Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

International Journal of Medical Informatics 142 (2020) 104248

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

Predicting the hotspots of age-adjusted mortality rates of lower respiratory infection across the continental United States: Integration of GIS, spatial statistics and machine learning algorithms

Abolfazl Mollalo[a,*], Behrooz Vahedi[b], Shreejana Bhattarai[c], Laura C. Hopkins[a], Swagata Banik[a], Behzad Vahedi[d]

[a] Department of Public Health and Prevention Science, School of Health Sciences, Baldwin Wallace University, Berea, OH, USA
[b] Department of Mathematics, University of Trento, Trento, Italy
[c] Department of Geography, University of Florida, Gainesville, FL, USA
[d] Department of Geography, University of Colorado Boulder, Boulder, CO, USA

ARTICLE INFO

Keywords:
Accuracy assessment
Decision trees
GIS
Hotspots
Lower respiratory infections
US

ABSTRACT

Objective: Although lower respiratory infections (LRI) are among the leading causes of mortality in the US, their association with underlying factors and geographic variation have not been adequately examined.
Methods: In this study, explanatory variables (n = 46) including climatic, topographic, socio-economic, and demographic factors were compiled at the county level across the continentalUS.Machine learning algorithms - logistic regression (LR), random forest (RF), gradient boosting decision trees (GBDT), k-nearest neighbors (KNN), and support vector machine (SVM) - were employed to predict the presence/absence of hotspots ($P < 0.05$) for elevated age-adjusted LRI mortality rates in a geographic information system framework.
Results: Overall, there was a historical shift in hotspots away from the western US into the southeastern parts of the country and they were highly localized in a few counties. The two decision tree methods (RF and GBDT) outperformed the other algorithms (accuracies: 0.92; F1-scores: 0.85 and 0.84; area under the precision-recall curve: 0.84 and 0.83, respectively). Moreover, the results of the RF and GBDT indicated that higher spring minimum temperature, increased winter precipitation, and higher annual median household income were among the most substantial factors in predicting the hotspots.
Conclusions: This study helps raise awareness of public health decision-makers to develop and target LRI prevention programs.

1. Introduction

Lower respiratory infections (LRI) are diseases of the lower respiratory tracts and include bronchitis, bronchiolitis, pneumonia, and recently emerged coronavirus (COVID-19). LRI are major public health concerns across the world ([1], [2], [3]), and are among the leading causes of mortality and morbidity in children and adults [4,5]. In 2016, LRI caused nearly 2.38 million deaths worldwide, including 652,572 children under five years old and 1,080,958 adults over 70 years old, making it the sixth leading cause of death for all ages [6].

LRI are the cause of a significant number of hospitalizations in developed countries [7]. In the US, LRI have been classified as the 7th leading cause of death and years of life lost [8]. In this country, bronchiolitis is the leading diagnosis of LRI in children younger than two years old, causing almost 150,000 annual hospitalizations [9]. Similarly, pneumonia is another most common reason for hospital admissions in the US that causes the most common severe bacterial infection in children [10]. However, with the success of the childhood vaccination programs such as the 7-valent and 13-valent pneumococcal conjugate vaccines, the proportion of elderly affected by LRI in the US has significantly declined [11].

Previous studies have shown that many socio-economic factors such as education level, income, and poverty [12] and environmental factors such as climate and air pollution ([13]; [14]) were significantly associated with LRI prevalence. Further, demographic factors such as age, gender, and race [15] and behavioral factors such as cigarette smoking [16] were correlated with LRI prevalence. Few studies have examined the spatial variation of LRI in small geographic regions. For example,

* Corresponding author at: Department of Public Health & Prevention Science, Baldwin Wallace University, 275 Eastland Road, Berea, OH, 44017, USA.
E-mail addresses: amollalo@bw.edu (A. Mollalo), b.vaheditorghabeh@studenti.unitn.it (B. Vahedi), s.bhattarai@ufl.edu (S. Bhattarai), lhopkins@bw.edu (L.C. Hopkins), sbanik@bw.edu (S. Banik), behzad@colorado.edu (B. Vahedi).

https://doi.org/10.1016/j.ijmedinf.2020.104248
Received 6 May 2020; Received in revised form 25 July 2020; Accepted 10 August 2020
Available online 22 August 2020

Beamer et al. [17] identified distinct patterns of significant spatial clusters for each LRI phenotype within Tucson, Arizona. Those clusters were associated with various community-level risk factors such as increased air pollution, poor housing conditions, and low socio-economic status. Beck et al. [18] conducted a study in Cincinnati, Ohio, to examine geographic variation of LRI hospitalization rates across Hamilton county using Getis-Ord Gi* statistic. They also examined whether such variation was correlated with socio-economic status using the non-parametric Kruskal-Wallis test. The results indicated a significant alteration in the median hospitalization rates by census tract quintile for both bronchiolitis and pneumonia. Further, socio-economic conditions had substantial influences on those hospitalization rates, and hotspots were located in the impoverished neighborhoods in the urban core.

In recent decades, the use of novel modeling techniques such as machine learning algorithms in public health studies, in particular, respiratory disease research has increased [19]. For instance, Heckerling et al. [20] trained a back-propagation artificial neural network (ANN) optimized by genetic algorithm to predict pneumonia among patients (n = 1044) with respiratory complaints from the University of Illinois and the University of Nebraska. A multitude of variables, such as demographics, symptoms, signs, and comorbidity with other respiratory diseases, including asthma and lung disease, were compiled to predict the presence or absence of pneumonia among the patients. The ANN model successfully predicted pneumonia on the test dataset with 93 % accuracy. In a case-control study in Taiwan, Kuo et al. [21] compared the performance of seven machine learning classifiers, including random forest and logistic regression, to predict hospital-acquired pneumonia among schizophrenic patients. Among the employed algorithms, random forest had the highest accuracy (93 %) in predicting pneumonia. Further, the significant predictors were clozapine use, clozapine prescription, and prescription duration.

While several studies have been conducted in smaller geographic regions, to our knowledge, no previous nationwide study has examined geographic variations of LRI mortality rates and their association with underlying factors across the US. Identifying hotspot(s) of LRI mortality rates (i.e., counties with higher than expected mortalities) and their presence or absence based on population-level underlying factors can help public health decision makers for targeted interventions at the national level. Thus, in this ecological study, we investigate the geographic variation of age-adjusted LRI mortality rates across the continental US from 1980 to 2014 using spatial statistics. Further, we employed several machine learning algorithms to predict hotspot(s) occurrence with potential risk factors in a geographic information system (GIS) framework.

