APPLICATION OF GRADIENT-BOOSTING AND DEEP LEARNING MODELS ON INFECTIOUS DISEASE OUTBREAKS

ABDULLAH, K-K.A.[1]*, SODIMU, S.M.[1], ODULE, T.J.[1] AND EZIMA, E.N.[2]
[1]Department of Mathematical Sciences, [2]Department of Biochemistry,
Olabisi Onabanjo University, Ago Iwoye, Nigeria

ABSTRACT

This study demonstrates the need for events on Internet news to limit the spread of infectious disease in sub-Sahara Africa. Evaluating the quality of surveillance system for preventing future trends with Internet health news, provides awareness and information in real-time. The dataset is modelled for weekly short-term outbreaks with 154,057,341 reported cases extracted from regional sub-Sahara Africa countries from 2010-2020. AdaBoost, Extreme Gradient Boost (XGBoost), Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) are employed for the evaluation of the internet health data. These models are adopted with varying Learning Rates (LR) to determine a better model for real-time predictions and more affected region. Comparative analysis for disease trends and predictions are performed using Mean Square Error and Mean Absolute Error. The results show there is spike in East region between 386 weeks, Central sub-Sahara has minimum values of 1.90e-06 and 7.49e-03 for Mean Square Error and Mean Absolute Error, respectively. The results show that XGBoost predicted more with low training time than CNN and LSTM, though, not affected with varying LR values while CNN predicted better with large LR of 0.1 with high training time. In conclusion, XGBoost works better at vary learning rate compared to deep learning models.

Keywords: Deep learning, Gradient-Boost, HealthMap, Infectious diseases
***Correspondence:** abdullah.adebisi@oouagoiwoye.edu.ng, +2348060046592

INTRODUCTION

Limiting infectious disease trend, prevention, and control strategies is a complex task in the context of emerging epidemics to make health-related decisions in real-time. Prediction intend to detect abnormal distribution of infectious diseases by selecting appropriate approaches for prevention and control of epidemics with decrease economic and social losses. In under-develop and developing Africa countries, the effect of epidemics can be terrible due to shortfall in health sector, poverty level and nutritional problem. The effect of disease outbreaks goes beyond human health but creates anxiety in the general populace and leads to national security problem, therefore, requires immediate and effective action [1]. Update the public and healthcare provider about disease outbreaks are mostly done by traditional surveillance system, this involves time-delay in processing information or incomplete report to make decisions. It is important to create a large datasets infectious disease predictions model to handle situations outbreaks in real-time which are difficult to cope with. Recently, early disease outbreaks detection can be discovered with Internet surveillance system or social network sites [2]. Studies have shown that large amount of information in real time can be found on Internet search queries to detect the spread of diseases, the studies can be found in [3, 4], HealthMap [5] etc. Furthermore, collection of real-time information on disease outbreaks from Internet news HealthMap database discovered that Internet data fulfilled the goal from search engine as [6].

Subsequently, models such as statistical linear techniques like Autoregressive Integrated Moving Average (ARIMA) have performed well with specified prediction time series [7] with different infectious disease [8, 9] but cannot accommodate many real applications [10]. Machine learning and deep learning model have gained much interest in the field of text analysis and data-driven with drawn attention and rival to classical statistical models on epidemics [11]. Machine learning models provide good accuracy in validating large datasets to identify disease outbreaks, also, studies have shown that using deep learning to predict diseases outbreaks yield better results when performed tasks that are difficult for conventional or statistical methods [12, 13].

