Makine Öğrenmesi Yaklaşımlarını Kullanarak Salgınları Erken Evrede Tespit Etme Alanındaki Eğilimler

Literatür Makalesi/Review Article

Ali Şenol[1]*, Yavuz Canbay[2], Mahmut Kaya[3]

[1]Bilgisayar Mühendisliği Bölümü, Mühendislik ve Doğa Bilimleri Fakültesi, Gaziantep İslam Bilim ve Teknoloji Üniversitesi, Gaziantep, Türkiye
[2]Bilgisayar Mühendisliği Bölümü, Mühendislik ve Mimarlık Fakültesi, Kahramanmaraş Sütçü İmam Üniversitesi, Kahramanmaraş, Türkiye
[3]Bilgisayar Mühendisliği Bölümü, Mühendislik Fakültesi, Siirt Üniversitesi, Siirt, Türkiye
ali.senol@gibtu.edu.tr, yavuzcanbay@ksu.edu.tr, mahmutkaya@siirt.edu.tr
(Geliş/Received:10.02.2021; Kabul/Accepted:19.08.2021)
DOI: 10.17671/gazibtd.878089

Özet— Tüm dünyayı etkisi altına alan COVID-19, salgınları erken dönemde tespit etmeye çalışan çalışmaların önemini ortaya koymaktadır. Herhangi bir salgın erken aşamada tespit edilebilirse, hastalığa yakalanan kişi sayısını azaltabilir ve gerekli tedavi daha erken sürede bulunabilir ve ek olarak tedavi masrafları da azaltılabilir. Salgınların erken aşamada tespit edilmesini sağlayan en önemli veri işleme yaklaşımlarından makine öğrenmesi, yeni gelen verileri, olayı veya durumu tahmin etmek için matematiksel modelleri ve istatistiksel yöntemleri kullanır. Makine öğrenmesi yaklaşımlarıyla, tıbbi veriler analiz edilerek ve işlenerek hastalıklar hakkında tahminlerde bulunulabilir. Çünkü daha önce toplanan hasta verileri, makine öğrenmesi yöntemleri kullanarak hastalıkların teşhis edilmesine imkân sağlayabilir. Hastalıkların yanı sıra, daha önce toplanan veriler kullanılarak salgınlar hakkında da tahminlerde bulunulabilir. Daha önce ortaya çıkan salgınların yeniden ortaya çıkışını tahmin etmek için denetimli öğrenme yaklaşımları olan Naive Bayes, Destek Vektör Makineleri (DVM), Karar Ağaçları (KA), Rastgele Orman (RO) ve Yapay Sinir Ağları (YSA) gibi birçok yaklaşım olsa da, temel bileşenler ve kümeleme analizi gibi denetimsiz öğrenme yaklaşımları da kullanılarak daha önce benzeri görülmemiş salgınlar tespit edilebilir. Bu çalışmada, bu alanda çalışmak isteyen araştırmacılara ışık tutmak amacıyla salgınları tespit etmeye yönelik geliştirilmiş olan makine öğrenmesi yaklaşımlarının ayrıntılı bir analizi sunulmaktadır.

Anahtar Kelimeler— makine öğrenmesi, salgın erken tespiti, pandemi, tahmin

Trends in Outbreak Detection in Early Stage by Using Machine Learning Approaches

Abstract— COVID-19 pandemic affecting the whole world, reveals the importance of the studies that trying to detect the outbreaks in early stage. If any outbreak can be detected in an early stage, the number of infected people can be reduced, the necessary treatment can be found and treatment expenses can be also reduced. The most important data processing approaches enabling to detect outbreaks in an early stage are machine learning approaches, which use mathematical models and statistical background. With machine learning techniques, medical data can be analyzed and processed to make predictions of illnesses. Because, previously collected patient datasets help to perform these predictions. Beside illnesses, outbreaks can be also predicted by using these collected datasets. Machine learning techniques enable us to process labelled and unlabelled datasets with the help of supervised and unsupervised approaches, respectively. Although there are many supervised learning approaches like Naïve Bayes (NB), Support Vector Machine (SVM), Decision Trees (DT), Random Forest (RF) and Artificial Neural Networks (ANN) to predict the emergence of the outbreaks that appeared before, it is also possible to detect any outbreak which are unprecedented before by using unsupervised learning approaches like principal component and cluster analysis. In this study, it is aimed to present a detailed analysis of machine learning approaches in outbreak detecting area to give a lead to the researchers who want to work in this area.

Keywords— machine learning, outbreak early detection, pandemic, prediction

1. INTRODUCTION

Nowadays, COVID-19 pandemic has clearly shown that global outbreaks can cause death of millions of people in a very short time. Because of globalization, any infectious illness can very easily spread among countries via mobility of people. SARS, H1N1 and final COVID-19 are some of the examples for this situation. COVID-19 pandemic has caused people to find new methods or tools that can diagnose any outbreak in early stage to find treatments and/or vaccines. Within the last decade, machine learning based illness and outbreak diagnosis have been widely studied by the researchers.

With the growth of the technology, the size of data transferred to computer environment has reached tremendous sizes. Therefore, classical data processing and analyzing approaches like data mining has been insufficient and intelligent approaches like machine learning are developed [1]. Machine learning techniques are a subset of artificial intelligence that use mathematical and statistical models based on statistics, data mining, pattern recognition and predictive analysis to learn from the past data to find patterns and make predictions about unknown events, data or conditions [2, 3].

Machine learning techniques are mostly used in many areas such as finance, education, healthcare etc. Especially, healthcare is the most popular area for machine learning to produce valuable outputs. Disease diagnosis [4], drug discovery [5, 6] and robot surgery [7, 8] are some of the areas that machine learning is applied. Besides, medical data analytics, which take the advantage of machine learning, aim to find the causes of diseases, to predict diseases and to prevent human from diseases. As can be seen in Fig. 1, there are four stages of medical care processes which are prevention of diseases, detection of health condition, diagnosis of diseases and treatment to cure any disease. In prevention stage, risks that may make people ill can be detected by using machine learning algorithms on gathered data. Data that received from medical or smart devices like wristbands could be processed and analysed with machine learning algorithms to detect any disease. In diagnostics stage, any disease could be diagnosed thanks to machine learning techniques by processing the data. Finally, by using machine learning approaches, treatment methods or drugs can be found by analysing the collected data [9].

