by removing the redundant, noisy, and irrelevant genes (features). From the literature review(see Discussion and outlook) it can be concluded that genetic algorithm (GA) showed that better performance than the other selection algorithms and thus, prominently used for feature selection from high dimensional data sets.

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 the investigation is carried out rigorously. The combination of 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 which 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 importance of the genes in the network. These genes can be used for diagnosis and drug targets. Our study presents

a theoretical framework to select diagnosis markers and drug targets by implementing ML techniques on sc-RNA-seq data.

Results

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/).[13] For the current study, we considered the 10X scRNA-seq data of the intraerythrocytic stages of P.falciparum 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 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 (ring, early trophozoite, late trophozoite, and schizont). Thus, we set out to utilize classification ML algorithms (Material and methods) which would allow us to predict the life cycle stage of a cell based on the gene expression pattern.

Data visualisation and dimensionality reduction of the scRNA-seq data

We used Seurat,[14] 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 V.M. et al.[13] , 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 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 standard dimensionional 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 dimensional 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 1 represents the UMAP projection and we notice four clusters for ring, early trophozoite, late trophozoite and schizont respectively.

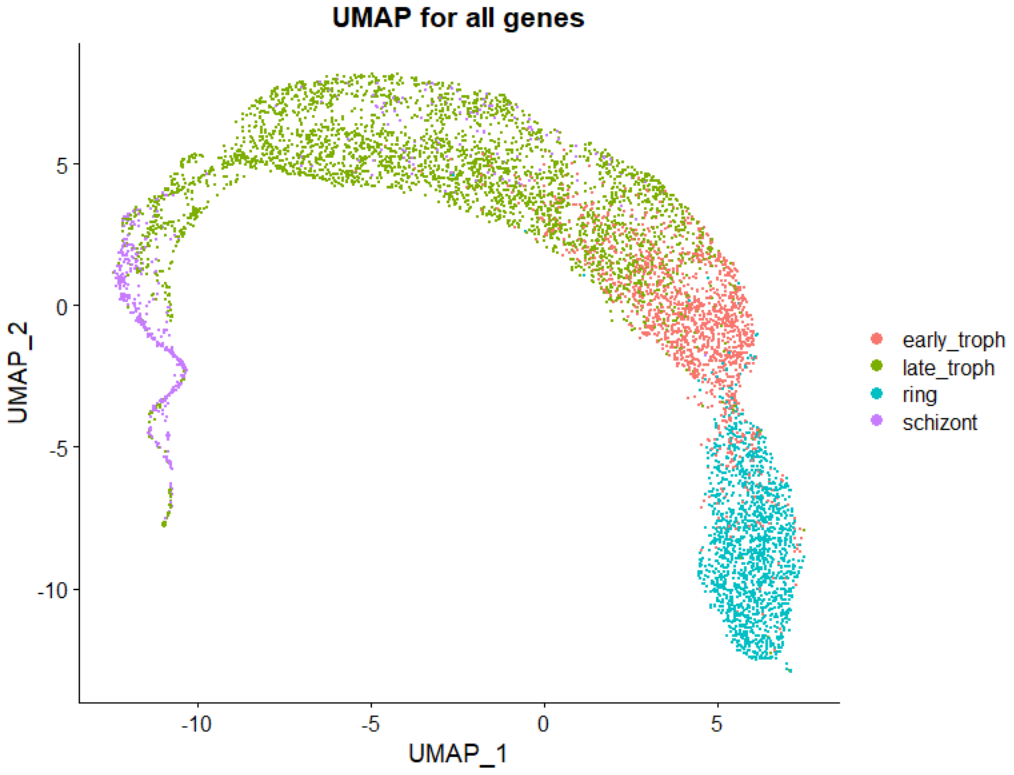

Figure 1: **Two dimensional visualization of the scRNA data set shows distinct cluster of life cycle stages**. UMAP of scRNA-seq counts of all 5066 features. Non-linear dimensional reduction of expression values of all genes in the dataset.The cell clusters are colored based on the blood cycle stages of *P.falciparum.*

Classification without feature selection

Next, we implement SVM, LR and RF algorithms to classify the cells into the four different stages by including all the genes. In order to calculate the prediction accuracy, we randomly selected 80 % of the cells as training set and 20 % of the cells are chosen for calculating the desired accuracy. Figure 2 shows the accuracy of SVM, LR, and RF. This is the baseline for our experiment. Without Feature selection, SVM performed best with classification accuracy measured this 91%. The least accurate algorithm was Random Forest (88%) whilst the level of performance of LR (90%) was acceptable.

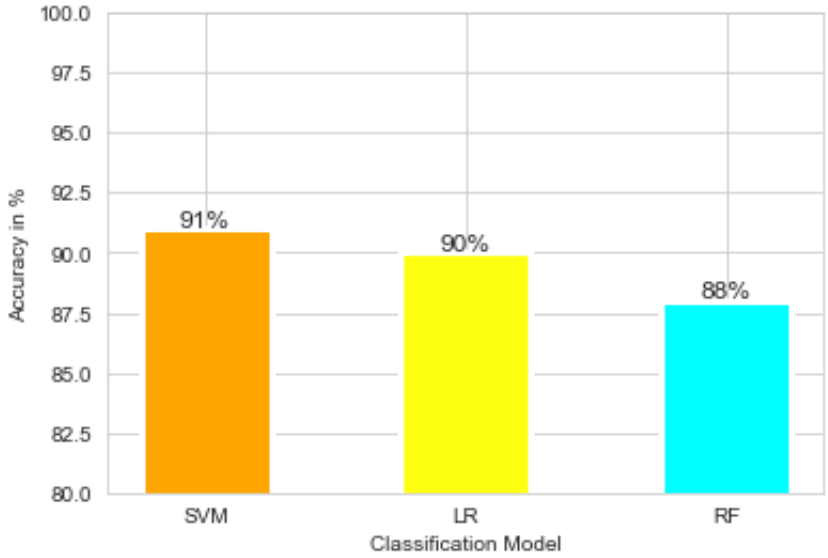

Figure 2: **Classification accuracy of different models without feature selection**. The classification accuracy are shown for different machine learning protocols namely SVM, LR and RF

As listed in Table 1, the following best F1 scores were determined: ring 96%, late_troph 94%, early_troph 86% and schizont 80%.

Features selection

From the data set, we observed that there are many genes which do not change expression levels across the life cycle stages. Thus, a feature selection algorithms would drastically

Table 1: F1 scores of different models of the different classes without Feature Selection

F1 scores(%) / Malaria Life Cycle Stage	SVM	LR	RF
early_troph	0.86	0.85	0.79
late_troph	0.94	0.91	0.91
ring	0.94	0.96	0.95
schizont	0.80	0.74	0.77

reduce the dimension and would allow us to extract important genes, the expression of which are responsible for the life cycle changes. In order to select the features, a genetic algorithm (GA) based– pipeline is implemented (Material and methods for details). The GA pipeline outputs the most optimal features and has removed the redundant features. Table 2 shows the details of the number of features. The initial number of features was **5066** out of which a subset of size **336** was selected using the GA. The dataset was reduced by **93.3%**.