These models extract meaningful patterns from complex datasets; translate into decisions and takes advantage of the growth in health-data to tackle the problem [14]. Extreme Gradient Boosting (XGBoost) and Adaptive-Boost (Adaboost) are decision tree ensemble classifiers based on a gradient boosting algorithm. XGBoost build an additive expansion of the objective function and control over fitting with more regularised model formalisation for the structured features [15, 16]. While, Adaptive-Boost (Adaboost) is applied on weak learner and train weight of each attributes to regulate the output in each iteration accuracy [17, 18], deep learning models like Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM) are used to compensate for the shortcomings of machine learning models. LSTM models adapt with number of parameters and capture local dependency patterns with optimised stepwise increase [19]. Bychkov *et al.*, used CNN and LSTM models to predict Colorectal Cancer on patients' data

with Area Under the Curve (AUC) as the metric [12]. However, it was demonstrated that gradient-boost and deep learning models generate better predictions performance than the conventional methods.

It is necessary to find models that can deal effectively with large amount of data with imbalance and complex data. In respect of this, ensemble gradient-boost models [9] and deep learning models are employed to randomize and generate different solutions with improve performance. This study is conducted on the information embedded in Internet news for infectious diseases. The approach involves using non-linear classification predictive models such as XGBoost, Adaboost, CNN, and LSTM to predict 2010-2020 trends of reported cases of infectious disease in sub-Sahara African using HealthMap. The data are extracted for each region of sub-Sahara Africa; Western, Central, Eastern, Southern Africa and these are used as input parameter to make decisions which are significantly and independently associated with infectious disease data extracted to predict disease trend, prevention, and control. The aim of this study is to use non-linear models for weekly predictions performance by tuning the optimization hyper-parameters like learning rate to demonstrate the effectiveness of the models using Mean Square Error (MSE) and Mean Absolute Error (MAE).

MATERIALS AND METHODS

Search strategy and data collection

The search by query on HealthMap site is conducted (http://healthMap.org), archive from 2010-2020. The search terms and keywords used on HealthMap are "infectious disease in sub-Sahara Africa" and "infectious disease in regional sub-Sahara Africa". Data pre-processing and normalization are performed using suitable attributes of reported cases extracted from HealthMap. However, redundant attributes or missing values are removed to enhance quality predictions. Thus, numbers of infectious disease cases reported on HealthMap sites are acquired on daily basis, combined to form weekly reports to identify short-term infectious disease activities trends.

A total number of 154,057,341 reported cases are extracted for four (4) regional sub-Sahara Africa countries which resulted into 4,694,475; 66,713,550; 66,562,795; 16,086,521 cases for Central, East, Southern and West Africa, respectively. Figure 1 shows spike in the East region between 386 weeks.

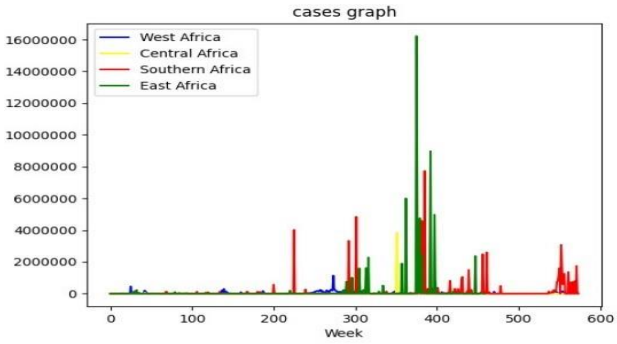


Figure 1: Weekly reported cases for infectious disease for four (4) regions of sub-Sahara Africa on HealthMap

Predictive models

In this study, the non-linear predictive models; Extreme Gradient Boost (XGBoost), Adaptive boost (Adaboost), Convolutional Neural Network (CNN), and Long-Short Term Memory (LSTM) are employed for modelling Internet news processed dataset. Given the dataset $d = \{x_i, y_i)$ for $i, \cdots, K$ and $y_1(x), y_2(x), \cdots, y_K(x)$ represent attributes with x input. The description of the predictive models as described in the subsection below:

Model 1: Adaptive Boost (AdaBoost)

It is an ensemble model that calculates the weighted average of weaker classifiers by summing the weighted predictions and put emphasis on selected features or pattern that are known as represented in Equation (1)

$$F_{Ad}(x) = sign\sum_{i=1}^{K} \alpha_i y_i(x) \quad (1)$$

Model 2: Extreme Gradient Boost (XGBoost)

