Research Article
Open Access

MJ Cell Press

MJ Mathematics and Computer Science

Available at:

https://mjcellpress.com/article/mjmcs03/

MJ - MATHEMATICS AND COMPUTER SCIENCE

Predicting the spread of Chikungunya using an ensemble regression approach: A case study of Chad, Brazil and Paraguay

Mohamed El Bachir[1], Ebenezer Maka Maka[2,3], Yannick Malong[2,3], Benjamin Garga[3], Daouda Hassana Daouda[1], Hamadjam Abboubakar[3,5*].

[1] *University of Ngaoundéré, Faculty of Science, Department of Mathematics and Computer Science, P.O. Box 454, Ngaoundéré, Cameroon.*

[2] *University of Douala, ENSPD, Department of Computer Engineering and Telecommunications, P.O. Box 2701, Douala, Cameroon.*

[3] *University of Douala, National Higher Polytechnic School of Douala,Laboratory of Computer Science Data Science and Artificial Intelligence, Douala, Cameroon.*

[4] *University of Ngaoundéré, ENSAI, Department of Electrical, Electronic and Automatic Engineering, P.O. Box 455, Ngaoundéré, Cameroon.*

[5] *University of Ngaoundéré, School of Geology and Mining Engineering, P.O. Box 115, Meiganga, Cameroon.*

**Corresponding author: h.abboubakar@gmail.com*

Key words

Chikungunya,
Grid Search,
Ensemble Regression,
Random Forest,
XGBoost,
Machine Learning,
Voting Regressor.

Abstract

The Chikungunya virus, primarily transmitted by female *Aedes aegypti* and *Aedes albopictus* mosquitoes, poses a growing global public health challenge due to its debilitating symptoms and rapid spread. Recent outbreaks in Southeast Asia, South America, and Central and East Africa highlight the difficulty of accurately predicting epidemics, given the complex interactions among environmental, climatic, and biological factors. Traditional epidemiological surveillance systems often remain insufficient for early outbreak detection. This study applies advanced machine learning techniques, specifically ensemble regression, to develop predictive models of Chikungunya epidemics in Chad, Brazil, and Paraguay. Random Forest and XGBoost regressors optimized via Grid Search are combined within a Voting Regressor ensemble framework. The ensemble model demonstrated superior predictive performance, achieving lower RMSE and MAE than individual models. At the 5% significance level, no statistically significant differences were observed between the Voting Regressor and XGBoost ($p = 0.2126$ and $p = 0.2081$, respectively) or Random Forest ($p = 0.2607$ and $p = 0.2997$, respectively), as determined by both the paired t-test and the Wilcoxon signed-rank test.

Received: 19.11.2025 Accepted: 13.01.2026 Published online: 15.01.2026

How to cite this article: El Bachir, M., Maka Maka, E., Malong, Y., Garga, B., Hassana Daouda, D., & Abboubakar, H. (2026). *Predicting the spread of Chikungunya using an ensemble regression approach: A case study of Chad, Brazil and Paraguay.* **MJ Mathematics and Computer Science, 2**(1), 32-63 . https://doi.org/10.63156/mjmcs03.

1 Introduction

Chikungunya is a viral disease transmitted to humans through the bite of infected mosquitoes, mainly *Aedes albopictus* and *Aedes aegypti* [1, 2]. On August 2, 2024, the European Centre for Disease Prevention and Control (ECDC) reported approximately 350,000 confirmed cases of Chikungunya virus (CHIKV) infection and more than 140 associated deaths worldwide. These cases were recorded across 21 countries in the Americas, Asia, Africa, and Europe [3]. The most common clinical manifestations include fever, skin rash, headache, myalgia, joint swelling, and persistent arthralgia. Although Chikungunya rarely leads to fatal outcomes, it can cause severe complications, particularly among elderly individuals and patients with underlying chronic conditions. Effective control of disease transmission therefore relies heavily on early detection and timely intervention.

In April 2005, CHIKV re-emerged in the Indian Ocean region, triggering a major outbreak of dengue-like illness in the Comoros Islands. Subsequent cases were rapidly reported in Mayotte, Mauritius, and La Réunion, with attack rates ranging from 35% to 75%, revealing substantial underdiagnosis and misclassification of cases in several affected areas. India confirmed widespread CHIKV circulation later in 2005, after more than three decades without reported cases, with the number of suspected infections exceeding 1.3 million, largely driven by the epidemic expansion in Sri Lanka and Southeast Asia. During this period, autochthonous transmission was documented for the first time in several countries, including Italy, France, New Caledonia, Papua New Guinea, Bhutan, and Yemen, following the introduction of the virus by viremic travelers into previously non-endemic regions. In response to the rapid global dissemination of CHIKV, the Pan American Health Organization and the Centers for Disease Control and Prevention intensified preparedness efforts for potential outbreaks in the Americas. The first local transmission in the Caribbean was reported in Saint Martin, and by July 2014, CHIKV had spread to more than 20 countries, resulting in over 440,000 reported cases (see Figure 1).

Figure 1: Chikungunya Virus (CHIKV) repartition in the western hemisphere [4].

The Centers for Disease Control and Prevention (CDC) have reported more than 230 imported cases of CHIKV

in the United States, underscoring the global public health threat posed by the disease, which has spread from Kenya to over 50 countries in less than a decade [4]. The different transmission routes are illustrated in Figure 2.

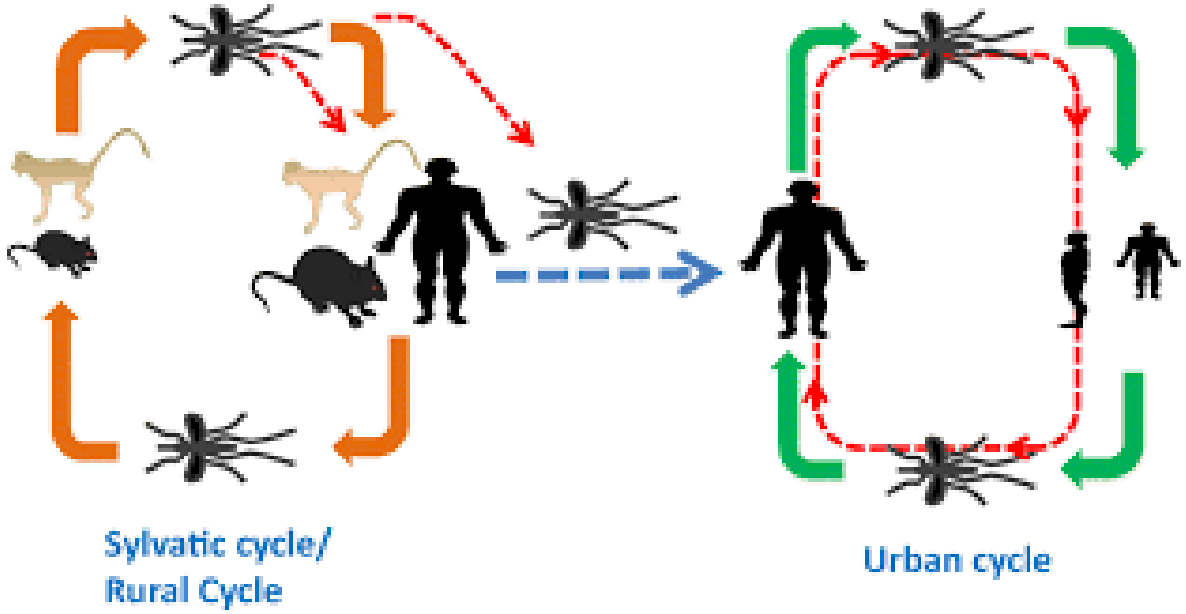

Figure 2: Chikungunya transmission routes [5].

The transmission of CHIKV occurs through two distinct cycles: the urban cycle and the sylvatic cycle. In Africa, the sylvatic cycle represents the predominant mode of transmission. In densely populated regions, however, CHIKV is mainly sustained through the urban cycle, in which humans act as the primary hosts and mosquitoes of the genus *Aedes* serve as vectors (see Figure 2). Nevertheless, *Ae. aegypti* remains a major viral vector, as illustrated by the 2013 outbreak in the Caribbean. Since 2005, several cases have been attributed to vertical transmission from mother to child, which has been reported to be particularly severe when maternal infection occurs within four days after delivery [6].

Most individuals fully recover from CHIKV infection. However, sporadic neurological, cardiac, and ophthalmic complications have been documented. Severe clinical manifestations are more frequently observed in infants and elderly individuals. Moreover, CHIKV infection may increase the risk of mortality among older patients with underlying comorbidities and in newborns infected during childbirth [7]. According to current evidence, individuals who recover from the disease are likely to develop long-lasting immunity against reinfection [8].

The combined implementation of medical interventions and vector control strategies is therefore essential for the effective management of Chikungunya outbreaks [9].

Artificial intelligence is increasingly applied across several scientific fields, including engineering, biology, and medicine [10–18]. Numerous studies have explored the use of machine learning models to predict the spread of vector-borne diseases such as dengue and chikungunya in Brazil [19–25]. For instance, da Silva *et al.* [24] proposed a machine learning-based framework for developing disease predictors capable of identifying both cases and affected

areas. To further enhance the identification of relevant risk factors, they also suggested the creation of an artificial expert committee based on metaheuristic techniques. In [25], the same authors provided a comprehensive review of the literature, focusing on strategies for forecasting arbovirus cases and identifying breeding sites.

The objective of this study is to employ ensemble regression techniques derived from artificial intelligence to develop a forecasting model for chikungunya. Each ensemble regression model is implemented using real epidemiological data from Brazil, Chad, and Paraguay. In addition, different data enrichment strategies are applied to increase the size and diversity of the datasets for Chad and Paraguay, with the aim of improving the accuracy of the predictions.

This work provides three methodological and empirical contributions that extend beyond the most common frameworks in the literature on arbovirus forecasting, particularly for chikungunya:

(i) **Multi-country comparative modeling (Africa & South America):** Most predictive studies focus on a single country, region, or even a single city, which limits the generalizability of their findings [26, 27]. Other multi-regional investigations emphasize risk mapping or virus co-circulation (often dengue/zika–chikungunya) rather than a systematic comparison of predictive models using uniform training and evaluation procedures [28, 29]. We propose a unified modeling and evaluation protocol applied to three contrasting epidemiological contexts (Chad, Brazil, and Paraguay). This framework enables us to (i) quantify the relative robustness of Random Forest, XGBoost, and Voting Regressor models across different climatic gradients and surveillance systems, and (ii) identify the factors whose predictive effects remain the most stable across countries. This comparative strategy directly addresses recent calls for more generalizable arbovirus forecasting frameworks in the context of climate change [30].

(ii) **Data augmentation strategy adapted to short epidemiological series:** Forecasting models are often constrained by the short and noisy nature of weekly chikungunya time series [31, 32]. To address this limitation, we introduce a controlled augmentation of time series data using overlapping sliding windows and seasonally stratified sub-series, thereby increasing the diversity of training trajectories while preserving temporal dependencies (lags and epidemic peaks). This approach is consistent with recent evidence showing that series-specific augmentation improves generalization in small biomedical datasets [31, 32]. To the best of our knowledge, such strategies remain rarely documented for chikungunya (in contrast to dengue) and offer valuable leverage for data-limited settings such as Chad.

(iii) **Explicit integration of climatic and epidemiological covariates with etiological lags:** Both theoretical and empirical studies indicate that temperature, precipitation, and humidity influence the vector competence of *Aedes* mosquitoes with measurable time delays (lags) [33–36]. Several investigations further confirm that incorporating climatic variables into epidemiological models improves the forecasting of

arboviral diseases [37, 38]. Our contribution consists in systematically linking climate data (temperature, relative humidity, and precipitation) with case series through (i) a set of lagged features (+1 to +12 weeks) and (ii) multi-scale aggregations (moving averages and maxima), within a unified multi-country evaluation framework. This design allows us to identify, in a comparable manner across countries, the climatic windows that provide the most stable predictive information, in line with recent mechanistic and empirical findings [28, 33].

By integrating (i) a multi-country comparative framework, (ii) data augmentation tailored to short epidemiological series, and (iii) rigorous climate–case cross-referencing with time lags, this study goes beyond the direct application of standard methods and offers elements of generalizability and data parsimony that are rarely combined in chikungunya research.

