Journal of Applied Statistics & Machine Learning
Vol. 3, Nos. 1-2, 2024, pp. 63-83, ISSN: 2583-2891

URL : www.blackroseindia.com/jasml

Prediction of Epidemics using Machine Learning Models: A Review

Puneet Misra[1], Sandesh Paul[2] and Shambhavi Mishra[3*]
[1]Assistant Professor, Department of Computer Science, University of Lucknow, Lucknow, Uttar Pradesh, India
[2]Research Scholar, Department of Statistics, University of Lucknow, Lucknow, Uttar Pradesh, India
[3]Assistant Professor, Department of Statistics, University of Lucknow, Lucknow, Uttar Pradesh, India
** Corresponding author: shambhavimishra.lko@gmail.com*

Received : 23 October 2024 • **Revised :** 22 November 2024;
Accepted : 10 December 2024 • **Published :** 29 December 2024

Abstract: The COVID-19 pandemic which began as an epidemic demonstrates the need for a better system and technologies to monitor and identify health emergencies before they become a disaster that claims human lives and damages the economy. The twenty-first century is known for technological advancements, particularly in the fields of Artificial Intelligence (AI) and Machine Learning (ML). A number of AI and ML-based algorithms and models have been developed to date for surveillance and precision decision-making in the healthcare domain. Machine Learning, as represented by Supervised Machine Learning, Unsupervised Machine Learning, Reinforcement Learning, and Semi-Supervised Machine Learning algorithms, has made considerable progress in the field of health care, However, there is still room for advancement. The purpose of this review paper is to identify the models developed for epidemic assessment and prediction, and simultaneously identify which areas of the healthcare system need improvement and how Machine Learning models can help.

Keywords: Pandemic, Epidemic, Artificial Intelligence, Machine Learning, Surveillance and Decision-making.

1. Introduction

Epidemics have been haunting human beings for centuries. In the past few decades, the outbreak of several diseases like the Black Death, Spanish flu, smallpox, Ebola, influenza, etc., and most recently Covid-19 aroused worldwide concern. A major public health issue is the prediction and analysis of epidemiological data. Several variables [1] are required to predict a future outbreak, apply preventative measures, and track disease outbreak progress. The health sector has recently received more attention

TO CITE THIS ARTICLE

Misra, P., Paul, S., & Mishra, S. (2024). Prediction of Epidemics using Machine Learning Models: A Review, *Journal of Applied Statistics & Machine Learning,* 3(1-2), pp. 63-83.

from machine-learning (ML) and Artificial Intelligence (AI) communities.AI [2] has numerous advantages, including flexibility, adaptability, pattern recognition, and quick computation and learning capabilities. AI aims [3] to create systems that replicate human behaviours, while ML allows systems to learn from fresh data without explicitly being programmed. Machine algorithms allow one to improve the predictive analytic accuracy for a certain task and develop new skills over time.

Electronic medical record (EMR) systems serve as the major sources of information in the twenty-first century. EMR was supposed to make clinical decision-making more efficient. Unfortunately, the digitization of medical data has resulted in "information overload" for healthcare practitioners. Because computers [4] can handle a broader range of variables, using predictive analytics via artificial intelligence (AI)/machine learning (ML) could improve our ability to discover clinically important patterns, such as those for epidemic diseases. Despite the significant performance of AI-based healthcare ML systems, researchers have focused in recent years on the interpretability, explainability, and trustworthiness of AI/ML. [5]

2. Objective

The purpose of this work is to illustrate epidemic prediction methods in the domain of healthcare. The study seeks to identify the need for a new machine-learning model capable of making accurate and dynamic predictions.

3. Techniques/Methods

Machine learning, as a topic of research, stands at the intersections of computer science, statistics, and a range of other disciplines concerned with automatic improvement over time, as well as inference and decision-making under uncertainty[6]**.** It is also known as predictive learning or statistical learning. Machine learning has tremendously influenced the way data-driven research is done today[7]. Machine learning is all about creating algorithms that allow the computer to learn. Learning is a process of finding statistical regularities or other patterns of data[8]. Several machine learning models have been developed to analyse the data complexity and extract useful information from the data. With time, certain modifications and rectifications are being done to improve the ability of models to learn the hidden pattern in the dataset, recognition of voice, and sentiment of social media posts. In machine learning [9, 10], representations and generalizations are used.

3.1. Components of Machine Learning Models

3.1.1. Supervised Machine Learning

A labelled training dataset is used in supervised machine learning algorithms to train the underlying algorithm first. This trained algorithm is then used to classify the unlabelled test dataset into similar groups. Supervised learning can be further categorized into two parts: Regression and Classification where the former can be used for prediction purposes and the latter to classify the output variables into two or more categories. The primary distinction between them is the outcome variable i.e. in classification the output variable is discrete whereas in the regression it is a real value or continuous number. Classification further can be classified into Binary classification and multiple classifications. In Binary classification, the output is categorised into two classes, like; yes/no, on the other hand, in multiple-class classification, the output is categorised into more than two classes.[7, 11]

3.1.2. Unsupervised machine learning

Unlike supervised machine learning, in unsupervised machine learning, there is no prior exercise to train the machine algorithm. Machine algorithms are just fed with unlabelled datasets based on their learning and experience in cluster groups. Since the machine is not familiar with the dataset, the output is also unknown to the data instructor. There are two important components of unsupervised learning,[7] transformation of dataset and clustering. Unsupervised dataset transformations are algorithms that produce a different representation of the data that can be easier to understand for humans or other machine learning algorithms than the original representation of the data. In contrast, clustering algorithms distribute data into distinct groups of similar features. For instance, google photos allow you to organize your pictures that might look like the same person

3.1.3. Reinforcement machine learning

Reinforcement machine learning follows a trial-and-error method to learn, which is based on the Rewards function. Rather than producing one output for one input, machine algorithms produce an output with an incentive or reward which helps algorithms to learn how a human does. Like in an online chess game, algorithms learn from past actions.

