



METHOD ARTICLE

REVISED Ensemble machine learning modeling for the prediction of artemisinin resistance in malaria [version 5; peer review: 1 approved, 2 approved with reservations]

Colby T. Ford^{1,2}, Daniel Janies¹¹Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, Charlotte, North Carolina, 28223, USA²School of Data Science, University of North Carolina at Charlotte, Charlotte, North Carolina, 28223, USA

v5 First published: 29 Jan 2020, 9:62
<https://doi.org/10.12688/f1000research.21539.1>
 Second version: 04 Feb 2020, 9:62
<https://doi.org/10.12688/f1000research.21539.2>
 Third version: 29 Apr 2020, 9:62
<https://doi.org/10.12688/f1000research.21539.3>
 Fourth version: 21 May 2020, 9:62
<https://doi.org/10.12688/f1000research.21539.4>
 Latest published: 25 Jun 2020, 9:62
<https://doi.org/10.12688/f1000research.21539.5>

Abstract

Resistance in malaria is a growing concern affecting many areas of Sub-Saharan Africa and Southeast Asia. Since the emergence of artemisinin resistance in the late 2000s in Cambodia, research into the underlying mechanisms has been underway.


The 2019 Malaria Challenge posited the task of developing computational models that address important problems in advancing the fight against malaria. The first goal was to accurately predict artemisinin drug resistance levels of *Plasmodium falciparum* isolates, as quantified by the IC₅₀. The second goal was to predict the parasite clearance rate of malaria parasite isolates based on *in vitro* transcriptional profiles.


In this work, we develop machine learning models using novel methods for transforming isolate data and handling the tens of thousands of variables that result from these data transformation exercises. This is demonstrated by using massively parallel processing of the data vectorization for use in scalable machine learning. In addition, we show the utility of ensemble machine learning modeling for highly effective predictions of both goals of this challenge. This is demonstrated by the use of multiple machine learning algorithms combined with various scaling and normalization preprocessing steps. Then, using a voting ensemble, multiple models are combined to generate a final model prediction.


Open Peer Review**Approval Status** ✓ ? ?

	1	2	3
version 5 (revision) 25 Jun 2020	✓ view		? view
version 4 (revision) 21 May 2020		? view	
version 3 (revision) 29 Apr 2020	✗ view		
version 2 (revision) 04 Feb 2020	✗ view		
version 1 29 Jan 2020			

1. **Sameer K. Antani** , National Institutes of Health, Bethesda, USA

Stefan Jaeger , National Institutes of Health, Bethesda, USA

2. **Jeremy Burrows** , Medicines for Malaria Venture (MMV), Geneva, Switzerland

3. **Alyssa E Barry** , Deakin University and Burnet Institute, Melbourne, Australia

Keywords

malaria, Plasmodium falciparum, machine learning, parallel computing, Apache Spark, big data, artemisinin, bioinformatics, DREAM Competition



This article is included in the **Artificial Intelligence and Machine Learning** gateway.

Myo Naung, Walter and Eliza Hall Institute,
University of Melbourne, Melbourne,
Australia

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Colby T. Ford (colby.ford@uncc.edu)

Author roles: **Ford CT:** Data Curation, Formal Analysis, Project Administration, Visualization, Writing – Original Draft Preparation; **Janies D:** Funding Acquisition, Investigation, Project Administration, Resources, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by the University of North Carolina at Charlotte Department of Bioinformatics and Genomics and the School of Data Science.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2020 Ford CT and Janies D. This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Ford CT and Janies D. **Ensemble machine learning modeling for the prediction of artemisinin resistance in malaria [version 5; peer review: 1 approved, 2 approved with reservations]** F1000Research 2020, 9:62
<https://doi.org/10.12688/f1000research.21539.5>

First published: 29 Jan 2020, 9:62 <https://doi.org/10.12688/f1000research.21539.1>

REVISED Amendments from Version 4

In this revision, we have addressed the latest reviewer's comments around the applicability of this work to the broader field of parasitology, also have also included some new work from Birnbaum *et al.* 2020. In addition, we also discuss the need for lab-based (*in vitro*) validation of these *in silico* findings, though this work helps to highlight the most probable/important things to test first.

It should be noted that there is some specific information that reviewers are asking about the input data that we do not have yet as this is part of a larger DREAM Challenge. Once this information is public, we will likely add it to this work as well.

Any further responses from the reviewers can be found at the end of the article

Introduction

Malaria is a serious disease caused by parasites belonging to the genus *Plasmodium* which are transmitted by *Anopheles* mosquitoes in the genus. The World Health Organization (WHO) reports that there were 219 million cases of malaria in 2017 across 87 countries¹. *Plasmodium falciparum* poses one of greatest health threats in Southeast Asia, being responsible for 62.8% of malaria cases in the region in 2017¹.

Artemisinin-based therapies are among the best treatment options for malaria caused by *P. falciparum*². The use of artemisinin in combination with other drugs, called artemisinin combination therapies, are the best treatment options today against malaria infections.

However, emergence of artemisinin resistance in Thailand and Cambodia in 2007 has been cause for research³. While there are polymorphisms in the kelch domain-carrying protein K13 in *P. falciparum* that are known to be associated with artemisinin resistance, the underlying molecular mechanism that confers resistance remains unknown⁴. In early 2020, Birnbaum *et al.* discovered that the highly-conserved gene *kelch13* is associated with a molecular mechanism that allows the parasite to feed on host erythrocytes by endocytosis of hemoglobin⁵. Given that artemisinin is activated by hemoglobin degradation products, these mutations can confer resistance to artemisinin.

The established pharmacodynamics benchmark for *P. falciparum* sensitivity to artemisinin-based therapy is the parasite clearance rate^{6,7}. Resistance to artemisinin-based therapy is considered to be present with a parasite clearance rate greater than five hours⁸. By understanding the genetic factors that affect resistance in malaria, targeted development can occur in an effort to abate further resistance or infections of resistant strains.

Previous research has shown success in applying similar machine learning methods in the explanation of genetic differences in plants⁹, fungi¹⁰, and even humans¹¹. Previous work in machine learning-based tropical disease research, including malaria and other diseases, has shown effective in drug discovery^{12,13} and in the understanding of degradomes¹⁴. Also, other machine learning work in malaria has focused on the identification and diagnosis of malaria using image classification¹⁵⁻¹⁷.

In this work, we create multiple machine learning-based models to address these issues around artemisinin resistance and parasite clearance. Given that the interpretation and analysis of many genes and their effects on resistance may be tedious, machine learning allows for a more power investigation into this relationship. Plus, we employ model explainability methods to help rank particular genes of interest in the malaria genome.

Prediction of artemisinin IC₅₀

First, we created a machine learning model to predict the IC₅₀ of malaria parasites based on transcription profiles of experimentally-tested isolates. IC₅₀, also known as the half maximal inhibitory concentration, is the drug concentration at which 50% of parasites die. This value indicates a population of parasites' ability to withstand various doses of antimalarial drugs, such as artemisinin.

Methods

Training data was obtained from the 2019 DREAM Malaria Challenge^{18,19}. The training data consists of gene expression data of 5,540 genes of 30 isolates from the malaria parasite, *Plasmodium falciparum*. For each malaria parasite isolate, transcription data was collected at two time points [6 hours post invasion (hpi) and 24 hpi], with and without treatment of dihydroartemisinin (the metabolically active form of artemisinin), each with a biological replicate. This yields a total of at eight data points for each isolate. The initial form of the training dataset contains 272 rows and 5,546 columns, as shown in [Table 1](#).

The transcription data was collected as described in [Table 2](#). The transcription data set consists of 92 non-coding RNAs (denoted by gene IDs that begins with 'MAL'), while the rest are protein coding genes (denoted by gene IDs that start with 'PF3D7'). The feature to predict is *DHA_IC50*.