The remainder of the paper is organized as follows. Section 2 presents the materials and methods. The results and their discussion are provided in Section 3. Finally, the paper concludes with a summary of the main findings and perspectives.

2 Material and methods

The materials (including Study areas, the hardware configuration and the software development environment) utilized to carry out this work are presented in this part. We will also go over the strategies and tactics required to comprehend this article.

2.1 Material

2.1.1 Study areas

A total of 927 cases were recorded on September 3, 2020; all patients were managed as outpatients, and no fatalities were reported. By that date, 13,488 cumulative cases had been documented, with no deaths recorded. Additional suspected cases were subsequently reported in Adré and Biltine (see Figure 3). As of October 2, 2020, 415 cases had been confirmed, including 247 in Abéché, 165 in Biltine, 3 in Gozbeida, and none in Abdi. No fatalities were reported at that time. Overall, one death and a total of 34,052 cases had been recorded by that date, with all patients receiving outpatient care [39].

The geographical distribution of chikungunya cases in Chad is illustrated in Figure 3, while the temporal evolution of daily cases and deaths is presented in Figure 4.

Figure 3: Region infected by CHIKV in Chad [39].

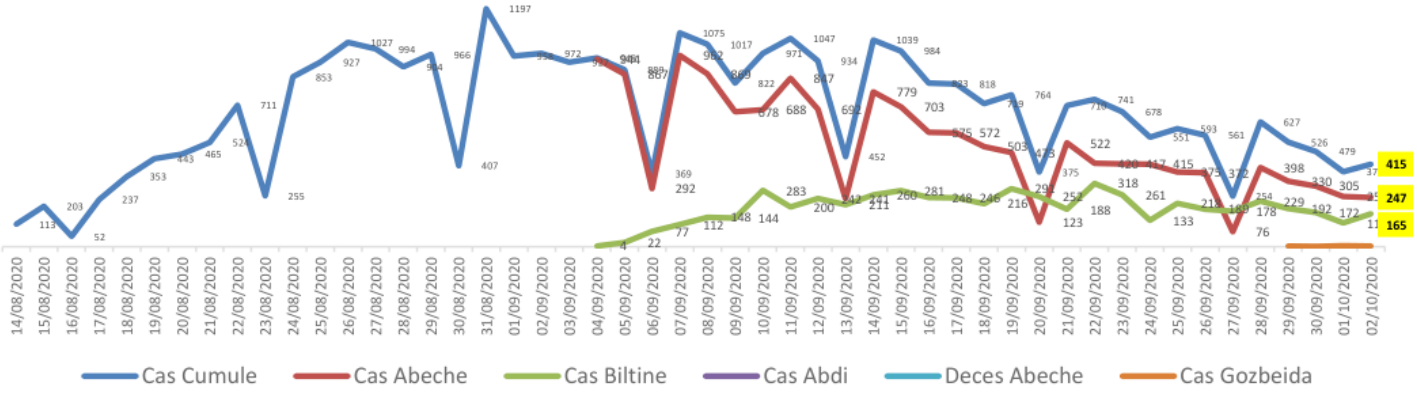

Figure 4: Daily evolution of Chikungunya cases and deaths in Chad [39].

Brazil: With a large population susceptible to CHIKV, a favorable climate, and abundant vector populations, Brazil is the largest and most populous country in Latin America. Since 2013, CHIKV has been locally transmitted across the country, with most of the initial cases reported in the northeastern region. Brazil has been the epicenter of the chikungunya epidemic in the Americas since 2016, with a total of 1,659,167 reported cases—the highest number recorded in the region. Unlike other countries and territories in the Americas, Brazil experiences chikungunya outbreaks on a yearly basis [40].

The cumulative number of chikungunya cases reported to the Brazilian Ministry of Health between March 2013 and June 2023 across the 26 states and the Federal District is presented in Figure 5. The spatial and temporal dispersion of chikungunya virus lineages throughout the Brazilian states and the Federal District is also illustrated.

Up to August 17, 2023, the years in which at least one viral genome was sequenced and deposited in GenBank were used to identify the circulating chikungunya lineages [40].

Chikungunya re-emerged in 2022–2023 after a relatively quiet period. Brazil was particularly affected, especially the state of Minas Gerais, where the incidence reached 395 cases per 100,000 inhabitants. Since its introduction into Brazil in 2014, the disease has progressively spread from the northeast to the southeast, with 3.6 million cases reported to PAHO/origin. In 2023, 30,724 cases—twice the number recorded in 2022—were reported within only 10 weeks in the southeastern region (see Figure 5). With epidemic peaks observed in 2018 and 2022, the virus reproduction rate has reached high levels, ranging from 1.5 to 2.5 [41].

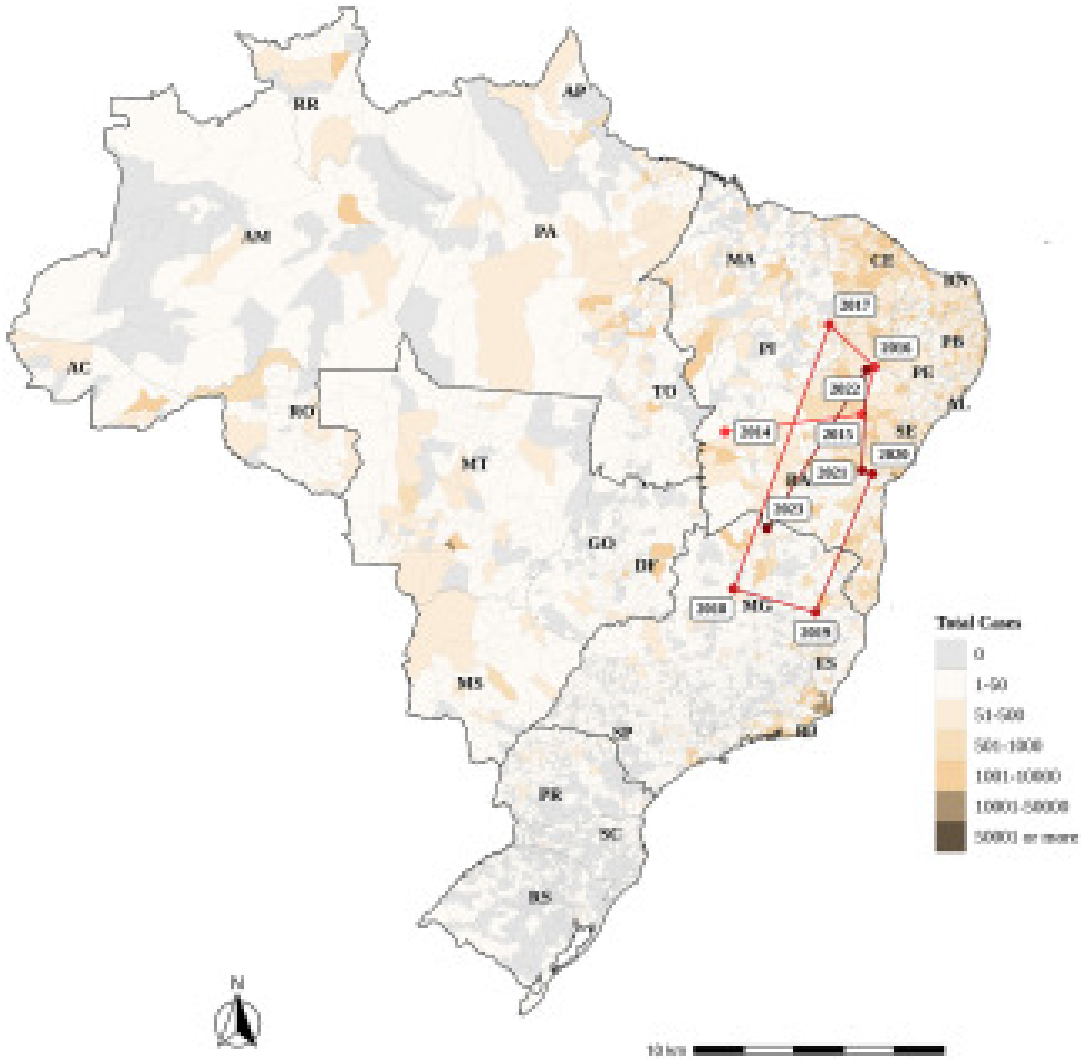

Figure 5: Map showing the annual epicentres of Chikungunya cases and the total number of cases by commune through 2023 [42]. AC=Acre, AL=Alagoas, AM=Amazonas, AP=Amapá, BA=Bahia, CE=Ceará, ES=Espírito Santo, DF=Distrito Federal (Federal district), GO=Goiás, MA=Maranhão, MG=Minas Gerais, MS=Mato Grosso do Sul, MT=Mato Grosso, PA=Pará, PB=Paraíba, PE=Pernambouc, PI=Piauí, PR=Paraná, RJ=Rio de Janeiro, RN=Rio Grande do Norte, RO=Rondônia, RR=Roraima, RS=Rio Grande do Sul, SC=Santa Catarina, SE=Sergipe, SP=São Paulo, TO=Tocantins, Km=kilometers, ECSA-American, East-Central-South-African-American sub-lineage

Paraguay: Since the discovery of indigenous illnesses in Paraguay in 2013, the virus has been found there annually. Four epidemic waves in 2015, 2016, 2018, and 2023 have been linked to potential CHIKV infections in Paraguay and have all occurred during the summer. A total of 118,179 suspected and confirmed infections, including 3,510 hospitalized case-patients and 46 fatalities, were reported between October 2, 2022, and April 10, 2023. Furthermore, 125 (43%) of the 294 suspected cases of acute meningoencephalitis that were recorded were related to CHIKV [43]. Paraguay experienced steady temperature increases during the past 40 years, with CHIKV recurrence in 2022 corresponding to the highest mean temperatures. Confirmed illnesses were limited to Central, Paraguarí, and Amambay districts (see figure 6). Panel (C) of figure 6 depicts the Chikungunya cases reported each week (gray area), standardized incidence per 100,000 people (blue line) and cumulative deaths.