Reinforcement machine [11] uses concepts from dynamical systems theory, specifically known as Markov decision processes. Markov decision processes are designed to incorporate only these three aspects—sensation, action, and goal—in their most basic forms, without underplaying any of them. Any approach that is well suited to solving such problems is considered a reinforcement learning method. Reinforcement learning possesses traits that are different from other machine learning algorithms, which is the trade-off between exploitation and exploration. To obtain the reward, the agent must exploit what it has already experienced, but it must also explore to make better future action selections.

3.1.4. Semi-supervised machine learning

Semi-supervised machine learning try to overcome the disadvantages of both supervised and unsupervised machine learning algorithms. Supervised machine learning needs a large amount of label data set which is costly and time-consuming whereas unsupervised makes clusters based on similarities in the data set which might be not effective and precise [12]. Thus, it is the amalgamation of supervised and unsupervised machine learning in which firstly, algorithms are used to train with mixed datasets i.e. labelled and unlabelled datasets. This mixed dataset comprises little labelled data and a significant amount of unlabelled data. One of the prominent examples of semi-supervised is a speech analysis.

4. Examples of some prominent machine learning models

4.1. Logistic Regression

Logistic Regression is an essential component of supervised machine learning. It is useful for describing the relationship between one or more explanatory variables (say,x) and the dependent variable(say, $\phi(x)$). Generally, the relationship between x and $\phi(x)$ is non-linear in logistic regression and practice, it has been seen that the value of $\phi(x)$ either increases or decreases continuously concerning the value of x. The most widely used basic form of logistic regression is given by-

$$\phi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} \quad \text{(i)}$$

where α and β are unknown parameters and beta reflects the rate of change in the curve. These unknown parameter in (i) [13], are estimated by a numerical implementation of the Maximum Likelihood estimation method. The outcome variable in logistic regression is binary or dichotomous, and the conditional distribution of the outcome variable follows the binomial distribution.

4.2. Naïve Bayes

Naive Bayes is a simple probabilistic classifier that predicts the class that optimizes the posterior probability using the Bayes theorem. The naive Bayes classifier is based on the notion that attributes are independent. Naive Bayes [7], models are extremely effective because they learn parameters by examining each feature separately and gathering straightforward per-class statistics from each feature.

The primary task in using [14], the Naïve Bayes classifier is to estimate the joint probability density function for each class, which is accomplished through multivariate normal distribution. If a training dataset comprises n points x_i ,i=1,2,..,n in a d-dimensional space, and y_i represents the class c_i,i=1,2,..,n for each point. It evaluates the posterior probability $P(c_i|x)$ for each class c_i and picks the class with the highest probability. The predicted class is given as

$$\hat{y}=\max[P(c_i \mid x)]$$

4.3. Support Vector Machine

Support vector machine is a supervised machine learning based on the maximum margin linear discriminants in which the objective is to determine the optimal hyperplane that maximizes the margin between the classes[14]. What distinguishes SVM from other machine learning methods is that it can use nonlinear relationships to map points to other dimensions and thus classify points that are not linearly separable[15]. Since training with large-scale datasets can become complicated and time-consuming, SVM is the best fit for small datasets and outperforms many algorithms in terms of accuracy. SVM uses a subset of training points to enhance classification efficiency. It will not perform well if the dataset contains noise. SVM can solve both linear and nonlinear problems, but nonlinear SVM is recommended over linear SVM because of its superior efficacy[16].

4.4. Decision Tree

Decision Tree widely used for classification and regression. A decision tree classifier is a recursive, partition-based tree model and this recursive technique yields a binary tree of decisions, each node of which contains questions. Each node in the decision tree denotes either a question or a terminal node (i.e., leaf) containing the answer. This is like building a hierarchical partition. The data is recursively partitioned until each leaf in the decision tree contains only one target value. Decision trees also can be used for regression problems. To make a prediction, we explore the Decision tree based on the tests in each node and determine the leaf into which the new data point falls. The prominent problem of the decision tree is overfitting. There are two methods to prevent overfitting: pre-pruning and post-pruning[7] [14].

4.5. Artificial Neural Networks

Artificial Neural Networks are a collection of machine learning algorithms which is based on the complex nature of human brains, which includes billions of interconnected neurons that process information simultaneously. An artificial neural network is made up of three layers: an input layer (or nodes), hidden layers (one or two or even three), and a final layer of output neurons. Each connection has attributed a weight, which is a numerical value[17]. Multiple approaches for classification and prediction utilising artificial neural network methods have been proposed in the last two decades. Their ability to execute mapping is one of the reasons for their extensive use[18].

