Türkiye Klinikleri Biyoistatistik Dergisi
Turkiye Klinikleri Journal of Biostatistics Turkiye Klinikleri J Biostat. 2022;14(2):129-36

ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH

DOI: 10.5336/biostatic.2022-91708

Comparative Analysis of Statistical and Supervised Learning Algorithms for Outbreak Detection in the Syndromic Surveillance of Influenza-Like Illness: A Methodological Research

Grip Benzeri Hastalıkların Sendromik Sürveyansında Salgın Tespiti için İstatistiksel ve Denetimli Öğrenme Algoritmalarının Karşılaştırmalı Analizi: Metodolojik Bir Araştırma

Hasan Ali ÖZKAN[a], Mehmet Kadri GOFRALILAR[a], Zeynep Filiz EREN[a]

[a]Department of Computer Engineering, Muğla Sıtkı Koçman University, Muğla, Türkiye

ABSTRACT Objective: Public health authorities monitor epidemiological syndromes to provide early alerts of anomalies. A variety of approaches are applied for effective surveillance systems for influenza like illness (ILI). The present study systematically scores the accuracy of algorithms used for automated and prospective infectious-disease-outbreak detection. Another objective is to improve the performance of machine-learning (ML) approaches through statistical learning. **Material and Methods:** In order to reflect various situations, the volume and the size of the outbreak is chosen different for each simulation. We simulate 20 yearly sets of "daily ILI visit" to emergency department (ED), which includes seasonal outbreaks as well as unusual outbreaks of varying duration and magnitude. We search which biosurveillance algorithms work best across hidden "unusual outbreaks". **Results:** In terms of timeliness, both settings of kNN (res-raw), RF (res-raw), and LR-raw have the best performance. All ML algorithms have sensitivity results greater than 0.90, where SVM-res (0.97), EWMA (0.96), CUSUM-moderate (0.95) are the best algorithms in terms of specificity. ML algorithms all give better performance with an integrated fitted regression model. The methods which have high sensitivity and specificity together is SVM-res (0.90 and 0.97), and LR-res (0.92 and 0.83). **Conclusion:** The results verified that ML algorithms integrated with statistical methods can be applied to daily ED data and can be used as a real-time surveillance method for prospective monitoring of ILI cases in the emergency setting. This study can contribute to the early detection of hidden unusual outbreaks for epidemiological studies.

Keywords: Public health surveillance; outbreak detection;
CUSUM; EWMA; machine learning algorithms

ÖZET Amaç: Halk sağlığı yetkilileri, sıradışı gözlemler oluşması durumunda, epidemiyolojik sürveyans ile erken uyarı elde etmeyi hedefler. İnfluenza benzeri hastalıklara (influenza-like illness, ILI) ait etkin sürveyans sistemleri için çeşitli yaklaşımlar bulunmaktadır. Bu çalışmada amaç, bulaşıcı hastalık ve salgınların otomatik ve ileriye dönük tespitinde kullanılan algoritmaların gücünü sistematik bir şekilde incelemektir. Bir diğer amaç ise, istatistiksel öğrenme yoluyla makine öğrenimi (machine-learning, ML) yaklaşımlarının performansını iyileştirmektir. **Gereç ve Yöntemler:** Yirmi adet bir yıl uzunluğunda "ILI"ye bağlı günlük acil servis ziyaretleri" türetilmiştir. Türetilen veriler, mevsimsel salgınların yanı sıra, değişik hacimde ve boyutta olağandışı salgınları da içermektedir. Gizli "olağandışı salgınların" tespitinde hangi biyo-gözetim algoritmalarının en iyi sonucu verdiği araştırılmıştır. **Bulgular:** Zamanlılık açısından, Knn (res-raw), RF (res-raw) ve LR-raw uygulamaları en iyi performansa sahiptir. Tüm ML algoritmaları 0,90'dan büyük duyarlılığa sahiptir. SVM-res (0.97), EWMA (0.96), CUSUM-moderate (0.95) özgüllük açısından en iyi algoritmalardır. ML algoritmalarının tümü, regresyon modeliyle entegre şekilde kullanıldığında daha iyi performans vermektedir. Duyarlılığı ve özgüllüğü aynı anda yüksek olan yöntemler SVM-res (0.90 ve 0.97) ve LR-res (0.92 ve 0.83)'tir. **Sonuç:** ML algoritmaları, istatistiksel yöntemlerle entegre edilerek günlük hasta verilerine uygulandığında yüksek performans göstermektedir. Gerçek zamanlı sürveyans sistemi geliştirirken kullanılacak olan algoritmalar, araştırmada hangi performans ölçüsünün önemli olduğuna göre seçilebilir. Çalışma, epidemiyolojik çalışmalarda, gizli olağandışı salgınların erken tespitine katkıda bulunacak niteliktedir.

