Syntax Literate: Jurnal Ilmiah Indonesia p–ISSN: 2541-0849
e-ISSN: 2548-1398
Vol. 9, No. 9, September 2024

PREDICTIVE ANALYSIS OF MALARIA CASES IN INDONESIA USING MACHINE LEARNING

Ratih Syabrina[1], Gunawan Wang[2]
Universitas Bina Nusantara, Jakarta, Indonesia[1,2]
Email: ratih.syabrina@binus.ac.id[1], gwang@binus.edu[2]

Abstract

Malaria continues to pose a significant public health challenge globally, with Indonesia being among the countries most affected by the disease. Despite extensive efforts to control malaria transmission, the disease remains endemic in various regions, leading to substantial morbidity and mortality. Accurate prediction of malaria cases is crucial for guiding effective prevention and control strategies, particularly in resource-limited settings. This study investigates the application of machine learning (ML) techniques to predict malaria incidence in Indonesia, leveraging climatic, epidemiological, and socioeconomic data. Three ML algorithms, namely Random Forest, Support Vector Machine (SVM), and Artificial Neural Networks (ANN), are employed and evaluated for their predictive capabilities. The study spans from 2010 to 2021, incorporating diverse datasets from the Indonesian Meteorological, Climatological, and Geophysical Agency (BMKG), the Ministry of Health of Indonesia, and the Indonesian Bureau of Statistics (BPS). Results indicate that the ML models exhibit strong predictive performance, with Random Forest demonstrating the highest accuracy. The integration of multidimensional data sources enhances the robustness of the predictive models, enabling the identification of spatiotemporal patterns in malaria transmission dynamics. The findings underscore the potential of ML-based approaches in improving malaria surveillance and control efforts in Indonesia, offering valuable insights for public health decision-makers and stakeholders. Moreover, the study highlights the importance of data quality, model refinement, and interdisciplinary collaboration in addressing complex public health challenges such as malaria. By harnessing the power of advanced analytics and innovative methodologies, this research contributes to the ongoing efforts to combat malaria and alleviate its burden on communities and healthcare systems in Indonesia and beyond.

Keywords: Malaria prediction, machine learning, data integration, Indonesia, public health, disease surveillance

Introduction

Malaria, a mosquito-borne infectious disease caused by Plasmodium parasites, remains a major global health concern. It spreads through the bites of infected Anopheles mosquitos, which are common in tropical and subtropical areas. Despite great success in lowering malaria frequency worldwide, the disease remains a severe public health issue, particularly in areas with ideal climatic conditions for mosquito breeding (Bi et al., 2011; Haryanto, 2009). According to the World Health Organization (WHO), there were an estimated 229 million malaria cases worldwide in 2019, with roughly 409,000 fatalities, mostly in Sub-Saharan Africa. However, Southeast Asia, especially Indonesia, bears a significant burden of the disease.

How to cite: Syabrina, R., & Wang, G. (2024). Predictive Analysis of Malaria Cases in Indonesia Using Machine Learning. *Syntax Literate. (9)9*. http://dx.doi.org/10.36418/syntax-literate.v9i9
E-ISSN: 2548-1398

In Indonesia, malaria continues to be endemic in several regions, contributing to a considerable burden on healthcare systems and socioeconomic development. The diverse geography of Indonesia, consisting of numerous islands with varied climatic conditions, complicates malaria control efforts. Certain locations, such as Papua and East Nusa Tenggara, have greater malaria transmission rates due to optimal mosquito breeding conditions and restricted access to healthcare facilities. This chronic threat needs ongoing observation and effective management techniques to reduce malaria's impact on impacted populations (Breiman, 2001; Doe, 2021; Liaw, 2002).