5. Review of Epidemics Prediction Models as Per diseases

5.1. Coronavirus disease (Covid -19)

Shuo Feng et al. [19] applied the SEIR (Susceptible-Exposed-Infected-Removed) model and AI model to analyse the epidemic trend in Wuhan(epicentre) and non-Wuhan(non-epicentre) areas respectively. Here AI model consists of DNN (Deep Neural Network) and RNN (Recurrent Neural Network). Factor (or parameters) that are taken here is epidemiological data of covid-19, migration index, population density, per capita GDP, the distance of each province from Wuhan and average temperature. They found that the SEIR model and AI model accurately anticipated the infection value of the epidemic condition in Wuhan and non-Wuhan provinces. Also mentioned

the reason behind choosing DNN and RNN over traditional methods of prediction because of their accuracy.

Another group of researchers Smita Rath et al.[20] uses linear regression and multiple linear regression to predict covid-19 cases. R-square values for linear regression and multiple regression were found to be 0.99 and 1, respectively. It demonstrates a good prediction model for forecasting new covid-19 instances and these models achieved outstanding accuracy.

In a similar course, Parul Arora et al. [21] implemented Deep Learning models, Recurrent neural networks based on Long short-term memory (LSTM) for the prediction of covid cases and analysis. Deep LSTM, convolutional LSTM, and bi-directional LSTM are tested on 32 states/unions, and the model with the highest accuracy is selected. They found that bi-directional LSTM achieved remarkable accuracy with an absolute error of 0.03 which outperformed all other prediction models. Convolutional LSTM yields the worst results.

Michal Wieczorek et al. [22] developed a seven layers Artificial Neural network that Nesterov-accelerated Adaptive Moment Estimation (NADAM) trained. They developed a model with a unified architecture that does not require changes for different countries and regions. In comparison with RNN, ANN performed 3% better than the RNN. One predictor model performs with very high accuracy for the majority of the region, which is around 87.70%.

Saleh I. ALzahrani et al.[23] applied four different prediction models namely, Autoregressive(AR), Moving average(MA), Autoregressive Moving average(ARMA) and Autoregressive Integrated Moving average(ARIMA) to forecast the number of confirmed covid cases in Saudi Arabia over the next four weeks. ARIMA outperformed other models, with an R-square of 0.99 and a Mean Absolute Error (MAE) of 17.93%. MA exhibits the worst efficiency, with an R-square of 0.46.

Ammar H. Elsheikh et al. [24] used a Deep Learning forecasting model to predict the outbreak of covid. Long short-term memory (LSTM) network was proposed as a deep learning model in their study. Further compared this model with Autoregressive Integrated moving average (ARIMA) and Nonlinear autoregressive artificial neural network (NARANN). The LSTM model outperforms ARIMA and NARANN in forecasting the prevalence of the outbreak. The RMSE of the forecasted data using LSTM was less than 11 and 28% of that of ARIMA and NARANN, respectively, indicating that the proposed method significantly outperforms other tested statistical and AI-based methods.

In other research work, a group of researchers [25] presented a new deep-learning approach: COVIDX-Net. The model comprises seven deep convolutional neural network architectures, including the modified Visual Geometry Group Network (VGG19), The Dense Convolutional Network (DenseNet121), the Inception network (InceptionV3), Residual Neural Network (ResNet), ResNetV2, Inception-ResNet-V2, Xception and MobileNetV2 model. This model detects COVID-19 in X-ray images automatically. Their research included 50 Chest X-ray images of which twenty-five were confirmed positive for COVID-19 cases. The proposed COVIDX-Net model confirmed that the best deep learning classifier performance scores are for the VGG19 and DenseNet121, with an accuracy of 90% each.

Mucahid Barstugan et al.[26] attempted to predict covid cases using Computed Tomography (CT) pictures in a similar situation as described above. Four different datasets were created for COVID-19 detection by extracting patches of varying sizes from 150 CT pictures. To improve classification accuracy, they use to feature an extraction process that was applied to patches. Further, the extracted features were classified using Support Vector Machines (SVMs). During the classification process, they used 2-fold, 5-fold, and 10-fold cross-validations. Researchers obtained the best classification accuracy of 99.68% using 10-fold cross-validation and the Grey-Level Size Zone Matrix (GLSZM) feature extraction method.

5.2. Influenza

Armin Spreco et al.[27] developed an integrated method for the detection and prediction of influenza. One's integrated detection and prediction method was among the first to be developed in naturally occurring local influenza epidemics. For detection, exponential regression is used, and linear regression is used for prediction. The performance evaluation is based on retrospective data. They discovered that the prediction module performed admirably in terms of peak activity timing and intensity.

J. Zhang and K. Nawata [28], In their study, used four distinct multi-step prediction algorithms: Multi-stage prediction (MSP), Adjusted multi-stage prediction (AMSP), Multiple single-output predictions (MSOP), and Multiple-output prediction (MOP) in the long short-term memory (LSTM). The results demonstrated that implementing multiple single-output predictions (MSOP) in a six-layer LSTM structure yielded the highest accuracy. To their knowledge, this is the first time LSTM has been used and improved for multi-step-ahead influenza outbreak prediction.

Gisele H.B. Miranadaet al.[29], introduce a method for predicting the weekly occurrence of influenza-like illness (ILI) in real-time using a dynamically calibrated compartmental SIR (Susceptible, Infected, Removal or Recovered) model. The findings show that the suggested method can be used to portray the overall behaviour of epidemics.

In other research work, Rui Yin et al.[30] proposed a weighted ensemble convolutional neural network (CNN) for predicting influenza virulence, named: VirPreNet. As the base model, they used ensemble CNN models. Their findings indicate that VirPreNet improves performance by aggregating each base model's prediction.