Table 2: Numbers of Features selected

	Number of Features
Full Dataset	5066
Features Selected after GA pipeline	336

Classification with feature selection

This section shows the detail of the classification results with feature selection stage. Figure 3 shows the accuracy of SVM, LR, and RF. SVM performed best with classification accuracy measured this 91%. RF and LR both gave this 90% accuracy. The coming sections shows the test results of the SVM, LR, and RF models respectively.

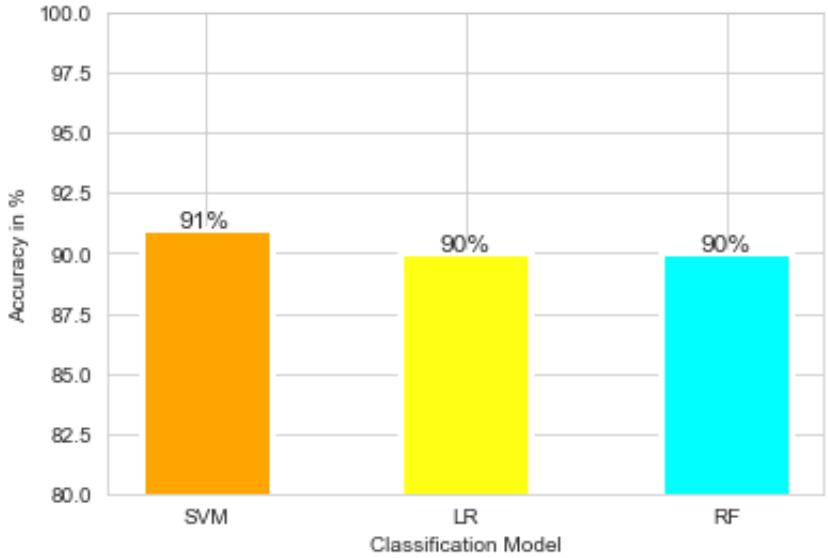

Figure 3: **Classification accuracy of different models with feature selection**. The classification accuracy are shown for different machine learning protocols namely SVM, LR and RF after selection of the 336 feature following genetic algorithm.

Using multiclass Support Vector Machine

Table 3 presents the precision, recall, and f1 scores of the SVM model for the four different classes. For late_troph and ring, we have achieved an f1 score of 0.93 and 0.94, respectively. We got an average f1 score for early_troph at 0.86. Schizont was the worst, at 0.78.

Table 3: Test results of SVM model with feature selection

Metric (%) / Malaria life cycle stage	precision	recall	F1-score
early_troph	0.92	0.82	0.86
late_troph	0.91	0.95	0.93
ring	0.90	0.99	0.94
schizont	0.92	0.67	0.78

Using Logistic Regression

Table 4 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 0.91 and 0.95, respectively. We got an average f1 score for early_troph at 0.87. Schizont was the worst, at 0.74.

Table 4: Test results of LR model with feature Selection

Metric (%) / Malaria Life Cycle Stage	precision	recall	F1-score
early_troph	0.87	0.86	0.87
late_troph	0.90	0.93	0.91
ring	0.95	0.96	0.95
schizont	0.83	0.66	0.74

Using Random Forest

Table 5 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 0.92 and 0.95, respectively. We got an average f1 score for early_troph at 0.83. Schizont was the worst at, 0.75.

Table 5: Test results of RF model with feature selection

Metric (%) / Malaria life cycle stage	precision	recall	F1-score
early_troph	0.91	0.77	0.83
late_troph	0.88	0.97	0.92
ring	0.93	0.97	0.95
schizont	0.93	0.63	0.75

Confusion matrix and mutual information between predicted and true labels for three models

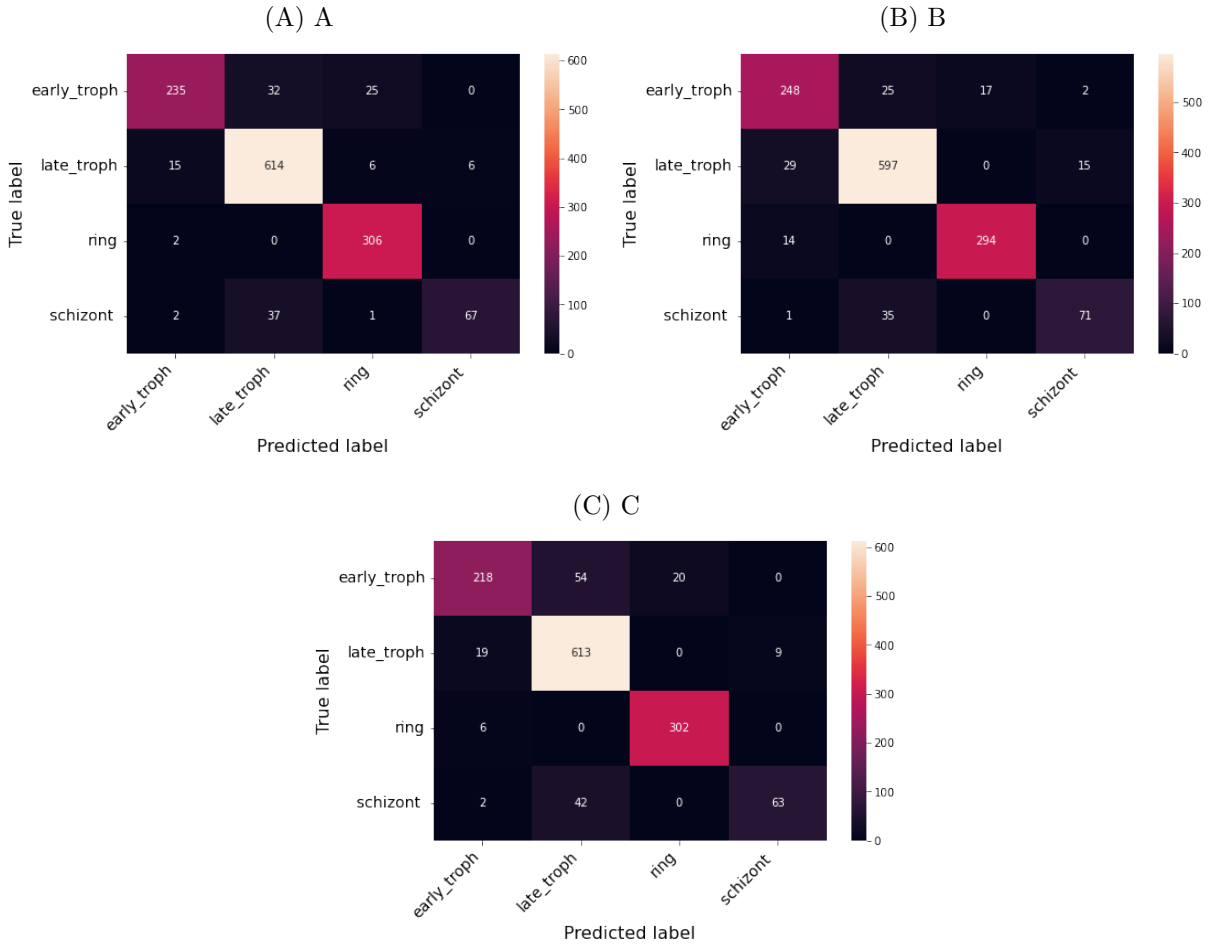

Figure 4: **Confusion matrix of different models show the prediction accuracy for different stages**. The heatmaps display the confusion matrix in predicting the four different stages as indicated after feature selection for three different models (A) SVM (B) LR (C) RF models

