# Computational Modeling and Artificial Intelligence for Infectious Disease Outbreak Risk Classification in Maceió, Alagoas, Brazil

**Samila Raphaela de Oliveira**[1], **Victor Luan de Lima Lemos**[1], **Tarsis Marinho de Souza**[1], **Cledja Karina Rolim da Silva**[1]

[1]Instituto Federal de Educação, Ciência e Tecnologia de Alagoas (IFAL)
Campus Arapiraca – Arapiraca – AL – Brasil

sro3@aluno.ifal.edu.br, vlll1@aluno.ifal.edu.br,
taris.souza@ifal.edu.br, cledja@ifal.edu.br

***Abstract.*** *Infectious diseases remain a major public health challenge and require effective strategies for surveillance, prevention, and control. In Brazil, despite the availability of epidemiological surveillance systems such as SINAN, the occurrence of outbreaks still reveals limitations in anticipating critical events. In this context, the early identification of risk patterns is essential to support timely public health action. This study analyzes and compares machine learning models for classifying outbreak risk using SINAN data from Maceió, Alagoas, with emphasis on five diseases: Hepatitis, Meningitis, Pertussis, Varicella, and Exanthematic Diseases. The proposed approach models risk as three levels (Normal, Attention, Outbreak) to support more interpretable and operationally useful surveillance outcomes.*

## 1. Introduction

Infectious diseases remain a major public health challenge, particularly in regions affected by socioeconomic inequality and limitations in health systems. Increasing human mobility and pathogen spread reinforce the need for timely public health responses [Jones et al. 2008, Balcan et al. 2009, Santangelo et al. 2023].

In Brazil, surveillance systems such as SINAN provide historical data on compulsory notification diseases, but are still mainly used for retrospective monitoring, limiting early intervention. Meanwhile, machine learning has expanded data-driven approaches in epidemiological surveillance, particularly for complex and temporally dependent patterns [Ludermir 2021, Santangelo et al. 2023].

Although epidemiological forecasting is commonly formulated as a regression problem, it may be sensitive to fluctuations, underreporting, and structural changes [Shaman and Karspeck 2012]. Therefore, this study models outbreak risk through a three-level classification approach to provide more interpretable results for surveillance.

The analysis focuses on Maceió, Alagoas, using historical SINAN data. Although the original dataset covered the entire state, the study was restricted to the municipality with the highest number of notifications to ensure a more consistent and interpretable time series. This work analyzes and compares machine learning models for outbreak risk classification to support epidemiological surveillance and public health decision-making.

## 2. Related Work

Studies by Shaman and Karspeck [2012], Santangelo et al. [2023], Zanardo et al. [2024], and Cabrera et al. [2022] explored infectious disease prediction through regression and classification approaches. While regression models estimate future case counts, they may be sensitive to fluctuations, delayed notifications, and heterogeneous epidemiological dynamics. Classification methods, on the other hand, identify risk levels and are more suitable for surveillance and early warning contexts [Du and Pang 2021, Gao et al. 2024]. In this scenario, *Logistic Regression* is commonly used as an interpretable baseline, whereas *Random Forest* and *Gradient Boosting* have shown strong performance in nonlinear problems [Breiman 2001, Friedman 2001].

In Brazil, databases such as DATASUS [Koike 2025] and SINAN provide important epidemiological data, although issues such as incomplete records, inconsistencies, and underreporting require careful preprocessing [Lima et al. 2021, Koike 2025, Silva et al. 2020, Borges et al. 2024]. Based on this literature, the present study adopts a classification-oriented approach to model outbreak risk in Maceió using historical SINAN data.

## 3. Proposed approach

This section presents the proposed approach, including data preparation, risk definition, modeling strategy, and evaluation criteria.

### 3.1. Data Source and Analytical Scope

The data used in this study were obtained from SINAN and initially covered the period from 2014 to 2024, comprising individual notifications of infectious diseases in the state of Alagoas. SINAN is one of the main epidemiological surveillance systems in Brazil, designed to register compulsory notification diseases and support control, prevention, and monitoring activities [Maia et al. 2019, Rocha et al. 2020, Souza and Silva 2020].

Although the original database contained records from the entire state, the analysis was restricted to the municipality of Maceió. The SINAN records included notification identifiers, temporal and geographic information, demographic variables, and disease-specific classification fields, but required filtering, transformation, and aggregation before being used in machine learning.