2. Material and methods

2.1. Data collection and preparation

Continental US age-adjusted mortality rates of LRI were obtained at the county level from Global Health Data Exchange (http://ghdx.healthdata.org/record/ihme-data/united-states-mortality-rates-county-1980-2014). The data were available for eight years: 1980, 1985, 1990, 1995, 2000, 2005, 2010, and 2014. The disease data were then spatialized at the county level in ArcGIS 10.7 (ESRI, Redlands, CA). The ESRI shapefile of the administrative boundary of US counties was obtained from Topologically Integrated Geographic Encoding and Referencing (TIGER)/Line US Census Bureau for the year 2018 (http://www.census.gov/).

Explanatory variables (n = 46) including climatic, topographic, socio-economic, and demographic factors were compiled at the county level across the continental US and stored in a file geodatabase in ArcGIS 10.7. The variables were selected according to either the previously published literature or domain knowledge.

Low and high air temperature can aggravate respiratory symptoms, particularly among individuals with preexisting conditions. Low air temperature can adversely impact epithelium by narrowing the respiratory airways and declining lung functions. In contrast, high air temperature can increase allergic illnesses possibly by increasing pollen production or extending the length of pollen season, which in turn can make the respiratory symptoms worse. Increased precipitation may facilitate the spread of respiratory diseases. Vitamin D, which is produced by sunlight exposure, may protect the human body against respiratory diseases. We obtained climate data including daily air temperature (°C), daily precipitation (mm), and daily sunlight (KJ/m^2) from the Centers for Disease Control and Prevention Wide-Ranging Online Data for Epidemiologic Research (CDC WONDER) database (http://wonder.cdc.gov/). Then, we aggregated the daily climate data for the spring (March 19-June 20), summer (June 20-September 22), autumn (September 22-December 21) and winter (December 21 to March 20) seasons (i.e., seasonal minimum and maximum temperature, seasonal average precipitation, and seasonal average sunlight).

The fine particulate matter (PM 2.5), which may contain soot, smoke, and dust, can get deep into human lungs and enter the bloodstream. According to Bowe et al. [22], exposure to high levels of PM 2.5 is associated with almost 200,000 deaths in the US. Moreover, cigarette smoking can damage human airways and the small air sacs in the lungs. Daily PM 2.5 air quality data was obtained from the CDC WONDER database. The mean values of PM 2.5 for the four seasons were computed for each county. Also, the data pertaining to cigarette smoking prevalence in the US for men and women were obtained from Dwyer-Lindgren et al. [23].

Respiratory infections are more complicated in infants and children living in high altitudes. During acute LRI, hypoxemia occurs more frequently in children at high altitudes, which may result in increased mortality [24]. Therefore, the topographic data (i.e., median altitude and slope) of US counties were also incorporated as explanatory variables. The altitude shuttle radar topography mission (STRM) digital elevation model with 30 m spatial resolution were obtained from the national map website (http://nationalmap.gov/). The altitude and slope values for counties were then quantified using zonal statistics function in ArcGIS Spatial Analyst extension.

Lower socio-economic status can be associated with unbalanced access to health care which in turn can lead to elevated mortality of diseases. A broad range of socio-economic and demographic variables including the proportion of the white and black population, median household income, poverty, unemployment rate, (lack of) health insurance, and the number of physicians per county was obtained from the US Census Bureau's American FactFinder (https://factfinder.census.gov/) and included in the file geodatabase. All data used in this study are publicly available from the above sources.

2.2. Spatial statistics

The spatial pattern of age-adjusted LRI mortality rates (i.e., clustered, dispersed, or random) across the continental US, were examined with global and local indices of spatial autocorrelation for every eight years of study. Moran's I and Getis-Ord General G were employed to investigate the extent to which the nearby counties had similar LRI rates. Moran's I is calculated using the following formula:

$$I = \frac{n \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} w_{ij} L_i L_j}{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \sum_{i=1}^{n} L_i^2} \quad (1)$$

where L_i and L_j are the deviations of LRI mortality rates from the average mortality rate for county i and county j, respectively; w_{ij} is a binary weight matrix between county i and county j based on the first-order Queen contiguity (i.e., each element in weight matrix is non-zero when the counties share borders of non-zero length); and n is the aggregate number of counties. The value of I ranges between -1 (negative spatial autocorrelations) and +1 (positive spatial autocorrelation), while values close to 0 indicate no spatial autocorrelation ([25], [26]).

Using the same notation as for Eq (1) Getis-Ord General G is computed as:

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}L_iL_j}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \quad (2)$$

A significant value of G indicates spatial clustering of LRI mortality rates. Both Moran's I and Getis-Ord General G statistics were calculated in ArcGIS 10.7.

Local measures of spatial autocorrelation such as Getis-ord G_i* also were applied to locate the identified spatial autocorrelations of LRI mortality rates ($P < 0.05$) as follows [27,28].

$$G_i^* = \frac{\sum_{j=1}^{n} w_{ij}L_j - \bar{L}\sum_{j=1}^{n} w_{ij}}{S\sqrt{\frac{[n\sum_{j=1}^{n} w_{ij}^2 - (\sum_{j=1}^{n} w_{ij})^2]}{n-1}}} \quad (3)$$

$$S = \sqrt{\frac{\sum_{j=1,j\neq i}^{n} C_j^2}{n-1} - \bar{C}^2} \quad (4)$$

A high positive and a high negative value of G_i^* imply hotspot and coldspot, respectively. However, the focus of this study is on mapping and analyzing the identified hotspots of LRI mortality rates for further modeling. More detailed information about the clustering and hotspot detection techniques have been published elsewhere ([29], [30]).

2.3. Machine learning modeling

Five different machine learning classifiers were employed to identify hotspot locations ($P < 0.05$) of the LRI age-adjusted mortality rates. The LRI mortality rate for the year 2014 was considered as dependent variable. The classifiers were vanilla logistic regression (LR), random forest (RF), gradient boosting decision trees (GBDT), k-nearest neighbors (KNN), and support vector machine (SVM). These classifiers were selected due to their successful performance in identifying intricate patterns in many binary classification applications ([31]; [32]). The scikit-learn Python package was used to develop the classifiers.

2.3.1. Logistic regression

LR, a linear function for binary classification, applies maximum likelihood estimation to minimize the errors after transforming the presence or absence of LRI hotspots into a logit variable [33]. The output of LR is the likelihood of LRI' hotspot occurrence, as a function of several exploratory variables and can be expressed as:

$$P = \frac{1}{1 + exp^{-z}} \quad (5)$$

Where P is the predicted likelihood of LRI hotspot occurrence bounded between 0 and 1; and z is a linear combination of the variables and its value varies between $-\infty$ and $+\infty$. More precisely:

$$z = \beta_0 + \beta_1X_1 + \beta_2X_2 + ...+\beta_nX_n \quad (6)$$

Where β_0 is the intercept and $\beta_i (i = 1,...,n)$ are the coefficients associated with the variables $X_i (i = 1,...,n)$. The detailed information about LR is provided by Hosmer and Lemeshow [34].