Figure 4 shows the confusion matrix of the three models. The confusion matrix for the SVM model shows that 32 samples were predicted as late_troph which should have been labelled as early_troph. Similarly, 37 samples were predicted as late_troph which were otherwise labelled as schizont. For late_troph class, 15 samples were misclassified as early_troph. For ring class, 2 samples were misclassified as early_troph.

The confusion matrix for the LR model shows that 25 samples were predicted as late_troph

which should have been labelled as early_troph. Similarly, 35 samples were predicted as late_troph which were otherwise labelled as schizont. For late_troph class, 29 samples were misclassified as early_troph. For ring class, 14 samples were misclassified as early_troph.

The confusion matrix for the RF model shows that 54 samples were predicted as late_troph which should have been labelled as early_troph. Similarly, 42 samples were predicted as late_troph which were otherwise labelled as schizont. For late_troph class, 19 samples were misclassified as early_troph. For ring class, 6 samples were misclassified as early_troph. These were some of the common misclassifications in all three models.

A comparison of classification without vs with feature selection.

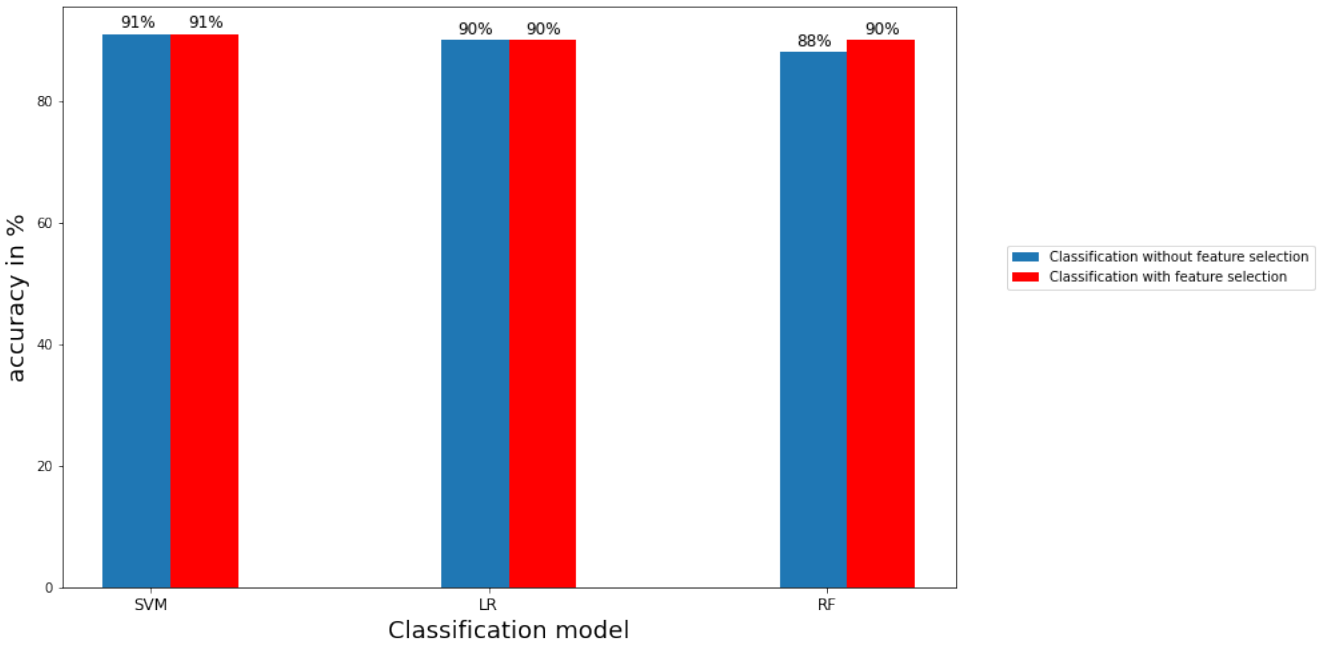

Figure 5: **Classification accuracy with feature selection vs without feature selection demonstrate the legitimacy of the selected features**. The bar graphs display a comparison between the values of accuracy for the three models and for classification with feature selection and without feature selection as indicated

Figure 5 shows the accuracy of classification without feature selection vs. classification with feature selection. We can see that we have reduced our feature set from 5066 to 336. Using these 336 features, we have achieved an accuracy of 91% in the SVM model, 90% in the LR

model, and 90% in the RF model. We have improved the accuracy of the RF model and have achieved similar results in the SVM and LR models.

For the SVM model, without feature selection, we got a f1 score of 0.86, 0.94, 0.94, and 0.80, where as with feature selection, we got a f1 score of 0.86, 0.93, 0.94, and 0.78 for early_troph, late_troph, ring, and schizont, respectively. For the LR model, without feature selection, we got a f1 score of 0.85, 0.91, 0.96, and 0.74, where as with feature selection, we got a f1 score of 0.87, 0.91, 0.95, and 0.74 for early_troph, late_troph, ring, and schizont, respectively. For the RF model, without feature selection, we got a f1 score of 0.79, 0.91, 0.95, and 0.77, where as with feature selection, we got a f1 score of 0.83, 0.92, 0.95, and 0.75 for early_troph, late_troph, ring, and schizont, respectively. We can see from the results that just by using the 336 features we have achieved similar f1 scores for all the four classes across all three models. This proves the robustness of the features selected from the GA pipeline. For the early_troph class, we have achieved the best f1 score of 0.87 from the LR model. For the late_troph class, we have achieved the best f1 score of 0.93 from the SVM model. For the ring class, we have achieved the best f1 score of 0.95 from both the SVM and RF models. Schizont was the worst, with an f1 score of 0.78 from the SVM model. This could be because less number of instances of the schizont class in the dataset compared to other classes.

We also calculated the mutual information (MI) between the predicted labels and the true labels of the three models using the joint probabilities from confusion matrix (see Subsection). For instance, C(1,1) of the confusion matrix represents the 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 6 shows the comparison of MI with and without feature selection. One of the advantages of displaying the 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.