### 3.2. Data Preparation

Data preparation involved filtering, monthly aggregation, and feature construction. Records without valid notification dates were removed, only notifications from Maceió were retained, and disease-specific confirmation criteria were applied according to SINAN classification rules.

To reduce instability, only diseases with at least 100 confirmed cases were included: Hepatitis, Meningitis, Pertussis, Varicella, and Exanthematic Diseases. The period from March 2020 to May 2022 was excluded because the COVID-19 pandemic altered notification dynamics and reduced comparability over time [Borges et al. 2024, Silva et al. 2020].

The filtered records were aggregated by disease and month, with missing months filled with zero to preserve temporal continuity [Antunes and Cardoso 2015]. From these

series, lagged values, moving averages, monthly and annual variation, and cyclical temporal features were derived for classification.

Continuous features were standardized with `StandardScaler`, fit on the training set only. PCA was deliberately not used: the 17 engineered features are individually interpretable in surveillance terms (e.g., `VAR_ANUAL` reads as a year-over-year trend), preserve temporal structure, and are already small relative to the sample size.

### 3.3. Risk Class Definition

The target variable was not directly available in SINAN and was therefore constructed statistically. Following historical baseline analysis used in epidemiological surveillance [Antunes and Cardoso 2015], each month was classified according to the expanding historical mean and standard deviation up to the previous month. Three risk levels were defined:

- **Normal**: observed number of cases less than or equal to the historical mean;
- **Attention**: observed number of cases above the historical mean and up to one standard deviation above it;
- **Outbreak**: observed number of cases greater than the historical mean plus one standard deviation.

This formulation avoids future information leakage. Crucially, the intermediate *Attention* class turns surveillance into a graded response rather than a binary alarm, providing an operational trigger between passive monitoring and full outbreak response, which is the central methodological contribution of this work.

### 3.4. Modeling Strategy

Three supervised models were evaluated: *Logistic Regression*, *Random Forest*, and *Gradient Boosting*. *Logistic Regression* was used as an interpretable baseline, while the ensemble models were selected because they usually perform well in nonlinear epidemiological classification tasks [Du and Pang 2021, Gao et al. 2024].

The observations were ordered chronologically, using about 80% of the oldest records for training and the most recent 20% for testing. Because the risk classes were imbalanced, especially the *Outbreak* class, SMOTE was applied only to the training set to reduce bias toward the majority class [Wang et al. 2021].

Hyperparameters were tuned through grid search with temporal cross-validation. Macro F1-score was adopted as the optimization criterion because it balances performance across all classes [Ahmad et al. 2022, Scursone et al. 2025].

Experiments used Python 3.11 with `scikit-learn` [Pedregosa et al. 2011] and `imbalanced-learn` (`random_state=42`); the grid explored 108, 324, and 16 combinations for GB, RF, and LR. Code and data can be found in: `https://github.com/victorflll/csbc-outbreak-risks`.

### 3.5. Evaluation Criteria

The models were evaluated using *Accuracy*, *Macro F1-score*, and *Recall*, *Precision*, and *F1-score* for the *Outbreak* class. These metrics capture both global performance and the practical need to identify critical states.

Predictive *uncertainty* was incorporated through the normalized entropy of class probabilities. In addition, anomaly detection methods were used as a complementary analytical layer: Isolation Forest and Local Outlier Factor were applied to identify unusual temporal patterns without redefining the outbreak target [Liu et al. 2008, Breunig et al. 2000].

## 4. Results and Discussion

This section discusses the experimental results, including a comparative analysis of the models, feature importance, uncertainty interpretation, and the epidemiological risk forecasts for 2026.

### 4.1. Comparative Performance of the Models

The models showed distinct performance patterns. Random Forest achieved the highest Accuracy and Macro F1-score, indicating the best overall classification performance. However, it also presented the highest average uncertainty, suggesting lower confidence in its predictions.

Gradient Boosting provided the best balance between performance and reliability. Although its global metrics were slightly below those of Random Forest, it achieved the highest recall for the *Outbreak* class and the lowest mean uncertainty. This is especially relevant for surveillance, where failing to detect an outbreak is more critical than producing some false alarms.