2.3.2. Random forest

RF developed by Breiman [35] is an ensemble learning method where a plethora of decision trees are produced based on bootstrap sampling. The input data are repeatedly split, based on many different generated classification trees. The final decision is made based on the maximum number of 'votes' obtained from individual trees ([36]; [37,38]). In this study, the number of trees was set to 1000. Also, the optimal number of layers from the root to the node of the trees was chosen using cross-validation from the set of {2, 3, 4}.

2.3.3. Gradient boosting decision trees

Similar to RF, GBDT is an ensemble method based on bootstrap sampling, which generates many decision trees. While RF uses the bagging method (e.g., equal probability of sample selection in each iteration), GBDT uses a boosting method (i.e., weighted (unequal) sample selection in each run). After each iteration, the weights are adjusted so that the higher weights will be assigned to the models with good performances (Friedman [39]).

Suppose x_i is a training sample, y_i is the associated label of x_i, and N is the number of training samples. For any training sample x_i, $F(x_i)$ is the classification (the ith decision tree) of x_i, and $L(y_i, F(x_i))$ is the loss between $F(x_i)$ and y_i. GBDT determines an optimal model such that $\sum_{i=1}^{n} L(y_i, F(x_i))$ is minimized. In the first step, the GBDT initialize the decision tree $F_0(x)$, then iteratively constructs m new trees. For each iteration, a negative gradient is computed and a new tree h(x) is added to reduce the residuals. The optimal model $F^*(x)$ can be calculated as follows:

$$F^*(x) = F_0(x) + v + \sum_{i=1}^{m} \rho_t{}^*h_t(x) \quad (7)$$

where m is the number of iterations; v controls the learning rate; ρ_t is the weight of $h_t(x)$ and $h_t(x)$ is the trained decision tree in the tth iteration [39].

2.3.4. K-nearest neighbors

The k-nearest neighbors classifier (k is a positive integer), is a non-parametric and distance-based algorithm that assigns a test sample to the class that is common among its k-nearest training samples. In other words, a county is classified as a hotspot of LRI if a majority of its neighboring counties are hotspots Peterson [40]. Using a random search algorithm, k = 10 was selected as the optimal number of nearest neighbors. Also, the explanatory variables are not involved in this algorithm.

The distance can be calculated in a variety of ways including Euclidean distance, Hamming distance, Manhattan distance and Minkowski Distance. We used Manhattan distance which yielded better results which is calculated as:

$$D_M = \sum_{i=1}^{n} |x_i - y_i| \quad (8)$$

where x and y are n -dimensional vectors such that $x = (x_1, x_2, ..., x_n)$ and $y = (y_1, y_2, ..., y_n)$.

2.3.5. Support vector machine

The SVM classifier, first proposed by Vapnik [41], uses robust statistical learning theory. Consider a dataset of high dimensional points, viewed as vector $\{x_i \in R^d: i = 1, ...n\}$, $d > 1$, where each point belongs to one of two classes defined by $\{y_i \in \{0,1\}: i = 1, ...n\}$. Here, y_i corresponds to the presence/absence of LRI hotspots. If we assume these points to be linearly separable (i.e., can be separated via a linear boundary), the goal of SVM is to find the d-dimentional hyperplane maximizing the margin (i.e., distance between the closest points or support vectors) as illustrated in Fig. 1 [42].

The hyperplane can be expressed as $d(x) = sgn(w.\ x_i + b)$, where w is the orientation of hyperplane and b is the offset of hyperplane from origin and sgn is sign function (i.e., sgn = +1 for presence and sgn = -1 for absence of LRI hotspot). SVM can work in the case where the points are not linearly separable by using a soft-margin. Soft margin allows a trade-off between the margin of separation and the miss-classification penalty. One form of which can be the aggregated distance of the miss-classified points to the separation hyperplane. The optimal separating hyperplane can be found using Lagrangian multipliers from:

$$Minimize \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j y_i y_j (x_i x_j) \quad (9)$$

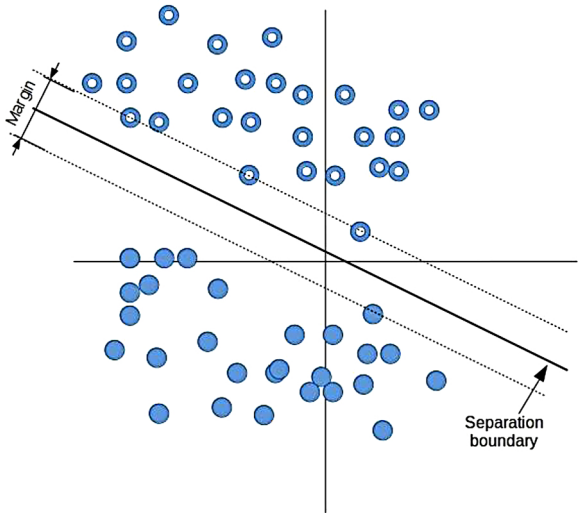

Fig. 1. Principle of linearly separable SVM using maximum margin.

$$\text{Subject to } \sum_{i=1}^{n} \alpha_i y_i = 0,\ 0 \le \alpha_i \le C \tag{10}$$

Where α_i are the Lagrange multipliers and the value of C or regularization shows a trade-off between maximizing the margin and minimizing the errors. Finally, w and b can be obtained as follows:

$$w = \sum_{i=1}^{n_{sv}} \alpha_i y_i x_i \tag{11}$$

$$b = y_i - w^T . x_i \tag{12}$$

Where n_{sv} is the number of support vectors placed on the margin lines.

Many real-world problems are nonlinear. In this case, SVM utilizes kernel functions to transform data into a higher dimensional space than the original dimension in which the input data can be separated by a linear boundary [43]. For non-linear separable cases, the above formula is extended using kernel function. This function maps the input dataset onto a higher dimensional feature space as shown in Fig. 2. The decision function is modified as:

$$f(x) = sign\left(\sum y_i \alpha_i K(X_i, X_j) + b\right) \tag{13}$$

Where $K(X_i, X_j)$ is a Gaussian radial basis function kernel as:

$$K(X_i, X_j) = \exp(-\gamma \|X_i - X_j\|^2) \tag{14}$$

Appropriate results highly depend on the selection of C and γ. Here, we used a grid search to find the optimum values for the two parameters. This method checks various combinations of C and γ in a range of pre-defined values (C between 0.5 and 20 with increments of 0.5 and γ between 0.005 and 1.0 with increments of 0.1). It should be noted that these ranges are boundaries of search space and have been chosen to cover a large enough space. For example, in our case, 20 is numerically large enough for C.