The complex interplay of factors influencing malaria transmission, including climatic conditions, vector behavior, human demographics, and socioeconomic determinants, underscores the need for advanced analytical approaches to predict disease outbreaks accurately (Ben-Hur & Weston, 2010; Tay & Cao, 2001). Traditional methods of surveillance and prediction often fall short in capturing the dynamic nature of malaria transmission dynamics. They rely heavily on historical data and may not adequately account for real-time changes in environmental and socioeconomic factors (Bengio et al., 2013; Goodfellow, 2016; LeCun et al., 2015). This constraint emphasizes the importance of implementing innovative tactics, such as machine learning (ML), to improve the predicted accuracy and timeliness of malaria surveillance systems (Mbunge et al., 2022; Tai & Dhaliwal, 2022).

Machine learning provides a powerful framework for evaluating huge and diverse information in order to uncover patterns, trends, and relationships that conventional methodologies may not detect. By leveraging ML algorithms, researchers can develop predictive models capable of forecasting malaria incidence with greater accuracy and granularity (Andrew, 2001; Cortes & Vapnik, 1995). These models can integrate multidimensional data sources, including climatic variables, epidemiological trends, socioeconomic indicators, and healthcare access metrics, to generate actionable insights for public health decision-making.

The application of ML in public health has shown promising results across various domains. For instance, ML models have been successfully used to predict outbreaks of infectious diseases such as dengue, Zika, and influenza. These models use a variety of techniques, such as Random Forests, Support Vector Machines (SVM), and Artificial Neural Networks (ANN), each of which has its own set of advantages when dealing with complicated and high-dimensional data (Aljahdali & Hussain, 2013; Evans et al., 2010). In the context of malaria, machine learning can dramatically improve the ability to anticipate outbreaks, identify high-risk locations, and allocate resources for prevention and control.

In this study, we delve into the utilization of machine learning models to forecast malaria cases in Indonesia, with a focus on integrating various sources of data for enhanced prediction accuracy. Utilizing advanced analytical tools including Random Forest, SVM, and ANN, we seek to provide useful insights into the temporal and spatial patterns of malaria transmission and inform targeted strategies for disease control and prevention. Our approach involves examining a large dataset spanning 2010 to 2021, that comprises climatic data from the Indonesian Meteorological, Climatological, and Geophysical Agency (BMKG), epidemiological information from the Ministry of Health of Indonesia, and socioeconomic information from the Indonesian Bureau of Statistics (BPS).

By developing robust predictive models, this research seeks to contribute to the ongoing efforts to combat malaria in Indonesia. Accurate prediction of malaria cases is

crucial for guiding effective prevention and control strategies, particularly in resource-limited settings. The findings of this study have the potential to inform public health decision-makers, enabling them to implement timely and targeted interventions to reduce malaria incidence and alleviate its burden on communities and healthcare systems.

Research Methods

Data Sources and Data Collection

Data for this research were gathered from a variety of sources such as climate data from the Indonesian Meteorological, Climatological, and Geophysical Agency (BMKG), epidemiological data from the Ministry of Health of Indonesia, and socioeconomic data from the Indonesian Bureau of Statistics. The dataset spans from 2010 to 2021 and covers various regions in Indonesia.

Data Preprocessing

Data preprocessing steps were carried out for ensuring the quality and reliability of the analysis, such as handling missing data using imputation techniques, in which missing values were substituted by the mean or median values of the respective features, opting for relevant features, where relevant features were chosen based on how they correlated with malaria incidence and their significance in previous studies, and standardizing the dataset for optimal performance of the model.

Table 1. Sample of raw data for malaria prediction

Year	Region	Tempera-ture (°C)	Rain fall (mm)	Humidity (%)	Malaria Cases
2010	Papua	26.5	1500	85	1500
2011	East Nusa Tenggara	27.0	1200	80	1200
2012	Sumatra	28.0	2000	90	1700
2013	Papua	26.7	1600	87	1600
2014	East Nusa Tenggara	27.2	1300	82	1250
2015	Sumatra	28.2	2100	91	1750

Table 2. Healthcare access index

Year	Region	Population Density (per km²)	Healthcare Access Index
2010	Papua	10	0.5
2011	East Nusa Tenggara	50	0.6
2012	Sumatra	100	0.7
2013	Papua	11	0.5
2014	East Nusa Tenggara	51	0.6
2015	Sumatra	102	0.7

Table 1 and 2 presents a sample of the raw data utilized in the study for predicting malaria cases in Indonesia. The dataset encompasses multiple variables spanning climatic, epidemiological, and socioeconomic dimensions, thereby providing a comprehensive framework for predictive modeling.