Data preparation

We used Apache Spark²⁰ to pivot the dataset such that each isolate was its own row and each of the transcription values for each gene and attributes (i.e. timepoint, treatment, biological replicate) combination was its own column. This exercise transformed the training dataset from 272 rows and 5,546 columns to 30 rows and 44,343 columns, as shown in [Table 3](#). We completed this pivot by slicing the data by each of the eight combinations of timepoint, treatment, and biological replicate, dynamically renaming the variables (genes) for each slice, and then joining all eight slices back together.

By using the massively parallel architecture of Spark, this transformation can be completed in a minimal amount of time on a relatively small cluster environment (e.g., <10 minutes using a 8-worker/36-core cluster with PySpark on Apache Spark 2.4.3).

Lastly, the dataset is then vectorized using the Spark `VectorAssembler`, and converted into a Numpy²¹-compatible array. Vectorization allows for highly scalable parallelization of the machine learning modeling in the next step.

Table 1. Initial IC₅₀ model training data format. Note that for Treatment, *UT* represents untreated samples and *DHA* represents samples treated with dihydroartemisinin.

Sample_Name	Isolate	Timepoint	Treatment	BioRep	Gene ₁	...	Gene ₅₅₄₀	DHA_IC50
isolate_01.24HR.DHA.BRep1	isolate_01	24HR	DHA	BRep1	0.008286	...	-2.48653	2.177
isolate_01.24HR.DHA.BRep2	isolate_01	24HR	DHA	BRep2	-0.87203	...	-1.79457	2.177
isolate_01.24HR.UT.BRep1	isolate_01	24HR	UT	BRep1	0.03948	...	-2.49517	2.177
isolate_01.24HR.UT.BRep2	isolate_01	24HR	UT	BRep2	0.125177	...	-1.73531	2.177
isolate_01.6HR.DHA.BRep1	isolate_01	6HR	DHA	BRep1	1.354956	...	-0.82169	2.177
isolate_01.6HR.DHA.BRep2	isolate_01	6HR	DHA	BRep2	-0.21807	...	-1.61839	2.177
isolate_01.6HR.UT.BRep1	isolate_01	6HR	UT	BRep1	1.31135	...	-2.62262	2.177
isolate_01.6HR.UT.BRep2	isolate_01	6HR	UT	BRep2	0.997722	...	-2.24719	2.177
...
isolate_30.6HR.UT.BRep2	isolate_30	6HR	UT	BRep2	-0.26639	...	-1.72273	1.363

Table 2. IC₅₀ training data information. (Adapted from Turnbull *et al.*, (2017) PLoS One²²).

	Training Set
Array	Bozdech
Platform	Printed
Plexes	1
Unique Probes	10159
Range of Probes per Exon	N/A
Average Probes per Gene	2
Genes Represented	5363
Transcript Isoform Profiling	No
ncRNAs	No
Channel Detection Method	Two Color
Scanner	PowerScanner
Data Extraction	GenePix Pro

Machine learning

We used the Microsoft Azure Machine Learning Service²³ as the tracking platform for retaining model performance metrics as the various models were generated. For this use case, 498 machine learning models were trained using various scaling techniques and algorithms. Scaling and normalization methods are shown in Table 14. We then created two ensemble models of the individual models using Stack Ensemble and Voting ensemble methods.

The Microsoft AutoML package²⁴ allows for the parallel creation and testing of various models, fitting based on a primary metric. For this use case, models were trained using Decision Tree, Elastic Net, Extreme Random Tree, Gradient Boosting, Lasso Lars, LightGBM, RandomForest, and Stochastic Gradient Decent algorithms along with various scaling methods from Maximum Absolute Scaler, Min/Max Scaler, Principal Component Analysis, Robust Scaler, Sparse Normalizer, Standard Scale Wrapper, Truncated Singular Value Decomposition Wrapper (as defined in Table 14). All of the machine learning algorithms are from the *scikit-learn* package²⁵ except for LightGBM, which is from the *LightGBM* package²⁶. The settings for the model sweep are defined in Table 4. The ‘Preprocess Data?’ parameter enables the scaling and imputation of the features in the data. Note that these models were evaluated using random sampling of the input training dataset provided by the DREAM Challenge, though the evaluation within the challenge was performed on an unlabelled testing dataset. The metrics in the Results section below reflect the evaluation on the sampled training data.

Once the 498 individual models were trained, two ensemble models (voting ensemble and stack ensemble) were then created and tested. The voting ensemble method makes a prediction based on the weighted average of the previous models’ predicted regression outputs whereas the stacking ensemble method combines the previous models and trains a meta-model using the elastic net algorithm based on the output from the previous models. The model selection method used was the Caruana ensemble selection algorithm²⁷.

Results

The voting ensemble model (using soft voting) was selected as the best model, having the lowest normalized Root Mean Squared Error (RMSE), as shown in Table 5. The top 10 models trained are reported in Table 6. Having a normalized RMSE of only 0.1228 and a Mean Absolute Percentage Error (MAPE) of 24.27%, this model is expected to accurately predict IC₅₀ in malaria isolates. See Figure 1 for a visualization of the experiment runs and Figure 2 for the distribution of residuals on the best model.

Prediction of resistance status

The second task of this work was to create a machine learning model that can predict the parasite clearance rate (fast versus slow) of malaria isolates. When resistance rates change in a pathogen, it can be indicative of regulatory

Table 3. Post-transformation format of the IC₅₀ model training data.

Isolate	DHA_IC50	hr24_trDHA_br1_Gene ₁	hr24_trDHA_br2_Gene ₁	...	hr6_trUT_br2_Gene ₅₅₄₀
isolate_01	2.177	0.008286	-0.87203	...	-2.24719
...
isolate_30	1.363	0.195032	0.031504	...	-1.72273

Table 4. Model search parameter setting for the IC₅₀ model search.

Parameter	Value
Task	Regression
Number of Iterations	500
Iteration Timeout (minutes)	20
Max Cores per Iteration	7
Primary Metric	Normalized Root Mean Squared Error
Preprocess Data?	True
k-Fold Cross-Validations	20 folds

Table 5. Model metrics of the final IC₅₀ ensemble model.

Metric	Value
Normalized Root Mean Squared Error	0.1228
Root Mean Squared Log Error	0.1336
Normalized Mean Absolute Error	0.1097
Mean Absolute Percentage Error	24.27
Normalized Median Absolute Error	0.1097
Root Mean Squared Error	0.3398
Explained Variance	-1.755
Normalized Root Mean Squared Log Error	0.1379
Median Absolute Error	0.3035
Mean Absolute Error	0.3035

Table 6. Top 10 training iterations of the IC₅₀ model search, evaluated by Root Mean Squared Error. Note that the top performing model (VotingEnsemble) is the final IC₅₀ model discussed in this paper.

