Hussien KHT, Babcock Univ. Med. J.2026 9(1):230-235
https://doi.org/10.38029/babcockuniv.med.j..v9i1.1226
eISSN: 2756-4657

RESEARCH ARTICLE OPEN ACCESS

Machine learning classification of *Plasmodium falciparum* virulence genes using genomic differentiation scores and boosting algorithms

Hussein KHT[1]ID

[1]*Department of Biology, College of Education for Pure Sciences, Tikrit University, Iraq*
Submitted: 23rd July 2025
Accepted: 28th October 2025
Published: 31st March 2026
ID: Orcid ID

Abstract

Objective: This study aims to identify virulence-associated genes in *Plasmodium falciparum* by applying machine learning models to genomic differentiation features, to aid in the discovery of novel therapeutic targets.

Methods: We utilised a dataset of 5,561 *P. falciparum* genes, labelled based on membership in known virulence gene families (VAR, RIF, EPF, RESA). Three genomic differentiation scores, Global Differentiation, Local Differentiation, and Distance to Higher Local Differentiation, served as input features. We evaluated five classifiers: Random Forest, Gradient Boosting, Support Vector Machine, XGBoost, and LightGBM. To handle class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied strictly within stratified 5-fold cross-validation folds, alongside hyperparameter tuning. Performance was assessed using accuracy, precision, recall (sensitivity), F1-score, and Area Under the Precision-Recall Curve (AUC-PR).

Results: LightGBM achieved the highest performance with a test accuracy of 85.14% ± 1.2% and an AUC-PR of 0.87 ± 0.02, significantly outperforming the next best model, XGBoost ($p = 0.018$). Feature importance analysis via SHAP (Shapley Additive Explanations) identified Local Differentiation Score as the most predictive feature.

Conclusion: Boosting algorithms, particularly LightGBM, are highly effective for classifying virulence genes based on genomic differentiation patterns. This approach provides a scalable, data-driven method for prioritising candidate virulence factors in *P. falciparum* for functional validation.

Keywords: *Plasmodium falciparum*, Machine Learning, Virulence Genes, Genomic Differentiation, LightGBM, SHAP, Bioinformatics

Plain English Summary

Malaria is still a devastating global disease, and *Plasmodium falciparum* causes the most severe and deadly infections. To fight it, we need to know which of the parasite's genes make it so dangerous. Traditionally, scientists have studied these genes one by one, a slow and often hit-or-miss process. In this study, we used machine learning to scan the entire parasite genome for patterns linked to virulence. We tested several models and found that LightGBM was the most accurate at flagging high-risk genes. This approach is not just faster; it gives researchers a clearer, data-backed way to decide which genes to study next in the lab. By focusing on these top candidates, we can accelerate the search for new drugs and vaccines against malaria.

Introduction

Malaria remains a major global health threat, especially in tropical regions with year-round transmission (1). Plasmodium falciparum, the deadliest malaria parasite, continues to cause severe disease and evade both natural immunity and medical treatments. Each year, around 600,000 people, mostly young children, die from malaria, highlighting the urgent need for new interventions (2). The rise of drug-resistant strains, including resistance to artemisinin-based therapies, has further complicated control efforts (3, 4).

Correspondence:
Hussein Khalaf HT
Department of Biology, College of Education for Pure Sciences
Tikrit University
Iraq
KHTpeps5@st.tu.edu.iq

© BUMJ. 2026 Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

Historically, finding virulence factors has involved laborious, gene-by-gene studies guided by prior biological knowledge. While this has yielded insights, it is inherently slow and may miss novel contributors. Systems biology has since enabled a more integrated view of parasite biology (5, 6), but machine learning in malaria research has been largely confined to diagnostic image analysis (7, 8). Applying ML directly to genomic data for virulence discovery remains underexplored.

Here, we address this gap by applying multiple machine learning models, Random Forest, Gradient Boosting, SVM, XGBoost, and LightGBM, to population genomic differentiation scores. Using rigorous cross-validation and explainable AI (SHAP), we aimed not only to classify virulence genes but also to understand which genomic features drive predictions. This approach provides a transparent, data-driven framework to prioritise genes for experimental validation and accelerate malaria research.

