BMC Medical Research Methodology

RESEARCH

Open Access

Combining machine learning and dynamic system techniques to early detection of respiratory outbreaks in routinely collected primary healthcare records

Dérick G. F. Borges[1,2†], Eluã R. Coutinho[1,3,4†], Thiago Cerqueira-Silva[1,6], Malú Grave[5], Adriano O. Vasconcelos[4], Luiz Landau[4], Alvaro L. G. A. Coutinho[4], Pablo Ivan P. Ramos[1], Manoel Barral-Netto[1,6], Suani T. R. Pinho[2], Marcos E. Barreto[1,7] and Roberto F. S. Andrade[1,2*]

Abstract

Background Methods that enable early outbreak detection represent powerful tools in epidemiological surveillance, allowing adequate planning and timely response to disease surges. Syndromic surveillance data collected from primary healthcare encounters can be used as a proxy for the incidence of confirmed cases of respiratory diseases. Deviations from historical trends in encounter numbers can provide valuable insights into emerging diseases with the potential to trigger widespread outbreaks.

Methods Unsupervised machine learning methods and dynamical systems concepts were combined into the Mixed Model of Artificial Intelligence and Next-Generation (MMAING) ensemble, which aims to detect early signs of outbreaks based on primary healthcare encounters. We used data from 27 Brazilian health regions, which cover 41% of the country's territory, from 2017-2023 to identify anomalous increases in primary healthcare encounters that could be associated with an epidemic onset. Our validation approach comprised (i) a comparative analysis across Brazilian capitals; (ii) an analysis of warning signs for the COVID-19 period; and (iii) a comparison with related surveillance methods (namely EARS C1, C2, C3) based on real and synthetic labeled data.

Results The MMAING ensemble demonstrated its effectiveness in early outbreak detection using both actual and synthetic data, outperforming other surveillance methods. It successfully detected early warning signals in synthetic data, achieving a probability of detection of 86%, a positive predictive value of 85%, and an average reliability of 79%. When compared to EARS C1, C2, and C3, it exhibited superior performance based on receiver operating characteristic (ROC) curve results on synthetic data. When evaluated on real-world data, MMAING performed on par with EARS C2. Notably, the MMAING ensemble accurately predicted the onset of the four waves of the COVID-19 period in Brazil, further validating its effectiveness in real-world scenarios.

[†]Dérick G. F. Borges and Eluã R. Coutinho contributed equally to this work.

*Correspondence:
Roberto F. S. Andrade
randrade@ufba.br
Full list of author information is available at the end of the article

Conclusion Identifying trends in time series data related to primary healthcare encounters indicated the possibility of developing a reliable method for the early detection of outbreaks. MMAING demonstrated consistent identification capabilities across various scenarios, outperforming established reference methods.

Keywords Syndromic surveillance, Outbreak detection, Primary healthcare data, Machine learning, Reproduction number

Background

A pandemic is a large-scale disruptive event caused by the emergence and rapid spread of a pathogen transmissible by some mechanism among members of a given population [1]. The establishment of an event of such magnitude is generally preceded, in the pre-pandemic period, by recurrent changes in the number of clinical records and their content – clinical characteristics and biomarkers. These changes can be identified and characterized accurately in several aspects, such as whether the increase in non-specific records is expected for some seasonal reason or by the selection and detailed analysis of cases not yet reported in the literature.

Over the years, epidemiological surveillance services have improved their methods of identifying new epidemics, showing that detecting the first cases and rapidly taking control measures can reduce the impact on affected populations [2, 3]. The increasingly faster conversion of indicative signs and their conversion into risk indicators has become a crucial mechanism for public health, requiring the continuous search for new methods to identify initial cases clinically and implement new approaches for treating routine data collected in health encounters [4].

A large number of methods to detect outbreaks from surveillance data [5–10] have been proposed. For instance, Unkel et al. [11] reported a non-exhaustive list of over 40 statistical methods for outbreak detection. A key challenge, however, lies in developing methods that offer a good balance between sensitivity and specificity to detect the vast majority of outbreaks without generating too many false positive alerts [12]. However, for reliable detection performance, the chosen method depends critically on the nature of the intended application [13].

This work focused on syndromic surveillance [14] using primary healthcare data (PHC). Two sets of risk analysis methods were devised to detect anomalies and dynamic properties of critical transitions extracted from the time-series of PHC data in Brazil. These data represent visits to PHC units of individuals presenting with symptoms of upper respiratory tract infection (URTI). While the approach developed here can be applied to diverse epidemiological scenarios and diseases, the current focus is justified by the impact of critical URTI-related epidemics and pandemics, such as the Spanish flu caused by an H1N1 virus in 1918, SARS-CoV and avian influenza (H5N1) in 2003, the emergence of a new H1N1 influenza virus strain in 2009, and the SARS-CoV-2 virus, which caused the recent COVID-19 pandemic in 2020 [15]. The investigations were carried out within the scope of the Alert-Early System of Outbreaks with Pandemic Potential (AESOP) project, described in [16].

We applied four unsupervised machine learning methods: Isolation Forest (ISF) [17, 18], Local Outlier Factor (LOF) [19], One Class Support Vector Machine (OCSVM) [20], and Copula-Based Outlier Detection (COPOD) [21] to study PHC data aggregated by epidemiological weeks between 2020–2023 and identify abrupt variations in the number of visits that can be characterized as *anomalies* and associated with an outbreak. We also applied a Next Generation Method (NGM) [22, 23], often used to describe epidemic events, to create an outbreak risk indicator based on an extended concept of reproduction number and its threshold value. Therefore, we introduce the **M**ixed **M**odel of **A**rtificial **I**ntelligence and **N**ext-**G**eneration (MMAING) ensemble and compare it with the Early Aberration Reporting System (EARS) method to measure the performance of outbreak risk indicators that AESOP may have anticipated.

This work is organized as follows: "Methods" section describes the methodology used, comprising the dataset and essential features of the chosen methods, along with the criteria and metrics used for evaluation. "Results" section presents results from MMAING over data from different Brazilian capital cities, comparative results with the COVID-19 pandemic period, and validation based on synthetic data. "Discussion" section discusses the results and limitations of our approach. Finally, "Conclusion" section brings some conclusions and highlights the relevance of our study.

Methods

The main steps in MMAING's workflow, from data preprocessing to outputs, are illustrated in Fig. 1.

In the pre-processing stage, the time series of URTI-related PHC data are obtained, consolidating an input database (green box in Fig. 1). The database is split into training and validation data sets as usual in unsupervised ML protocols (orange and blue boxes). At the same time,

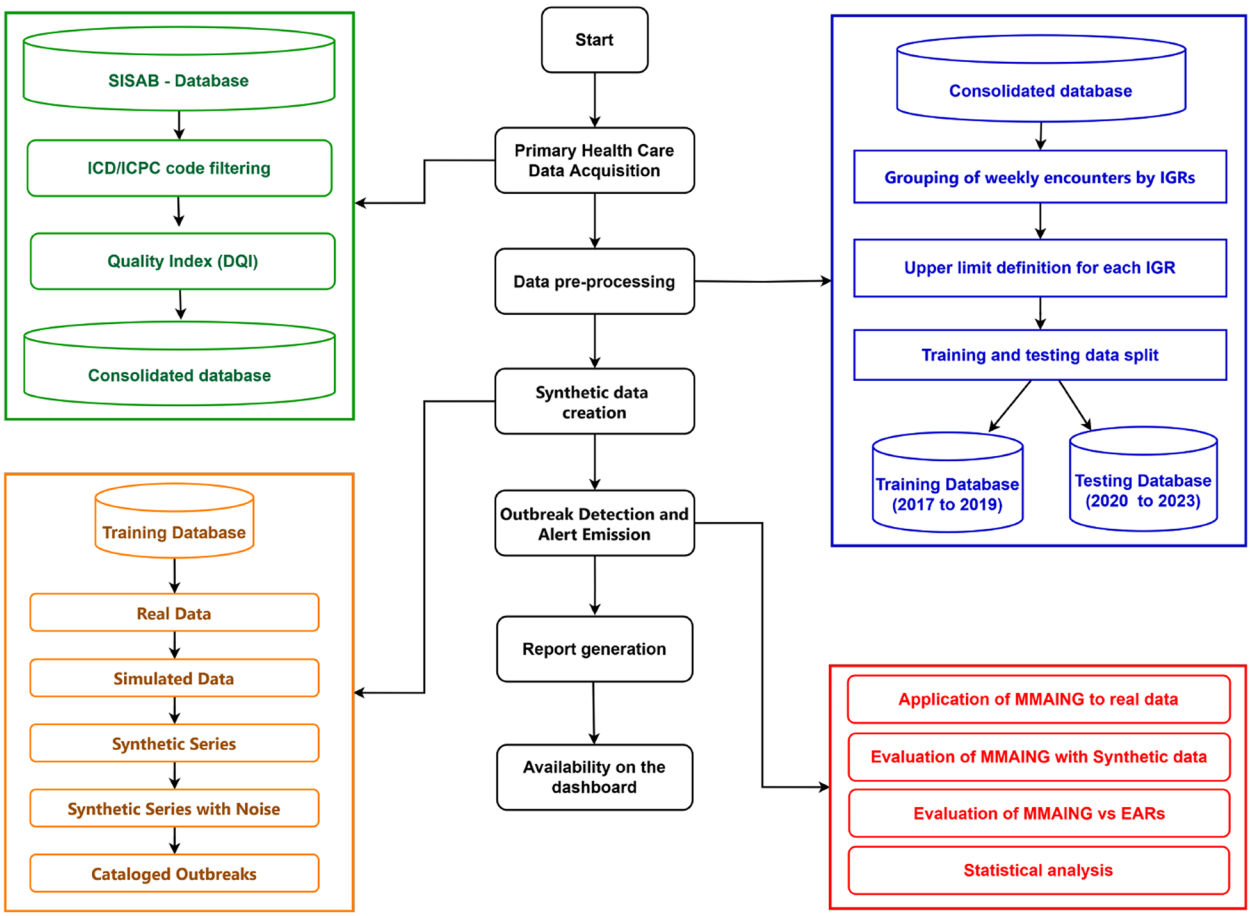

Fig. 1 Simplified MMAING workflow with main steps in the central column and details in the lateral ones. The green box represents the initial stages related to data acquisition from Brazil's National Primary Health Information System (SISAB), filtering of diseases related to URTI using the International Classification of Diseases (ICD-10) and the International Classification of Primary Health Care (ICPC-2) codes, and Data Quality Index (DQI) evaluation. The blue box identifies the pre-processing stages, such as grouping the data at municipality level by Imediate Geographic Regions (IGR), calculation of the upper limit, and data splitting. The orange box indicates the stages for generating and cataloging the synthetic series. The red box describes the stages for Outbreak Detection, EWS Emission, and the comparison of MMAING with EARS on real and synthetic data. The reports and dashboard access are currently of exclusive use by the responsible health teams at municipal, state and national level

a full database is used to implement the NGM. The training datasets are also used to generate synthetic series to enable quantitative evaluation of the models through statistical metrics.

Outbreak detection and early warning signals (EWS) issuance by MMAING (red box) are conducted to identify anomalous patterns that occurred in Brazilian state capitals from 2020 to 2023, drawing comparisons to the events occurred during the whole COVID-19 pandemic period between 2020 and 2022. Detection of possible outbreaks in original and synthetic data and comparison with EARS results are evaluated using statistical metrics.