Iteration	Preprocessor	Algorithm	Normalized RMSE
498		VotingEnsemble	0.12283293
370	SparseNormalizer	RandomForest	0.132003138
432	StandardScalerWrapper	LightGBM	0.133180215
240	SparseNormalizer	RandomForest	0.133779391
430	StandardScalerWrapper	RandomForest	0.137084337
65	SparseNormalizer	RandomForest	0.13884791
56	SparseNormalizer	RandomForest	0.14417843
68	MaxAbsScaler	ExtremeRandomTrees	0.151925822
470	StandardScalerWrapper	RandomForest	0.152262231
181	MinMaxScaler	LightGBM	0.15279075

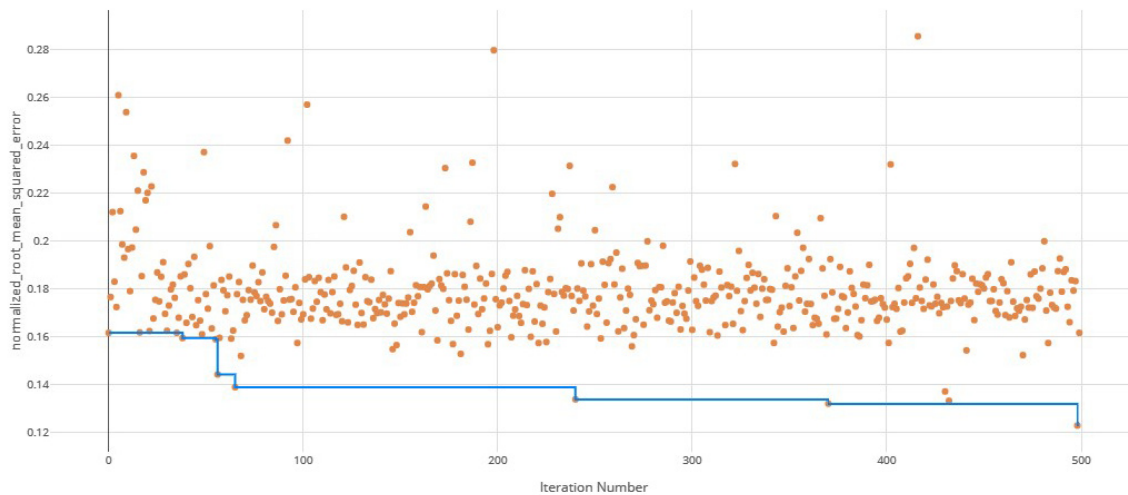


Figure 1. Root Mean Squared Error (RMSE) by iteration of the IC₅₀ model search. Each orange dot is an iteration with the blue line representing the minimum RMSE up to that iteration.

changes in the pathogen’s genome. These changes can be exploited for the prevention of further resistance spread. Thus, a goal of this work is to understand genes important in the prediction of artemisinin resistance. The relationship of this use case to the first is that parasite clearance is a measure of the effectiveness of a treatment regimen. While the first use case looked at the drug concentration, this use case looks into the speed at which the parasites are cleared as a result of a standard treatment.

Methods

An *in vivo* transcription data set from Mok *et al.*, (2015) Science²⁸ was used to predict the parasite clearance rate of malaria parasite isolates based on *in vitro* transcriptional profiles (see Table 8).

The training data consists of 1,043 isolates with 4,952 genes from the malaria parasite *Plasmodium falciparum*. For each malaria parasite isolate, transcription data was collected for various *PF3D7* genes. The form of the training dataset contains 1,043 rows and 4,957 columns, as shown in Table 7. The feature to predict is *ClearanceRate*.

Data preparation

The training data for this use case did not require the same pivoting transformations as in the last use case as each record describes a single isolate. Thus, only the vectorization of the data was necessary, which was performed using the Spark VectorAssembler and then converted into a Numpy-compatible array²². Note that this vectorization only kept the numerical columns, which excludes the Country, Kmeans_Grp, and Asexual_stage_hpi_ attributes as they are either absent or contain non-matching factors (i.e. different set of countries) in the testing data.

Machine learning

Once the 98 individual models were trained, two ensemble models (voting ensemble and stack ensemble) were then created and tested as before. Model search parameters are shown in Table 9.

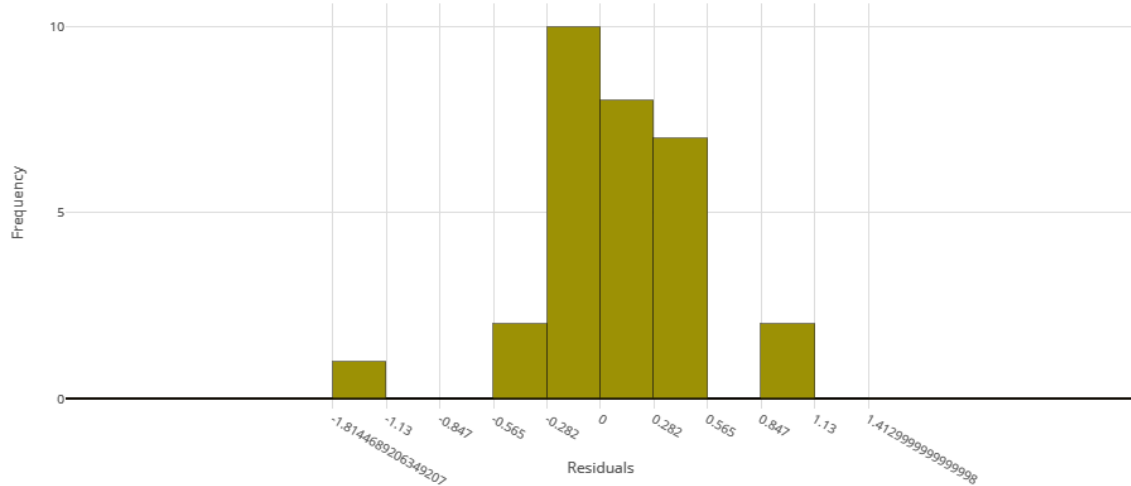


Figure 2. Model residuals of the final IC₅₀ ensemble model.

Table 7. Format of the clearance rate model training data.

Sample_Names	Country	Asexual_stage hpi_	Kmeans_Grp	PF3D7_0100100	...	PF3D7_1480100	ClearanceRate
GSM1427365	Bangladesh	20	B	0.226311	...	-0.64171	Fast
...
GSM1427537	Cambodia	12	C	0.81096	...	-1.72825	Slow
...
GSM1428407	Vietnam	8	A	0.999095	...	NaN	Fast

Table 8. Training dataset information from Mok et al., 2015²⁸.

	Training Set
Number of isolates	1043
Isolate collection site	Southeast Asia
Isolate collection years	2012–2014
Sample type	<i>in vivo</i>
Synchronized?	Not synchronized
Number of samples per isolate	1
Additional attributes	~18 hpi, Non-perturbed, No replicates

Table 9. Model search parameter settings for the clearance rate model search.

Parameter	Value
Task	Regression
Number of iterations	100
Iteration timeout (minutes)	20
Max cores per iteration	14
Primary metric	weighted area under the receiver operating characteristic curve (AUC)
Preprocess data?	True
k-Fold cross-validations	10 folds

Results

The voting ensemble model (using soft voting) was selected as the best model, having the highest area under the receiver operating characteristic curve (AUC), as shown in [Table 11](#). The top 10 of the 100 models trained are reported in [Table 10](#). Having a weighted AUC of 0.87 and a weighted F1 score of 0.80, this model is expected to accurately predict isolate clearance rates. A confusion matrix of the predicted results versus actuals is shown in [Table 12](#). See [Figure 3](#) for a visualization of the experiment runs and see [Figure 4](#) and [Figure 5](#) for the ROC and Precision-Recall curves on the best model. Note that these models were evaluated using random sampling of the input training dataset provided by the DREAM Challenge, though the evaluation within the challenge was performed on an unlabelled testing dataset. The metrics in the Results section below reflect the evaluation on the sampled training data.

Note that the averages reported in [Figure 4](#) and [Figure 5](#) are defined as follows:

- ‘micro’: Computed globally by combining the true positives and false positives from each class at each cutoff.
- ‘macro’: The arithmetic mean for each class. This does not take class imbalance into account.
- ‘weighted’: The arithmetic mean of the score for each class, weighted by the number of true instances in each class (support).

Feature importance

Feature importances were calculated using mimic-based model explanation of the ensemble model²⁹. The mimic explainer works by training global surrogate models to mimic blackbox models (i.e. complex models that are difficult to explain). The surrogate model is an interpretable model, trained to approximate the predictions of a black box model as accurately as possible³⁰. In [Figure 6](#) and [Table 13](#), the feature importance values for each class (“Slow”, “Fast”, and NULL) are shown. This shows which genes are important in the prediction of clearance rate.