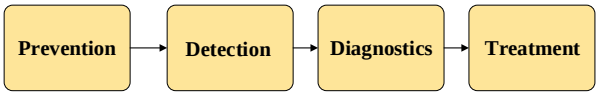

Figure 1. The stages of medical care [9]

The major contributions of this work can be listed as follows:

1. To present up-to-date information in the field of outbreak prevention, detection, diagnostics, and treatment trends with machine learning approaches,
2. To shed light on researchers who want to work in this area.

2. MACHINE LEARNING ALGORITHMS

Mainly, machine learning approaches are divided into four categories which are supervised, unsupervised, semi-supervised and reinforcement learning [10, 11]. These approaches are explained below and Fig. 2 shows us how supervised, unsupervised and semi-supervised learning methods work.

2.1. Supervised Machine Learning Algorithms

Supervised machine learning algorithms, which are known as classification approaches, use class labels of past data to predict any new arrived data [12]. Nearest Neighbour, NB, DT, Linear Regression (LR), SVM, and ANN are mostly used classification algorithms. These algorithms try to find patterns in dataset and create models. These models can make predictions about class label of a test data in the light of the past data. In classification problems, collected dataset is divided into two subsets which are known as train and test. Supervised learning algorithms use train dataset to train the model and then test dataset to test the model [13, 14].

2.2. Unsupervised Machine Learning Algorithms

Unsupervised machine learning algorithms, which are commonly known as clustering algorithms, does not require to know the class labels of any data [15]. Because, similarities of the elements of dataset are calculated and then the dataset is divided into various parts which are known as clusters according to these similarities. K-means, Principle Component Analysis (PCA) and Apriori algorithm for association rules are some examples of unsupervised machine learning approaches.

2.3. Semi-supervised Machine Learning Algorithms

These machine learning algorithms combine the advantages of supervised and unsupervised learning approaches [16]. Train dataset consist of some labelled and unlabelled data. The labelled data is very efficient to improve the accuracy of the model. Semi-supervised learning allows us to process the dataset without any selection between supervised and unsupervised machine learning algorithms. Self-training, Semi-supervised SVM and Co-training are some examples for semi-supervised learning algorithms, and speech analysis, internet content classification and protein sequence classification are some of its usage areas.

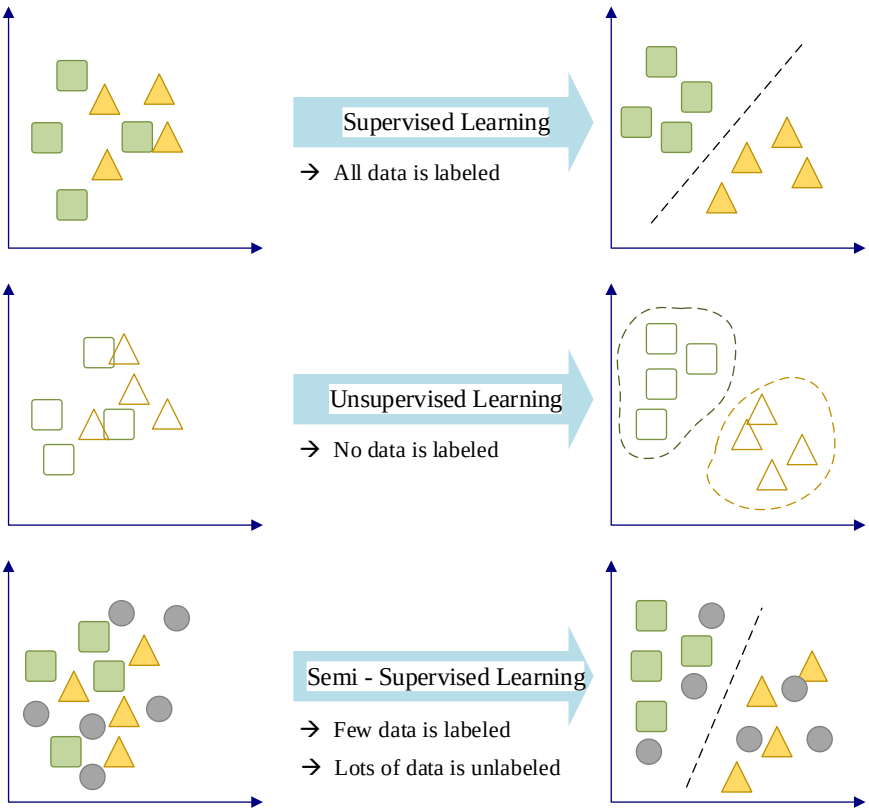

Figure 2. Supervised, unsupervised and semi–supervised learning

2.4. Reinforcement Machine Learning Algorithms

Unlike classical learning methods, reinforcement learning is used without any prior knowledge and in the conditions that classical methods do not work [17]. As can be seen in Fig. 3, reinforcement learning system includes two components; the agent and the environment [18]. The agent deals with the learning and decision process. The environment, on the other hand, is in contact with the agent and consists of everything that it cannot change. In this learning method, to select the best action, reward or punishment mechanism is used. Q-learning, SARSA, Markov Decision Process are some examples of this type of learning, and game programming, robotic, business strategy planning and disease diagnosis are some of its usage areas.

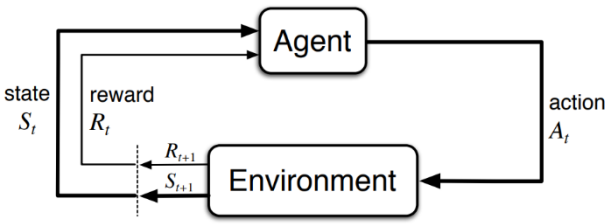

Figure 3. The agent–environment interaction in reinforcement learning [19]

2.5. Methods to Evaluate the Success of Machine Learning Algorithms

There are many proposed approaches in the literature to evaluate the success of machine learning algorithms. Mainly, these approaches are divided into two parts which are internal methods and external methods. In the internal methods, which are used for clustering techniques in general, the clustering success is calculated based on similarity among the data in the same cluster, and dissimilarity among the data in the different clusters via some metrics [1]. Silhouette Index, Dunn Index, Davies-Bouldin Index, Calinski-Harbasz are some of the metrics used for these purposes. On the other hand, in external methods, the success assessment is done based on comparison of actual class labels and the labels that the algorithm assigned to each data. Purity, Accuracy, Precision, Recall, F1-Score, Ran Index (RI), Normalized Mutual Information (NMI), Matthew's Correlation Coefficient (MCC), Receiver Operating Characteristic (ROC) curve, Mean Absolute Error (MAE) and Root-Mean Squared Error (RMSE) are some examples of external methods.