Data sources

We used data from Brazil's National Primary Health Information System (SISAB - Portuguese acronym for Sistema de Informação em Saúde para a Atenção Básica), which contain data on all encounters from publicly funded primary health care (PHC) facilities. The public healthcare in Brazil is universal and covers the entire population, and around 75% of the population only use the public healthcare [24]. Each encounter in PHC healthcare facilities is coded using *International Classification of Diseases* (ICD-10) and *International Classification of Primary Health Care* (ICPC-2). The original data are organized by municipality (identified by name and code), year, epidemiological week. This study, covering data for the period 2017–2023, is based on a data subset that is obtained by extracting only those entries with codes related to 50 specific upper respiratory tract infection (URTI) conditions. After this, the data is organized into time series, where each data point corresponds to

the weekly count PHC visits in that municipality related to URTIs. The detailed list of the 50 ICD-10 and ICAP-2 codes can be found in the Supplementary Material (MS) Table A1. Before being used in the analyses, the data undergoes a quality test, which takes into account the completeness, timeliness and consistency of the records. This test assigns to each municipality its weekly Data Quality Index (DQI) indicating whether these criteria were satisfied or not. Analyses are performed only for those pieces of data with a positive DQI evaluation [25].

Data from the 5,570 Brazilian municipalities were grouped into 510 *Immediate Geographic Regions* (IGRs) for a precise analysis of the risk of epidemic outbreaks. As defined by the *Brazilian Institute of Geography and Statistics* (IBGE), an IGR is a group of municipalities that have, as their primary references, an urban network and a local urban center where the nearby population seeks goods, services, and work. We used the URTI time series of 27 IGRs corresponding to capital cities, covering 41% of the country's inhabitants, offering a broad and diversified view of the evolution of respiratory syndromes during the epidemiological weeks of the period under study.

Study design

We combined four unsupervised methods (ISF, LOF, OCSVM, COPOD) commonly used for anomaly detection with the deterministic NGM that accounts for dynamic causes, underlying formation, and outbreak propagation. These five methods were aggregated into an ensemble (MMAING), representing an integrative methodology to identify emerging epidemic patterns through the detection of EWS. The association between series anomalies and EWS requires that the former be restricted to upward trends. The final indication of pandemic risk (*yes/no*) is based on a voting system that takes into account the equally weighted *yes/no* EWS indications provided by each of the methods individually. The final prediction is the option receiving the mojority of votes, which can unambiguously be applied for an odd number of methods [26, 27].

To improve the efficiency in detecting EWS among different models, where the identification of detected points can be counterproductive, we propose a strategy consisting of establishing a time-dependent upper limit $\epsilon(x(t))$ based on confidence intervals [28], and assigning a positive risk of outbreak only to those values of t exceeding such threshold, i.e., $x(t) > \epsilon(x(t))$. Mathematically, this threshold is defined as

$$\epsilon(x) = \bar{x} + \frac{z\sigma}{\sqrt{n}}, \tag{1}$$

where $\bar{x}, z$, σ and n denote the sample mean, the critical value for a given confidence level, which we assumed to be z=1.96 to warrant a 95% of confidence interval. The sample standard deviation, and the sample size, respectively. This threshold is crucial for identifying potential EWS, but its effectiveness depends on the sampling strategy adopted. We applied two strategies:

- A *moving window scheme* averages the data from a recent past of w weeks, removing short-term fluctuations. This approach proves to be extremely useful when there is a need to monitor recent trends in the data set [29].
- Seasonally adjusted regimes considering cyclical patterns or trends from previous years. This choice is essential for events impacted by external variables, such as climate change, holidays, and other periodic events.

Our methodology considered both strategies, with $n = 5$ for the moving window scheme and references to the same epidemiological weeks from 2017 to 2019 for the seasonal scheme. These strategies facilitate the detection of increased patterns at specific time intervals, ensuring that only EWS points exceeding the threshold value are considered while improving precision by avoiding false positives. The indication of a potential outbreak is followed by validation against already established syndromic surveillance methods, such as EARS and synthetic data.

Synthetic data

To evaluate the models used in this study, we created 27 synthetic series based on real data from each of the 27 IGRs of Brazilian capital cities, as illustrated in Fig. 2. We introduced noise into the generated series to simulate periods of abnormal behavior, which we defined as outbreaks, and recorded these disturbed intervals to identify and catalog these occurrences, similarly to the approach reported by Neill [30]. This procedure allowed the evaluation of the models using statistical metrics. To do so, we used real data recorded between 2017 and 2019, before the COVID-19 pandemic, and carried out the following procedures:

Real series smoothing Let X_t, where t is a discrete variable ranging from $t_1 = 1$ to $t_N = N$, be a time series representing a value recorded for an epidemiological week. The moving average M_i at each time t_i is obtained by convolving the data $X_{t_i} \equiv X_i$ with a moving window of any width $1 < w \leq N$, whereby in this work we have considered $w = 8$ weeks. The average over past neighboring points is a smoothed series with $N - 7$, centered at the points at t_i, $i = 4, 5, 6...N - 4$, and expressed by

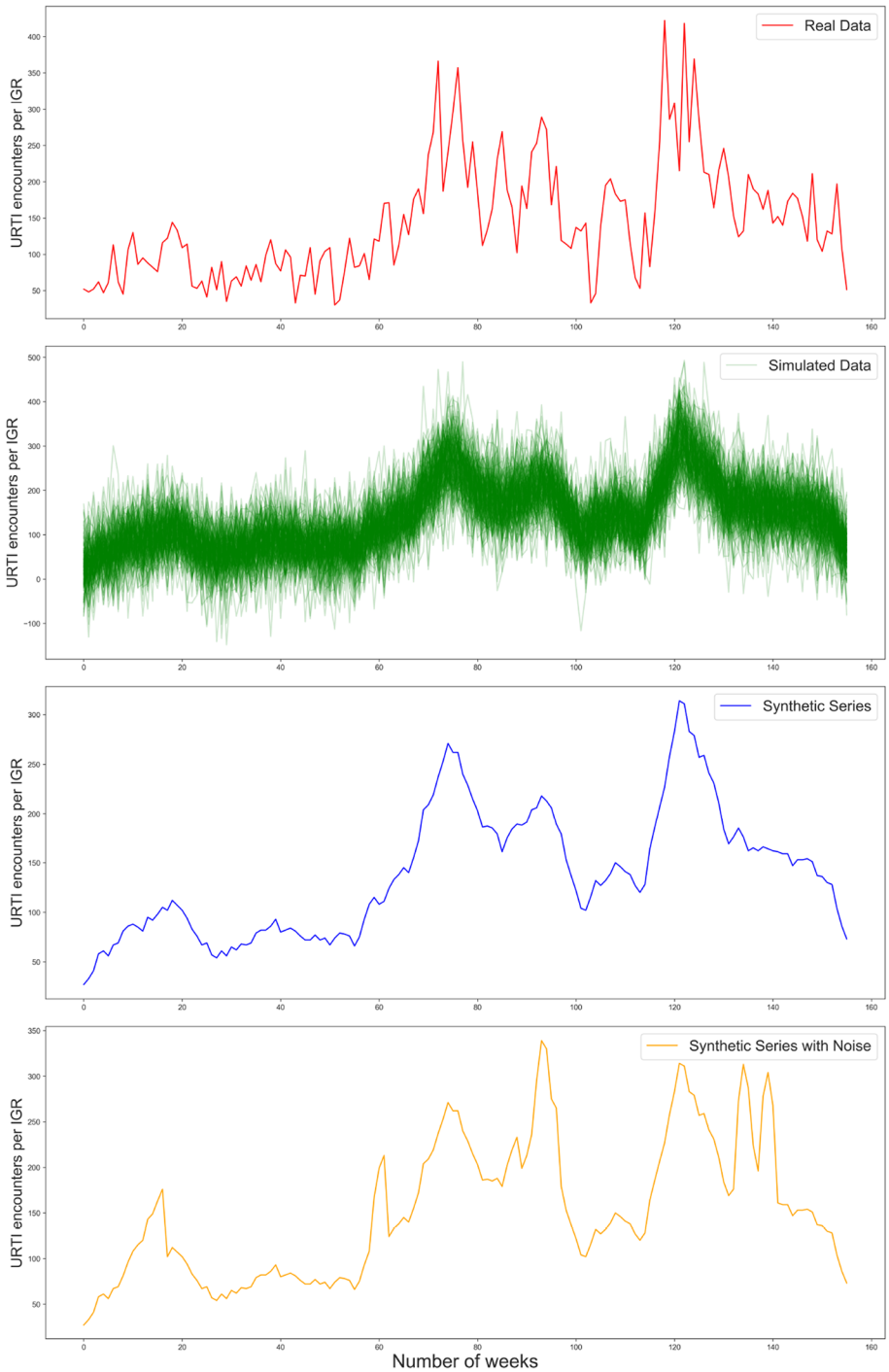

Fig. 2 Process of synthetic data generation: Red - real data from an IGR; Green - set of 200 simulated series; Blue - synthetic time series by averaging over green curves; Orange - synthetic time series with superimposed noise

$$M_i = \frac{1}{8} \sum_{t=i-3}^{i+4} X_i. \tag{2}$$

Synthetic data generation The moving average converts the original series of integer number is into a series of continuous rational numbers. Therefore, for each time t_i, a normal continuous distribution $\mathcal{N}(\mu, \sigma)$ was used to generated synthetic data $\bar{X}_i \sim \mathcal{N}(\mu, \sigma)$ as the average of $(n = 200)$ simulations, whereby

$$\mathcal{N}(\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_t-\mu}{\sigma}\right)^2\right] \quad (3)$$

The mean μ of each distribution was set to that of the smoothed series M_i multiplied by a factor F_M ranging from 0.8 to 1.2. The standard deviation σ was taken as that of the real series, multiplied by a factor F_D ranging from 0.2 to 0.8. This procedure allowed the generation of a data series following a distribution similar to the original series with superimposed small changes.

Addition of random noise To test the ability of MMAING and other methods to detect specific events, we generated the series Z_i to simulate the emergence of outbreak periods by adding random fluctuations to the series $\bar{X}_i$. Each of the $K \in [1,6]$ simulated outbreaks, which may last from 4 to 10 weeks, consists of a set of $4 \leq P_k \leq 10$ consecutive points, where P_k is randomly chosen from a uniform distribution $U(4, 10)$. Its starting time t_k is also randomly chosen from $U(1, N-10)$. This procedure does not exclude the possibility that a specific time t_j be included in two or more sets P_k. This way, after starting with the series $\bar{X}_i$, a total number $L = \sum_{k=1}^{K} P_k$ of interventions is required to obtain Z_i. Each intervention ℓ in the interval $[1, L]$ amounts to evaluating $\tilde{\epsilon_\ell}$ to be added to a well defined $\bar{X}_\ell$, in such a way that

$$Z_\ell = \bar{X}_\ell + \tilde{\epsilon_\ell}. \quad (4)$$

For a given series $\bar{X}_i$, the value of each $\tilde{\epsilon_\ell}$ is obtained by a fixed procedure, which is independent of the value of ℓ and of the interval P_k where $\bar{X}_\ell$ is located. It requires the random choice of two parameters from two previously evaluated sets, namely the amplitude $\mathbf{A} = \{A_1, A_2, ..., A_{10}\}$ and frequency $\mathbf{f} = \{f_1, f_2, ..., f_8\}$, where $A_j \sim U[50, B]$, with B as an integer number defined as $B = (max(\bar{X}\_min(\bar{X}))/2$, so that the set $\mathbf{A}$ depends on the specific series, and $f_j \sim U[0,1]$ does not explicitly depends on the series $\bar{X}_i$. Finally, with the help of $\mathbf{A}$ and $\mathbf{f}$, we define

$$\tilde{\epsilon_\ell} = A_\ell|(\sin(2\pi f_\ell(t_k - t_\ell) + \frac{1}{10^4}|, \quad (5)$$

where the inclusion of a small non-zero constant in the definition of $\tilde{\epsilon}_\ell$ seeks to realistically reproduce the initial behavior of an outbreak, playing an essential role to simulate an actual event.