Studies performed by, Taichi Murayama et al.[31] used a Robust two-stage influenza prediction model. They used an autoregressive model in the initial stage and an LSTM model in the second phase. The first model forecasts future Influenza Like illness (ILI) rates/patient numbers based on historical data, while the second model forecasts sudden outbreaks using user-generated data(UGC). The study was conducted in two countries, the United States and Japan. Their proposed model achieved an accuracy of.935 and.914 for the respective country data and thus findings indicate that the proposed model is the best for seasonal flu prediction.

Again Taichi Murayama et al.[32] proposed a method for predicting the geographical distribution of influenza patients by using commuting data that utilises a graph convolutional network extension (GCN) model. Compare this proposed model with Vector autoregression (VAR), LSTM, and CNNRNN-Res. Ones GCN-based model performed better than other models.

5.3. Malaria

Godson Kalipe et al. [33] used a variety of machine learning and deep learning models for the Malaria Outbreak and analysis, including KNN, Random Forest, SVM, XGboost, ANN, and Naive Bayes. A comparison of these models was also presented. Six years of data from various health centres were used for the research. The accuracy, precision, error rate, recall, and Matthews correlation coefficient were used to evaluate the performance of these models. For this particular use case, XGBoost outperforms all models in terms of accuracy 96.26%, recall 93.89%, and precision 91.82%.

Deep learning-based smartphone application [34] implemented for malaria parasite detection in thick smear images. A dataset of 1819 images from 150 patients was used for this purpose. To create parasite candidates, an intensity-based Iterative

Global Minimum Screening (IGMS) initially performs a quick screening of an entire thick smear image. Each candidate is then classified as a parasite or non-parasite by a customised CNN model. The customised CNN model yields an AUC score of 98.39% on average and a standard deviation of 0.18%, which shows its robustness and efficacy. The average accuracy of the customised CNN model is 93.46%, the specificity is 94.33%, the F-score is 93.40%, the sensitivity is 92.59%, and the precision is 94.25%

Pallavi Mohapatra et al.[35] performed a comparative analysis to determine the best model among malaria prediction accuracy methodologies in various climate conditions of Odisha State. The Waikato Environment for Knowledge Analysis (WEKA) is used in this objective, a collection of machine learning algorithms. WEKA was used in conjunction with two classifier techniques: MLP and J48. The J48 cross-validation approach performs better, but MLP performs even worse in forecasting malaria occurrences.

Matheus Félix Xavier Barboza et al.[36] proposed machine learning and deep learning model to estimate the prevalence of malaria cases in the states of Amazon. They used and compared random forest, long short-term memory (LSTM), and gated recurrent unit (GRU) models and their finding indicates that the LSTM design performed better in clusters with less variability in the number of cases, but the GRU performs better in clusters with high variability.

In the same course David Harvey et al.[37] introduced the first data-driven malaria epidemic early warning system capable of forecasting the 13-week case rate in a primary care facility. They train a combination of Gaussian Processes and Random Forest Regressors on the extraordinarily high-fidelity data of infant visits to estimate the weekly number of malaria cases over 13 weeks. Discovered that when it comes to raising an alert, the algorithm has 30% precision and more than 99% recall. For the high alert level, this jumps to more than 99% precision and 5% recall.

Eric Kamana et al.[38] first time using the LSTM sequence to sequence (LSTMseq2seq) model to study the impact of climate change on the re-emergence of malaria cases. Based on the influence of climatic conditions, the introduced LSTMSeq2Seq model considerably improved the prediction of malaria re-emergence. The LSTMSeq2Seq model has achieved a prediction accuracy of 87.3% on average.

5.4. Other epidemics

Satya Ganesh Kakarla et al. [39] used a weather-integrated multiple machine learning model to predict Dengue incidences. They performed their research using

Epidemiological and Meteorological data from 2003 to 2007 with statistical, machine learning and deep learning models. LSTM, which is a deep learning model, achieved RMSE =0.345 and R-square=0.9, the best among all other methods of prediction.

In another research work, Sumiko Anno et al.[40] developed an Early Warning System (EWS) for dengue spatiotemporal dengue fever hotspots based on climate. They used a machine learning algorithm to look for parameters that have a spatiotemporal link with dengue disease outbreaks. The machine learning model is built on a deep AlexNEt model that was trained via transfer learning and achieved 100% accuracy on an 8-fold cross-validation test dataset.

Van-Hau Nguyen et al. [41] designed deep learning models for dengue fever prediction using lagged Dengue Fever incidence and meteorological variables. They carried out their research using time series data. They applied attention-enhanced LSTM (LSTM-ATT), CNN, Transformer, and LSTM. Also compared it to the conventional machine learning model. Their findings suggest that LSTM-ATT was effective in predicting Dengue Fever occurrences.

Similarly, Samrat Kumar Dey et al.[42] attempted to forecast dengue illness in 11 districts of Bangladesh using medical records, socio-economic and metrological data, and machine learning algorithms. Machine learning models, Multiple Linear Regression (MLR), and Support Vector Regression (SVR) were used. MLR and SVR achieved 67% and 75% accuracy, respectively.

Qanita Bani Baker et al. [43] used sentiment analysis of Arabic tweets to anticipate epidemics such as influenza. For such purpose, they applied machine learning models like naïve Bayes, support vector machine, Decision trees and K-nearest neighbour. The results show that among these three models, Naïve Bayes and K nearest neighbour performed well with an accuracy of 89.06% and 86.43% respectively.