Table 10. Top 10 training iterations of the clearance rate model search.
 Note that the top performing model (VotingEnsemble) is the clearance rate model discussed in this paper.

Iteration	Preprocessor	Algorithm	Weighted AUC
98		VotingEnsemble	0.870471056
99		StackEnsemble	0.865215516
65	StandardScalerWrapper	LogisticRegression	0.86062304
33	StandardScalerWrapper	LogisticRegression	0.859881677
97	StandardScalerWrapper	LogisticRegression	0.858791006
44	StandardScalerWrapper	LogisticRegression	0.856105491
73	StandardScalerWrapper	LogisticRegression	0.855502817
17	RobustScaler	SVM	0.855452622
43	StandardScalerWrapper	LogisticRegression	0.855368394
61	RobustScaler	LogisticRegression	0.854357599

Table 11. Model metrics of the final clearance rate ensemble model.

Metric	Accuracy
f1_score_macro	0.6084
AUC_micro	0.9445
AUC_macro	0.8475
recall_score_micro	0.8101
recall_score_weighted	0.8101
average_precision_score_weighted	0.8707
weighted_accuracy	0.8585
precision_score_macro	0.6217
precision_score_micro	0.8101
balanced_accuracy	0.6027
log_loss	0.4455
recall_score_macro	0.6027
precision_score_weighted	0.8
AUC_weighted	0.8705
average_precision_score_micro	0.8911
f1_score_weighted	0.8019
f1_score_micro	0.8101
norm_macro_recall	0.354
average_precision_score_macro	0.7344
accuracy	0.8101

Table 12. Confusion matrix of clearance rate predictions versus actual.

Class		Prediction		
		Fast (ID: 0)	Slow (ID: 1)	Null (ID: 2)
Actual	Fast (ID: 0)	661	74	0
	Slow (ID: 1)	115	184	0
	Null (ID: 2)	6	3	0

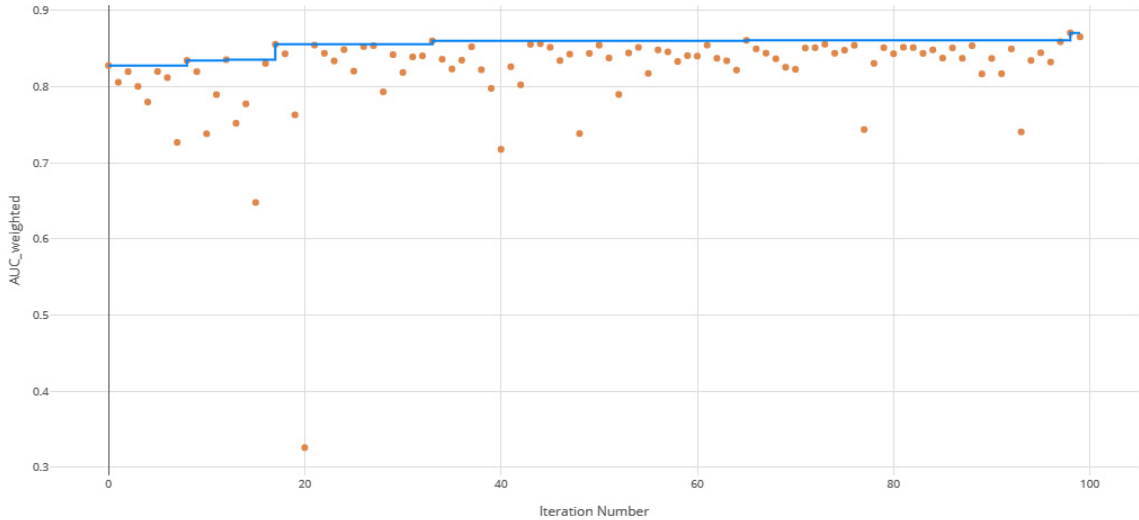


Figure 3. Area under the receiver operating characteristic curve (AUC) by iteration of the clearance rate model. Each orange dot is an iteration with the blue line representing the maximum AUC up to that iteration.

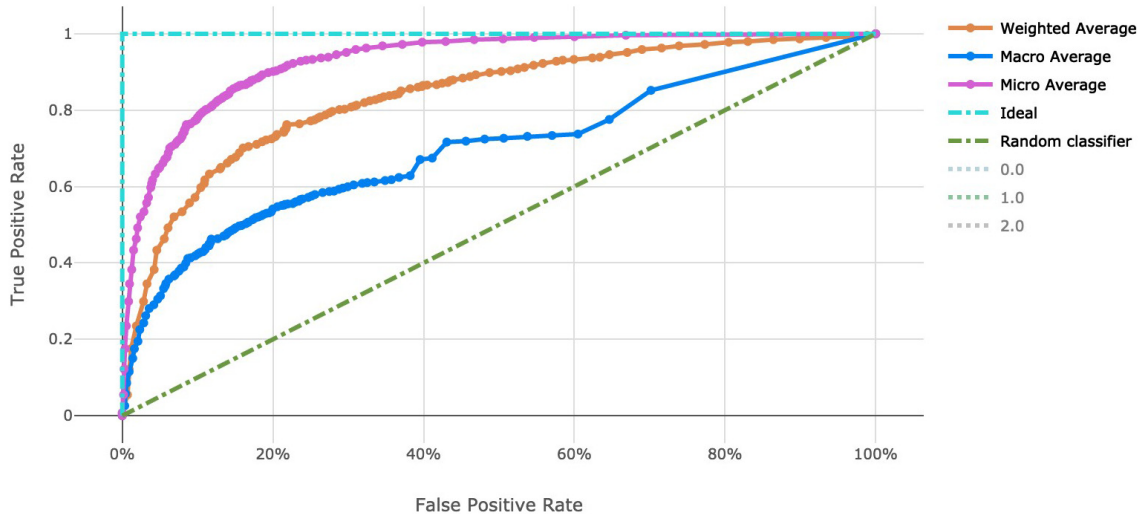


Figure 4. Receiver operating characteristic curve of the clearance rate model.

The mimic explainer was opted over other traditional methods such as principal component analysis (PCA) because of its ability to provide clearer interpretations into the features’ importance. PCA occludes the true values of individual features by summarising multiple features together. Given that insights into particular genes’ importance on resistance were desired here, the mimic explainer provides this output in a more straightforward manner.

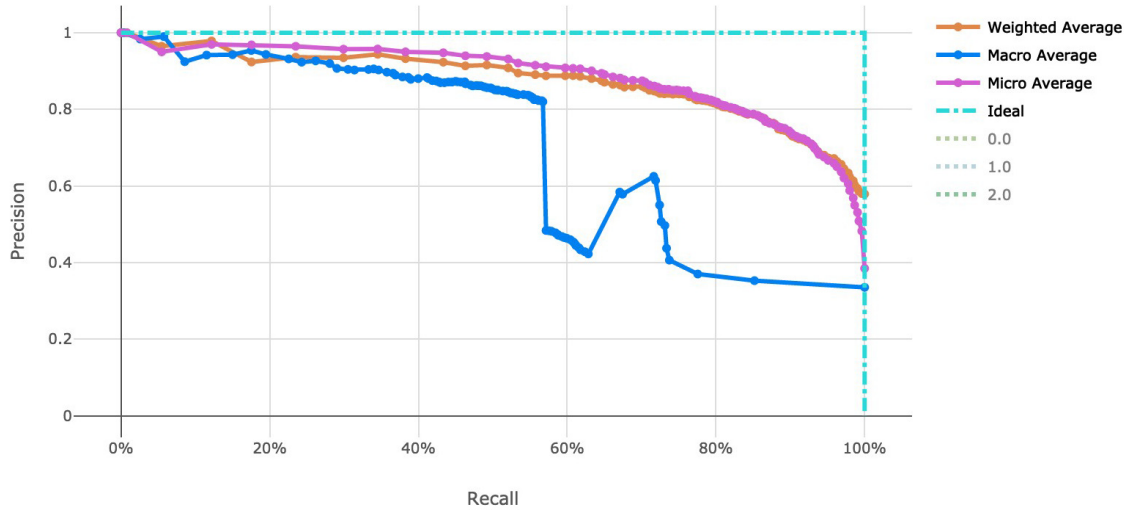


Figure 5. Precision-Recall curve of the clearance rate model.