Cataloguing of abnormal occurrences introduced by addition of noise To verify whether the models would be able to detect new variations and allow an analysis through statistical metrics, we first identified the abnormal signals presented in the series $\bar{X}_i$ without noise, which may have originated from the preservation of the existing trend in the original series. Then, we recorded the perturbed periods by adding noise ($\tilde{\epsilon}$) to create a reference dataset, identifying the start and end points of the outbreak, which will be the detection target.

MMAING ensemble

In machine learning (ML), an ensemble combines multiple models aiming to achieve higher precision and robustness while reducing variation and bias compared to single, isolated models. Among the existing approaches, hard voting [27, 31] stands out as one of the simplest and most effective. This specific technique consists of training a plurality of different models and subsequently employing a model, which in the context of this work is called MMAING, to integrate the results of the base models. The original data set is used to train the base models, whose results serve as input for MMAING, culminating in the decision about issuing an EWS or not, as illustrated in Fig. 1.

MMAING's innovation is materialized through a previously unexplored combination of four ML models and one NGM model and a selective strategy that chooses signals from three of them based on a majority voting system that guarantees an equal contribution of each model. Therefore, MMAING not only simplifies the interpretation of results through an integrative voting process but also reinforces the reliability of detections, establishing a robust consensus among its methods.

Machine learning methods

Unsupervised machine learning models are commonly applied to anomaly detection over time series data, as they can learn patterns from the data and thus spot those points that do not fit any patterns as anomalies [32, 33]. MMAING incorporates the four ML methods that will be detailed in the sequence. They perform better in detecting anomalies as compared to other unsupervised methods targeted to predict continuous outcomes in regression problems, like OLS, LASSO, SVR, Boosting, etc.

Isolation Forest (ISF): a tree-based algorithm that uses the concept of *isolation* to calculate an abnormality score for each point and employs a recursive partitioning process to isolate outliers. ISF creates a random forest, where each decision tree randomly selects a feature and chooses a threshold value within the minimum and maximum range of that feature to split the data. This process is repeated until all the data points are isolated. Anomalies are identified based on a score value obtained

by the average length of paths in the forest trees divided by a normalization factor. Abnormal data points tend to be isolated more quickly, resulting in shorter paths from the root compared to *standard* data points [17, 18].

Local Outlier Factor (LOF): an algorithm designed to identify outliers in diverse datasets, including time series. LOF works by evaluating the local density (i.e., how tightly packed) of data points compared to the density of their neighbors. Consequently, a point surrounded by many nearby points will have a high local density, while an isolated point will exhibit a low local density. If the local density of a point is significantly lower relative to its neighbors, it is classified as an outlier [19].

One-Class Support Vector Machine (OCSVM): an algorithm aiming to find an optimal hyperplane that maximizes the separation between examples of different classes in a high-dimensional space. This process involves projecting the input data into a high-dimensional space via a kernel function and creating a hyperplane to determine whether a new observation is within the hyperplane (not an anomaly) or outside it (an anomaly) [20].

Copula-based outlier detection (COPOD): it represents a significant innovation in the field of multivariate outlier detection, moving away from conventional cluster-centric approaches. Unlike these approaches, COPOD is based on copulas [34], a mathematical function to model dependencies (correlation) between variables. Detecting outliers with COPOD involves a three-step process: i) calculating empirical cumulative distribution functions based on the data, ii) generating the empirical copula function from these functions, and iii) using the empirical copula to estimate the tail probabilities and quantifying these probabilities as data anomaly scores [21]. Tail probabilities correspond to the probability of a data point falling at the extremes of distributions, which are the least frequent areas and, therefore, most susceptible to outliers. After estimating the tail probabilities, COPOD transforms them into anomaly scores representing the degree of abnormality of each point, and those points with high abnormality degrees are flagged as potential outliers.

Next Generation method

The Next Generation Method (NGM) provides reliable information on the evolution of epidemic outbreaks, including the reproduction number R_0, which measures the average number of new infections caused by an infected individual at the beginning of an epidemic process when the whole population has no protection to that infection [35, 36]. Recently, that procedure was generalized in terms of the generation time interval distribution [22, 37], for any value of $t > 0$. Its implementation requires the time series of confirmed cases $B(t)$ in a given population, from which one may infer the number of individuals who become infected and the number of others to whom they can transmit the pathogens. $R(t)$ estimates are based on classical compartmental models (SIR, SEIR, and more complex ones) [22, 23, 38] or on approaches using data of confirmed cases without further consideration of a model dynamics [39–42].

As in other problems of population dynamics, the estimation of $R(t)$ needs the idea of generation time, so that the variables of the dynamical system have to depend on both time t and age τ. For epidemiological systems, τ means the age of infection [43], and as a consequence $R(t)$ represents the average number of people that an individual infected at time t can infect during his entire infectious period τ. The implementation of NGM to accurately describe the spread of the agent in the population requires the accurate identification of each individual who becomes infected and the number of contacts with other individuals to whom the infected person can transmit the pathogens.

Using a generation interval distribution, also called generation time distribution, $g(\tau)$, which considers the time that an individual who visited a PHC unit, if infected, takes to generate a new infection, we define $R(t)$ as:

$$R(t) = \frac{a(t)}{\int_0^\infty a(t-\tau)g(\tau)d\tau} \tag{6}$$

where, $a(t)$ is the number of visits and $g(\tau)$ is the normalized generation time distribution.

According to [23], the generation time interval distribution $g(\tau)$ for the SEIR compartmental model is given by:

$$g(\tau) = \frac{\epsilon e^{-\kappa\tau} + \frac{\kappa}{\gamma-\kappa}[e^{-\kappa\tau} - e^{-\gamma\tau}]}{\frac{\epsilon}{\kappa} + \frac{1}{\gamma}} \tag{7}$$

for which the parameters ϵ, κ and γ represent, respectively, factor of pre-symptomatic infection, inverse of latency period, and removal rate. In this study, given the lack of knowledge about the disease with respiratory symptoms and therefore its dynamics, we defined κ and γ as equal ($\kappa = \gamma + \delta$ with $\delta \to 0$), and $\epsilon = 0$ as in the simple version of SEIR model. Thus, $g(\tau)$ reduces to:

$$g(\tau) = e^{-\gamma\tau}\gamma^2\tau \tag{8}$$

To be conveniently used, this form of the distribution $g(\tau)$ needs to be converted to the corresponding

version valid for discrete time and then normalized [44]. For this, we will discretize this distribution by initially considering a geometric progression associated with SEIR model. Replacing τ by $(n-1)$ in (8), the general term of the geometric progression is given by:

$$g(n) = (n-1)g_1 q^{(n-1)}, \tag{9}$$

where $q = e^{-\gamma}$, $t = 0$ corresponds to $n = 1$. Based on the sum $S(m)$ of the first m terms of the function $g(n)$, the first term g_1 (the normalization factor) is obtained (for the intermediate steps of the calculation see Appendix B.1), leading to:

$$g_1 = \frac{(1-q)^2}{q[1+(m-1)q^m - mq^{m-1}]} \tag{10}$$

Therefore, we are applying the method based on the Next Generation Matrix technique (NGM) discretized and normalized in terms of a geometric progression, and, as in [23] we take into account the dynamics through $g(\tau)$ and the data through $a(t)$.

In order to promote a more realistic potential detection, MMAING adds a new perspective to this metric by including the estimation of $\hat{R}(t)$, which is defined in a similar way as $R(t)$, with the difference that it now depends on the weekly incidence of registered encounters at PHC posts. It corresponds to an extension to the approach developed in [23], which amounts to using the series of URTI-related PHC encounters to evaluate $\hat{R}(t)$, an analogous of usual $R(t)$ that might assess the risk of epidemic outbreaks. If we use the same expressions for evaluating $R(t)$ in [23] as well as similar transmission dynamics based on the SEIR model, some key mathematical properties of $R(t)$ are also expected to hold for $\hat{R}(t)$, e.g. an increase or decrease of its values when the number of encounters increases or decreases.

By contrast, the clear-cut epidemiological meaning of $R(t)$ [44, 45] (whether smaller, equal or larger than 1) obtained from a series of confirmed cases cannot be transferred to $\hat{R}(t)$ in a straightforward way, but it becomes necessary to calibrate the threshold value of $\hat{R}(t)$ that better measures the actual outbreak risk. Because the number of confirmed infection cases is usually smaller than the number of encounters, it is expected that $\bar{R}$, a threshold value of $\hat{R}(t)$ defining actual epidemic risk, will be larger than 1. Indeed, our estimates have shown that $\langle \hat{R}(t) \rangle$, the average of $\hat{R}(t)$ taken over values of $\hat{R}(t) > 1$ for the national PHC time series is ~ 1.24. Additionally, the analyses of the corresponding series for each Brazilian state consistently indicate that $1.2 < \langle \hat{R}(t) \rangle < 1.3$. Therefore, we established the criterion for outbreak risk by the condition $\hat{R}(t) > 1.25 \equiv \bar{R}$. According to that criterion, the NGM is applied carrying to the ensemble information from both dynamical process and on the data.

A somewhat similar analysis was used in a study on the transmissivity of Ebola virus disease based on confirmed cases, where it was observed that a suitable threshold value $\bar{R}$ for the usual $R(t)$, set in terms of its past median, could forecast spreading trends within 1–2 weeks [46]. These results were confirmed one year later in a similar study using COVID-19 data [10].

Early Aberration Reporting System (EARS)

EARS [6] is available in its three variants - C1, C2, and C3 - mainly used to monitor weekly syndromic counts. These methods are helpful when limited baseline data is available. The variants are based on the Shewhart procedure that uses a moving sample mean and a sample standard deviation to standardize each observation [47]. By default, variant C1 calculates the sample mean and standard deviation using data from the seven weeks before the current observation; C2 is similar but considers a two-week delay; and C3 combines the results obtained with C2 for the current and two previous weeks, as detailed discussed in [6, 47]. In this work, however, we extended the period to eight weeks to better adjust the data. EWS for the different variants are produced when the corresponding statistics C1 or C2 exceed three sample standard deviations above the sample mean or if C3 exceeds two sample standard deviations above the sample mean [6, 47].

MMAING configuration

As mentioned earlier, MMAING is an ensemble of models that utilizes a voting system to detect EWS linked to potential outbreaks in syndromic time-series data. Data recorded between 2017 and 2019 was used for training due to the stability observed in the number of encounters throughout epidemiological weeks to ensure sensitivity to abnormal changes. Subsequently, we applied these models to real datasets recorded between 2020 and 2023, aiming to empirically analyze the anticipation of outbreak periods during the COVID-19 pandemic. Additionally, we evaluated the performance of the proposed ensemble with synthetic data.

