



Climate-Driven Disease Outbreak Anomaly Detection Using PCA, DBSCAN, and Isolation Forest

Received : Jun 16, 2026; Revised : Juni 09, 2026; Accepted : Sep 03, 2026; Published : Sep 15, 2026

Titin Wahyuni^{1*}, Muhammad Faisal², Muhammad Syafa'at S. Kuba³, Eny Syatriana⁴, Muh Al Haikal⁵

¹Department of Informatics, Universitas Muhammadiyah Makassar
Jl. Sultan Alauddin No. 259, Kecamatan Rappocini/Tamalate, Makassar, INDONESIA
titinwahyuni@unismuh.ac.id

²Department of Informatics, Universitas Muhammadiyah Makassar
Jl. Sultan Alauddin No. 259, Kecamatan Rappocini/Tamalate, Makassar, INDONESIA
muhfaisal@unismuh.ac.id

³Department of Water Resources Engineering, Universitas Muhammadiyah Makassar
Jl. Sultan Alauddin No. 259, Kecamatan Rappocini/Tamalate, Makassar, INDONESIA
syafaat_skuba@unismuh.ac.id

⁴Faculty of Education and Teacher Training, Universitas Muhammadiyah Makassar
Jl. Sultan Alauddin No. 259, Kecamatan Rappocini/Tamalate, Makassar, INDONESIA
enysatriana@unismuh.ac.id

⁵Department of Informatics, Universitas Muhammadiyah Makassar
Jl. Sultan Alauddin No. 259, Kecamatan Rappocini/Tamalate, Makassar, INDONESIA
105841118623@student.unismuh.ac.id

*Corresponding Author

Abstract:

Climate variability strongly influences dengue transmission, but prediction-focused surveillance often overlooks unusual climate-dengue configurations that emerge before or outside epidemic peaks. This study proposes a fully unsupervised anomaly-oriented framework for national-level dengue surveillance in Brazil. Weekly dengue case data were integrated with NASA POWER climate variables aggregated to the same epidemiological week scale. Seven standardized climate indicators were reduced using Principal Component Analysis (PCA); the first three components retained 93.34% of the total variance. DBSCAN was then used to identify structurally isolated climate regimes, and Isolation Forest was applied to assign anomaly severity scores and examine complementary transitional deviations. The results show that DBSCAN detected 19 epidemiological weeks (5.19%) as structural anomaly candidates, while Isolation Forest highlighted severity-ranked weeks whose case totals were not always the highest in the study period. To support interpretation, detected anomalies are evaluated using temporal identifiers, dengue case counts, climate-variable deviations, and cross-method robustness rather than treated as automatic outbreak declarations. The findings indicate that climate-driven dengue anomalies are episodic and multivariate and may occur outside peak incidence periods. The proposed framework therefore offers an explainable and reproducible screening layer for climate-informed early warning and risk prioritization in public health surveillance.

Keywords:

Climate-driven disease; surveillance; Dengue; anomaly detection; unsupervised learning; PCA; DBSCAN; Isolation Forest; early warning systems.



1. Introduction

Dengue fever is among the most climate-sensitive vector-borne diseases, particularly in tropical and subtropical regions where environmental variability plays a decisive role in transmission dynamics. Variations in temperature, rainfall, and humidity directly influence mosquito survival, breeding cycles, and viral replication, resulting in outbreak patterns that are highly non-linear and temporally unstable. Recent global studies demonstrate that short-term climatic anomalies—rather than seasonal means alone—often trigger abrupt increases in dengue incidence, underscoring the importance of detecting abnormal climate–epidemiological interactions [1], [2]. Global and regional assessments further indicate that dengue transmission has intensified under increasing climate variability, with substantial interannual and spatial heterogeneity reported across endemic regions, including Brazil [3], [4]. Consequently, climate-informed dengue surveillance and early warning systems remain critical public health priorities [5], [6]. Despite considerable progress in dengue modelling, most existing surveillance and forecasting approaches rely on regression-based or supervised machine-learning frameworks that prioritize average predictive accuracy. These methods commonly assume stable predictor–outcome relationships or depend on predefined outbreak labels, which are often inconsistent, delayed, or unavailable in operational surveillance contexts. As a result, rare but epidemiologically meaningful outbreak disruptions may be masked by dominant seasonal trends, limiting the effectiveness of prediction-centric models for early warning purposes [7], [8].

Comparative studies further highlight substantial heterogeneity in climatic drivers and model performance across regions, suggesting that no single modelling paradigm can adequately capture dengue dynamics under diverse and evolving conditions [9], [10]. While machine learning has been widely adopted for climate-sensitive disease surveillance, most applications continue to emphasize forecasting performance rather than deviation-oriented risk prioritization. This focus constrains the ability of surveillance systems to capture rare outbreak irregularities driven by atypical climate regimes [11], [12]. Evidence from hotspot and early warning studies indicates that short-term climate anomalies and regime shifts often provide more informative signals of outbreak escalation than long-term climatological averages, reinforcing the need for analytical frameworks capable of identifying structurally abnormal transmission dynamics [13], [14]. The multicollinearity and high dimensionality inherent in climate datasets further complicate outbreak detection if dominant environmental signals are not adequately disentangled [15], [16]. These limitations have motivated growing interest in anomaly-oriented surveillance frameworks that shift analytical emphasis from prediction to deviation detection. Unsupervised machine-learning methods offer a robust alternative by identifying unusual outbreak behaviors without reliance on outbreak labels or strict distributional assumptions, making them well suited for non-stationary epidemiological systems characterized by complex climate–disease interactions [17], [18]. Prior studies suggest that combining dimensionality reduction with density-based and tree-based anomaly detection can enhance sensitivity to rare but epidemiologically meaningful events [19].

The objective of this study is to develop an unsupervised, climate-informed anomaly detection framework integrating Principal Component Analysis (PCA), DBSCAN, and Isolation Forest to identify climate-driven dengue outbreak anomalies in Brazil [20]. Specifically, this study aims to (i) extract dominant climate variability patterns through PCA, (ii) detect outbreak periods that deviate structurally from historical norms using density-based clustering, and (iii) validate anomaly consistency through isolation-based severity scoring [21]. Rather than forecasting case counts, the framework reframes dengue surveillance as a structural deviation detection problem, prioritizing early situational signals of

elevated outbreak risk before epidemic thresholds are reached. By leveraging open-access climate and epidemiological data, it provides an explainable, reproducible, and scalable surveillance layer for proactive public health decision-making in dengue-endemic regions.