Anahtar kelimeler: Halk sağlığı sürveyansı; salgın tespiti;
CUSUM; EWMA; makine öğrenmesi yöntemleri

Correspondence: Zeynep Filiz EREN
Department of Computer Engineering, Muğla Sıtkı Koçman University, Muğla, Türkiye
E-mail: zeynepfilizeren@mu.edu.tr

Peer review under responsibility of Turkiye Klinikleri Journal of Biostatistics.

Received: 01 Jun 2022 ***Received in revised form:*** 10 Jun 2022 ***Accepted:*** 10 Jun 2022 ***Available online:*** 13 Jun 2022

In public health monitoring, epidemiological surveillance has gained more importance in the recent years. Due to the increase in the world population and the increase in environmental risks, infections are spreading rapidly.[1] Early detection of emerging infectious diseases provides timely information in case of a new epidemic or bioterrorism and reduces the impact of these diseases with early intervention in epidemics. Health authorities establish automatic and prospective surveillance systems to detect outbreaks and pandemics. For this purpose, various approaches are applied to public health data. For a complete review of statistical methods and implementations see.[2-4]

Influenza like illness (ILI) and respiratory tract infections, such as coronavirus disease-2019 (COVID-19), are among the most common acute infections worldwide which leads to considerable morbidity, a considerable economic burden to health care and days lost from work and school.[5] Seasonal trends may result in extensive wait times and overcapacity in emergency departments (ED), which directly upsets operational expenses and quality of service. Early outbreak detection methods can help improve management of ED crowding when effective intervention approaches and staffing response are applied. Real-time syndromic surveillance systems have the potential to detect outbreaks of ILI before conventional diagnosis and laboratory-based surveillance identifies them.[6]

A number of outbreak detection methods, such as statistical process control methods, time series methods, and machine learning techniques have been developed for effective temporal disease monitoring.[1,7-9] Automated disease surveillance systems use variants of these methods to detect natural and human made outbreaks. The aim of this paper is to compare the performance of six outbreak detection algorithms with a total of 11 settings in monitoring daily syndromic data which is simulated as to represent a real-life syndromic activity. The comparisons indicate optimal candidate algorithms for real-time monitoring. The results of this study may be useful for designing early warning system for epidemiological surveillance.

MATERIAL AND METHODS

In this section, the algorithms used in the study are briefly described. These algorithms provide the public health authorities the first indication of such anomalies, raising an alarm if epidemic data becomes abnormal.

STUDY OVERVIEW

The proposed approach for detecting respiratory outbreaks consists of three stages: i) We generate 20 yearly sets of "daily time series data" for daily ILI visits at an ED. Ten sets are used for training and 10 for testing. ii) First, a regression model with autoregressive integrated moving average (ARIMA) error terms is fitted to predict daily pediatric ED visits for respiratory syndromes, and calculate the associated residuals. Then we monitor the residuals with residual EWMA and CUSUM charts to obtain signals for unusual respiratory outbreaks. iii) We use four machine learning algorithms for classification and determine the signals for abnormal respiratory outbreaks. These steps are described in Section 2.3.

All algorithms are conducted on a PC with 2.3 GHz Dual-Core Intel core i5 processor and 8 GB of RAM. The surveillance models were constructed using R programming language (release 4.2.0) and with packages e1071, caret, randomForest and forecast.[10-13]

DATA

We generated data, based on the simulation algorithm developed by Noufaily et al.[14] Compatible with this study, we generated 20 yearly sets of "ED logs of daily ILI visits", which includes a total of 295,564 patients with an average of 40 daily counts, with range [0, 314]. We used a negative binomial model with mean μ and variance $\emptyset\mu$, where $\emptyset$ is a dispersion parameter ($\emptyset \geq 1$).