Furthermore, to assess MMAING's adaptability and versatility in scenarios that may demand moderate or high rigor (i.e., larger or smaller number of false positive EWS), we adopted two distinct configurations, which distinguish themselves by the values of a few specific parameters in each method. The parameter values adopted in each configuration, named "balanced" (BLCf) and "strict" (STCf), are indicated in Table 1. In both cases, the adopted configurations were used to run all ML models were based on a total of 810 synthetic series obtained

Table 1 MMAING parameterization settings

Method	Parameter	BLCf	STCf
ISF	n_{est}	500	400
	$\mathcal{C}$	0.4	0.3
LOF	n_{nei}	500	300
	$\mathcal{C}$	0.4	0.3
OCSVM	ν	0.8	0.5
	kernel	RBF	RBF
	γ	0.001	0.001
COPOD	$\mathcal{C}$	0.4	0.3
NGM	$\bar{R}$	1.25	1.30
	γ	0.2	0.2

from 30 independent series for each of the actual data sets for the 27 IGRs. BLCf settings may decrease precision and increase sensitivity, resulting in more EWS and consequently increasing the number of false positives. On the other hand, the STCf aims to enhance precision and reduce the false positive rate, which may result in failing to issue true EWS.

BLCf uses parameters that seek a compromise between sensitivity and specificity. For example, for the ISF and LOF methods, the number of estimators (n_{est}) and neighbors (n_{nei}) are set to 500, while the contamination $\mathcal{C}$ is set to 0.4, indicating that we expect about 40% of anomalous points. The strict configuration means a reduction in the number of ISF trees and the number of LOF neighbors to 400 and 300, respectively, as well as a reduction in contamination to 0.3.

The OCSVM method uses a higher ν value in BLCf (0.8) compared to STCf (0.5), indicating greater flexibility in class separation. The use of the Radial Basis Function (RBF) kernel and the value of γ are kept consistent between configurations, suggesting that the shape of the decision boundary and the complexity of the model are considered adequate in both cases. COPOD and NGM also present different parameter values between configurations. COPOD maintains consistency with ISF and LOF concerning contamination, while NGM adjusts the threshold $\bar{R}$ to reflect the desired stringency of detection.

In outbreak detection contexts, these configurations can be adapted with specific objectives to monitor public health data. While the sensitivity measure is important if a clear and consistent EWS is needed, in scenarios where surges occur quite frequently, precision (or positive predictive value) measures the probability of an EWS being true, particularly when surges do not occur frequently. For each of these measures, the user may want to define settings for the algorithms and/or prioritize which measures are most important for their surveillance needs [48].

Training and validation

To assess MMAING's effectiveness, we adopted three distinct approaches. The first involved using PHC data to conduct an empirical analysis, comparing the outbreak detection results obtained during the COVID-19 pandemic with the official reports from the Brazilian government [49], which detail the different waves of the pandemic. The second approach aimed to evaluate the detection capability of MMAING under two distinct configurations using categorized synthetic data. Finally, the third approach used real and synthetic data to compare MMAING and EARS models. To achieve this, we employed statistical metrics such as Probability of Detection (POD), Sensitivity, Specificity, Positive Predictive Value (PPV), and F1 [12, 50], under the following assumptions:

- POD evaluation: For each scenario and each period of the current week, if an EWS is generated at least once between the start and end of an outbreak, the outbreak is considered detected [8, 48]. POD is an event-based sensitivity (i.e., the entire outbreak interval is counted as a single observation for the sensitivity measurement), thus corresponding to the proportion of outbreaks detected with the total number of synthetic replicas.
- For sensitivity analysis, True Positive (TP), True Negative (TN), False Negative (FN), and False Positive (FP) detected events are defined, respectively, as weeks with a surge and issued a warning, weeks with no surges nor warning, weeks with a surge but no warning, and weeks with no surge but an issued warning.
- All metrics are defined as usual: $Sensitivity(Se) = TP/(TP + FN)$, $Specificity = TN/(TN + FP)$; Positive Predictive Value: $PPV = TP/(TP + FP)$, and F1 measurement, defined as the harmonic mean of sensitivity and PPV: $F1 = 2 \times (Se \times PPV)/(Se + PPV)$. Finally, the average reliability is defined as the average of the previous five metrics.

Codes

All calculations were perfomed using the Python language. For the ML methods, we used the Scikit-learn library, including the Isolation Forest, One Class SVM and Local Outlier Factor algorithms. We also used the PyOD library for COPOD and sklearn.metrics for the analyses. The NGM code has been created by the authors. All codes are available at https://github.com/cidacslab/MMAING.

Results

To assess the proposed ensemble, we used four complementary strategies comprising (i) data from IGRs of Brazilian capital cities in the period 2020 to 2023; (ii) data for a reference period of the COVID-19 pandemic between 2020 and 2022; (iii) a quantitative comparison between EWS generated by MMAING and those produced by traditional syndromic surveillance methods, between 2020 and 2023; and (iv) the use of manually labeled synthetic data. Results using the "balanced" BLCf were obtained in all four sub-sections, while STCf was used only in "Quantitative analysis based on synthetic data" section to compare two sets of MMAING results for synthetic series.

EWS for Brazilian capitals (2020–2023)

Figure 3 highlights results from URTI time series from three strategically chosen state capitals within the vast geography of Brazil: a) Belém (IGR 150001), capital of Pará, located in the north region; b) Belo Horizonte (IGR 310001), capital of Minas Gerais, located in the southeast region; and c) Porto Alegre (IGR 430001), capital of Rio Grande do Sul, located in the extreme south. This selection aims to illustrate both the diverse spectrum of epidemiological patterns that occurred over time as well as reflect the varied dynamics of URTIs in different regions of the country.

The obtained EWS for the selected Brazilian capitals, depicted in Fig. 3, fulfils the role of anticipating warning signals indicated by the increasing number of encounters that may correspond to a confirmed infection of a given respiratory syndrome. In addition, the SM sections C.2, C.3, and C.4 provide an extended analysis of other state capitals in Brazil, with figures and tables that outline EWS points for these regions, offering a comprehensive and comparative view of performance.

Analysis of EWS during the COVID-19 period (2020–2022)

Figure 4 exemplarly details some aspects of the pandemic in Brazil extending from 2020 to 2022. This period, which can be divided into four sub-intervals, was strategically chosen as it provides a set of URTI episodes with historical significance, allowing for a deep and informative analysis of a real pandemic.

We illustrate the analysis for Belo Horizonte (IGR 310001) as a reference. Our objective was to verify the applicability of MMAING in identifying the main trends and waves of the pandemic, as reflected in the time series of URTI encounters in PHC by the IGR. To achieve this objective, MMAING was trained with PHC data from the pre-COVID period (2017 to 2019) and subsequently applied and validated on data from the pandemic period (2020 to 2022). The trends and patterns detected for all IGRs are indicated in the SM section C.2.

As shown in Fig. 4, the pandemic started with the confirmation of the first COVID-19 case in Brazil on February 26, 2020, during epidemiological week number 9 [49, 51, 52]. MMAING detected consecutive EWS in weeks 7 and 8 and weeks 10 to 12 of 2020; this time window extends from February 9 to March 21, 2020. The analysis indicated that subsequent EWS, in weeks 27 to 29 of 2020, preceded the peak of the first wave of the disease, recorded between weeks 29 and 30, as presented in [49, 51].

The transition to the subsequent interval was signaled by EWS in weeks 46, 48, and 49 of 2020, coinciding with the emergence of the Gamma variant (November), which became the main variant in Brazilian territory two months later [49]. In 2021, four EWS between weeks 9 and 12 were detected before the peak of the second wave, which occurred between weeks 13 and 14.

The model also identified two EWS in weeks 45 and 47 of 2021, which anticipated the start of the third wave (week 49). This wave, driven by the Omicron variant, is marked by a drastic increase in COVID-19 cases from December 2021, with repercussions in January 2022, culminating in a peak between epidemiological weeks 5 and 6 of 2022 [49, 51]. Additionally, in November 2021 (weeks 43 to 48), a new subvariant (BQ.1) was identified, with significant growth over other circulating Omicron sublineages, leading to increased cases at the beginning of December. The EWS in weeks 40 and 42 preceded the new growth in the number of cases in December, as well as the emergence of the new sublineage, thereby reinforcing the effectiveness of MMAING in anticipating epidemiological trends. The official end of the COVID-19 pandemic announced by the Ministry of Health was effective from week 21 of 2022.

Nevertheless, a new wave marked by reinfections of Omicron and its sublineages was observed, with peaks in June and December. June, which begins in epidemiological week 22 of the beginning of the fourth wave, was marked by a significant increase in cases and deaths due to the BA.4 and BA.5 sublineages, responsible for 79% of positive COVID-19 tests [53]. According to our analysis, MMAING correctly signaled the beginning of the fourth wave with EWS between epidemiological weeks 18 and 21.

Our comprehensive analysis of the 27 IGRs revealed significant anticipation patterns throughout the pandemic period. It is important to note that PHC data moderately reflect the timeline of COVID-19 in Brazil. The number of encounters in PHC during the waves, as defined by official reports, varied significantly across

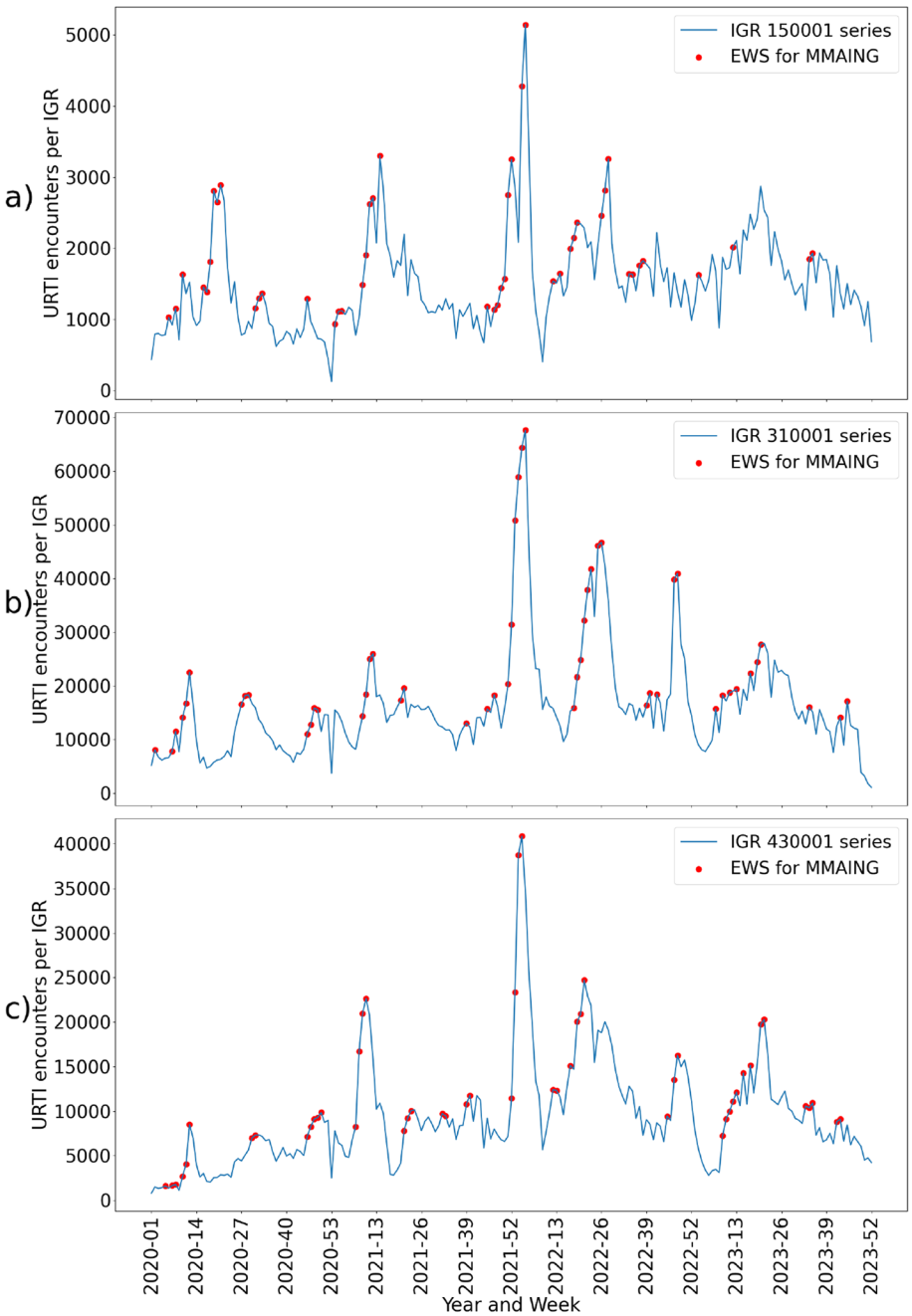

Fig. 3 Number of URTI encounters in three Brazilian capitals, from 2020 to 2023, with EWS issued by MMAING indicated by red dots. **a** Belém (north region, IGR 150001); **b** Belo Horizonte (southeast region, IGR 310001); and **c** Porto Alegre (south region, IGR 430001)

the IGRs. A plausible explanation for this phenomenon is the lack of standardization in public health recommendations in Brazil during the pandemic [52]. Despite data limitations, early detection across multiple IGRs demonstrated the broad applicability and utility of syndromic PHC data. We found that 16 (59.3%) of the IGRs provided EWS for the first pandemic wave, while 21 (77.8%) did so for the second wave. Similarly, 14 (48.2%) of the IGRs anticipated the third wave, and a significant majority of 25 (92.6%) anticipated the fourth wave. In all cases, EWS were detected between 0 and 4 weeks before the actual onset of each wave, with the best response observed between 1 to 2 weeks in advance, as detailed in the SM C.2. These findings highlight the MMAING's