3. BACKGROUND INFORMATION ABOUT OUTBREAKS

This section introduces some background information about outbreaks which are briefed below.

3.1. Definition of Outbreak/Epidemic/Pandemic

Outbreak is the unexpected rise of in the number of any illness cases. It may take a few weeks or years and may spread to narrow area or wider area. On the other hand, epidemic is a more infectious disease that spread wider area when compared to outbreak. As for the pandemic, it can spread to countries or continents, and it is very difficult to protect people from it and to predict what an outbreak will evolve over time. Therefore, it is necessary to be careful against all kinds of outbreaks [20].

Various statistical and medical parameters are used to diagnose outbreaks. Some significant parameters are listed below [21]:

- Daily death count
- Number of germ carrier
- Incubation period
- Environmental parameters like, weather condition, wind, temperature
- Transmission rate
- Mobility
- Geographical location
- Age and gender
- Highly and least vulnerable population
- Underlying disease
- Report time
- Strategic policies and many more

To be successful when struggling against any outbreak, a systematic and collaborative way is always required. It is impossible for any country to overcome worldwide pandemic alone. As the COVID-19 pandemic experience has taught us, to overcome any pandemic it is required to struggle against it globally. To do this, as seen in the Table 1, its all stages like definition, confirmation, reporting must be managed collaboratively [22].

Table 1. Outbreak milestones [22]

Outbreak levels	Definition
Starting date of outbreak	The date when symptoms start
Detection date of outbreak	The date when predefined threshold is exceeded. Threshold value refers to minimum number of cases for particular diseases
Reporting date of outbreak	The date when the outbreak is reported to local, national or international authorities
Laboratory confirmation date of outbreak	The date when the disease is confirmed by laboratory results
Response date of public health	The date when the authorities take action to stop or take under control the outbreak
First public communication date	The date that information about the outbreak is shared via local, national or international medias like TVs, radios or presses etc.

3.2. Data Collection

The data belonging to any outbreak may be gathered from various sources which are shown in Fig. 4 [23]. This data is called as *background data* till any outbreak is defined. After defining the outbreak, this data is started to be called as *surveillance data* which describes the data of patients who may be infected by the outbreak. Since there is a relationship between outbreaks and time, surveillance data is in the form of time series. After collecting data, outbreak detection systems which are called as *bio-surveillance system*, takes the control and analyses surveillance data and tries to find the outbreaks patterns.

When any outbreak is defined, the data about three components are gathered, which are *time* when it is started, *place* where it is started and *people* who are infected. Firstly, outbreak curve according to the date (hourly, daily or weekly) is drawn to illustrate the timeline of the outbreak to understand its time dependency clearly. Then, its starting place is defined to take measures like water and food security and keep away people from risky area. Besides, infected people are identified by age, sex and profession to define risky people. All of these data are collected from and shared to World Health Organization's (WHO) partners to struggle against the outbreak [24].

3.3. Organizations and Programmes to Straggle against Outbreaks

WHO is the most operative international organization that tries to find out outbreaks in early stage. It collaborates with local health agencies to detect outbreaks. For example, in 2002, WHO collaborated with Iraqi Communicable Disease Control Center, and Republican Institute of Public Health, Belgrade to detect outbreaks. Similarly, in January 2003, WHO collaborated with the French Institut de Veile Sanitaire for the designing the specifications of a computerised early warning system. In 2004, WHO collaborated with United Nations agencies, nongovernmental organizations and Federal Ministry of Health and State Ministries of Health to detect infectious illnesses in North, South and West Darfur states [25].

The European Center for Disease Prevention and Control is another organization that operates to detect and struggle against outbreaks. Besides, the International Committee of the Red Cross is also an organization that struggles against outbreaks. In 2005, WHO revised International Health Regulation (IHR) to create a better network for more efficient disease and outbreak reporting. This caused to improve new networks like ProMEDmail and the Global Public Health Intelligence Network (GPHIN). In 2014, Global Health Security Agenda was established "to ensure safe and secure of the world against infectious disease threats and to promote global health security as an international security priority" [26].

The Global Outbreak Alert and Response Network (GOARN), which struggles against infectious disease outbreaks, natural disasters and other humanitarian emergencies, is a network of partners of WHO [25]. Academic and scientific institutions, medical and surveillance initiatives, laboratory networks, the Red Cross and United Nations (UN) organizations are some of

partners of the GOARN. UN Cluster system, which is constructed by the UN in 2005, aims to improve capacity for responding to humanitarian emergencies [27]. The system includes different partners to struggle against humanitarian emergencies threats.

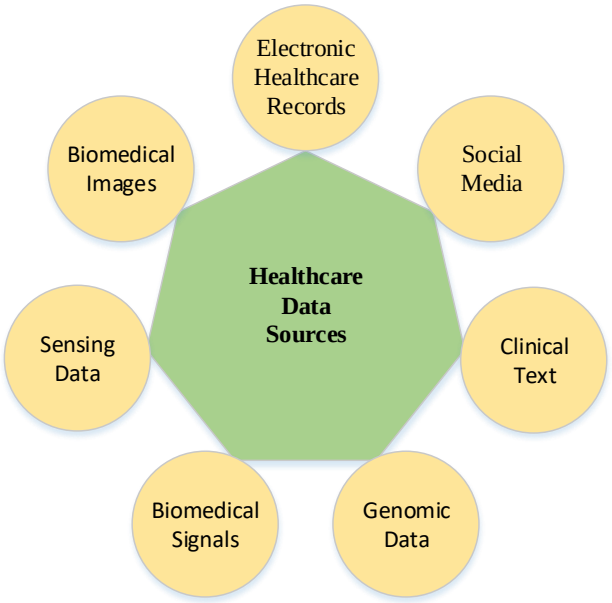

Figure 4. Surveillance data sources [23]

3.4. Outbreak Early Warning Systems

Early warning systems are the instruments to inform people about the risks that are oncoming hazardous events occur, thereby they ensure that measures are taken to reduce potential harm, and sometimes provide opportunities to prevent the dangerous event from occurring [28]. Disease or outbreak early warning systems are the systems that warn the authorities who are responsible from public health to take the required measures [15]. To take any outbreak under the control, public health agencies conduct disease surveillance, which is known as epidemiologic surveillance, gather and analysis human data to detect any outbreak in early stage to reveal the characteristics of the outbreak. The main objectives of epidemiologic surveillance are outbreak early event detection and outbreak situational awareness.