We take the start day of the ILI season as September 1, since ILI in Türkiye onset in September, peak until March, and offset in May.[15]

The simulations are set to reflect the ED report data based on 7 days of the week. We have 7,306 days (5,218 weekdays, 2,088 weekends) over 20 years, of which 132 are public holidays. For each yearly set, public holidays are defined manually, which is consistent with the last 20 years. The simulation study includes 3 steps. For each step, weekdays include a lower volume of reports where weekends and public holidays almost double that volume.

1. *Baseline data* without any outbreak is generated.

2. *Seasonal outbreak data* is generated, which has a range of 8 weeks on average. This data mimics the "seasonal influenza" which we expect to experience every year. Here, the outbreak sizes are simulated using a Poisson distribution and the start day is determined randomly according to a lognormal distribution.

3. *Spiked outbreak data* is generated, similar to step 2. This data mimics the unexpected situations such as a new pandemic/epidemic like H1N1, COVID-19; or bio-terrorist incidences which are hidden in seasonal outbreak periods.

The simulated data captures a wide range of data structures, to identify which algorithms work best across the daily syndromic surveillance for ILI with different outbreak period, volume and size. Out of 20 yearly sets of "daily simulated data", 10 sets are shown in Figure 1 explicitly as: (a) the seasonal outbreaks, (b) the spiked outbreaks and (c) the total cases.

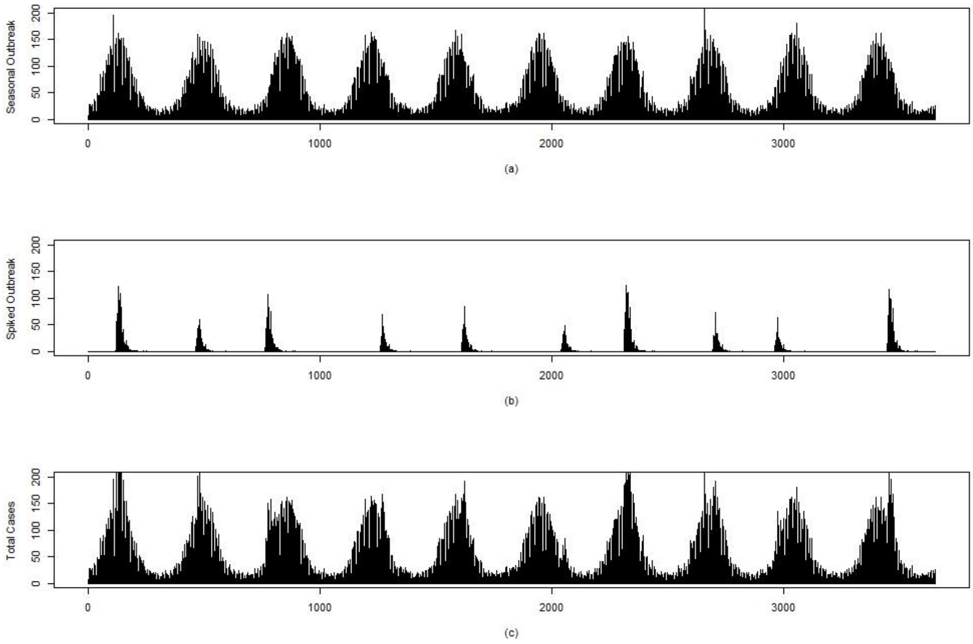

FIGURE 1: Simulated daily data which includes 3,653 days of (a) the seasonal outbreaks, (b) the spiked outbreaks and (c) total cases.

THE ALGORITHMS

This section describes the algorithms we implemented and compared in this study. These algorithms provide public health authorities with the first indication of unusual activity.

REGRESSION MODEL WITH ARIMA ERROR TERMS AND APPLICATION TO THE EMPIRICAL DATA

As the daily ILI counts constitute a time series with seasonal variation, we use a regression model with ARIMA error terms to fit the data. In order to better meet the assumptions, the response variable can be transformed. The general form of the model is given as:

$$Z_t = \mu_t + \varepsilon_t,\ t = 1, 2, \mathrm{K} \qquad (1)$$

where $Z_t = \log(Y_t)$ is the transformed response (Y_t). Here, Y_t is the daily ED visits of ILI cases, μ_t is the mean of response day t, which depends on a set of predictor variables, and ε_t is the error term which follows an ARIMA(p, d, q) process, where parameters (p, d, q) are non-negative integers, p is the order of the autoregressive part, d is the degree of differencing part, and q is the order of the moving-average part of the model.