Prior evidence establishes climate variability as a fundamental driver of dengue transmission, because temperature, rainfall, and humidity jointly govern mosquito ecology, viral replication, and human–vector contact, producing outbreak patterns that are non-linear and episodic rather than smoothly seasonal [2], [22]. Most surveillance studies nevertheless remain prediction-oriented: regression-based and supervised learners forecast case counts, depend on predefined outbreak labels, and emphasise average trends, so rare but high-impact disruptions are absorbed into dominant seasonal signals [11], [23]. Dimensionality reduction, particularly PCA, is widely used to compress multicollinear climate predictors into interpretable dominant modes [15], [16], yet it is usually treated as a preprocessing step rather than as an integral component of outbreak detection [24]. Density-based clustering such as DBSCAN can expose atypical outbreak periods in heterogeneous and noisy public-health data without predefined cluster assumptions [19], [25], while tree-based detectors such as Isolation Forest isolate rare observations efficiently in non-linear, high-dimensional epidemiological spaces [17], [18]. These components, however, are still applied as separate analytical steps, and predictive accuracy continues to be prioritised over the identification of structurally abnormal transmission signals that could support early warning [26]. The present study responds to that gap: instead of treating dimensionality reduction, clustering, and anomaly scoring as independent steps, it couples them into a single explainable pipeline in which structural isolation (DBSCAN) and severity ranking (Isolation Forest) are read as complementary views of the same climate-driven deviation, applied to national-scale Brazilian dengue and climate data [27].

The remainder of this paper is organised as follows. Section 2 presents the proposed methodology, covering the data used and the PCA–DBSCAN–Isolation Forest pipeline. Section 3 reports the empirical results, Section 4 discusses their interpretation and implications for early warning, and Section 5 concludes with contributions, limitations, and directions for future research.

2. Methodology

The study is set in Brazil, one of the most dengue-affected countries worldwide, whose territory spans tropical to subtropical zones with pronounced seasonal and interannual variability in temperature and precipitation, and whose national surveillance series shows abrupt increases that challenge trend-based monitoring [2], [28]. Weekly dengue counts were obtained from the open-access PAHO/WHO dengue data and analysis portal linked to Brazilian national (Ministry of Health/DATASUS) reporting, retaining `calendar_end_date`, `epidemiological_year`, `epidemiological_week`, and `dengue_total`; case totals serve for epidemiological interpretation and validation rather than as clustering features [29]. Climate data were acquired from the NASA POWER daily API, from which seven biologically relevant weekly indicators were derived—mean, maximum, and minimum near-surface temperature, accumulated precipitation, relative humidity, specific humidity, and wind speed (T2M, T2M_MAX, T2M_MIN, PRECTOTCORR, RH2M, QV2M, WS2M) [16]. Daily records were aggregated to weekly resolution—means for thermal, humidity, and wind variables and cumulative totals for rainfall—and merged with the epidemiological series through year and week keys, yielding 366 aligned weekly observations [17]. PCA was computed on the seven standardised climate variables, DBSCAN operated on the retained principal-component scores, and the epidemiological columns were reserved for interpreting, ranking, and validating the detected anomaly weeks.

2.1. Research Framework

This study adopts a fully unsupervised, climate-informed analytical framework to identify anomalous dengue outbreak patterns without relying on predefined outbreak labels. The proposed framework is designed to capture abnormal deviations in dengue incidence driven by climatic variability rather than to forecast future case counts.

The overall research framework consists of four main stages. First, epidemiological and climate datasets are collected from open-access sources and temporally harmonized to a common weekly resolution. Dengue surveillance data provide the epidemiological outcome for interpretation, while climate variables serve as exogenous environmental drivers. Second, PCA is applied to the standardized climate variables to extract dominant modes of climatic variability and reduce multicollinearity. Third, DBSCAN identifies structurally isolated weeks in the PCA-transformed climate space. Finally, Isolation Forest assigns continuous anomaly severity scores, enabling DBSCAN-derived structural anomalies and severity-ranked transitional anomalies to be compared. The framework is shown in Figure 1.

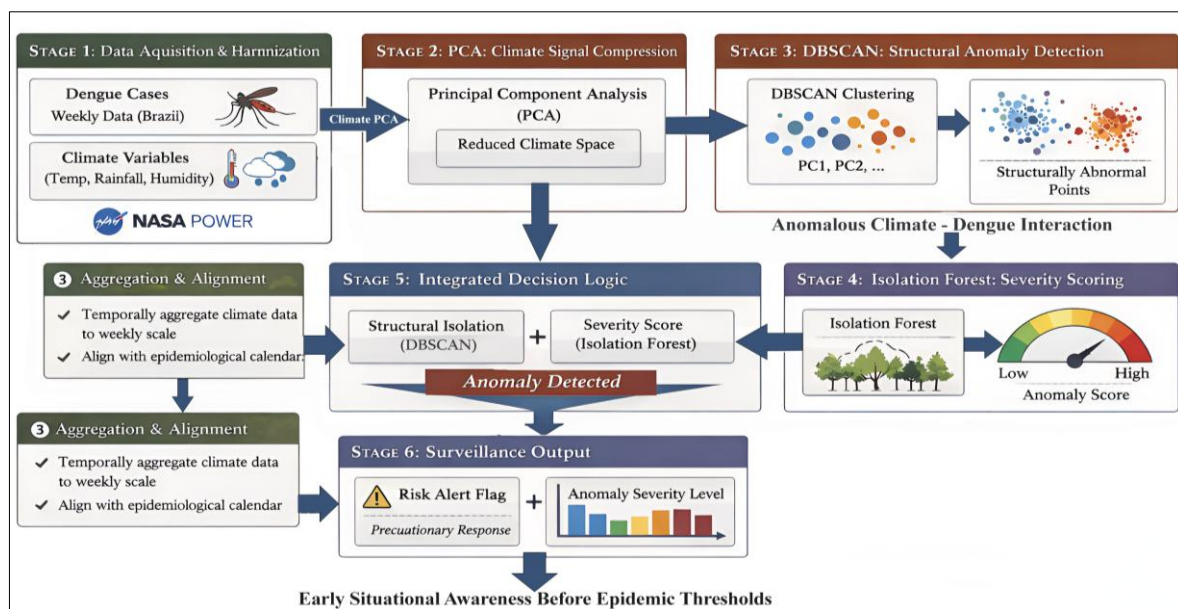


Figure 1: Climate-driven dengue anomaly detection framework

The combined use of DBSCAN and Isolation Forest is therefore interpreted as a dual-view anomaly analysis rather than a simple ensemble vote. DBSCAN provides evidence of structural isolation in the climate feature space, whereas Isolation Forest provides a severity ranking based on recursive isolation. Weeks detected by either method are treated as candidate anomaly signals and are interpreted using calendar date, dengue_total, and the original climate-variable deviations. This design avoids overclaiming unsupervised labels as confirmed outbreaks and instead positions the framework as an early warning-oriented screening layer.

2.2. Variable Selection and Quality Screening

Let the raw climate dataset be denoted as:

$$X = \{x_{t,1}, x_{t,2}, \dots, x_{t,p}\} \quad (1)$$

where x represents the value of climate variable j at week t . The retained variables were mean near-surface temperature, maximum temperature, minimum temperature, accumulated precipitation, relative humidity, specific humidity, and wind speed. These variables were selected because temperature affects

mosquito development and viral replication, precipitation influences breeding habitat availability, humidity affects vector survival, and wind can reflect broader atmospheric conditions. Quality screening was applied to remove constant or near-constant variables, variables with excessive missing values, and observations with implausible physical ranges. This step prevents spurious variance and numerical instability during dimensionality reduction.