This is a decision tree ensemble classifier based on a gradient boosting algorithm with high scalability. The model minimises the loss $L(y, f(x))$ with λ to control the complexity of the tree as depicted in Equation (2a &2b)

$$L_{XGB} = \sum_{k=1}^{K} L(y_i, f(x_i)) + \sum_{i=1}^{N} \Omega(\hbar_n) \quad (2a)$$

$$\Omega(\hbar) = \lambda t + \frac{1}{2}\eta\|w\|^2 \quad (2b)$$

Where L is the training loss function while Ω is the regularization term which reduce the step size, t is the size of leaves of the trees and w is the output of the leaves.

Model 3: Convolutional Neural Network (CNN**)**

The prediction occurs where the inputs (*x*) of the network are passed to the hidden layer (*h*) and output of each node serves as input to the other hidden layer until the finally output layer (*y*). The parameters weight (w), *x* is the input vector, and bias (b) is

randomly initialised as given in Eq. (3) while Rectified Linear Unit (ReLU) is added to the network in the hidden layers as depicted in Eq. (4) to capture complex relationship by improving the learning speed of the deep learning model but cause vanishing gradient problem.

$$f(w,x) = f(w_i^T x + b_i) \tag{3}$$

$$f(\mathrm{Re}LU(h_{i,k})) = \max(0, h_{i,k}) \tag{4}$$

Model 4: Long Short-Term Memory (LSTM)

LSTM solves vanishing gradient problem due to the memory cell connected through each layer. The cell contains gates that manage the cell state and output so that the input on the hidden layer (*h*) can increase or decrease during the forward and backward process. The responsive of input decreases over time (*t*) as new input are forgotten but due to the memory cell, forget gate is open and input gate is blocked, memory cell continues to remember the previous input. Therefore, three (3) gates are represented as input (*i*), output (*o*) and forget (*f*) gates control non-linear activation function (δ) . The equations for forget, memory cell, input, and output information at time-step *t*are represented in Eq. (5) as follows:

$$\left.\begin{aligned} f_t &= \delta(w_f x_t + \tau_f h_{t-1} + b_f) \\ i_t &= \delta(w_i x_t + \tau_i h_{t-1} + b_i) \\ o_t &= \delta(w_o x_t + \tau_o h_{t-1} + b_o) \\ \tilde{\beta}_t &= \tanh(w_{\tilde{\beta}} x_t + \tau_{\tilde{\beta}} h_{t-1} + b) \\ \beta_t &= f_t^o \beta_{t-1} + i_t^o \tilde{\beta}_t \end{aligned}\right\} \tag{5}$$

where

w =weighted matrix

τ = recurrent connection between previous and current hidden layer

$\tilde{\beta}$ = candidate hidden state computed based on the current input and the previous hidden state

β = internal memory unit

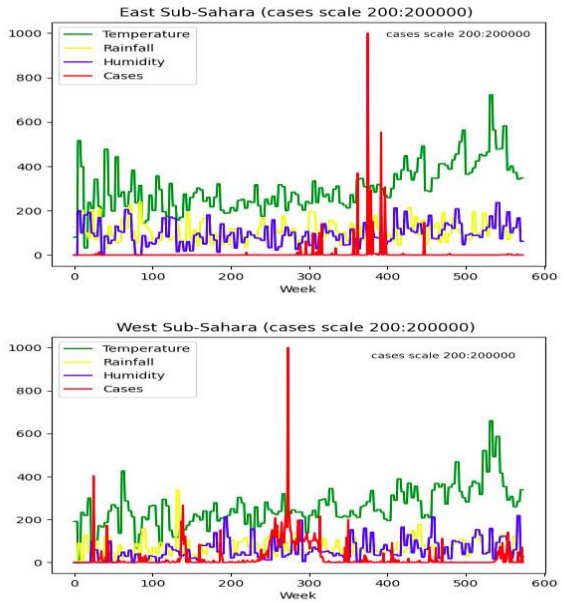
Implementation for the parameters setting

