Review of: "A Graphical User Interface Based on Logistic Regression Approach for Malarial Detection"

Dr Monirul Islam[1]

1 International University of Business Agriculture and Technology

Potential competing interests: No potential competing interests to declare.

Title: "A Graphical User Interface Based on Logistic Regression Approach for Malarial Detection"

This study addresses the critical global health issue of malaria, a deadly communicable disease transmitted by mosquitoes. The investigation focuses on employing machine learning (ML) techniques to predict malaria presence in individuals, providing a potentially valuable tool for early diagnosis and management of the disease. The study is particularly relevant given the high mortality rate associated with malaria and the ongoing challenge it poses to healthcare authorities worldwide.

The primary objective of this study is to compare the performance of three ML-based techniques—Logistic Regression (LR), Support Vector Machine (SVM), and Random Forest (RF)—in predicting malaria presence using a database of 350 records. The performance metrics considered include classification accuracy, precision, recall, and F-score. Additionally, the study aims to develop a graphical user interface (GUI) based on the best-performing technique to facilitate malaria detection.

Review Report

Abstract

Authors should provide more specific details about the key findings and their implications, enhancing its clarity and relevance to readers in the abstract.

Introduction

The introduction effectively underscores the global health challenge posed by malaria, highlighting its prevalence and mortality rates. However, there are some inaccuracies that need correction, and additional context is required to enhance clarity and precision.

Inaccurate Statistic: The statement "Billions of people in the world die from this disease annually" is incorrect? According to WHO, malaria causes approximately 608,000 deaths annually, not billions!

Global Threat: It's noted that almost 50% of the global population is at risk of malaria, emphasizing the widespread nature of the threat.

Research Gap:

The research gap is not clearly identified. To identify a research gap is more important for authors and for the readers. Then address the research gap accordingly, which will bolster the research work and the selection of the methodology.

Methodology:

For international readers, methodology is more important than the results, so methodology should be more understandable, concise, and simple, but innovative, which will support the originality of the work. It is very difficult to understand the originality of this paper. Authors should clearly describe the originality of the paper.

Overview of Mathematical Modeling:

The historical perspective on mathematical models for understanding malaria transmission is well presented, showing the evolution from the SIR model to more sophisticated approaches. The discussion on the reproductive number (R0) effectively explains its significance in measuring transmission intensity and its correlation with socioeconomic factors.

Include a brief summary of the dataset's features (attributes) and their types (categorical, numerical). Mention any additional preprocessing steps (e.g., normalization, standardization) if they were applied to the dataset.

Machine Learning Approaches

The introduction transitions smoothly into the relevance of machine learning (ML) models in predicting and managing malaria, mentioning various ML approaches used in previous studies and indicating a broad research interest in this area. Comparison of ML Techniques; Evaluating LR, SVM, and RF based on classification accuracy, precision, recall, and F-score using a dataset of 350 records.

Diverse ML Techniques Importance of Early Prediction: Emphasizing how early prediction models can enhance prevention and control efforts.

Detailed Explanation of Dataset: Elaborate on the dataset's features, the source of the data, and any preprocessing steps.

Clarify ML Techniques Selection: Explain why LR, SVM, and RF were chosen over other potential techniques.

Discuss GUI Features: Provide more details on the GUI's design, user interface, and functionality.

Correct Statistical Inaccuracy: Replace "Billions of people in the world die from this disease annually" with accurate figures.

Expand Literature Review: Include more recent studies and a broader range of ML techniques to provide a more comprehensive background, in which you can also justify the use of machine learning techniques in your study.

Feature Reduction Technique (FRT):

Provide more details on how the correlation threshold of 0.6 was determined. Is it based on prior studies, or was it

selected empirically?

Mention if any domain knowledge or expert opinion was used to retain certain features despite their correlation values.

Classification Algorithms:

Include a rationale for choosing these three specific algorithms. For instance, explain why these algorithms are suitable for your dataset and the problem at hand.

Support Vector Machine (SVM):

Include information on the kernel used (e.g., linear, polynomial, RBF) and the rationale behind the choice. Describe the parameters and any regularization techniques applied.

Logistic Regression (LR):

Emphasize the reasons for using logistic regression in this context, particularly its interpretability and efficiency. Mention any regularization (e.g., L1, L2) used to prevent overfitting.

Random Forest (RF):

Clarify why exactly 10 decision trees were used. Was this number chosen based on cross-validation or another method? Include information on how the performance of the RF model was validated (e.g., cross-validation, holdout set).

Evaluation Metrics:

Discuss the possibility of using additional evaluation metrics (e.g., ROC-AUC, Matthews correlation coefficient) to provide a more comprehensive assessment of model performance. Can you include this in your study?

General Comments:

The explanation of stratified sampling for dealing with class imbalance is excellent. It would be beneficial to include the specific metrics used to evaluate the models (e.g., accuracy, precision, recall, F-score) and how these metrics address the imbalance issue.