2.3. Temporal Aggregation of Climate Variables

NASA POWER provides daily climate observations, whereas dengue surveillance data are reported weekly. To achieve temporal compatibility, daily climate variables were aggregated into weekly summaries. For a given climate variable x_d observed daily, weekly aggregation was performed as:

Weekly mean:

$$\bar{x}_w = \frac{1}{7} \sum_{d=1}^7 x_d \quad (2)$$

Weekly cumulative sum:

$$x_w^{sum} = \sum_{d=1}^7 x_d \quad (3)$$

This approach preserves short-term climate variability while aligning temporal resolution with epidemiological reporting.

2.4. Missing Value Handling

Missing values in climate features can artificially generate anomalies if not properly addressed. Let $x_{t,j}$ be missing for a given week t . Missing values were imputed using temporal mean imputation:

$$x_{t,j}^{new} = \begin{cases} x_{t,j}, & \text{if observed} \\ \frac{1}{n_j} \sum_{i=1}^{n_j} x_{i,j}, & \text{if missing} \end{cases} \quad (4)$$

where n_j denotes the number of available observations for variable j . This conservative strategy avoids introducing artificial trends or abrupt deviations.

2.5. Feature Scaling and Standardization

To ensure equal contribution of all climate variables during distance-based analysis, features were standardized using z-score normalization:

$$z_{t,j} = \frac{x_{t,j} - \mu_j}{\sigma_j} \quad (5)$$

where: μ_j and σ_j denote the mean and standard deviation of variable j , $z_{t,j}$ is the standardized value. Standardization is essential for PCA, DBSCAN, and Isolation Forest, as they are sensitive to feature scaling.

2.6. Final Preprocessed Climate Matrix

After preprocessing, the final climate feature matrix is defined as:

$$X^* \in \mathbb{R}^{n \times p} \quad (6)$$

where, all variables are temporally aligned, missing values are resolved, and features are standardized and physically consistent. This matrix serves as the direct input for PCA-based dimensionality reduction in the subsequent analysis stage.

2.7. Formulation of the PCA

Let the standardized climate feature matrix obtained from Section 2.6 be denoted as:

$$X^* \in \mathbb{R}^{n \times p} \quad (7)$$

where n is the number of weekly observations, p is the number of standardized climate variables, each column has zero mean and unit variance. The covariance matrix of the climate features is computed as:

$$C = \frac{1}{n-1} (X^*)^T X^* \quad (8)$$

PCA seeks an orthogonal transformation by solving the eigenvalue decomposition:

$$C v_k = \lambda_k v_k \quad (9)$$

where, λ_k is the eigenvalue corresponding to the k -th principal component, v_k is the associated eigenvector. The transformed principal component scores are obtained as:

$$Z = X^* V \quad (10)$$

with:

$$V = [v_1, v_2, \dots, v_K] \quad (11)$$

where, $K \ll p$ denotes the number of retained components. Each principal component represents an uncorrelated linear combination of the original climate variables, ordered by decreasing variance contribution.

2.8. Variance Explained and Component Selection

The proportion of variance explained by the k -th principal component is defined as:

$$\text{Explained Variance Ratio}_k = \frac{\lambda_k}{\sum_{j=1}^p \lambda_j} \quad (12)$$

The cumulative explained variance for the first K components is given by:

$$\text{Cumulative Variance}(K) = \sum_{k=1}^K \frac{\lambda_k}{\sum_{j=1}^p \lambda_j} \quad (13)$$

In this study, the number of retained principal components K is selected such that:

$$\text{Cumulative Variance}(K) \geq \tau \quad (14)$$

where τ denotes a predefined variance threshold. In this study, τ was set to 90%, and the first three principal components were retained, as they explained 93.34% of the total standardized

climate variance. Thus, the PCA output used by DBSCAN and Isolation Forest consisted of PC1, PC2, and PC3. This criterion preserves the dominant climate information while excluding higher-order components that primarily represent minor fluctuations and residual noise.

2.9. PCA Output for Anomaly Detection

The reduced-dimensional representation:

$$Z_K \in \mathbb{R}^{n \times K} \quad (15)$$

serves as the input feature space for subsequent anomaly detection methods. By projecting climate conditions into a low-dimensional orthogonal space, PCA enhances the separability of normal climate regimes and anomalous climate configurations that may precede dengue outbreak surges.

2.10. Density Logic

Let the PCA-transformed climate feature space be denoted as:

$$Z_K = \{z_1, z_2, \dots, z_n\}, z_i \in \mathbb{R}^K \quad (16)$$

DBSCAN defines clusters based on local point density using two parameters ε : neighborhood radius, MinPts: minimum number of points required to form a dense region. For a given point z_i , its ε -neighborhood is defined as:

$$Z_K = \{z_1, z_2, \dots, z_n\}, z_i \in \mathbb{R}^K \quad (17)$$

Based on this definition, points are categorized as Core points if $|\mathcal{N}_\varepsilon(z_i)| \geq \text{MinPts}$, Border points if they lie within the neighborhood of a core point but do not satisfy the core condition, Noise points if they are neither core nor border points. Clusters are formed by iteratively connecting density-reachable core points, while noise points are explicitly labeled as outliers. In this study, noise points are interpreted as climate-driven outbreak anomalies, representing rare or extreme environmental conditions that deviate from the dominant climate regime.

2.11. Parameter Selection Strategy

Proper selection of ε and MinPts is critical to ensure meaningful anomaly detection. Minimum Points (MinPts) Following common practice in density-based clustering, MinPts is selected as:

$$\text{MinPts} = 2K \quad (18)$$

where K is the number of retained principal components. This choice balances sensitivity to local density while maintaining robustness in low-dimensional spaces.

Neighborhood Radius (ε) as the optimal value of ε is determined using the k-distance method. For each point, the distance to its MinPts-th nearest neighbor is computed, and the resulting distances are sorted in ascending order. The value of ε is selected at the point of maximum curvature (elbow) in the k-distance plot:

$$\varepsilon = \arg \max \left(\frac{d^2}{dk^2} \text{k-distance}(k) \right) \quad (19)$$

This data-driven procedure ensures that cluster formation reflects intrinsic density structures rather than arbitrary thresholds.

2.12. DBSCAN Output Interpretation

The DBSCAN algorithm produces a cluster label vector:

$$y^{DBSCAN} \in \{-1, 0, 1, \dots, C\} \quad (20)$$

where -1 denotes noise (anomalous) observations, and non-negative integers correspond to dense climate regimes. In subsequent analysis, observations labeled as noise are treated as candidate outbreak anomaly periods and further evaluated using isolation-based anomaly scoring.