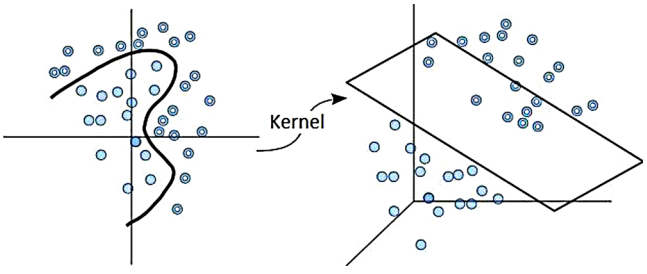

Fig. 2. A non-linear boundary in the input space (left) and a maximum margin hyperplane in feature space (right).

2.4. Accuracy assessment

To employ the algorithms, 70 % and 30 % of the dataset were randomly selected for training and test dataset, respectively. A randomized search algorithm for tuning hyper-parameters in each classification algorithm was used. L1 regularization (LASSO) was used to reduce the complexity of the model and to avoid overfitting. This is done by penalizing small weights to zero, leading to a sparser model.

The performances of the classifiers were assessed with several metrics: overall accuracy ($\frac{T_P + T_N}{T_P + T_N + F_P + F_N}$), precision ($\frac{T_P}{T_P + F_P}$), recall ($\frac{T_P}{T_P + F_N}$), F1-score ($2*\frac{Precision * recall}{Precision + recall}$), false positive rate or FPR ($\frac{F_P}{T_N + F_P}$) and area under ROC (receiver operating characteristic) curve (ROC AUC). In the above formulas, T_P, T_N, F_P, and F_N represent the number of true positives, true negatives, false positives, and false negatives, respectively.

The area under the precision-recall curve (PR AUC), which shows the tradeoff between precision and recall of different thresholds, was also measured because the classes were imbalanced (Goutte & Gaussier [44]). All evaluation metrics were computed on the test dataset.

3. Results

The null hypothesis of complete spatial randomness was rejected for all study years based on Moran's I (range: 0.36 – 0.61; p-values < 0.001) and General G (range: 0.0018 – 0.0019; p-values < 0.001) statistics. The z-scores of both statistics almost consistently increased to large values from 1980 to 2014, indicating highly significant clustering (Table 1). Clustering was minimal from 1980 to 1990, but sharply and consistently increased thereafter.

In the earlier years of the study period (1980–1985), the identified hotspots of the LRI mortality rates by Getis-Ord Gi* hotspot detection technique were mostly concentrated in the western US. In contrast, from 1990 to 2000, these hotspots became less prominent, while LRI hotspots shifted toward the southeastern parts of the US (Fig. 3). These counties continue to represent hotspots through the remaining periods.

In total, 118 counties (3.8 % of US counties) were persistently identified as (part) of LRI hotspots (Fig. 4). Among these were counties in Georgia (n = 49), Kentucky (n = 25), and Virginia (n = 22) that were persistently affected, and accounted for 81.3 % of total persistent hotspot counties.

All the classification algorithms predicted the hotspots of LRI mortality rates with relatively high accuracy (≥ 0.84); however, GBDT and RF were the most accurate models (0.92) (Table 2). Precision-recall plots of the employed models (Fig. 5) showed that GBDT had the highest PR AUC - indicating the largest values of both precision and recall for different cut-off values.

GBDT achieved the highest F1- score (85 %) and PR AUC (84 %), compared to the other models, while the LR model had the worst performance (Table 2). Also, the results of RF were slightly better than KNN and SVM. Overall, of the employed machine learning algorithms,

Table 1
Results of the global Moran's I and General G statistic of age-adjusted LRI mortality rates, continental US, 1980-2014.

Year	Index		Z-score		Type of distribution	P-value
	Moran's I	General G	Moran's I	General G		
1980	0.38	0.0019	36.31	8.27	Clustered	~ 0
1985	0.36	0.0019	34.59	8.40	Clustered	~ 0
1990	0.37	0.0019	35.04	9.57	Clustered	~ 0
1995	0.41	0.0018	39.50	12.10	Clustered	~ 0
2000	0.49	0.0018	47.00	15.50	Clustered	~ 0
2005	0.53	0.0018	51.06	18.81	Clustered	~ 0
2010	0.58	0.0018	55.79	22.24	Clustered	~ 0
2014	0.61	0.0018	58.35	24.68	Clustered	~ 0

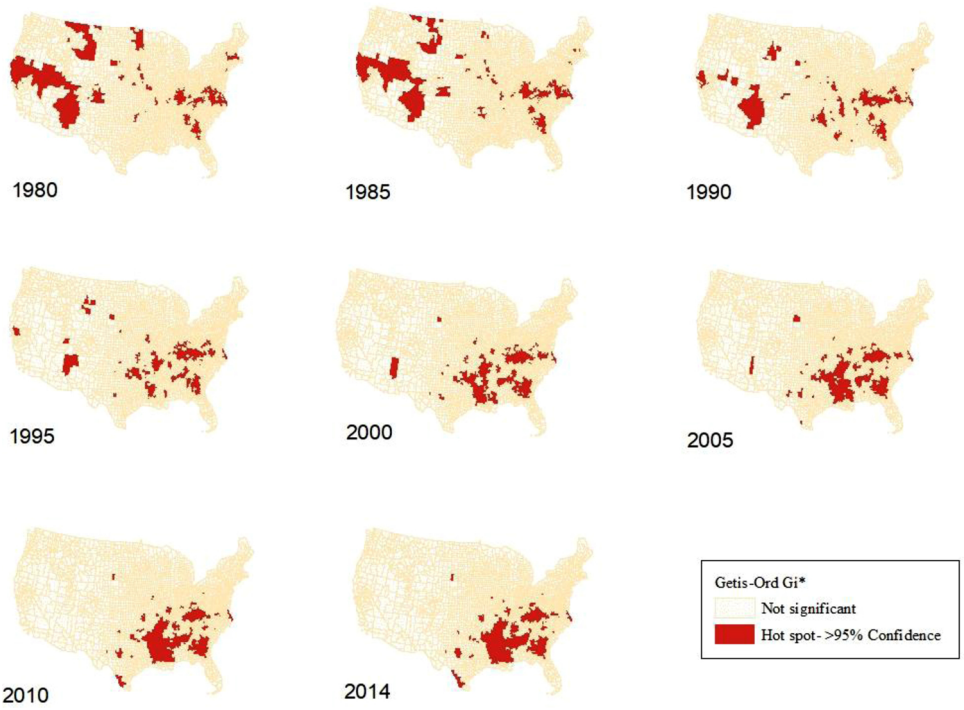

Fig. 3. Location of hotspots of LRI mortality rates in the continental US using Getis-Ord Gi* hotspot detection technique, 1980-2014.