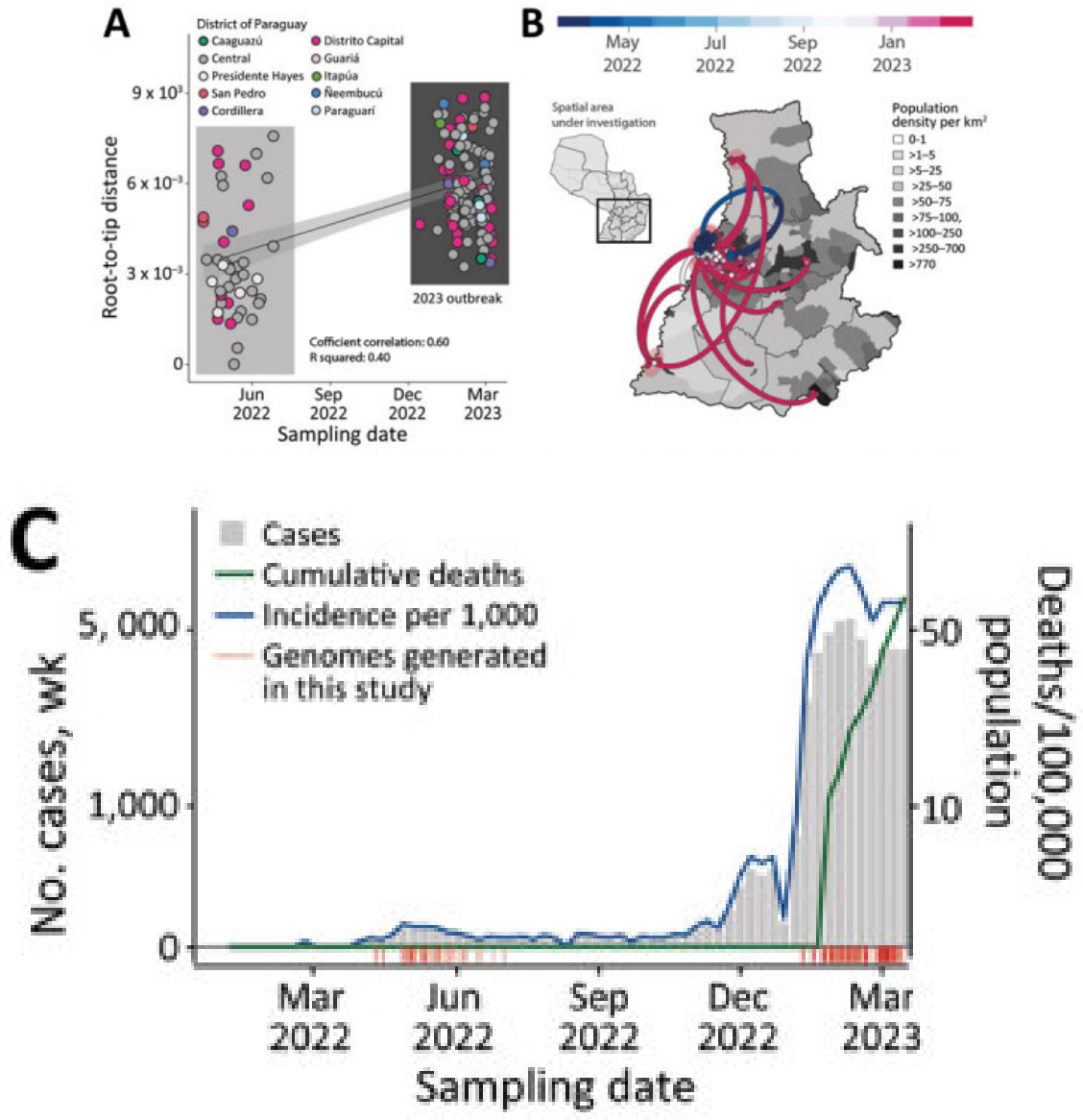

Figure 6: The Chikungunya cases declared every week in Paraguay: A) With 90% confidence intervals, the study looks at the genetic distances between the virus samples and the dates of sampling. The colors indicate the samples' original locations; B) The map illustrates the geographic spread of the CHIKV ECSA virus in Paraguay, with the circles denoting significant phylogenetic locations; C) Evolution of the disease [44]

2.2 Methods

2.2.1 Data collection

The data used in this work come from different sources:

Epidemiological data

- Chad: During the 2020 Chikungunya outbreak, epidemiological statistics were taken from a World Health Organization (WHO) study [39]. This dataset covers the period from August 12, 2020 to November 10, 2020;
- Brazil: data was collected from mendeley website[1] [45]. Clinical, sociodemographic, and laboratory data pertaining to patients with confirmed cases of dengue and Chikungunya are presented in this dataset. It covers the period from 2013 to 2021;
- Paraguay: Data was collected via the PAHO website[2], which reports Chikungunya cases in real time, with weekly records varying between 2013 and 2017.

Climatic data Climate data for Paraguay and Chad were obtained from weatherandclimate[3], while for Brazil were uploaded to Kaggle[4] collected by INMET (National Institute of Meteorology of Brazil), corresponding to the same time intervals as the cases of Chikungunya in each country, namely:

- Chad: in the cities of Biltine, Abeche and Abdi;
- Brazil: in the cities of Amapá, Bahia, Ceará, Espírito Santo, Federal District, Goiás, Maranhao, Minas Gerais, Mato Grosso do Sul, Mato Grosso, Pará, Paraíba, Pernambuco, Piauí, Paraná, Rio de Janeiro, Rio Grande do Norte, Rondônia, Roraima, Rio Grande do Sul, Santa Catarina, Sergipe, São Paulo and Tocantins;
- Paraguay: asuncion and central.

These datasets include variables such as temperature, humidity, and precipitation, which are essential for understanding how meteorological conditions influence disease transmission.

[1] https://data.mendeley.com/datasets/2d3kr8zynf/2
[2] https://www3.paho.org/data/index.php/en/mnu-topics/chikv-en/550-chikv-weekly-en.html
[3] http://weatherandclimate.com/
[4] https://www.kaggle.com/datasets/gregoryoliveira/brazil-weather-information-by-inmet

2.2.2 Data Exploration and Preparation

The analysis of climatic variables—namely temperature, precipitation, and humidity—constituted the first step of the data exploration process. The objective of this phase was to examine the distribution of climate data and its temporal variability.

To address missing data in the datasets from Chad and Paraguay, several imputation and data completion strategies were considered, including the KNN imputer [46, 47] and data augmentation techniques [48]. The k-Nearest Neighbors (k-NN) algorithm is a non-parametric supervised learning method used for both classification and regression tasks. It operates by identifying the k nearest data points (neighbors) to a given observation and making predictions based on their corresponding values [47]. In this study, k-NN was applied to complete missing values in the datasets [46]. Data augmentation refers to the artificial expansion of a dataset through the modification of existing samples or the generation of new synthetic observations. This strategy enhances model robustness and generalization by providing a more diverse and representative training set [48].

Remark 1. *With regard to the data augmentation strategy, synthetic observations were generated by injecting Gaussian noise into the numerical variables. Specifically, for each target country (Chad and Paraguay), a KNN imputer with $k = 5$ was first applied to fill in missing values for temperature, humidity, precipitation, and case counts. Subsequently, 100 additional samples were created by randomly selecting existing observations and applying proportional noise to each variable, using a noise level factor of 0.05. This procedure yields realistic synthetic data while increasing the overall diversity of the dataset. This approach was preferred over alternative techniques such as SMOTE or simple bootstrapping, as the present study addresses a regression problem rather than a classification task with class imbalance. The objective here is to enrich a limited dataset with consistent and plausible values rather than to balance categorical outcomes [49].*

A method called feature engineering (FE) involves taking raw data and turning it into new variables, or features. We employed this method in our study to enhance the Brazilian epidemiological data [50, 51]. In fact, we came across a dataset containing individuals who tested positive for both dengue and Chikungunya while gathering data on cases of the disease. The data of patients who tested positive for Chikungunya was then cross-referenced with the dates on which the illness was discovered. As a result, our epidemiology database was enhanced and a new variable (feature) linking each instance to a particular date could be created. Figure 7 illustrates the importance of these techniques in model performance.

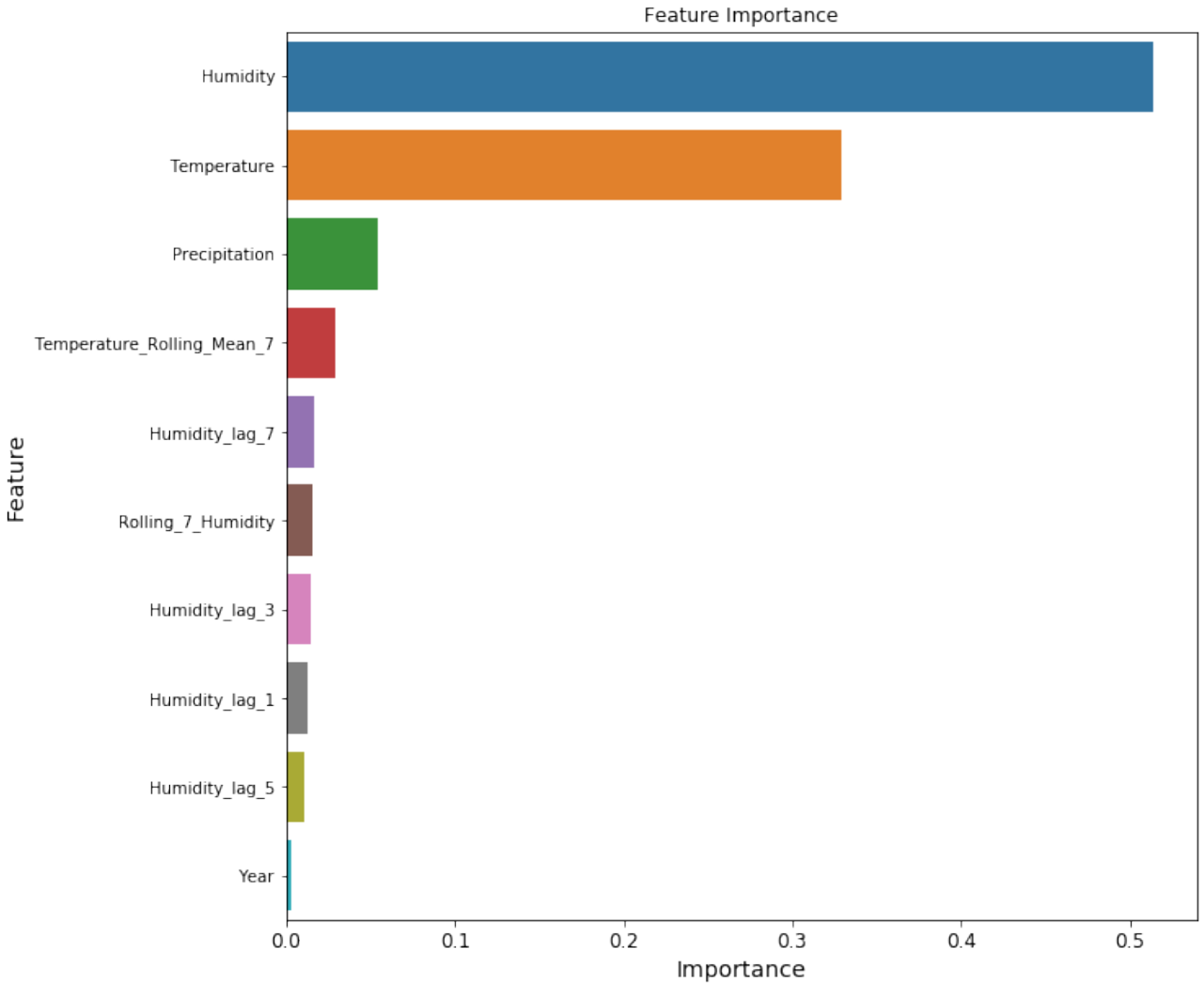

Figure 7: Feature performances.

2.2.3 Data Exploration and preparation

Here, we will thoroughly examine our data. The headings of the various datasets will be shown first, followed by an analysis of how Chikungunya has changed throughout the years in the countries under study. Lastly, we will draw attention to the characteristics' relationships, specifically those between the epidemiological and climatic data.

An identical modeling pipeline was applied to all three countries (Chad, Brazil, and Paraguay). For each dataset, the same preprocessing procedures (missing data imputation, normalization), feature engineering strategy (lagged epidemiological variables and climatic covariates), machine learning models (Random Forest, XGBoost, and Voting Regressor), hyperparameter optimization via grid search, and evaluation metrics: Mean Absoluet Error (MAE), Root Mean Square Error (RMSE), and R^2) were used. This unified pipeline guarantees that performance differences across countries reflect underlying epidemiological and climatic heterogeneity rather than methodological bias.

Data headers: Figure 8 illustrates the main variables available in climate datasets, which include date, temperature, humidity, and precipitation.

```
In [6]: df_climate['brazil'].head()
```

Out[6]:

	Date	Temperature	Humidity	Precipitation
0	2013-01-01	26.750000	75.541667	5.358333
1	2013-01-02	26.041667	76.333333	1.716667
2	2013-01-03	25.875000	76.708333	1.804167
3	2013-01-04	26.250000	77.583333	2.087500
4	2013-01-05	26.625000	76.833333	2.708333

Figure 8: Example of climate data headers (Brazil case)

Figure 9 shows the main variables present in the epidemiological datasets, covering Chikungunya cases over time.

```
In [17]: df_cases['brazil'].head()
```

Out[17]:

	Date	Cases
0	2012-12-30	120
1	2012-12-31	258
2	2013-01-01	376
3	2013-01-02	1139
4	2013-01-03	1463

Figure 9: Example of epidemiological data headers (Brazil case)

So, when we combine the epidemiological data with the climate data, the header of the resulting dataset is shown in figure 10.

	Date	Temperature	Humidity	Precipitation	Cases
0	2013-01-01	26.750000	75.541667	5.358333	376
1	2013-01-02	26.041667	76.333333	1.716667	1139
2	2013-01-03	25.875000	76.708333	1.804167	1463
3	2013-01-04	26.250000	77.583333	2.087500	1628
4	2013-01-05	26.625000	76.833333	2.708333	1099

Figure 10: Example of the final dataset combining climatic and epidemiological data (Brazil case)

2.2.4 Illustration of the evolution of Chikungunya

The time-evolution of the CHIKV at Chad, Brazil and Paraguay is depicted. respectively, in figures 11, 12, and 13.

