Case-Based Reasoning Model for Predicting the Malaria Cases

Konan N'gatta Aimé Kouassi[1], Koffi Kouakou Ive Arsene[2], Gooré Bi Tra[3]
Dept. Data Science and Artificial Intelligence Laboratory (LASDIA),
Institut National Polytechnique Felix, HOUPHOUET Boigny (INP-HB), Yamoussoukro, Côte d'Ivoire[1,3]
Dept. Computer Science Research and Telecommunication (LARIT),
Institut National Polytechnique Felix, HOUPHOUET Boigny (INP-HB), Yamoussoukro, Côte d'Ivoire[2]

***Abstract*—Malaria remains a major public health issue in Ivory Coast, where the need for accurate and interpretable predictive models is critical for effective disease control. While most existing approaches prioritize predictive accuracy over interpretability, this study addresses the need for explainable models suitable for deployment in resource-limited public health settings. The study used a dataset covering multiple regions of Côte d'Ivoire over the period 2019–2023 and including climatic variables such as temperature, humidity, and precipitation. Seven conventional machine learning models (CatBoost, XGBR, RFR, DTR, SVM, KNN, and Linear Regression) were compared with three proposed Case-Based Reasoning variants (CBR_0, CBR_1, and CBR_2), which differ in their similarity-weighting strategies and correction constant. The results show that CBR_2 achieved the best predictive performance, with RMSE = 0.081, MAE = 0.057, and $R^2 = 72.40\%$, followed by the Random Forest Regressor. A Wilcoxon signed-rank test confirmed a statistical significant difference of this permofance ($W = 18029$, $p = 1.02 \times 10^{-20}$). Beyond predictive accuracy, qualitative criteria including explainability, transparency, and adaptability were evaluated, further highlighting the superiority of CBR_2 over conventional black-box models. These findings highlight the potential of Case-Based Reasoning for epidemiological forecasting and decision support in malaria control.**

Keywords*—Malaria prediction; case-based reasoning; machine learning; epidemiological forecasting; explainable artificial intelligence

I. Introduction

Malaria remains a major public health problem, particularly in sub-Saharan Africa, where it continues to cause thousands of deaths each year despite considerable efforts in prevention and treatment [1]. According to the official WHO report dated December 4, 2025, 282 million cases of malaria were reported in 80 countries worldwide, resulting in 610,000 deaths. 95% of the cases and 95% of the deaths were recorded in the WHO African Region. This disease has disastrous consequences in Ivory Coast, where it is the leading cause of death among children under 5 and the main reason for medical consultations at health facilities across the country.

According to Dr. Tedros Adhanom Ghebreyesus, Executive Director of the WHO since 2017: "Malaria elimination is achievable in many countries, provided that sustained investments are made and efforts are stepped up to reach those most at risk". Modern tools, such as artificial intelligence models, are therefore essential to ensure better monitoring of the disease and to support effective malaria control and elimination efforts.

A major challenge is accurately predicting malaria cases to optimize the allocation of often limited medical resources and identify high-risk areas [2]–[6] Such predictions also support the development of early warning systems for timely intervention and improved disease control [7], [8]

Although several machine learning (ML) models have been proposed to predict malaria incidence [2]–[4], these approaches share common limitations. First, most of them require large volumes of training data, which are rarely available in resource-limited settings [9]. Second, and more critically, they predominantly rely on black-box architectures whose internal decision logic is opaque, making it difficult for health professionals and policy-makers to trust and act upon their outputs [10], [11]. Despite significant advances in Explainable Artificial Intelligence (XAI), a significant gap remains between the development of explainable methods and their actual applicability in operational contexts [12], most current XAI approaches rely on post-hoc techniques (e.g., SHAP) that explain predictions after the fact without guaranteeing that these explanations accurately reflect the model's internal workings [10]. Crucially, no existing study in the literature has incorporated an inherently explainable model such as Case-Based Reasoning (CBR) to predict the number of malaria cases in a given locality.

Furthermore, according to Caro-Martinez et al. [12], in healthcare it is crucial to understand how an artificial intelligence model generates its predictions, as this is essential for ensuring its reliable and trustworthy use. It not only helps build user confidence but also facilitates scientific validation, error detection, and continuous improvement of the model. In addition, Pradeep et al. [10] show that in many sensitive fields such as healthcare, finance, and the justice system, automated decisions must be transparently justified in order to meet ethical and regulatory requirements and facilitate their integration into decision-support systems. Thus, there is a lack of inherently explainable models capable of balancing predictive performance with transparency [10].

With this in mind, this study focuses on Case-Based Reasoning (CBR) as a natively explainable paradigm for malaria case prediction, examining both its theoretical foundations and its practical applicability in a real-world epidemiological context.

The CBR approach appears to be a particularly relevant alternative for addressing the limitations of current XAI approaches [11]. Unlike post-hoc methods, CBR is inherently

explainable and transparent, as it relies on a reasoning mechanism based on similar past cases [11]. Each prediction is thus supported by references to situations that have already been observed, providing a natural way to answer the questions "why" and "how" a decision was made. This form of explanation, based on concrete examples, is more intuitive and accessible to non-expert users, particularly business decision-makers [11]. In short, by combining traceability, transparency, and ease of explanation, CBR remains a promising approach for designing artificial intelligence systems that are understandable, high-performing, and operational in critical fields such as healthcare [10], [13].

In practical applications, CBR offers several advantages, notably its inherent interpretability: each prediction is justified by reference to previous cases, making the model intuitive for non-technical users [13]. Furthermore, thanks to its adaptability, the system can be expanded by adding new cases to the knowledge base without requiring complete retraining [9]. Finally, CBR performs well with small datasets, unlike ML models that require large amounts of data [9]. Despite these advantages, no existing study in the literature incorporates a CBR approach to predict the number of malaria cases in a given locality.

However, several challenges and limitations associated with the CBR approach must be acknowledged. In particular, prediction accuracy depends heavily on how the case base is organized [11], [14]. Furthermore, determining an appropriate similarity metric for heterogeneous data (numeric, categorical, and temporal) remains a challenge [11], [15]. Additionally, traditional CBR does not always account for how case characteristics evolve over time [16], [17].

The objective of this study is therefore to propose an enhanced CBR-based approach that is explainable, transparent, and effective to improve the prediction of malaria case counts at the locality level in Côte d'Ivoire, thereby addressing both the predictive limitations of classical CBR and the opacity of existing ML-based forecasting models.