Prediction of hand, foot, and mouth disease epidemics in Japan, Kazuhiro Yoshida et al.[44] employed LSTM which is called RNN. The LSTM model was trained on weekly hand foot and mouth disease data. The output of research shows that LSTM can predict the future epidemic patterns of hand foot and mouth disease.

A table has been produced about the models used for disease outbreak

Table 1: Summary of Reviewed papers

Sr. No	*Epidemics*	*Objective*	*Data Sources*	*Algorithms*	*Performance/ Findings*	*Year*	*References*
1.	Covid-19	Prediction and analyse the epidemic tend	epidemiological data of COVID-19, Baidu migration project, population density, per capita GDP	SEIR model And Deep Neural Networks (DNN) and Recurrent neural networks (RNN)	Models were effective in predictive covid-19 cases	2020	[19]
2.	Covid-19	Prediction and analysis of covid -19 cases	Time series data of covid-19 from the Ministry of Health and Family Welfare of India	RNN based Deep Long-short-term memory (LSTM), Convolutional LSTM, Bidirectional LSTM	Bidirectional LSTM achieved remarkable accuracy with an absolute error < 0.03	2020	[21]
3.	Covid-19	Develop a model for covid-19 forecasting.	Dataset provided by Johns Hopkins University	Seven layers ANN, RNN	ANN outperformed RNN, with accuracy= 87.70%	2020	[22]
4.	Covid-19	Forecasting outbreak of covid-19	Data provided by ministry of Saudi Arabia	Deep learning model LSTM, Autoregressive Integrated moving average (ARIMA), Nonlinear autoregressive artificial neural network (NARANN)	The proposed deep learning model outperformed others' forecasting model	2020	[24]
5.	covid-19	To assist radiologists to automatically diagnose COVID-19 in X-ray images.	Fifty chest X-ray images including 25 confirmed positive COVID-19 cases	COVIDX-Net: Seven different architectures of deep convolutional neural network models-(VGG19), (DenseNet121), (InceptionV3), (ResNet),ResNetV2,Inception-ResNet-V2, Xception and MobileNetV2 model.	Best deep learning classifier is VGG19 and DenseNet121, with an accuracy of 90% each	2021	[25]

Sr. No	*Epidemics*	*Objective*	*Data Sources*	*Algorithms*	*Performance/ Findings*	*Year*	*References*
6.	covid-19	applied feature extraction process to computed tomography (CT) images and further used SVM to classification.	The data set consists of 150 CT abdomen pictures from the fifty-three affected patients.	Support Vector Machine (SVM), feature extraction process: Grey Level Co-occurrence Matrix (GLCM), Local Directional Pattern (LDP), Grey Level Run Length Matrix (GLRLM), Grey-Level Size Zone Matrix (GLSZM), and Discrete Wavelet Transform (DWT)	SVM accuracy=99.68% with Grey-Level Size Zone Matrix (GLSZM) feature extraction method.	2021	[26]
7.	Influenza	Multi-step-ahead time series prediction for Influenza outbreak	US flu data from the 40th week of 2002 to the 30th week of 2017, collected from Portal of the Centre for Disease Control and Prevention (CDC)	Four different multi-step LSTM prediction algorithms: Multi-stage prediction (MSP), Adjusted multi-stage prediction (AMSP), Multiple single-output predictions (MSOP), and Multiple-output prediction (MOP)	MSOP in a six-layer LSTM structure yielded the highest accuracy	2017	[28]
8.	Influenza	predicting the weekly occurrence of influenza-like illness (ILI) in real-time	weekly influenza-like illness data in Belgium throughout the seasons 2010-2011 to 2015-2016	SIR (Susceptible, Infected, Removal or Recovered) model	suggested model performed well	2019	[29]
9.	Influenza	prediction of influenza A virus	not mentioned	VirPreNet:-weighted ensemble convolutional neural network (CNN)	VirPreNet improves performance	2020	[30]

Sr. No	*Epidemics*	*Objective*	*Data Sources*	*Algorithms*	*Performance/ Findings*	*Year*	*References*
10.	Influenza	Two-stage influenza model: robust The first stage uses AR and at second stage uses LSTM for the prediction	Centre for Disease Control and Prevention (CDC) , US. Infectious Disease Weekly Report (IDWR), Japan. Google Trend (GT) data	Auto Regressive, LSTSM model	accuracy for US=.935 and Japan= .914	2019	[31]
11.	Influenza	using GCN model and commuting data to predict regional influenza	National Institute of Infectious Diseases (NIID)	graph convolutional network extension (GCN) model, Vector autoregression (VAR), LSTM, and CNN-RNN-Res.	proposed GCN model outperformed all other models	2020	[32]
12.	Malaria	utilizing machine learning and deep learning to predict epidemics	National Vector Borne Disease Control Program, Indian meteorological Centre, and Cyclone Warning Centre, India	KNN, Random Forest, SVM, XGboost, ANN, and Naive Bayes	XGBoost outperforms all models with accuracy of 96.26%, recall of 93.89%, and precision of 91.82%.	2018	[33]
13.	Malaria	deep learning based mobile application for detection of malaria parasite	not mentioned	customised CNN model	accuracy= 93.46%, specificity= 94.33%, F-score = 93.40%, sensitivity 92.59%, precision =94.25%	2020	[34]
14.	Malaria	finding a suitable machine learning model for malaria prediction	National Vector Borne Disease Control Programme, India	Multi-Layer prediction (MLP) model and J48 in WEKA	J48 performs better	2021	[35]