The parameters used includes weekly reported regional cases of infectious diseases on Influenza, Malaria, Tuberculosis, Cholera, SARS, HIV-AIDS, Ebola, Dengue, Measles, Zika, Brucellosis, temperature, humidity and rainfall which are fixed for all the four (4) regions of sub-Sahara Africa. The proposed predictive models considered 200 epochs, batch size (8), dropout rate (0.2), number of hidden layers is 2 while the number of nodes in hidden layer is 64 for CNN and LSTM. Apparently, the learning rate varies from 0.00001, 0.0001, 0.001, 0.01, 0.1 for all the models with different training time (*t*). Table 1 shows the details of the four (4) regions of sub-Sahara Africa datasets.

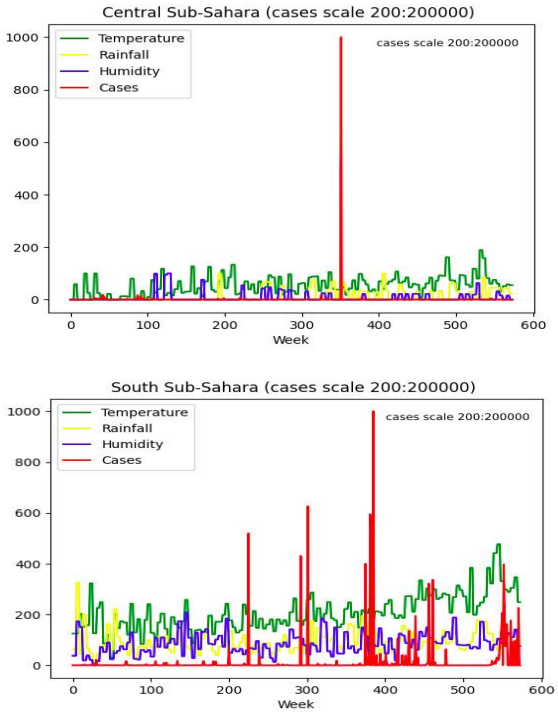
Table 1: The detail description of datasets for four (4) regions of sub-Sahara Africa

Sub-Sahara Region	Training	Test	Total Dataset
Central	3,755,580	938,895	4,694,475
Southern	53,250,236	13,312,559	66,562,795
East	53,370,840	13,342,710	66,713,550
West	12,869,216.8	3,217,304.2	16,086,521

However, the dataset is split into a training and test set of ratios 80:20 respectively to have better predictions. Also, the effect of integrating exogenous data with reported cases are analysed in Figure 2(a-d) as shown. The Figure 2a-d is scaled to 200:200000 for infectious diseases cases to accommodate the effect of exogenous data like humidity, rainfall and temperature. Higher humidity correlated with higher infectious diseases cases like dengue fever, malaria etc., in all the four regions. While longer season of mild temperature lead to transmission of vector-borne diseases. Consequently, rainfall increases infectious diseases with pathogens like malaria. There is spike in week 376, 274, 352, 386 for West, East, Central and Southern for Figure 2 (a-d), however, the overall trend was consistent in the remaining weeks.