The predictors of this model are determined as day-of-the-week, month-of-the-year, holiday and trend effects. Dummy variables are set for the day-of-the-week effect as $D_1, \ldots, D_6$ by setting Wednesday as the reference day, month-of-the-year effect as $M_1, \ldots, M_{11}$ by setting June as the reference month, and holiday as H for all national holidays. The dummies are indicators, where for example $D_i = 1$ on the relevant weekday and $D_i = 0$ otherwise. The first and second order sine and cosine functions are incorporated in the model to count the seasonal effect in. Second order trigonometric terms are included to show the lingering effect appears after each respiratory outbreak. Finally, trend effect is included with the term t and final model is presented as follows:

$$\begin{aligned} Z_t = {} & \beta_0 + \sum_{i=1}^{6} \beta_i D_i + \sum_{j=1}^{11} \beta_{6+j} M_j + \beta_{18} H + \beta_{19} t + \\ & \beta_{20} \sin(\frac{2\pi t}{365.25}) + \beta_{21} \cos(\frac{2\pi t}{365.25}) + \beta_{22} \sin(\frac{4\pi t}{365.25}) + \beta_{23} \cos(\frac{4\pi t}{365.25}) + \varepsilon_t \end{aligned} \qquad (2)$$

After fitting the regression model with an ARIMA error term to test data, the residuals can be acquired. As fitting this model yields identically distributed and independent residuals, the standardized residuals (r_t) can be used to construct different surveillance algorithms to monitor daily respiratory counts. In this study, EWMA and CUSUM settings are implemented with standardized residual data, where ML algorithms are implemented with i) raw data, ii) standardized residual data.

RESIDUAL EWMA

EWMA charts are widely used for monitoring epidemics.[16] The residual EWMA statistic (E_t) implemented to this problem is a weighted average of all previous observed y_t values as:

$$E_t = \lambda\, y_t + (1 - \lambda)\, E_{t-1} \text{ for } t = 1,2, \ldots, \qquad (3)$$

where y_t is the observation on day t, $0 < \lambda \leq 1$, and $E_0 = 0$ (as starting value). The smoothing constant, λ defines the level of memory of past counts. When choosing the threshold, h, one needs to decide between protection from false alarms and the ability to detect real changes quickly. It is suggested to use a threshold between 6 and 7. We take $\lambda = 0.2$ and $h = 6.5$ in this study

In an EWMA chart, if E_t exceeds the defined threshold (h), a signal is obtained.

RESIDUAL CUSUM

The residual CUSUM implemented to this problem is an upper-sided standardized CUSUM with a CUSUM statistic at time t:

$$C_t = \max\,(0; C_{t-1} + r_t \ - k) \qquad (4)$$

Here, $C_0 = 0$ is the initial value and k is the reference value used to design a CUSUM statistic which is sensitive to a specified mean shift size. If C_t exceeds a defined threshold (h), a signal is obtained. The essential idea is that an outbreak will result in larger than expected positive forecast errors and their values will accumulate in the CUSUM statistic and eventually result in an outbreak signal.

When applying the residual CUSUMs, the practitioner may apply variants of the monitoring approach with different design settings. We used aggressive and routine settings with different values of parameters k and h.[17]

Typically, resetting after every signal (always-reset rule) is used to reset the statistics. In this study, we use this general approach along with a more complicated approach to reset C_t, and call it "auto-reset rule." Auto-resetting can be completed using a linear regression with sliding windows plugin proposed by.[18] The method basically regresses $C_t, C_{t-1}, \dots, C_{t-v}$ on $t, t-1, \dots, t-v$. If the estimated slope of the regression model is less than zero and statistically significant, then a downward trend is said to be detected and expressed as the end of an outbreak. C_t is then reset to zero after these conditions are satisfied. In this study, s is chosen as 7 days.

MACHINE LEARNING ALGORITHMS

Machine learning is a cluster of scientific methods that tries to discover a pattern in the input data after preprocessing step. In this study, we used supervised learning since we generate a model from the input data to classify the observations according to a label (whether there is an unusual outbreak or not).

There are various algorithms under this category, but according to previous studies in this area, support vector machine (SVM), logistic regression (LR), k-nearest neighbors (kNN), and random forest (RF) are the most widely used predicting algorithms among others.[9]

A brief description of these algorithms are presented in this section. More comprehensive review on supervised machine learning algorithms can be found in.[19]

SVM finds a hyperplane that distinctly classifies the data points with the largest margin.