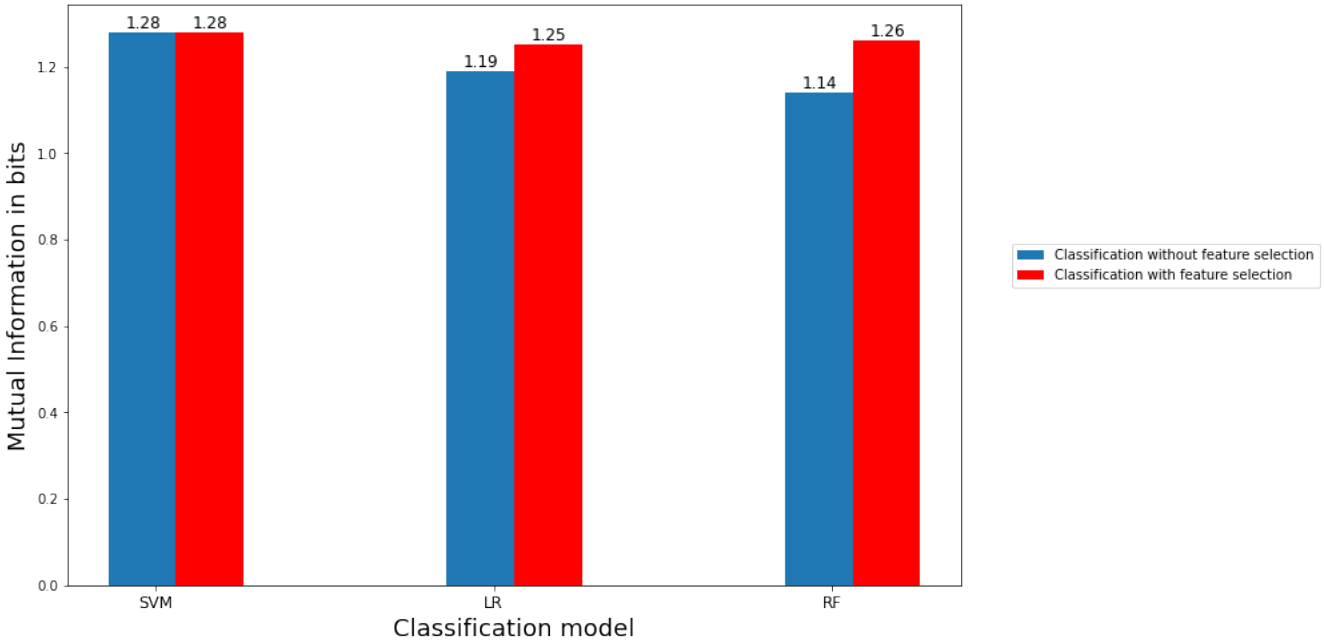

Figure 6: **Mutual information with and without feature selection**. The bar graphs display a comparison between the values of mutual information in bits between predicted and actual labels for the three models and for classification with feature selection and without feature selection as indicated

Classification with randomly selecting 336 features

In order to test whether the GA based feature selection algorithm is able to select the features appropriately, We randomly chose 336 features from our dataset and evaluated the prediction accuracy using the SVM, LR, and RF models. We achieved accuracy of 0.83, 0.83, and 0.83 for the models. Table 6shows the f1 scores of the different classes for the three models.

Table 6: F1 scores of different models of the different classes with randomly selecting 336 features

F1 scores / Malaria Life Cycle Stage	SVM	LR	RF
early_troph	0.69	0.70	0.68
late_troph	0.86	0.85	0.87
ring	0.89	0.89	0.89
schizont	0.77	0.72	0.75

The accuracy and the f1 scores of this experiment were lower compared to the classification results with feature selection using the GA pipeline. This results demonstrate the legitimacy of the feature selection method.

Construction and analysis of protein-protein interaction network

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.[15] Protein-protein interaction networks (PPIN) are graphs of the physical 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 336 proteins in Plasmodium falciparum. We used the Search Tool for the Retrieval of Interacting Genes/Proteins database (STRING 11.0b)[16] 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 0.90). The network construction shows a set of highly connected modules (Figure 7).

The topological analysis of the PPI network

Various topological measures are generally used to evaluate the both global and node characteristics in the PPI networks, including degree (k), between centrality (BC), eccentricity, closeness centrality (CC), eigenvector centrality (EC), and clustering coefficient.[17] Here, highest degree nodes are identified using degree distribution. Additionally, We have used Markov Clustering Algorithm (MCL) to find clusters in the network. Among these clusters, we identified clusters which also contains the node with highest degree and high betweennss centrality.

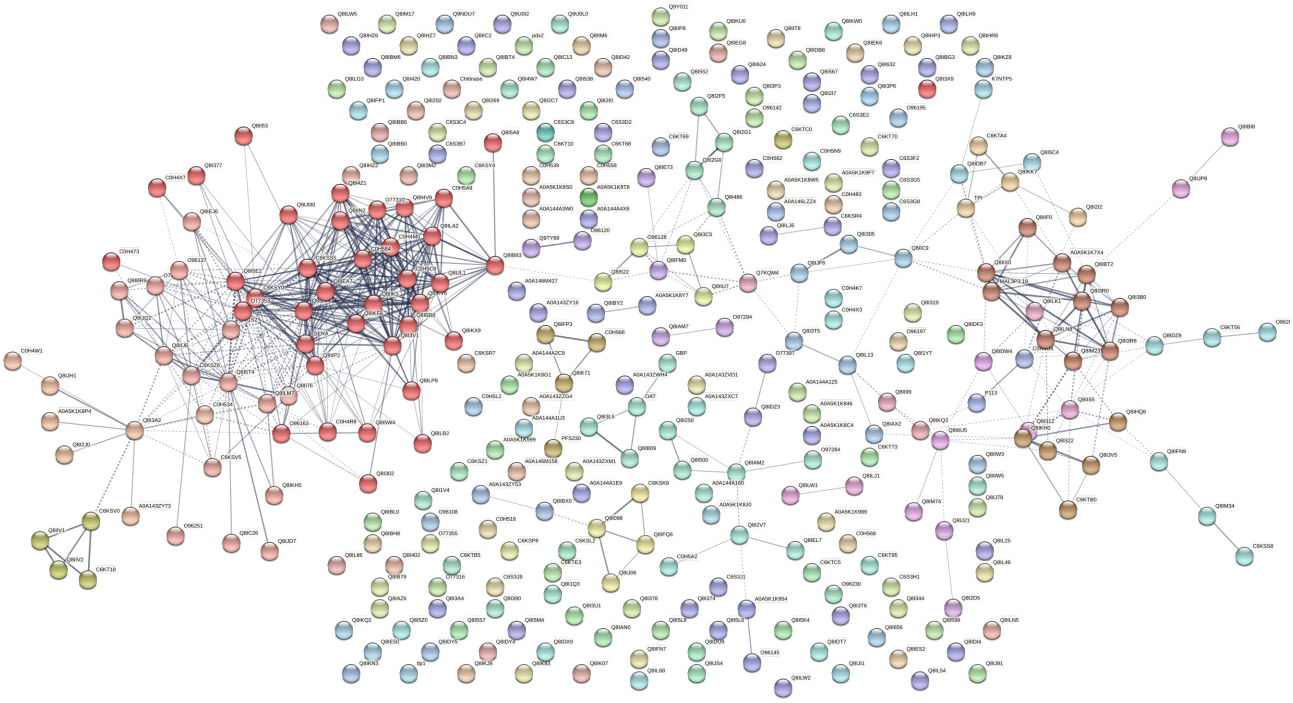

Figure 7: **Protein-protein interaction network exhibits different clusters**. The graph shows the protein protein interaction network of the 336 proteins selected by the feature selection method. The different colors indicate different identified clusters.