In the remainder of this study, Section II presents a summary of previous work on predicting the number of malaria cases. Section III outlines the classical CBR approach. Section IV describe the proposed CBR approach. Section V presents the experimental comparison with classical CBR and ML models using data from Côte d'Ivoire. The results are discussed in Section VI. Section VII concludes the study with a summary and outlook for future work.

II. Related Work

This section provides background on malaria case prediction using data analysis and machine learning approaches, and introduces the classical CBR approach.

A. Statistical Models

Classical statistical approaches have been widely applied to malaria time series forecasting. Linear regression and ARIMA models have been used to model temporal trends in case counts [18]–[20], while their seasonal extensions (SARIMA) have been adopted to capture recurrent epidemic cycles [21]. These models have demonstrated their ability to reflect seasonal patterns and weather-driven fluctuations in malaria incidence [3], [22].

However, statistical models suffer from several structural limitations. First, they rely on restrictive assumptions linearity, stationarity, and variable independence that are rarely met in real epidemiological data [23]. Second, and more critically, they are inherently unable to capture the nonlinear relationships and complex interactions among the climatic, environmental, and socioeconomic determinants of malaria transmission [22], [24], [25]. These shortcomings motivated the shift toward machine learning approaches, discussed in the following section.

B. Machine Learning Approaches

Several ML models have been developed to predict malaria case counts, with the aim of anticipating disease progression and facilitating more efficient allocation of medical resources [2]. The studies reviewed in this section are organized around three main axes: 1) the types of ML models employed, 2) the input features and data sources used, and 3) the limitations identified.

1) Machine learning models for malaria prediction: A wide range of ML algorithms has been applied to malaria case prediction. Ensemble methods have received considerable attention: Random Forest models have been used by [18] and [26] to capture non-linear relationships between environmental variables and malaria incidence, achieving competitive accuracy across different geographic settings. Gradient Boosting methods, including XGBoost and LightGBM, have been explored by [27], who demonstrated their superiority over single-tree models in terms of predictive performance on imbalanced epidemiological datasets. Kernel-based methods such as Support Vector Machines (SVM) have also been applied [28], particularly in contexts where the training dataset is limited in size. More recently, deep learning architectures have emerged in related infectious disease prediction tasks: for instance, [29] proposed CONV-FAN-POX, a convolutional deep learning framework for monkeypox detection, illustrating the growing trend toward deep feature extraction in epidemiological surveillance.

Beyond individual models, ML approaches have been shown to serve broader public health objectives. In [5], [6], and [30], ML models have helped public health authorities develop dynamic budgets, identify high-risk areas, and implement early warning systems.

2) Limitations of existing studies: Despite these advances, the existing literature reveals several recurring and significant limitations.

a) High data requirements: Most ML models, particularly deep learning architectures and ensemble methods, require large volumes of labeled training data [1], [9]. This constitutes a critical bottleneck in sub-Saharan Africa, where epidemiological surveillance systems are often incomplete or insufficiently digitized [4].

b) Lack of explainability and transparency: The models used in the reviewed studies are predominantly *black-box* models that do not provide intelligible justifications for their predictions [11], [28]. This opacity is particularly problematic

in healthcare contexts, where clinicians and policy-makers need to understand and trust the reasoning behind a prediction [12], [31]. Although some studies apply post-hoc explanation techniques such as SHAP or LIME, these methods do not guarantee that the explanation faithfully reflects the model's internal logic [10] and are often difficult to interpret for non-expert users [11].

c) Poor adaptability to temporal dynamics: Most ML models reviewed are trained statically and do not incorporate mechanisms for continuous updating as new epidemiological data become available [16], limiting their reliability for operational forecasting.

d) Limited geographic transferability: Many models were developed and validated in specific geographic contexts, and their performance in different settings such as Côte d'Ivoire has rarely been assessed [2], [3].

e) Absence of inherently explainable approaches: Crucially, no study in the reviewed literature has proposed an inherently explainable model for malaria case prediction. The CBR paradigm, which is natively interpretable and well suited to small datasets, has not been explored in this context, a gap that the present study aims to fill.

C. Introduction to Case-Based Reasoning

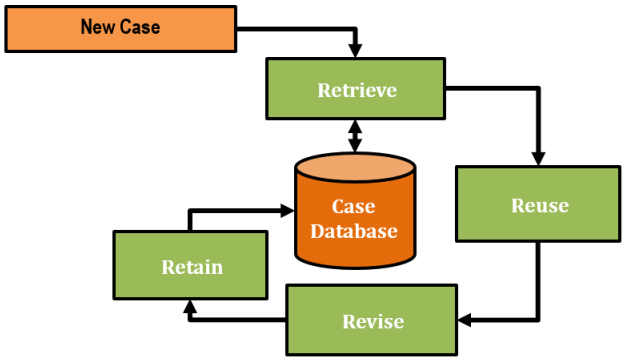

Fig. 1. The CBR life cycle.

Case-Based Reasoning (CBR) is an artificial intelligence method that solves new problems by drawing on concrete past experiences, known as cases, which contain a description of the problem, the solution applied, and the results obtained. The most similar cases are retrieved from a case base and their solutions are adapted to address the new problem. This enables intuitive and easily explainable reasoning, analogous to the way humans solve problems by drawing on past experiences. As shown in Fig. 1, the CBR life cycle comprises five sequential tasks [32]:

- Define: a new problem is described in order to compare it to past problems stored in the case base.
- Retrieve: using a similarity measure, the most similar problems in the case base are searched and selected.
- Reuse: knowledge related to the selected cases is reused to propose an initial solution.
- Revise: if the initial solution is not suitable, it is revised to make it more appropriate.
- Retain: the new problem and its associated solution are stored in the case base.

Major challenges associated with this approach include structuring the case database, selecting appropriate similarity metrics, and accounting for changes in the target variable over time (as discussed in Section I). To address these shortcomings, an improved CBR approach for predicting the number of malaria cases is proposed in the following section.

III. Proposed CBR Approach

This section presents the enhanced Case-Based Reasoning (CBR) approach proposed for predicting the number of malaria cases. The classical CBR cycle Retrieve, Reuse, Revise, Retain is extended through two key contributions: 1) a regional clustering mechanism that restricts similarity search to epidemiologically homogeneous groups, and 2) an error correction constant that compensates for temporal drift in case characteristics. The rationale for each design choice is provided alongside its mathematical formulation. Fig. 2 provides an overview of the proposed architecture.