Logistic Regression showed the weakest results, with lower Macro F1-score and lower recall for the *Outbreak* class. This suggests that a linear model was not sufficient to capture the nonlinear temporal patterns of the epidemiological series, which is consistent with the stronger performance of ensemble methods [Breiman 2001, Friedman 2001, Hastie et al. 2009], see Table 1.

**Table 1. Performance of the evaluated models.**

| Model | Acc. | $F1_M$ | $Rec._O$ | $Prec._O$ | Unc. |
|---|---|---|---|---|---|
| RF | 0.8167 | 0.7569 | 0.7895 | 0.7143 | 0.6590 |
| GB | 0.7833 | 0.7317 | 0.8421 | 0.7273 | 0.0863 |
| LR | 0.7000 | 0.6287 | 0.6842 | 0.7222 | 0.4152 |

**Table 2. Final hyperparameters per model (Grid Search, TimeSeriesSplit, macro F1).**

| Model | Hyperparameters |
|---|---|
| Gradient Boosting (selected) | `n_estimators=200`, `max_depth=7`, `learning_rate=0.05`, `min_samples_split=5`, `min_samples_leaf=2`, `loss=log_loss`, `random_state=42` |
| Random Forest | `n_estimators=200`, `max_depth=10`, `min_samples_split=2`, `min_samples_leaf=1`, `class_weight=balanced_subsample`, `random_state=42` |
| Logistic Regression | `C=1`, `penalty=l2`, `solver=lbfgs`, `class_weight=balanced`, `max_iter=2000` |

These results reinforce that global accuracy alone is not sufficient for evaluating epidemiological classifiers. In practice, the best model is not necessarily the one with the highest overall score, but the one that best identifies critical situations with stable and interpretable predictions.

### 4.2. Confusion Patterns and Feature Importance

The confusion matrices revealed important differences in the way each model handled the three risk classes. Random Forest was effective in identifying the *Normal* class but showed more confusion between *Attention* and *Outbreak*. Gradient Boosting, in turn, produced fewer false negatives for the *Outbreak* class, which makes it more appropriate for early warning use. Logistic Regression exhibited more frequent confusion between *Normal* and *Attention*, suggesting insufficient flexibility to model the progression of risk states, as shown in Figure 1.

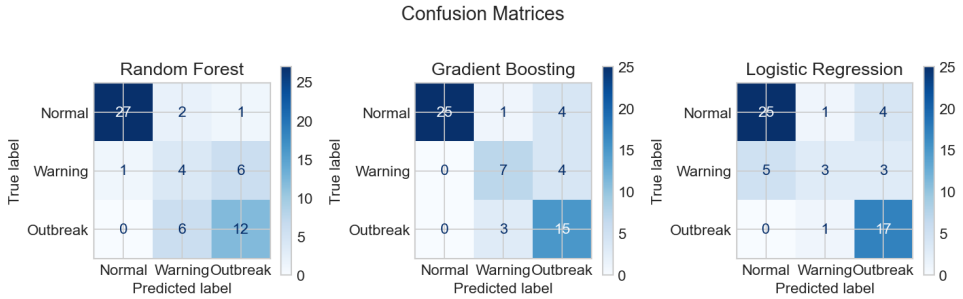


**Figure 1. Confusion matrices for Random Forest, Gradient Boosting, and Logistic Regression.**

The feature importance analysis, based on the selected Gradient Boosting model, indicated that temporal variation variables were the most relevant predictors. Annual variation, monthly variation, and lagged counts were among the most influential features. This suggests that the classifier learned to recognize changes in trajectory rather than relying only on absolute case levels. Such a result is epidemiologically coherent, since outbreak situations are often marked by abrupt shifts relative to previous patterns, as illustrated in Figure 2.

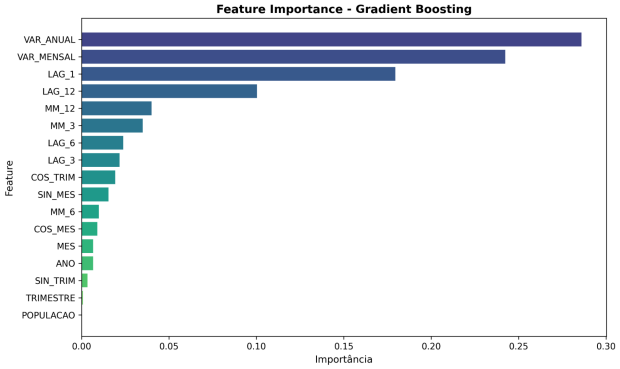


**Figure 2. Feature importance obtained from the Gradient Boosting model.**

Together, the confusion and importance analyses support the selection of Gradient Boosting as the most suitable model for the problem. It not only performed well, but did so in a way that is more consistent with surveillance priorities.