This PPIN is composed of of 334 nodes with number of edges: 571, average node degree: 3.42, avg. local clustering coefficient: 0.304, expected number of edges: 519 PPI enrichment p-value: 0.013. We can see that proteins in red cluster have highest degree and high betweenness centrality. So, we can consider red cluster as disease module. We have analysed other topological properties like degree, BC, eccentricity, CC, EC, clustering coefficient, etc of this Red cluster using Gephi.[18]

The proteins in Table 7 from red cluster have high degree and betweenness centrality (BC). In this cluster, we have, number of nodes: 37, number of edges: 272, average node degree: 14.7 avg. local clustering coefficient: 0.79, expected number of edges: 128 PPI enrichment p-value: ¡ 1.0e-16. We can see that this cluster has lesser nodes with high interaction and high clustering coefficient. So, this is a small world network. We can see from the above table that C6KSS5, Q8I2V4, C6KSY0 and Q8IKV6 has highest degree with high BC .We considered these four proteins as the hubs or bottlenecks as this nodes has high degree

Table 7: Topological analysis of the PPI network of the selected proteins

Proteins name	Degree	betweeness centrality	Description
C6KSS5	27	78.99	conserved Plasmodium protein, unknown function
Q8I2V4	27	73.99	Regulator of chromosome condensation-PP1-interacting protein
C6KSY0	26	124.65	AP2 domain transcription factor, putative(1. regulation of transcription, DNA-templated 2. DNA-binding transcription factor activity)
Q8IKV6	26	65.30	similar to Rhoptry neck protein 2(function unknown)
Q8IIK5	26	36.76	Rhoptry neck protein 4 (Antigen binding)
Q8IEA7	26	35.61	Uncharacterized protein
Q8I3V1	24	106.02	Glideosome-associated protein 40, putative
C0H5C8	24	41.29	Protein kinase domain-containing protein (protein kinase activity and ATP binding)
C0H584	24	30.43	Protein MAM3, putative (magnesium ion homeostasis).

(k) and BC. We have chosen 7 more proteins that has high degree and BC to consider them backbone of the PPIN. These proteins are C6KSS5, Q8I2V4, C6KSY0, Q8IKV6, Q8IIK5, Q8IEA7, Q8I3V1, C0H5C8, C0H584. These 9 proteins are highly connected in PPIN and has control over the network. Next we have checked for the functionality of all these proteins.

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 336 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 8) 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 336 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 8b 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.

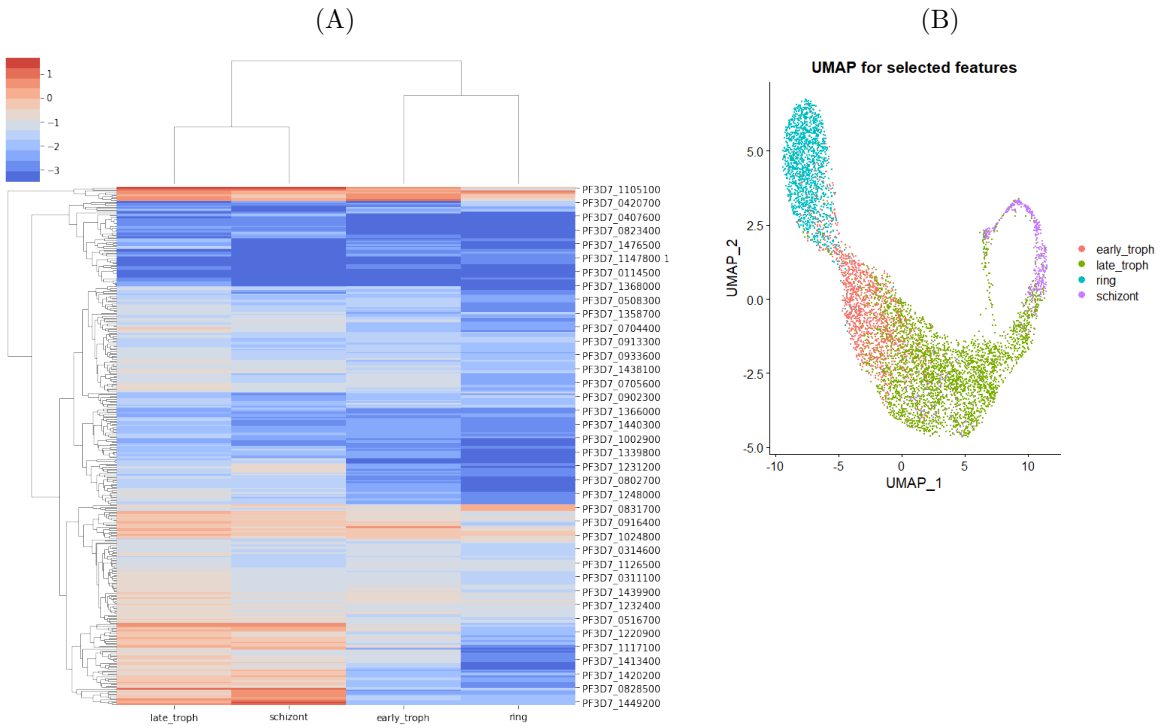

Figure 8: **The expression profiles are distinct among the stages**. (A) Expression Profile of the selected genes across the different stages. The heatmap shows the average RNA-count of the selected 336 genes across the different stages as indicated. A hierarchical clustering is performed on the expression levels in order group genes with similar expression patterns indicated by the dendogram. (B) UMAP of scRNA-seq counts of the selected features.Non-linear dimensional reduction of expression values of the selected 336 features.The cell clusters are colored based on the blood cycle stages of *P.falciparum.*

Discussion and outlook

The present study has proposed a two-stage model for feature selection and classification which has been shown to improve the classification of the different stages of the Malaria Life Cycle. This was achieved by removing the irrelevant features from the total data set considered for analysis. 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.

Various machine learning (ML) approaches have been proposed to yield more accurate results. Karthik and Sudha ,[19] reviewed ML methods for classifying gene expression model or computational analytical structure for complicated diseases, by identifying several differentially expressed gene techniques. Authors in [20] used the Convolutional Neural Networks (CNNs) based deep learning models for attribute extraction and categorization. For achieving higher categorization accuracy, they selected certain dominating features including size, colour, shape and cell count from the images. Similarly, a more effective two stage approach based on CNNs on a larger dataset was also proposed by.[21] It remains an especially challenging task to distinguish the multiple growth stages of parasites. Seng et al.[22] 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.