4. MACHINE LEARNING BASED OUTBREAK PREDICTION

Machine learning algorithms enables us to analyze surveillance data and compare observed results with expected values to find differences. To determine which difference is significant, generally a threshold value is used and when the threshold value is exceeded the system warns the expert about results. As can be seen in the Fig. 5, in the surveillance systems, firstly the data about the symptoms of events is aggregated. Then, to make the data suitable for processing operation, it is pre-processed by means of completing missing data, deleting unnecessary attributes and merging the data that aggregated from different sources. In the next step, a machine learning algorithm is used to make prediction about emergence of the outbreak.

Figure 5. Prediction model of outbreaks with machine learning techniques

The main machine learning algorithms to detect outbreaks can be classified into four parts as follows:

- *Environmental conditions based diagnosis systems;* predict outbreaks according to environmental factors like climates factors, weather conditions, etc.
- *Human data based outbreak diagnosis systems;* are mostly based on clinical data which are collected by health organizations like hospitals or clinics.
- *Specific outbreak diagnosis aimed systems;* aim to detect any outbreak which was diagnosed before. In these systems, symptoms are known and specifications of the outbreak are defined.
- *Unknown outbreak diagnosis aimed systems*; try to detect any unknown and mostly new outbreaks. Since no class labels are available, unsupervised machine learning algorithms are preferred.

There are various machine learning based outbreak prediction models as given in the Table 2, Table 3, Table 4, and Table 5. Most of them try to predict or detect any outbreak by using gathered data before. This means that many of them try to detect or predict known outbreaks like Ebola, Cholera, Swine Fever, H1N1 Influenza, Dengue Fever, Zika, Oyster Norovirus, etc. [29-31]. However, there is a need for approaches that could predict never seen outbreaks. Some of machine learning based outbreak prediction approaches can be summarized as follows.

4.1. Malaria Outbreak Prediction

Malaria is a highly fatal infectious disease transmitted to humans by mosquitoes biting humans. According to WHO, the number of deaths due to malaria in 2019 is 409000 [32]. Some machine learning methods [33-35] have been used to detect malaria outbreak in recent years. In [33], authors used SVM and ANN to predict malaria outbreak in early stage by using the data all 35 districts of Maharashtra, from 2011 to 2014. Temperature, humidity, monthly rainfall, total number of positive cases and total number of Plasmodium Falciparum cases are used as parameters. Root Mean Squared Error (RMSE) and Receiver Operating Characteristic (ROC) were used to compare accuracies of the algorithms. According to the results, SVM performed better than other methods in predicting the outbreak. In another study, Modu et al. [34] proposed an intelligent malaria outbreak early warning mobile application by using climatic factors and machine learning algorithms. Several machine learning algorithms were used to find the

most successful algorithm for the system. According to the obtained results, it was seen that SVM gave the most successful results. In 2020, Cömert et al. [35] proposed a method that defines malaria outbreak by using DT based algorithm. The dataset used in the experiments contains the features of maximum temperature, minimum temperature, humidity, rainfall amount, positive case, and Plasmodium Falciparum rate. As mentioned in the paper, the results showed that the proposed models could detect malaria outbreak with 100% accuracy on the test data.

Table 2. Comparison of the studies using machine learning techniques for malaria outbreak detection

Work	Disease / Outbreak	Year	Aim(s)	ML Algorithm(s)	Data / Place	Type	Highest Result / Success
[33]	Malaria	2015	Predicting malaria in early stage	SVM and ANN	Maharashtra state	Supervised	Predicting the malaria 15-20 days before rise
[34]	Malaria	2017	Malaria outbreak early warning mobile application	Several ML algorithms like SVM, KNN, NB, etc.	Twitter data	Supervised	Accuracy: 99%
[35]	Malaria	2020	To detect malaria outbreak	Random decision tree, LR and Gaussian process method	Maharashtra State	Supervised	Accuracy: 100%

4.2. Dengue Outbreak Prediction

Dengue fever is a type of mosquito-borne viral infection and 100-400 million people get this type of viral infection every year. With early diagnosis of the disease and proper medical care, it is possible to reduce the death rate less than one percent [36]. Machine learning methods can be used to diagnose this kind of disease. For this purpose, Zhu et al. [37] proposed a delay permutation entropy based SVM approach to find whether there was a correlation between the climate and dengue fever outbreak. To increase the accuracy of the model, they used global climate data beside the local data. In another study, Iqbal and Islam [38] used various machine learning algorithms which are LogitBoost, kNN, ANN, SVM, NB, DT, LR to predict dengue outbreak. The dataset used in the experiments is collected from different patients. According to the results LogitBoost was the best one which classified the test data with 92% accuracy.

Anno et al. [39] proposed a machine learning based deep AlexNet model to improve an early warning system which used climatic data in Taiwan. Reason to use this model was to find whether there was a correlation between climatic factors and dengue fever. According to the experimental results, their model yielded 100% accuracy on test dataset of longitude-time sea surface temperature images. Raja et al. [40] improved a Bayesian Network based dengue outbreak prediction model which is artificial intelligence driven, and used predictors of weather variables and vector indices sourced from the Ministry of Health. An improved model was used to predict the dengue fever outbreak in around the Klang Valley, Malaysia. The model predicts the outbreak nearly 79%-84% accuracy.

Amin et al. [41] collected tweets about dengue and flu infection. Between August 2018 and December 2019, 598911 tweets were received regarding these infections. LSTM with Word2Vec technique was used for infection detection. The proposed approach has achieved more successful results than known machine learning methods in the early detection of the disease.

Benedum et al. [42] aimed to use the advantage of machine learning methods for the detection of dengue outbreaks with early warning systems. In their study, the authors compared the capabilities of regression and time series models using dengue surveillance and weather data. As a result of the experiments performed for outbreak prediction, the ARIMA model has achieved very successful results for long-term predictions. When dengue surveillance, population, temporal and weather data were used in outbreak estimation, random forest (RF) method achieved more successful results than Poisson regression and ARIMA models.

4.3. Prediction of Various Outbreaks

Tapak et al. [31] proposed a machine learning based model to predict the time of future influenza outbreak. They used SVM, ANN and RF to model illnesses and detect outbreaks. The test results showed that ANN gives the best accuracy with 89.9%. The authors successfully predicted the date of infestation of various alien species and the new H1N1 influenza pandemic in 2009 by using their proposed method.

Foodborne is a type of illness that results from spoiled food, bacteria or viruses. Sadilek et al. [43] proposed a machine learning based model which aims to detect foodborne illness using anonymous and aggregated web search and location data. According to the results, it was reported that the proposed method produced successful results and may be integrated into existing inception protocols.