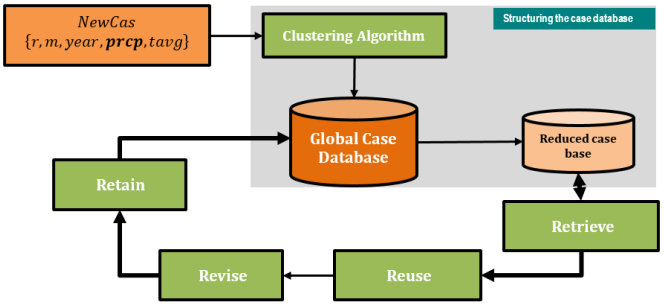

Fig. 2. Illustration of the proposed CBR approach.

A. Notation and Preliminary Definitions

To ensure clarity throughout this section, Table I summarizes the main symbols used in the proposed model.

TABLE I. Summary of Mathematical Notation

Symbol	Definition
Ω	Global case base (all historical cases)
ϕ	Reduced case base (regional cluster)
Cas_t	Target case (new case to predict)
Cas_p^i	i-th previous case in ϕ
att_i^t, att_i^p	Value of attribute i for target and previous case
$sim_i(\cdot,\cdot)$	Local similarity for attribute i
$Sim(\cdot,\cdot)$	Global (weighted) similarity
w_i^*	Normalized weight of attribute i
W_j	Prediction weight of the j-th neighbor
m	Number of nearest neighbors retrieved
k	Total number of attributes
n	Number of instances in Ω
cst	Temporal correction constant
$Pred(\text{Cas}_t)$	Predicted number of malaria cases

B. Case Structuring

In a CBR system, a *case* encapsulates a past experience comprising three elements: a problem description, the solution

applied, and the outcome obtained [32]. Accurate and consistent case representation is a prerequisite for reliable similarity computation and solution adaptation. Three properties are particularly critical in this context:

- Retrieval relevance and solution adaptation: Homogeneous case representation improves similarity computation and enables the retrieval of the most contextually relevant historical cases [9].
- Explainability: A clear and structured case representation makes the system's reasoning transparent and its predictions interpretable by non-expert users [9].
- Integration of domain knowledge: A well-defined structure allows epidemiological knowledge (e.g., seasonality, regional characteristics) to be embedded directly into the case representation [11].

Each past and new case is represented as a feature vector containing six attributes selected on the basis of their established influence on malaria transmission [5]:

$$\text{Cas} = \{r,\ m,\ year,\ prcp,\ tavg,\ hmdt\} \tag{1}$$

where, r denotes the geographic region, m the month, $year$ the year, $prcp$ the precipitation, $tavg$ the average temperature, and $hmdt$ the humidity. The target variable to be predicted is the number of malaria cases associated with each case. The choice of these features reflects the climatic and spatio-temporal determinants of malaria incidence: temperature and precipitation directly influence mosquito breeding cycles, humidity affects vector survival, while region and month capture spatial heterogeneity and seasonal periodicity respectively.

C. Case Base Structuring via Regional Clustering

In a naive CBR system, similarity is computed against the entire case base Ω, which introduces two problems: 1) irrelevant cases from geographically distant regions pollute the similarity ranking, and 2) local epidemiological characteristics specific to each region such as hydrography, vegetation density, and population structure are not accounted for. To address these issues, Ω is partitioned into regional clusters $\phi_1, \phi_2, \ldots, \phi_k$ using the attribute r (region) as the grouping criterion. During retrieval, only the cluster ϕ_i corresponding to the target case's region is searched, thereby reducing the search space and improving the contextual relevance of retrieved cases.

Formally, the clustering operation partitions Ω as follows:

$$\Omega = \bigcup_{i=1}^{k} \phi_i, \quad \phi_i \cap \phi_j = \emptyset \ \forall i \neq j \tag{2}$$

where, each cluster ϕ_i groups all cases belonging to region r_i. Algorithm 1 formalizes this partitioning procedure, and Fig. 3 illustrates the resulting case base structure.

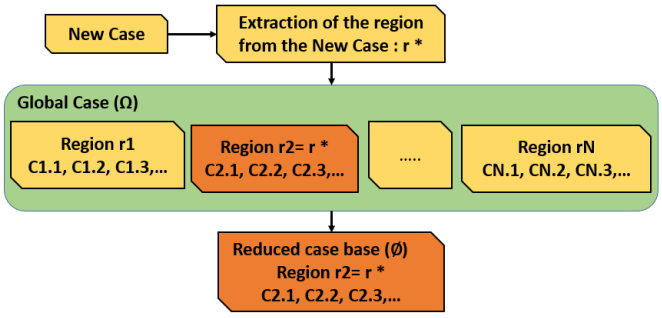

Fig. 3. Process of case base reduction via regional clustering.

Algorithm 1 Case Base Partitioning by Region

Require: $\Omega = \{(X_i, Y_i)\}_{i=1}^{N}$: Global case base
Require: $R = \{r_1, r_2, \ldots, r_p\}$: Set of regions in Ω
Ensure: Regional clusters $\phi_1, \phi_2, \ldots, \phi_k$
  Initialize an empty set of clusters $\phi \leftarrow \emptyset$
  for each region $r_i \in R$ **do**
    if cluster ϕ_k associated with r_i exists **then**
      Add case to ϕ_k
    else
      Create new cluster $\phi_k \leftarrow \{r_i\}$
    end if
  end for
  return $\phi_1, \phi_2, \ldots, \phi_k$

D. Similarity Computation

Similarity computation is the core mechanism of CBR retrieval. Given a target case Cas_t and a previous case $\text{Cas}_p^i \in \phi$, the overall similarity $Sim(\text{Cas}_t, \text{Cas}_p^i)$ is obtained through a two-step process: 1) computation of a *local similarity* for each attribute, and 2) aggregation into a *global similarity* via a weighted sum. The local similarity metric is chosen according to the attribute type categorical, numerical, or cyclic as detailed below.