Numerous ML approaches have been proposed in the literature to enhance gene expression data classification such as clustering, classification, dimensional reduction, among others[23] . 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 random forests classifier in.[24,25] Hossain et al.[26] designed an effective variational quantum circuit (VQC-based) approach to recognize the existence of malaria from RBC images through the classification of optimized feature set extracted from

them. Murad et al.[27] used algorithms based on multifilter and hybrid approaches to feature selection leading to accuracy in excess of 90%. Mei et al.[28] suggested a dimensionality reduction method for classifying tumour gene expression data. Arowolo et al.[29] and Li et al.[30] 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.[31] 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.[32] Cheng-Lung Huang[33] 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.[34] 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[35] used Particle Swarm optimisation and GA in order to perform highly accurate classification. The authors in[36] achieved a classification accuracy of 90.32 % using GA for feature selection and a SVM Classifier.

For further research the hybrid methods for feature selection, impact of parameter fine tuning on various algorithms'level and use of other methods including Ensemble Learning may be attempted. Next, we have constructed protein-protein interaction network between our proteins (which we get after using feature selection algorithm) and we have done network analysis on this protein-protein interaction network using STRING 11.0b and Gephi software. Various topological measures are estimated to evaluate the node characteristics in the PPI networks, including degree (k), between centrality (BC), eccentricity, closeness centrality (CC), eigenvector centrality (EC), and clustering coefficient.[17] We found degree and betweeness centrality of each protein though this calculation. Proteins having high de-

gree and betweeness centrality tend to assert more control over the network function. These proteins can be considered as drug targets for future studies.

Material and Methods

System Design

The proposed classification technique comprises of two stages, namely:

1. Dimensionality Reduction with Feature Selection

2. Classification

Dimensionality Reduction with Feature Selection

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 high computational time. Dimensionality reduction is a technique that removes redundant features that hinder performance. We have used the feature selection[37] dimensionality reduction technique.

Let X be the initial m dimensional set of features, defined by the equation X = x(i)— i = 1,2,...m where x(i) are the defined features and m are the genes. The process of feature selection generates Y(i)—i = 1,2,...,p where Y(i) represents the new subset of features and p is now the number of features in the subset with p ¡= m. There are three types of feature selection methods - Filter, Wrapper, and Embedded approaches[38,39] .

Figure 9 shows the the proposed computational model framework. The feature selection stage uses GA for dimensionality reduction.Genetic Algorithm (GA) is a metaheuristic, evolutionary, stochastic optimization algorithm inspired by the process of natural selection. GAs are commonly used to generate optimum solutions to problems employing three biologically inspired operators selection, crossover and mutation.

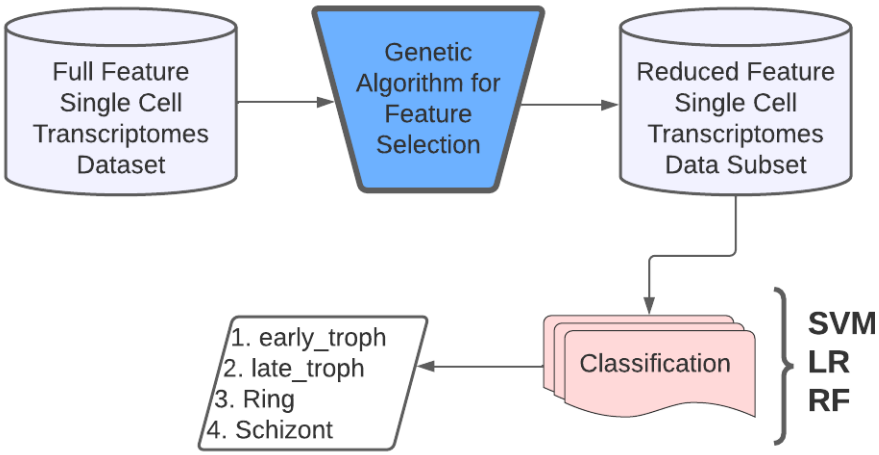

Figure 9: Proposed computational model framework

ruled Comment/* */

Classification

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 or integer or real-valued. In the field of machine learning, the 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. Figure 10 shows the pseudo code of the entire pipeline. We have used three classification algorithms in our research viz. Support Vector Machines, Logistic Regression and Random Forest.

Support vector machines (SVM) are one of the most popular, robust, non-probabilistic, binary, linear and non-linear classifier, supervised learning algorithms that analyze data for classification or regression analysis in machine learning. Based on a set of categorized training data, an SVM training model assigns new unseen examples to either of the trained categories, creating the decision boundaries (or hyperplanes) that can segregate n-dimensional space into classes.

Logistic regression (LR) is another powerful supervised ML algorithm used for binary classification problems that can be generalized to multiclass classification. A logistic function is used to model the probability of a discrete outcome based on an input variable. LR is an extensively employed algorithm for classification problems in industry owing to its high simplicity and efficiency particularly for linearly separable classes.

Random forests (RF) are an ensemble learning classification method and work by constructing a multitude of decision trees at training time. For classification jobs, the RF output is the class selected by most trees. Ensemble learning combines many classifiers to provide solutions to complex problems. In machine learning RFs also assist in reducing the training set overfitting by decision trees and also increases precision.

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

Begin: General Steps for feature selection using GA

Initialization:
t ← 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} ←newly created individuals
    t ← t + 1
end
Solution#1 ←Individuals with maximum fitness

Function EVALUATE_FITNESS(P):
    clf ← Random_Forest_Classifier()
    for each individual i ϵ P do
        fitness(i) ← 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 10: Psuedocode of the proposed method

Experimentation

The following sections will introduce each of the followed steps in detail. A summary of the implementation of the entire pipeline is depicted in Figure 11.