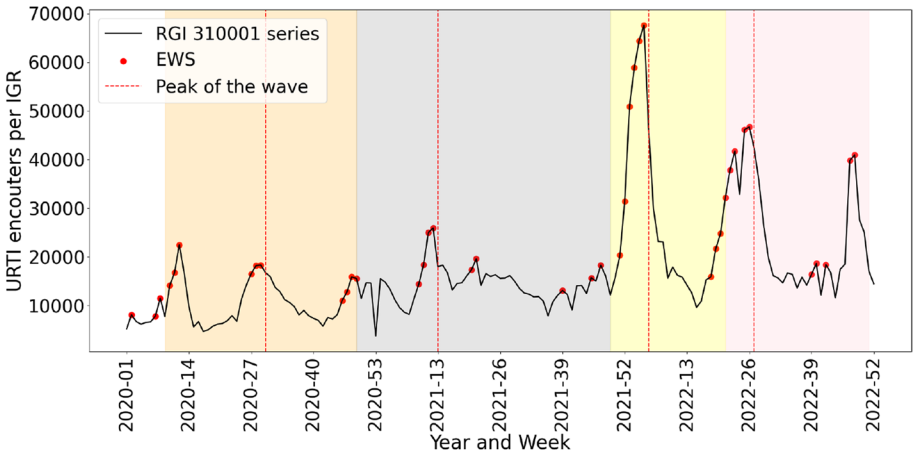

Fig. 4 Details of MMAING's results for URTI encounters in Belo Horizonte (as already displayed in Fig. 3-b), restricted to the COVID-19 period (2020–2022) and split into four successive time intervals shaded by different colors: initial outbreak (orange), second wave, marked by the arrival of Gamma variant (gray), third wave, influenced by Omicron variant (yellow), and fourth wave (pink), due to reinfections of Omicron and its sublineages. EWS points are indicated in red

capacity to provide timely and crucial information, proving effective in different scenarios. Consequently, the potential of PHC data for epidemiological surveillance is underscored.

Quantitative analysis based on synthetic data

Using synthetic data, it is possible to evaluate the capability of MMAING and test the robustness of the analyses, ensuring that the proposed methods are applicable to a wide range of scenarios. Figure 5 presents two synthetic series based on the Belo Horizonte time series (IGR 310001) as a reference, illustrating the EWS issued by MMAING. Each series presents distinct patterns, resulting from the random insertion of outbreaks (green) that vary in position and size.

Building on this, we proceed to evaluate the performance of MMAING under two different configurations, presented in Table1, and examine its adaptability to various scenarios, by employing a blind testing approach, which is characterized by splitting datasets for training and validation. The model is trained exclusively with actual data, whereas the validation is performed on a set of simulated time series. In this set, each sample consists of the average of synthetic series superimposed to localized high intensity random noise to simulate outbreaks, as detailed in "Synthetic data" section. This approach ensures that model performance evaluation is carried out under unbiased conditions, reflecting its ability to generalize to new data. To this end, 30 simulations were generated from each of the 27 IGR time series, resulting in 810 unique simulated time series, applying the two test configurations.

The results for the 27 IRGs synthetic series are then grouped by five Brazilian Geographic Regions: 7 for the North (N); 9 for the Northeast (NE); 4 for the Southeast (SE); 3 for the South (S) and finally, 4 for the Central-West (CW). The average results for BLCf and STCf are presented in Table 2.

Based on the results for the two configurations, we observed that MMAING presented a slight variation between PPV and specificity, while the main differences were for POD, sensitivity, and F1 score. It can be seen that the Central-West region stood out when evaluating each separate region, presenting the best evaluation rates. In both configurations, the results were higher than those from other regions and the general averages. On the other hand, the South Region obtained the lowest rates in both configurations, revealing a performance below average values.

The average results for BLCf – 0.86 for POD, 0.85 for PPV, 0.59 for sensitivity, 0.68 for F1 score, and 0.98 for specificity – indicated that MMAING held a high probability of outbreak detection, reasonable sensitivity, and a good PPV rate, suggesting that most EWS had a considerable probability of being true and detected correctly. As for STCf, it can be seen that the average result among regions was 0.73 for POD, 0.86 for PPV, 0.44 for sensitivity, 0.57 for F1 score, and 0.99 for specificity, highlighting that MMAING presented a more rigorous behavior, maintaining greater precision but reducing detection probability and sensitivity. This behavior ensures that any issued EWS is likely to be true. However, not all existing EWS in the series would be detected, providing a reliable and precise detection system.

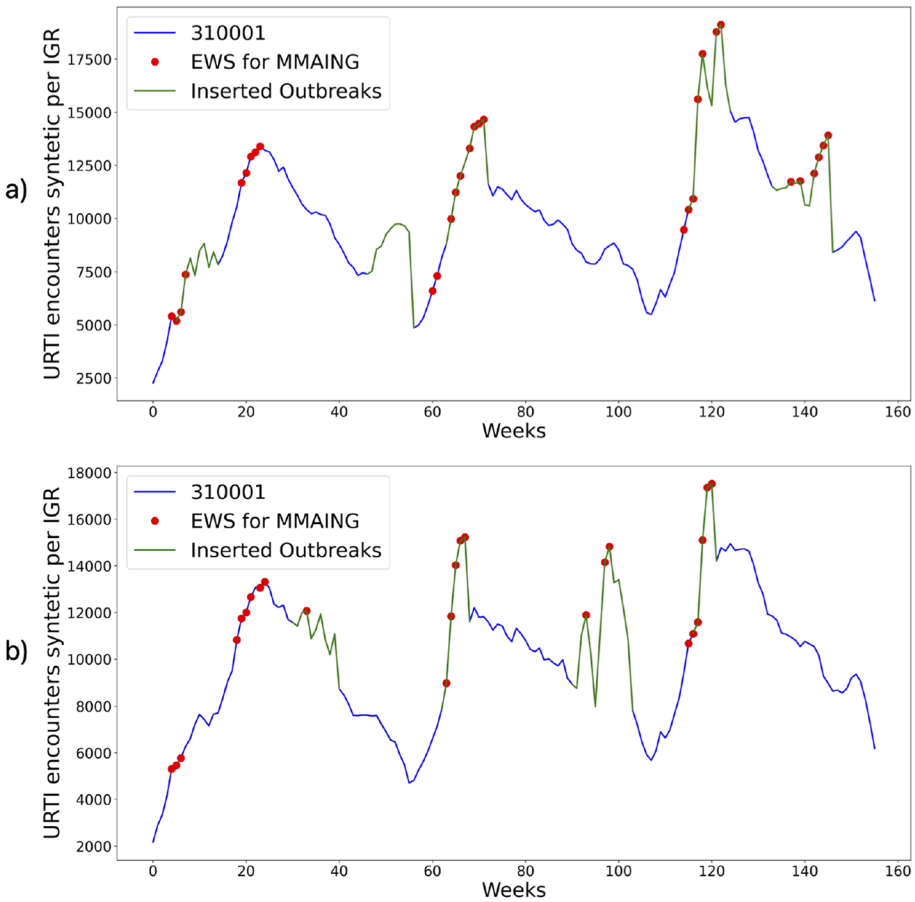

Fig. 5 Details of the MMAING results for synthetic URTI encounters in Belo Horizonte. In **a**, 5 outbreaks were inserted, and in **b**, 4 outbreaks, both highlighted in green. The early warning signals (EWS) issued by MMAING are represented by the red points

Table 2 Average performance metrics for MMAING – BLCf (top) and STCf (bottom)

Region	POD	PPV	Sensitivity	F1	Specificity
MMAING – BLCf					
North	0.83	0.85	0.56	0.67	0.98
Northeast	0.87	0.86	0.57	0.68	0.98
Southeast	0.86	0.87	0.58	0.69	0.98
South	0.83	0.77	0.57	0.64	0.97
Central-West	0.91	0.85	0.63	0.72	0.98
27 IGRs	0.86	0.85	0.59	0.68	0.98
MMAING – STCf					
North	0.76	0.86	0.44	0.56	0.99
Northeast	0.72	0.87	0.43	0.56	0.99
Southeast	0.77	0.86	0.47	0.60	0.99
South	0.67	0.75	0.36	0.47	0.98
Central-West	0.75	0.89	0.47	0.60	0.99
27 IGRs	0.73	0.87	0.44	0.57	0.99

The validation of MMAING was carried out by analyzing the synthetic series generated from real series. This approach ensured the selection of adequate configurations for the model, especially in BLCf parametrization, thus providing an adequate adaptation to its degree of applicability.

For further analysis, the results obtained for configurations 1 and 2 over the 30 simulations, grouped by the five regions of Brazil and individual synthetic IGRs, can be consulted in the SM section C.1.

Comparative analysis with EARS

We compared MMAING with EARS, a widely used method for detecting outbreaks and monitoring of weekly syndromic counts. We applied both methods to all the 27 Immediate Geographic Regions (IGRs) of Brazilian capital states, as they reflect syndromic activity in varied regional scenarios. The comparison aimed to evaluate our proposed ensemble's agreement and relative effectiveness.

Table 3 Average coincidence between MMAING and EARS in the period 2020 to 2023 for all 27 IGRs

Coincidence (%)			
EARS	**C1**	**C2**	**C3**
MMAING	63.41	66.23	39.65

Table 4 Annual counts of EWS detected by MMAING and EARS C2 from 2020 to 2023 (PHC data)

Year	MMAING	EARS C2	Coinciding EWS	Coincidence (%)
2020	337	389	251	74.48
2021	283	246	166	58.66
2022	345	372	255	73.91
2023	243	201	128	52.67

Table 5 Annual counts of EWS detected by MMAING and EARS from 2020 to 2023, per IGR (PHC data)

IGR	MMAING	EARS C2	Coinciding EWS	Coincidence (%)
150001	46	39	22	47.82
310001	51	53	39	76.48
430001	51	51	31	60.78
27 IGRs	1208	1208	800	66.23

Table 3 summarizes average coincidence between scores issued by MMAING and EARS variants, with C2 outperforming its counterparts. Tables 4 and 5 detail the comparison of MMAING and EARS C2 based on primary healthcare data across all years and for the selected IGRs, respectively.