1) Local similarity metrics:

a) Categorical attribute Region (r): Since region is a nominal attribute with no inherent ordering, a binary metric is used [33]. Two cases are fully similar if they belong to the same region, and fully dissimilar otherwise.

$$sim(r_t, r_p^i) = \begin{cases} 1 & \text{if } r_t = r_p^i \\ 0 & \text{if } r_t \neq r_p^i \end{cases} \tag{3}$$

b) Numerical attribute year ($year$): A normalized numerical distance is used to assign higher similarity to temporally proximate cases, reflecting the assumption that recent epidemiological patterns are more relevant than distant ones [34]:

$$sim(y_t, y_p^i) = 1 - \frac{|y_t - y_p^i|}{\max_{p \in \phi} |y_t - y_p^i|} \tag{4}$$

c) Cyclic attribute month (m): Month is a periodic variable: the distance between December ($m = 12$) and January ($m = 1$) should be 1, not 11. A circular distance is therefore adopted to account for the periodic nature of the annual cycle:

$$sim(m_t, m_p^i) = 1 - \frac{\min(|m_t - m_p^i|,\ 12 - |m_t - m_p^i|)}{6} \quad (5)$$

The denominator 6 corresponds to the maximum possible circular distance between two months, ensuring the similarity is bounded in $[0, 1]$.

d) Numerical attributes temperature ($tavg$), Precipitation ($prcp$), Humidity ($hmdt$): These three continuous attributes are normalized to $[0, 1]$ via Min-Max scaling prior to similarity computation. The normalized absolute difference directly yields a bounded similarity in $[0, 1]$ [33]:

$$sim(tavg_t, tavg_p^i) = 1 - |tavg_t - tavg_p^i| \quad (6)$$
$$sim(prcp_t, prcp_p^i) = 1 - |prcp_t - prcp_p^i| \quad (7)$$
$$sim(hmdt_t, hmdt_p^i) = 1 - |hmdt_t - hmdt_p^i| \quad (8)$$

The same functional form is applied to these three attributes because they share the same scale after normalization, and smaller absolute differences reflect greater contextual similarity for malaria transmission conditions.

2) Global similarity: The global similarity between the target case Cas_t and a previous case Cas_p is defined as the weighted sum of all local similarities:

$$Sim(\text{Cas}_t, \text{Cas}_p) = \sum_{i=1}^{k} w_i^* \times sim_i(att_i^t, att_i^p) \quad (9)$$

In the absence of domain-specific evidence prioritizing one attribute over another, equal weights are assigned to all k attributes to avoid the introduction of subjective bias:

$$w_i = 1, \quad \forall i \in \{1, 2, \ldots, k\} \quad (10)$$

To ensure that the global similarity remains in $[0, 1]$, the weights are normalized:

$$w_i^* = \frac{w_i}{\sum_{j=1}^{k} w_j} = \frac{1}{k}, \quad \forall i \in \{1, 2, \ldots, k\} \quad (11)$$

It follows that $\sum_{i=1}^{k} w_i^* = 1$, which guarantees $Sim(\text{Cas}_t, \text{Cas}_p) \in [0, 1]$ given that each $sim_i \in [0, 1]$.

E. Prediction of Malaria Cases

Once the m most similar cases $\{\text{Cas}_j^p\}_{j=1}^m$ are retrieved from ϕ, the predicted number of malaria cases is estimated as a similarity-weighted average of the target values of the retrieved neighbors:

$$Pred(\text{Cas}_t) = \sum_{j=1}^{m} W_j \times Pred(\text{Cas}_j^p) \quad (12)$$

where, the prediction weight W_j of each neighbor is proportional to its global similarity to the target case:

$$W_j = \frac{Sim(\text{Cas}_t, \text{Cas}_j^p)}{\sum_{j=1}^{m} Sim(\text{Cas}_t, \text{Cas}_j^p)} \quad (13)$$

This weighted averaging scheme assigns greater influence to the most similar historical cases, while ensuring that $\sum_{j=1}^{m} W_j = 1$, so the prediction remains within the range of observed values.

F. Temporal Correction Constant

A fundamental limitation of classical CBR is the implicit stationarity assumption: the relationship between features and the target variable is assumed to remain constant over time. In the malaria context, this assumption is violated due to evolving climatic conditions, changes in healthcare interventions, and long-term trends in case counts. To compensate for this temporal drift, a correction constant cst is introduced, estimated as the mean absolute year-on-year variation in malaria case counts across all instances in Ω:

$$cst = \frac{1}{n} \sum_{i=1}^{n} \left| \text{NbreCas}_i^l - \text{NbreCas}_i^{l+1} \right| \quad (14)$$

where, NbreCas_i^l and NbreCas_i^{l+1} denote the number of malaria cases for instance i in years l and $l + 1$ respectively. The corrected prediction is then:

$$Pred(\text{Cas}_t) = \sum_{j=1}^{m} W_j \times Pred(\text{Cas}_j^p) + cst \quad (15)$$

The constant cst captures the average magnitude of inter-annual fluctuations, providing a simple, data-driven estimate of temporal drift that requires no additional parameters. However, since absolute differences are used, the direction of the trend is not encoded: the same correction is applied regardless of whether cases are increasing or decreasing. This limitation which may cause overestimation during sustained decline or underestimation during sustained growth is acknowledged and identified as a direction for future work (see Section VI).

G. Complete Model Formulation

Substituting Eq. (9), (13), and (14) into Eq. (15), the complete prediction equation of the proposed CBR model is:

$$Pred(\text{Cas}_t) = \sum_{j=1}^{m} \left(\frac{\sum_{i=1}^{k} w_i^* \cdot sim_i(att_i^t, att_i^p)}{\sum_{j=1}^{m} Sim(\text{Cas}_t, \text{Cas}_j^p)} \right) \times Pred(\text{Cas}_j^p) + \frac{1}{n} \sum_{i=1}^{n} \left| \text{NbreCas}_i^l - \text{NbreCas}_i^{l+1} \right| \tag{16}$$

This formulation integrates three components: 1) a *similarity-weighted retrieval* from a regionally clustered case base, 2) a *cyclic similarity metric* suited to the seasonal nature of malaria data, and 3) a *temporal correction* that reduces the bias introduced by epidemiological drift over time.