Ensure that all the steps mentioned in the methodology are systematically connected and the flow is logical. This will help in understanding how each step contributes to the overall analysis.

Overall, with the suggested additions and clarifications, it can be even more robust and informative in the methodology section and data preprocessing, feature reduction, and model selection.

Experimental Setup:

The experimental setup section of your report is critical as it describes how you plan to evaluate the performance of the machine learning models. Here is a detailed review of this section with suggestions for improvement and clarification:

Performance Metrics:

Ensure that the formulas for the performance metrics are clearly formatted. Using LaTeX or another typesetting tool can improve readability.

Briefly explain the significance of each metric in the context of your study. For instance, why is the F1-score particularly useful in cases of imbalanced datasets?

K-Fold Cross-Validation:

Explain why these specific values of k were chosen. Are they based on standard practice, or were they selected through preliminary experiments?

Discuss the potential impact of different values of k on the bias-variance trade-off. For example, lower values of k might result in higher variance, whereas higher values of k might reduce variance but increase computational cost.

Include a brief discussion on the computational resources required for running k-fold cross-validation, especially for higher values of k, given that you have 350 records.

Tools and Implementation:

- Mention the specific libraries and tools used in Python for implementing the machine learning models and performance evaluation. For instance, scikit-learn for model building and valuation, pandas and numpy for data manipulation, etc.
- If any custom scripts or additional tools were used for data preprocessing, feature selection, or visualization, mention them as well.

General Comments

Ensure that all steps and processes are logically connected. Clearly show how the performance metrics and cross-validation techniques fit into the overall experimental workflow.

Include any assumptions made during the experiments, such as the distribution of data or any constraints faced.

Overall, the experimental setup section is well-structured and covers the essential aspects needed to evaluate the machine learning models. By incorporating the suggested clarifications and additional

Results and Discussion:

The results and discussion section is crucial as it presents the findings of your study and interprets the significance of these findings. Authors are requested to address the following suggestions for improvement and clarification:

Ensure that the formatting of the tables is consistent and clear. Proper alignment and spacing will enhance readability. It would be helpful to include a brief description or legend for each table to explain the metrics (e.g., CA for classification accuracy, s.d. for standard deviation).

Interpretation of Results:

Clearly highlight the best performing model in each table.

Consider providing a brief summary of the key findings from each table directly under the table for quick reference.

Detailed Analysis:

Provide a more detailed analysis of why logistic regression outperformed the other models. Discuss potential reasons such as the nature of the data, the linear separability of the classes, and the model's robustness.

Standard Deviation Insight:

Emphasize the importance of the low standard deviation in logistic regression's performance, indicating its stability.

Practical Implications:

Discuss the practical implications of these findings. For instance, how does the high accuracy and reliability of logistic regression impact its use in real-world malaria prediction?

Comparative Insights:

Provide insights into the performance of SVM and RF. Mention any observed strengths or weaknesses, and discuss potential areas for improvement.

Graphical User Interface (GUI):

Description: Include a detailed description of the GUI, explaining its functionality and how it can be used for malaria prediction.

Screenshot: Ensure that Figure 2 (the GUI screenshot) is clear and properly referenced in the text. Describe the key features visible in the screenshot.

General Comments

Flow and Coherence: Ensure a smooth flow from presenting results to discussing their implications. This helps in maintaining coherence and making the section more engaging.

Overall, the results and discussion section is well-structured and provides a detailed comparison of the performance of the models. By incorporating the suggested clarifications, detailed analysis, and visual aids, this section can be more informative and impactful.

Conclusion and Future Scope:

The conclusion and future scope section succinctly summarizes the findings of the study and outlines potential directions for further research. Here is a detailed review with suggestions for improvement and clarification:

Conclusion:

Highlight Key Findings: Emphasize the key findings more explicitly. For instance, you could state the specific performance metrics (accuracy, precision, recall, F1-score) where logistic regression outperformed the other algorithms.

Reiteration of Results:

Briefly reiterate the specific numerical results (mean accuracy, precision, recall, F1-score) for logistic regression to reinforce the conclusion.

Significance of Findings:

Discuss the practical significance of these findings. Explain how the superior performance of logistic regression can impact malaria prediction and potential clinical applications.

Future Scope:

Detailed Future Work: Provide more details on the proposed future work. For instance, specify what kind of additional data will be collected (e.g., more patient records, different regions, additional features) and how this data will enhance the study.

Exploration of Other Classifiers:

Mention specific classifiers that could be explored in future work (e.g., neural networks, gradient boosting machines, ensemble methods) and justify why they might be beneficial.

Advanced Techniques:

Consider mentioning advanced techniques that could be incorporated in future work, such as deep learning, ensemble learning, or hybrid models, and their potential advantages.

Implementation and Deployment: Mention any plans for the practical implementation and deployment of the model, such as developing a more robust and user-friendly GUI, integrating the model into healthcare systems, or conducting real-world trials. Specifically, where your results and or your techniques can be used.

Overall Comments: Major revision is needed.