For actual data, we compared MMAING and the three variations of EARS to identify the occurrence of coinciding EWS between 2020 and 2023 for all 27 IGRs. We detailed the congruence of EWS detected by EARS C2, organized by years, presenting concordance rates of 74.48% for the year 2020, 58.66% for 2021, 73.91% for 2022, and 52.67% for 2023 (Table 4). These results highlight the models' ability to consistently signalize the potential emergence of health problems and associated annual events, illustrating the coincidence between EWSs issued by MMAING and C2. Then, we counted the number of EWS that coincide and overlap for both methods in the 27 IGRs, as summarized in Table 5 for three selected IGRs: Belém (IGR 150001), Belo Horizonte (IGR 310001) and Porto Alegre (IGR 430001).

Furthermore, Fig. 6 illustrates the distribution of EWS detected for both methods for these three IGRs. In each case, two consecutive graphs show EWS

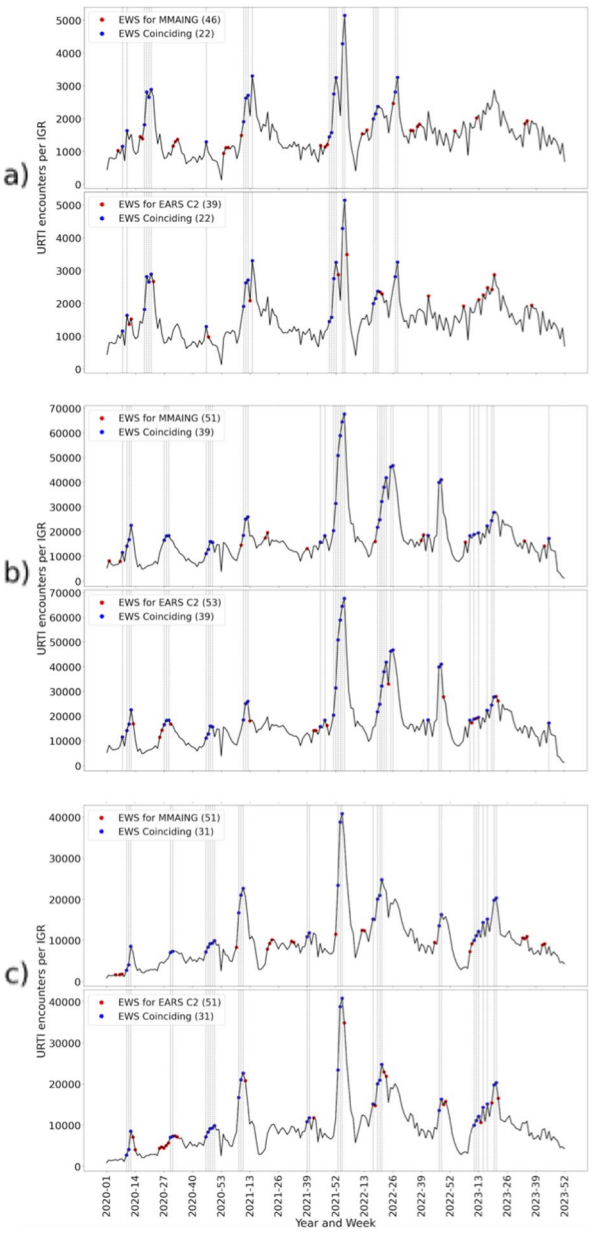

Fig. 6 EWS events for three IGRs: **a** Belém (IGR 150001); **b** Belo Horizonte (IGR 310001) and **c** Porto Alegre (IGR 430001) from 2020 to 2023. The top and bottom graphs indicate EARS and MMAING detections. Blue and red markers indicate events by both (EWS Coinciding) or just one method. The sum of red and blue markers corresponds to the total of EWS events

detections for MMAING (top) and EARS CS (bottom). Blue markers highlight 'EWS Coinciding' events when both models detected corresponding events guided by vertical dashed lines. The combination of red and blue markers provides an overview of total EWS. In the relationship between EWS (blue) detected by EARS C2 and the total (blue+red) from MMAING, we observed

that Belém recorded a correspondence of 47.82%, Belo Horizonte with 76.48% and Porto Alegre with 60.78%. These percentages emphasize the precision and variation in detecting EWS across models for different locations, highlighted in Table 5. See the SM section C.3 for details on other EARS variants.

Table 6 Average performance of MMAING and EARS (C1, C2, and C3) for IGRs of Belém (150001), Belo Horizonte (310001), and Porto Alegre (430001)

IGR	POD	PPV	Sensitivity	F1	Specificity
MMAING					
150001	0.75	0.85	0.50	0.62	0.99
310001	0.81	0.88	0.57	0.68	0.99
430001	0.87	0.83	0.59	0.68	0.98
27 IGRs	0.86	0.85	0.59	0.68	0.98
EARS C1					
150001	0.80	0.97	0.50	0.64	1.00
310001	0.80	0.97	0.53	0.68	1.00
430001	0.73	0.99	0.49	0.64	1.00
27 IGRs	0.83	0.97	0.53	0.68	1.00
EARS C2					
150001	0.80	0.86	0.54	0.66	0.99
310001	0.80	0.90	0.56	0.67	0.99
430001	0.73	0.93	0.52	0.65	0.99
27 IGRs	0.80	0.86	0.55	0.66	0.99
EARS C3					
150001	0.72	0.64	0.38	0.46	0.96
310001	0.71	0.67	0.41	0.50	0.97
430001	0.65	0.79	0.43	0.54	0.98
27 IGRs	0.75	0.66	0.40	0.49	0.97

Next, we simultaneously used synthetic data to evaluate the performance of the four methods - MMAING, EARS C1, C2, and C3 - in the 810 simulated series of single scenarios. We detail the results for the same three IGRs and the overall average of the 27 IGRs, as presented in Table 6. We observed that all methods, except EARS C3, performed similarly. MMAING presented better POD and sensitivity than the EARS variants; however, it lost PPV for EARS C1 and maintained close PPV with EARS C2. All methods had close F1 scores and specificities. MMAING was superior to EARS C3 in all metrics.

Figure 7 shows the distribution of performance metrics – POD, PPV, sensitivity, F1 score, and specificity – for the considered methods. Notably, MMAING demonstrated consistency across almost all metrics, sustaining itself with high scores and, with few variations, good detection probability, suggesting a robust balance between true positives and negatives EWS. In contrast, the EARS variants exhibited greater variations in their metrics, with EARS C1 and C2 possibly offering a trade-off between correctly identifying the positive EWS (PPV) and capturing as many positive EWS as possible (sensitivity), respectively. EARS C3 is distinguished by below-average metrics, except for its highly concentrated specificity, which may be preferable when false positive costs are pronounced. Analysis of the F1 score, which harmonizes PPV and sensitivity, suggests that all models maintained a moderately high level of balanced performance, with MMAING offering a slight advantage in consistency.

Figure 8 illustrates the Receiver Operating Characteristic (ROC) curve of the four detection models - MMAING, EARS C1, C2, and C3. MMAING proved robust, with an average Area Under the Curve (AUC) of 0.78 and an effective balance between sensitivity and

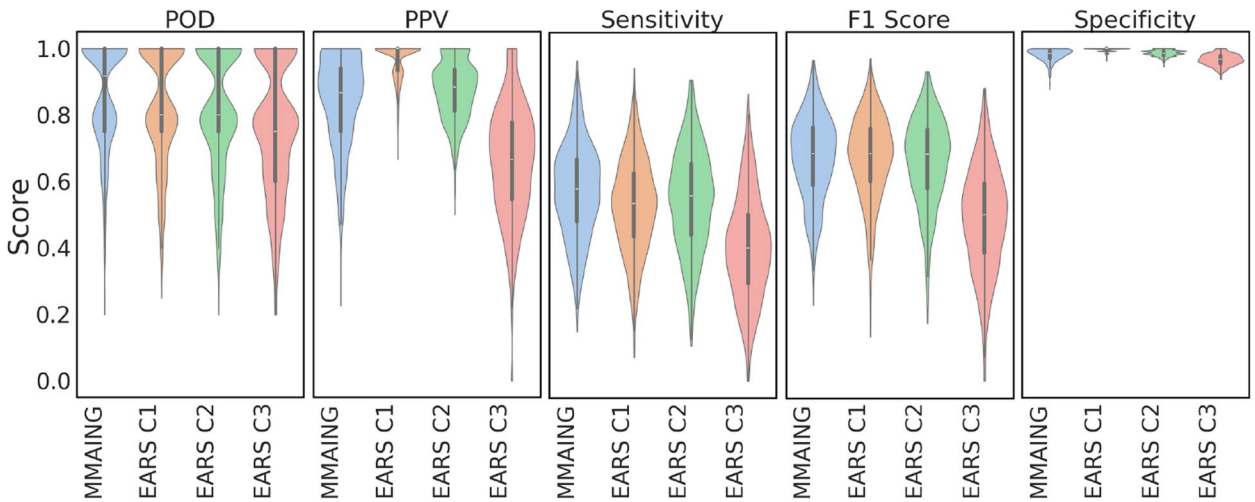

Fig. 7 Distribution of evaluated scores across the different models considering 30 simulations for 27 IGRs of Brazilian capital states

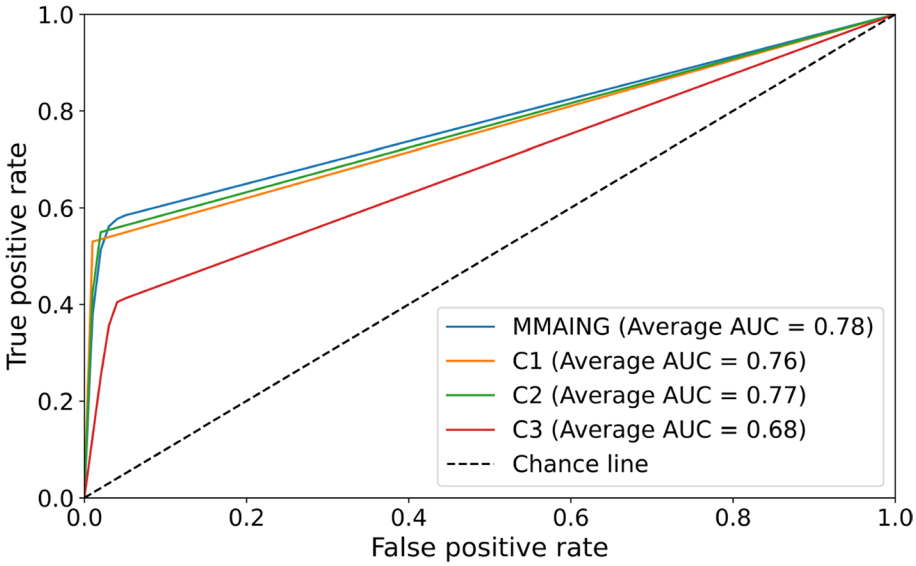

Fig. 8 Discriminatory capacity of MMAING and EARS variations (C1, C2 and C3) illustrated by their ROC curves

specificity, suggesting a consistent ability to discriminate between positive and negative classes. EARS C1 and C2 models, with AUCs of 0.76 and 0.77, respectively, performed similarly to MMAING, indicating that they also have good classification accuracy. However, the slight difference in AUC suggests that MMAING was slightly superior in overall performance. EARS C3, with an AUC of 0.69, presented a clear decrease in performance compared to the others, suggesting lower classification accuracy and a tendency to have a higher rate EWS of false positives.