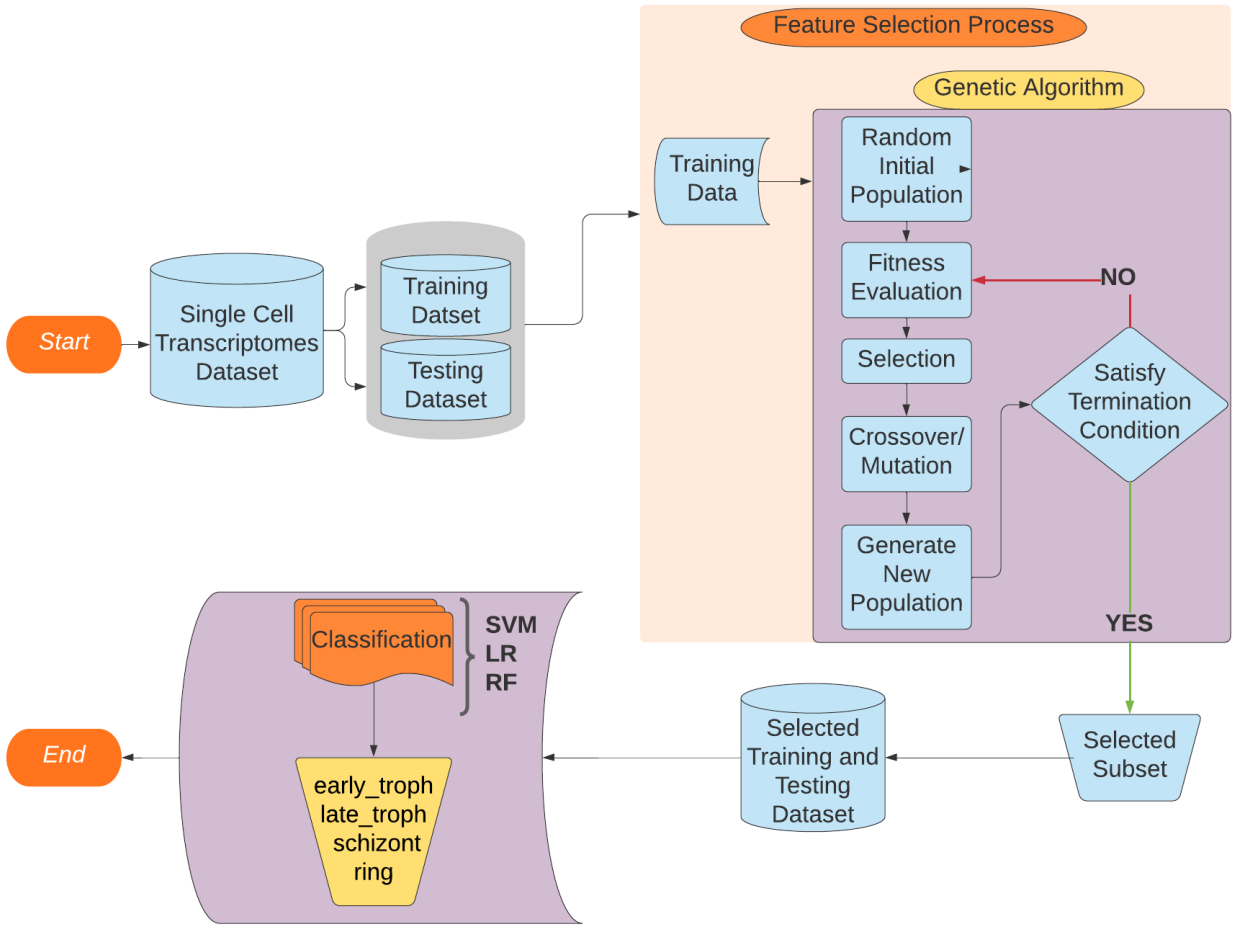

Figure 11: Flowchart of entire pipeline

Data Preparation

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.

Genetic Algorithm

Genetic Algorithm (GA) is a stochastic evolutionary optimization technique. It starts with an initial randomized set of population of features (500 here) and then creates another population using subsets of the available features whose individuals are evaluated using the Random Forest predictive model for the target task. The tournament 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. The individuals of this subset are stored in the Hall of Fame which is continuously sorted so as to have the first element with maximum fitness value so far. This process is iterated to yield the optimum features for the set termination criteria (maximum generations = 100, if no change in the best-fitted individual for 20 generations). Few other GA modelling parameters used in our research are - uniform crossover probability of 0.5, flip-bit-mutation probability of 0.2.

Classification Process

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

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:[40]

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

where $P(X_1, X_2)$ is the joint distribution of the two variables. $P(X_1)$ and $P(X_2)$ is the marginal distribution of the two variables. It's a dimensionless quantity that's measured in bits. Each element of 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 probability of true label $P(y)$ and conditional probability $P(y'|y)$. $P(y')$ is given by sum of joint probability over true label y. We have used this to find the $I(y; y')$.

Data and Software Availability

The data are freely accessible as a processed dataset through a user-friendly web interface (www.sanger.ac.uk/science/tools/mca/mca/).[13] Our dataset has 5066 rows and 6737 columns. Each row corresponds to single cell and each column corresponds to a gene. We have 5066 feature in our dataset. We have four different malaria life cycle stages (early troph, late troph, ring, 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 Genetic Algorithm 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[14] 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 336 proteins in Plasmodium falciparum. We have used the Search Tool for the Retrieval of Interacting Genes/Proteins database (STRING 11.0b)[16] to construct the PPI network associated with these proteins. STRING software `https://string-db.org/` 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 0.90). Various topological measures are generally used to evaluate the both global and node characteristics in the PPI networks, including degree (k), between centrality (BC), eccentricity, closeness centrality (CC), eigenvector centrality (EC), and clustering coefficient.[17] Here, highest degree nodes are identified using degree distribution. Additionally, We have used Markov Clustering Algorithm (MCL) (using STRING) to find clusters in the network. Among these clusters, we identified red cluster which contain the node with highest degree and high betweennss centrality. We have analysed topological properties like degree, BC, eccentricity, CC, EC, clustering coefficient, etc of the Red cluster using Gephi[18] software.

Acknowledgement

Authors thank the Department of Biotechnology (No. BT/RLF/Re-entry/32/2017), Government of India for funding.

References

(1) Abbas, N.; Saba, T.; Rehman, A.; Mehmood, Z.; Kolivand, H.; Uddin, M.; Anjum, A. Plasmodium life cycle stage classification based quantification of malaria parasitaemia in thin blood smears. *Microscopy research and technique* **2019**, *82*, 283–295.

(2) Chappell, L.; Ross, P.; Orchard, L.; Russell, T. J.; Otto, T. D.; Berriman, M.;

Rayner, J. C.; Llinás, M. Refining the transcriptome of the human malaria parasite Plasmodium falciparum using amplification-free RNA-seq. *BMC genomics* **2020**, *21*, 1–19.

(3) Poostchi, M.; Silamut, K.; Maude, R. J.; Jaeger, S.; Thoma, G. Image analysis and machine learning for detecting malaria. *Translational Research* **2018**, *194*, 36–55.

(4) Ngara, M.; Palmkvist, M.; Sagasser, S.; Hjelmqvist, D.; Björklund, Å. K.; Wahlgren, M.; Ankarklev, J.; Sandberg, R. Exploring parasite heterogeneity using single-cell RNA-seq reveals a gene signature among sexual stage Plasmodium falciparum parasites. *Experimental cell research* **2018**, *371*, 130–138.

(5) Sa, J. M.; Cannon, M. V.; Caleon, R. L.; Wellems, T. E.; Serre, D. Single-cell transcription analysis of Plasmodium vivax blood-stage parasites identifies stage-and species-specific profiles of expression. *PLoS biology* **2020**, *18*, e3000711.

(6) Reid, A. J.; Talman, A. M.; Bennett, H. M.; Gomes, A. R.; Sanders, M. J.; Illingworth, C. J.; Billker, O.; Berriman, M.; Lawniczak, M. K. Single-cell RNA-seq reveals hidden transcriptional variation in malaria parasites. *elife* **2018**, *7*, e33105.