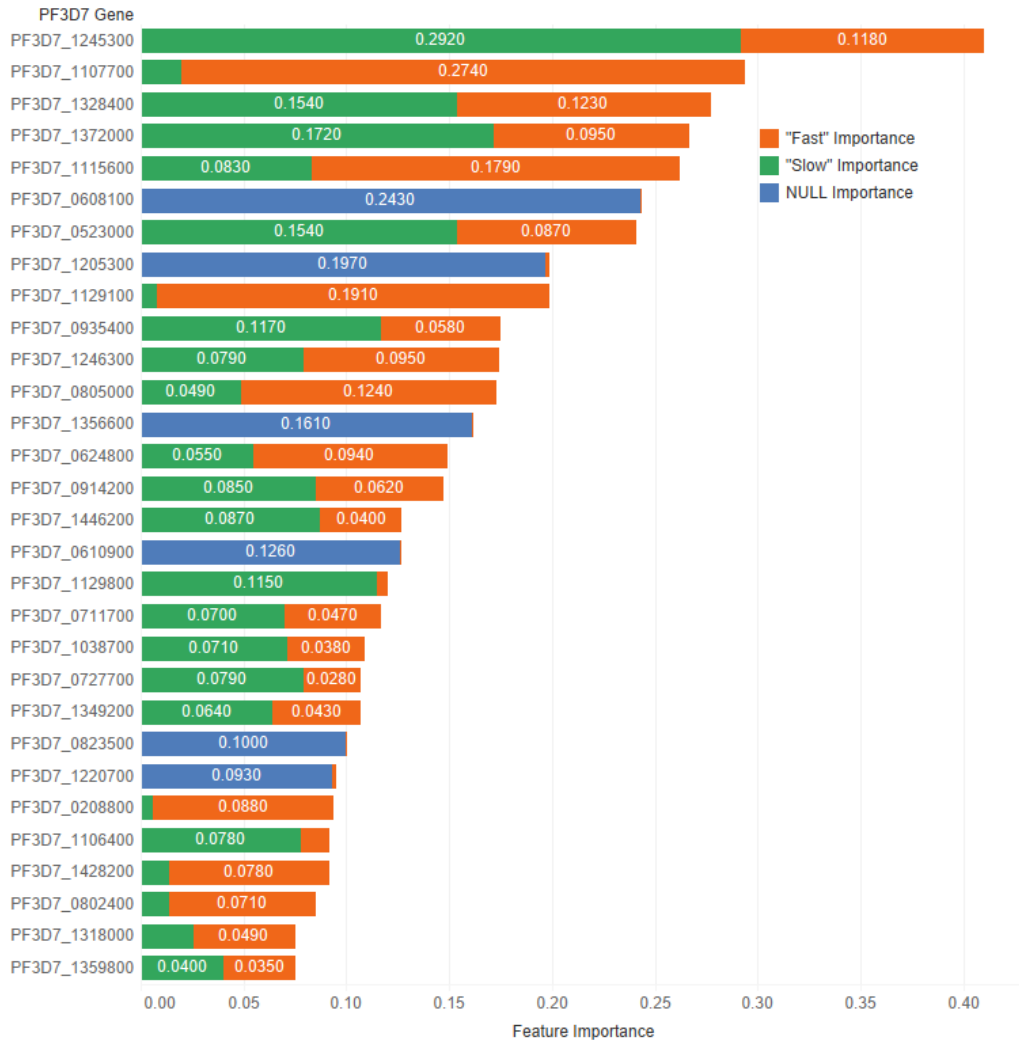


Figure 6. Derived feature importances using the black box mimic model explanation of the clearance rate model. (Shown: Top 30 genes.)

Table 13. Top 10 PF3D7 genes (features) in predicting clearance rate.

Rank	PF3D7 Gene	Slow Importance	Fast Importance	NULL Importance	Overall Importance
1	PF3D7_1245300	0.292	0.118	0.000	0.410
2	PF3D7_1107700	0.020	0.274	0.000	0.294
3	PF3D7_1328400	0.154	0.123	0.000	0.277
4	PF3D7_1372000	0.172	0.095	0.000	0.267
5	PF3D7_1115600	0.083	0.179	0.000	0.262
6	PF3D7_0608100	0.000	0.000	0.243	0.243
7	PF3D7_0523000	0.154	0.087	0.000	0.241
8	PF3D7_1205300	0.000	0.002	0.197	0.199
9	PF3D7_1129100	0.008	0.191	0.000	0.199

Table 14. Scaling function information for machine learning model search³¹.

Scaling and Normalization	Description
StandardScaleWrapper	Standardize features by removing the mean and scaling to unit variance
MinMaxScaler	Transforms features by scaling each feature by that column's minimum and maximum
MaxAbsScaler	Scale each feature by its maximum absolute value
RobustScaler	This Scaler features by their quantile range
PCA	Linear dimensionality reduction using singular value decomposition of the data to project it to a lower dimensional space
TruncatedSVDWrapper	This transformer performs linear dimensionality reduction by means of truncated singular value decomposition. Contrary to PCA, this estimator does not center the data before computing the singular value decomposition. This means it can efficiently work with sparse matrices.
SparseNormalizer	Each sample (each record of the data) with at least one non-zero component is re-scaled independently of other samples so that its norm (L1 or L2) equals one

Discussion

By using distributed processing of the data preparation, we can successfully shape and manage large malaria datasets. We efficiently transformed a matrix of over 40,000 genetic attributes for the IC_{50} use case and over 4,000 genetic attributes for the resistance rate use case. This was completed with scalable vectorization of the training data, which allowed for many machine learning models to be generated. By tracking the individual performance results of each machine learning model, we can determine which model is most useful. In addition, ensemble modeling of the various singular models proved effective for both tasks in this work. While the number of training observations for each use case stand to be improved, the usage of adequate cross-validation can help to stabilize the risk of over fitting models to such a small dataset. Also note that there is an imbalance in the number of samples in each class in the clearance rate experiment, which stands to be remedied in future work. There are over double the number of “Fast” clearance rate isolates compared to “Slow”. This can be seen in the variation in model performance as indicated by the macro average Precision-Recall curve (Figure 5).

The resulting model performance of both the IC_{50} model and the clearance rate model show relatively adequate fitting of the data for their respective predictions. While additional model tuning may provide a lift in model performance, we have demonstrated the utility of ensemble modeling in these predictive use cases in malaria. In both models, we show that IC_{50} and clearance rate can be effectively predicted using transcriptomic analysis data with machine learning. By extension, this is also predicting the phenotypic result of the genetic variations among the samples as is relates to resistance.

In a broader sense for the field parasitology, this exercise helps to quantify the importance of genetic features, spotlighting potential genes that are significant in artemisinin resistance. The merit of this work showcases the utility of machine learning to assist in the understanding of the underlying genetic/transcriptomic mechanisms that affect drug performance.

Specific examples include PF3D7_1245300, the most important feature in predicting slow parasite clearance. PF3D7_1245300 is the gene that codes for the NEDD8-conjugating enzyme UBC12 (UniProt ID: Q8I4X8), a ligase used in the ubiquitin conjugating pathway. Another example, PF3D7_1107700 is the most important gene for fast clearance rate. PF3D7_1107700 (UniProt ID: Q8IIS5) is important in the regulation of the cell cycle, specifically in the maturation of ribosomal RNAs and in the formation of the large ribosomal subunit. Future *in vitro* experiments of this *in silico* work should be performed to validate these findings. While biological confirmations of these genetic factors are needed, this analysis helps to rank the most probable factors by importance, therefore reducing the *in vitro* work to be performed.