H. Proposed Model Algorithm

Algorithm 2 Enhanced CBR Model for Malaria Case Prediction

Require: $\Omega = \{(X_i, Y_i)\}_{i=1}^{N}$: Global case base
Require: $Train \subset \Omega$: Training set (X_i, Y_i)
Require: $Test \subset \Omega$: Test set (X_i)
Require: $\text{Cas}_t \in Test$: Target case $\{r, m, year, prcp, tavg, hmdt\}$
Require: $w_1^*, \ldots, w_k^*$: Normalized attribute weights
Require: $\{\phi_i\}_{i=1}^{p}$: Regional clusters
Require: m: Number of nearest neighbors
Step 1: Partition Ω into regional clusters $\phi_1, \ldots, \phi_k$ (Algorithm 1)
Step 2: Identify ϕ_i such that $\text{Cas}_t.r \in \phi_i$
Step 3: Apply Min-Max scaling to all attributes in ϕ_i
Step 4: Compute $sim_i(att_i^t, att_i^p)$ for each attribute and each $\text{Cas}_p \in \phi_i$
Step 5: Compute $Sim(\text{Cas}_t, \text{Cas}_p)$ via Eq. (9)
Step 6: Select the m most similar cases $\{\text{Cas}_j^p\}_{j=1}^{m}$
Step 7: Compute $Pred(\text{Cas}_t)$ via Eq. (16)
return $Pred(\text{Cas}_t)$

The proposed algorithm (Algorithm 2) consists of seven steps : regional partitioning of the case base, selection of the relevant cluster, attribute normalization, local and global similarity computation, retrieval of the m nearest neighbors, and final prediction incorporating the temporal correction constant.

IV. Experimentation

This section describes the dataset used to validate the proposed approach, the preprocessing steps applied, and the models and evaluation metrics employed.

A. Description of the Dataset

The data were collected from two institutions. Epidemiological data were obtained from the Ministry of Health of Côte d'Ivoire, which monitors malaria cases across all national health facilities. Monthly time series were constructed for each of the 33 health regions, reporting the number of severe malaria cases recorded between 2019 and 2023. Climate variables, including temperature and precipitation, were provided by the Airport, Aviation, and Meteorological Operations and Development Company (SODEXAM).

B. Preprocessing

Categorical variables (months and regions) were encoded using the `OneHotEncoder()` function from the `scikit-learn` library. The `MinMaxScaler()` method was applied to normalize the temperature variable, and the `RobustScaler()` function was used for the precipitation variable. These normalizations and encodings are necessary to provide machine learning models with a clean dataset that promotes effective learning.

- OneHotEncoder(): This function transforms categorical variables into binary numerical variables. Each category is converted into a separate column containing only the values 0 or 1.

$$x_i = \begin{cases} 1 & \text{if the category is present} \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

- MinMaxScaler(): This method normalizes data by scaling all values into the range $[0, 1]$, preventing attributes with large numerical values from excessively influencing the model [1].

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{18}$$

where, x is the original value, $x_{\min}$ and $x_{\max}$ are the minimum and maximum values respectively, and x' is the normalized value. This normalization is particularly suited to distance- and similarity-based methods.

C. Experimental Setup

The dataset was partitioned using a single temporal split, with 80% of the data (2019–2022) used for training and the remaining 20% (2023) reserved for testing. This chronological split was deliberately chosen over k-fold cross-validation for two reasons. First, malaria incidence data exhibits strong temporal dependencies and seasonal autocorrelation; a randomized fold assignment would mix past and future observations, introducing data leakage and producing overly optimistic performance estimates. Second, the evaluation is designed to simulate a realistic forecasting scenario in which a model trained on historical epidemiological data is used to predict future case counts for an unseen period. In this setting, preserving the temporal order of observations is methodologically essential. This approach is consistent with established practice in time-series forecasting evaluation. The features used were region, year, month, temperature, and precipitation. The objective was to predict the number of malaria cases.

Several supervised learning models were selected for comparison due to their widespread use in the literature and their recognized performance in regression tasks: CatBoost, XGBoost Regressor (XGBR), Random Forest Regressor (RFR), Decision Tree Regressor (DTR), Support Vector Machine (SVM), k-Nearest Neighbors (KNN), and linear regression, as well as the three proposed CBR variants CBR_0, CBR_1, and CBR_2.

- CBR_2: This approach corresponds to the full proposed CBR method (see Section III). It incorporates the regional clustering algorithm (Algorithm 1), cyclic distance-based similarity, and the correction constant.
- CBR_1: This variant applies the same principles as CBR_2, using cyclic distance-based similarity, but without the clustering algorithm or the correction constant.
- CBR_0: This minimalist variant applies neither the clustering algorithm, the correction constant, nor the cyclic distance; numerical distance is used for the month attribute instead.

D. Evaluation Metrics

Two types of criteria were used to evaluate the models.

The quantitative criteria are:

- RMSE: the square root of the mean squared error; it heavily penalizes large errors and is sensitive to outliers [35].

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{19}$$

- MAE: the average of the absolute errors between actual and predicted values; it is easy to understand and penalizes all errors equally [36].

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{20}$$

- $\mathbf{R^2}$: measures the proportion of variance explained by the model [36].

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{21}$$

The qualitative criteria are:

- Explainability (Expl): A model's ability to provide understandable justifications for its predictions [31].
- Transparency (Transp): A model's ability to make its internal workings accessible and interpretable [31].
- Adaptability (Adapt): A model's ability to incorporate new data without requiring a full retraining process [9].

V. Results

A. Training Set Performance

Table II reports the performance of ML models on the training set. KNN achieves perfect scores (RMSE = MAE = 0.00, $R^2 = 100\%$), followed by RFR (RMSE = 0.023, MAE = 0.016, $R^2 = 96.54\%$). These results suggest overfitting, particularly for KNN, as such scores are rarely reproducible on unseen data.

B. Test Set Performance

Table III presents the comparative evaluation on the test set, combining quantitative metrics (RMSE, MAE, R^2) and qualitative criteria (Explainability, Transparency, Adaptability).

CBR-based models achieve the best generalization performance. CBR_2 records the lowest errors (RMSE = 0.081, MAE = 0.057) and the highest coefficient of determination ($R^2 = 72.40\%$), followed by CBR_1 ($R^2 = 60.04\%$) and CBR_0 ($R^2 = 57.70\%$). This progressive improvement reflects the cumulative effect of the enhancements introduced at each stage. Among ML models, RFR and KNN show acceptable performance ($R^2 > 60\%$), while CatBoost, XGBR, DTR, SVM, and linear regression perform poorly ($R^2 < 40\%$), with linear regression recording the worst results (RMSE = 0.154, $R^2 = 2.18\%$).

Only CBR models (CBR_0, CBR_1, CBR_2) satisfy all three qualitative criteria ("Yes"). All ML models are classified "No", as their use requires specialized machine learning expertise, making them unsuitable for direct deployment by health personnel in Côte d'Ivoire.