(7) Walzer, K. A.; Fradin, H.; Emerson, L. Y.; Corcoran, D. L.; Chi, J.-T. Latent transcriptional variations of individual Plasmodium falciparum uncovered by single-cell RNA-seq and fluorescence imaging. *PLoS genetics* **2019**, *15*, e1008506.

(8) Poran, A.; Nötzel, C.; Aly, O.; Mencia-Trinchant, N.; Harris, C. T.; Guzman, M. L.; Hassane, D. C.; Elemento, O.; Kafsack, B. F. Single-cell RNA sequencing reveals a signature of sexual commitment in malaria parasites. *Nature* **2017**, *551*, 95–99.

(9) Rawat, M.; Srivastava, A.; Johri, S.; Gupta, I.; Karmodiya, K. Single-Cell RNA Sequencing Reveals Cellular Heterogeneity and Stage Transition under Temperature Stress in Synchronized Plasmodium falciparum Cells. *Microbiology spectrum* **2021**, *9*, e00008–21.

(10) Pradhan, M. Evolutionary computational algorithm by blending of PPCA and EP-Enhanced supervised classifier for microarray gene expression data. *IAES International Journal of Artificial Intelligence* **2018**, *7*, 95.

(11) Jain, D.; Singh, V. An efficient hybrid feature selection model for dimensionality reduction. *Procedia Computer Science* **2018**, *132*, 333–341.

(12) Singh, D. A. A. G.; Leavline, E. J.; Priyanka, R.; Priya, P. P. Dimensionality reduction using genetic algorithm for improving accuracy in medical diagnosis. *International Journal of Intelligent Systems and Applications* **2016**, *8*, 67.

(13) Howick, V. M.; Russell, A. J.; Andrews, T.; Heaton, H.; Reid, A. J.; Natarajan, K.; Butungi, H.; Metcalf, T.; Verzier, L. H.; Rayner, J. C., et al. The Malaria Cell Atlas: Single parasite transcriptomes across the complete Plasmodium life cycle. *Science* **2019**, *365*, eaaw2619.

(14) Butler, A.; Hoffman, P.; Smibert, P.; Papalexi, E.; Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology* **2018**, *36*, 411–420.

(15) Silverbush, R., Dana Sharan A systematic approach to orient the human protein–protein interaction network. *Nature Communications* **2019**,

(16) Szklarczyk D, N. K. L. D. K. R. P. S. D. N. L. M. F. T. B. P. J. L. v. M. C., Gable AL The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* **2021**, *49*, 91–107.

(17) Albert, R.; Barabási, A.-L. Statistical mechanics of complex networks. *Reviews of modern physics* **2002**, *74*, 47.

(18) Bastian, M.; Heymann, S.; Jacomy, M. Gephi: An Open Source Software for Exploring and Manipulating Networks. 2009; `http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154`.

(19) Karthik, S.; Sudha, M. 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* **2018**, *8*, 182–191.

(20) Chima, J. S.; Shah, A.; Shah, K.; Ramesh, R. Malaria Cell Image Classification Using Deep Learning. *International Journal of Recent Technology and Engineering* **2020**, *8*, 5553–59.

(21) Arshad, Q. A.; Ali, M.; Hassan, S.-u.; Chen, C.; Imran, A.; Rasul, G.; Sultani, W. A dataset and benchmark for malaria life-cycle classification in thin blood smear images. *Neural Computing and Applications* **2022**, *34*, 4473–4485.

(22) Li, S.; Du, Z.; Meng, X.; Zhang, Y. Multi-stage malaria parasite recognition by deep learning. *GigaScience* **2021**, *10*, giab040.

(23) AROWOLO, M. O.; ADEBIYI, M. O.; NNODIM, C. T.; ABDULSALAM, S. O.; ADEBIYI, A. A. 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)* **2021**, *18*, 9849–11.

(24) Li, Q.; Dong, B.; Wang, D.; Wang, S. Identification of Secreted Proteins From Malaria Protozoa With Few Features. *IEEE Access* **2020**, *8*, 89793–89801.

(25) Mishra, S. K. Human Malaria Detection and Stage Classification using Random Forest Classifier.

(26) Hossain, M. M.; Rahim, M. A.; Bahar, A. N.; Rahman, M. M. Automatic malaria disease detection from blood cell images using the variational quantum circuit. *Informatics in Medicine Unlocked* **2021**, *26*, 100743.

(27) Al-Rajab, M.; Lu, J.; Xu, Q. A framework model using multifilter feature selection to enhance colon cancer classification. *Plos one* **2021**, *16*, e0249094.

(28) Mei, Q.; Zhang, H.; Liang, C. A discriminative feature extraction approach for tumor classification using gene expression data. *Current Bioinformatics* **2016**, *11*, 561–570.

(29) Arowolo, M. O.; Adebiyi, M. O.; Adebiyi, A. A.; Okesola, O. J. A hybrid heuristic dimensionality reduction methods for classifying malaria vector gene expression data. *IEEE Access* **2020**, *8*, 182422–182430.

(30) Li, J.; Zhao, Z.; Zhou, L.; Wang, Y. Y-SPCR: A new dimensionality reduction method for gene expression data classification. 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2019; pp 401–408.

(31) Rokach, L. Genetic algorithm-based feature set partitioning for classification problems. *Pattern Recognition* **2008**, *41*, 1676–1700.

(32) Zhang, Z.; Yang, P. An ensemble of classifiers with genetic algorithmbased feature selection. *The IEEE intelligent informatics bulletin* **2008**, *9*, 18–24.

(33) Huang, C.-L.; Wang, C.-J. A GA-based feature selection and parameters optimization-for support vector machines. *Expert Systems with applications* **2006**, *31*, 231–240.

(34) Chuang, L.-Y.; Ke, C.-H.; Chang, H.-W.; Yang, C.-H. A two-stage feature selection method for gene expression data. *OMICS A journal of Integrative Biology* **2009**, *13*, 127–137.

(35) Li, S.; Wu, X.; Tan, M. Gene selection using hybrid particle swarm optimization and genetic algorithm. *Soft Computing* **2008**, *12*, 1039–1048.

(36) Mohamad, M. S.; Deris, S.; Illias, R. M. 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* **2005**, *5*, 91–107.

(37) Almugren, N.; Alshamlan, H. A survey on hybrid feature selection methods in microarray gene expression data for cancer classification. *IEEE access* **2019**, *7*, 78533–78548.

(38) Jabeen, A.; Ahmad, N.; Raza, K. *Classification in BioApps*; Springer, 2018; pp 133–172.

(39) Sahu, B.; Dehuri, S.; Jagadev, A. A study on the relevance of feature selection methods in microarray data. *The Open Bioinformatics Journal* **2018**, *11*.

(40) Cover, T. M. *Elements of information theory*; John Wiley & Sons, 1999.