### 4.3. Uncertainty Interpretation

A central contribution of this work is the explicit inclusion of uncertainty in model evaluation. The uncertainty comparison, illustrated in Figure 3, showed that Random Forest,

despite its strong global performance, produced much less concentrated class probabilities than Gradient Boosting. This means that its predictions were often less reliable from an operational perspective.

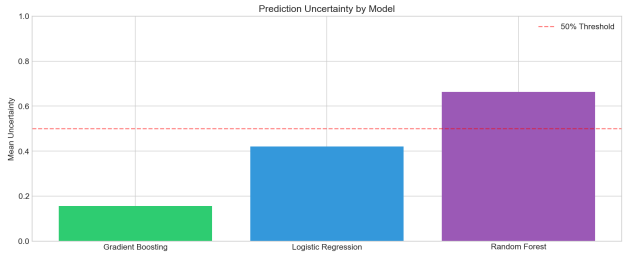


**Figure 3. Average predictive uncertainty for the evaluated models.**

This result is particularly relevant because surveillance decisions are not based solely on whether a classifier is correct, but also on how confidently it indicates a critical scenario. In this sense, the analysis confirms that predictive uncertainty is not a secondary diagnostic measure, but an important component of model selection in public health.

However, the correspondence was not absolute: some anomalous periods were not classified as outbreaks, and not all outbreak states were highly anomalous. This confirms that anomaly detection should be interpreted as complementary rather than substitutive. It identifies unusual behavior in the feature space, whereas the outbreak target is defined by statistical deviation relative to the historical baseline [Liu et al. 2008, Breunig et al. 2000].

### 4.4. Risk Forecasts for 2026

After model selection, outbreak risk forecasts were generated for 2026. The results indicate epidemiological stability for *Pertussis*, *Exanthematic Diseases*, and *Varicella*, which remained in the *Normal* class throughout the year. In contrast, Hepatitis and Meningitis showed more variable patterns, with predominance of *Attention* months and isolated *Outbreak* classifications. This reinforces the importance of disease-specific monitoring.

Meningitis showed the most critical pattern, with a pronounced increase in outbreak probability in May and adjacent *Attention* periods. Hepatitis also displayed relevant fluctuations, with predominance of *Attention* for 11 months. By contrast, the other diseases maintained consistently low outbreak probability.

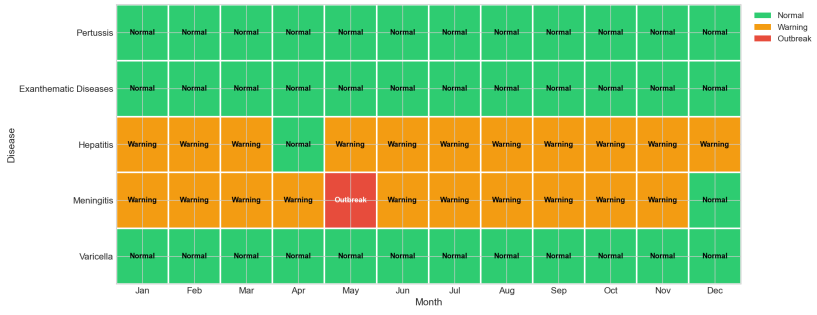


**Figure 4. Risk calendar for 2026, showing the monthly classification of each disease in Maceió.**

Figure 4 summarizes the forecasted epidemiological scenario for 2026. *Outbreak* months are concentrated mainly in Meningitis and Hepatitis, while the remaining diseases remain stable over the year. The predominance of *Attention* in some periods is also relevant, since it supports gradual intervention rather than only binary responses. Thus, the risk calendar provides a compact and operational summary for epidemiological surveillance.