C. Overall Ranking

Table IV provides a combined ranking based on both quantitative and qualitative scores. CBR_2 achieves the lowest total score (6), ranking first overall. RFR ranks second (score = 12), despite failing all qualitative criteria. CBR_1 and KNN share third place (score = 15), while linear regression ranks last (score = 34). These results confirm that CBR_2 achieves the best balance between predictive accuracy and operational transparency. Fig. 4 - 7 illustrate these results graphically.

D. Statistical Validation

A Wilcoxon signed-rank test was conducted to assess whether the superiority of CBR_2 over RFR is statistically significant. The test yields $W = 18{,}029$ and $p = 1.02 \times 10^{-20}$, well below the 0.05 threshold, confirming that the performance advantage of CBR_2 reflects a genuine difference in generalization capacity and is not an artefact of the temporal split.

TABLE II. Performance of ML Models on the Training Set

Model	RMSE	MAE	R^2 (%)
CatBoost	0.079	0.058	59.38
XGBR	0.098	0.073	37.37
RFR	0.023	0.016	96.54
DTR	0.099	0.075	36.16
SVM	0.102	0.061	32.35
KNN	0.000	0.000	100
Reg Lin	0.118	0.084	9.83

TABLE III. PERFORMANCE OF THE VARIOUS MODELS ON THE TEST SET

Quantitative Metrics				Qualitative Criteria		
Model	RMSE	MAE	R^2 (%)	Expl	Transp	Adapt
CBR_0	0.100	0.072	57.70	Yes	Yes	Yes
CBR_1	0.097	0.069	60.04	Yes	Yes	Yes
CBR_2	0.081	0.057	72.40	Yes	Yes	Yes
CatBoost	0.119	0.084	38.42	No	No	No
XGBR	0.140	0.099	15.84	No	No	No
RFR	0.094	0.067	62.22	No	No	No
DTR	0.138	0.099	18.03	No	No	No
SVM	0.140	0.092	15.45	No	No	No
KNN	0.096	0.068	60.47	No	No	No
Reg Lin	0.154	0.108	2.18	No	No	No

TABLE IV. MODEL RANKING ON THE TEST SET

Quantitative Ranking				Qualitative Ranking				
Model	RMSE	MAE	R^2	Expl	Transp	Adapt	Total	Rank
CBR_0	5	5	5	1	1	1	18	4
CBR_1	4	4	4	1	1	1	15	3
CBR_2	1	1	1	1	1	1	6	1
CatBoost	6	6	6	2	2	2	24	5
XGBR	8	8	8	2	2	2	30	7
RFR	2	2	2	2	2	2	12	2
DTR	7	8	7	2	2	2	28	6
SVM	8	7	9	2	2	2	30	7
KNN	3	3	3	2	2	2	15	3
Reg Lin	9	9	10	2	2	2	34	8

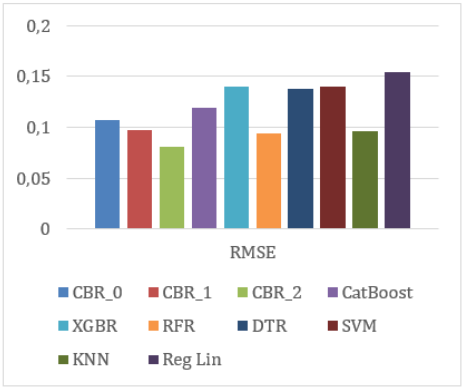

Fig. 4. Comparison of model RMSE on the test set.

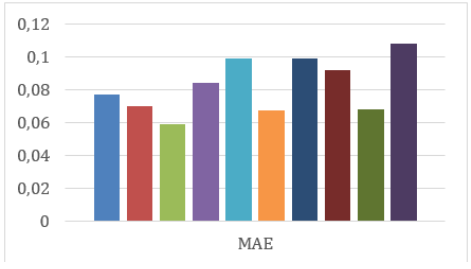

Fig. 5. Comparison of model MAE on the test set.

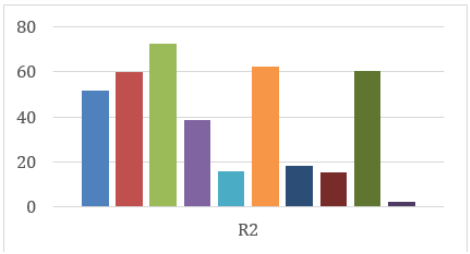

Fig. 6. Comparison of model R^2 on the test set.

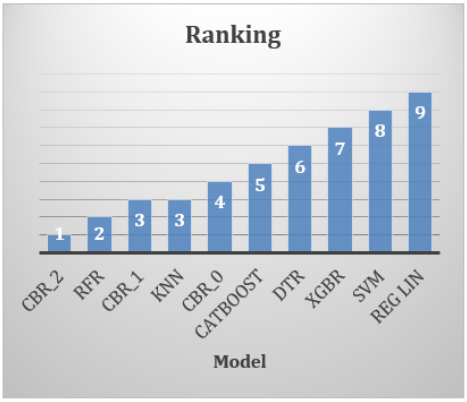

Fig. 7. Overall model ranking based on quantitative and qualitative criteria.

VI. DISCUSSION

A. Interpretation of CBR Performance

The superiority of CBR_2 over all other models results directly from three cumulative design improvements: the adoption of a cyclic similarity measure suited to seasonal epidemiological patterns, the optimization of the case base via clustering, and the introduction of a temporal correction constant. Each transition from CBR_0 to CBR_1, then to CBR_2 confirms that these choices jointly address the core challenge of malaria forecasting: capturing seasonality while maintaining generalization ability.

The significant performance drop observed for ML models between training and test sets reflects a classic overfitting phenomenon, particularly pronounced for KNN and RFR. This suggests that these models memorize training patterns rather than learn transferable relationships between climatic variables and malaria incidence, a limitation that CBR's instance-based reasoning naturally avoids.

B. The Added Value of Qualitative Criteria

Quantitative metrics alone are insufficient for evaluating models intended for operational deployment in public health. The qualitative evaluation reveals a structural divide: CBR models are natively explainable, transparent, and adaptable, whereas all ML models require specialized knowledge to interpret and maintain. In a context such as Côte d'Ivoire, where health personnel are not ML experts, this distinction is operationally decisive. A model that health workers can understand, question, and trust is more likely to influence decisions than a high-performing black box [12], [31].