Figures 2a & b: Effect of integrating exogenous data with Internet Health Data for East and West sub-Sahara



Figures 2c & d: Effect of integrating exogenous data with Internet Health Data for Central & South sub-Sahara

Predictive models analysis

In the predictive evaluation metrics, the training set is represented as $x_I \in \Re^n, i = 1, \cdots K$while testing set is represented as $x_i' \in \Re^n, i = 1, \cdots, K'$. However, the two (2)metrics; Mean Absolute Error: (MAE) and Mean Square Error (MSE) are used for evaluation as depicted in Eq. (6) and Eq. (7) with Average loss= $\frac{1}{K'} \ell((x^{\bullet})^T \phi(x_i'), (y_i')$ as generalization performance. The models utilised Python package for the implementation.

$$MSE = \frac{1}{K} \sum_{i=1}^{K} (o_i - \rho_i)^2 \quad (6)$$

$$MAE = \frac{1}{K} \sum_{i=1}^{K} \left| \frac{(o_i - \rho_i)}{o_i} \right| \quad (7)$$

where o_i, ρ_i, $\hat{o}_i$, and $\hat{\rho}_i$ are the actual cumulative cases, predicted cumulative cases, average cumulative cases and average predicted cumulative cases, respectively, and *K* is the total number of data point.

2 RESULTS AND DISCUSSION

Experimental results

In this experiment, extensive comparative analysis on XGBoost, AdaBoost, CNN and LSTM are used in studying the performance of predictions of infectious diseases due to the real-time datasets extracted from Internet HealthMap. In the test sets, five (5) varying Learning Rates (LR) of 0.00001, 0.0001, 0.001, 0.01, 0.1 are used in the models with fixed parameters for epochs, batch size, dropout, hidden layers and nodes for Gradient Boost and Deep Learning on East, West, Southern and central sub-Sahara Africa respectively. The experiments are evaluated using the MSE and MAE as depicted in Table 2(a-d) and 3(a-d).

Table 2a: Predictive performance of the Gradient Boost Models with different LR for East sub-Sahara Africa

LR	XGBOOST			ADABOOST		
	MSE	**MAE**	**Time**	**MSE**	**MAE**	**TIME**
0.000001	5.09e-03	7.49e-02	0.072	3.26e-03	5.98e-02	0.122
0.00001	5.09e-03	7.49e-02	0.065	3.25e-03	5.56e-02	0.136
0.0001	5.09e-03	7.49e-02	0.059	3.28e-03	5.54e-02	1.218
0.001	5.09e-03	7.49e-02	0.086	3.41e-03	8.39e-02	0.136
0.01	5.09e-03	7.49e-02	0.104	3.80e-03	1.14e-01	0.118

Table 2b: Predictive performance of the Gradient Boost Models with different LR for West sub-Sahara Africa

LR	XGBOOST			ADABOOST		
	MSE	**MAE**	**Time**	**MSE**	**MAE**	**TIME**
0.000001	1.30e-02	2.22e-01	0.060	1.39e-02	2.30e-01	0.122
0.00001	1.30e-02	2.22e-01	0.062	1.39e-02	2.30e-01	0.105
0.0001	1.30e-02	2.22e-01	0.088	1.39e-02	2.30e-01	0.118
0.001	1.30e-02	2.22e-01	0.078	1.36e-02	2.27e-01	0.110
0.01	1.30e-02	2.22e-01	0.085	1.14e-02	2.06e-01	0.127

Table 2c: Predictive performance of the Gradient Boost Models with different LR for South sub-Sahara Africa

LR	XGBOOST			ADABOOST		
	MSE	**MAE**	**Time**	**MSE**	**MAE**	**TIME**
0.000001	1.53e-01	8.37e-01	0.110	9.93e-02	5.15e-01	0.123
0.00001	1.53e-01	8.37e-01	0.076	9.93e-02	5.18e-01	0.103
0.0001	1.53e-01	8.37e-01	0.082	9.92e-02	5.18e-01	0.101
0.001	1.53e-01	8.37e-01	0.058	9.83e-02	5.19e-01	0.102
0.01	1.52e-01	8.37e-01	0.057	9.65e-02	5.20e-01	0.111

Table 2d: Predictive performance of the Gradient Boost Models with different LR for Central sub-Sahara Africa