These two examples of important genes identified here along with the other may one day be the target for future drugs or may prove integral in the overall understanding of how resistance works in *P. falciparum*. The utility of these models will help in directing development of alternative treatments or coordination of combination therapies in resistant infections and provides an example of the usage of machine learning in the identification of important genetic feature in infectious disease research.

Preprint

An earlier version of this article can be found on bioRxiv (doi: [10.1101/856922](https://doi.org/10.1101/856922)).

Data availability

Underlying data

The challenge datasets are available from Synapse (<https://www.synapse.org/>; Synapse ID: [syn18089524](https://www.synapse.org/#!Synapse:syn18089524)). Access to the data requires registration and agreement to the conditions for use at: <https://www.synapse.org/#!Synapse:syn18089524>.

Challenge documentation, including the detailed description of the Challenge design, data description, and overall results can be found at: <https://www.synapse.org/#!Synapse:syn16924919/wiki/583955>.

Whole genome expression profiling of artemisinin-resistant Plasmodium falciparum field isolates, Accession number GSE59099: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE59099>.

Zenodo: colbyford/malaria_DREAM2019: Ensemble Machine Learning Modeling for the Prediction of Artemisinin Resistance in Malaria - Initial Code Release for Research Publication (F1000). <https://doi.org/10.5281/zenodo.3590459>³².

This project contains the following underlying data:

- /SubChallenge1/data/sc1_X_train.pkl (Pickle file of the SubChallenge 1 independent variables, pivoted by Timepoint, Treatment, and BioRep.)
- /SubChallenge1/data/sc1_y_train.pkl (Pickle file of the SubChallenge 1 dependent variable, DHA_IC50.)
- /SubChallenge2/data/sc2_X_train.pkl (Pickle file of the SubChallenge 2 independent variables.)
- /SubChallenge2/data/sc2_y_train.pkl (Pickle file of the SubChallenge 2 dependent variable, ClearanceRate.)

Data are available under the terms of the Creative Commons Zero “No rights reserved” data waiver (CC0 1.0 Public domain dedication).

Software availability

- Source code available from: https://github.com/colbyford/malaria_DREAM2019
- Archived source code at time of publication: <https://doi.org/10.5281/zenodo.3590459>³²
- License: GPL-3.0

References

1. **Fact sheet about malaria.** World Health Organization. 2019.
[Reference Source](#)
2. **Guidelines for the treatment of malaria.** World Health Organization. 2015.
[Reference Source](#)
3. Dondorp AM, Nosten F, Yi P, *et al.*: **Artemisinin resistance in *Plasmodium falciparum* malaria.** *N Engl J Med.* 2009; **361**(5): 455–467.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
4. Ouattara A, Kone A, Adams M, *et al.*: **Polymorphisms in the K13-propeller gene in artemisinin-susceptible *Plasmodium falciparum* parasites from Bougoula-Hameau and Bandiagara, Mali.** *Am J Trop Med Hyg.* 2015; **92**(6): 1202–1206.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
5. Birnbaum J, Scharf S, Schmidt S, *et al.*: **A kelch13-defined endocytosis pathway mediates artemisinin resistance in malaria parasites.** *Science.* 2020; **367**(6473): 51–59.
[PubMed Abstract](#) | [Publisher Full Text](#)
6. Saralamba S, Pan-Ngum W, Maude RJ, *et al.*: **Intrahost modeling of artemisinin resistance in *Plasmodium falciparum*.** *Proc Natl Acad Sci U S A.* 2011; **108**(1): 397–402.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
7. White NJ: **The parasite clearance curve.** In: *Malar J.* 2011; **10**: 278.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
8. Ashley EA, Dhorda M, Fairhurst RM, *et al.*: **Spread of artemisinin resistance in *Plasmodium falciparum* malaria.** *N Engl J Med.* 2014; **371**(5): 411–423.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
9. Kell DB, Darby RM, Draper J: **Genomic computing. Explanatory analysis of plant expression profiling data using machine learning.** *Plant Physiol.* 2001; **126**(3): 943–951.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
10. Clare A: **Machine learning and data mining for yeast functional genomics.** PhD thesis, University of Wales, Aberystwyth, 2003.
[Reference Source](#)
11. Lee S, Kerns S, Ostrer H, *et al.*: **Machine Learning on a Genome-wide Association Study to Predict Late Genitourinary Toxicity After Prostate Radiation Therapy.** *Int J Radiat Oncol Biol Phys.* 2018; **101**(1): 128–135.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
12. Grapov D, Fahrman J, Wanichthanarak K, *et al.*: **Rise of deep learning for genomic, proteomic, and metabolomic data integration in precision medicine.** *OMICS.* 2018; **22**(10): 630–636.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
13. Ekins S, de Siqueira-Neto JL, McCall LI, *et al.*: **Machine Learning Models and Pathway Genome Data Base for *Trypanosoma cruzi* Drug Discovery.** *PLoS Negl Trop Dis.* 2015; **9**(6): e0003878.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
14. Kuang R, Gu J, Cai H, *et al.*: **Improved prediction of malaria degradomes by supervised learning with svm and profile kernel.** *Genetica.* 2009; **136**(1): 189–209.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
15. Das DK, Ghosh M, Pal M, *et al.*: **Machine learning approach for automated screening of malaria parasite using light microscopic images.** *Micron.* 2013; **45**: 97–106.
[PubMed Abstract](#) | [Publisher Full Text](#)
16. Liang Z, Powell A, Ersoy I, *et al.*: **Cnn-based image analysis for malaria diagnosis.** In: *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM).* IEEE, 2016; 493–496.
[Publisher Full Text](#)
17. Poostchi M, Silamut K, Maude RJ, *et al.*: **Image analysis and machine learning for detecting malaria.** *Transl Res.* 2018; **194**: 36–55.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
18. Davis S, Button-Simons K, Bensellak T, *et al.*: **Leveraging crowdsourcing to accelerate global health solutions.** *Nat Biotechnol.* 2019; **37**(8): 848–850.
[PubMed Abstract](#) | [Publisher Full Text](#)
19. Ghouila A, Siwo GH, Entfellner JD, *et al.*: **Hackathons as a means of accelerating scientific discoveries and knowledge transfer.** *Genome Res.* 2018; **28**(5): 759–765.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
20. Zaharia M, Xin RS, Wendell P, *et al.*: **Apache spark: A unified engine for big data processing.** *Commun ACM.* 2016; **59**(11): 56–65.
[Publisher Full Text](#)
21. van der Walt S, Colbert SC, Varoquaux G: **The numpy array: A structure for efficient numerical computation.** *Comput Sci Eng.* 2011; **13**(2): 22–30.
[Publisher Full Text](#)
22. Turnbull LB, Siwo GH, Button-Simons KA, *et al.*: **Simultaneous genome-wide gene expression and transcript isoform profiling in the human malaria parasite.** *PLoS One.* 2017; **12**(11): e0187595.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
23. **Microsoft Azure Machine Learning Service.** 2019.
[Reference Source](#)
24. Microsoft: **Azure Machine Learning AutoML Core version 1.0.79.** 2019.
[Reference Source](#)
25. Pedregosa F, Varoquaux G, Gramfort A, *et al.*: **Scikit-learn: Machine learning in Python.** *J Mach Learn Res.* 2011; **12**: 2825–2830.
[Reference Source](#)
26. Ke G, Meng Q, Finley T, *et al.*: **Lightgbm: A highly efficient gradient boosting decision tree.** In: I. Guyon, U. V. Luxburg, S. Bengio, *et al.* editors, *Advances in Neural Information Processing Systems.* Curran Associates, Inc. 2017; **30**: 3146–3154.
[Reference Source](#)
27. Caruana R, Niculescu-Mizil A, Crew G, *et al.*: **Ensemble selection from libraries of models.** In: *Proceedings of the Twenty-first International Conference on Machine Learning, ICML '04,* New York, NY, USA, 2004; 18.
[Publisher Full Text](#)
28. Mok S, Ashley EA, Ferreira PE, *et al.*: **Drug resistance. Population transcriptomics of human malaria parasites reveals the mechanism of artemisinin resistance.** *Science.* 2015; **347**(6220): 431–435.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
29. Lundberg SM, Lee S: **A unified approach to interpreting model predictions.** In: I. Guyon, U. V. Luxburg, S. Bengio, *et al.*, editors, *Advances in Neural Information Processing Systems.* Curran Associates, Inc., 2017; **30**: 4765–4774.
[Reference Source](#)
30. Molnar C: **Interpretable Machine Learning.** 2019.
[Reference Source](#)
31. Microsoft: **Microsoft Azure Machine Learning - AutoML Preprocessing.** 2019.
[Reference Source](#)
32. Ford C. colbyford/malaria_DREAM2019: **Ensemble Machine Learning Modeling for the Prediction of Artemisinin Resistance in Malaria - Initial Code Release for Research Publication (F1000).** 2019.
<http://www.doi.org/10.5281/zenodo.3590459>