Sr. No	*Epidemics*	*Objective*	*Data Sources*	*Algorithms*	*Performance/ Findings*	*Year*	*References*
15.	Malaria	Prediction of malaria using deep learning models	Sistema de Informação de Vigilância Epidemiológica de Malária (SIVEP-MALARIA), Brazil	random forest, long short-term memory (LSTM), and gated recurrent unit (GRU)	the LSTM design performed better in clusters with less variability in the number of cases, but the GRU performs better in clusters with high variability	2021	[36]
16.	Malaria	develop early warning system for malaria	IeDA database	combination of Gaussian Processes and Random Forest Regressors	when it comes to raising an alert, the algorithm has 30% precision and more than 99% recall. For the high alert level, this jumps to more than 99% precision and 5% recall.	2021	[37]
17.	Malaria	Predicting the impact of climate change on the re-emergence of malaria cases	collected monthly malaria cases in all thirty-one provinces in China from January 2004 to December 2016	LSTM sequence to sequence (LSTMseq2seq) model	accuracy of 87.3% on average	2022	[38]
18.	Dengue	weather -integrated multiple machine learning model for Dengue prediction	Integrated Disease Surveillance Programme (IDSP), Directorate of Health Services, Kerala. Indian Meteorological Department (IMD), Pune and National Oceanic and Atmospheric Administration (NOAA), USA	vector auto regression (VAR), support vector regression (SVR), generalized boosted regression (GBM), and long short-term memory (LSTM)	LSTM best among all other model, achieved RMSE =0.345 and R-square=0.9	2022	[39]

Sr. No	*Epidemics*	*Objective*	*Data Sources*	*Algorithms*	*Performance/ Findings*	*Year*	*References*
19.	Dengue	developed an early warning system for Dengue Fever	Taiwan CDC 2016, Japan JAXA, NOAA, USA.	CNNs	accuracy=100%	2019	[40]
20.	Dengue	forecasting dengue fever based on climate data	National Institute of Hygiene and Epidemiology (NIHE), Vietnam	attention-enhanced LSTM (LSTM-ATT), CNN, Transformer, and LSTM	LSTM-ATT was effective in predicting Dengue Fever	2021	[41]
21.	Dengue	Prediction of dengue incidents using hospitalized patients, metrological and socio-economic data	www.bmd.gov.bd, www.bbs.gov.bd	Multiple Linear Regulation (MLR), and Support Vector Regression (SVR)	accuracy-MLR = 67% and SVR=75%	2022	[42]
22.	hand, foot, and mouth	Prediction of hand, foot, and mouth disease epidemics	(https://www.niid.go.jp/niid/en/idwr-e.html	LSTM	LSTM can predict the future epidemic patterns	2022	[44]

5. Discussion

AI/ML applications are becoming more prevalent in the health sector, and they and emerging daily. Researchers are developing several models for tracking, monitoring, and forecasting epidemics. In a research, Michal Wieczorek et al. [22] developed a seven-layer ANN to forecast the covid-19 epidemic. Shuo Feng et al. [19] applied SIER and AI models for covid-19 prediction stating that AI model outperformed the traditional model. For the same epidemic, a study performed by Ammar H. Elsheikh et al. [24],claimed that LSTM, based on deep learning performed better than AI models. However [25] developed a deep learning model, COVIDX-Net for covid-19 prediction from X-ray and also [21], has developed a model for prediction of COVID-19 epidemics that shows remarkable efficiency. J. Zhang and K. Nawata [28] in their studies stated that four-layer LSTM is effective in the prediction of an influenza outbreak. Godson

Kalipe et al. [33] stated that XGBoost is best for prediction of Malarial outbreak in comparison of KNN, Random Forest, SVM, XGboost, ANN, and Naive Bayes. Parul Arora et al. [21] compared RNN based on LSTM and stated that bi-directional LSTM is better than convolutional LSTM and Deep learning for prediction of disease outbreak. Rui Yin et al.[30] in their study indicated that weighted convolutional CNN is best for the prediction of Influenza. Kumar Dey et al.[42] found that SVR is more effective in prediction than MLR. Other researchers [34, 36, 39] have also stated that LSTM performs well. This survey demonstrates that AI/ML is effective in predicting epidemics. AI/ML models with additional ingredients [19, 20, 21] produce considerable improvements. However, it should be noted that every prediction model proposed or previously constructed to anticipate epidemics has limitations.

6. Conclusion

Timely predictions of epidemics are the need of the hour. Creating a new machine learning model that assists health practitioners in making more accurate and precise decisions reduces the later catastrophe caused by infectious diseases. This paper's findings imply that currently, existing machine learning algorithms for epidemic prediction have specific limits and constraints such as working with small size dataset, assumptions regarding long term predction etc. The findings also suggest that combining additional high-level problem-independent algorithmic frameworks with Machine learning models can improve epidemic forecasting skills. This review paper also emphasises the importance of developing machine learning models that are compatible with different locations, regions, and countries in the future.

References

[1] T. Britton and G. S. Tomba, "Estimation in emerging epidemics: Biases and remedies," *J. R. Soc. Interface*, vol. 16, no. 150, 2019, doi: 10.1098/rsif.2018.0670.