Overall, ROC curve analysis suggested that MMAING may be the preferred choice for applications that balance correctly identifying positive EWS and preventing false positives.

Discussion

Outbreak detection algorithms play a crucial role in health surveillance, monitoring, and providing EWS to the risk of infectious diseases spreading, as exemplified during the COVID-19 pandemic. MMAING, which is a combination of ML algorithms for anomaly detection and mechanistic description by NGM using compartment models, produced relevant results based on PHC data and with improved reliability of EWS. The MMAING ensemble benefits from the individual ability of unsupervised ML models to detect anomalies in a diverse range of healthcare data, without the need to previously label the time series. This represents an advantage with respect to the use of labeled data in supervised learning for surveillance purposes, where there is not enough time to perform this task manually and where a quick response is crucial. By combining such models and adopting a voting mechanism, MMAING favors collective outlier patterns produced by the different algorithms over any single individual pattern that could go unnoticed as an outbreak [7]. The inclusion of dynamic information on a possible outbreak through the integration of NGM results into the ensemble decision, the constraint of a data-dependent upper limit (Eq. 2.2), and the choice of different configurations based on proper parameter values of each algorithm increase MMAING's overall reliability.

Regarding limitation aspects resulting from used data, like zero inflation, under reporting, and others, it is important to emphasize that, in general, the PHC data sets received directly from the MoH are not sparse. In addition, by explicitly restricting the analyzed data to the subset with positive DQI, greatly reduces the possibility of facing zero inflation issues. Finally, in the current work we consider aggregated data at the IGR level, which further reduces the mentioned problems.

MMAING results for all the 27 IGRs synthetics based on the "balanced" BLCf – probability of detection of 86%, PPV of 85%, and average reliability of 79% – suggested that our integrated methodology, if adequately configured and trained, can reliably forecast coming epidemics threats (see Table 2).

Using the same results, Fig. 4 highlights that MMAING could detect the emergence of all COVID-19 waves. For the initial one, it generated EWS even before the first case was confirmed, which aligns with recent findings [52]. In the sequence, MMAING identified a new wave corresponding to the introduction of the Gamma variant, characterized by the collapse of the health system

and local health crises [49]. The same happened with the early signaling of the third wave driven by the Omicron variant. Finally, MMAING correctly signaled the beginning of the fourth wave associated with a new subvariant (BQ.1) of Omicron and other circulating Omicron sublineages.

Besides the validation by comparison of results for two different data sets related to the same sequence of events, the systematic use of synthetic data provided quantitative measures that indicate MMAING's high reliability based on a set of metrics, namely POD, PPV, sensitivity, F1 score, and specificity. Remarkably, the compromise between POD, PPV, and sensitivity provides a good criterion for verifying the confidence of the results, as shown for the 27 IGRs of state capitals in the five geographic regions of Brazil. In general, MMAING stood out in terms of POD and sensitivity, with average values of 0.86 and 0.59, respectively, which were higher than other models, and presented satisfactory values for the other metrics: PPV of 0.84, F1 score of 0.68, and specificity of 0.98. These results are comparable to those of other models used in public health monitoring [10, 12, 27, 48, 54]

The fact that the random input of outbreaks on the synthetic series still depends on the actual data explains a slight but noticeable score variability across Brazil's geographic regions (see Table 2). The highest and lowest scores were obtained for the Central-West and South regions, also shown in Table 2; it is not clear how to explain this behavior, except for the small number of states in each geographic region (4 and 3 respectively), which enhanced fluctuations, and that MMAING had difficulties detecting small amplitude noise. Indeed, two IGRs in the Central-West (520001 and 530001) and one in the South (430001) regions led to results above the national average.

MMAING's reliability was also checked by comparing its results with those of the EARS method using real-world and simulated data. In the first case, the analysis of PHC data by year (2020–2023) and by IGR resulted in a high average coincidence with EARS C2 of 66.23%, with emphasis on years 2021 and 2022, corresponding to the peak and start of the declining COVID-19 period. The coincidences ranged from 52% to 75% per year and approximately from 47% to 84% per geographic region. As previously observed [52], EWS issued by EARS for 116 IGRs in Brazil ranged from 60% to 68%, which indicates coherence between EARS and MMAING methods. Regarding EARS C1 and C3 variations, their coincidence measures with MMAING stayed around 52 to 72% and 26 to 51% per year and 51 to 84% and 19 to 57% per geographic region, respectively.

A comparison between MMAING and EARS variants using real data sets revealed promising results as well as the potential applicability of the MMAING ensemble. Additionally, the effectiveness of the model was evaluated in a controlled environment with synthetic data, whose performance analysis was conducted based on the metrics shown in Table 6. POD guarantees the coverage of all possible outbreaks; sensitivity can identify outbreaks reliably and consistently in cases of frequent outbreaks; specificity is critical to reducing false EWS; PPV confirms the precision in signaling an outbreak [48], and the F1 score balances PPV and sensitivity, offering a harmonic measure of overall performance.

No model is generally better across all performance metrics (see Fig. 7 for all the 27 IGRs of Brazilian capital states). MMAING was more timely and had a slightly higher POD, which may be appropriate for early warning systems as it enables prompt implementation of effective interventions at the onset of an outbreak. Furthermore, it had the highest sensitivity combined with good PPV, high specificity, and satisfactory F1 score rate, which makes it more reliable in detecting actual outbreaks with a smaller margin of false EWS.

EARS C2 offered a balance between POD and PPV. It was the model that came closest to MMAING, presenting a balanced approach, sensitive enough to detect outbreaks with confidence but also specific enough to keep false EWS to an acceptable minimum.

Furthermore, Fig. 8 shows MMAING as the most balanced model with an average AUC of 0.78, which indicates good performance in distinguishing between correct detections and false EWS in terms of outbreak detection, being more reliable than the EARS C3, which has the worst average AUC of 0.68, partially more efficient than the C1 and C2 variants, with AUCs of 0.76 and 0.77, respectively. This comparative analysis highlighted that the choice between MMAING and EARS variants must be guided by application requirements, weighing each model's benefits and limitations against performance metrics relevant to a particular scenario.

The proposed method has limitations, the major one being its difficulty in identifying outbreaks of extremely low magnitude. Another limitation of our approach is that MMAING needs to consider the spatial spread of epidemics, although multiple locations can be analyzed separately, as performed in this work. In theory, the general framework used in MMAING could be extended to metapopulation models and incorporating spatial dependence [23].

To the best of our knowledge, this is the first time such different methods have been combined, which is a timely contribution to establishing more robust decision-making mechanisms. The proposed MMAING ensemble optimizes the balance between sensitivity and

specificity when issuing EWS, representing a significant innovation in the field of syndromic surveillance.

We hope that MMAING will form a valuable complement to existing point-source outbreak detection methods such as the EARS [6], modified Farrington [8], WSARE [7], ASMODEE [10], RAMMIE [9] and Generalized Linear Models [55].

Conclusion

The MMAING ensemble, proposed here as a method to detect EWS based on primary health care data, has proved robust according to measures commonly used to assess unsupervised learning models. By leveraging information on both data as well as the underlying dynamic transmission process, and employing different unsupervised machine learning methods integrated with a next-generation method, it offers a new methodological perspective that enriches the toolkits available for outbreak detection.

In this study, we highlight the effectiveness of MMAING in detecting EWS for epidemic outbreaks using primary healthcare data. The model outperformed other epidemiological surveillance methods, demonstrating robustness when using synthetic and actual data compared with EARS C1, C2, and C3 models.

While an AUC below 0.8 may be considered moderate in some fields, it is important to contextualize these results within outbreak detection, where data are noisy, and perfect classification is rarely achievable. We emphasize that MMAING's AUC of 0.79 is slightly below 0.8, representing a meaningful improvement over existing early detection systems.

Integrating unsupervised machine learning with dynamic systems techniques has proven to be an effective, robust, and promising approach in the surveillance field capable of anticipating epidemiological events. The results emphasize MMAING's ability to discern patterns in time series, which is essential for establishing a reliable protocol and anticipating outbreaks, significantly contributing to public health responses.

Therefore, MMAING emerges as a truly innovative tool in epidemiological surveillance, offering a more proactive and efficient approach that is sure to pique the interest of public health professionals and researchers.

Further questions related to the possible usefulness of MMAING in analyzing other scenarios based not only on primary health care or whether it can be a starting point for a broader spreading model are worthy of being developed. Indeed, it is necessary to recognize that outbreaks are influenced by external factors such as seasonality, climate, and mobility. Exploring causal relationships between these factors and anomaly signals could enhance outbreak prediction and interpretation. Future work could integrate environmental and behavioral data to assess whether observed anomalies align with known causal drivers of epidemic dynamics. For instance, weather influence becomes very important for the case of arboviruses outbreaks, is currently being considered, and will be reported in a next work.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-025-02542-0.

Supplementary Material 1.

Materials availability

Online Supplementary Material file provides detailed information on obtained results for all investigated IGRs.

Code availability

Code is available at https://github.com/cidacslab/MMAING.

Authors' contributions

RFSA, MEB, STRP, and ALGAC conceptualized the original work. MEB, RFSA, STRP, ERC, and DGFB proposed and developed the methodology. ERC, DGFB, and TCS implemented the codes and obtained raw results. ERC, DGFB, TCS, MEB, RFSA, STRP, ALGAC, ERC, DGFB, LL, and MG discussed, interpreted, and attested reliability of results. ERC and DGFB structured the original draft of the manuscript. MEB, RFSA, STRP, TCS, ALGAC, LL, MBN, PIPR, TCS substantively revised the manuscript. All authors read and approved the final manuscript.

Funding

This study is part of the Alert-Early System of Outbreaks with Pandemic Potential (AESOP;http://aesop.health) program funded by Fundação Oswaldo Cruz and the Rockefeller Foundation (grant 2023 PPI 007 awarded to MB-N). Additional support was provided by the National Institute of Science and Technology for Complex Systems (INCT-SC Brazil), the Rio de Janeiro State Research Support Foundation (FAPERJ), and the Postgraduate Program in Civil Engineering at the Federal University of Rio de Janeiro - UFRJ - COPPE. ALGAC, LL, PIPR, STRP, RFSA and MB-N are research fellows from the National Council for Scientific and Technological Development (CNPq), which also granted a PhD scholarship to DGFB. MEB is supported by The Royal Society (UK). TC-S acknowledges funding from the Royal Society (UK) (NIF/R1/231435). The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Data availability

Primary healthcare (PHC) data was provided by the AESOP project. Aggregated IGR data and synthetic data can be made available upon request to the authors. Codes are available at https://github.com/cidacslab/MMAING.

Declarations

Ethics approval and consent to participate

The study is based on secondary, aggregated, non-identified data, and was approved by the Ethical Review Board of Oswaldo Cruz Foundation - Fiocruz Bahia Regional Office, CAAE 61444122.0.0000.0040.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details
[1]Center of Data and Knowledge Integration for Health (CIDACS), Instituto Gonçalo Moniz, Fiocruz Bahia, Salvador, Brazil. [2]Physics Institute, Federal University of Bahia (UFBA), 40170-115 Salvador, Bahia, Brazil. [3]Department of Computing, Fluminense Federal University, 28895-532 Rio das Ostras RJ, Brazil. [4]Department of Civil Engineering, COPPE/Federal University of Rio de Janeiro, Rio de Janeiro, Brazil. [5]Department of Civil Engineering, Fluminense Federal University, Niterói, Brazil. [6]Medicine and Precision Public Health Laboratory (MeSP2), Instituto Gonçalo Moniz, Fiocruz Bahia, Salvador, Brazil. [7]Department of Statistics, London School of Economics and Political Science, London, UK.