Oyster norovirus is one of the leading types of epidemics caused by oysters collected in coastal waters contaminated with sewage waters. Chenar and Deng [44] proposed an ANN based model to predict the oyster norovirus outbreak. They used six environmental data which are solar radiation, water temperature, wind, gage height, salinity and rainfall to train the model. They found that the oyster norovirus outbreak could be predicted in 2 days with the model before occurrence. The overall accuracy of the

model was 99.83%. Chenar and Deng [45] also proposed a genetic programming based approach to predict the oyster norovirus on the same data. They used RF and binary LR in the framework of genetic algorithm to construct the model. According to the results, their model was characterized with the area under the Receiver Operating Characteristic curve of 0.86

Table 3. Comparison of the studies using machine learning techniques for dengue outbreak detection

Work	Disease / Outbreak	Year	Aim(s)	ML Algorithm(s)	Data / Place	Type	Highest Result / Success
[37]	Dengue	2016	To find the correlation between the climate and dengue fever outbreak	Delay permutation Entropy and SVM	Hong Kong Department of Health data	Supervised	Accuracy: 63.4%
[38]	Dengue	2019	To predict dengue outbreak	LogitBoost, kNN, ANN, SVM, NB, DT	Dengue disease dataset	Supervised	Accuracy: 92%
[39]	Dengue	2019	To improve an early warning system for dengue outbreak	Deep learning	Climatic and meteorological data of Taiwan	Supervised	Accuracy: 100%
[40]	Dengue	2019	To predict the Aedes outbreak	Bayesian Network	The data that collected from the Ministry of Health, Malaysia	Supervised	Accuracy: 84%
[41]	Dengue	2020	Early detection of dengue outbreak in tweets	Deep Learning	Social media data	Unsupervised	Accuracy: 94.5%
[42]	Dengue	2020	To predict dengue case counts and outbreaks	RF, Poisson Regression, Logistic regression, and ARIMA	Dengue surveillance and weather data for Peru, Puerto Rico, and Singapur	Supervised	MCC: Peru :0.26, Puerto Rico: 0.53, Singapur: 0.14

Table 4. Comparison of the studies using machine learning techniques for various outbreaks' prediction

Work	Disease / Outbreak	Year	Aim(s)	ML Algorithm(s)	Data / Place	Type	Highest Result / Success
[31]	Influenza	2019	To predict the time of future influenza outbreak	SVM, ANN, RF	Dataset of weekly illness frequencies in Iran	Supervised	Accuracy: 89.9%
[43]	Foodborne	2018	To find unhealthy restaurants	FINDER	Web search and location data	Supervised	Accuracy: 74%
[44]	Oyster norovirus	2018	To predict the oyster norovirus outbreak	ANN	The Northern Gulf of Mexico coast	Supervised	Accuracy: 99.83%
[45]	Oyster norovirus	2018	To predict the oyster norovirus outbreak	RF and Binary Logistic Regression with framework of genetic programming	The Northern Gulf of Mexico coast	Supervised	ROC curve: 0.86
[47]	H1N1 flu	2018	To find the future distribution of species	Neural Network	Several cities in Japan	Supervised	Requirement of less info and applicability to all organisms
[48]	Swine fever	2020	To predict African swine fever outbreak	Random forest, ANN, SVM, AdaBoost, C4.5, and NB	African swine fever outbreak data and the WorldClim database meteorological data	Supervised	Accuracy: 98.24%
[49]	Cardiovascular diseases	2019	To predict the outbreak of cardiovascular diseases	SVM, KNN, NB, SVC with RBF kernel algorithm etc.	Data of Institute of Clinical Physiology of the National Research Council (Italy) and the National Institute of Diabetes and Digestive and Kidney Diseases repository (USA)	Supervised	Accuracy: 95.25%
[50]	Vaccine-derived poliovirus (VDPV)	2020	To predict vaccine-drived poliovirus outbreak	WOA-RVFL	Various UCI datasets	Supervised	Accuracy: 98.3%
[51]	Any unknown infectious diseases	2016	To predict any unknown infectious disease or outbreak	Sentiment analysis	Social media data	Unsupervised	F1-Score: 72%

H1N1 influenza virus is an important pandemic disease that started in 2009 and caused 284500 deaths worldwide [46]. Koike and Morimoto [47] used neural network to predict the expansion of the geographical range of non-indigenous species. According to their work, if the current distribution of species were known, their future distribution would also be known. They found that there was a correlation between the actual and predicted infestation dates. Liang et al. [48] used ANN, SVM, AdaBoost, C4.5, NB, RF with CfsSubset Evaluator-Best First feature selection method to construct an African swine fever outbreak prediction model. Experimental results showed that RF with CfsSubset Evaluator-Best First feature selection method was the best algorithm that reached highest accuracy.

Machine learning methods are also very successful in detecting the outbreak of cardiovascular diseases [49], vaccine-drived polyvirus outbreak [50] and any unknown infectious outbreak [51]. Mezzatesta et al. [49] used several algorithms to predict the outbreak of cardiovascular diseases. They used the data obtained from Data of Institute of Clinical Physiology of the National Research Council (Italy) and the National Institute of Diabetes and Digestive and Kidney Diseases (USA) repositories to test the model. Experimental results showed that non-linear SVC with RBF kernel algorithm, which was optimized with GridSearch, presented 65.25% accuracy. Hemedan et al. [50] proposed a random vector functional link (RVFL) networks and whale optimization algorithm (WOA) method to predict vaccine-derived polivirus outbreak. In the study, WOA was used to find the optimum parameters for configuration of the algorithm to increase the accuracy of the approach. The results showed that the approach can detect the vaccine-derived poliovirus surveillance in high accuracy and higher performance than traditional RVFL. Lim et al. [51] used an unsupervised sentiment analysis on Twitter data to predict any unknown infectious disease or outbreak. They collected users' expressions about symptoms and then created symptom weighting vector and time period. Finally, results about infectious disease or outbreak are retrieved by using the model. It is a successful approach based on unsupervised sentiment analysis using user, text and temporal information in social media data.

4.4. COVID-19 Outbreak Prediction

COVID-19 is an infectious disease that emerged on December, 2019 in Wuhan. The disease has been spreading ever since and has turned into a pandemic that affects millions of people worldwide. There exist many studies in the literature employing machine learning methods to detect COVID-19 spread. Liu et al. [52] proposed a method to predict COVID-19 for several parts of China by using machine learning. They gathered the data from Chinese Center Disease for Control and Prevention, COVID-19-related internet search activity from Baidu, news media activity reported by Media Cloud, and daily forecasts of COVID-19 activity from a special epidemic model. Then, they clustered data into 32 groups and trained the model for each group. They calculated correlation matrix to find similarity of outbreak patterns. Finally, they used LASSO multi-variable regularized linear model to make prediction. According to the results that given by the authors, the model can make prediction 2 days ahead of current time.