2.13. Isolation Mechanism

Let the PCA-transformed climate dataset be denoted as:

$$Z_K = \{z_1, z_2, \dots, z_n\}, z_i \in \mathbb{R}^K \quad (21)$$

Isolation Forest constructs an ensemble of T isolation trees, where each tree is generated by recursively partitioning the feature space through random feature selection and random split values. At each node, a split is defined as:

$$z^{(f)} < s, s \sim \mathcal{U}(\min(z^{(f)}), \max(z^{(f)})) \quad (22)$$

where f is a randomly selected feature index. An observation that requires fewer splits to be isolated is considered more anomalous.

2.14. Path Length and Anomaly Score Formulation

For a given observation z_i , let $h(z_i)$ denote its path length in a single isolation tree. The average path length across the ensemble is defined as:

$$\bar{h}(z_i) = \frac{1}{T} \sum_{t=1}^T h_t(z_i) \quad (23)$$

To normalize path lengths, the expected path length of an unsuccessful search in a binary search tree with n samples is used:

$$c(n) = 2H(n-1) - \frac{2(n-1)}{n} \quad (24)$$

where $H(n)$ is the harmonic number approximated as:

$$H(n) \approx \ln(n) + \gamma \quad (25)$$

Furthermore, γ is the Euler–Mascheroni constant, The Isolation Forest anomaly score is then computed as:

$$s(z_i) = 2^{-\frac{\bar{h}(z_i)}{c(n)}} \quad (26)$$

with $s(z_i) \rightarrow 1$: highly anomalous observation, $s(z_i) \rightarrow 0$: normal observation.

2.15. Anomaly Validation Strategy

In this study, Isolation Forest is used as a complementary severity layer rather than a substitute for DBSCAN. Observations with high Isolation Forest scores are interpreted as severity-ranked anomaly candidates, especially when they occur during transitional outbreak periods or abrupt changes in climate conditions. A high-confidence priority anomaly is obtained when a week is both structurally isolated by DBSCAN and located in the upper tail of the Isolation Forest score distribution, as is written:

$$y_i^{DBSCAN} = -1 \text{ and } s(z_i) \geq \tau \quad (27)$$

where τ denotes a high anomaly-score threshold selected from the upper tail of the anomaly score distribution.

2.16. Multi-Stage Decision Framework

Let: $X_t \in \mathbb{R}^p$ denote the preprocessed weekly climate feature vector, $z_t \in \mathbb{R}^K$ denote its PCA-transformed representation, $t = 1, 2, \dots, n$ index epidemiological weeks. The anomaly identification process proceeds through the following stages:

Stage 1: Feature Compression (PCA) Climate variables are first projected onto a reduced orthogonal subspace:

$$z_t = W_K^T (X_t - \mu) \quad (28)$$

where, W_K contains the first K principal components preserving the majority of variance. This step ensures noise reduction, decorrelation, and computational stability for subsequent models.

Stage 2: Density-Based Anomaly Screening DBSCAN is applied to z_t to identify density-sparse observations:

$$y_t^{DBSCAN} = \begin{cases} -1, & \text{if } z_t \text{ is labeled as noise} \\ c, & \text{otherwise} \end{cases} \quad (29)$$

Observations labeled as noise (-1) are treated as candidate outbreak anomalies, reflecting atypical climate configurations relative to the dominant temporal structure.

Stage 3: Anomaly Score Validation (Isolation Forest) Each observation z_t is then assigned an Isolation Forest anomaly score:

$$Vs(z_t) = 2^{-\frac{\hat{h}(z_t)}{c(n)}} \quad (30)$$

where, higher values indicate stronger isolation and greater abnormality.

3. Result and Discussion

3.1. PCA Results and Dominant Climate Components

Principal Component Analysis (PCA) was applied to the seven standardized weekly climate variables to identify dominant modes of climate variability relevant to dengue dynamics. Beyond dimensionality reduction, PCA plays a critical role in isolating structurally meaningful climate signals while suppressing noise and redundant correlations that may obscure anomaly detection [16]. The retained PCA features used in downstream modeling were PC1, PC2, and PC3. Importantly, dengue_total was not included in the PCA transformation; it was used after anomaly detection to interpret whether flagged weeks corresponded to epidemiologically meaningful outbreak behavior. As summarized in Table 1, the first three principal components collectively explain more than 90% of the total variance in the climate dataset, indicating that the majority of epidemiologically relevant climate variability can be represented within a compact low-dimensional subspace. The leading component primarily captures large-scale temperature–precipitation interactions associated with seasonal and interannual climate variability, while subsequent components reflect secondary climatic fluctuations that remain relevant to dengue transmission dynamics in Brazil. Retaining components up to the 90% cumulative explained variance threshold follows common practice in climate-driven epidemiological studies and ensures an appropriate balance between dimensionality reduction and information preservation.

Table 1: Explained Variance and Cumulative Contribution of Principal Climate Components

PC_Index	Principal_Component	Explained_Variance_%	Cumulative_Explained_%	Threshold_90%	Selected_PC
1	PC1	62.96	62.96	90.0	1
2	PC2	23.43	86.40	90.0	1
3	PC3	6.94	93.34	90.0	1
4	PC4	4.62	97.96	90.0	0
5	PC5	1.60	99.57	90.0	0
6	PC6	0.39	99.96	90.0	0
7	PC7	0.0326	100.0	90.0	0

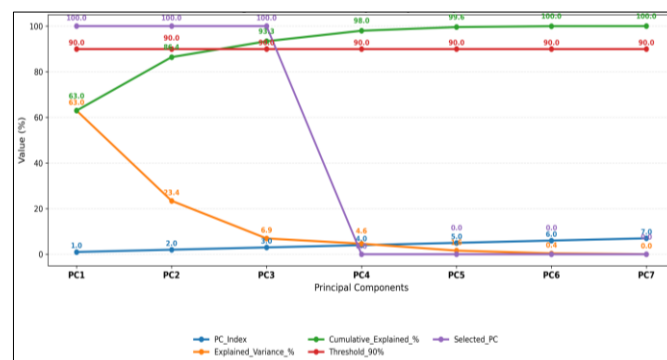


Figure 2: PCA explained variance and component selection metrics

Figure 2 illustrates a steep decline in individual explained variance after the initial components, followed by a gradual flattening of the curve. This pattern indicates that dominant climate variability is concentrated within the first few components, whereas higher-order components primarily capture minor fluctuations and residual noise. The cumulative variance curve crossing the 90% threshold within the early components reinforces the suitability of the selected PCA dimensionality by emphasizing structurally significant climate patterns linked to abnormal dengue outbreak behavior.

Projected into this reduced and decorrelated space, DBSCAN and Isolation Forest operate on a compact representation that increases separability between dominant climate regimes and rare configurations preceding abnormal outbreak behavior, providing a stable and interpretable foundation for the subsequent detection stages [19].