C. Consistency with the Literature

The competitive performance of RFR ($R^2 = 62.22\%$) aligns with findings from Jinad et al., Lucas et al., and Monteiro et al., who identify Random Forest as one of the most effective ML models for malaria prediction [18], [26], [37]. However, the CBR_2 model proposed in this study surpasses RFR both quantitatively and qualitatively. The Wilcoxon signed-rank test ($W = 18{,}029$ and $p = 1.02 \times 10^{-20}$) confirms that this advantage is statistically significant, reflecting a genuine difference in generalization capacity.

D. Practical Implications

The CBR_2 model offers concrete value for public health decision-making. Its case-based reasoning mechanism enables health authorities to trace the historical cases that drove each prediction, facilitating early detection of high-risk periods and supporting timely preventive interventions. Its adaptability the ability to incorporate new cases without full retraining makes it particularly suited to operational surveillance settings where data are updated continuously and resources are limited.

E. Limitations of the Study

Despite these results, several limitations must be acknowledged.

1) Correction constant design: The error correction constant quantifies historical fluctuations independently of trend direction. It may therefore lead to overestimation during sustained decline periods or underestimation during sustained growth phases.

2) Temporal validation: The evaluation relies on a single 80/20 temporal split. Although the Wilcoxon test confirms statistical significance, this scheme does not fully assess the model's stability across varied epidemiological periods.

3) Geographic scope: The model was validated exclusively on data from Côte d'Ivoire. Its transferability to other sub-Saharan African contexts with different climatic profiles and data quality remains to be assessed.

4) Feature set: The current model relies on a fixed set of climatic and epidemiological variables. The potential contribution of socioeconomic or behavioral features has not been explored.

F. Future Work

These limitations suggest four concrete directions. First, the correction constant should be replaced by a trend-aware mechanism such as a directional drift estimator to improve predictions during monotonic phases. Second, a walk-forward validation framework across multiple temporal windows should be adopted to rigorously assess temporal stability. Third, the approach should be evaluated across diverse geographic contexts in sub-Saharan Africa. Fourth, hybrid CBR-ML architectures that combine the interpretability of CBR with the feature-learning capacity of ML models should be explored.

VII. Conclusion

This study proposed an enhanced Case-Based Reasoning (CBR) approach for predicting malaria cases in Côte d'Ivoire, addressing two major limitations of existing models: the opacity of machine learning black-box models and the stationarity assumption of classical CBR. The proposed model, CBR_2, incorporates three key contributions : regional clustering, cyclic similarity measurement, and a temporal correction constant which jointly improve both predictive accuracy and operational interpretability.

When evaluated against seven ML models (CatBoost, XGBR, RFR, DTR, SVM, KNN, and linear regression) and two CBR baselines (CBR_0, CBR_1), CBR_2 achieves the best overall performance on the test set (RMSE = 0.081, MAE = 0.057, $R^2 = 72.40\%$), it outperformed the strongest ML competitor, RFR ($R^2 = 62.22\%$), with statistical significance confirmed by a Wilcoxon signed-rank test ($p = 1.02 \times 10^{-20}$). Beyond quantitative metrics, CBR_2 is the only model to satisfy all three qualitative criteria explainability, transparency, and adaptability making it directly deployable by health personnel without machine learning expertise. This dual advantage places CBR_2 first in the overall ranking with a score of 6, ahead of RFR (12) and CBR_1 (15).

These results confirm that it is possible to achieve competitive predictive performance while maintaining full interpretability, a critical requirement for AI adoption in public health. They also highlight the largely untapped potential of CBR in epidemiological forecasting, a field currently dominated by black-box approaches. In practice, CBR_2 can support early warning systems, epidemic anticipation, and healthcare resource allocation in resource-limited settings.

Three directions are identified for future work, improving the temporal correction mechanism, strengthening the validation framework, and extending the approach to broader geographic contexts in sub-Saharan Africa.

Acknowledgment

This work was partially supported by the international network of french-speaking engineering universities and schools RESCIF.

We would also like to thank the Ministry of Health of Côte d'Ivoire and SODEXAM for providing the data used in this study.

References

[1] E. Naroum, E. M. Maka, H. Abboubakar, P. Dayang, A. B. Bamana, B. Garga, H. D. Daouda, M. Bakouri, and I. Khan, "Comparative analysis of deep learning and machine learning techniques for forecasting new malaria cases in cameroon's adamaoua region," *Intelligence-Based Medicine*, vol. 11, p. 100220, 2025.

[2] E. Mbunge, R. C. Millham, M. N. Sibiya, and S. Takavarasha, "Application of machine learning models to predict malaria using malaria cases and environmental risk factors," in *2022 Conference on Information Communications Technology and Society (ICTAS)*. Durban, South Africa: IEEE, March 2022, pp. 1–5.

[3] M. T. Pillay *et al.*, "Utilizing a novel high-resolution malaria dataset for climate-informed predictions with a deep learning transformer model," *Scientific Reports*, vol. 13, no. 1, p. 23091, December 2023.

[4] D. Harvey, W. Valkenburg, and A. Amara, "Predicting malaria epidemics in Burkina Faso with machine learning," *PLOS ONE*, vol. 16, no. 6, p. e0253302, June 2021.

[5] M. A. Sehlabana, D. Maposa, A. Boateng, and S. Das, "Integrating climate and environmental data with Bayesian models for malaria prediction," *Statistics, Optimization and Information Computing*, vol. 14, no. 5, pp. 2930–2956, September 2025.

[6] A. K. Verma and V. Kuppili, "Data-oriented neural time series with long short-term memories (LSTM) for malaria incidence prediction in Goa, India," in *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. Kanpur, India: IEEE, July 2019, pp. 1–6.

[7] O. Khan, J. O. Ajadi, and M. P. Hossain, "Predicting malaria outbreak in The Gambia using machine learning techniques," *PLOS ONE*, vol. 19, no. 5, p. e0299386, May 2024.

[8] A. Singh, M. Mehra, A. Kumar, M. Niranjannaik, D. Priya, and K. Gaurav, "Leveraging hybrid machine learning and data fusion for accurate mapping of malaria cases using meteorological variables in western India," *Intelligent Systems with Applications*, vol. 17, p. 200164, February 2023.

[9] D. Doyle, A. Tsymbal, and P. Cunningham, "A review of explanation and explanation in case-based reasoning," 2003.

[10] P. Pradeep, M. Caro-Martínez, and A. Wijekoon, "Empowering explainable artificial intelligence through case-based reasoning: A comprehensive exploration," *IEEE Transactions on Knowledge and Data Engineering*, vol. 37, no. 12, pp. 7120–7139, December 2025.