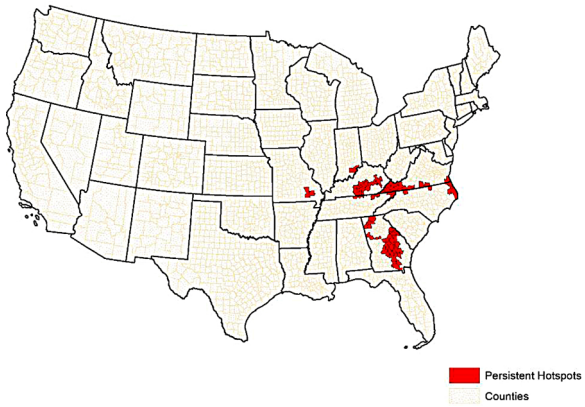

Fig. 4. Location of counties that were persistently identified as hotspots of LRI mortality rates by Getis-Ord Gi* hotspot detection technique, 1980-2014.

Table 2
Evaluation metrics associated with each of the employed machine learning classifiers.

	Accuracy	Precision	Recall	F1-Score	ROC AUC	PR AUC	FPR
Classifier							
LR	0.84	0.75	0.87	0.78	0.86	0.72	0.17
RF	0.92	0.87	0.82	0.84	0.82	0.83	0.03
GBDT	0.92	0.87	0.83	0.85	0.83	0.84	0.04
KNN	0.90	0.84	0.8	0.82	0.8	0.82	0.05
SVM	0.91	0.83	0.86	0.84	0.86	0.82	0.07

the decision trees (i.e., GBDT and RF) yielded a more accurate predictions.

The contributions of variables were analyzed for the GBDT and RF models (Fig. 6). The results of the GBDT model indicated that spring minimum temperature, winter precipitation, and median household income had the greatest positive influence in predicting the hotspots.

4. Discussion

In this study, we integrated spatial statistical tools with machine learning classifiers in a GIS platform to identify hotspots of the LRI mortality rates across the continental US and to identify the most substantial LRI-associated environmental and socio-economic factors. Given the lack of nationwide spatial analysis and modeling of LRI, our modeling framework can be applied as a general protocol specifically to more prevalent respiratory diseases in the US such as asthma, chronic obstructive pulmonary disease, pneumonia and COVID-19 to support public health decision makings at the national level. Overall, there was a historical shift in hotspots away from the western US into the southeastern parts of the country, and the hotspots were highly localized in a few counties. Environmental factors contributed most strongly to these hotspots, while economic and social factors seem to be of secondary significance.

According to Fischer et al. [45], advanced computational models can translate the occurrence of infectious diseases into decision-support tools. Unlike traditional models, machine learning algorithms can quantify the association between infectious disease and explanatory variables, even with incomplete or noisy data [26] in a shorter time period and less costs.

Moran's *I* and General G statistics confirmed that LRI mortality rates are spatially clustered ($P < 0.001$) across the continental US. Counties with high mortality rates tend to locate closer together than expected by chance. Using Getis-Ord Gi*, we identified several hotspots across the continental US. Additionally, spatial-temporal analysis of the clusters found a notable geographic shift in the location of hotspots from the west coast to the southeast of the US during the study period. The spatial pattern and shift in the locations of hotspots over time may partially reflect the vast differences in LRI mortality rates by drivers of geographic patterns, including environment, socio-economic and behavior factors. It may also be attributed to the health disparities or improved health care quality such as PCV7 and PCV13 vaccination programs during the study period. The latter is consistent with the substantial global decline of Streptococcus pneumonia - the leading cause of LRI mortality - as estimated by GBD 2016 Lower Respiratory

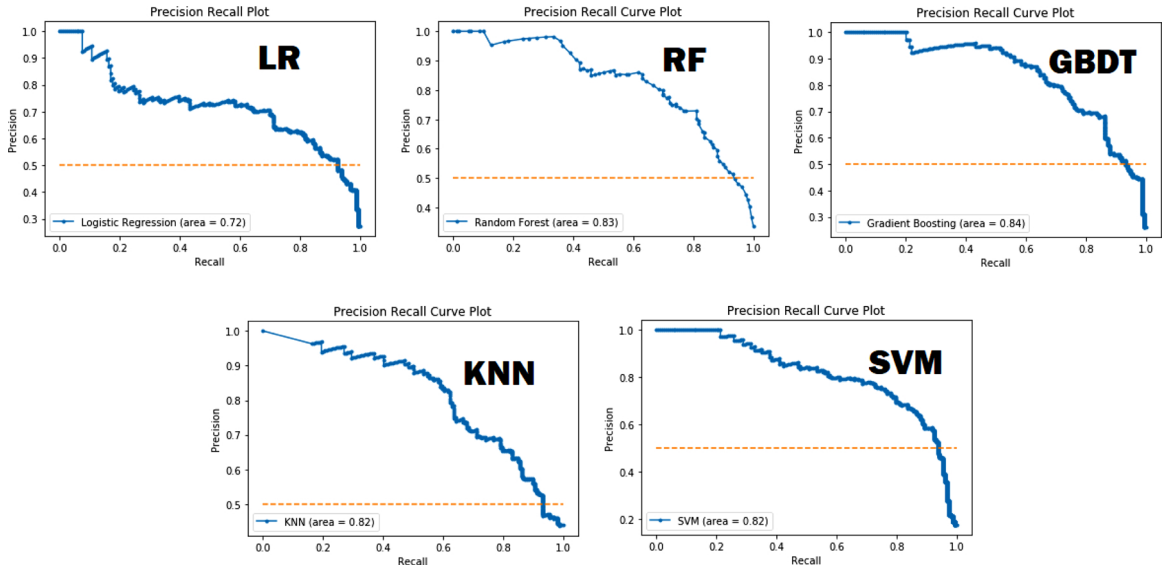

Fig. 5. Results of the precision-recall curve for employed machine learning classifiers. The orange dash line annotates the average precision.

Infections Collaborators [46]. Moreover, some states (including Georgia, Kentucky, and Virginia) and counties included persistent hotspots, suggesting targeting resources and policy interventions in these areas.

All the classifiers showed a considerable accuracies; however, due to the imbalanced dataset, in general, ensemble decision trees outperformed the (complex) SVM or traditional and frequently applied LR. Additionally, although SVM was slightly less accurate compared to the decision trees, it is less interpretable, slower to run, and more susceptible to overfitting. Allyn et al. [47] developed LR, RF, GBDT, SVM, and Naïve Bayes Model to predict the mortality of 4676 patients after elective cardiac surgery from December 2005 to December 2012. Their results showed RF outperformed the other classifiers (AUC = 0.788). Our results are also in agreement with the findings of Churpek et al. [48], who compared LR, tree-based models, KNN, SVM, and neural networks. Their findings showed that RF was the most accurate classifier (AUC = 0.801), followed by the gradient boosting machine (AUC = 0.794).