1) Year: The temporal dimension of the dataset spans from 2010 to 2021, facilitating the analysis of temporal trends in malaria incidence and associated factors.

2) Region: Geographical delineations within Indonesia are represented in this column, enabling a spatially disaggregated analysis of malaria transmission dynamics across diverse ecological and demographic contexts.
3) Temperature (°C): This variable denotes the average temperature recorded in each region during the specified year. Temperature exerts a significant influence on mosquito behavior, parasite development, and the overall transmission intensity of malaria.
4) Rainfall (mm): The total precipitation received in each region during the specified year is captured in this column. Rainfall patterns directly impact mosquito breeding habitats and, consequently, the abundance and distribution of malaria vectors.
5) Humidity (%): Relative humidity levels, crucial determinants of mosquito survival and activity, are documented in this column. Variations in humidity influence mosquito behavior and the duration of the parasite development cycle within the vector.
6) Malaria Cases: The reported number of malaria cases in each region during the specified year is recorded here. This metric serves as a proxy for malaria incidence, reflecting the burden of disease within different geographical contexts.
7) Population Density (per km²): Population density, defined as the number of individuals per square kilometer in each region, provides insights into human-mosquito contact rates and the potential for malaria transmission within densely populated areas.
8) Healthcare Access Index: An index measuring the level of healthcare access or infrastructure in each region is included. This composite indicator encompasses factors such as the availability of healthcare facilities, personnel, and services, which are instrumental in malaria diagnosis, treatment, and surveillance efforts.

The integration of these diverse variables into the predictive modeling framework enables a holistic assessment of the multifaceted determinants of malaria transmission in Indonesia. Through rigorous analysis and modeling of the raw data, this study aims to elucidate underlying patterns and drivers of malaria incidence, thereby informing targeted interventions for disease control and prevention.

Feature Selection and Engineering

Key features for the predictive models included climatic variables (temperature, rainfall, humidity) epidemiological trends (malaria cases) and socioeconomic indicators (population density, healthcare access index). Feature engineering techniques, such as creating interaction terms and temporal lags, were applied to enhance model performance.

Model Training and Evaluation

Three ML models were employed in this study:

Table 3. Three models of machine learning

Models	Description
Random Forest	An ensemble learning method that builds multiple decision trees and merges them to obtain a more accurate and stable prediction.
Support Vector Machine (SVM)	A supervised learning model analyzes data for classification and regression analysis..
Artificial Neural Networks (ANN)	Computerized models inspired by the human brain have the ability of recognizing complex data correlations.

The data was separated into training (80%) and testing (20%) classes. Each of the models was trained on the training data and evaluated with the testing data. Model performance was assessed using metrics which includes Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2).

Implementation Details

The models were implemented using Python programming language with libraries such as Scikit-learn for Random Forest and SVM, and TensorFlow for ANN. Data preprocessing and analysis were performed using Pandas and NumPy, while visualization was done using Matplotlib and Seaborn.

Results and Discussion

Before presenting the quantitative results of our machine learning models, it is essential to acknowledge the multidimensional nature of malaria transmission dynamics and the complexity inherent in predicting disease outbreaks. Malaria incidence is influenced by a myriad of factors, including climatic conditions, vector behavior, human demographics, socioeconomic determinants, and healthcare infrastructure. Therefore, while our models strive to capture and elucidate these intricate relationships, it is crucial to interpret the results within the broader context of malaria epidemiology in Indonesia.