LR	XGBOOST			ADABOOST		
	MSE	**MAE**	**Time**	**MSE**	**MAE**	**TIME**
0.000001	2.86e-05	3.39e-02	0.072	2.20e-06	1.06e-02	0.095
0.00001	2.85e-05	3.39e-02	0.071	2.20e-06	1.06e-02	0.099
0.0001	2.86e-05	3.39e-02	0.076	2.20e-06	1.07e-02	0.094
0.001	2.86e-05	3.39e-02	0.058	2.03e-06	9.96e-03	0.101
0.01	2.86e-05	3.39e-02	0.070	2.68e-06	1.18e-02	0.091

Table 3a: Predictive performance of the Deep Learning Models with different LR for East sub-Sahara Africa

LR	CNN			LSTM		
	MSE	**MAE**	**TIME**	**MSE**	**MAE**	**TIME**
0.000001	3.29e-03	3.25e-02	12.628	1.67e-02	3.94e-01	43.914
0.00001	3.29e-03	3.25e-02	12.745	5.71e-03	1.98e-01	43.530
0.0001	3.29e-03	3.25e-02	12.721	1.01e-02	2.92e-01	43.401
0.001	3.29e-03	3.25e-02	12.875	3.61e-03	8.13e-02	43.844
0.01	3.29e-03	3.25e-02	12.702	3.29e-03	3.25e-02	43.908

Table 3b: Predictive Performance of the Deep Learning Models with different LR for West sub-Sahara Africa

LR	CNN			LSTM		
	MSE	MAE	TIME	MSE	MAE	TIME
0.000001	1.38e-02	2.01e-01	12.244	1.68e-02	3.02e-01	44.237
0.00001	1.38e-02	2.01e-01	12.605	1.84e-02	3.99e-01	44.070
0.0001	1.38e-02	2.01e-01	12.704	1.94e-02	4.09e-01	44.179
0.001	1.38e-02	2.01e-01	12.495	2.21e-02	4.31e-01	44.099
0.01	1.38e-02	2.01e-01	12.239	1.62e-02	3.75e-01	43.899

Table 3c: Predictive Performance of the Deep Learning Models with different LR for Central sub-Sahara Africa

LR	CNN			LSTM		
	MSE	**MAE**	**TIME**	**MSE**	**MAE**	**TIME**
0.000001	1.90e-06	7.49e-03	5.639	4.04e-02	4.51e-01	20.141
0.00001	1.90e-06	7.49e-03	5.592	1.04e-02	2.73e-01	18.803
0.0001	1.90e-06	7.49e-03	5.681	2.11e-02	3.09e-01	19.803
0.001	1.90e-06	7.49e-03	5.822	3.31e-03	1.19e-01	18.773
0.01	1.90e-06	7.49e-03	5.461	3.21e-02	3.44e-01	19.783

Table 3d: Predictive Performance of the Deep Learning Models with different LR for South sub-Sahara Africa

LR	CNN			LSTM		
	MSE	**MAE**	**TIME**	**MSE**	**MAE**	**TIME**
0.000001	1.06e-01	5.29e-01	11.850	8.33e-02	5.64e-01	41.266
0.00001	1.06e-01	5.28e-01	11.709	8.50e-02	5.51e-01	40.387
0.0001	1.06e-01	5.28e-01	11.642	7.97e-02	6.01e-01	41.152
0.001	1.06e-01	5.28e-01	11.541	9.08e-02	5.28e-01	40.769
0.01	1.06e-01	5.28e-01	11.497	8.80e-02	5.35e-01	41.550