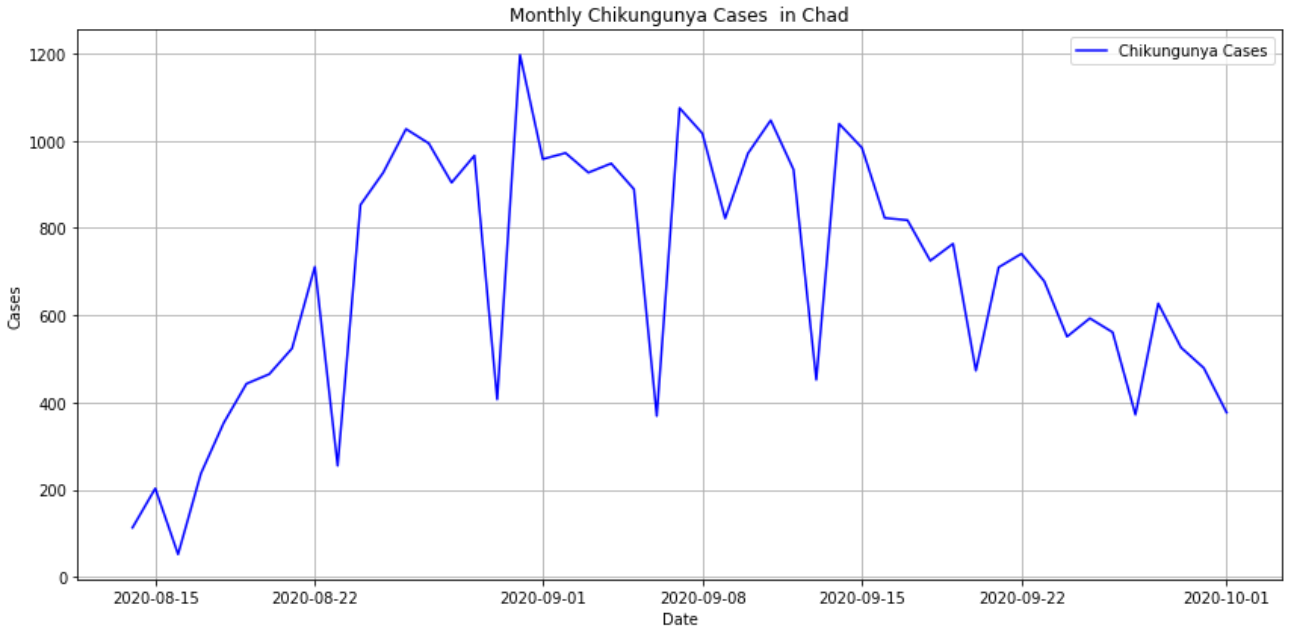

Figure 11: Historical evolution of reported Chikungunya cases in Chad over time, illustrating disease incidence

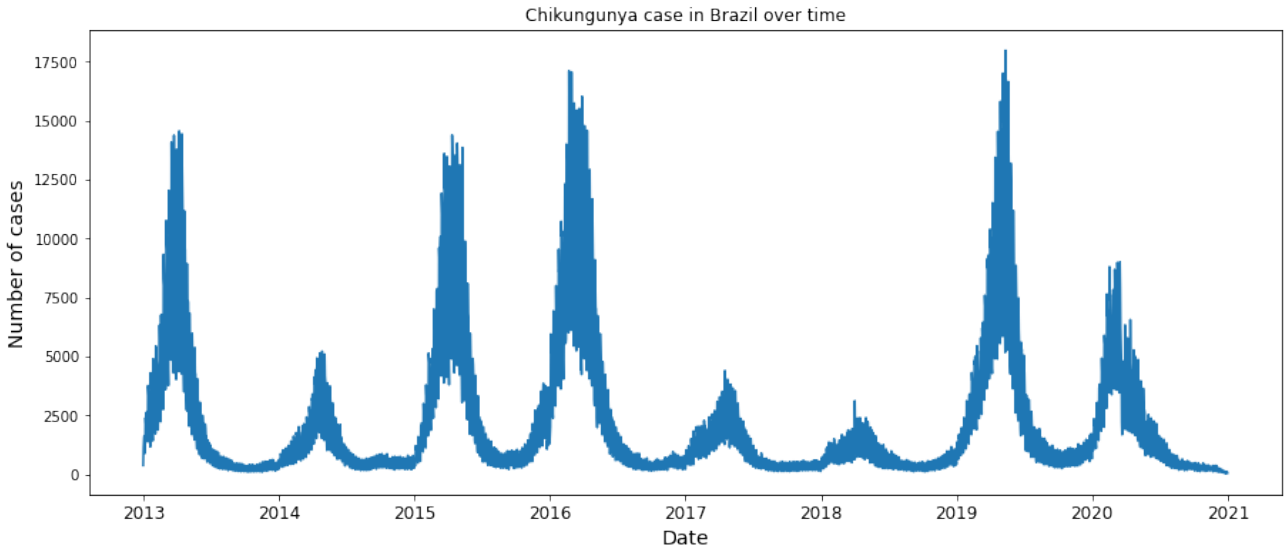

Figure 12: Historical evolution of reported Chikungunya cases in Brazil over time, illustrating disease incidence.

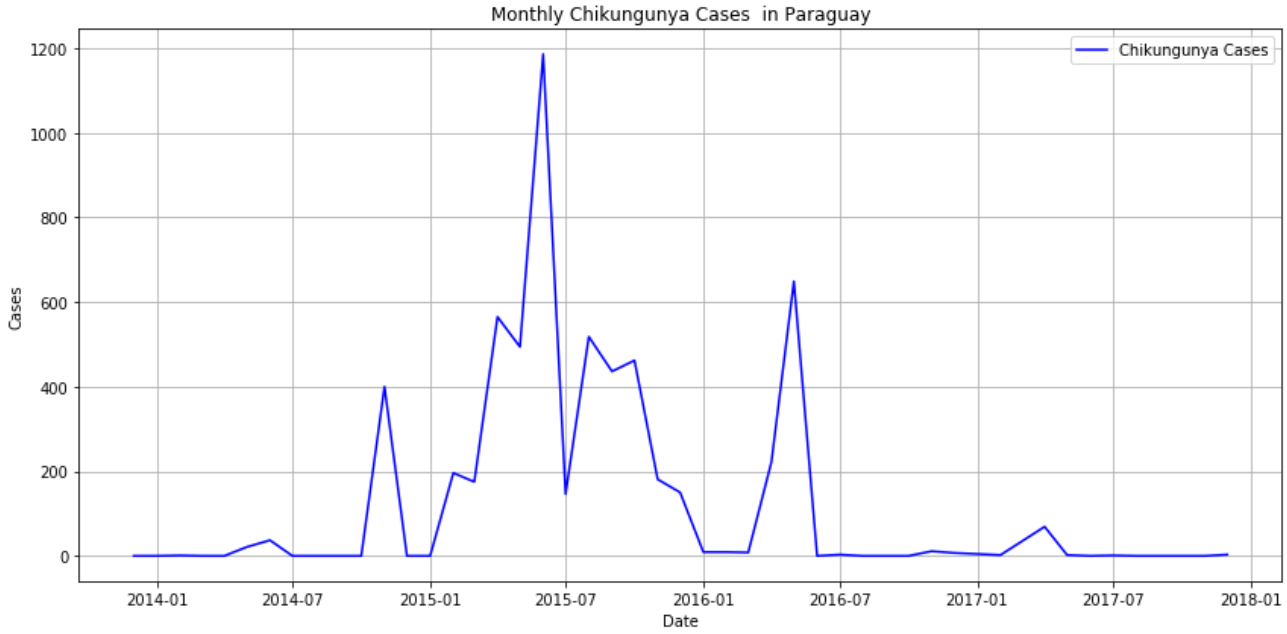

Figure 13: Historical evolution of reported Chikungunya cases in Paraguay over time, illustrating disease incidence.

Correlation analysis The following analysis explores the relationship between the number of Chikungunya cases and climatic variables.

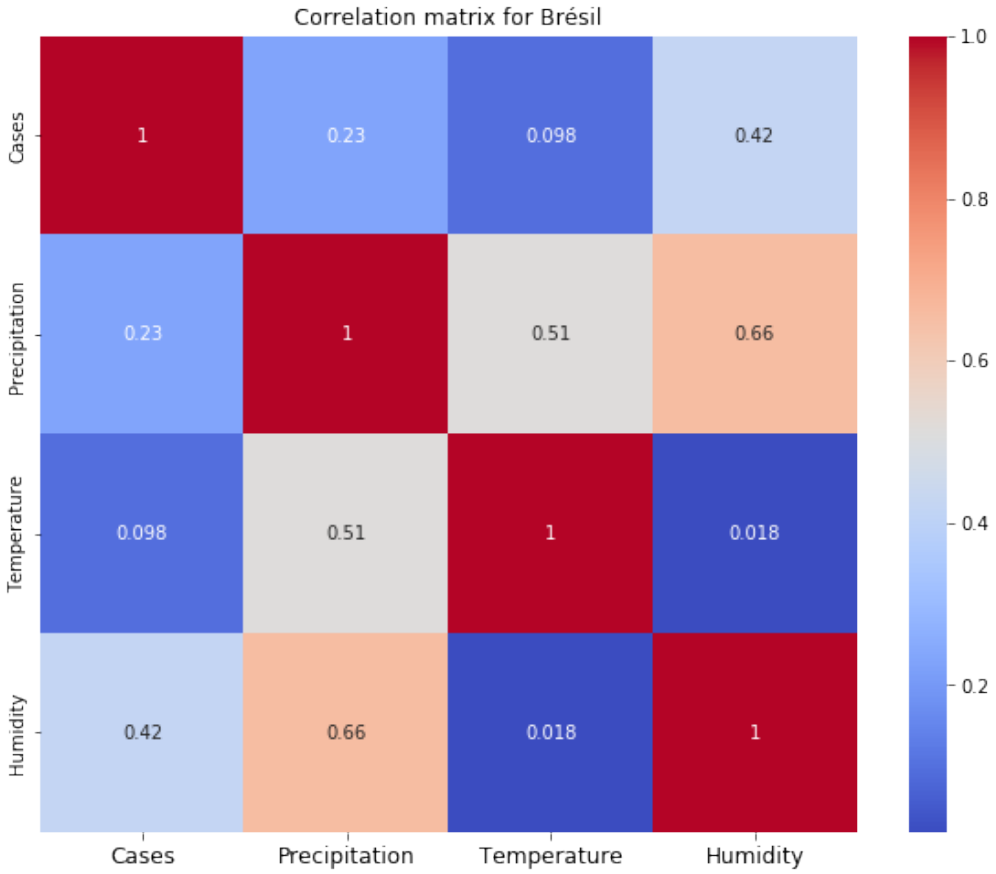

Figure 14: orrelation between climate variables and number of cases.

Anomalies, such as sharp variations or gaps in the records, were also brought to light by data mining and required particular methodological changes to guarantee the accuracy of subsequent analysis.

2.3 Predictive models

In order to forecast Chikungunya cases, we choose to employ multiple supervised regression models. Using an ensemble model, this method enables us to assess the effectiveness of several algorithms and get predictions that are more reliable. The following models were selected: Random Forest Regressor, XGBoost Regressor with Grid Search, and Ensemble Model (Voting Regressor).

2.3.1 Theoretical Foundations of the predictive models

Random Forest: Using a bootstrap sample and a subset of variables, Random Forest [52] builds a forest of B regression trees $h(x, \theta_b)$. The average of each individual prediction is the final forecast:

$$\hat{f}_{\mathrm{RF}}(x) = \frac{1}{B} \sum_{b=1}^{B} h(x, \theta_b).$$

This bagging technique keeps bias low while lowering variance. Despite the large number of noisy variables, the model remains consistent and robust to dimensionality [53].