LR is used for binary classification problems. LR can be explained as linear regression applied to classification problems.

kNN is based on placing unknown samples into the class of their nearest neighbors. The majority voting kNN rule generalizes this concept by finding the kNN and choosing the class that is most frequent among them.

RF is a successful classification and regression method. It combines several randomized decision trees and aggregates their predictions by averaging.

RESULTS

PERFORMANCE MEASURES

The performance of the algorithms are evaluated according to their capability of detecting the spiked outbreaks in the presence of the usual seasonal outbreaks. The performance comparison of algorithms is implemented using timeliness, sensitivity, specificity, power of detection (POD) and positive predictive value (PPV or precision) metrics. For each metric, we take 10 sets of simulations into account.

Timeliness is the average time delay, an algorithm gives a signal from the onset of the spiked outbreak period. Timeliness is 0 if the outbreak is detected on the first day and 1 when the outbreak is not detected at all.

$$Timeliness = \frac{\sum_{sim=1}^{10}(t_{outbreak\ detection} - t_{onset})/(total\ spiked\ outbreak\ days)}{10} \qquad (5)$$

Sensitivity is the number of signals in outbreak period divided by the total outbreak days:

$$Sensitivity = \frac{\#\ of\ signals\ among\ spiked\ outbreak\ period}{\#\ of\ outbreak\ days} \qquad (6)$$

Specificity is the number of no-signal days divided by the total non-outbreak days:

$$Specificity = \frac{\#\ of\ no\text{–}signals\ among\ non\text{–}spiked\ outbreak\ period}{\#\ of\ non\text{–}outbreak\ days} \qquad (7)$$

POD is the probability of detecting the outbreak, by having a signal at least once during a spiked outbreak:

$$POD = \frac{\#\ of\ spiked\ outbreaks\ flagged\ at\ least\ once}{10} \qquad (8)$$

PPV is the proportion of detected outbreaks (true positives):

$$PPV = \frac{\#\ of\ true\ positives}{\#\ of\ positives} \qquad (9)$$

SIMULATION RESULTS

We implemented EWMA and 2 settings of CUSUM algorithms (CUSUM moderate, CUSUM aggressive) to standardized residuals of the fitted regression model; then four machine learning algorithms are applied to both standardized residuals data (SVM-res, LR-res, kNN-res, and RF-res) and raw data which includes daily counts of respiratory patients (SVM-raw, LR-raw, kNN-raw, and RF-raw). Totally, we have 11 algorithms applied to 10 yearly sets of simulations.

Figure 2 shows the average detection capability of algorithms, denoted with their settings, in terms of 6 performance metrics. In terms timeliness, both settings of kNN, RF and LR-raw have the best performance (≅ 0); and are followed by CUSUM settings, EWMA and so on. All ML algorithms have sensitivity results greater than 0.90 (with RF and kNN settings maximum), where CUSUM aggressive, CUSUM moderate and EWMA have lower sensitivity measures as 0.68, 0.46 and 0.56, respectively. SVM-res (0.97), EWMA (0.96), CUSUM-moderate (0.95) are the best algorithms in terms of specificity, followed by CUSUM-aggressive (0.87), LR-res (0.83), RF-res (0.61), etc. For specificity, ML algorithms give better performance when they are applied to the residuals data.

The PPV of both settings of SVM and LR give the best performance (≥0.98). The POD results of each algorithm is equal to 1, except SVM-res (0.90), RF-raw (0.30), and SVM-raw (0.30).

After implementing the fitted regression model and calculating the residual data, the time difference of the algorithms are recorded in terms of seconds. Both settings of kNN algorithms are the fastest ones (0.02 seconds), followed by LR, CUSUM, SVM and RF settings with a maximum of 7.68 seconds.