## 5. Conclusions and Future Work

This study applied machine learning models to classify infectious disease outbreak risk in Maceió, Alagoas, using historical SINAN data. Instead of predicting the exact number of future cases, a three-level classification approach (Normal, Attention, and Outbreak) was adopted, proving more suitable for epidemiological surveillance.

Among the evaluated models, Gradient Boosting showed the best balance between outbreak sensitivity, predictive stability, and low uncertainty. Random Forest also achieved strong overall performance, but with higher uncertainty, while Logistic Regression was less effective in capturing nonlinear temporal patterns in the data.

The results also showed that combining classification with uncertainty analysis improves epidemiological risk interpretation. In addition, the 2026 forecasts and risk calendar demonstrated the practical value of the proposed approach for disease monitoring and surveillance planning.

Future work includes incorporating external variables (climate, demographics, and mobility), prospective validation in real surveillance scenarios, and integrating predictive uncertainty into risk-based decision strategies.

## References

Ahmad, G. N. et al. (2022). Efficient medical diagnosis of human heart diseases using machine learning techniques with and without gridsearchcv. *IEEE Access*, 10:80151–80173.

Antunes, J. L. F. and Cardoso, M. R. A. (2015). Uso da análise de séries temporais em estudos epidemiológicos. *Epidemiologia e Serviços de Saúde*, 24(3):565–576.

Balcan, D. et al. (2009). Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences*, 106(51):21484–21489.

Borges, P. K. d. O. et al. (2024). Impacto da covid-19 sobre doenças de notificação compulsória: um estudo de série temporal. *Revista da Escola de Enfermagem da USP*, 58:e20240098.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Breunig, M. M. et al. (2000). Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, pages 93–104.

Du, L. and Pang, Y. (2021). A novel data-driven methodology for influenza outbreak detection and prediction. *Scientific Reports*, 11(1):13275.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, pages 1189–1232.

Gao, S. et al. (2024). Early detection of disease outbreaks and non-outbreaks using incidence data. *arXiv preprint arXiv:2404.08893*.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2 edition.

Jones, K. E. et al. (2008). Global trends in emerging infectious diseases. *Nature*, 451(7181):990–993.

Koike, M. (2025). Datasus: Uma ferramenta essencial para a saúde pública no brasil. *Arquivos Brasileiros de Cardiologia*, 122(2):e20250123.

Lima, E. C. A., Oliveira, J. P., and Santos, M. R. (2021). Qualidade da informação no sistema de informação de agravos de notificação: uma revisão integrativa. *Cadernos de Saúde Pública*, 37(6):e00123420.

Liu, F. T., Ting, K. M., and Zhou, Z.-H. (2008). Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422.

Ludermir, T. B. (2021). Inteligência artificial e aprendizado de máquina: estado atual e tendências. *Estudos Avançados*, 35:85–94.

Maia, D. A. B. et al. (2019). Avaliação da implantação do sistema de informação de agravos de notificação em pernambuco, 2014. *Epidemiologia e Serviços de Saúde*, 28:e2018187.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Rocha, M. S., Bartholomay, P., et al. (2020). Notifiable diseases information system (sinan): main characteristics of notification and data analysis related to tuberculosis. *Epidemiologia e Serviços de Saúde*, 29(1):e2019017.

Santangelo, O. E. et al. (2023). Machine learning and prediction of infectious diseases: a systematic review. *Machine Learning and Knowledge Extraction*, 5(1):175–198.

Scursone, G. F. et al. (2025). Hyperparameter optimization of xgboost on air pollution and respiratory health data. *Studies in Health Sciences*, 6(4):e21945–e21945.

Shaman, J. and Karspeck, A. (2012). Forecasting seasonal outbreaks of influenza. *Proceedings of the National Academy of Sciences*, 109(50):20425–20430.

Silva, G. D. M. d. et al. (2020). Identificação de microrregiões com subnotificação de casos de tuberculose no brasil, 2012 a 2014. *Epidemiologia e Serviços de Saúde*, 29:e2018485.

Souza, L. M. and Silva, C. A. (2020). Uso do sinan como ferramenta para a vigilância epidemiológica no brasil. *Revista Brasileira de Epidemiologia*, 23:e200045.

Wang, S. et al. (2021). Research on expansion and classification of imbalanced data based on smote algorithm. *Scientific Reports*, 11(1):24039.