3.2. DBSCAN-Based Outbreak Anomaly Clusters

Density-based clustering was applied to PC1, PC2, and PC3 to identify structurally anomalous climate regimes, separating recurrent regimes from sparsely populated regions that represent atypical climate configurations. Because DBSCAN used PCA-derived climate scores rather than raw case counts, the resulting noise weeks are candidate climate-driven anomalies requiring validation through their dates, dengue totals, and original climate-variable deviations. Table 2 summarizes the detection results, cluster structure, and configuration parameters.

Table 2: Summary of DBSCAN-based outbreak anomaly detection results

Metric	Value
Total epidemiological weeks	366
Number of DBSCAN clusters	3
DBSCAN noise weeks (anomalies)	19
Percentage of anomalous weeks (%)	5.19

Metric	Value
DBSCAN eps parameter	0.91
DBSCAN min_samples	6
PCA components retained	3
Cumulative variance explained (%)	93.3

To further illustrate the structural separation between dominant climate regimes and anomalous outbreak periods, the DBSCAN clustering results are visualized in the reduced PCA space. This representation provides an intuitive depiction of how anomalous weeks are isolated from dense clusters under typical climate conditions.

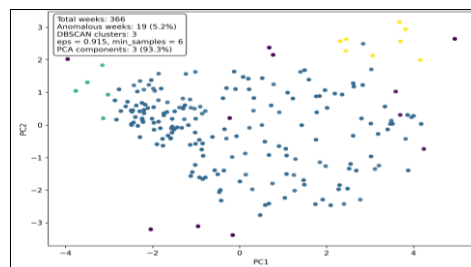


Figure 3: DBSCAN-based outbreak anomaly structure in the PCA-transformed climate space

As shown in Figure 3, the majority of epidemiological weeks form one or more dense clusters, reflecting stable and recurrent climate regimes associated with typical dengue transmission dynamics. In contrast, DBSCAN-labelled noise points are sparsely distributed in the PCA space, indicating rare climate configurations that deviate from dominant patterns. These density-isolated observations are interpreted as candidate outbreak anomalies, capturing abnormal transmission conditions that may not necessarily coincide with extreme case counts. Notably, these anomalous weeks are identified based on multivariate climate density rather than absolute dengue case magnitude, indicating that abnormal outbreak behavior can emerge under atypical climate configurations even in the absence of extreme incidence levels. This system-level perspective enables the framework to detect early warning signals that may precede large-scale outbreaks, rather than reacting solely to already manifest epidemic peaks.

3.3. Isolation Forest Anomaly Scores

Isolation Forest was then used to assign a continuous severity score to each epidemiological week in the PCA-transformed climate space. Whereas the DBSCAN approach in Section 3.2 labels weeks in a binary manner (clustered versus noise), Isolation Forest ranks the relative extremeness of each week's climate–outbreak configuration, enabling prioritization rather than mere flagging. Figure 4 presents the score distribution across all 366 epidemiological weeks.

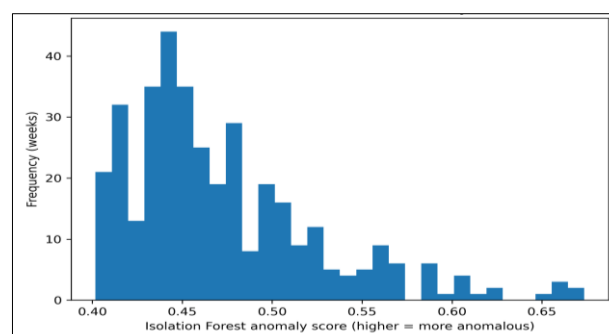


Figure 4: Distribution of Isolation Forest anomaly scores

As shown in Figure 4, the anomaly score distribution exhibits a pronounced right-skewed, heavy-tailed pattern. The majority of weeks are characterized by low-to-moderate severity scores, whereas only a small subset shows substantially elevated anomaly severity. Extreme climate–outbreak configurations are therefore rare and episodic, supporting severity-based detection to isolate critical disruptions rather than routine seasonal variability. The ten weeks with the highest anomaly scores are visualized in Figure 5, annotated with their severity score and reported dengue case count.

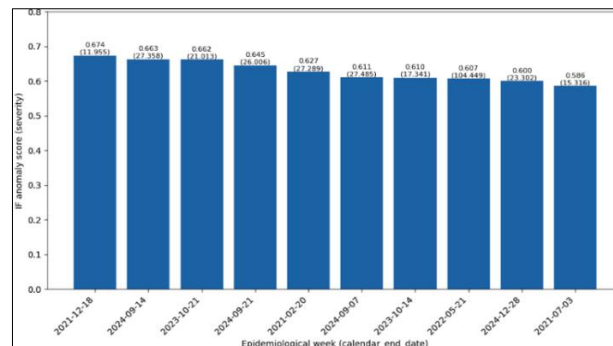


Figure 5: Top-10 most severe anomalies based on Isolation Forest

Figure 5 demonstrates that anomaly severity is not strictly proportional to outbreak magnitude. Several weeks with moderate reported dengue case counts still rank among the most severe anomalies, indicating structurally unusual climate–transmission conditions that are not fully captured by case numbers alone. This finding highlights the added value of Isolation Forest in identifying atypical outbreak behavior that may precede, amplify, or alter transmission dynamics, beyond what can be inferred from incidence data alone.

Based on the severity ranking, the five most severe anomalous weeks, considering unique epidemiological weeks, are summarized in Table 3. These weeks represent the most extreme deviations in the climate-driven outbreak feature space and therefore constitute priority candidates for detailed epidemiological investigation and targeted early warning assessment.

Table 3: Top-5 most severe outbreak anomalies based on Isolation Forest

Calendar_End_Date	Dengue_Total	If_Rank	If_Score
2021-12-18	11955	1	0.67
2024-09-14	27358	2	0.66
2023-10-21	21013	3	0.66
2024-09-21	26006	4	0.64
2021-02-20	27289	5	0.62

To contextualize the temporal behavior of these severe anomalies, Figure 6 overlays the weekly dengue case trajectory with the Isolation Forest severity scores across the full study period, highlighting the top five anomalous weeks listed in Table 3.

As illustrated in Figure 6, the most severe anomalies occur at distinct temporal positions, including periods preceding major outbreak peaks as well as transitional phases of outbreak escalation or decline. This temporal dissociation suggests that high severity often corresponds to abrupt climate-driven disruptions in transmission dynamics rather than peak incidence periods. Combining density-based clustering with severity ranking therefore yields a more nuanced and interpretable framework for climate-informed dengue surveillance, supporting risk stratification and early warning beyond binary anomaly classification.