The findings of decision trees indicated that higher spring temperature and increased precipitation during winter are among the most substantial predictors of the presence or absence of the hotspots. The contribution of these environmental factors is most likely due to the changes in the epidemiology of weather-sensitive pathogens and host immune response, which can, in turn, lead to respiratory infections

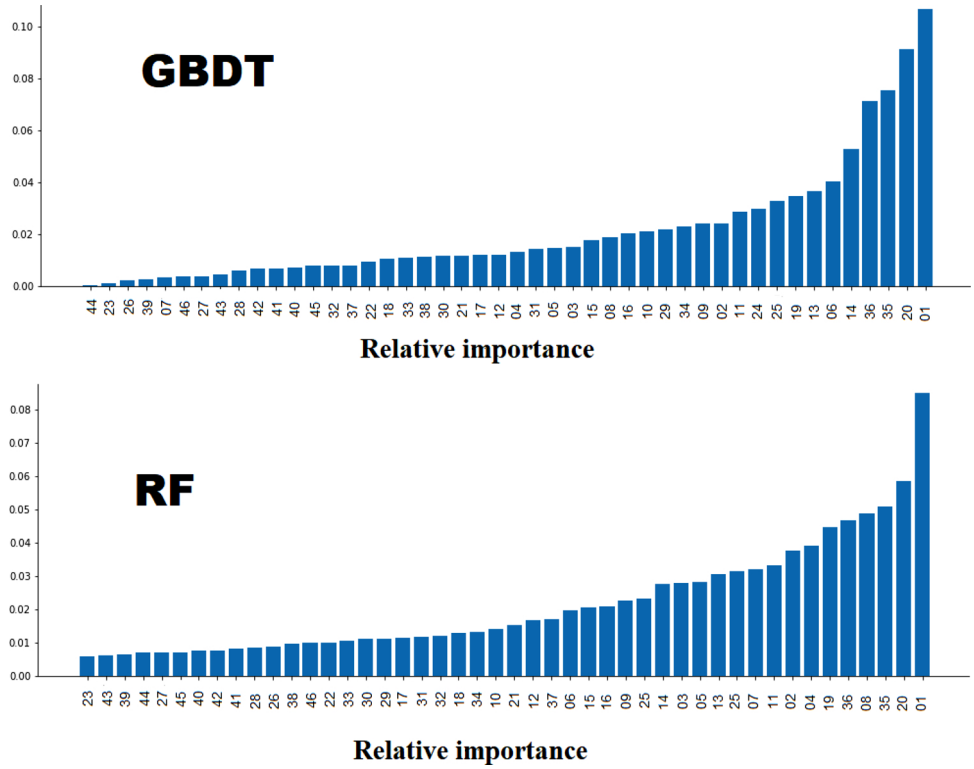

Fig. 6. Relative variable importance analysis using the gradient boosting and random forest decision trees. A detailed description of x-axis codes is provided in Supplementary Material.

[49]. Other studies show that respiratory infections are seasonal, especially during winter and rainy months. Seasonality may play a role due to the proximity of people in enclosed environments during cold temperature weather, which can facilitate the spread of infections during those seasons. For example, Thomas et al. [50] found that RSV infection was more prevalent in children during the winter months in Canada. In Malaysia, LRI was positively correlated with the monthly number of rainy days but negatively associated with the monthly mean temperature [51]. A study conducted in Pakistan showed that LRI cases were more frequent in months when the minimum temperature was lower [52], however, in Brazil, statistically significant associations were found between viral LRI and increasing temperature and decreasing humidity [53]. Inconsistent findings may be due to different studied organisms or different spatial units of analysis. For example, from county-level studies, one can not draw a conclusion at the individual level due to ecological fallacy. Moreover, age is a potential confounder that needs to be adjusted, particularly in studying mortality rates of diseases, to avoid distorting the relationship.

The findings of decision trees also implied that the economic status such as median household income and the higher proportion of the population living below the poverty line (according to the definition of US census Bureau (https://www.census.gov/) were among substantial socio-economic factors in describing LRI hotspots. Although we cannot provide an explicit explanation for economic factors, poor access to basic treatments is a plausible explanation. The findings were consistent with a large body of literature worldwide. LRI was found predominantly in the disadvantaged populations in South Auckland, New Zealand [54]. These populations were living in areas in the bottom quintile for socio-economic deprivation and with high rates of smoke exposure and poor living conditions. Similarly, impoverished children living in informal households without electricity and running water had approximately four times higher LRI mortality rates in South Africa [55].

There are several limitations of the current research study. First, the variables incorporated in the machine learning models undergoes several transformations and are susceptible to measurement or analysis errors. Also, neglecting the role of spatial autocorrelation, especially in sparse data, may produce biased estimates of the importance of variables. Another limitation is attributed to the selection of spatial scale. The values within each county are uniform, but there might be sharp contrasts between neighboring sub-counties, however, the choice of the spatial unit was dictated by the available data. Future studies should analyze and predict hotspots of LRI at the sub-county level, such as zip code or census tract levels, for targeted human interventions, particularly for Virginia, Kentucky, and Georgia, which were persistently identified as LRI hotspots. Additionally, future LRI studies should incorporate the concentration of other criteria air pollutants such as ground ozone, Sulphur oxides, lead, carbon monoxide, and nitrogen oxides as they may cause serious damages to internal organs especially to lungs which can lead to a higher mortality of LRI.

To our knowledge, this is the first study that incorporated national datasets on the LRI mortality rate using machine learning algorithms. Despite the above limitations, these findings have important public health implications. Predicting why the counties with high LRI mortality rates cluster geographically can be helpful further to reduce mortality in these regions. Moreover, the results of decision tree modeling can provide insight for future research geared toward identifying contributing factors such as median household income and climate factors to elevated LRI mortality rates. Despite significant efforts for mitigating mortality of LRI, there are many clustered counties, particularly in Georgia, Kentucky, and Virginia, where LRI mortality rates have remained elevated for the past 35 years.

CRediT authorship contribution statement

Abolfazl Mollalo: Conceptualization, Writing - original draft, Data curation, Formal analysis, Writing - review & editing. **Behrooz Vahedi:** Formal analysis. **Shreejana Bhattarai:** Writing - review & editing. **Laura C. Hopkins:** Writing - review & editing. **Swagata Banik:** Writing - review & editing. **Behzad Vahedi:** Conceptualization, Writing - review & editing.

Declaration of Competing Interest

The authors report no declarations of interest.

Acknowledgments

The first author would like to thank Professor Gregory Glass for kindly reviewing the earlier version of the manuscript. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.ijmedinf.2020.104248.