Kumar and Hembram [53] presented a statistical model to predict the spread rate of COVID-19 in China and Italy. They utilized the models of Logistic equation, Weibull equation, and the Hill equation. Data analysis was performed to find out the effect of environmental parameters like temperature, humidity and wind on spread of COVID-19. The results showed that there was a strong relation among COVID-19 spread, wind speed and humidity. However, contrary to popular belief, it was found that there was no relation between temperature and COVID-19 spread.

Silva et al. [54] used k-means, which is an unsupervised machine learning algorithm, to find the best number of clusters for COVID-19 pandemic and as a result they determined the number of clusters as four. Ardabili et al. [55] compared machine learning algorithms and soft computing models to determine which algorithm is better for predicting COVID-19 outbreak. According to the obtained results Multi Layered Perceptron (MLP) and Adaptive Network-based Fuzzy Inference System (ANFIS) were the algorithms that showed best results.

In another study, Fong et al. [56] tried to predict COVID-19 pandemic. They employed SVM, Fast Decision Tree, and LR and according to the experimental results their novel algorithm which is named polynomial neural network with corrective feedback (PNN+cf) was the algorithm that reached highest accuracy value.

Karadayi et al. [57] proposed a deep hybrid learning framework to detect anomalies in unlabelled data. As a case study, they used the data of the Italian Department of Civil Protection. According to the experimental studies, their model achieves significant improvements on unlabelled data even if being of data very noisy and high contamination ratio. Khakharia et al. [58] proposed a prediction model to forecast the COVID-19 pandemic by using various machine learning algorithms. They used data of 10 highly and densely populated countries like China, Germany, Philippines, Ethiopia etc. The best result was obtained as 99.93% by employing Auto-Regressive Moving Average (ARMA) on Ethiopia data.

Behnam and Jahanmahin [59] used machine learning approaches to predict the spread of the COVID-19 outbreak and they made prediction about end of it in Iran. The study aims to predict the short-term, medium-term and long-term spread of the outbreak. In the presented data analytics approach, it has been concluded that the best one among the proposed methods to detect the spread of the pandemic is the Gaussian functions curve. Among the recommendations made in the study, it is emphasized that transparent data is needed to accurately predict the

pandemic process and that making country-specific assessments gives healthier results.

Gatta et al. [60] proposed an Epidemiological Neural Network model to predict the spread of the COVID-19 pandemic and to analyze the effects of applied quarantine strategies on the spread of the pandemic. In the study, spatial and temporal features were examined by using graphic convolutional neural networks and long-short-term memory networks together. The proposed model was compared with the real pandemic data in Italy and the spread of the outbreak was analyzed with the Epidemiological Neural Network model. Tiwari et al. [61] used machine learning methods to analyze the future growth and effects of the COVID-19 pandemic. For this purpose, Naive-Bayes, SVM and Linear Regression algorithms were applied on a real-time dataset. In the study, the Naive Bayes method achieved more successful results than other methods.

Table 5. Comparison of the studies using machine learning techniques for COVID-19 outbreak detection

Work	Disease / Outbreak	Year	Aim(s)	ML Algorithm(s)	Data / Place	Type	Highest Result / Success
[52]	COVID-19	2020	To predict COVID-19 pandemic	Clustering based machine learning model	Several sources like, internet, news, China Centers for Disease Control and Prevention COVID-19 data etc.	Supervised	Predicts 2 days ahead of current time
[53]	COVID-19	2020	To predict spread rate of COVID-19	Logistic equation, Weibull equation and Hill equation	COVID-19 data of WHO that given on website for China and Italy	Statistical model	It is determined that data of infected people were fitted with Gaussian distribution with the peak at ~40 days for the countries.
[54]	COVID-19	2020	To find the best number of clusters for COVID-19	k-Means	World Health Organization COVID-19 Situation Reports	Unsupervised	Determination of suitable cluster numbers
[55]	COVID-19	2020	To find which one is better for predicting COVID-19: machine learning algorithms or soft computing models	Several ML algorithms like MLP, ANFIS, GA etc.	Italy, Germany, Iran, USA, and China COVID-19 data	Supervised	Accuracy: 99.9%
[56]	COVID-19	2020	To find the best machine learning approach for prediction of COVID-19	Polynomial neural network with corrective feedback (PNN+cf)	COVID-19 data of China and Italy	Supervised	PNN+cf provided minimum RMSE error compared to other approaches
[57]	COVID-19	2020	To detect anomalies in unlabelled data	Deep Learning	The Italian Department of Civil Protection	Unsupervised	Their model achieves significant improvements on unlabelled data
[58]	COVID-19	2020	To developed an outbreak prediction system for COVID-19	Several ML algorithms	10 populated countries like India, Pakistan, Indonesia etc.	Supervised	Accuracy: 99.93%
[59]	COVID-19	2021	To predict the spread and end of the COVID-19 outbreak	Several ML algorithms	Iran COVID-19 data	Unsupervised	Predicts peak of outbreak and end of outbreak
[61]	COVID-19	2021	To analyze the future growth and effects of the COVID-19 outbreak	Naive Bayes, Linear Regression, SVM	Worldwide COVID-19 data	Supervised	MAE: 488806.74

5. CHALLENGES IN OUTBREAK DETECTION AND IT'S FUTURE

While trying to improve a model for outbreak detection there are several problems to be overcome. One of these problems is parameter settings. In the machine learning area, this is the most common problem and there exist some approaches such as random search, grid search and Bayesian optimization to find the best parameters. The second challenge is the lack of proper data. In some cases, quality and quantity of the data is insufficient. The data may be in unstructured form and preprocessing operations may be requires. Another problem for this area is to track the people who infected or contacted. Because, information and data about such people affects the success of the model. In the light of these information, we can sum up some challenges that obstacle the success of machine learning based models to predict outbreaks as follows [21]:

- Tracking of infected/contacted people

- Longer incubation period causes to limited time for detection of outbreak
- Lack of proper data
- Overfitting of the data
- Plenty of data obstructs to select correct data
- Improper algorithm selection and size of the features (feature selection problem)
- Model complexity
- Specifying the timeline for outbreaks
- Detection of unknown outbreaks

In the outbreak prediction area, two approaches are commonly used. The first one is to predict the time of the emergence of any outbreak that had appeared before, and the other one is to detect any outbreak that had not detected before. With the impact of the COVID-19 pandemic, it is expected that more studies will be conducted on the monitoring and prediction of outbreaks in the future. Since it has the ability to extract valuable information from data, more studies based on machine learning are expected to be proposed.