[11] ——, "A practical exploration of the convergence of case-based reasoning and explainable artificial intelligence," *Expert Systems with Applications*, vol. 255, p. 124733, December 2024.

[12] M. Caro-Martínez *et al.*, "iSee: A case-based reasoning platform for the design of explanation experiences," *Knowledge-Based Systems*, vol. 302, p. 112305, October 2024.

[13] J. M. Schoenborn, R. O. Weber, D. W. Aha, J. Cassens, and K.-D. Althoff, "Explainable case-based reasoning: a survey," in *AAAI-21 workshop proceedings*, 2021.

[14] N. B. Doğan, B. U. Ayhan, G. Kazar, M. Saygili, Y. E. Ayözen, and O. B. Tokdemir, "Predicting the cost outcome of construction quality problems using case-based reasoning (CBR)," *Buildings*, vol. 12, no. 11, p. 1946, November 2022.

[15] W. Li, W. K. Härdle, and S. Lessmann, "A data-driven case-based reasoning in bankruptcy prediction," November 2022.

[16] G. A. Pérez-Pérez, M. F. Valdez-Ávila, M. G. Orozco-del Castillo, C. Bermejo-Sabbagh, and J. A. Recio-García, "CBR-FoX: A case-based reasoning software tool for auditing time series predictions," *SoftwareX*, vol. 32, p. 102450, December 2025.

[17] F. Sørmo, J. Cassens, and A. Aamodt, "Explanation in case-based reasoning–perspectives and goals," *Artificial Intelligence Review*, vol. 24, no. 2, pp. 109–143, October 2005.

[18] K. H. D. C. Monteiro *et al.*, "Integrating machine learning and spatial clustering for malaria case prediction in Brazil's Legal Amazon," *BMC Infectious Diseases*, vol. 25, no. 1, p. 802, June 2025.

[19] E. K. Panzi, N. I. Kandala, E. L. Kafinga, B. M. Tampwo, and N.-B. Kandala, "Forecasting malaria morbidity to 2036 based on geo-climatic factors in the Democratic Republic of Congo," *International Journal of Environmental Research and Public Health*, vol. 19, no. 19, p. 12271, September 2022.

[20] M. Wang *et al.*, "A novel model for malaria prediction based on ensemble algorithms," *PLOS ONE*, vol. 14, no. 12, p. e0226910, December 2019.

[21] S. W. Jalloh, B. Malenje, H. Imboga, and M. H. Hodges, "Forecasting malaria cases using climate variability in Sierra Leone," *Malaria Journal*, vol. 24, no. 1, p. 158, May 2025.

[22] J. X. Zhai *et al.*, "Development of an empirical model to predict malaria outbreaks based on monthly case reports and climate variables in Hefei, China, 1990–2011," *Acta Tropica*, vol. 178, pp. 148–154, February 2018.

[23] B. O. Nyawanda *et al.*, "Forecasting malaria dynamics based on causal relations between control interventions, climatic factors, and disease incidence in western Kenya," *Journal of Global Health*, vol. 14, p. 04208, October 2024.

[24] B. Shi, S. Lin, Q. Tan, J. Cao, and A. Et, "Inference and prediction of malaria transmission dynamics using time series data," *Infectious Diseases of Poverty*, 2020.

[25] P. Taconet *et al.*, "Data-driven and interpretable machine-learning modeling to explore the fine-scale environmental determinants of malaria vectors biting rates in rural Burkina Faso," *Parasites & Vectors*, vol. 14, no. 1, p. 345, June 2021.

[26] S. Jinad, M. O. Kama, M. Baba-Adamu, A. C. Ikegwu, and O. Machete, "Analysis of malaria and climate in Damaturu City of Nigeria using predictive supervised learning," *Discover Sustainability*, vol. 6, no. 1, p. 1239, November 2025.

[27] M. S. Rahman and M. A. B. Shiddik, "Unraveling global malaria incidence and mortality using machine learning and artificial intelligence-driven spatial analysis," *Scientific Reports*, vol. 15, no. 1, p. 28334, August 2025.

[28] G. J. Gbaguidi, N. Topanou, W. L. Filho, and G. K. Ketoh, "Towards an intelligent malaria outbreak warning model based intelligent malaria outbreak warning in the northern part of Benin, West Africa," *BMC Public Health*, vol. 24, no. 1, p. 450, February 2024.

[29] Ç. N. Tülü and Y. Kaya, "Convolutional fourier analysis network (conv-fan-pox): A novel time–frequency approach for medical image analysis," *Biomedical Signal Processing and Control*, vol. 117, p. 109698, 2026.

[30] L. Zou, L. Xia, L. Hou, X. Zhao, and D. Yin, "Data-efficient reinforcement learning for malaria control," May 2021.

[31] J.-B. Lamy, B. Sekar, G. Guezennec, J. Bouaud, and B. Séroussi, "Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach," *Artificial Intelligence in Medicine*, vol. 94, pp. 42–53, March 2019.

[32] A. Aamodt and E. Plaza, "Case-based reasoning: Foundational issues, methodological variations, and system approaches," *AI communications*, vol. 7, no. 1, pp. 39–59, 1994.

[33] T. W. Liao, Z. Zhang, and C. R. Mount, "Similarity measures for retrieval in case-based reasoning systems," *Applied Artificial Intelligence*, vol. 12, no. 4, pp. 267–288, 1998.

[34] A. Sylla, T. Coudert, and L. Geneste, "A case-based reasoning (CBR) approach for engineer-to-order systems performance evaluation," *IFAC-PapersOnLine*, vol. 54, no. 1, pp. 717–722, January 2021.

[35] M. F. X. Barboza *et al.*, "Prediction of malaria using deep learning models: A case study on city clusters in the state of Amazonas, Brazil, from 2003 to 2018," *Revista da Sociedade Brasileira de Medicina Tropical*, vol. 55, pp. e0420–2021, 2022.

[36] M. T. Pillay, M. T. Q. Le, Y. Takamatsu, T. V. Phong, N. Kgalane, and N. Minakawa, "Application of a Temporal Fusion Transformer and long-term climate and disease data to assess the predictive power and understand the drivers for malaria and dengue," *International Journal of Environmental Research and Public Health*, vol. 23, no. 1, p. 75, January 2026.

[37] T. C. D. Lucas *et al.*, "Improving disaggregation models of malaria incidence by ensembling non-linear models of prevalence," *Spatial and Spatio-temporal Epidemiology*, vol. 41, p. 100357, June 2022.