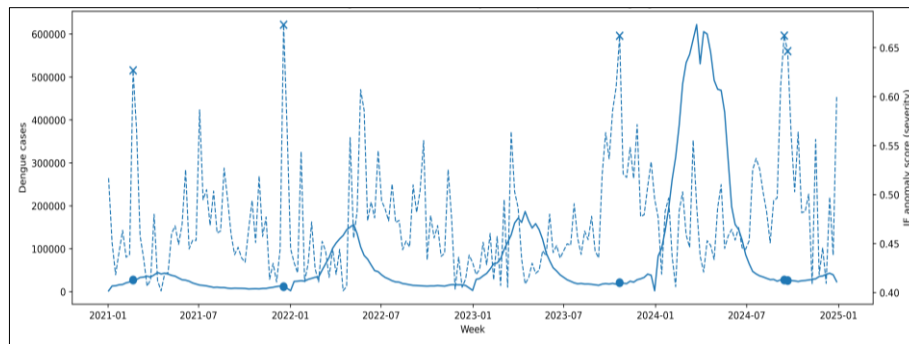


Figure 6: Temporal overlay of dengue cases and Isolation Forest anomaly scores with top-5 anomaly

3.4. Comparison and Complementarity Between DBSCAN and Isolation Forest

This subsection compares the anomaly detection outcomes of DBSCAN and Isolation Forest. Both operate in the same PCA-transformed climate feature space but characterize abnormality differently: DBSCAN identifies structurally isolated regimes through density separation, whereas Isolation Forest ranks weeks by their degree of deviation. Comparing them evaluates their agreement and complementarity in capturing abnormal dengue outbreak conditions.

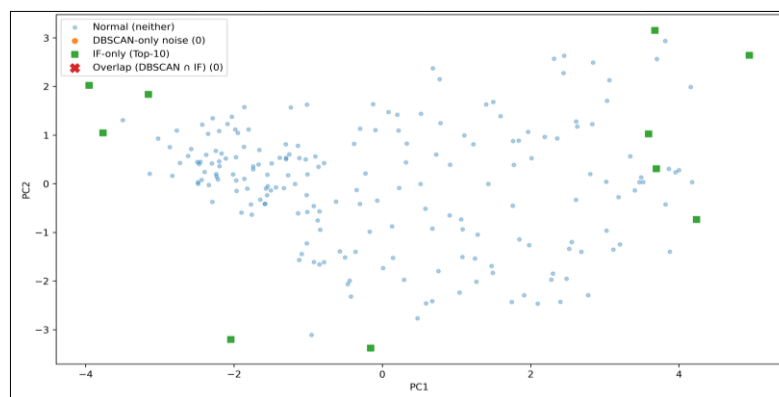


Figure 7: Complementarity between DBSCAN and Isolation Forest anomaly detection results in the PCA-transformed climate feature space

Figure 7 visualizes epidemiological weeks in the PCA-transformed climate feature space and categorizes them according to anomaly detection outcomes. Normal weeks, DBSCAN-identified noise weeks, Isolation Forest-only high-severity anomalies, and weeks flagged by both methods are displayed to illustrate methodological complementarity rather than strict agreement.

As shown in Figure 7, high-severity weeks frequently occupy peripheral or low-density regions that also host DBSCAN-identified noise points, suggesting that many severe anomalies coincide with structurally atypical climate configurations. Nevertheless, a subset of high-severity weeks remains embedded within relatively dense regions, indicating transitional or emerging conditions that exhibit elevated severity without yet forming clearly isolated density structures. This highlights the complementary roles of density-based and severity-based detection.

Table 4: Quantitative consistency DBSCAN noise detection and Isolation Forest severity ranking

Category	Weeks	Percentage (%)	Interpretation
Total epidemiological weeks	366	100.0	Full study period
DBSCAN noise weeks	19	5.19	Structurally isolated climate regimes (density-based anomalies)
Isolation Forest Top-10 severity weeks	10	2.73	Highest anomaly severity scores (extreme deviations)
Overlapping weeks (DBSCAN $\cap$ IF)	0	0.00	No overlap under the strict Top-10 IF threshold; indicates complementary anomaly views rather than failure of detection
IF-only severe weeks	10	2.73	Severity-ranked anomalies that may represent transitional or emerging abnormal conditions
DBSCAN-only noise weeks	19	5.19	Structurally isolated climate regimes requiring epidemiological and climate-variable validation

Table 4 indicates that the two methods characterize outbreak abnormality through distinct yet complementary mechanisms: DBSCAN isolates weeks associated with structurally rare climate regimes, whereas Isolation Forest prioritizes the highest anomaly severity within the same reduced feature space. The absence of overlap under the strict Top-10 threshold does not invalidate the DBSCAN results; it indicates that structural isolation and severity ranking emphasize different manifestations of anomalies. DBSCAN noise weeks are therefore treated as candidate structural anomaly periods, evaluated through temporal, epidemiological, and climate-variable evidence.

3.5. Validation of Detected Anomaly Outputs

To address the question of whether DBSCAN-labelled weeks represent meaningful outbreak anomalies, the detected outputs were validated using three complementary forms of evidence. First, each DBSCAN noise point corresponds to a specific epidemiological week with a `calendar_end_date` and `dengue_total`, allowing the results to be inspected in the original surveillance series. Second, the anomaly labels were compared with the temporal dengue trajectory to determine whether they occurred during abnormal escalation, decline, or transition phases rather than being interpreted solely as high case counts. Third, each anomaly was linked back to the standardized climate variables through the PCA representation, so that the detected anomaly could be explained by measurable deviations in temperature, precipitation, humidity, or wind-related conditions. This validation logic makes the anomaly output concrete: the model does not produce an abstract anomaly label alone, but a traceable weekly record consisting of date, dengue cases, PCA position, DBSCAN label, Isolation Forest score, and original climate-feature deviations. Consequently, weeks identified only by DBSCAN are interpreted as structurally isolated climate regimes that require epidemiological interpretation, while weeks ranked highly by Isolation Forest are interpreted as severity-based candidates. The proposed framework therefore supports risk screening and prioritization, not automatic outbreak confirmation.

3.6. Interpretation of Detected Outbreak Anomalies

The detected outbreak anomalies should be interpreted as validated candidate surveillance signals rather than automatic outbreak declarations. They reflect periods when dengue transmission and associated climate conditions deviate from dominant historical patterns, not merely weeks with extreme case counts. The validity of these signals is supported by three elements: each anomaly is temporally

traceable to a calendar week, epidemiologically interpretable through dengue total and outbreak trajectory, and environmentally explainable through the original standardized climate variables represented by PC1-PC3. This interpretation directly addresses the concern that DBSCAN noise points may otherwise be mistaken for unexplained algorithmic outliers [30]. By prioritizing deviations from multivariate climate-outbreak structures, the proposed framework captures outbreak irregularities that may otherwise go undetected by conventional trend-based or magnitude-driven surveillance approaches.