Received: 7 May 2024 Accepted: 26 March 2025
Published online: 16 April 2025

References

1. Madhav N, Oppenheim B, Gallivan M, Mulembakani P, Rubin E, Wolfe N. 17. In: Pandemics: Risks, Impacts, and Mitigation. 2017. pp. 315–345. https://doi.org/10.1596/978-1-4648-0527-1_ch17.
2. Buehler JW, Hopkins RS, Overhage JM, Sosin DM, Tong V. Framework for evaluating public health surveillance systems for early detection of outbreaks: recommendations from the CDC Working Group. MMWR Recomm Rep. 2004;53(RR-5):1–11.
3. Wagner M, Tsui F, Cooper G, Espino JU, Harkema H, Levander J, et al. Probabilistic, Decision-theoretic Disease Surveillance and Control. Online J Public Health Inform. 2011;3(3):e61012.
4. Chiolero A, Buckeridge D. Glossary for public health surveillance in the age of data science. J Epidemiol Community Health. 2020;74(7):612–6.
5. Farrington CP, Andrews NJ, Beale AD, Catchpole MA. A Statistical Algorithm for the Early Detection of Outbreaks of Infectious Disease. J R Stat Soc Ser A Stat Soc. 1996;159(3):547–63.
6. Hutwagner L, Thompson W, Seeman GM, Treadwell T. The bioterrorism preparedness and response early aberration reporting system (EARS). J Urban Health. 2003;80:i89–96.
7. Wong WK, Moore A, Cooper G, Wagner M. What's strange about recent events (WSARE): an algorithm for the early detection of disease outbreaks. J Mach Learn Res. 2005;6:1961–98.
8. Noufaily A, Enki DG, Farrington P, Garthwaite P, Andrews N, Charlett A. An improved algorithm for outbreak detection in multiple surveillance systems. Stat Med. 2013;32(7):1206–22.
9. Morbey RA, Elliot AJ, Charlett A, Verlander NQ, Andrews N, Smith GE. The application of a novel 'rising activity, multi-level mixed effects, indicator emphasis' (RAMMIE) method for syndromic surveillance in England. Bioinformatics. 2015;31(22):3660–5.
10. Jombart T, Ghozzi S, Schumacher D, Taylor TJ, Leclerc QJ, Jit M, et al. Real-time monitoring of COVID-19 dynamics using automated trend fitting and anomaly detection. Philos Trans R Soc B. 1829;2021(376):20200266.
11. Unkel S, Farrington CP, Garthwaite PH, Robertson C, Andrews N. Statistical Methods for the Prospective Detection of Infectious Disease Outbreaks: A Review. J R Stat Soc Ser A Stat Soc. 2011;175(1):49–82.
12. Bédubourg G, Le Strat Y. Evaluation and comparison of statistical methods for early temporal detection of outbreaks: A simulation-based study. PLoS ONE. 2017;12(7):e0181227.
13. Faverjon C, Berezowski J. Choosing the best algorithm for event detection based on the intended application: a conceptual framework for syndromic surveillance. J Biomed Inform. 2018;85:126–35.
14. May L, Chretien JP, Pavlin JA. Beyond traditional surveillance: applying syndromic surveillance to developing settings - opportunities and challenges. BMC Public Health. 2009;9(1):242.
15. Petersen E, Koopmans M, Go U, Hamer DH, Petrosillo N, Castelli F, et al. Comparing SARS-CoV-2 with SARS-CoV and influenza pandemics. Lancet Infect Dis. 2020;20(9):e238–44.
16. Ramos PIP, Marcilio I, Bento AI, Penna GO, de Oliveira JF, Khouri R, et al. Combining digital and molecular approaches using health and alternate data sources in a next-generation surveillance system for anticipating outbreaks of pandemic potential. JMIR Public Health Surveill. 2024;10:e47673.
17. Liu FT, Ting KM, Zhou ZH. Isolation Forest. In: 2008 Eighth IEEE International Conference on Data Mining, 2008. pp. 413–422. https://doi.org/10.1109/ICDM.2008.17.
18. Liu FT, Ting KM, Zhou ZH. Isolation-based anomaly detection. ACM Trans Knowl Discov Data. 2012;6(1):1–39.
19. Breunig MM, Kriegel HP, Ng RT, Sander J. LOF: identifying density-based local outliers. SIGMOD Rec. 2000;29(2):93–104.
20. Schölkopf B, Platt JC, Shawe-Taylor J, Smola AJ, Williamson RC. Estimating the support of a high-dimensional distribution. Neural Comput. 2001;13(7):1443–71.
21. Li Z, Zhao Y, Botta N, Ionescu C, Hu X, COPOD: copula-based outlier detection. In: 2020 IEEE international conference on data mining (ICDM). IEEE; 2020. pp. 1118–23.
22. Wallinga J, Lipsitch M. How generation intervals shape the relationship between growth rates and reproductive numbers. Proc R Soc B Biol Sci. 2007;274(1609):599–604.
23. Jorge DCP, Oliveira JF, Miranda JGV, Andrade RFS, Pinho STR. Estimating the effective reproduction number for heterogeneous models using incidence data. R Soc Open Sci. 2022;9(9):220005.
24. eGestor Informação e Gestão da Atenção Basica. Cobertura da Atenção Primária. 2024. https://sisaps.saude.gov.br/sistemas/esusaps/. Accessed Jan 2024.
25. Florentino PTV, Bertoldo-Junior J, Barbosa GCG, et al. Impact of Primary Health Care Data Quality on Infectious Disease Surveillance in Brazil: Case Study, JMIR Public Health Surveill 2025;11:e67050. https://doi.org/10.2196/67050. PMID: 39983017 PMCID: 11870279.
26. Lam L, Suen S. Application of majority voting to pattern recognition: an analysis of its behavior and performance. IEEE Trans Syst Man Cybern Part A Syst Hum. 1997;27(5):553–68.
27. Texier G, Allodji RS, Diop L, Meynard JB, Pellegrin L, Chaudet H. Using decision fusion methods to improve outbreak detection in disease surveillance. BMC Med Inform Decis Mak. 2019;19(1):1–11.
28. Schruben L. Confidence interval estimation using standardized time series. Oper Res. 1983;31(6):1090–108.
29. Bhaskaran K, Gasparrini A, Hajat S, Smeeth L, Armstrong B. Time series regression studies in environmental epidemiology. Int J Epidemiol. 2013;42(4):1187–95.
30. Neill DB. An empirical comparison of spatial scan statistics for outbreak detection. Int J Health Geogr. 2009;8:1–16.
31. Wolpert DH. Stacked generalization. Neural Netw. 1992;5(2):241–59.
32. Goldstein M, Uchida S. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. PLoS ONE. 2016;11(4):e0152173.
33. Ersoy P. Evolution of Outlier Algorithms for Anomaly Detection. Manch J Artif Intell Appl Sci. 2021;2(1).
34. Durante F, Sempi C. Copula theory: an introduction. In: Copula Theory and Its Applications: Proceedings of the Workshop Held in Warsaw, 25-26 September 2009. Springer; 2010. pp. 3–31.
35. Diekmann O, Heesterbeek JAP. Mathematical epidemiology of infectious diseases: model building, analysis and interpretation, vol 5. John Wiley & Sons; 2000.
36. Van den Driessche P, Watmough J. Reproduction numbers and subthreshold endemic equilibria for compartmental models of disease transmission. Math Biosci. 2002;180(1–2):29–48.
37. Nishiura H, Chowell G. The effective reproduction number as a prelude to statistical estimation of time-dependent epidemic trends. Math Stat Estimation Approaches Epidemiol. 2009:103–121.
38. Champredon D, Dushoff J. Intrinsic and realized generation intervals in infectious-disease transmission. Proc R Soc B Biol Sci. 1821;2015(282):20152026.
39. Huber JH, Johnston GL, Greenhouse B, Smith DL, Perkins TA. Quantitative, model-based estimates of variability in the generation and serial intervals of Plasmodium falciparum malaria. Malar J. 2016;15(1):1–12.
40. Lessler J, Ott CT, Carcelen AC, Konikoff JM, Williamson J, Bi Q, et al. Times to key events in Zika virus infection and implications for blood donation: a systematic review. Bull World Health Organ. 2016;94(11):841.
41. Park SW, Champredon D, Dushoff J. Inferring generation-interval distributions from contact-tracing data. J R Soc Interface. 2020;17(167):20190719.
42. Nishiura H, Linton NM, Akhmetzhanov AR. Serial interval of novel coronavirus (COVID-19) infections. Int J Infect Dis. 2020;93:284–6.
43. Kermack WO, McKendrick AG. A contribution to the mathematical theory of epidemics. Proc R Soc Lond Ser A Containing Pap Math Phys Character. 1927;115(772):700–21.

44. Fraser C. Estimating individual and household reproduction numbers in an emerging epidemic. PLoS ONE. 2007;2(8):e758.
45. Proverbio D, Kemp F, Magni S, Gonçalves J. Performance of early warning signals for disease re-emergence: a case study on COVID-19 data. PLoS Comput Biol. 2022;18(3):e1009958.
46. Jombart T, Jarvis CI, Mesfin S, Tabal N, Mossoko M, Mpia LM, et al. The cost of insecurity: from flare-up to control of a major Ebola virus disease hotspot during the outbreak in the Democratic Republic of the Congo, 2019. Eurosurveillance. 2020;25(2):1900735.
47. Fricker RD Jr, Hegler BL, Dunfee DA. Comparing syndromic surveillance detection methods: EARS'versus a CUSUM-based methodology. Stat Med. 2008;27(17):3407–29.
48. Noufaily A, Morbey RA, Colón-González FJ, Elliot AJ, Smith GE, Lake IR, et al. Comparison of statistical algorithms for daily syndromic surveillance aberration detection. Bioinformatics. 2019;35(17):3110–8.
49. Observatório Fiocruz COVID-19. Boletim especial - Balanço de dois anos da pandemia Covid-19: Janeiro de 2020 a Janeiro de 2022. Rio de Janeiro; 2022. https://www.arca.fiocruz.br/handle/icict/55828 Accessed Jan 2024.
50. Powers D. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. J Mach Learn Technol. 2011;2(1):37–63.
51. Moura EC, Cortez-Escalante J, Cavalcante FV, Barreto ICdHC, Sanchez MN, Santos LMP. Covid-19: evolução temporal e imunização nas três ondas epidemiológicas, Brasil, 2020–2022. Rev Saude Publica. 2022;56:105.
52. Cerqueira-Silva T, Marcilio I, de Araújo Oliveira V, Florentino PTV, Penna GO, Ramos PIP, et al. Early detection of respiratory disease outbreaks through primary healthcare data. J Glob Health. 2023;13.
53. todos pela saúde I. Em duas semanas, identificação de BA.4 e BA.5 passa de 44% para 79,3% das amostras positivas de SARS-CoV-2. 2022. https://www.itps.org.br/pesquisa/monitoramento-das-variantes-do-sars-cov-2. Accessed Jan 2024.
54. Craig AT, Leong RNF, Donoghoe MW, Muscatello D, Mojica VJC, Octavo CJM. Comparison of statistical methods for the early detection of disease outbreaks in small population settings. IJID Reg. 2023;8:157–63.
55. Zareie B, Poorolajal J, Roshani A, Karami M. Outbreak detection algorithms based on generalized linear model: a review with new practical examples. BMC Med Res Methodol. 2023;23(1):235.

Publisher's Note