6. CONCLUSION AND FUTURE WORKS

COVID-19 pandemic has been shown that prediction of outbreaks on the time can rescue many people's lives. Machine learning based approaches are the most important of works that try to detect outbreak in early stage. These algorithms are the most suitable approaches for predicting outbreaks, because of learning ability from the past. There are various machine learning algorithms that try to detect outbreak as presented in this work. However, it cannot be said that there is an approach satisfying all the needs in this area. Because, there are various problems and obstacles to be overcome like data procurement, preprocessing and isolation from unauthorized ones. Therefore, it can be said that there is a long way to go in this area. Besides, detection of an outbreak that unknown and never seen before is also an open area to research.

REFERENCES

[1] A. Şenol, H. Karacan, "A Survey on Data Stream Clustering Techniques", *European Journal of Science and Technology*, 13, 17-30, 2018.

[2] S. Messaoud, et al., "A Survey on Machine Learning in Internet of Things: Algorithms, Strategies and Applications", *Internet of Things*, 12, 2020.

[3] T. Meng, et al., "A Survey on Machine Learning for Data Fusion", *Information Fusion*, 57, 115-129, 2020.

[4] C. Chen, "A Hybrid Intelligent Model of Analyzing Clinical Breast Cancer Data Using Clustering Techniques with Feature Selection", *Applied Soft Computing*, 20, 4-14, 2014.

[5] J. Vamathevan, et al., "Applications of Machine Learning in Drug Discovery and Development", *Nature Reviews Drug Discovery*, 18(6), 463-477, 2019.

[6] C. Réda, E. Kaufmann, A. Delahaye-Duriez, "Machine Learning Applications in Drug Development", *Computational and Structural Biotechnology Journal*, 18, 241-252, 2020.

[7] M.T. Thai, et al., "Advanced Intelligent Systems for Surgical Robotics", *Advanced Intelligent Systems*, 2(8), 2020.

[8] M. Jahanbani Fard, et al., "Machine Learning Approach for Skill Evaluation in Robotic-Assisted Surgery", **Proceedings of the World Congress on Engineering and Computer Science**, San Fransisco, October 19-21, 2016.

[9] A. Smiti, "When Machine Learning Meets Medical World: Current Status and Future Challenges", *Computer Science Review*, 37, 2020.

[10] J. Friedman, T. Hastie, R. Tibshirani, "The Elements of Statistical Learning", *Springer Series in Statistics New York*, 1, 2001.

[11] G. James, **An Introduction to Statistical Learning**, Springer, 2013.

[12] P. Cunningham, M. Cord, S.J. Delany, **Supervised Learning, in Machine Learning Techniques for Multimedia**, Springer, 21-49, 2008.

[13] T. Uyar, K. Karaca Uyar, Emre Yağlı, "Gözetimli Makine Öğrenmesiyle Noktalama ve Etkisiz Kelime Sıklıkları Kullanarak Yazar Tanıma", *Bilişim Teknolojileri Dergisi*, 14(2), 183-190, 2021.

[14] A. Özgür, H. Erdem, "Saldırı Tespit Sistemlerinde Kullanılan Kolay Erişilen Makine Öğrenme Algoritmalarının Karşılaştırılması", *Bilişim Teknolojileri Dergisi*, 5(2), 41-48, 2012.

[15] H.M. Abbas, M.M. Fahmy, "Neural Networks for Maximum Likelihood Clustering", *Signal Process*, 36(1), 111-126, 1994.

[16] M.R. Ackermann, et al., "StreamKM++: A Clustering Algorithm for Data Streams", *Journal of Experimental Algorithmics*, 17, 1-30, 2012.

[17] L.P. Kaelbling, M.L. Littman, A.W. Moore, "Reinforcement Learning: A Survey", *Journal of Artificial Intelligence Research*, 4(1), 237–285, 1996.

[18] R. Nian, J. Liu, B. Huang, "A Review on Reinforcement Learning: Introduction and Applications in Industrial Process Control", *Computers & Chemical Engineering*, 139, 2020.

[19] R.S. Sutton, A.G. Barto, **Reinforcement Learning: An Introduction**, MIT Press, 2018.

[20] D. Grennan, "What Is a Pandemic?", *JAMA*, 321(9), 910, 2019.

[21] G. R. Shinde, et al., "Forecasting Models for Coronavirus Disease (COVID-19): A Survey of the State-of-the-Art", *SN Computer Science*, 1(4), 197, 2020.

[22] M.S. Smolinski, A.W. Crawley, J.M. Olsen, "Finding Outbreaks Faster", *Health Security*, 15(2), 215-220, 2017.

[23] T. Van-Dai, L. Chuan-Ming, G.W. Nkabinde, "Big data stream computing in healthcare real-time analytics", **IEEE International Conference on Cloud Computing and Big Data Analysis**, 37-42, Chengdu, China, 2016.

[24] İnternet: Outbreak Investigation, https://www.who.int/hac/techguidance/training/outbreak%20investigation_en.pdf, 07.08.2020.

[25] İnternet: Toward the Development of Disease Early Warning Systems, https://www.ncbi.nlm.nih.gov/books/NBK222241/, 03.12.2020.

[26] A. Abdeslam, F. El Bouanani, H. Ben-azza, "Four Parallel Decoding Schemas of Product Block Codes", *Transactions on Networks and Communications*, 2, 49-69, 2014.

[27] M. Ahmed, "Buffer-based Online Clustering for Evolving Data Stream", *Information Sciences*, 489, 113-135, 2019.

[28] M. Hahsler, M. Bolaños, "Clustering Data Streams Based on Shared Density between Micro-Clusters", *IEEE Transactions on Knowledge and Data Engineering*, 28(6), 1449-1461, 2016.

[29] N. Agarwal, et al., "Data Mining Techniques for Predicting Dengue Outbreak in Geospatial Domain Using Weather Parameters for New Delhi, India", *Current Science*, 114, 2281-2291, 2018.

[30] S.A., Balamurugan, M.S.M. Mallick, Chinthana, "Improved prediction of dengue outbreak using combinatorial feature selector and classifier based on entropy weighted score based optimal ranking", *Informatics in Medicine Unlocked*, 20, 2020.