References

[1] P.V. Dasaraju, C. Liu, Infections of the respiratory system, Medical Microbiology, School of Medicine, University of Texas Medical Branch at Galveston, Galveston, TX, 1996 Accessed online from https://www.ncbi.nlm.nih.gov/books/NBK8142/#top on 6/13/2019.
[2] A. Mollalo, K.M. Rivera, B. Vahedi, Artificial neural network modeling of novel coronavirus (COVID-19) incidence rates across the Continental United States, Int. J. Environ. Res. Public Health 17 (12) (2020) 4204.
[3] A. Mollalo, B. Vahedi, K.M. Rivera, GIS-based spatial modeling of COVID-19 incidence rate in the continental United States, Sci. Total Environ. 728 (2020) 138884.
[4] V. Rahmanian, M. Shakeri, H. Shakeri, A.S. Jahromi, A. Bahonar, A. Madani, Epidemiology of influenza in patients with acute lower respiratory tract infection in south of Iran (2015-2016), Acta Fac. Med. Naissensis 36 (1) (2019) 27–37, https://doi.org/10.2478/afmnai-2019-0003.
[5] R.E. Malosh, E.T. Martin, J.R. Ortiz, A.S. Monto, The risk of lower respiratory tract infection following influenza virus infection: a systematic and narrative review, Vaccine 36 (1) (2018) 141–147, https://doi.org/10.1016/j.vaccine.2017.11.018.
[6] C. Troeger, B. Blacker, I.A. Khalil, P.C. Rao, J. Cao, S.R.M. Zimsen, et al., Estimates of the global, regional, and national morbidity, mortality, and aetiologies of lower respiratory infections in 195 countries, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016, Lancet Infect. Dis. 18 (11) (2018) 1191–1210, https://doi.org/10.1016/S1473-3099(18)30310-4.
[7] P. Torzillo, J. Dixon, K. Manning, S. Hutton, M. Gratten, L. Hueston, et al., Etiology of acute lower respiratory tract infection in Central Australian Aboriginal children, Pediatr. Infect. Dis. J. 18 (8) (1999) 714–721, https://doi.org/10.1097/00006454-199908000-00012.
[8] C.J.L. Murray, A.H. Mokdad, K. Ballestros, M. Echko, S. Glenn, H.E. Olsen, et al., The state of US health, 1990-2016: burden of diseases, injuries, and risk factors among US states, JAMA – J. Amer. Med. Assoc. 319 (14) (2018) 1444–1472, https://doi.org/10.1001/jama.2018.0158.
[9] K. Hasegawa, Y. Tsugawa, D.F.M. Brown, J.M. Mansbach, C.A. Camargo, Trends in bronchiolitis hospitalizations in the United States, 2000-2009, Pediatrics 132 (1) (2013) 28–36, https://doi.org/10.1542/peds.2012-3877.
[10] S.S. Huang, K.M. Johnson, G.T. Ray, P. Wroe, T.A. Lieu, M.R. Moore, et al., Healthcare utilization and cost of pneumococcal disease in the United States, Vaccine 29 (18) (2011) 3398–3412, https://doi.org/10.1016/j.vaccine.2011.02.088.
[11] J.M. Walter, R.G. Wunderink, Severe respiratory viral infections: new evidence and changing paradigms, Infect. Dis. Clin. North Am. 31 (3) (2017) 455–474, https://doi.org/10.1016/j.idc.2017.05.004.
[12] M. Sonego, M.C. Pellegrin, G. Becker, M. Lazzerini, Risk factors for mortality from acute lower respiratory infections (ALRI) in children under five years of age in low and middle-income countries: a systematic review and meta-analysis of observational studies, PLoS One 10 (1) (2015).
[13] S. Lapeña, M.B. Robles, L. Castañón, J.P. Martínez, S. Reguero, M.P. Alonso, I. Fernández, Climatic factors and lower respiratory tract infection due to respiratory syncytial virus in hospitalised infants in northern Spain, Eur. J. Epidemiol. 20 (3) (2005) 271–276.
[14] M. Mirsaeidi, H. Motahari, M. Taghizadeh Khamesi, A. Sharifi, M. Campos, D.E. Schraufnagel, Climate change and respiratory infections, Ann. Am. Thorac. Soc. 13 (8) (2016) 1223–1230.
[15] X. Wang, Y. Guo, G. Li, Y. Zhang, D. Westerdahl, X. Jin, et al., Spatiotemporal analysis for the effect of ambient particulate matter on cause-specific respiratory