Open Peer Review

Current Peer Review Status:   

Version 5

Reviewer Report 11 July 2022

<https://doi.org/10.5256/f1000research.27621.r139927>

© 2022 Barry A et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

 **Alyssa E Barry** 

IMPACT, School of Medicine, Deakin University and Burnet Institute, Melbourne, Victoria, Australia

Myo Naung

Walter and Eliza Hall Institute, University of Melbourne, Melbourne, Australia

This is commendable work by the authors making use of two publicly available datasets – the 2019 DREAM Malaria Challenge and an *in vivo* transcription data set from Mok *et al.*, (2015) to create a confident machine learning model predicting IC₅₀ (the rate at which parasites respond to artemisinin) and the transcriptional features (gene expression) involved in fast vs slow parasite clearance rates. Source codes were also made available to public for reproducibility. The manuscript would benefit from more structure in the “**methods**” and “**results**” sections to present clearer analysis workflow and results. Adding more explanation and discussion of biological significance from the generated models should improve the quality of the manuscript. There are a few specific suggestions for the authors in a revision of their manuscript below.

Major

The strongest suggestion is to re-structure the methods section. Rather than having separate sections (“method”, “data preparation”, “results”) for each machine learning exercise, I suggest merging some of these paragraphs. For instance, the entire method section describing model generation (page 3-5) to predict IC₅₀ can be possibly renamed as “machine learning method to predict IC₅₀”. Along this line, the paragraph titled “Prediction of artemisinin IC₅₀” should merge with this section. A similar arrangement is also suggested for paragraphs from page 5-7. Adding a generic workflow figure should clarify some of these issues. Both of the results sections (page 5 & 8) should come after methods.

Authors used datasets specific to *Plasmodium falciparum* malaria and thus title should reflect about *P. falciparum* malaria. The results section paragraph (page 8) should be expanded to include more explanation of results from figure-6, which can be connected to the previously known attributes of these top-30 hits as discussed in the 4th paragraph of discussion (page 13).

Minor

- Abstract 2nd paragraph– indicate ‘2019 DREAM Malaria Challenge’
- Methods typo - “This yields a total of at eight data points for each isolate.”
- Description of table-3 (page 4) such as number of rows, columns in training data can be moved to the Github page. The same suggestions for table-1 and table-7.
- A simple definition or explanation of “Caruana ensemble” selection algorithm (page 5) should be added.
- Mean Absolute Percentage Error (MAPE) of algorithms should be added to table-6 (page 6) because it is mentioned in the results section (page 5) that the voting ensemble model has chosen as the best model based on RMSE and MAPE.
- Paragraph starting with “Note that the” (page 8) should be considered as figure legends for figure 4 and 5.
- Paragraph explaining the use of mimic explainer over PCA (page 10) should be considered to move to discussion possibility merging with 4th

Is the rationale for developing the new method (or application) clearly explained?

Yes

Is the description of the method technically sound?

Yes

Are sufficient details provided to allow replication of the method development and its use by others?

Partly

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: malaria, genomics, computational biology

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Reviewer Report 10 July 2020

<https://doi.org/10.5256/f1000research.27621.r65540>

© 2020 Antani S. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Sameer K. Antani 

Communications Engineering Branch, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

The authors have updated the article but there is limited update on machine learning elements, or it is not apparent from the web-based interface. I am willing to accept the article related to prior comments, and also recognizing that the work is limited by the data from the DREAM challenge.

Is the rationale for developing the new method (or application) clearly explained?

Partly

Is the description of the method technically sound?

Partly

Are sufficient details provided to allow replication of the method development and its use by others?

Partly

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

Partly

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Partly

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 4

Reviewer Report 22 June 2020

<https://doi.org/10.5256/f1000research.26770.r63887>

© 2020 Burrows J. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Jeremy Burrows 

Medicines for Malaria Venture (MMV), Geneva, Switzerland

Page 3: Artemisinin-based therapies are described as being among the best treatment options for falciparum malaria. ACTs are the mainstay therapy and are, definitively, the best treatment options. This should be altered.

Page 3: The underlying biology of artemisinin partial resistance is becoming clearer – the authors should cite Kelch13-defined endocytosis pathway mediates artemisinin resistance in malaria parasites¹.

Page 3: In terms of predicting the IC50 of DHA on parasites, the authors really need to comment on the importance of time point and whether the parasites are synchronous or asynchronous. Usually growth inhibition assays are 48h-72h with asynchronous parasites and these result in virtually no differences in the IC50s between WT and highly resistant K13 mutant strains – indeed, this is why artemisinin partial resistance took so long to be identified. Table 1 does show the timepoint (which is good) but the synchronicity of the isolate is not mentioned. Also did the group include well characterized control lab-adapted strains (both resistant and WT)? What is the range of IC50s in the data set?

Table 1 – Abbreviations need to be described. I know what DHA is, but some readers may not. What is UT?

The computational discussion is beyond me, but I was trying to work out exactly what the authors were claiming. Is the conclusion of the first step that the IC50 can be predicted based on the transcriptomic analysis, given full genomic information of an isolate? If so, that could be interesting in predicting phenotype from genotype (in the absence of phenotypic data), but if, on the other hand, it simply confirms resistance will be evident when certain mutations are involved, then that is not so helpful as we know that already. Can the authors very clearly, in layman's terms, explain what value their model offers to the parasitology community? The same points relate to parasite clearance rate? If the model is simply telling us what we know already then that is significantly less useful or interesting than if it predicts things that we do not yet know. Some very clear explanations of the hypotheses and conclusions are needed for non-computational parasitologists to understand the merit of this work. This is not approvable without such clarity, on the assumption that other reviewers with computational expertise have approved the underlying methods.

The identification of PF3D71245300, a NEDD8-conjugating enzyme UBC12 and PF3D71107700 seem to me to be predicted genes for slow and fast clearance are the main conclusions from this work and there should be a stronger statement with respect to the need for follow-up biology to confirm these.

I would like to have seen a plot of predicted vs actual IC50 and predicted vs actual clearance rate

in a form that is easily interpretable (perhaps it's there for those in the 'know'). I was still left unclear as to how good the models were; the authors described them as 'adequate' which sounds rather underwhelming.

In short – the work may have merit, but it is not communicated in a form that makes it clear what the added value is to use the model and what the actual quality and impact is of the model.