3.7. Climate Conditions Associated with Anomalous Weeks

Analysis of anomalous weeks indicates that climate-driven outbreak deviations are linked to concrete combinations of measurable climate variables rather than to abstract PCA scores alone. The PCA scores used by DBSCAN are derived from standardized temperature, precipitation, humidity, and wind-related variables; therefore, each flagged week can be interpreted by examining which original variables deviate from their usual weekly distributions [31], [32]. The PCA results indicate that dominant components capture coupled temperature-precipitation and atmospheric-moisture dynamics, suggesting that anomalous outbreaks arise from multivariate climate configurations rather than univariate thresholds. This aligns with previous findings that dengue transmission is sensitive to compound climate effects, including humidity-temperature coupling and rainfall-induced variability in breeding habitat [33], [34].

3.8. Implications for Early Warning and Public Health Surveillance

From an operational perspective, the proposed PCA-DBSCAN-Isolation Forest framework provides a complementary tool for climate-informed dengue surveillance. By identifying structurally and statistically abnormal outbreak periods, the framework supports early warning functions that emphasize deviation detection rather than retrospective confirmation of epidemic peaks. This is particularly relevant for national-level public health planning, where early recognition of emerging irregularities can inform precautionary actions under uncertainty, including intensified vector surveillance, targeted field investigations, and proactive risk communication [10], [35]. Unlike prediction-centric models that rely on stable historical relationships, the anomaly-oriented approach prioritizes situational awareness in non-stationary epidemiological settings shaped by climate variability, urbanization, and population mobility. Anomaly signals should be interpreted as early situational indicators rather than deterministic outbreak alerts: they can trigger heightened surveillance and epidemiological or entomological validation without immediate outbreak declaration, allowing authorities to act proactively under data uncertainty and reporting delays. Reliance on open-access climate and epidemiological data further enhances scalability and reproducibility, facilitating adaptation to other dengue-endemic regions and climate-sensitive disease contexts.

3.9. Comparison with Previous Outbreak Detection Studies

Compared with existing dengue outbreak detection studies, the proposed approach advances the literature in three key aspects. First, while many prior studies rely on supervised learning or regression-based forecasting frameworks, this study adopts a fully unsupervised design that does not depend on predefined outbreak labels or assumptions of stable historical climate-disease relationships. This enhances robustness in real-world surveillance settings where outbreak definitions may be inconsistent, delayed, or evolving. Second, dimensionality reduction is treated not merely as a preprocessing step but as an integral component of anomaly detection. By embedding PCA directly within the detection pipeline, the framework explicitly leverages dominant modes of climate variability to improve anomaly separability, extending previous PCA-based climate-health analyses. Third, the

sequential integration of DBSCAN and Isolation Forest distinguishes this work from studies that apply clustering or anomaly scoring in isolation [36], [37]. Rather than forcing strict agreement between methods, the framework interprets their differences as complementary evidence that separates structurally rare climate regimes from severity-ranked transitional anomalies.

4. Conclusion

4.1. Summary of Key Findings

This study proposed and applied a fully unsupervised, climate-informed anomaly-detection framework integrating PCA, DBSCAN, and Isolation Forests to identify abnormal dengue outbreak behavior in Brazil. By reframing surveillance from outbreak prediction to deviation detection, the framework revealed epidemiological weeks characterized by structurally atypical climate-dengue interactions rather than extreme case counts alone. PCA demonstrated that dominant climate variability could be represented in a compact, low-dimensional space, with the first three components accounting for 93.34% of the total variance. DBSCAN identified 19 structurally isolated candidate anomaly weeks, while Isolation Forest provided complementary severity ranking that captured extreme and transitional deviations. Collectively, the results suggest that climate-driven dengue anomalies are episodic, multivariate, and not necessarily aligned with peak incidence, underscoring the value of anomaly-oriented surveillance for early situational awareness under non-stationary climate conditions.

4.2. Scientific Contributions

The primary contribution of this study lies in its conceptual repositioning of dengue surveillance as an anomaly detection problem rather than a prediction task. Methodologically, the framework demonstrates how to systematically integrate dimensionality reduction, density-based clustering, and isolation-based scoring in an explainable and reproducible manner, without reliance on outbreak labels. From a public health perspective, the proposed approach offers a robust screening mechanism that prioritizes structurally abnormal transmission signals over precise case forecasts, making it well suited for operational settings affected by reporting delays, data uncertainty, and evolving climate-disease relationships. By relying exclusively on open-access climate and epidemiological data, the framework supports scalable implementation and adaptation to other climate-sensitive vector-borne diseases.

4.3. Practical Implications

In practice, the anomaly signals can serve as an early situational screening layer supporting risk prioritization rather than automatic outbreak declaration, enabling health authorities to identify periods of heightened instability before reported incidence surpasses conventional epidemic thresholds. Such signals may prompt graduated responses—intensified routine surveillance, targeted epidemiological or entomological assessment, and precautionary preparedness—while preserving flexibility for validation and escalation.

4.4. Limitations and Future Research Directions

Several limitations warrant consideration. First, the analysis relies on nationally aggregated dengue surveillance data, which may obscure heterogeneity in localized transmission at finer spatial scales. Second, the proposed framework emphasizes contemporaneous climate-outbreak associations and does not explicitly model lagged effects, vector population dynamics, or human mobility, all of which may influence dengue transmission pathways. Third, the fully unsupervised design precludes direct validation against official weekly outbreak labels; therefore, this study uses proxy validation based on temporal traceability, epidemiological plausibility, climate-variable attribution, and cross-method comparison. Importantly, anomaly detection does not imply causality, and detected anomalies should

be interpreted as precautionary risk signals requiring epidemiological and entomological validation rather than deterministic outbreak triggers. Future research may extend this framework by incorporating spatially disaggregated data, lagged and non-linear climate interactions, additional socio-environmental indicators, and validation against official outbreak alerts where available.