XGBoost: With regularization and second-order approximation, XGBoost is an enhanced variant of gradient boosting [54]. The forecast at each iteration t is:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i),$$

and the objective to minimise is:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} \ell\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t),$$

with the regularization :

$$\Omega(f) = \gamma T + \tfrac{1}{2}\lambda \sum_{j=1}^{T} w_j^2.$$

By second-order Taylor approximation:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2 \right] + \Omega(f_t),$$

where g_i and h_i are the first and second derivatives of the loss with respect to the previous prediction. The optimal value of the weight of a leaf j is written as:

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}.$$

The separation gain for a split is given by:

$$\mathcal{L}_{\text{split}} = \frac{1}{2}\left[\frac{(\sum_{i \in I_L} g_i)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{(\sum_{i \in I_R} g_i)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{(\sum_{i \in I} g_i)^2}{\sum_{i \in I} h_i + \lambda} \right] - \gamma.$$

By considering regularization and scalable architecture, this formulation makes it possible to construct trees efficiently.

Voting Regressor: The Voting Regressor combines the predictions of multiple base regression models $\{M_1, M_2, \ldots, M_$ [55, 56]. Two aggregation strategies are commonly used:

- **Simple averaging**:

$$\hat{y}_{\text{vote}} = \frac{1}{m} \sum_{k=1}^{m} \hat{y}_k, \tag{1}$$

- **Weighted averaging**:

$$\hat{y}_{\text{vote}} = \sum_{k=1}^{m} w_k \hat{y}_k, \qquad \text{with} \sum_{k=1}^{m} w_k = 1, \tag{2}$$

where w_k denotes the weight assigned to model M_k, allowing stronger models to contribute more to the final prediction.

The architecture of the Ensemble model is shown in figure 15.

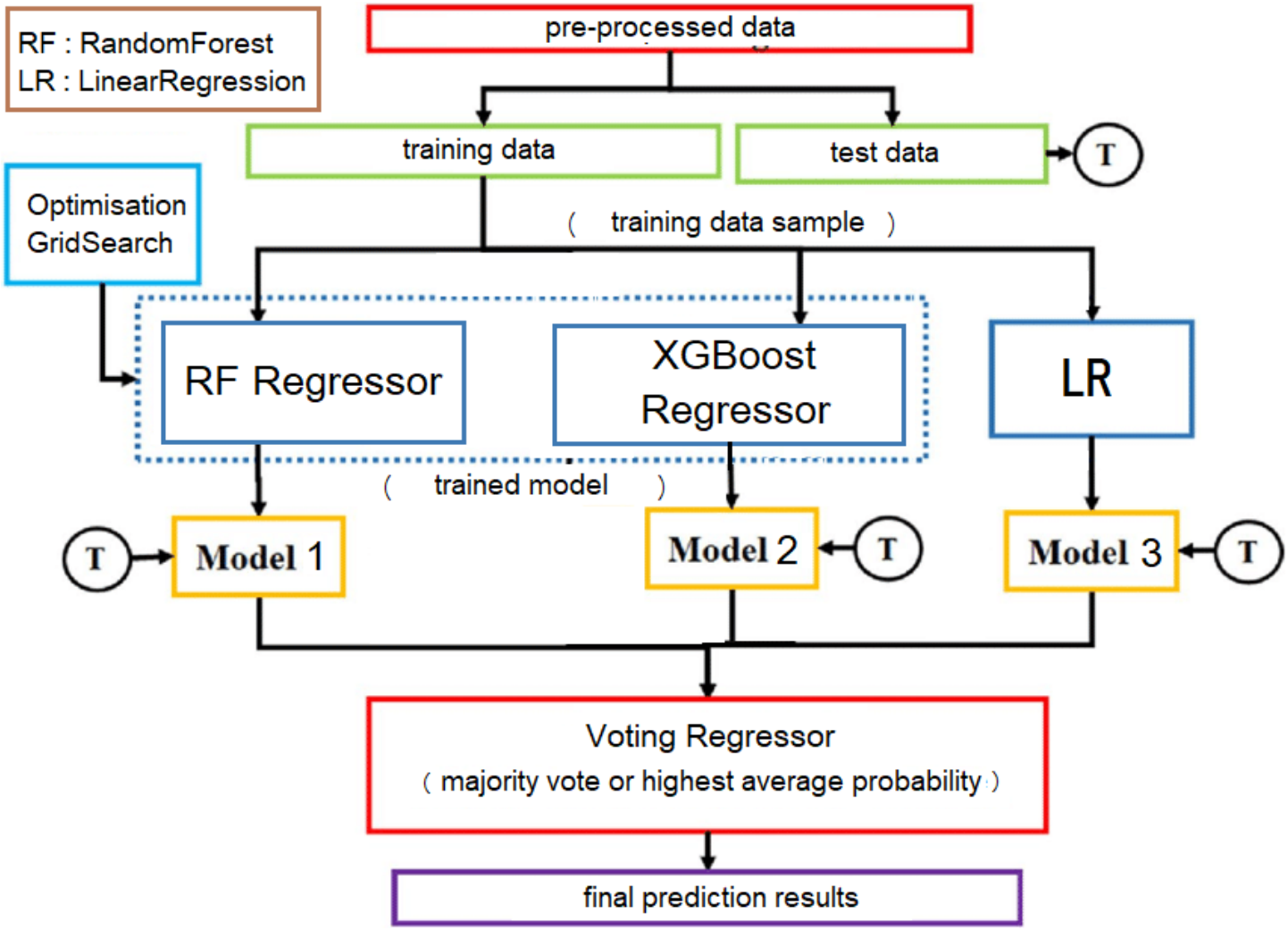

Figure 15: Architecture of the Ensemble models.

2.4 Evaluation metrics

Predictive models are evaluated using three performance criteria: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and coefficient of determination (R^2). In the following, n represents the total number of Chikungunya cases, y_i denotes the actual value of Chikungunya, and $\hat{y}_i$ represents the anticipated value of Chikungunya.

2.4.1 Mean Absolute Error (MAE)

The mean absolute error calculates the average of the absolute values of the variances between the expected and actual values.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|. \tag{3}$$

2.4.2 Root Mean Square Error (RMSE)

The Root Mean Square Error measures the difference between the predicted and actual values of a model. It is commonly used in the fields of regression and time series forecasting, especially to assess the accuracy of forecasting models. The expression is as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{4}$$

2.4.3 The coefficient of determination (R^2)

The coefficient of determination R^2 is a statistical metric that determines how much of a dependent variable's variation can be predicted by one or more independent variables in a regression model. It shows how well the model accounts for variations in the observed data. The expression is as follows:

$$R^2 = 1 - \frac{SSR}{SST}, \tag{5}$$

where

$$SSR = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{6}$$

is the sum of the squares of the residuals, representing the variance not explained by the model, and

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2, \tag{7}$$

is the sum total of squares, representing the total variance of the data.

3 Results and Discussion

This part focuses on the examination of the forecast outcomes that are derived from the application of machine learning techniques to climate data and past Chikungunya cases in three countries: Brazil, Paraguay, and Chad. The objective is to provide a thorough presentation of the findings, suggest a thorough discussion, and point out any flaws or restrictions.

Table 1: Description of optimized hyperparameters for Random Forest Regressor with GridSearch

	Hyperparameters			
Country	n_estimators	max_depth	min_samples_leaf	min_samples_split
Chad	200	10	2	2
Brazil	200	20	2	2
Paraguay	200	None	1	5

Table 2: Description of optimized hyperparameters for the XGBoost Regressor with GridSearch

	Hyperparameters			
Country	n_estimators	max_depth	subsample	learning_rate
Chad	300	5	0.8	0.2
Brazil	100	5	1.0	0.2
Paraguay	100	7	0.1	0.2

3.1 Results

3.1.1 Cross-validation method (Training and Test)

The dataset, consisting of 366 instances for Chad and 1826 instances for Brazil and Paraguay, including climate data and information on Chikungunya cases, was divided into two parts for training. First, the data was shuffled and then separated into two sets: 80% of the data was used for training (training set) and 20% was reserved for testing (testing set).

3.1.2 Choice of Hyperparameters for Ensemble Models

The hyperparameters of the models used to determine their performance are detailed below.

3.1.3 Random Forest Regressor

This section details the choice of specific hyperparameters for the Random Forest Regressor and their impact on prediction accuracy. These hyperparameters were adjusted using the grid search method to optimize the performance of the model illustrated in Table 1.

3.1.4 XGBoost Regressor

The hyperparameters used for the XGBoost Regressor are listed in table 2.

3.1.5 Results obtained by our models

In this section, we present our results through a table, followed by an in-depth discussion. Subsequently, we illustrate the various performance measures using various graphs.

Table 3: Performance indicators (MAE, RMSE, R^2 score) of the models for Brazil, Chad and Paraguay.

Country	**Models**	**MAE**	**RMSE**	R^2 **Score**
Brazil	Linear Regression	1178.99	1591.25	0.447
	Random Forest Regressor	897.17	1477.20	0.523
	XGBoost Regressor	827.16	1459.15	0.535
	Voting Regressor	**840.14**	**1387.23**	**0.65**
Chad	Linear Regression	56.06	98.34	0.198
	Random Forest Regressor	47.81	80.90	0.457
	XGBoost Regressor	55.17	101.13	0.152
	Voting Regressor	**50.52**	**93.02**	**0.282**
Paraguay	Linear Regression	67.15	84.68	0.332
	Random Forest Regressor	35.31	61.28	0.650
	XGBoost Regressor	40.32	71.95	0.517
	Voting Regressor	**40.37**	**62.25**	**0.59,97**

The table 3 below illustrates the metrics for evaluating the results of our work.

The results show that the overall model, in particular the Voting Regressor, outperformed the other models overall in terms of accuracy. Reduce errors (MAE and RMSE) and increase the R^2 score, indicating a better explanation of the variance in the data.

For Paraguay, the Voting Regressor obtained a low RMSE (62.25) and the best R^2 score (0.5957), outperforming the Linear Regressor and XGBoost Regressor models. In Brazil, although the XGBoost Regressor model performed very well (RMSE of 1459.15 and R^2 of 0.65), the Voting Regressor managed to obtain comparable results with a particularly low RMSE of 1387.23. For Chad, the performance of the models was weaker overall, but the Voting Regressor still showed a slight improvement over the other models, although the R^2 score remained low (0.282).

In conclusion, the ensemble models, and particularly the Voting Regressor, proved to be the best at predicting cases of Chikungunya in the three countries studied, although with more reliable predictions for Brazil and Paraguay than for Chad, which is due to the quantity of data available for this country.

Figure 16 shows, in terms of RMSE and coefficient of determination R^2, the level of performance of each of the models.