[2] G. Kumar, K. Kumar, and M. Sachdeva, "The use of artificial intelligence based techniques for intrusion detection: A review," *Artif. Intell. Rev.*, vol. 34, no. 4, pp. 369–387, 2010, doi: 10.1007/s10462-010-9179-5.

[3] N. K. Tran *et al.*, "Evolving Applications of Artificial Intelligence and Machine Learning in Infectious Diseases Testing," *Clin. Chem.*, vol. 68, no. 1, pp. 125–133, 2021, doi: 10.1093/clinchem/hvab239.

[4] G. Gauglitz, "Artificial vs. human intelligence in analytics," *Anal. Bioanal. Chem.*, vol. 411, no. 22, pp. 5631–5632, 2019, doi: 10.1007/s00216-019-01972-2.

[5] K. Rasheed, A. Qayyum, M. Ghaly, A. Al-Fuqaha, A. Razi, and J. Qadir, "Explainable, trustworthy, and ethical machine learning for healthcare: A survey," *Computers in Biology and Medicine*, vol. 149. 2022. doi: 10.1016/j.compbiomed.2022.106043.

[6] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," vol. 349, no. 6245, 2015.

[7] A. Pajankar and A. Joshi, *Introduction to Machine Learning with Scikit-learn*. 2022. doi: 10.1007/978-1-4842-7921-2_5.

[8] V. Nasteski, "An overview of the supervised machine learning methods," *Horizons.B*, vol. 4, pp. 51–62, 2017, doi: 10.20544/horizons.b.04.1.17.p05.

[9] H. O. Alanazi, A. H. Abdullah, and K. N. Qureshi, "A Critical Review for Developing Accurate and Dynamic Predictive Models Using Machine Learning Methods in Medicine and Health Care," *J. Med. Syst.*, vol. 41, no. 4, 2017, doi: 10.1007/s10916-017-0715-6.

[10] S. D. Villalba and P. Cunningham, "An evaluation of dimension reduction techniques for one-class classification," *Artif. Intell. Rev.*, vol. 27, no. 4 SPEC. ISS., pp. 273–294, 2007, doi: 10.1007/s10462-008-9082-5.

[11] R. S. Sutton and A. G. Barto, *Reinforcement learning, Second Edition in progress*. 2012. [Online]. Available: http://incompleteideas.net/sutton/book/the-book.html%5Cnhttps://www.dropbox.com/s/f4tnuhipchpkgoj/book2012.pdf

[12] Y. C. A. P. Reddy, P. Viswanath, and B. E. Reddy, "Semi - supervised learning : a brief review," vol. 7, pp. 81–85, 2018.

[13] A. Agresti, *An Introduction to Categorical Data Analysis: Second Edition*. 2006. doi: 10.1002/0470114754.

[14] M. J. Zaki and M. J. Meira, *Data Mining and Analysis: Fundamental Concepts and Algorithms*. 2013. [Online]. Available: https://books.google.com.tr/books?id=Gh9GAwAAQBAJ&lpg=PR9&dq=Data Mining and Analysis: Foundations and Algorithms&hl=tr&pg=PR9#v=onepage&q=Data Mining and Analysis: Foundations and Algorithms&f=false

[15] G. S. Fu, Y. Levin-Schwartz, Q. H. Lin, and D. Zhang, "Machine Learning for Medical Imaging," *J. Healthc. Eng.*, vol. 2019, no. 1, pp. 505–515, 2019, doi: 10.1155/2019/9874591.

[16] V. Mishra, Y. Singh, and S. Kumar Rath, "Breast Cancer detection from Thermograms Using Feature Extraction and Machine Learning Techniques," *2019 IEEE 5th Int. Conf. Converg. Technol. I2CT 2019*, 2019, doi: 10.1109/I2CT45611.2019.9033713.

[17] J. Bell, "Chapter 5 - Artificial Neural Networks," *Mach. Learn. Hands-On Dev. Tech. Prof.*, pp. 91–116, 2014.

[18] H. O. Alanazi, A. H. Abdullah, and K. N. Qureshi, "A Critical Review for Developing Accurate and Dynamic Predictive Models Using Machine Learning Methods in Medicine and Health Care," *J. Med. Syst.*, vol. 41, no. 4, Apr. 2017, doi: 10.1007/s10916-017-0715-6.

[19] S. F. Id, Z. Feng, C. Ling, C. Chang, and Z. F. Id, "Prediction of the COVID-19 epidemic trends based on SEIR and AI models," pp. 1–15, 2021.

[20] S. Rath, A. Tripathy, and A. R. Tripathy, "Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model," *Diabetes Metab. Syndr. Clin. Res. Rev.*, vol. 14, no. 5, pp. 1467–1474, 2020, doi: 10.1016/j.dsx.2020.07.045.

[21] P. Arora, H. Kumar, and B. Panigrahi Ketan, "Prediction and Analysis of COVID-19 Cases using Regression Models: A Descriptive Case Study of India," *J. Comput. Sci.*, vol. 18, no. 10, pp. 968–978, 2022, doi: 10.3844/jcssp.2022.968.978.

[22] M. Wieczorek, J. Siłka, and M. Woźniak, "Neural network powered COVID-19 spread forecasting model," *Chaos, Solitons and Fractals*, vol. 140, 2020, doi: 10.1016/j.chaos.2020.110203.