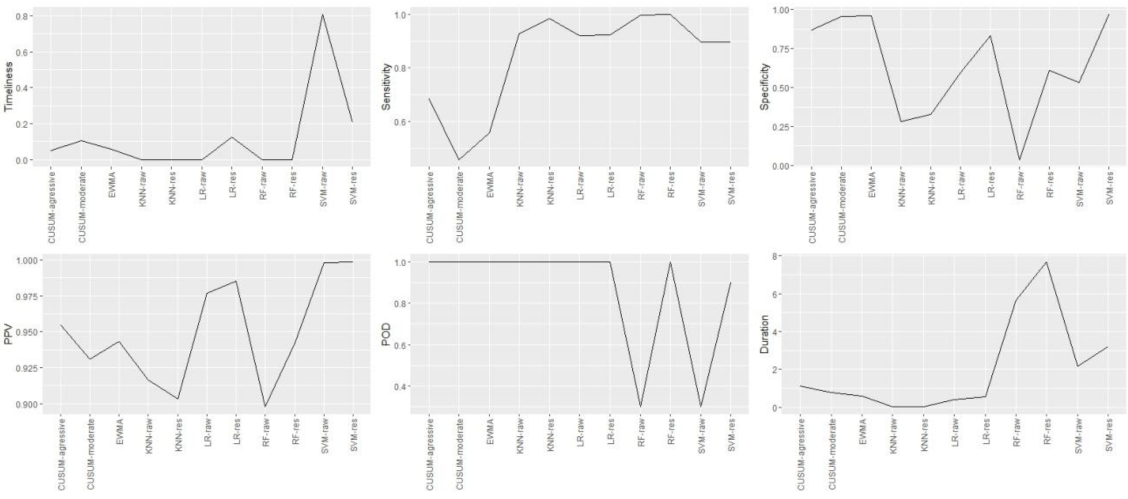

FIGURE 2: Average timeliness, sensitivity, specificity, PPV, POD and duration metrics for 10 sets of simulated data.
PPV: Positive predictive value; POD: Power of detection.

DISCUSSION

The performance results indicate that all the approaches studied here can detect the start date of the unusual outbreaks with a considerably high performance. kNN and RF settings have high sensitivity but poor specificity, where CUSUM settings and EWMA have lower sensitivity with high specificity. The methods which have high sensitivity and specificity together is SVM-res (0.90 and 0.97), and LR-res (0.92 and 0.83). POD and PPV results are similar and considerably high for each algorithm, except for SVM-raw and RF-raw. kNN, LR and CUSUM settings are the fastest ones, with minimum execution time.

It has been shown that when integrated with the fitted regression model, CUSUM settings have higher accuracy.[17] We can also conclude that, when the algorithms are applied to the "residuals of fitted regression model" instead of raw data, they all give better results. However, there is no "one" algorithm which is better across all detection measures. When developing a surveillance system, algorithm/algorithms can be chosen according to which aspects of detection are more important in that system as discussed in.[1] Moreover, if a single algorithm has to be chosen, SVM-res is the one, which did not get any poor result from any one of the performance metrics.

There are increasing number of local systems that collect and monitor data at a county level, city level, or even hospital level. These systems collect clinical data, usually at a daily rate, including emergency department chief complaints and admissions.[20] In this paper, we provide a simulation study to assess the detection capabilities of surveillance algorithms. This study can assist daily surveillance practitioners and decision makers in deciding which algorithm would be more beneficial based on their needs.

LIMITATIONS

Since the data is simulated as ILI admissions at a local ED, the limitations of this work may include the lack of demonstration of regional-nationwide generalizability. However, this framework can easily be used to generate similar models for other hospitals and medical centers using their local data. Another limitation is that the method relies on 20 sets of yearly data. The clinical professionals may study the modeling results to see if any adjustment is needed for better modeling and can update their baseline accordingly.

CONCLUSION

In this study, an approach for modeling and monitoring daily ED visits of ILI cases is discussed. We performed this study to evaluate the performance of different statistical and machine learning algorithms when detecting hidden unusual outbreaks, such as COVID-19. Proposed framework consists of 2 main stages: modeling and monitoring. After fitting a regression model with ARIMA error terms to the daily ED counts, the monitoring stage of the approach is conducted. EWMA and 2 different CUSUM settings and are applied to standardized residuals data and then SVM, LR, kNN and RF methods are applied to both residual data of fitted regression model; and raw data. Our results show that the performance of ML algorithms can be improved when integrated with statistical approaches. The results also verified that kNN and RF settings (for high sensitivity); CUSUM and EWMA (for high specificity); SVM-res and LR-res (for high sensitivity and specificity together), can be preferred as a real-time surveillance method for prospective monitoring of respiratory cases in the emergency setting.

Source of Finance

During this study, no financial or spiritual support was received neither from any pharmaceutical company that has a direct connection with the research subject, nor from a company that provides or produces medical instruments and materials which may negatively affect the evaluation process of this study.