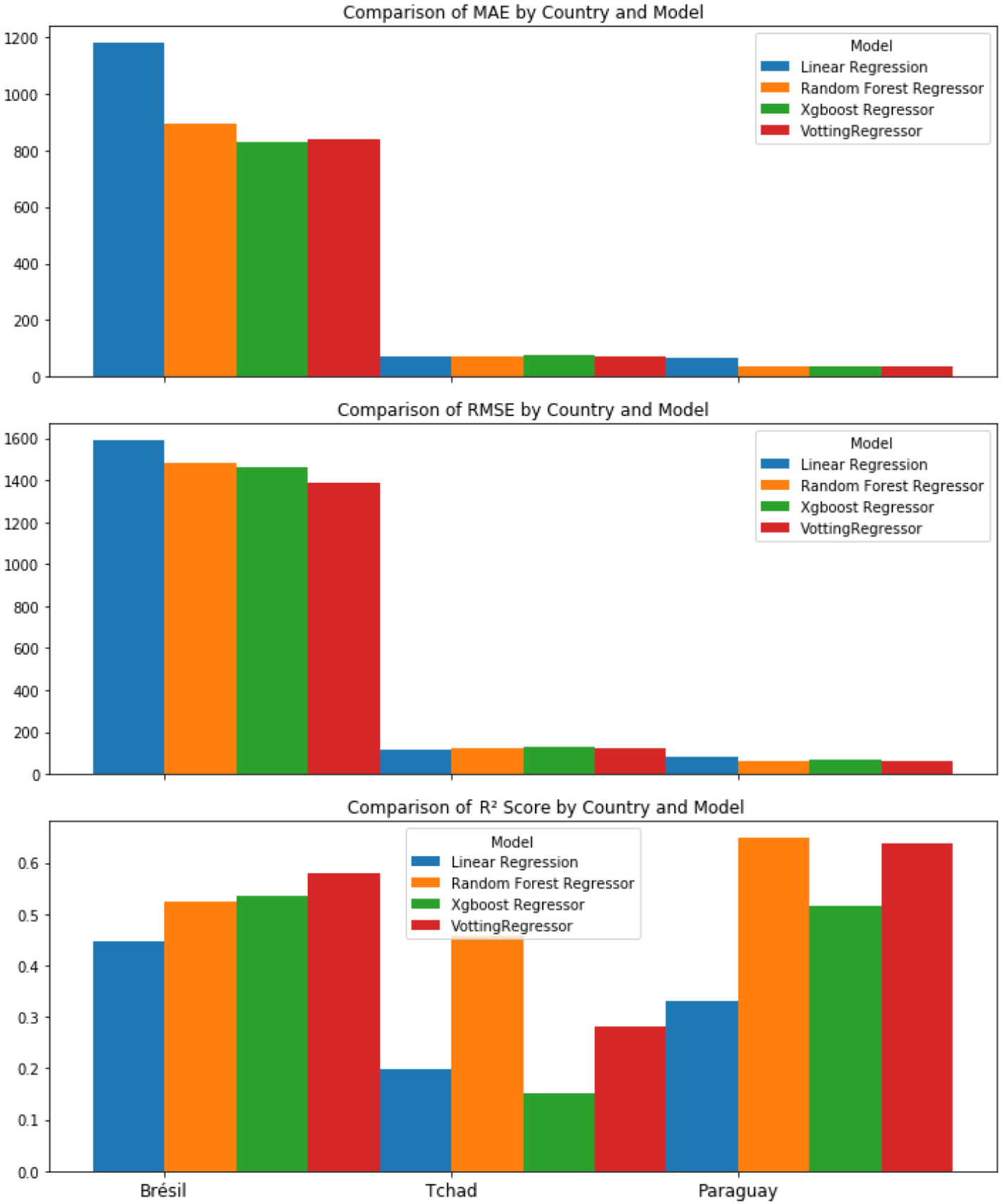

Figure 16: Performance comparison of Linear Regression, Random Forest, XGBoost, and Voting Regressor across Brazil, Chad, and Paraguay using MAE, RMSE, and R^2 metrics (top to bottom).

3.1.6 Prediction

Here we present the test predictions in each country for the overall model (VotingRegressor). Figures 17, 18, and 19 illustrate the Chikungunya prediction phase in Brazil, Chad, and Paraguay.

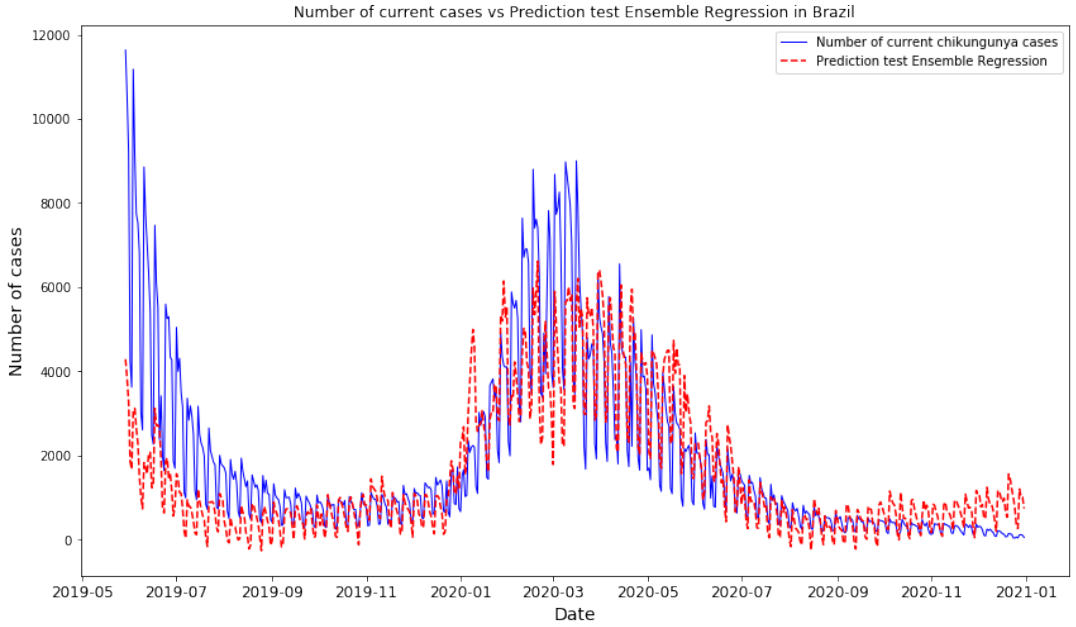

Figure 17: Comparison between observed Chikungunya cases and ensemble regression predictions in Brazil over the test period.

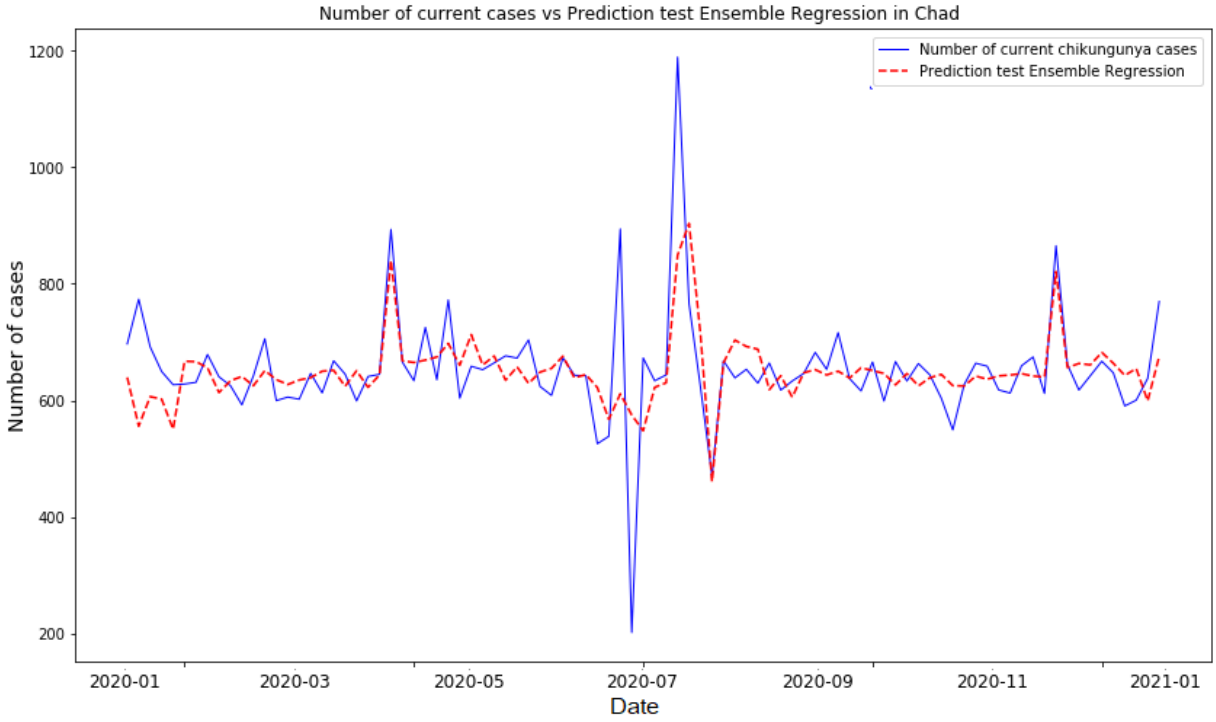

Figure 18: Comparison between observed Chikungunya cases and ensemble regression predictions in Chad over the test period.

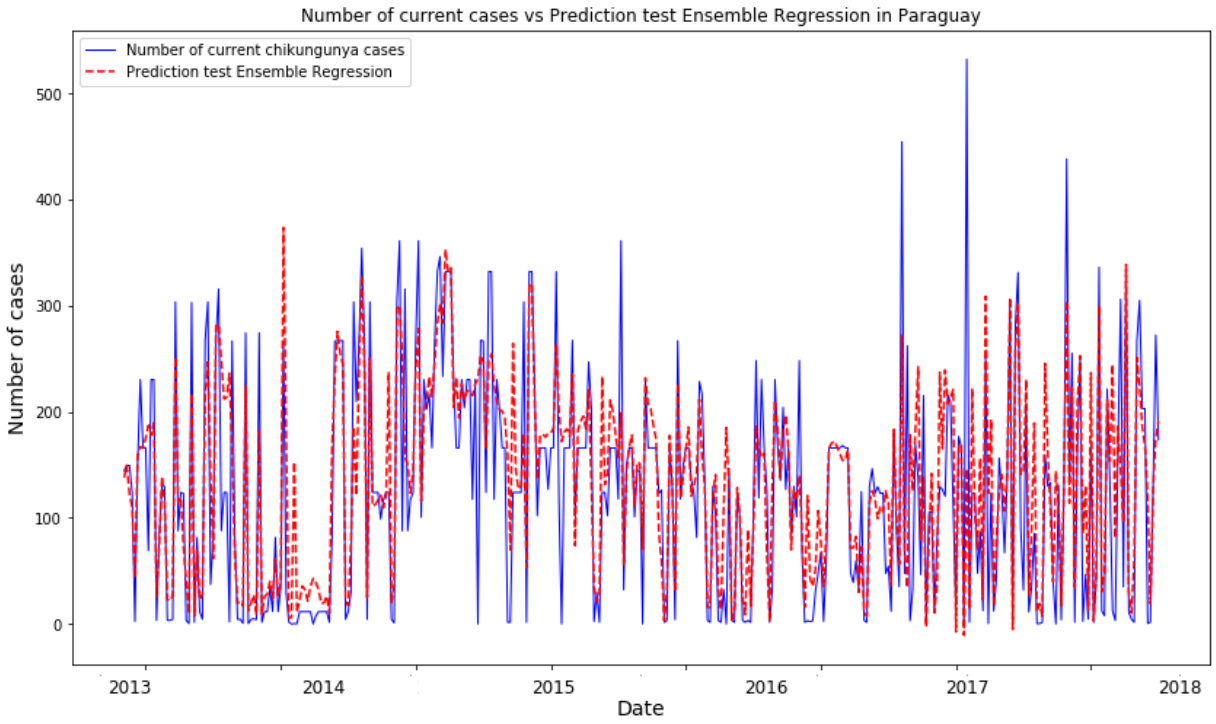

Figure 19: Comparison between observed Chikungunya cases and ensemble regression predictions in Paraguay over the test period.

Figure 20 illustrates the 3-year forecast for Chikungunya in Brazil.