[23] S. I. Alzahrani, I. A. Aljamaan, and E. A. Al-Fakih, "Forecasting the spread of the COVID-19 pandemic in Saudi Arabia using ARIMA prediction model under current public health interventions," *Journal of Infection and Public Health*, vol. 13, no. 7. pp. 914–919, 2020. doi: 10.1016/j.jiph.2020.06.001.

[24] A. H. Elsheikh *et al.*, "Deep learning-based forecasting model for COVID-19 outbreak in Saudi Arabia," *Process Saf. Environ. Prot.*, vol. 149, pp. 223–233, 2021, doi: 10.1016/j.psep.2020.10.048.

[25] E. E. Hemdan, "COVIDX-Net : A Framework of Deep Learning Classifiers to Diagnose COVID-19 in X-Ray Images".

[26] M. Barstugan, U. Ozkaya, and S. Ozturk, "Coronavirus (COVID-19) Classification using CT Images by Machine Learning Methods," no. 5, pp. 1–10.

[27] A. Spreco, O. Eriksson, Ö. Dahlström, B. J. Cowling, and T. Timpka, "Integrated detection and prediction of influenza activity for real-time surveillance: Algorithm design," *J. Med. Internet Res.*, vol. 19, no. 6, pp. 1–21, 2017, doi: 10.2196/jmir.7101.

[28] J. Zhang and K. Nawata, "Multi-step prediction for influenza outbreak by an adjusted long short-term memory," 2018.

[29] G. H. . Miranda, J. M. Baetens, N. Bossuyr, and E. Al., "Real-time prediction of influenza outbreaks in Belgium _ Elsevier Enhanced Reader.pdf."

[30] R. Yin, Z. Luo, P. Zhuang, Z. Lin, and C. K. Kwoh, "VirPreNet: A weighted ensemble convolutional neural network for the virulence prediction of influenza A virus using all eight segments," *Bioinformatics*, vol. 37, no. 6, pp. 737–743, 2021, doi: 10.1093/bioinformatics/btaa901.

[31] T. Murayama, N. Shimizu, S. Fujita, S. Wakamiya, and E. Aramaki, "Robust two-stage influenza prediction model considering regular and irregular trends," *PLoS One*, vol. 15, no. 5, pp. 1–14, 2020, doi: 10.1371/journal.pone.0233126.

[32] T. Murayama, N. Shimizu, S. Fujita, S. Wakamiya, and E. Aramaki, "Predicting regional influenza epidemics with uncertainty estimation using commuting data in Japan," *PLoS One*, vol. 16, no. 4 April, pp. 3–5, 2021, doi: 10.1371/journal.pone.0250417.

[33] G. Kalipe, V. Gautham, and R. K. Behera, "Predicting Malarial Outbreak using Machine Learning and Deep Learning Approach: A Review and Analysis," in *Proceedings - 2018 International Conference on Information Technology, ICIT 2018*, Dec. 2018, pp. 33–38. doi: 10.1109/ICIT.2018.00019.

[34] F. Yang *et al.*, "Deep Learning for Smartphone-Based Malaria Parasite Detection in Thick Blood Smears," vol. 24, no. 5, pp. 1427–1438, 2020.

[35] P. Mohapatra, N. K. Tripathi, I. Pal, and S. Shrestha, "Determining suitable machine learning classifier technique for prediction of malaria incidents attributed to climate of Odisha," *Int. J. Environ. Health Res.*, vol. 32, no. 8, pp. 1716–1732, 2022, doi: 10.1080/09603123.2021.1905782.

[36] M. F. X. Barboza, "Real-time prediction of influenza outbreaks in Belgium," *Rev. Soc. Bras. Med. Trop.*, vol. 55, no. April, pp. 1–9, 2022, doi: 10.1590/0037-8682-0420-2021.

[37] D. H. Id, W. Valkenburg, and A. Amara, "Predicting malaria epidemics in Burkina Faso with machine learning," pp. 1–16, 2021.

[38] E. Kamana, J. Zhao, and D. Bai, "Predicting the impact of climate change on the re-emergence of malaria cases in China using LSTMSeq2Seq deep learning model : a modelling and prediction analysis study," 2022, doi: 10.1136/bmjopen-2021-053922.

[39] S. G. Kakarla, P. Krishna, K. Hari, P. Vavilala, G. Sumanth, and B. Boddeda, "Weather integrated multiple machine learning models for prediction of dengue prevalence in India," no. 0123456789, 2022.

[40] S. Anno and T. Hara, "View of Spatiotemporal dengue fever hotspots associated with climatic factors in Taiwan including outbreak predictions based on machine-learning. pdf."

[41] V. N. Id *et al.*, "PLOS NEGLECTED TROPICAL DISEASES Deep learning models for forecasting dengue fever based on climate data in Vietnam," pp. 1–22, 2022.

[42] S. Kumar *et al.*, "PLOS ONE Prediction of dengue incidents using hospitalized patients, metrological and socio- economic data in Bangladesh : A machine learning approach," no. August 2021, pp. 1–17, 2022.

[43] F. Shatnawi and Q. Ba. BAker, "Detecting Epidemic Diseases Using Sentiment Analysis of Arabic Tweets," vol. 26, no. 1, pp. 50–70, 2020.

[44] K. Y. Id, T. Fujimoto, M. Muramatsu, and H. Shimizu, "Prediction of hand , foot , and mouth disease epidemics in Japan using a long short-term memory approach," vol. 71, pp. 1–16, 2022.