[31] L. Tapak, et al., **Comparative Evaluation of Time Series Models for Predicting Influenza Outbreaks: Application of Influenza-Like Illness Data from Sentinel Sites of Healthcare Centers in Iran.** BMC Research Notes, 2019.

[32] İnternet: Malaria, https://www.who.int/news-room/fact-sheets/detail/malaria, 30.01.2021.

[33] V. Sharma, "Malaria Outbreak Prediction Model Using Machine Learning", *International Journal of Advanced Research in Computer Engineering & Technology*, 9(3), 99-102, 2016.

[34] B. Modu, et al., "Towards a Predictive Analytics-Based Intelligent Malaria Outbreak Warning System", *Applied Sciences*, 7(8), 2017.

[35] G. Comert, N. Begashaw, A. Turhan-Comert, "Malaria Outbreak Detection with Machine Learning Methods", *Biorxiv*, 2020.

[36] İnternet: Dengue and severe dengue, https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue, 29.01.2021.

[37] G. Zhu, J. Hunter, Y. Jiang, "Improved Prediction of Dengue Outbreak Using the Delay Permutation Entropy", **IEEE International Conference on Internet of Things and IEEE Green Computing and Communications and IEEE Cyber, Physical and Social Computing and IEEE Smart Data**, 828-832, Chengdu, China, 2016.

[38] N. Iqbal, M. Islam, "Machine Learning for Dengue Outbreak Prediction: A Performance Evaluation of Different Prominent Classifiers", *Informatica*, 43, 363-371, 2019.

[39] S. Anno, et al., "Spatiotemporal Dengue Fever Hotspots Associated with Climatic Factors in Taiwan Including Outbreak Predictions Based on Machine-Learning", *Geospatial Health*, 14(2), 2019.

[40] R. Dhesi Baha, et al., "Artificial Intelligence Model as Predictor for Dengue Outbreaks", *Malaysian Journal of Public Health Medicine*, 19(2), 2019.

[41] S. Amin, et al., "Detecting Dengue/Flu Infections Based on Tweets Using LSTM and Word Embedding", *IEEE Access*, 8, 189054-189068, 2020.

[42] C.M. Benedum, et al., "Weekly Dengue Forecasts in Iquitos, Peru; San Juan, Puerto Rico; and Singapore", *PLOS Neglected Tropical Diseases*, 14(10), 2020.

[43] A. Sadilek, et al., "Machine-Learned Epidemiology: Real-Time Detection of Foodborne Illness at Scale", *NPI Digital Medicine*, 1, 2018.

[44] S. S. Chenar, Z. Deng, "Development of Artificial Intelligence Approach to Forecasting Oyster Norovirus Outbreaks Along Gulf of Mexico Coast", *Environment International*, 111, 212-223, 2018.

[45] S. S. Chenar, Z. Deng, "Development of Genetic Programming-Based Model for Predicting Oyster Norovirus Outbreak Risks", *Water Research*, 128, 20-37, 2018.

[46] F. S. Dawood, et al., "Estimated Global Mortality Associated with the First 12 Months of 2009 Pandemic Influenza a H1N1 Virus Circulation: A Modelling Study", The Lancet infectious diseases, 12(9), 687-695, 2012.

[47] F. Koike, N. Morimoto, "Supervised Forecasting of the Range Expansion of Novel Non-Indigenous Organisms: Alien Pest Organisms and the 2009 H1N1 Flu Pandemic", *Global Ecology and Biogeography*, 27(8), 991-1000, 2018.

[48] R. Liang, et al., "Prediction for Global African Swine Fever Outbreaks Based On A Combination of Random Forest Algorithms and Meteorological Data", *Transboundary and Emerging Diseases*, 67, 935-946, 2019.

[49] S. Mezzatesta, et al., "A Machine Learning-Based Approach for Predicting the Outbreak of Cardiovascular Diseases in Patients on Dialysis", *Computer Methods and Programs in Biomedicine*, 177, 9-15, 2019.

[50] A. A. Hemedan, et al., "Prediction of the Vaccine-derived Poliovirus Outbreak Incidence: A Hybrid Machine Learning Approach", *Scientific Reports*, 10(1), 5058, 2020.

[51] S. Lim, C.S. Tucker, S. Kumara, "An Unsupervised Machine Learning Model For Discovering Latent Infectious Diseases Using Social Media Data", *Journal of Biomedical Informatics*, 66, 82-94, 2017.

[52] D. Liu, "A Machine Learning Methodology for Real-Time Forecasting of the 2019-2020 COVID-19 Outbreak Using Internet Searches, News Alerts and Estimates From Mechanistic Models", *Arxiv*, 1-23, 2020.

[53] J. Kumar, K. Hembram, "Epidemiological Study of Novel Coronavirus (COVID-19)", *Arxiv*, 1-9, 2020.

[54] R. Fray da Silva, et al., "Unsupervised Machine Learning and Pandemics Spread: The Case of COVID-19", **SBCAS**, 2020.

[55] S.F. Ardabili, et al., "COVID-19 Outbreak Prediction with Machine Learning", *Algorithms*, 13(10), 1-36, 2020.

[56] S. Fong, et al., "Finding an Accurate Early Forecasting Model from Small Dataset: A Case of 2019-nCoV Novel Coronavirus Outbreak", *International Journal of Interactive Multimedia and Artificial Intelligence*, 6, 132-140, 2020.

[57] Y. Karadayı, M.N. Aydin, A. Selçuk, "Unsupervised Anomaly Detection in Multivariate Spatio-Temporal Data Using Deep Learning: Early Detection of COVID-19 Outbreak in Italy", *IEEE Access*, 8, 164155-164177, 2020.

[58] A. Khakharia, et al., "Outbreak Prediction of COVID-19 for Dense and Populated Countries Using Machine Learning", *Annals of Data Science*, 8, 1-19, 2020.

[59] A. Behnam, R. Jahanmahin, "A data analytics approach for COVID-19 spread and end prediction (with a case study in Iran)", *Modelling Earth Systems and Environment*, 1-11, 2021.

[60] V. La Gatta, et al., "An Epidemiological Neural network exploiting Dynamic Graph Structured Data Applied to the COVID-19 Outbreak", *IEEE Transactions on Big Data*, 7(1), 45-55, 2020.

[61] D. Tiwari, et al., "Pandemic Coronavirus Disease (Covid-19): World Effects Analysis And Prediction Using Machine-Learning Techniques", *Expert Systems*, 1-20, 2021.