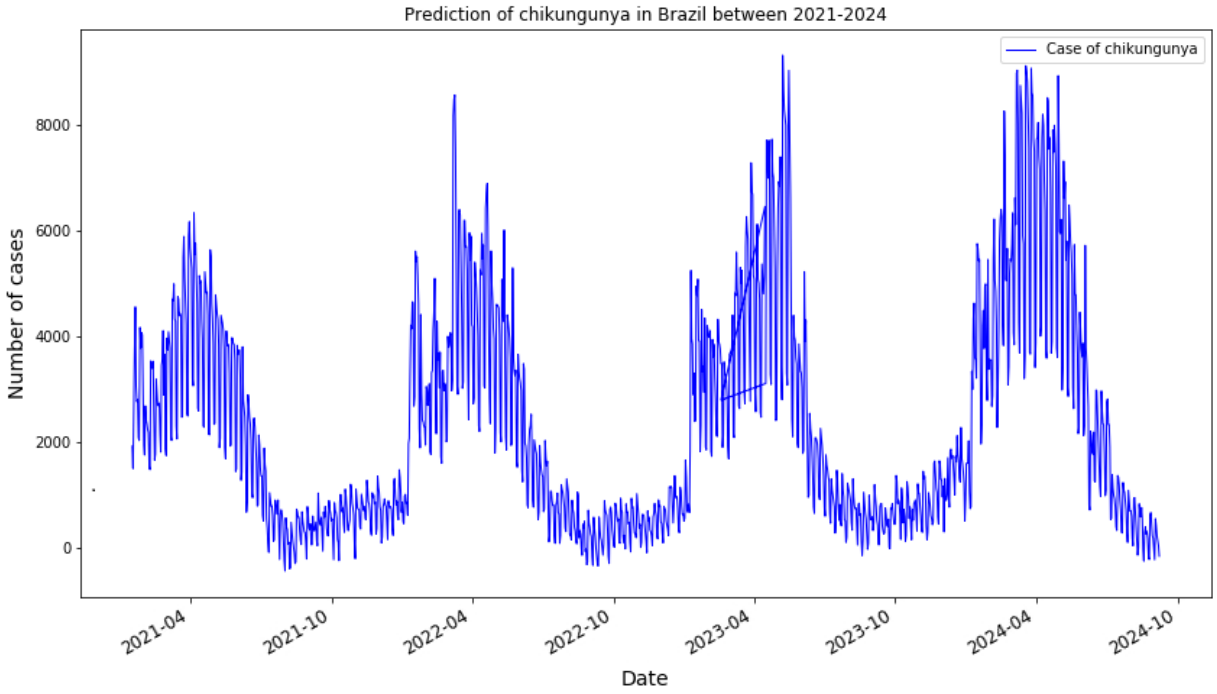

Figure 20: Forecasting of Chikungunya cases in Brazil from 2021 to 2024.

3.2 Wilcoxon statistical test and paired t-test analysis

The paired t-test did not reveal statistically significant differences between Voting Regressor and XGBoost ($p = 0.2126$) or Random Forest ($p = 0.2607$) (see Figure 21). Similarly, the Wilcoxon signed rank test confirmed that the improvements observed in the Voting Regressor compared to XGBoost ($p = 0.2081$) and Random Forest ($p = 0.2997$) were not significant at the 5% level. These results indicate that, although the VotingRegressor achieved slightly better error measures, its superiority cannot be statistically demonstrated. However, its consistently competitive performance across different countries suggests that the ensemble strategies remain useful in stabilizing predictions in heterogeneous epidemiological and climatic contexts.

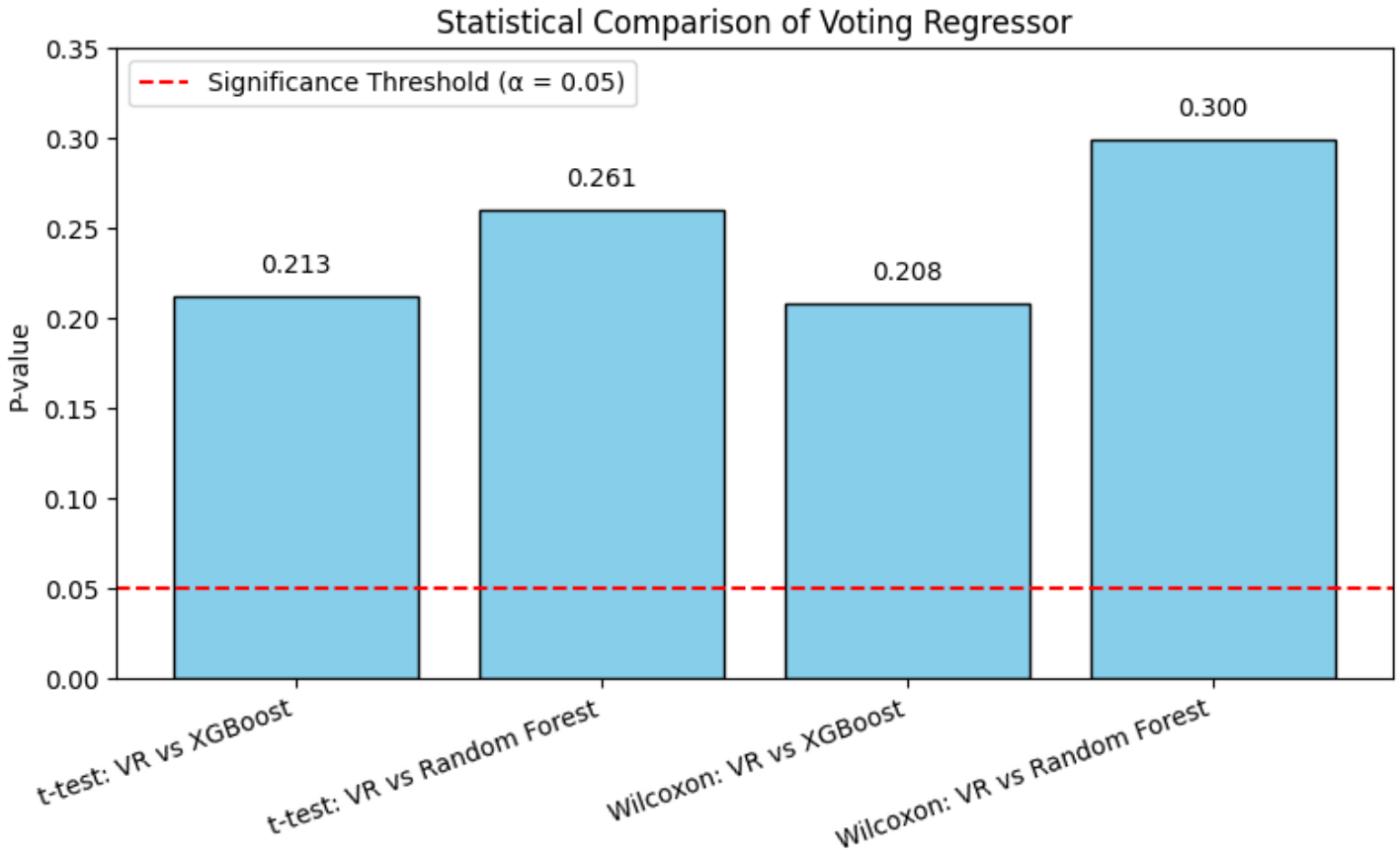

Figure 21: Statistical Comparison of voting regressor (case of brazil)

3.3 Discussion

Our findings demonstrate the competitive performance of regression models applied to epidemiological and climatic data related to chikungunya. With a low MAE and a comparatively reduced RMSE, the Voting Regressor ensemble model—which integrates the predictions of the Linear Regressor, Random Forest Regressor, and XGBoost Regressor—outperformed the individual models overall. In the case of Paraguay, the Voting Regressor achieved an MAE of 40.37, representing a clear improvement over the Linear Regressor, which yielded an MAE of 67.15. The RMSE of the Voting Regressor was 62.25, indicating strong predictive performance despite the complexity of

the epidemiological and climatic datasets. The R^2 score of 0.59 further shows that this model explains nearly 60% of the variance in the data for this region, highlighting the crucial role played by data augmentation techniques [48] and KNN-based imputation [47] in achieving these results.

For Brazil, the Voting Regressor exhibited comparable robustness, with an R^2 score of 0.65—higher than those obtained with the Random Forest and XGBoost regressors—and a lower RMSE (1387.23) than the other models. These findings suggest that the Voting Regressor effectively combines the strengths of individual algorithms to produce more balanced and reliable predictions.

In Chad, although the overall predictive performance of all models was lower, the Voting Regressor remained the best-performing approach, achieving an MAE of 50.52 and an RMSE of 93.02. The R^2 score of 0.282, the highest among the models tested for this country, indicates that the ensemble model better captures data variability despite the limitations associated with the relatively small dataset. Here again, data augmentation and KNN imputation strategies played a critical role in improving model performance.

In this context, the reduced predictive accuracy observed in some settings, such as Chad, may be partly attributed to the limited availability of explanatory variables, underscoring the need for a more comprehensive integration of climatic and epidemiological predictors. Moreover, particular attention should be paid to the effects of meteorological factors, including humidity and temperature. For example, while higher humidity appears to be associated with increased chikungunya incidence, rising temperatures may be linked to a decline in reported cases. These seemingly paradoxical relationships provide valuable perspectives for refining future predictive models.

4 Conclusion

The main objective of this study was to develop a predictive model for chikungunya using ensemble regression approaches derived from artificial intelligence, by examining the influence of climatic variables on disease transmission in Brazil, Paraguay, and Chad through advanced regression techniques. Epidemiological data were obtained from the PAHO real-time surveillance platform (for Paraguay), WHO reports (for Chad), and the Mendeley database (for Brazil), while climatic data were collected from reliable sources such as the Weather and Climate website. The models selected for this study included the Random Forest Regressor and the XGBoost Regressor optimized via grid search, as well as an ensemble model (Voting Regressor) combining Linear Regression, Random Forest Regressor, and the optimized XGBoost Regressor. Among these approaches, the Voting Regressor ensemble model achieved the best overall performance, yielding the lowest MAE, relatively low RMSE, and satisfactory predictive accuracy (65% for Brazil, 28.2% for Chad, and 59.97% for Paraguay). At the 5% significance level, the paired t-test and the Wilcoxon signed-rank test revealed no statistically significant differences between the Voting Regressor and XGBoost ($p = 0.2126$ and $p = 0.2081$, respectively) or Random Forest ($p = 0.2607$ and $p = 0.2997$, respectively). Although the Voting Regressor exhibited slightly improved error metrics, this advantage could not

be demonstrated statistically. Nevertheless, its consistent performance across different countries suggests that ensemble strategies are effective for stabilizing predictions in heterogeneous epidemiological and climatic contexts. The limitations of this study indicate that climatic variables alone are insufficient to fully explain the observed variability in chikungunya incidence across the investigated countries. Future research should therefore incorporate additional environmental and socio-ecological factors and develop hybrid models that combine multiple machine learning algorithms to further refine forecasting accuracy. Recommendations include strengthening data collection systems, adopting intervention strategies adapted to climatic variability, and enhancing community awareness of hygiene practices and chikungunya prevention. Furthermore, to ensure that methodological improvements are meaningful, the systematic application of statistical validation procedures—such as the paired t-test and the Wilcoxon signed-rank test—remains essential.

Acknowledgment

The authors express their sincere gratitude to the King of Skolabwandjan for his invaluable support and continuous assistance, which greatly contributed to the successful completion of this work.

Funding

Not Applicable

Availability of data and materials

Data used in this work are available from the corresponding author on a reasonable request.

Declarations

Disclosure statement : The authors report there are no competing interests to declare.

Conflicts of interest : The authors do not have any conflict or competing interests.

References

[1] Pan American Health Organization. Chikungunya - PAHO/WHO — Pan American Health Organization. https://www.paho.org/en/topics/chikungunya, Accessed 12/June/2024.

[2] Thomas E Morrison. Reemergence of chikungunya virus. *Journal of virology*, 88(20):11644–11647, 2014.

[3] European Centre for Disease Prevention and Control . Chikungunya worldwide overview. https://www.ecdc.europa.eu/en/chikungunya-monthly, Accessed 10/June/2024.

[4] Christophe N Peyrefitte, Dominique Rousset, Boris AM Pastorino, Regis Pouillot, Maël Bessaud, Fabienne Tock, Helene Mansaray, Olivier L Merle, Aurelie M Pascual, Christophe Paupy, et al. Chikungunya virus, cameroon, 2006. *Emerging infectious diseases*, 13(5):768–771, May 2007.

[5] José V.J. Silva, Louisa F. Ludwig-Begall, Edmilson F. de Oliveira-Filho, Renato A.S. Oliveira, Ricardo Durães-Carvalho, Thaísa R.R. Lopes, Daisy E.A. Silva, and Laura H.V.G. Gil. A scoping review of chikungunya virus infection: epidemiology, clinical characteristics, viral co-circulation complications, and control. *Acta Tropica*, 188:213–224, 2018.

[6] Vaishnavi K Ganesan, Bin Duan, and St Patrick Reid. Chikungunya virus: pathophysiology, mechanism, and modeling. *Viruses*, 9(12):368, 2017.

[7] World Health Organization. Chikungunya. https://www.who.int/health-topics/chikungunya#tab=tab_1, Accessed 10/june/2024.