DISCUSSION

The actual number of infectious diseases reported cases are presented in Table 2. However, the predicted values of the four (4) regional sub-Sahara shows that, the results are close to the target values due to small number of MAE and MSE generated as shown in the Table 3 & 4 for Gradient Boosting and Deep learning respectively. According to Table 3(a-d), as the number of cases increases, MAE and MSE reduces with increase training time but as the cases reduced in Central sub-Sahara, the evaluation metrics of MAE and MSE reduces with increase training time for both Adaboost and XGBoost. As the reported cases become smaller for West sub-Sahara in Table 3b compared to East and Southern Sub-Sahara, the MAE and MSE become higher with low training time for XGBoost and Adaboost respectively. However, the metrics of MAE and MSE for XGBoost shows better results with low training time. However, with constant values for each LR, this indicated that XGBoost is not affected with different LR values. In Table 4 (a-d), the predicted value is close to target when the value of the reported cases become smaller in Central sub-Sahara for CNN compared to Adaboost with higher training time. For LSTM, the result become better when the value of LR is higher with higher training time. These predictive models have great performances for early warning. Although, the effect of climatic factors such as temperature, humidity and rainfall of infectious diseases are not considered, the predicted results may have bias to some degree. Consequently, correlation analysis of the searched keywords is based on vocabularies data on Internet HealthMap.

CONCLUSION

The study investigated the prediction of infectious disease based on Internet news data using gradient boost and deep learning models. The study presented

Internet news data by integrating climatic factor to study the effect on the extracted search query data in sub-Sahara Africa. The models based on Internet news data showed the best accuracy is achieved with gradient boost more efficiently than XGBoost as well as deep learning prediction models. This study can be extended by combining climatic factors to the independent variable in predicting improve reliability of the model. Also, several models can be combined as ensemble model to improve the reliability of the model.

ACKNOWLEDGEMENT

The information used in the experiment can be found in Googletrend, HealthMap, Tensorflow, Keras, Python. Acknowledge goes to the Dr. Adeeko Faizal Oluseun, of Department of Outpatient, Olabisi Onabanjo University Teaching Hospital (OOUTH), Sagamu, Ogun State for his Medical Advice on the write up.

Declaration of Competing Interest

The author(s) declares that there is no competing financial interests or personal relationships that influence the work in this paper.

REFERENCES

1. LAU, J.T.F., GRIFFITHS, S., CHOI, K.C. & TSUI, H.Y. (2010). Avoidance behaviors and negative psychological responses in the general population in the initial stage of the H1N1 pandemic in Hong Kong, BMC *Infectious Diseases*, **10**(139). doi:10.1186/1471-2334-10-139.
2. CHRISTAKI, E. (2015). New technologies in predicting, preventing and controlling emerging infectious diseases, *Taylor and Francis Group, LLC, Virulence*, **6**(6): 558-565
3. JACKOWAY, A., SAMET, H. & SANKARANARAYANAN, J. (2011). Identification of live news events using twitter", In: Proceedings of the 3rd ACM Sigspatial International Workshop on Location-Based Social Networks. ACM, 25–32.
4. ZHANG, Y., MILINOVICH, G., XU, Z., BAMBRICK, H., MENGERSEN, K., TONG, S. & HU, W. (2017). Monitoring Pertussis Infections Using Internet Search Queries, *Science Report*, **7**(10437). doi:10.1038/s41598-017-11195-z.
5. BROWNSTEIN, J.S., FREIFELD, C.C., REIS, B.Y. & MANDL, K.D. (2008). Surveillance Sans Frontières: Internet based emerging infectious disease intelligence and the HealthMap project, *PLoS Medicine*, e151 **5**(7): 1019-1024. doi:10. 1371/journal.pmed.0050151.
6. SAMARAS, L., GARCIA-BARRIOCANA, E. & SICILIA, M.A. (2019). Syndromic surveillance models using Web data: a systematic review, Book by Lytras M., Sarirete A. Innovation in Health Informatics, 1st Edition, A Smart Healthcare Primer, Elsevier Science Publishing Co Inc., Imprint by Academic Press Inc 13.11.2019. ISBN: 9780128190432, ISBN10: 0128190434, ISBN13: 9780128190432, Chapter 2, pp 39–77. https://doi.org/10.1016/B978-0-12-819043-2.00002-2
7. CASCIO, A. BOSILKOVSKI, M. RODRIGUEZ-MORALES, A-J. & PAPPAS, G. (2011). The socio-ecology of zoonotic infections. Clinical microbiology and infection. *The European Society of Clinical Microbiology and Infectious Diseases*, **17:** 336-342.
8. JOHANSSON, M.A., REICH, N.., HOTA, N.A., BROWNSTEIN, J.S. & SANTILLANA, M. (2016). Evaluating the performance of infectious disease forecasts: A comparison of climate-driven and seasonal dengue forecasts for Mexico. *Scientific Reports* | **6**: 33707 | doi: 10.1038/srep33707
9. LIU, L., LUAN, R.S., YIN, F., ZHU, X.P. & LU, Q. (2016). Predicting the incidence of hand, foot and mouth disease in Sichuan province, China using the ARIMA model. *Epidemiology and Infection*, **144**: 144–151, 2016. https://doi.org/10.1017/s0950268815001144
10. POLGREEN, P.M., CHEN, Y., PENNOCK, D.M., NELSON, F.D. & WEINSTEIN, R.A. (2008). Using internet searches for influenza surveillance. *Clinical Infectious Diseases*, **47**(11): 1443–1448.
11. AHMED, N.K., ATIYA, A.F., EL-GAYAR, N. & EL-SHISHINY, H. (2010). An empirical comparison of machine learning models for time series forecasting. *Econometric Reviews*, **29:** 5-6.
12. BYCHKOV, D., LINDER, N., TURKKI, R., NORDLING, S., KOVANEN, P.E., VERRILL, C., WALLIANDER, M., LUNDIN, M., HAGLUND, C. & LUNDIN, J. (2018). Deep learning-based tissue analysis predicts outcome in colorectal cancer. *Scientific Reports*, **8**: 3395. doi;10.1 038/s41598-018-21758-3.
13. ESTEVA, A., KUPREL, B., NOVOA, R.A., KO, J., SWETTER, S.M., BLAU, H.M. & THRU, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Springer Nature*, **542:** 115-118. [CrossRef] [PubMed]