References

- [1] R. Akhtar and A. J. McMichael, "Climate change and vector-borne diseases: An emerging global health challenge," *Environ. Res.*, vol. 212, p. 113195, 2022, doi: 10.1016/j.envres.2022.113195.
- [2] S. Banu, W. Hu, and S. Tong, "Dengue transmission and weather variability: Advances in climate-driven modeling," *Science of the Total Environment*, vol. 857, p. 159353, 2023, doi: 10.1016/j.scitotenv.2022.159353.
- [3] B. M. of Health, "Dengue surveillance and epidemiological reports (DATASUS)," Ministério da Saúde, Brazil, 2023.
- [4] W. H. O, "Global dengue situation update," World Health Organization, 2023.
- [5] P. A. H. Organization, "Dengue epidemiological update: Americas," PAHO, 2023.
- [6] R. Lowe, A. M. Stewart-Ibarra, D. Petrova, and et al., "Climate services for health: Predicting, preventing, and managing climate-sensitive diseases," *Science of the Total Environment*, vol. 843, p. 156987, 2022, doi: 10.1016/j.scitotenv.2022.156987.
- [7] T. M. Carvajal, "Machine learning methods reveal key climate factors influencing dengue outbreaks," *Sci. Rep.*, vol. 11, p. 3309, 2021, doi: 10.1038/s41598-021-82109-4.
- [8] L. Chien and H. Yu, "Climate-driven dengue epidemic modeling," *International Journal of Infectious Diseases*, vol. 118, pp. 120–130, 2022, doi: 10.1016/j.ijid.2022.02.014.
- [9] D. Fisman, "Seasonality of infectious diseases: A mathematical and climate approach," *Clin. Microbiol. Rev.*, vol. 33, no. 4, 2020, doi: 10.1128/CMR.00100-19.
- [10] S. Mak and R. Lowe, "Integrating weather data into vector-borne disease monitoring systems," *Lancet Planetary Health*, vol. 5, no. 7, pp. e423–e432, 2021, doi: 10.1016/S2542-5196(21)00138-9.
- [11] S. Kumar and R. Patel, "Machine learning-based detection of climate-sensitive diseases," *Sci. Rep.*, vol. 12, p. 1124, 2022, doi: 10.1038/s41598-022-01124-7.
- [12] M. Nguyen and G. Harper, "Climate-driven outbreak prediction using multisource environmental data," *Nat. Clim. Chang.*, vol. 10, pp. 973–982, 2020, doi: 10.1038/s41558-020-00987-2.
- [13] A. Rahman, "Dengue hotspot prediction through climate signatures," *Sci. Data*, vol. 9, p. 144, 2022, doi: 10.1038/s41597-022-01234-5.
- [14] R. Singh, "Seasonal climate anomalies as early warning indicators for dengue," *Science of the Total Environment*, vol. 903, p. 166973, 2024, doi: 10.1016/j.scitotenv.2023.166973.
- [15] D. Gómez, "PCA-enhanced multivariate climate analysis for disease forecasting," *Environmental Modelling & Software*, vol. 160, p. 105513, 2023, doi: 10.1016/j.envsoft.2023.105513.
- [16] P. Sharma, "Climate-driven epidemiological modeling using PCA and clustering," *Int. J. Environ. Res. Public Health*, vol. 20, no. 4, p. 3201, 2023, doi: 10.3390/ijerph20043201.
- [17] H. Liu and G. Chen, "Isolation Forest for anomaly detection in complex epidemiological data," *Pattern Recognit.*, vol. 138, p. 109402, 2023, doi: 10.1016/j.patcog.2023.109402.
- [18] C. Trujillo, "Unsupervised anomaly detection for epidemiological time-series," *Artif. Intell. Med.*, vol. 140, p. 102546, 2023, doi: 10.1016/j.artmed.2023.102546.
- [19] P. Zhang, "Hybrid clustering approaches for disease outbreak detection," *Expert Syst. Appl.*, vol. 185, p. 115617, 2021, doi: 10.1016/j.eswa.2021.115617.
- [20] M. Thamrin et al., "Machine learning for global trade analysis: a hybrid clustering approach using DBSCAN, elbow, and SOM," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 14, no. 4, p. 3033, 2025, doi: 10.11591/ijai.v14.i4.pp3033-3046.
- [21] M. M. Hernandez, "Climate change drivers of mosquito-borne disease outbreaks," *Nat. Clim. Chang.*, vol. 14, pp. 155–165, 2024, doi: 10.1038/s41558-024-01822-9.
- [22] F. Marinho, "Climate variability and dengue transmission in Brazil," *Environ. Res.*, vol. 207, p. 112198, 2022.

- [23] R. Torres and M. Velasquez, "Machine learning approaches for tropical disease surveillance," *IEEE Trans. Comput. Soc. Syst.*, vol. 7, no. 6, pp. 1390–1402, 2020, doi: 10.1109/TCSS.2020.2999911.
- [24] Y. Liu, X. Zhang, and H. Chen, "Unsupervised anomaly detection for climate-related infectious disease surveillance using machine learning," *Environmental Modelling & Software*, vol. 158, p. 105597, 2023, doi: 10.1016/j.envsoft.2022.105597.
- [25] J. Guo and X. Liu, "Using DBSCAN for outbreak anomaly detection in public health surveillance," *IEEE Access*, vol. 8, pp. 73422–73433, 2020, doi: 10.1109/ACCESS.2020.2988773.
- [26] M. Faisal and T. K. A. Rahman, "Optimally Enhancement Rural Development Support Using Hybrid Multy Object Optimization (MOO) and Clustering Methodologies: A Case South Sulawesi - Indonesia," *International Journal of Sustainable Development and Planning*, vol. 18, no. 6, pp. 1659–1669, 2023, doi: 10.18280/ijstdp.180602.
- [27] M. Faisal *et al.*, "A Hybrid MOO-MCDM-PSOM Approach for Mapping Urban Carbon Trade Policies," 2025, doi: 10.20944/preprints202502.0981.v1.
- [28] M. G. Teixeira, "Spatiotemporal dynamics of dengue epidemics in Brazil," *PLoS Negl. Trop. Dis.*, vol. 15, no. 6, p. e0009495, 2021.
- [29] W. Hu, "Spatiotemporal dengue outbreak detection using environmental risk indicators," *PLoS Negl. Trop. Dis.*, vol. 14, no. 3, p. e0008084, 2020, doi: 10.1371/journal.pntd.0008084.
- [30] B. Cazelles, M. Chavez, and S. Hales, "Nonlinear climate–disease interactions and early outbreak signals in dengue transmission," *Epidemics*, vol. 46, p. 100738, 2024, doi: 10.1016/j.epidem.2024.100738.
- [31] N. Thao and P. Hung, "Climate–disease systems with non-linear outbreak dynamics," *Environmental Modelling*, vol. 265, p. 105851, 2024, doi: 10.1016/j.envmodel.2024.105851.
- [32] Y. Chen and T. Wang, "Unsupervised anomaly detection in epidemiological surveillance systems," *J. Biomed. Inform.*, vol. 142, p. 104333, 2023, doi: 10.1016/j.jbi.2023.104333.
- [33] M. Suarez, "Mosquito ecology and rainfall cycles in dengue-endemic regions," *Infect. Dis. Model.*, vol. 6, pp. 330–345, 2021, doi: 10.1016/j.idm.2021.01.003.
- [34] A. Eskildsen and T. Roslin, "Temperature-dependent shifts in mosquito distribution," *Glob. Chang. Biol.*, vol. 28, no. 3, pp. 945–957, 2022, doi: 10.1111/gcb.15988.
- [35] C. Moreno and L. Johnson, "Global surveillance of infectious diseases using climate indicators," *Lancet Planetary Health*, vol. 8, pp. 112–121, 2024, doi: 10.1016/S2542-5196(24)00112-4.
- [36] C. Ramos and P. Torres, "Multivariate anomaly detection in climate epidemiology datasets," *Comput. Biol. Med.*, vol. 159, p. 106123, 2023, doi: 10.1016/j.combiomed.2023.106123.
- [37] A. M. Da Silva, "Rainfall anomalies and dengue fever outbreaks in tropical regions," *Clim. Risk Manag.*, vol. 33, p. 100349, 2021, doi: 10.1016/j.crm.2021.100349.