mortality in Beijing, China, Environ. Sci. Pollut. Res. - Int. 23 (11) (2016) 10946–10956.
[16] C.T. McEvoy, E.R. Spindel, Pulmonary effects of maternal smoking on the fetus and child: effects on lung development, respiratory morbidities, and life long lung health, Paediatr. Respir. Rev. 21 (2017) 27–33.
[17] P.I. Beamer, N. Lothrop, Z. Lu, R. Ascher, K. Ernst, D.A. Stern, et al., Spatial clusters of child lower respiratory illnesses associated with community-level risk factors, Pediatr. Pulmonol. 51 (6) (2016) 633–642, https://doi.org/10.1007/978-3-319-46720-7.
[18] A.F. Beck, T.A. Florin, S. Campanella, S.S. Shah, Geographic variation in hospitalization for lower respiratory tract infections across one county, JAMA Pediatr. 169 (9) (2015) 846–854, https://doi.org/10.1097/CCM.0b013e31823da96d.Hydrogen.
[19] C.E. Reid, M. Jerrett, I.B. Tager, M.L. Petersen, J.K. Mann, J.R. Balmes, Differential respiratory health effects from the 2008 northern California wildfires: a spatio-temporal approach, Environ. Res. 150 (2016) 227–235.
[20] P.S. Heckerling, B.S. Gerber, T.G. Tape, R.S. Wigton, Use of genetic algorithms for neural networks to predict community-acquired pneumonia, Artif. Intell. Med. 30 (1) (2004) 71–84.
[21] K.M. Kuo, P.C. Talley, C.H. Huang, L.C. Cheng, Predicting hospital-acquired pneumonia among schizophrenic patients: a machine learning approach, BMC Med. Inform. Decis. Mak. 19 (1) (2019) 42.
[22] B. Bowe, Y. Xie, Y. Yan, Z. Al-Aly, Burden of cause-specific mortality associated with PM2. 5 air pollution in the United States, JAMA Network Open 2 (11) (2019) e1915834-e1915834.
[23] L. Dwyer-Lindgren, A.H. Mokdad, T. Srebotnjak, A.D. Flaxman, G.M. Hansen, C.J.L. Murray, Cigarette smoking prevalence in US counties: 1996-2012, Popul. Health Metr. 12 (1) (2014) 1–13, https://doi.org/10.1186/1478-7954-12-5.
[24] S. Niermeyer, P.A. Mollinedo, L. Huicho, Child health and living at high altitude, Arch. Dis. Child. 94 (10) (2009) 806–811, https://doi.org/10.1136/adc.2008.141838.
[25] A. Mollalo, A. Alimohammadi, M. Khoshabi, Spatial and spatio-temporal analysis of human brucellosis in Iran, Trans. R. Soc. Trop. Med. Hyg. 108 (11) (2014) 721–728.
[26] A. Mollalo, L. Mao, P. Rashidi, G.E. Glass, A GIS-Based artificial neural network model for spatial distribution of tuberculosis across the Continental United States, Int. J. Environ. Res. Public Health 16 (1) (2019) 157.
[27] T.H. Grubesic, R. Wei, A.T. Murray, Spatial clustering overview and comparison: accuracy, sensitivity, and computational expense, Ann. Assoc. Am. Geogr. 104 (6) (2014) 1134–1156, https://doi.org/10.1080/00045608.2014.958389.
[28] J. Aldstadt, Spatial clustering, in: M.M. Fischer, A. Getis (Eds.), Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications, Springer-Verlag, Berlin, 2010, pp. 279–300.
[29] A. Mollalo, A. Alimohammadi, M.R. Shirzadi, M.R. Malek, Geographic information system-based analysis of the spatial and spatio-temporal distribution of zoonotic cutaneous leishmaniasis in Golestan Province, North-East of Iran, Zoonoses Public Health 62 (1) (2015) 18–28.
[30] A. Mollalo, J.K. Blackburn, L.R. Morris, G.E. Glass, A 24-year exploratory spatial data analysis of Lyme disease incidence rate in Connecticut, USA, Geospat. Health 12 (2) (2017) 588.
[31] S.A. Naghibi, K. Ahmadi, A. Daneshi, Application of support vector machine, random forest, and genetic algorithm optimized random forest models in groundwater potential mapping, Water Resour. Manag. 31 (9) (2017) 2761–2775.
[32] P. Thanh Noi, M. Kappas, Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using Sentinel-2 imagery, Sensors 18 (1) (2018) 18.
[33] N. Bailey, T. Clements, J.T. Lee, S. Thompson, Modelling soil series data to facilitate targeted habitat restoration: a polytomous logistic regression approach, J. Environ. Manage. 67 (4) (2003) 395–407.
[34] D.W. Hosmer, S. Lemeshow, Applied Logistic Regression, John Wiley & Sons, New York, 2000.
[35] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[36] H. Bostrom, Estimating class probabilities in random forests, Sixth International Conference on Machine Learning and Applications (ICMLA 2007), IEEE, 2007, pp. 211–216.
[37] T. Hastie, R. Tibshirani, J. Friedman, Random forests, The Elements of Statistical Learning, Springer, New York, NY, 2009, pp. 587–604.
[38] A. Mollalo, A. Sadeghian, G.D. Israel, P. Rashidi, A. Sofizadeh, G.E. Glass, Machine learning approaches in GIS-based ecological modeling of the sand fly *Phlebotomus papatasi*, a vector of zoonotic cutaneous leishmaniasis in Golestan province, Iran, Acta Trop. 188 (2018) 187–194.
[39] J.H. Friedman, Stochastic gradient boosting, Comput. Stat. Data Anal. 38 (4) (2002) 367–378.
[40] L.E. Peterson, K-nearest neighbor, Scholarpedia 4 (2) (2009) 1883.
[41] V. Vapnik, Principles of risk minimization for learning theory, Advances in Neural Information Processing Systems, (1992), pp. 831–838.
[42] H. Yoon, S.C. Jun, Y. Hyun, G.O. Bae, K.K. Lee, A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer, J. Hydrol. 396 (1-2) (2011) 128–138.
[43] B. Scholkopf, A.J. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond, MIT press, 2001.
[44] C. Goutte, E. Gaussier, A probabilistic interpretation of precision, recall and F-score, with implication for evaluation, European Conference on Information Retrieval, Springer, Berlin, Heidelberg, 2005, pp. 345–359.
[45] L.S. Fischer, S. Santibanez, R.J. Hatchett, D.B. Jernigan, L.A. Meyers, P.G. Thorpe, M.I. Meltzer, CDC grand rounds: modeling and public health decision-making, Morbid. Mortal. Weekly Rep. 65 (48) (2016) 1374–1377.
[46] GBD 2016 Lower Respiratory Infections Collaborators, Estimates of the global, regional, and national morbidity, mortality, and aetiologies of lower respiratory infections in 195 countries, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016, Lancet Infect. Dis. 18 (11) (2018) 1191.
[47] J. Allyn, N. Allou, P. Augustin, I. Philip, O. Martinet, M. Belghiti, et al., A comparison of a machine learning model with EuroSCORE II in predicting mortality after elective cardiac surgery: a decision curve analysis, PLoS One 12 (1) (2017).
[48] M.M. Churpek, T.C. Yuen, C. Winslow, D.O. Meltzer, M.W. Kattan, D.P. Edelson, Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards, Crit. Care Med. 44 (2) (2016) 368.
[49] M.Z. Hossain, H. Bambrick, D. Wraith, S. Tong, A. Khan, S. Hore, W. Hu, Sociodemographic, climatic variability and lower respiratory tract infections: a systematic literature review, Int. J. Biometeorol. 63 (2 PG-209–219) (2019) 209–219, https://doi.org/10.1007/s00484-018-01654-1.
[50] E. Thomas, M.J. Margach, C. Orvell, B. Morrison, E. Wilson, Respiratory syncytial virus subgroup B dominance during one winter season between 1987 and 1992 in Vancouver, Canada, J. Clin. Microbiol. 32 (1) (1994) 238–242.
[51] P.W.K. Chan, F.T. Chew, T.N. Tan, K.B. Chua, P.S. Hooi, Seasonal variation in respiratory syncytial virus chest infection in the tropics, Pediatr. Pulmonol. 34 (1) (2002) 47–51, https://doi.org/10.1002/ppul.10095.
[52] V. Erling, F. Jalil, L.Å. Hanson, S. Zaman, The impact of climate on the prevalence of respiratory tract infections in early childhood in Lahore, Pakistan, J. Public Health Med. 21 (3) (1999) 331–339, https://doi.org/10.1093/pubmed/21.3.331.
[53] R.Q. Gurgel, P.G. De Matos Bezerra, M. Do Carmo Menezes Bezerra Duarte, A.Á. Moura, E.L. Souza, L.S. Da Silveira Silva, et al., Relative frequency, possible risk factors, viral codetection rates, and seasonality of respiratory syncytial virus among children with lower respiratory tract infection in Northeastern Brazil, Medicine (United States) 95 (15) (2016) 1–8, https://doi.org/10.1097/MD.0000000000003090.
[54] A.A. Trenholme, E.J. Best, A.M. Vogel, J.M. Stewart, C.J. Miller, D.R. Lennon, Respiratory virus detection during hospitalisation for lower respiratory tract infection in children under 2 years in South Auckland, New Zealand, J. Paediatr. Child Health 53 (6) (2017) 551–555, https://doi.org/10.1111/jpc.13529.
[55] H.K. Hutton, H.J. Zar, A.C. Argent, Clinical features and outcome of children with severe lower respiratory tract infection admitted to a pediatric intensive care unit in South Africa, J. Trop. Pediatr. 65 (1) (2019) 46–54, https://doi.org/10.1093/tropej/fmy010.