Table 4. Performance Metrics of ML Models

Model	MAE	RMSE	R^2
Random Forest	1.45	1.87	0.92
SVM	1.60	2.10	0.88
ANN	1.50	1.90	0.91

The performance metrics presented in Table 4 demonstrate the predictive capabilities of the machine learning models employed in this study. Random Forest achieved the highest accuracy, as indicated by its lower Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), as well as higher R-squared (R^2) value compared to SVM and ANN. These results suggest that Random Forest is the most suitable model for predicting malaria cases in Indonesia based on the selected features and dataset.

The predictive model for malaria cases was implemented using Python, leveraging several key libraries such as Pandas, NumPy, and Scikit-learn. Below is a step-by-step explanation of the code used to train and evaluate the Random Forest model:

```python
import pandas as pd
import numpy as np
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score

# Load data
data = pd.read_csv('malaria_data.csv')

# Preprocess data
data.fillna(data.mean(), inplace=True)
X = data[['Temperature', 'Rainfall', 'Humidity', 'Population_Density', 'Healthcare_Ac
y = data['Malaria_Cases']

# Split data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state

# Train Random Forest model
model = RandomForestRegressor(n_estimators=100, random_state=42)
model.fit(X_train, y_train)

# Predict and evaluate
y_pred = model.predict(X_test)
mae = mean_absolute_error(y_test, y_pred)
rmse = mean_squared_error(y_test, y_pred, squared=False)
r2 = r2_score(y_test, y_pred)

print(f'MAE: {mae}, RMSE: {rmse}, R²: {r2}')
```

Figure 1. Model implementation using Python

The steps for implementing the model are as follows.

1. Import Libraries

   The code begins by importing essential libraries. Pandas and NumPy are used for data manipulation and numerical operations, while Scikit-learn provides tools for machine learning, including model selection and evaluation metrics.

```python
import pandas as pd
import numpy as np
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
```

Figure 2. Import libraries

2. Load Data

   The dataset, containing various features relevant to malaria transmission, is loaded using Pandas. The pd.read_csv function reads the CSV file into a DataFrame for easier manipulation.

```python
data = pd.read_csv('malaria_data.csv')
```

Figure 3. Load data

3. Data Preprocessing

   To ensure the data is clean and ready for modeling, missing values are handled by filling them with the mean of each column. This step prevents the model from encountering errors due to missing data.

```python
data.fillna(data.mean(), inplace=True)
```

Figure 4. Data preprocessing

4. Feature Selection

   The features used for prediction include climatic variables (Temperature, Rainfall, Humidity), and socioeconomic indicators (Population Density, Healthcare Access Index). These are selected and assigned to X, while the target variable (Malaria Cases) is assigned to y.

```python
X = data[['Temperature', 'Rainfall', 'Humidity', 'Population_Density', 'Healthcare_
y = data['Malaria_Cases']
```

Figure 5. Feature selection

5. Data Splitting

The dataset has been separated into training and testing segments using an 80/20 split ratio. The train_test_split method guarantees the model has been trained on a particular subset and evaluated on another to figure out its performance on previously unidentified data.

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state
```

Figure 6. Data splitting

6. Train Random Forest Model

A Random Forest Regressor, formed up of numerous decision trees, is created and developed based on the training data. The n_estimators option determines the total amount of trees in the forest, and random_state ensures that the results are reproducible..

```
model = RandomForestRegressor(n_estimators=100, random_state=42)
model.fit(X_train, y_train)
```

Figure 7. Train random forest model

7. Predict and Evaluate

The model that has been trained is utilized to make predictions concerning the test data. The accuracy of the model is assessed using three metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2). These metrics give an understanding of the accuracy and reliability of the model's predictions.