References

1. Birnbaum J, Scharf S, Schmidt S, *et al.*: A Kelch13-defined endocytosis pathway mediates artemisinin resistance in malaria parasites. *Science*. 2020; **367** (6473): 51-59

Is the rationale for developing the new method (or application) clearly explained?

No

Is the description of the method technically sound?

Partly

Are sufficient details provided to allow replication of the method development and its use by others?

Partly

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

Partly

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Drug discovery, malaria, parasitology (not computational modelling).

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 22 Jun 2020

Colby Ford, University of North Carolina at Charlotte, USA

Thank you for your review. We have added the additional context about ACTs, and the *kelch13* gene from the Birnbaum paper. In addition, we have included information about how this work is of merit and applicable to the broader field of parasitology. We also included information in the discussion about the need for biological (*in vitro*) validation of these findings, but that this work helps to "bubble up" the most probable/important things to test first.

As for the specific questions about the data used in this study, we are still waiting on the overall DREAM Challenge write up and release to occur, which should contain much more in-depth information about the lab procedures (timepoints, test, etc.) and data collection. We are open to adding this information into our paper as well once we can get it from the DREAM Challenge.

Though only the ROC Curves are shown in the paper, all of the figures for model performance are in the GitHub repository. The plot of actual vs. predictive performance (a.k.a. calibration curve) is here:

https://github.com/colbyford/malaria_DREAM2019/blob/master/SubChallenge2/model/amls_model_7-31-19/Calibration.PNG

Competing Interests: No competing interests were disclosed.

Version 3

Reviewer Report 18 May 2020

<https://doi.org/10.5256/f1000research.25874.r62868>

© 2020 Antani S et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The author(s) is/are employees of the US Government and therefore domestic copyright protection in USA does not apply to this work. The work may be protected under the copyright laws of other jurisdictions when used in those jurisdictions.



Stefan Jaeger 

National Library of Medicine, National Institutes of Health, Bethesda, USA

Sameer K. Antani 

Communications Engineering Branch, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

The authors have addressed several, but not all of the reviewers' comments. The description of the state-of-the-art could be stronger. For example, the authors should discuss the status quo in machine learning for malaria drug-resistance detection, and the status/results of the DREAM Competition in particular, including the context of the data used. Some questions remain regarding the computation of averages in the ROC and precision-recall curves in Figures 4 and 5, see for example the bump in the latter (blue curve). The authors have not explained Figure 6, as reviewers asked them to do (labeling and legend fonts are too small). The authors also don't explain how they do the testing (size of test set, evaluation scheme etc.) I am not sure about the usefulness of Figure 2 - Sensitivity and specificity are more intuitive measures than squared errors.

Is the rationale for developing the new method (or application) clearly explained?

Partly

Is the description of the method technically sound?

Partly

Are sufficient details provided to allow replication of the method development and its use by others?

Partly

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

Partly

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: machine learning, artificial intelligence, data science, malaria screening

We confirm that we have read this submission and believe that we have an appropriate level of expertise to state that we do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 19 May 2020

Colby Ford, University of North Carolina at Charlotte, USA

We appreciate the reviewer's comments and have made some updates to the manuscript to reflect some figure quality issues and to address some points of confusion.

In this revision, we have addressed the reviewer's comments around the precision-recall curve and the ROC curve by better explaining the variation in the P-R curve and defining the micro, macro, and weighted average metrics shown in the figures. In addition, we have replaced the feature importance bar chart (Figure 6) with a higher quality version, which should be much more readable. We have also better described the model evaluation process.

Note: This article is part of a larger DREAM Challenge, from which a larger compilation manuscript will be written at a later date. As such, we cannot yet publish the data used in this work. We also cannot control the data used in this work as we were to use the data provided by the 2019 Malaria DREAM Challenge. Thus, we should not be evaluated on the lack of data, lack of a public testing dataset (which will be published in the future parent article), lack of a public evaluation scheme, the status/results of the DREAM Competition, or the full context of the data used.

Competing Interests: No competing interests.

Version 2

Reviewer Report 17 March 2020

<https://doi.org/10.5256/f1000research.24636.r60584>

© 2020 Antani S et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The author(s) is/are employees of the US Government and therefore domestic copyright protection in USA does not apply to this work. The work may be protected under the copyright laws of other jurisdictions when used in those jurisdictions.



Stefan Jaeger 

National Library of Medicine, National Institutes of Health, Bethesda, USA

Sameer K. Antani 

Communications Engineering Branch, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

The authors present a machine learning approach for detecting malaria drug-resistance based on genetic attributes. To this end, they train many different models, which they combine with known ensemble methods like voting. The detection of malaria drug resistance is an important medical problem and the application of machine learning in this context deserves further exploration. However, the paper has several shortcomings that the authors need to address:

- The author should provide a better description of the state-of-the-art and existing literature at the beginning of their paper.
- Also demonstrate the need for such an approach. It is implicitly suggested, but greater clarity is needed on what gaps this approach fills. This can be addressed through previous bullet also.
- The overall structure of the paper lacks clarity and concrete results. The authors claim that their exercise helps to “quantify the importance of genetic features, spotlighting potential genes that are significant in artemisinin resistance. The utility of these models will help in directing development of alternative treatment or coordination of combination therapies in resistant infections.” However, the experimental validation of these statements is insufficient, and the derived feature importance need to be discussed in more detail to convince the reader. In this context, Figure 6 need to be explained and discussed. What is the black block mimic model? Why has it been chosen by the authors for ranking features? In what way do other feature ranking schemes like PCA differ?

- The paper describes two experiments: a regression experiment with the IC50 value as target, and a classification experiment with three different parasite clearance rates. However, both experiments need further justification. In the first experiment, the number of rows (patterns) seems to be very small compared to the number of features (genetic attributes), which makes over-training very likely. The authors need to comment on this and address the issue if possible. In the second experiment, it is unclear how the three different clearance rates relate to drug-resistance. What is the correlation between these classes and drug-resistance? Why have the authors trained many more models for the first experiment?
- Listing of source code for formatting data is unnecessary and not suitable for a research paper. They have provided links to their code so including it in the paper seems superfluous, unless they want to make a point about it, which is absent. Further, that their example output after vectorization contains NaNs does not inspire confidence in the quality of the code; and, obviously needs further discussion.
- The authors also cite that an earlier version of this article is available on bioRxiv. They should include discussion on what improvements are in this work that substantially improve over that.

Is the rationale for developing the new method (or application) clearly explained?

Partly

Is the description of the method technically sound?

Partly

Are sufficient details provided to allow replication of the method development and its use by others?

Partly

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: machine learning, artificial intelligence, data science, malaria screening

We confirm that we have read this submission and believe that we have an appropriate level of expertise to state that we do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 09 Apr 2020

Colby Ford, University of North Carolina at Charlotte, USA

We sincerely appreciate the reviewers' feedback on this work and have improved the article based on your recommendations.

We have addressed each comment as follows in the article:

- Added additional examples of ML-based work in genomics, other tropical diseases, and in malaria.
- Added a brief explanation about the utility of this approach and its benefit over manual analysis.
- Addressed the reviewers' questions in the article around the explainability and black box methods and gave examples of the role of certain important genes identified here.
- Addressed the small observation size and the training of many models in the article and have better explained the relationship between drug resistance and parasite clearance rates. Further information on this data can be found in Mok et al., 2015. Also, the reason the second use case has fewer models trained was due to performance and risk of overfitting. However, with more observations, the machine learning modeling performance may increase with additional training and tuning time.
- The example code segments have been removed from the article.
- For the data quality concern, this is the data provided by the DREAM competition, thus isn't something we can control.
- The previous version on bioRxiv is nearly identical and was published there until the gateway was set up on F1000 and the publication embargo was lifted.

Competing Interests: No competing interests.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research