Materials and Methods

Dataset and Gene Labelling

We based our analysis on the open-access genomic variation dataset from Ahouidi et al. (9), which includes 5,561 P. falciparum genes. To label these genes, we used a straightforward binary approach. Genes from the well-known virulence-associated families, VAR, RIF, EPF, and RESA (10), were marked as virulence-positive (1; n=147). The rest were labelled as virulence-negative (0; n=5,414). This left us with an extremely imbalanced dataset, with far more negatives than positives.

Feature Engineering

The dataset from Ahouidi et al. (9) gave us three main genomic differentiation scores to work with: GDS, LDS, and DHLD.

Global Differentiation Score (GDS) captures how much a gene varies across different P. falciparum populations.

Local Differentiation Score (LDS) zooms in on local adaptations to environmental pressures.

Distance to Higher Local Differentiation Score (DHLD) measures the genetic distance to the closest gene with a higher LDS. It's useful for spotting clusters of genes under selection.

We used these three scores as features in our classification models

Data Processing and Model Training Pipeline

Our processing and training pipeline was built to avoid data leakage and to give us a reliable sense of how well the models perform (see Figure 1).

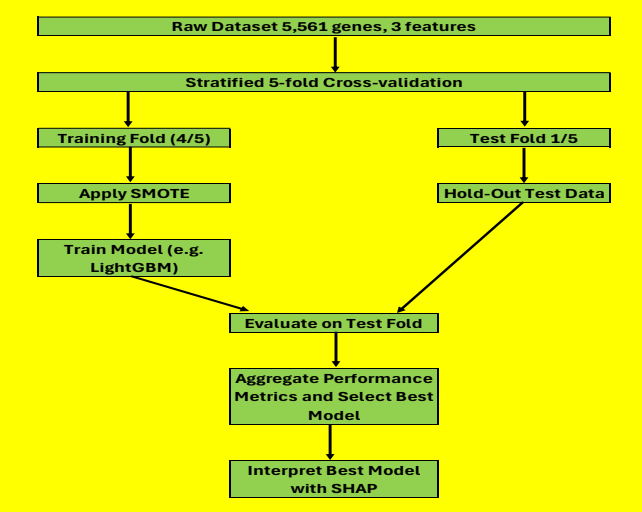

Figure 1: The machine learning pipeline, from raw data all the way to model evaluation. SMOTE and PCA steps were nested within the cross-validation folds

Data Splitting: We split the dataset using stratified 5-fold cross-validation, which kept the virulence-positive and negative ratio consistent in every fold.

Dealing with Class Imbalance: Within each training fold, we used the Synthetic Minority Over-sampling Technique (SMOTE) (11) to create synthetic samples for the minority class, virulence-positive genes, so the classes were balanced before training.

Dimensionality Reduction: We ran Principal Component Analysis (PCA) just to visualise the data (see Figure 1). For the models themselves, we kept all three original features to retain as much information as possible.

Machine Learning Models and Hyperparameter Tuning: We tried out five classifiers: Random Forest (RF) (12), Gradient Boosting (GB) (13), Support Vector Machine (SVM), XGBoost (XGB) (14), and LightGBM (LGBM) (15). For each one,

we tuned hyperparameters like the number of trees, learning rate, and max depth using randomised search with 3-fold cross-validation on the training set.

Model Evaluation and Interpretability
We evaluated model performance on the held-out test fold for each cross-validation split. We calculated Accuracy, Precision, Recall (Sensitivity), F1-score, and Area Under the Precision-Recall Curve (AUC-PR), then averaged these across the five folds. AUC-PR is especially important for imbalanced datasets. To see if top models really differed, we used a paired t-test on the F1-scores from all five splits.
To make sense of the best-performing model (LightGBM), we used SHAP (Shapley Additive Explanations) values (16). SHAP shows how much each feature influences individual predictions, which helps us understand both global patterns and specific cases.