14. SILVER, D., SCHRITTWIESER, J., SIMONYAN, K., ANTONOGLOU, L., HUANG, A., GUEZ, A., HUBERT, T., BAKER, L., LAI, M., BOLTON, A., CHEN, Y., LILLICRAP, T., HUI, F., SIFRE, L., DRIESSCHE, G., GRAEPEL, T.T. & HASSABIS, D. (2017). Mastering the game of Go without human knowledge. *Nature*, **550**: 354–359.
15. CHEN, T. & GUESTRIN, C. (2016). XGBoost. In Proceedings of the 22nd ACM SIGKDD, International Conference Knowledge Discovery and Data Mining—KDD '16, San Francisco, CA, USA, 13–17th August, 785–794.
16. TIANQI, A. & CARLOS, G. (2016). Xgboost: A scalable tree boosting system, In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining New York, NY, USA, ACM, KDD, **16:** 785–794.
17. MA, S. & DU, T. (2010). Improved Adaboost face detection, In Measuring Technology and Mechatronics Automation (ICMTMA), 2010 *International Conference on Institute of Electrical and Electronics Engineers (IEEE)*, **2:** 434–437.
18. ZHANG, H., XIE, Y. & XU, C. (2011). Classifier Training Method for Face Detection Based on Adaboost. *In Transportation, Mechanical, and Electrical Engineering (TMEE), International Conference on. IEEE*, 731–734.
19. VOLKOVA, S., AYTON, E., PORTERFIELD, E.K. & CORLEY, C.D. (2017). Forecasting influenza-like illness dynamics for military populations using neural networks and social media. *PLoS ONE*, **12**(12): https://doi.org/10.1371/journal. pone.0188941
20. JEROME, H.F. (2001). Greedy function approximation: A Gradient Boosting Machine. *The Annals of Statistics*, **29**(5): 1189 – 1232.