[8] Heidi Auerswald, Camille Boussioux, Saraden In, Sokthearom Mao, Sivuth Ong, Rekol Huy, Rithea Leang, Malen Chan, Veasna Duong, Sowath Ly, et al. Broad and long-lasting immune protection against various chikungunya genotypes demonstrated by participants in a cross-sectional study in a cambodian rural community. *Emerging microbes & infections*, 7(1):1–13, 2018.

[9] Hamadjam Abboubakar, Albert Kouchéré Guidzavaï, Joseph Yangla, Irépran Damakoa, and Ruben Mouangue. Mathematical modeling and projections of a vector-borne disease with optimal control strategies: A case study of the chikungunya in chad. *Chaos, Solitons & Fractals*, 150:111197, 2021.

[10] Abdulazeez AlSajri AlSajri and Nour Hariry. The future of surgery: Robotic-assisted procedures and their impact. *SHIFAA*, 2024:52–55, Mar. 2024.

[11] Kani Djoulde, Boukar Ousman, Abboubakar Hamadjam, Laurent Bitjoka, and Clergé Tchiegang. Classification of pepper seeds by machine learning using color filter array images. *Journal of Imaging*, 10(2):41, 2024.

[12] Francis Yongwa Dtissibe, Ado Adamou Abba Ari, Hamadjam Abboubakar, Arouna Ndam Njoya, Alidou Mohamadou, and Ousmane Thiare. A comparative study of machine learning and deep learning methods for flood forecasting in the far-north region, cameroon. *Scientific African*, 23:e02053, 2024.

[13] Lukas Fischer, Lisa Ehrlinger, Verena Geist, Rudolf Ramler, Florian Sobiezky, Werner Zellinger, David Brunner, Mohit Kumar, and Bernhard Moser. Ai system engineering—key challenges and lessons learned. *Machine Learning and Knowledge Extraction*, 3(1):56–83, 2020.

[14] Benjamin Garga, Hamadjam Abboubakar, Rodrigue Saoungoumi Sourpele, David Libouga Li Gwet, and Laurent Bitjoka. Pollen grain classification using some convolutional neural network architectures. *Journal of Imaging*, 10(7):158, 2024.

[15] George Benneh Mensah, Maad M. Mijwil, Mostafa Abotaleb, Guma Ali, and Pushan Kumar Dutta. High performance medicine: Involving artificial intelligence models in enhancing medical laws and medical negligence matters a case study of act, 2009 (act 792) in ghana. *SHIFAA*, 2025:1–6, Jan. 2025.

[16] Esaie Naroum, Ebenezer Maka Maka, Hamadjam Abboubakar, Paul Dayang, Appolinaire Batoure Bamana, Benjamin Garga, Hassana Daouda Daouda, Mohsen Bakouri, and Ilyas Khan. Comparative analysis of deep learning and machine learning techniques for forecasting new malaria cases in cameroon's adamaoua region. *Intelligence-Based Medicine*, page 100220, 2025.

[17] James Phoenix and Mike Taylor. *Prompt Engineering for generative AI. "* O'Reilly Media, Inc.", 2024.

[18] Chen Wang, Ling-han Song, Zhou Yuan, and Jian-sheng Fan. State-of-the-art ai-based computational analysis in civil engineering. *Journal of Industrial Information Integration*, 33:100470, 2023.

[19] Kirstin Roster, Colm Connaughton, and Francisco A Rodrigues. Machine-learning–based forecasting of dengue fever in brazilian cities using epidemiologic and meteorological variables. *American Journal of Epidemiology*, 191(10):1803–1812, 2022.

[20] Kirstin Roster, Colm Connaughton, and Francisco A Rodrigues. Predicting dengue fever in brazilian cities. *bioRxiv*, pages 2021–02, 2021.

[21] Lucas M Stolerman, Pedro D Maia, and J Nathan Kutz. Forecasting dengue fever in brazil: An assessment of climate conditions. *PloS one*, 14(8):e0220106, 2019.

[22] Clarisse Lins de Lima, Ana Clara Gomes da Silva, Giselle Machado Magalhães Moreno, Cecilia Cordeiro da Silva, Anwar Musah, Aisha Aldosery, Livia Dutra, Tercio Ambrizzi, Iuri VG Borges, Merve Tunali, et al. Temporal and spatiotemporal arboviruses forecasting by machine learning: a systematic review. *Frontiers in Public Health*, 10:900077, 2022.

[23] Cecilia Cordeiro da Silva, Clarisse Lins de Lima, Ana Clara Gomes da Silva, Giselle Machado Magalhães Moreno, Anwar Musah, Aisha Aldosery, Livia Dutra, Tercio Ambrizzi, Iuri Valério Graciano Borges, Merve

Tunali, et al. Forecasting dengue, chikungunya and zika cases in recife, brazil: a spatio-temporal approach based on climate conditions, health notifications and machine learning. *Research, Society and Development*, 10(12), 2021.

[24] Cecilia Cordeiro da Silva, Clarisse Lins de Lima, Ana Clara Gomes da Silva, Giselle Machado Magalhães Moreno, Anwar Musah, Aisha Aldosery, Livia Dutra, Tercio Ambrizzi, Iuri VG Borges, Merve Tunali, et al. Spatiotemporal forecasting for dengue, chikungunya fever and zika using machine learning and artificial expert committees based on meta-heuristics. *Research on Biomedical Engineering*, 38(2):499–537, 2022.

[25] Clarisse Lins de Lima, Ana Clara Gomes da Silva, Cecilia Cordeiro da Silva, Giselle Machado Magalhães Moreno, Abel Guilhermino da Silva Filho, Anwar Musah, Aisha Aldosery, Livia Dutra, Tercio Ambrizzi, Iuri Valério Graciano Borges, et al. Intelligent systems for dengue, chikungunya, and zika temporal and spatio-temporal forecasting: a contribution and a brief review. *Assessing COVID-19 and other pandemics and epidemics using computational modelling and data analysis*, pages 299–331, 2022.

[26] Melissa Baptista and et al. Predictable chikungunya infection dynamics in brazil. *PLoS Neglected Tropical Diseases*, 2022.

[27] Ana Silva and et al. Spatial and temporal dynamics of chikungunya incidence in brazil. *Diseases*, 2024.

[28] Jane P. Messina and et al. The overlapping global distribution of dengue, chikungunya, zika and yellow fever. *Nature Communications*, 2025.

[29] Tania Carvajal and et al. A nationwide joint spatial modelling of simultaneous epidemics of dengue, chikungunya and zika in colombia. *BMC Infectious Diseases*, 2025.

[30] Sadie J. Ryan and et al. Priorities for modelling arbovirus transmission under climate change. *Trends in Ecology & Evolution*, 2025.

[31] Guilherme Salles and et al. Synthetic data generation methods in healthcare: A review on open challenges and future perspectives. *Patterns*, 2024.

[32] Vinicius Maragno and et al. Overview of data augmentation techniques in time series analysis. Preprint, 2024.

[33] Erin A. Mordecai and et al. Climate predicts geographic and temporal variation in mosquito-borne disease dynamics on two continents. *Nature Communications*, 11(1):1–13, 2020.

[34] Lung-Chang Chien and et al. Long-term effects of climate factors on dengue fever over a 40-year period. *BMC Public Health*, 2024.

[35] Abdallah M. Samy and et al. Effects of enso and dipole mode index on chikungunya incidence. *Scientific Reports*, 2020.

[36] World Health Organization. Using climate to predict infectious disease outbreaks: A review. Technical report, WHO, 2004.

[37] Jesse K. Davis and et al. Improving the prediction of arbovirus outbreaks: a comparison of climate-driven models for west nile virus. *Acta Tropica*, 185:242–250, 2018.

[38] Multiple authors. Assessing dengue forecasting methods: A comparative study of case-only versus climate-augmented models. medRxiv preprint, 2024.

[39] Organisation Mondiale de la Santé. *Évaluation externe conjointe des principales capacités Règlement sanitaire international du Tchad: rapport de mission, 24-28 juillet 2023.* Organisation Mondiale de la Santé, 2024.

[40] William M. de Souza, Guilherme S. Ribeiro, Shirlene T.S. de Lima, Ronaldo de Jesus, Filipe R.R. Moreira, Charles Whittaker, Maria Anice M. Sallum, Christine V.F. Carrington, Ester C. Sabino, Uriel Kitron, Nuno R. Faria, and Scott C. Weaver. Chikungunya: a decade of burden in the americas. *The Lancet Regional Health - Americas*, 30:100673, 2024.

[41] Iasmim Ferreira de Almeida, Claudia Torres Codeço, Raquel Martins Lana, Leonardo Soares Bastos, Sara de Souza Oliveira, Danielle Andreza da Cruz Ferreira, Vinicius Barbosa Godinho, Thais Irene Souza Riback, Oswaldo Gonçalves Cruz, and Flavio Codeço Coelho. The expansion of chikungunya in brazil. *The Lancet Regional Health–Americas*, 25, 2023.

[42] Iasmim Ferreira de Almeida, Claudia Torres Codeço, Raquel Martins Lana, Leonardo Soares Bastos, Sara de Souza Oliveira, Danielle Andreza da Cruz Ferreira, Vinicius Barbosa Godinho, Thais Irene Souza Riback, Oswaldo Gonçalves Cruz, and Flavio Codeço Coelho. The expansion of chikungunya in brazil. *The Lancet Regional Health–Americas*, 25, 2023.

[43] Marta Giovanetti, Cynthia Vazquez, Mauricio Lima, Emerson Castro, Analia Rojas, Andrea Gomez de la Fuente, Carolina Aquino, Cesar Cantero, Fatima Fleitas, Juan Torales, et al. Rapid epidemic expansion of chikungunya virus east/central/south african lineage, paraguay. *Emerging Infectious Diseases*, 29(9):1859, 2023.

[44] Marta Giovanetti, Cynthia Vazquez, Mauricio Lima, Emerson Castro, Analia Rojas, Andrea Gomez de la Fuente, Carolina Aquino, Cesar Cantero, Fatima Fleitas, Juan Torales, et al. Rapid epidemic expansion of chikungunya virus east/central/south african lineage, paraguay. *Emerging infectious diseases*, 29(9):1859, 2023.

[45] Thomás Oliveira, Igor Teixeira, Leonides Medeiros Neto, Theo Lynn, Sebastião Silva Neto, Vanderson Sampaio, and Patricia Endo. Arbovirus clinical data, brazil, 2013–2020. https://doi.org/10.17632/2d3kr8zynf.2, 2021. Mendeley Data, V2.

[46] Turki Aljrees. Improving prediction of cervical cancer using knn imputer and multi-model ensemble learning. *Plos one*, 19(1):e0295632, 2024.

[47] Oliver Kramer and Oliver Kramer. K-nearest neighbors. *Dimensionality reduction with unsupervised nearest neighbors*, pages 13–23, 2013.

[48] Loris Nanni, Michelangelo Paci, Sheryl Brahnam, and Alessandra Lumini. Comparison of different image data augmentation approaches. *Journal of imaging*, 7(12):254, 2021.

[49] Abdulaziz Aborujilah, Rasheed Mohammad Nassr, Tawfik Al-Hadhrami, Mohd Nizam Husen, Nor Azlina Ali, Abdulaleem Al Othmani, and Mustapha Hamdi. Comparative study of smote and bootstrapping performance based on predication methods. In *International conference of reliable information and communication technology*, pages 3–9. Springer, 2020.

[50] Carlton Reid Turner, Alfonso Fuggetta, Luigi Lavazza, and Alexander L Wolf. Feature engineering [software development]. In *Proceedings Ninth International Workshop on Software Specification and Design*, pages 162–164. IEEE, 1998.

[51] Carlton Reid Turner, Alfonso Fuggetta, Luigi Lavazza, and Alexander L Wolf. A conceptual basis for feature engineering. *Journal of Systems and Software*, 49(1):3–15, 1999.

[52] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[53] Erwan Scornet, Gérard Biau, and Jean-Philippe Vert. Consistency of random forests. *The Annals of Statistics*, 43(4), 2015.

[54] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[55] Kun An and Jiang Meng. Voting-averaged combination method for regressor ensemble. In *International Conference on Intelligent Computing*, pages 540–546. Springer, 2010.

[56] Shikun Chen and Wenlong Zheng. Rrmse-enhanced weighted voting regressor for improved ensemble regression. *PloS one*, 20(3):e0319515, 2025.