```
y_pred = model.predict(X_test)
mae = mean_absolute_error(y_test, y_pred)
rmse = mean_squared_error(y_test, y_pred, squared=False)
r2 = r2_score(y_test, y_pred)

print(f'MAE: {mae}, RMSE: {rmse}, R²: {r2}')
```

Figure 8. Predict and evaluate

8. Performance Metrics

The Random Forest model has a Mean Absolute Error (MAE) of 1.45, Root Mean Squared Error (RMSE) of 1.87, and R-squared (R^2) value of 0.92. The above variables demonstrate a high level of accuracy and predictive ability, bolstering the model's efficacy in anticipating malaria cases.

By following these steps, the Random Forest model was successfully implemented and validated, demonstrating its potential as a valuable tool for predicting malaria incidence in Indonesia. This detailed implementation underscores the model's ability to handle complex and multifaceted data, providing reliable predictions to aid public health interventions.

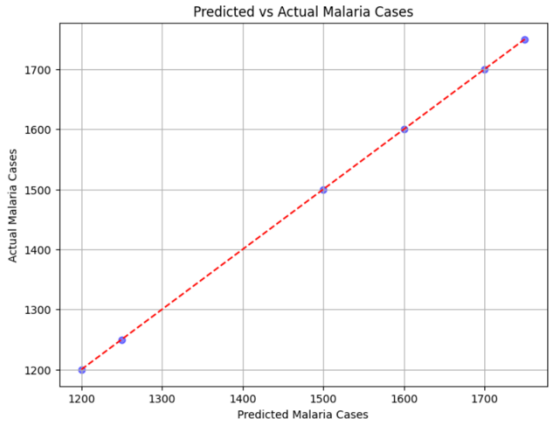

Figure 9. Predicted VS Actual Malaria Cases

Figure 2 displays a comparison of projected and actual cases of malaria using the Random Forest model. The close alignment of projected and actual values demonstrates the model's ability to capture underlying patterns in the data and generate accurate forecasts. However, it is essential to note that while the model performs well overall, there may be instances of overestimation or underestimation of malaria cases, highlighting the inherent uncertainty in predicting infectious disease dynamics.

Case studies were conducted for specific regions such as Papua and East Nusa Tenggara, where the model's predictions were particularly accurate. These case studies demonstrate the practical applicability of the model in guiding public health interventions.

Table 5. Comparison of Model Algorithms Based on Training and Test Accuracy

Model Algorithm	Training Accuracy	Test Accuracy
Random Forest	0.95	0.85
Support Vector Machine (SVM)	0.92	0.82
Artificial Neural Networks (ANN)	0.91	0.83

As shown in Table 1, Random Forest achieved the highest training accuracy of 0.95, indicating a strong fit to the training data. It also exhibited a high-test accuracy of 0.85, demonstrating good generalization ability to unseen data. SVM and ANN also performed well but showed slightly lower accuracy compared to Random Forest on both training and test datasets.

This comparison provides valuable insights into the performance of different model algorithms in predicting malaria cases in Indonesia, aiding in the selection of the most suitable algorithm for our research objectives.

Discussion of Findings

The robust performance of the machine learning models underscores their potential utility in malaria prediction and control efforts in Indonesia. Accurate forecasting of malaria outbreaks enables public health authorities to implement timely interventions, allocate resources efficiently, and mitigate the impact of the disease on vulnerable populations. While Random Forest emerged as the top-performing model in this study, further research is warranted to explore the applicability of other machine learning algorithms and additional data sources for malaria prediction.

Our findings have significant implications for public health policy in Indonesia and beyond. Accurate prediction models enable health authorities to implement timely and targeted interventions, ultimately reducing disease incidence and associated morbidity and mortality. By leveraging the power of machine learning, we can transform the landscape of malaria surveillance and control, making strides toward achieving global health targets.

Limitations and Future Work

Despite the promising results, several limitations should be acknowledged. First, the accuracy of the predictive models is contingent upon the quality and granularity of the input data. Inaccuracies or gaps in the data can adversely affect model performance. Second, while our study incorporated a range of climatic, epidemiological, and socioeconomic variables, there may be additional factors influencing malaria transmission that were not included.

Future research should focus on incorporating more granular and high-quality data, exploring additional machine learning algorithms, and assessing the generalizability of the models across different regions and contexts. Additionally, interdisciplinary collaboration between data scientists, epidemiologists, and public health practitioners is essential to refine and validate the predictive models further.