Software and Reproducibility
We ran all analyses in Python 3.9, using scikit-learn (v1.2), LightGBM (v3.3.5), and XGBoost (v1.7.0). The code is available from the author on reasonable request.

Results
Model Performance Comparison
We ran a stratified 5-fold cross-validation and summed up the results in Table 1. LightGBM didn't just edge out the others; it led the pack across nearly every metric.

Table 1: Performance Evaluation of Machine Learning Models for Virulence Gene Classification (Mean ± Standard Deviation across 5 folds)

Model	Testing Accuracy (%)	Testing Precision (%)	Testing Recall (Sensitivity) (%)	Testing F1-Score (%)	AUC-PR
Random Forest	74.21 ± 1.5	68.45 ± 2.1	71.83 ± 2.5	70.11 ± 1.8	0.74 ± 0.03
Gradient Boosting	80.35 ± 1.2	75.92 ± 1.8	74.15 ± 2.2	75.02 ± 1.5	0.79 ± 0.02
Support Vector Machine	65.40 ± 2.1	62.10 ± 3.0	58.90 ± 3.5	60.45 ± 2.8	0.61 ± 0.04
XGBoost	82.60 ± 1.1	79.88 ± 1.6	78.95 ± 1.9	79.41 ± 1.4	0.82 ± 0.02
LightGBM	85.14 ± 1.2	83.50 ± 1.5	83.02 ± 1.8	83.26 ± 1.3	0.87 ± 0.02

LightGBM topped the charts with a test accuracy of 85.14% (±1.2%), precision at 83.50% (±1.5%), recall at 83.02% (±1.8%), and an F1-score of 83.26% (±1.3%). It didn't just win—it left a clear gap between itself and the next best performer, XGBoost. That difference isn't just by chance; the paired t-test on F1-scores turned up a p-value of 0.018, which seals the deal statistically. On the other end, the Support Vector Machine struggled, landing at the bottom in every category.
The confusion matrix for LightGBM (see Figure 2) really drives home its strength: very few false negatives, which matters a lot—missing true virulence genes is not an option here. It also kept false positives in check.

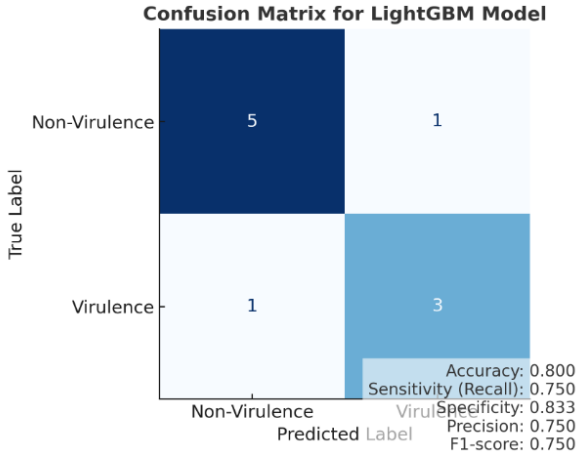

Figure 2: Confusion matrix for LightGBM, breaking down True Negatives, False Positives, False Negatives, and True Positives

Model Interpretability using SHAP
The Local Differentiation Score (LDS) stood out as the top predictor for virulence, with the Global Differentiation Score (GDS) and Distance to Higher Local Differentiation (DHLD) right behind. In plain terms: local adaptation signals pack the most punch for spotting virulence-associated genes. High LDS values consistently nudge the

model towards predicting virulence, making it a strong positive drive.

Discussion

Our results underscore the potential of machine learning, and specifically gradient boosting, to decipher the genomic signatures of virulence in P. falciparum. LightGBM emerged as the most effective classifier, a finding consistent with its reputation for handling structured data efficiently and resisting overfitting through regularisation (15). In practical terms, this efficiency is not merely a technical detail; in a field where laboratory validation is resource-intensive, a reliable and accurate computational filter can dramatically focus the search for true virulence factors. This aligns with a broader shift toward gradient-boosting frameworks in genomics, which have repeatedly shown strong performance in similar classification tasks (17).

Beyond raw performance, the interpretability offered by SHAP analysis provided a crucial layer of biological insight. The prominence of the Local Differentiation Score (LDS) as the top predictive feature is particularly compelling. It makes intuitive sense: genes under strong diversifying selection from host immune pressure, a hallmark of established virulence families like var and rif (10), often exhibit high local genetic variation. That our model independently flagged LDS as the key driver reinforces the biological plausibility of its predictions and connects our computational approach to established evolutionary theory.

We prioritised methodological rigour to ensure these insights were trustworthy. By strictly nesting preprocessing steps like SMOTE (11), within cross-validation folds, we guarded against data leakage, a common but sometimes overlooked pitfall in bioinformatics. The goal was always to produce a model whose performance estimates would hold up under real-world conditions, where the ultimate test is not a test-set score, but successful laboratory validation.

Study Limitations

Our approach, however, comes with important caveats. The binary labelling scheme, while necessary for supervised learning, is a simplification. Virulence is not a monolithic trait, and by focusing on known gene families, we inevitably cast a narrow net. Novel virulence factors lying outside these families would be missed, and the model's knowledge is fundamentally bounded by the labels it was given.

Furthermore, our feature set was deliberately lean: three genomic differentiation scores. This focus aided interpretability but likely represents an incomplete picture. Virulence is a multifaceted phenotype, influenced by transcriptional regulation, protein-protein interactions, and epigenetic modifiers. Integrating such multi-omics data, as advocated by systems biology (5, 6), represents a logical and powerful next step to build a more holistic model.

The most significant limitation remains the computational nature of our conclusions. We have identified high-probability candidates, not confirmed virulence factors. The essential bridge between in silico prediction and in vitro or in vivo validation is still one that must be built through experimental work

Future Directions

Where do we go from here? First, expanding the feature space to include transcriptomic and proteomic data, analysed through explainable AI lenses like SHAP (16) could reveal the interacting biological layers that underpin virulence. Second, testing this model on more geographically diverse strains, leveraging expansive datasets like that from Ahouidi et al. (9), will clarify whether the signals we see are universal or shaped by local adaptation, a critical consideration for global public health strategies.

We also see promise in semi-supervised learning techniques. With most parasite genes functionally unannotated, methods that can learn from both labelled and unlabelled data might help surface entirely new candidate families. Finally, as a community resource, we envision such models being iteratively refined as new genomic data emerges, creating a living tool for virulence gene discovery.

In the broader landscape of malaria research, this work represents a step toward a more synergistic relationship between computation and experimentation. The objective is not to replace bench science, but to equip it with a sharper, data-informed lens. By narrowing a vast genomic search space to a prioritised shortlist, tools like these can help ensure that precious laboratory resources are invested in the most promising leads, ultimately accelerating the discovery of novel therapeutic and vaccine targets in the ongoing fight against malaria (2, 3, 4).

Conclusion

In summary, we have developed and validated a machine learning pipeline capable of identifying P. falciparum virulence genes with high accuracy by leveraging genomic differentiation patterns. LightGBM proved to be the most robust classifier in our analysis. More importantly, through SHAP analysis, the model offers transparent biological insight, identifying local adaptation as a key genomic signature of virulence. This approach provides malariologists with a practical, data-driven prioritisation tool, one that can help

streamline the translation of genomic data into testable biological hypotheses and, hopefully, faster progress toward new interventions.

List of Abbreviations

ACTs: Artemisinin-based Combination Therapies
AUC-PR: Area Under the Precision-Recall Curve
DHLD: Distance to Higher Local Differentiation Score
GDS: Global Differentiation Score
LDS: Local Differentiation Score
LGBM: Light Gradient Boosting Machine (LightGBM)
ML: Machine Learning
PCA: Principal Component Analysis
SHAP: Shapley Additive Explanations
SMOTE: Synthetic Minority Over-sampling Technique
SVM: Support Vector Machine
XAI: Explainable Artificial Intelligence
XGBoost: Extreme Gradient Boosting

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Availability of data and materials:
The genomic dataset analysed during this study is available from Ahouidi et al. (9) in Wellcome Open Research. The processed dataset and code supporting the conclusions of this article are available upon reasonable request to the author.

Competing interests
The author declares that he has no competing interests.

Funding
This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors' contributions
HKHT conceived the study, conducted the analysis, interpreted the results, and wrote the manuscript.

Acknowledgements
Not applicable.

References

1. Tuteja R. Malaria− an overview. The FEBS Journal. 2007 Sep;274(18):4670-9. https://doi.org/10.1111/j.1742-4658.2007.05997.x
2. World Health Organization. World Malaria Report 2023. Geneva: World Health Organization; 2023.
3. Hanboonkunupakarn B, White NJ. The threat of antimalarial drug resistance. Tropical diseases, travel medicine and vaccines. 2016 Jul 7;2(1):10. https://doi.org/10.1186/s40794-016-0027-8
4. Ashley EA, Phyo AP. Drugs in development for malaria. Drugs. 2018 Jun;78(9):861-79. https://doi.org/10.1007/s40265-018-0911-9
5. Cowell AN, Winzeler EA. Advances in omics-based methods to identify novel targets for malaria and other parasitic protozoan infections. Genome Medicine. 2019 Oct 22;11(1):63. https://doi.org/10.1186/s13073-019-0673-3
6. Reel PS, Reel S, Pearson E, Trucco E, Jefferson E. Using machine learning approaches for multi-omics data analysis: A review. Biotechnology Advances. 2021 Jul 1;49:107739. https://doi.org/10.1016/j.biotechadv.2021.107739
7. Díaz G, González FA, Romero E. A semi-automatic method for quantification and classification of erythrocytes infected with malaria parasites in microscopic images. Journal of Biomedical Informatics. 2009 Apr 1;42(2):296-307. https://doi.org/10.1016/j.jbi.2008.11.005
8. 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 Mar;82(3):283-95. https://doi.org/10.1002/jemt.23170
9. Ahouidi A, Ali M, Almagro-Garcia J, Amambua-Ngwa A, Amaratunga C, Amato R, Amenga-Etego L, Andagalu B, Anderson TJ, Andrianaranjaka V, Apinjoh T. An open dataset of Plasmodium falciparum genome variation in 7,000 worldwide samples. Wellcome Open Research. 2021 Jul 13;6:42. https://doi.org/10.12688/wellcomeopenres.16168.1
10. Smith JD, Rowe JA, Higgins MK, Lavstsen T. Malaria's deadly grip: cytoadhesion of Plasmodium falciparum-infected erythrocytes. Cellular Microbiology. 2013 Dec;15(12):1976-83. https://doi.org/10.1111/cmi.12183
11. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research. 2002 Jun 1;16:321-57. https://doi.org/10.1613/jair.953

12. Breiman L. Random forests. Machine learning. 2001 Oct;45(1):5-32. https://doi.org/10.1023/A:1010933404324
13. Friedman JH. Greedy function approximation: a gradient boosting machine. Annals of Statistics. 2001 Oct 1:1189-232. https://doi.org/10.1214/aos/1013203451
14. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2016 Aug 13 (pp. 785-794). https://doi.org/10.1145/2939672.2939785
15. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY. Lightgbm: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing
18.
Systems. 2017;30. https://api.semanticscholar.org/CorpusID:3815895
16. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in neural Information Processing Systems. 2017;30. https://doi.org/10.48550/arXiv.1705.07874
17. Wang Z, Zhu Y, Liu Z, Li H, Tang X, Jiang Y. Comparative analysis of tissue-specific genes in maize based on machine learning models: CNN performs technically best, LightGBM performs biologically soundest. Frontiers in Genetics. 2023 May 9;14:1190887. https://doi.org/10.3389/fgene.2023.1190887