Conflict of Interest

No conflicts of interest between the authors and/or family members of the scientific and medical committee members or members of the potential conflicts of interest, counseling, expertise, working conditions, share holding and similar situations in any firm.

Authorship Contributions

Idea/Concept: *Zeynep Filiz Eren;* ***Design:*** *Zeynep Filiz Eren;* ***Control/Supervision:*** *Zeynep Filiz Eren;* ***Data Collection and/or Processing:*** *Mehmet Kadri Gofralılar;* ***Analysis and/or Interpretation:*** *Hasan Ali Özkan;* ***Literature Review:*** *Zeynep Filiz Eren;* ***Writing the Article:*** *Zeynep Filiz Eren;* ***Critical Review:*** *Zeynep Filiz Eren.*

REFERENCES

1. 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. [Crossref] [PubMed] [PMC]
2. Buckeridge DL. Outbreak detection through automated surveillance: a review of the determinants of detection. J Biomed Inform. 2007;40(4):370-9. [Crossref] [PubMed]
3. Burkom HS, Murphy SP, Shmueli G. Automated time series forecasting for biosurveillance. Stat Med. 2007;26(22):4202-18. [Crossref] [PubMed]
4. 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. [Crossref] [PubMed]
5. Mofijur M, Fattah IMR, Alam MA, Islam ABMS, Ong HC, Rahman SMA, et al. Impact of COVID-19 on the social, economic, environmental and energy domains: lessons learnt from a global pandemic. Sustain Prod Consum. 2021;26:343-59. [Crossref] [PubMed] [PMC]
6. van-Dijk A, Aramini J, Edge G, Moore KM. Real-time surveillance for respiratory disease outbreaks, Ontario, Canada. Emerg Infect Dis. 2009;15(5):799-801. [Crossref] [PubMed] [PMC]
7. Zacher B, Ullrich A, Ghozzi S. Supervised learning for automated infectious-disease-outbreak detection. Online J Public Health Inform. 2019;11(1). [Crossref]
8. Jafarpour N, Izadi M, Precup D, Buckeridge DL. Quantifying the determinants of outbreak detection performance through simulation and machine learning. J Biomed Inform. 2015;53:180-7. [Crossref] [PubMed]
9. Cabatuan M, Manguerra M. Machine learning for disease surveillance or outbreak monitoring: a review. 2020 IEEE 12th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM), Dec 3-7, 2020. IEEE; 2020. p.1-6. [Crossref]
10. Meyer D, Dimitriadou E, Hornik K, Leisch F, Weingessel A. e1071: Misc functions of the Department of Statistics (e1071), TU Wien. R Packag version. 2014;1(3). [Link]
11. Kuhn M. Building predictive models in R using the caret package. J Stat Softw. 2008;28(5):1-26. [Crossref]
12. Liaw A, Wiener M. Classification and regression by randomForest. R news. 2002;2(3):18-22. [Link]
13. Hyndman RJ, Khandakar Y. Automatic time series forecasting: the forecast package for R. J Stat Softw. 2008;27(3):1-22. [Crossref]
14. 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. [Crossref] [PubMed]
15. Kanra G, Tezcan S, Yilmaz G; Turkish National Respiratory Syncytial Virus (RSV) Team. Respiratory syncytial virus epidemiology in Turkey. Turk J Pediatr. 2005;47(4):303-8. [PubMed]
16. Sparks RS, Keighley T, Muscatello D. Optimal exponentially weighted moving average (EWMA) plans for detecting seasonal epidemics when faced with non-homogeneous negative binomial counts. J Appl Stat. 2011;38(10):2165-81. [Crossref]
17. Hagen KS, Fricker RD, Hanni KD, Barnes S, Michie K. Assessing the Early Aberration Reporting System's ability to locally detect the 2009 influenza pandemic. Stat Polit Policy. 2011;2(1). [Crossref]
18. De Oca VM, Jeske DR, Zhang Q, Rendon C, Marvasti M. A cusum change-point detection algorithm for non-stationary sequences with application to data network surveillance. J Syst Softw. 2010;83(7):1288-97. [Crossref]
19. Singh A, Thakur N, Sharma A. A review of supervised machine learning algorithms. 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom) March 16-18, 2016. Ieee; 2016. p.1310-5.
20. Shmueli G, Burkom H. Statistical challenges facing early outbreak detection in biosurveillance. Technometrics. 2010;52(1):39-51. [Crossref]