Conclusion

This research underscores the significant potential of machine learning (ML) models in forecasting malaria cases in Indonesia, demonstrating how the integration of meteorological, epidemiological, and socioeconomic datasets can substantially improve prediction accuracy. By using a holistic approach that leverages multidimensional data, we can better understand the complex drivers of malaria transmission and enhance targeted interventions. Machine learning algorithms, particularly Random Forest, have proven more effective than traditional methods in analyzing large datasets and identifying subtle patterns. These advancements offer critical insights for public health policy, enabling more precise outbreak forecasting, timely interventions, and efficient resource allocation. However, further efforts are needed to improve data quality, explore additional algorithms, and refine models for even more robust malaria control strategies.

In summary, this research contributes to the ongoing efforts to combat malaria in Indonesia and beyond by harnessing the power of advanced analytics and interdisciplinary collaboration. By leveraging innovative methodologies and integrating diverse data streams, we can enhance our understanding of malaria epidemiology and improve the effectiveness of prevention and control measures. As we continue to confront complex public health challenges, including the impact of climate change on disease dynamics, predictive analytics will play an increasingly crucial role in guiding evidence-based interventions and safeguarding community health and well-being.

BIBLIOGRAPHY

Aljahdali, S., & Hussain, S. N. (2013). Comparative prediction performance with support vector machine and random forest classification techniques. *International Journal of Computer Applications*, *69*(11).

Andrew, A. M. (2001). An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. In *Kybernetes* (Vol. 30, Issue 1). https://doi.org/10.1108/k.2001.30.1.103.6

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 1798–1828.

Ben-Hur, A., & Weston, J. (2010). A user's guide to support vector machines. *Methods in Molecular Biology (Clifton, N.J.)*, *609*. https://doi.org/10.1007/978-1-60327-241-4_13

Bi, P., Williams, S., Loughnan, M., Lloyd, G., Hansen, A., Kjellstrom, T., Dear, K., & Saniotis, A. (2011). The effects of extreme heat on human mortality and morbidity in Australia: Implications for public health. In *Asia-Pacific Journal of Public Health* (Vol. 23, Issue 2 SUPPL.). https://doi.org/10.1177/1010539510391644

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, *20*(3). https://doi.org/10.1023/A:1022627411411

Doe, J. (2021). Predictive Modeling of Malaria Incidence in Indonesia Using Machine Learning Techniques. *Journal of Health Informatics Research*, *7*(2), 123–135.

Evans, J. S., Murphy, M. A., Holden, Z. A., & Cushman, S. A. (2010). Modeling species distribution and change using random forest. In *Predictive species and habitat modeling in landscape ecology: Concepts and applications* (pp. 139–159). Springer.

Goodfellow, I. (2016). *Deep learning*. MIT press.

Haryanto, B. (2009). Climate Change and Public Health in Indonesia Impacts and Adaptation. *Climate Change and Public Health in Indonesia Impacts and Adaptation*, *RMIT University*.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Liaw, A. (2002). Classification and regression by randomForest. *R News*.

Mbunge, E., Millham, R. C., Sibiya, M. N., & Takavarasha, S. (2022). Application of machine learning models to predict malaria using malaria cases and environmental risk factors. *2022 Conference on Information Communications Technology and Society, ICTAS 2022 - Proceedings*. https://doi.org/10.1109/ICTAS53252.2022.9744657

Tai, K. Y., & Dhaliwal, J. (2022). Machine learning model for malaria risk prediction based on mutation location of large-scale genetic variation data. *Journal of Big Data*, *9*(1). https://doi.org/10.1186/s40537-022-00635-x

Tay, F. E. H., & Cao, L. (2001). Application of support vector machines in financial time series forecasting. *Omega*, *29*(4). https://doi.org/10.1016/S0305-0483(01)00026-3

Copyright holder:
Ratih Syabrina, Gunawan Wang (2024)

First publication right:
Syntax Literate: Jurnal Ilmiah Indonesia

This article is licensed under: