Ontology-Based Data Collection for a Hybrid Outbreak Detection Method Using Social Media

Ghazaleh Babanejaddehaki, Aijun An, and Heidar Davoudi

***Abstract*— Given the persistent global challenge presented by rapidly spreading diseases, as evidenced notably by the widespread impact of the COVID-19 pandemic on both human health and economies worldwide, the necessity of developing effective infectious disease prediction models has become of utmost importance. In this context, the utilization of online social media platforms as valuable tools in healthcare settings has gained prominence, offering direct avenues for disseminating critical health information to the public in a timely and accessible manner. Propelled by the ubiquitous accessibility of the internet through computers and mobile devices, these platforms promise to revolutionize traditional detection methods, providing more immediate and reliable epidemiological insights. Leveraging this paradigm shift, our proposed framework harnesses Twitter data associated with infectious disease symptoms, employing ontology to identify and curate relevant tweets. Central to our methodology is a hybrid model that integrates XGBoost and Bidirectional Long Short-Term Memory (BiLSTM) architectures. The integration of XGBoost addresses the challenge of handling small dataset sizes, inherent during outbreaks due to limited time series data. XGBoost serves as a cornerstone for minimizing the loss function and identifying optimal features from our multivariate time series data. Subsequently, the combined dataset, comprising original features and predicted values by XGBoost, is channeled into the BiLSTM for further processing. Through extensive experimentation with a dataset spanning multiple infectious disease outbreaks, our hybrid model demonstrates superior predictive performance compared to state-of-the-art and baseline models. By enhancing forecasting accuracy and outbreak tracking capabilities, our model offers promising prospects for assisting health authorities in mitigating fatalities and proactively preparing for potential outbreaks.**

***Index Terms*— Outbreak detection, machine learning, deep learning, time series, ontology, social media.**

I. Introduction

THE integration of the social media is pivotal for disease monitoring and prediction, enabling the seamless inclusion of low-cost and time-sensitive information sources into established surveillance systems. Research has underscored the potential benefits for public health officials in leveraging social media platforms and search engines to enhance their capabilities in monitoring and responding to disease outbreaks [1], [2], [3], [4]. There are more social media applications available today than ever before, and they are experiencing growth in user bases across various demographics. Social networking sites have evolved into quick and inexpensive ways to share ideas and access public information. Moreover, unreported health information and health-related opinions can be found on social media.

Twitter is used extensively in research because it is the most widely used social networking site for the dissemination of information and ideas, including studies on climate change communication, sentiment analysis during natural disasters, tracking public response to political campaigns, analyzing public health awareness during disease outbreaks, and understanding consumer behavior and preferences in the fashion industry [2], [5], [6], [7], [8]. It is an effective resource for tracking trends for a variety of reasons. First, the high frequency of posted messages enables minute-by-minute analysis. Second, Twitter posts are more descriptive and public than search engine logs, analyses can be conducted by examining user profile data, which encompasses demographic data and specific information. Third, Twitter users vary in age, including not only young people but also middle-aged and technologically savvy seniors [9].

According to studies [3], [10], [11], [12], [13], [14], [15], combining social media data with traditional surveillance information such as clinical, hospital, and laboratory reports can aid in the detection of outbreaks and the prevention of severe consequences. However, the majority of these studies used the disease's name as a keyword to collect data from social media, as well as sentiment analysis and supervised machine learning methods to detect or track outbreaks. Detecting the signals for unknown diseases by focusing on specific disease names or common symptoms is not always effective. For example, prior to January 2020, most people with symptoms such as fever and sore throat were diagnosed with seasonal flu. However, as more symptoms were reported, a new coronavirus closely related to SARS-CoV (now called SARS-CoV-2) was found to be the cause of the symptoms; this virus was given the name COVID-19 [16]. Thus, outbreaks could be detected more quickly and accurately if data collection is expanded to include a broader range of symptoms

Manuscript received 15 May 2024; accepted 6 August 2024. Date of publication 13 August 2024; date of current version 14 October 2024. This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) via an Emerging Infectious Diseases Modeling Initiative (EIDM) Grant. *(Corresponding author: Ghazaleh Babanejaddehaki.)*

Ghazaleh Babanejaddehaki and Aijun An are with the Department of Electrical Engineering and Computer Science, York University, Toronto, ON M3J 1P3, Canada (e-mail: ghazalba@yorku.ca; aan@yorku.ca).

Heidar Davoudi is with the Faculty of Science, Ontario Tech University, Toronto, ON L1G 0C5, Canada (e-mail: heidar.davoudi@ontariotechu.ca).

Digital Object Identifier 10.1109/TNB.2024.3442912

1558-2639 © 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Ontario Tech University. Downloaded on November 10,2026 at 19:29:46 UTC from IEEE Xplore. Restrictions apply.

related to various infectious diseases, encompassing diverse semantic names for these symptoms. Ontology as a knowledge base can provide comprehensive symptom names.

Various techniques and analyses were employed to identify outbreaks. Numerous multidisciplinary studies, based on traditional compartmental susceptible-infected-removed (SIR) models, have extensively examined the effectiveness of various public intervention measures and control strategies. These studies aim to better understand how different interventions, such as vaccination campaigns, social distancing measures, and travel restrictions, impact the spread and control of infectious diseases within a population. References [17] and [18]. Time Series Analysis (TSA) is also one of the most common types of analysis used for outbreak detection and case prediction in the health domain using both traditional surveillance data and social media data sources. TSA can extract information from past time periods and use that information to make predictions about present and future time periods. Univariate and multivariate analyses of temporal data have been used in studies of disease outbreaks to predict the number of new cases and fatalities. To predict or track outbreaks, these studies employ various statistical, deep learning, machine learning, and hybrid models [19], [20], [21], [22], [23], [24], [25], [26].

Regression analysis is among the numerous statistical techniques used to model the spread of infectious diseases. XGBoost, one of the most effective regression models, is an optimized distributed gradient boosting library that can examine the significance of all input features and perform well with small datasets. It has been shown to be a trustworthy and effective regression machine learning problem solver [27], [28], [29], [30], [31].

Moreover, while traditional statistical models assume linearity and stationarity, which may not hold true for infectious diseases with complex, nonlinear, and nonstationary dynamics, machine learning models can capture complex and nonlinear patterns and may be more suitable for modeling infectious diseases with complex dynamics and interactions. As a result, most of the previous studies focusing on traditional statistical models were unable to guarantee the development of an accurate and robust infectious disease forecasting model, as they either reported moderately unsatisfactory results [32] or their predictions still did not align well with ground-truth data [33], [34].

In contrast, deep learning techniques have recently been successfully applied to a variety of challenging real-world prediction problems, including time-series forecasting [35], [36], [37]. Deep learning approaches are commonly regarded as the most effective method for dealing with the noisy and chaotic nature of time series forecasting problems because it generates more accurate predictions. Long-term short-term memory (LSTM) is one of the most effective and widely used deep learning techniques. However, the size of the dataset has a significant impact on the performance of deep learning models, and the majority of infectious disease outbreak time series data sets are relatively small.

As a result, a time-series model that employs both deep learning and non-deep-learning machine learning models together may improve prediction accuracy. In this paper, we propose a hybrid and novel machine-learning-based model. It utilizes retrieved texts from Twitter, focusing on an infectious disease ontology and symptoms associated with six distinct infectious diseases. These texts are transformed into a time series dataset, enabling us to predict disease outbreaks. Our model handles multivariate time series data, allowing us to forecast the number of infected cases for each disease and accurately predict the onset of outbreaks.

The proposed model tackles the challenge of potentially limited training data for the Bi-LSTM through a novel combined training approach leveraging the strengths of XGBoost. First, an XGBoost model is trained on the original dataset to capture the relationships between historical symptom-related tweets and infected cases. This knowledge is then integrated into the training process. We create a new, richer dataset for Bi-LSTM training by combining the original features (daily tweet counts for symptoms), actual infected cases, and predicted infected cases from the trained XGBoost model. By training the Bi-LSTM on this combined dataset, it can learn from both historical tweet patterns and the initial predictions from XG-Boost, potentially refining the final forecast for infected cases. This approach allows the Bi-LSTM to benefit from XGBoost's expertise while mitigating the limitations of potentially small training data. The main contributions of this study are as follows:

- We use disease ontology to identify infectious disease symptoms and extract the data from Twitter.
- We predict the infected number of cases with a novel hybrid model based on tweets related to the symptoms of infectious diseases.
- We predict the contaminated cases earlier than what the health authorities have reported.
- We leverage the information from past outbreaks in each region to identify potential future outbreaks. By analyzing historical data, we can discern patterns and trends that aid in the detection and prediction of forthcoming outbreaks.
- Our method outperforms ten other machine learning methods used for infectious disease case prediction.

The structure of the paper is as follows: Section II describes related work, while Section III describes the preliminaries. Section IV explains the proposed model. Section V details the evaluation and experiments. The paper concludes in Section VI.

II. Related Work

This section presents a concise review of related research on outbreak detection and the use of machine-learning algorithms for predicting positive cases and deaths from infectious diseases. Increasingly, researchers and medical professionals are turning to social media for various purposes, including professional development, networking, health campaigns, and disease monitoring and prediction [38].

Several review articles [3], [6], [15], [39], [40], [41] have examined the role of social media platforms in disease surveillance. They found that social media offers advantages in terms of speed, accuracy, and cost-effectiveness, recommending its integration with existing surveillance systems. Additionally,

Authorized licensed use limited to: Ontario Tech University. Downloaded on November 10,2026 at 19:29:46 UTC from IEEE Xplore. Restrictions apply.

some studies have explored how social media enhances health protection, public health messaging, and lessons learned from health campaigns.

The importance of predicting infectious disease outbreaks has been demonstrated by various research studies [10], [11], [19], [20], [21], [22], [24], [25], [31], [42], [43], [44]. These studies employed machine learning techniques, including classification, regression algorithms, and statistical methods, to track and predict outbreaks. By leveraging data from diverse sources, such as social media and health authorities, these methods effectively mitigate the impact of infections on public health using univariate or multivariate time series data.

In our study, we compare the performance of our proposed method with several baselines and state-of-the-art models, including Vector Autoregression (VAR), eXtreme Gradient Boosting (XGBoost), Long Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory (Bi-LSTM), N-Beats, FaceBookProphet, Temporal Fusion Transformer (TFT), and DeepAR. We are going to mention some of the recent works that have used these models.

The vector autoregressive (VAR) [45] model has proven to be especially useful for forecasting and describing the dynamic behavior of economic and financial time series. Goto et al. [46] used time series analyses based on VAR to investigate the impact of meteorological variables on dengue incidence in Sri Lanka. To predict positive COVID-19 cases, Wang et al. [47] proposed a validated Vector Autoregression (VAR) time series model.

XGBoost (eXtreme Gradient Boosting) [48], [49] is commonly used for feature selection. Lv et al. [50] predicted the number of Hemorrhagic fever with renal syndrome (HFRS) cases in mainland China using the XGBoost model from 2004 to 2018. Luo et al. [31] compared LSTM and XGBoost models to study the future trend of COVID-19 in the United States, finding that the XGBoost model showed better performance. Similarly, Alim et al. [51] and Fang et al. [29] compared the accuracy of the autoregressive integrated moving average (ARIMA) model and the XGBoost model in predicting COVID-19 occurrences in the United States and brucellosis in mainland China. In both studies, the XGBoost model demonstrated superior predictive capability. Guo et al. [30] proposed a mixed framework of XGBoost-LSTM for COVID-19 prediction. They used XGBoost to predict transmission trends based on key infection features and LSTM to predict transmission fluctuations using infection-related multiple time series features. The combination of the two models achieved more accurate predictions of infectious disease spread. In contrast, our proposed method in this paper involves the combination of XGBoost and LISM during the training phase, while Guo et al.'s method combines the separately trained models in the prediction phase, the model using a weighted average or another appropriate statistical method to arrive at a final prediction of disease transmission to achieve a more accurate prediction of infectious disease spread. The XGBoost model captures the key features of infectious diseases, while the LSTM model considers the influence of multiple factors on the spread of infectious diseases. Additionally, their method uses a time series clustering algorithm to improve predictions in different regions based on population flow.

In the field of deep learning, LSTM is an artificial Recurrent Neural Network (RNN) architecture. It is an effective algorithm for building sequential time series models. Deep-bidirectional LSTM [52] is an extension of the previously described LSTM models that apply two LSTMs to the input data [53]. For predicting the number of Covid-19 infected cases there are research works e.g. in India, Arura et al. [54] employed recurrent neural network (RNN)-based variants such as deep LSTM, convolutional LSTM, and Bi-LSTM. Wang et al. [49] used LSTM networks and rolling updating mechanisms to incorporate new forecasting outcomes into model training for the subsequent iteration of different countries. Hasan [55] found that a hybrid model composed of ensemble empirical mode decomposition (EEMD) and artificial neural network (ANN) outperformed traditional statistical analysis.

Temporal Fusion Transformer (TFT) [56] is an attention-based, performance- and interpretability-optimized Deep Neural Network. Multi-Horizon Quantile Recurrent Forecaster (MQRNN) [10] is another DNN implementation with good performance, but it provides no insight into feature interpretability. Caldas and Soares [57] utilized TFT to forecast the volume of non-urgent patients in Portuguese public hospitals by Health Regional Areas (HRA). Zhou et al. [58] used TFT for Covid-19 forecasting as one of the baselines.

Prophet is a time series forecasting model that Facebook created initially to address time series issues in the business world. Although there are many different approaches to predicting business outcomes, many business issues have a number of features in common, like seasonal effects [59]. The approach makes use of a simple time-series model [60] with three key elements: trend, seasonality, and holidays. Prophet uses a piece-wise linear model and a saturating growth model to predict trends. For growth forecasts, a model that resembles population growth models in natural ecosystems [61] is utilized, where nonlinear growth encounters a saturation point at a carrying capacity. An effective and frequently helpful solution is a piece-wise model with a constant growth rate for forecasting issues where this saturating point is never reached. Prophet uses the Fourier series to create a flexible model of periodic effects that can capture seasonality [62], but to account for holidays, one needs to have a predetermined list of past and future holiday events. Since holiday effects are presumed to be independent, adding them to the model becomes easy [59]. Recent studies by Sah et al. [44] and Papastefanopoulos et al. [25] used Prophet to predict the Covid-19 pandemic.

DeepAR is a probabilistic forecasting technique based on autoregressive recurrent neural networks. DeepAR provides three key advantages: In the beginning, it generates probabilistic predictions in the form of Monte Carlo samples that can be used to determine consistent quantile estimates for all sub-ranges in the forecasting scope. Second, because Gaussian noise is not assumed, a wide range of likelihood functions can be supported, allowing the user to select the

Authorized licensed use limited to: Ontario Tech University. Downloaded on November 10,2026 at 19:29:46 UTC from IEEE Xplore. Restrictions apply.

function that best matches the statistical characteristics of the data. Finally, because it is based on deep learning, it can learn complex dependencies and seasonal behaviors with little manual assistance [63]. There are some recent works using DeppAR to predict the active cases [25], [64], [65]

N-Beats model functions generally in a similar manner to conventional decomposition methods, like the seasonality-trend-level approach [66]. The model is extremely quick to train and can be applied with little if any, adjustment to a wide range of target domains in addition to benefits for interpretability and accuracy, as measured on several well-known datasets [67] and the recent works that used N-Beats for Covid19 outbreak detection [25], [64], [68], [69], [70].

III. Preliminaries

In this section, we present the necessary definitions and mathematics behind the two models that are used throughout the proposed hybrid method.

A. XGBoost Method

XGBoost is a boosting algorithm based on the evolution of gradient boosting decision tree (GBDT) algorithm, which has achieved remarkable results in practical applications due to its high accuracy, fast speed and unique information processing scheme. Compared with traditional statistical models, it has advantages in predicting nonlinear data [50], [71], [72].

XGBoost, a kind of boosting algorithm, which assembles multiple learning algorithms to achieve a better predictive performance than any of the constituent learning algorithms alone, has excelled in many fields. Compared with the traditional GBDT algorithm, XGBoost applies a second-order Taylor expansion to the loss function and simultaneously implements the first derivative and the second derivative. In addition, a regularization term is added to the objective function, which improves the generalizability of a single tree and reduces the complexity of the objective function. In short, XGBoost has attracted the attention of researchers due to its fast speed, excellent classification effect, and ability to allow custom loss functions [50].

The classification and regression tree (CART) algorithm, first proposed by Breiman [73], refers to the general term of a classification tree and regression tree. The CART classification tree introduces the Gini coefficient to replace the information gain or information gain rate. The regression tree adopts different methods to evaluate the effect, including the prediction error (mean squared error, log error, etc.). Therefore, the node is no longer a category but a numerical value. The XGBoost algorithm is mainly composed of two parts: the decision tree algorithm and gradient boosting algorithm. Gradient boosting is an excellent technique for constructing prediction models and a representative algorithm for boosting. The theory of boosting is to establish weak evaluators individually and iteratively integrate multiple weak evaluators. The gradient boosting tree uses the CART algorithm as the main structure. The XGBoost model also draws on the idea of random forest, introducing row sampling and column sampling that can reduce the amount of calculation and prevent overfitting. Moreover, it introduces the early-stopping mechanism to prevent overfitting. XGBoost aims to reduce the regularized objective in the following manner:

$$obj(\theta) = \sum_i L(y_i, \hat{y}_i^{(t)}) + \sum_k \Omega(f_t) \quad f_k \in F \tag{1}$$

where L is the training loss function that measures the deviation between the value predicted by our model and the actual value y_i, Ω is the regularization function that measures the complexity of the model to prevent overfitting, f is a function in the functional space F, and F is the set of all possible regression trees. In order to minimize the objective function, XGBoost uses parameters to find an optimal tree structure employing a greedy search algorithm.

B. Bidirectional LSTM Method

RNN has been used for problems involving sequential time series and temporal dependencies [74]. An unfurled RNN is capable of processing recent input using historical data. In the meantime, one of the RNN variations can train the longterm dependency data, which is a challenge for RNN. With the help of memory cells, a hidden layer unit that Hochreiter and Schmidhuber [75] predicted will be employed in LSTM, RNN's limitations have been removed. The input gate, output gate, and forget gate are the three gates that govern the selfconnections in memory cells, which store and control the network's temporal state [76].

Input and output gates are used to regulate how memory cell inputs and outputs are distributed throughout the rest of the network. Additionally, the memory cell now incorporates a forget gate, which passes output data with large weights from one neuron to the next. Information that has high activation results is retained in memory cells; if the input unit has high activation, then the information is stored there. Additionally, if the output unit is highly activated, the information will be passed to the following neuron. Otherwise, high weight input data is stored in a memory cell.

Standard RNNs simply consider the inputs in one direction and ignore any future knowledge. By taking into account both the past and the future, the bidirectional LSTM (Bi-LSTM) recovers the entire temporal information of time t. This divides the standard RNN's hidden neurons into forward and backward states, with the forward state neurons not coupled to the backward state neurons and vice versa. Figure 1 displays the BiLSTM's time-step fundamental structure in its unfolded state. This structure is comparable to the conventional unidirectional RNN without the backward states.

The Bi-LSTM structure shown in Phase 2 of Figure 1 can be expressed by the following equation:

$$h_i = LSTM(h_i, \leftarrow) \oplus LSTM(h_i, \rightarrow) \tag{2}$$

where h_i represents the output at time step i, LSTM$(x_i, \leftarrow)$ denotes the LSTM operation on the input x_i in the forward direction, and LSTM$(x_i, \rightarrow)$ denotes the LSTM operation on the input x_i in the backward direction, and $\oplus$ represents the concatenation of the outputs from both forward and backward directions, capturing the bidirectional context.

Authorized licensed use limited to: Ontario Tech University. Downloaded on November 10,2026 at 19:29:46 UTC from IEEE Xplore. Restrictions apply.

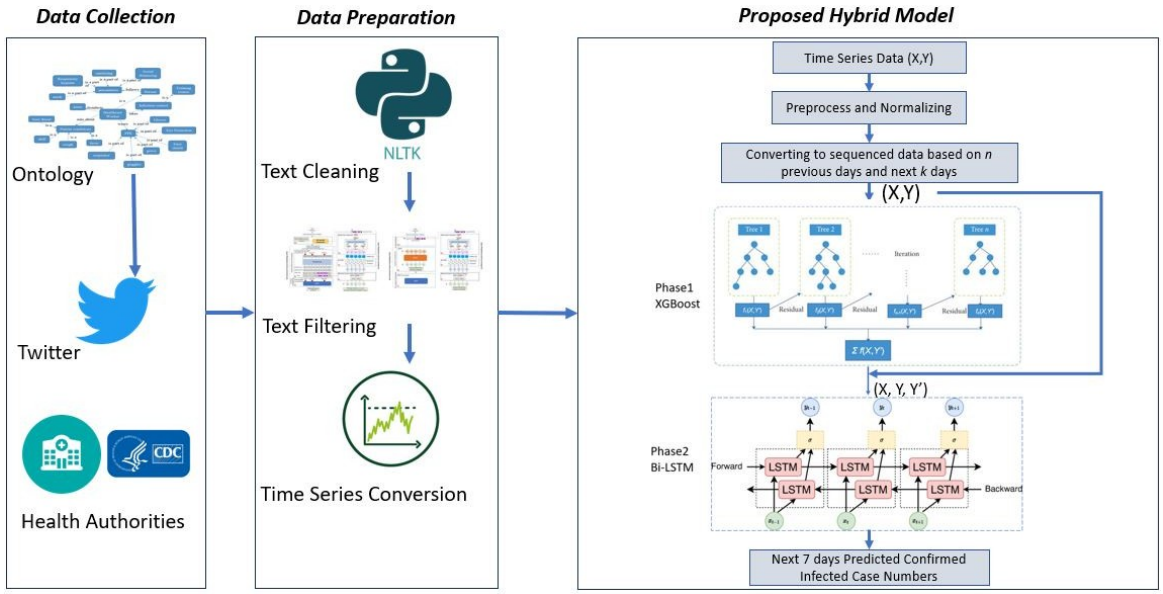


Fig. 1. Proposed outbreak detection method steps.

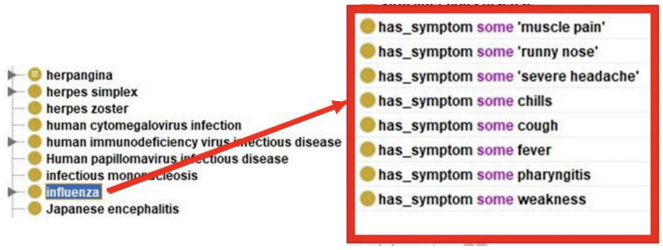


Fig. 2. Viral infectious disease ontology.

IV. Proposed Method

The proposed methodology includes three components: *data collection, data preparation, and the hybrid machine-learning model.* Figure 1 depicts the steps of each component of the proposed method for detecting six infectious diseases. Each component will be discussed in detail in the following subsections.

A. Data Collection

One novel aspect of this research is the use of infectious disease symptoms frequency in daily tweets to detect outbreaks rather than disease names, as in most cases for new infectious diseases, such as the recent Covid-19, we may not have a name for them in the early stages. Another novel aspect of this work is the use of ontology, which allows access to a comprehensive list of infectious disease symptoms with similar semantic names, such as fatigue and tiredness. In this study, we employ a disease ontology [77], [78] concentrating on viral infectious diseases and using SPARQL Query to extract the related symptoms in the symptom subclass connected to diseases with has symptom relationship as depicted in Figure 2. This ontology includes 104 symptoms for different infectious diseases. As our focus will be on the duration between 2012 to 2021, six popular infectious diseases (Ebola, Zika, MERS, H1N1, Chikungunya, Covid19) outbreaks were selected from this duration in this study. The identified symptoms related to these diseases are 65, which will be used as keywords to collect tweets from Twitter. We also collected the daily number of cases for each of these diseases from health authorities [79], [80]. Referring to our first contribution, we will conduct experiments to evaluate the efficacy of the proposed model by comparing its results on three different collections of tweets. The first group will employ ontology to obtain tweets mentioning infectious disease symptoms (i.e., using 65 keywords). In addition, we compiled two additional sets of data from Twitter: one set centered on disease names and common symptoms mentioned in health authority websites (totaling 18 keywords) [79], [80], and the other set centered on disease names alone (totaling 6 keywords).

B. Data Preparation

After obtaining the tweets, some data preparation is required. In the first stage, natural language processing techniques [81] are used to delete and remove stop words, sparse keywords, and specific terms that do not carry vital information. Then all punctuation, accent marks, and other diacritics were removed. After that, the tweet text was transformed into token words. These preprocessing steps were incorporated to increase the proposed model's efficiency and processing speed.

The second stage is to filter out noisy and unrelated tweets about the sickness. In this step, we used some of the strategies described in [82] and [83] to filter out tweets that are not about the disease and tweets that contain sarcasm. The final stage involves converting these cleaned disease-related tweets into a time series. The final output format will be a numerical dataset that aggregates the daily number of tweets mentioning each keyword by country. In the first dataset, we may have a list of synonyms for the same symptom, such as "fatigue", "tiredness" and "exhaustion" from the ontology we used. In the final step of creating a time series dataset, we combine these synonyms under the same category and count them all together to make the dataset more efficient by decreasing the number of

Authorized licensed use limited to: Ontario Tech University. Downloaded on November 10,2026 at 19:29:46 UTC from IEEE Xplore. Restrictions apply.

dimensions for our predictive models. For instance, the sum of daily tweets about fatigue, tiredness, and exhaustion will be counted under the concept of "fatigue". The dataset will consist of a time series of the number of daily tweets of symptoms as independent variables and the number of confirmed infected cases as a dependent variable. We first transform it into a supervised learning dataset and then generate the sequences needed to predict the number of infected people for the next k days based on the n previous days' number of tweets of symptoms. Each row in the supervised learning dataset will consist of (n*m) independent variables related to (m) symptoms experienced over the previous (n) days between time (t-n) and (t-1) and the related dependent variable of the number of confirmed infected cases at the time (t).

C. Proposed Hybrid Model

We introduce a novel hybrid model that combines the predictive capabilities of XGBoost and Bidirectional LSTM (BiLSTM) to forecast confirmed infected cases. This innovative approach aims to capitalize on the strengths of each model while mitigating the challenges associated with limited data availability.

1) Hybrid XGBoost-BiLSTM Proposed Method: In this section, we depict the operation of the proposed model. In the first phase, the training dataset (X, Y) is utilized in a supervised learning framework, where the independent variables (X) comprise the count of daily tweets from the previous (n) days for each symptom, while the corresponding dependent variables (Y) represent the number of infected cases for the next k days. Specifically, each row in the training dataset encompasses ($n \times m$) independent variables linked to (m) symptoms experienced over the preceding n days, spanning from time ($t - n$) to ($t - 1$), alongside the corresponding dependent variable indicating confirmed cases from time t to $t + k - 1$. This dataset is then employed to train the XGBoost. For instance, we may use the number of tweets related to symptoms from the previous 10 days to predict the number of confirmed cases for the following 7 days.

Our proposed model employs an integrated training strategy to enrich the learning process of the Bi-LSTM. This strategy involves the creation of a new data frame that incorporates various components crucial for training:

- Original Features (X): Daily tweet counts for each symptom recorded over the preceding n days.
- Actual Infected Cases (Y): The number of infected cases for the subsequent k days.
- Predicted Infected Cases from X GBoost (Y').

By combining these elements into a unified training dataset, we provide the Bi-LSTM with a richer source of information for learning. This enables the model to glean insights from historical tweet patterns represented by X, incorporate actual future infected cases from Y, and refine its predictions based on initial forecasts generated by X GBoost (Y').

Our hybrid model leverages the expertise of XGBoost in discerning complex relationships between features and infected cases. This allows the Bi-LSTM to capitalize on the knowledge captured by XGBoost, thereby enhancing its forecasting accuracy. Furthermore, by merging Y and Y' with the original features (X), our approach addresses the inherent challenge of limited training data for the Bi-LSTM. This integration of richer information during training empowers the Bi-LSTM to make more informed predictions about future infected cases. In summary, our proposed hybrid model with integrated training offers a promising solution for forecasting infected cases based on historical tweet data. By seamlessly integrating the strengths of both XGBoost and Bi-LSTM, our model aims to deliver more accurate and reliable predictions, ultimately contributing to the advancement of outbreak detection and management strategies.

2) Evaluation Metrics: We evaluate the predictive effect of our model using two popular forecasting parameters, such as Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE), as follows:

$$RMSE = \sqrt{(\frac{1}{n})\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{3}$$

$$MAPE = \frac{100\%}{n}\sum_{t=1}^{n}|\frac{\hat{y}_i - y_i}{y_i}| \tag{4}$$

where n is the number of observations, and $\hat{y} - y_i$ is the error between the forecasted value and the actual value. RMSE is a frequently used measure of the differences between values predicted by a model or an estimator and the values observed. It is the square root of the average squared error. MAPE quantifies accuracy as a percentage, which can be calculated as a cumulative absolute percent error for each time frame as the actual values minus the predicted values divided over the actual values. Therefore, it represents the average error as a percentage.

V. Experimental Evaluation

This study aims at proposing a novel approach to predicting the number of positive cases using tweets related to infectious disease symptoms. We used the Python language for implementation and Twitter API for collecting the tweets. The experiments are carried out in the Google Colaboratory using Python 3.6, PyTorch 1.6, 16G memory, NVIDIA Corporation TU104GL [Tesla T4] GPU and the Spyder environment with open-source libraries such as Tensorflow, Pandas, Numpy, and Keras for different models implementation. We also used Protege for querying and extracting information from our ontology.

A. Data

We use social media data from Twitter, one of the most popular social media and network platforms with about 340 million users in more than 150 countries. We collected Twitter data for the duration of 2012-2021, using 65 symptoms derived from the ontology as keywords related to six popular disease outbreaks (Ebola, Zika, MERS, H1N1, Chikungunya, Covid19) and another two sets of datasets will be using 18 keywords including disease names and the most common symptoms reported by health authorities, and the last dataset will include just 6 keywords focusing on the disease names.

Authorized licensed use limited to: Ontario Tech University. Downloaded on November 10,2026 at 19:29:46 UTC from IEEE Xplore. Restrictions apply.

TABLE I
SAMPLE DISEASE OUTBREAKS TIME SERIES DATA

Country	Diseases	Start Range	End Range	Number of Days
Guinea	Ebola	25 March 2014	13 April 2016	750
Brazil	Zika	1 January 2015	31 July 2016	577
Saudi Arabia	MERS	23 April 2014	31 August 2015	495
India	H1N1	1 January 2015	31 December 2015	365
Colombian Caribbean	Chikungunya	6 December 2013	31 October 2014	329
USA	Covid-19	19 January 2020	30 April 2020	101
Canada	Covid-19	28 January 2020	30 April 2020	93

Six attributes were used in the subsequent tweet analysis: source, username, date tweet, language, and location.

Data cleaning is the first step in Twitter data preprocessing. We used the 'source' attribute in Twitter data to remove bot posted tweets. Li et al. [84] developed a list of sources (e.g., sent from an iPhone) to aid in the identification of human generated tweets. We examined each tweet's 'source' value and excluded tweets which sources were not listed [84]. Then we used NLP to clean the data and [82], [83] to filter out tweets that were unrelated to the disease or contained sarcasm. For sarcasm detection, this approach utilizes Affective and Contextual Embeddings ACE 1 and ACE 2. When fed a text passage (a collection of sentences, or "document" for short) the models attempt to determine whether or not the text is sarcastic. Each model's architecture depends on two subcomponents: a) an embedding of emotional features, and b) an embedding of contextual features. To obtain representations of a document's emotional qualities, the former makes use of a Bi-LSTM with multi-head attention neural architecture [85]. For the latter, a BERT model is used. In ACE 1, we combine the two by training a brand-new BERT model with the addition of affective feature embeddings into the input sequence of BERT to obtain task-specific embeddings. After training on labeled sarcasm detection data, ACE 2 combines the two components in a fully connected layer using a softmax to produce a classifier.

The next step is to determine the location from which each tweet was sent. To geo-reference tweets, three common metadata sources are used: geotagged location (tweet from), user profile address (user from), and content mentioned place (tweet about). The final stage for all three sets of datasets with and without ontology is to convert these cleaned diseaserelated tweets to time series, as explained in Section IV. The format will be a numerical dataset that counts the number of daily tweets related to each keyword, broken down by country.

The symptoms and disease names are considered independent variables in our multivariate time series dataset. The dependent variable is the number of confirmed cases retrieved from health authorities [79], [80]. Table I shows the diseases and the time spans we used to train and test our models.

B. Experiments

For all baselines, the state of the arts and the proposed hybrid model, we used the aforementioned historical three sets of data to make the number of infected case predictions. The preprocessed dataset was divided into an 80% training and 20% test set. Min-max normalization was used to normalize the time series data within the range [0, 1]. MinMaxScaler preserves the shape of the original data distribution. It has no effect on the information contained in the original data. After predicting time series on testing data, this normalization must be reversed so that the predicted data are similar to the original testing time series data, though some models, such as DeepAR, have automatic scaling that we do not need to apply manually. Furthermore, for the VAR model, we must first make the data stationary, as we know that most outbreak datasets are not stationary for a variety of reasons, the most common of which is that we have trends for them, as shown in Figure 5 for the Covid-19 dataset for the USA based on duration mentioned in Table I. Augmented Dickey-Fuller (ADF) test was used to test and confirm its non-stationarity, and by applying the BoxCox transformation, the original data became less volatile and more stationary.

The proposed method will be compared with the nine baselines shown in Table II. The selection of optimal hyperparameters is critical for obtaining a high-accuracy forecasting model [86]. Deep learning models typically include a number of hyperparameters. We use the Bayesian optimization algorithm to fine-tune the hyperparameters of each model during the model training process. Although grid search is a popular strategy for determining optimal hyperparameters, fine-tuning deep learning methods requires more computational time and resources. This is because the grid search method considers all parameter combinations exhaustively, necessitating more computational resources, particularly in the case of deep learning.

The primary reason for using Bayesian hyperparameter optimization is that it does not consider all hyperparameter combinations, requiring less training time and resources. Bayesian hyperparameter optimization predicts good tuning parameters using Bayesian models based on Gaussian processes [86]. Wu et al. [86] found that the Bayesian optimization-based method could find the best hyperparameters for popular machine learning algorithms. The Bayesian optimization technique [87], [88] is used to tune the hyperparameters in all of the experiments in this study, in accordance with [84], [89], [90]. The Bayesian optimization algorithm determines the appropriateness of each model based on the error in the validation data.

Table II depicts the domain of all hyperparameters used in the models that were implemented. The window size hyperparameter used in transforming input time series into suitable sequenced data for each dataset, the window size number has a significant impact on obtaining models with minimal error

Authorized licensed use limited to: Ontario Tech University. Downloaded on November 10,2026 at 19:29:46 UTC from IEEE Xplore. Restrictions apply.

TABLE II
HYPERPARAMETER RANGES

Hyperparameter	Range
Common hyperparameters	Window size (Number of past days): [2, 3, 5, 7, 10, 11, 12, 13, 14, 15], Forecast length: 7 Learning rate: [0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05] Dense activation function: [ReLU, Linear], Output activation function: [ReLU, Linear]
LSTM, Bi-LSTM	Activation function: [ReLU, Linear] Dropout rate: [0.0, 0.1, 0.2, 0.3, 0.4, 0.5] Number of Units: [4, 8, 16, 32, 64, 128] Layers:[1,2,3,4], Epochs:[100,200,300,400,500]
XGBoost	n_estimators:[100,200,300]
VAR and LSTM	Differencing:[1, 2, 3, 4, 5]
N-BEATS	Epochs:[100,200,300], Batch: [32, 64, 125, 512], LWIDTH: [32, 64, 125, 512]
FaceBook Prophet	Growth: ['linear', 'logistic'] , Changepoint_prior_scale: [0.1, 0.2, 0.3, 0.5, 0.8]
Temporal Fusion Transformer or TFT (Google)	LSTMLAYERS: [2,4,6,8,10], Dropout rate: [0.0, 0.1, 0.2, 0.3, 0.4, 0.5] Batch: [32, 64, 125, 512], Epochs:[100,200,300,400,500], Hidden_size: [16, 20, 32, 64, 125, 512]
DeepAR (Amazon)	Num_layers : [2, 4, 6, 8] , Num_cells: [21, 64, 128, 512] Cell_type : [LSTM, RNN, GRU] , Batch: [32, 64, 125, 512]

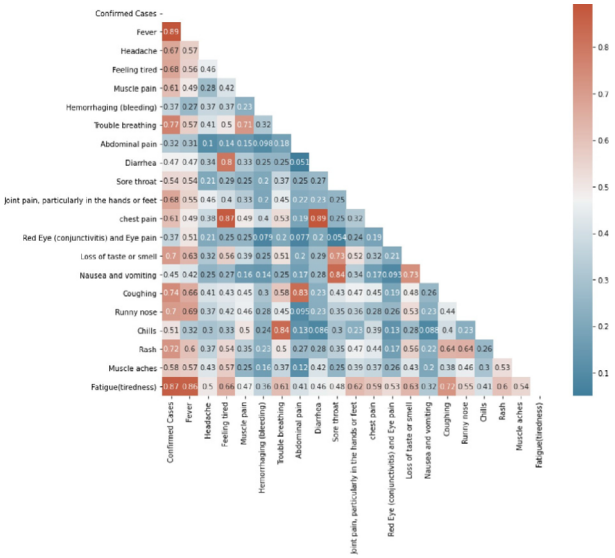


Fig. 3. Covid-19 correlation heatmap.

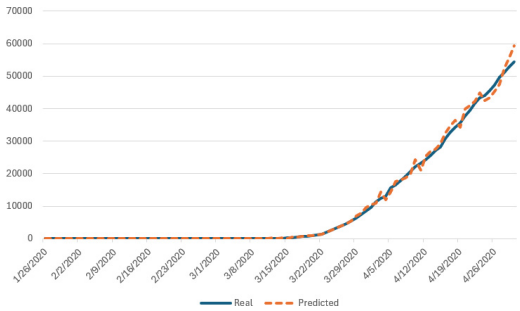


Fig. 4. Predicted and the real number of positive cases for Covid-19 Canada.

in forecasting future values. Furthermore, the models used in this study have various key hyperparameters that influence the forecasting accuracy of the obtained models.

C. Results

In this section, we share our analysis and results based on contributions 1 to 5. As the results for all six diseases were similar, we opted to visualize only one of the diseases in some of our experiments to maintain simplicity and accommodate space limitations. For the first step analysis, we conducted an analysis shown in Figure 3, the heatmaps depict the correlation between the number of daily tweets for each symptom or disease name and the number of daily confirmed cases for COVID-19 in the first datasets. It is evident that there is a high correlation between the daily confirmed case numbers and the symptoms related to Covid-19, such as fever (0.89), fatigue (0.87), and difficulty breathing (0.77), whereas some symptoms that are uncommon for COVID-19, such as abdominal pain (0.32) and red eyes (0.37) have a lower correlation. According to the results of our experiment, this pattern repeated for all six diseases, and there is a strong correlation between specific symptoms of each disease and daily case numbers.

In our first contribution, we utilized three datasets collected through various data collection methods, as described in Section IV. The algorithm's performance was evaluated for each dataset using two metrics, as outlined in Table III. The results indicate that the first dataset exhibited the highest performance. This outcome is attributed to the utilization of the ontology, which enables a broader coverage of keywords associated with symptoms and facilitates the extraction of more relevant tweets. This expanded coverage and enhanced tweet extraction are pivotal in improving the accuracy of the model.

In our second contribution, we presented the predicted number of positive cases generated by our hybrid model and compared them with the actual case numbers reported by Canadian health authorities, as depicted in Figure 4. The visualization reveals a significant alignment between the predicted and actual case numbers.

The third contribution is exemplified in Figure 5, which portrays our results and analysis contrasting the predicted number of positive cases generated by our hybrid model with the actual cases reported by health authorities for COVID-19 in the USA. Given our focus on producing a lucid graph for

Authorized licensed use limited to: Ontario Tech University. Downloaded on November 10,2026 at 19:29:46 UTC from IEEE Xplore. Restrictions apply.

TABLE III
PERFORMANCE OF PROPOSED HYBRID MODEL WITH DIFFERENT DATASETS

Models	Metrics	Ebola	Zika	MERS	H1N1	Chikungunya	Covid-19
Dataset 1	MAPE	**0.05**	**0.08**	**0.07**	**0.08**	**0.06**	**1.06**
	RMSE	**239.03**	**251.11**	**382.15**	**353.97**	**372.64**	**2347.19**
Dataset 2	MAPE	0.11	0.23	0.18	0.25	0.21	1.73
	RMSE	565.27	816.94	785.38	895.23	803.92	3964.89
Dataset 3	MAPE	2.05	2.37	2.58	2.42	2.39	4.03
	RMSE	4943.73	5744.31	6785.38	6695.23	6603.92	11064.52

TABLE IV
PERFORMANCE OF DIFFERENT MODELS

Models	Metrics	Ebola	Zika	MERS	H1N1	Chikungunya	Covid-19
VAR	MAPE	5.15	5.73	7.82	7.92	6.15	6.05
	RMSE	12440.72	14646.05	30891.35	28810.71	23823.89	24610.21
XGBoost	MAPE	2.59	1.52	1.60	1.61	2.11	3.13
	RMSE	6214.42	3744.56	3767.93	3762.85	5016.33	6750.61
Unidirectional LSTM	MAPE	2.01	2.25	3.73	4.22	6.12	6.67
	RMSE	4794.37	5664.51	8985.38	10165.27	14786.94	15760.96
Bi-directional LSTM	MAPE	1.44	1.02	1.63	2.74	4.75	5.10
	RMSE	3385.45	2388.06	3842.78	6554.61	11459.22	11840.43
VAR and LSTM	MAPE	1.84	1.32	2.74	3.90	4.62	5.19
	RMSE	4376.56	3208.35	6572.33	9386.06	11144.18	12066.63
NBEATS	MAPE	0.93	0.77	1.02	2.58	3.91	4.91
	RMSE	2095.14	1674.04	2303.63	6162.36	9417.53	11360.58
FaceBook Prophet	MAPE	1.04	1.13	2.04	2.74	4.07	5.62
	RMSE	2378.63	2692.62	4856.68	6554.68	9806.93	13146.68
Temporal Fusion Transformer or TFT (Google)	MAPE	0.80	0.88	0.75	1.07	3.42	3.23
	RMSE	1755.75	1993.93	1589.21	2387.32	8223.64	7018.07
DeepAR (Amazon)	MAPE	0.92	0.98	0.95	1.96	2.95	3.98
	RMSE	2070.37	2276.51	2121.36	4634.74	7076.81	8985.52
XGBoost and BiLSTM	MAPE	**0.03**	**0.06**	**0.05**	**0.06**	**0.06**	**1.02**
	RMSE	**237.02**	**217.95**	**275.19**	**279.47**	**375.70**	**2284.73**

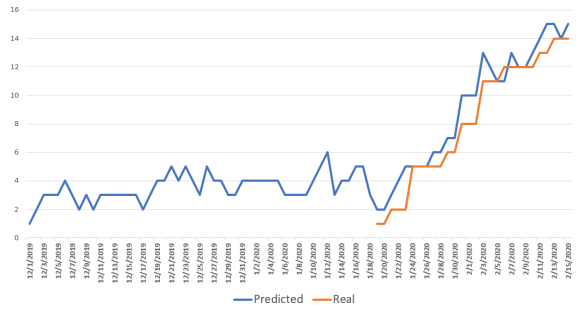


Fig. 5. Covid-19 first dataset heatmap.

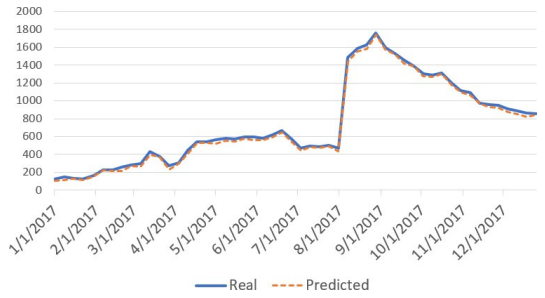


Fig. 6. Predicted and the real number of positive cases for Covid-19 USA.

comparison between predicted and reported case numbers by the CDC, we limited the visualization to a brief timeframe of approximately one month. As depicted in Figure 5, there is a remarkable alignment between the predicted and actual case counts, demonstrating the efficacy of the proposed hybrid model in identifying infected cases approximately four to six weeks earlier based on symptom-related tweets and Twitter conversations pertaining to each disease.

In our proposed method, constituting our fourth contribution, we aimed to pinpoint outbreaks in specific regions by leveraging data from prior occurrences. Our findings entailed training the model for H1N1 outbreaks in India during 2015, and through the analysis of tweets spanning from 2012 to 2022, we successfully identified an outbreak in India in 2017. Figure 6 illustrates the forecasted and actual counts of positive cases reported by the India NCDC [91] for the H1N1 outbreak in 2017. Given the NCDC's weekly reporting, we constructed the graph to display both predicted and actual values based on the weekly incidence of positive cases.

The fifth contribution and analysis revolve around the findings of the investigation into whether or not our proposed

Authorized licensed use limited to: Ontario Tech University. Downloaded on November 10,2026 at 19:29:46 UTC from IEEE Xplore. Restrictions apply.

method can improve forecasting accuracy over the baselines and state-of-the-art models shown in Table IV. We conducted our experiments using the data from the six previously mentioned diseases. All experiments are performed ten times, and the average performance measures are reported using the RMSE and MAPE metrics. In the presented models for each disease, the models use some window size as a number of previous days to predict the next 7 days' number of positive cases, for example, based on the behavior of our multivariate time series for COVID-19 the best window size is using the past 10 days value to predict the number of positive cases in the next 7 days. As shown in Table IV, our proposed hybrid model outperforms other models in both metrics.

Based on our results, XGBoost outperforms other deep learning models, excluding our proposed model, when the dataset size is small, as in the case of COVID-19. However, when the dataset size is larger, as in the case of Ebola, XGBoost outperforms fewer of them. This may be due to the regularisation property of XGBoost, which also prevents overfitting for small dataset sizes. Additionally, datasets with a longer duration may exhibit different behavior in terms of their population, geographical location of the country, lockdown rules, and other factors that affect the algorithm's performance. Our proposed model leverages XGBoost's ability to handle small datasets. Furthermore, by incorporating the initial forecasts generated by XGBoost into the training process, the BiLSTM can refine its predictions based on these insights. This refinement process allows the Bi-LSTM to adjust its predictions in light of the additional information provided by XGBoost, leading to more accurate forecasts.

VI. Conclusion

Twitter serves as an invaluable repository of information concerning epidemics and contagious diseases, sourced from online discussions. Harnessing the wealth of data available on this platform holds the potential to not only forecast future pandemics but also gain profound insights into ongoing outbreaks. This study underscores the pivotal role of social media in augmenting traditional surveillance methods, such as monitoring testing center data, hospital admissions, and contact tracing. Research has shown a significant correlation between social media discourse and epidemiological trends, enhancing the effectiveness of disease surveillance efforts.

In this investigation, we leveraged Twitter data spanning nearly a decade, from 2012 to 2021, focusing specifically on discussions related to symptoms associated with infectious diseases. This rich dataset formed the comerstone of our novel approach, where we developed a hybrid model meticulously designed to predict infection case numbers and proactively anticipate future outbreaks. By employing an ontology comprising a comprehensive array of infectious disease keywords, we meticulously identified and extracted relevant symptom data, laying the foundation for our predictive framework.

The results of our analysis unveiled the remarkable efficacy of the proposed hybrid model, outperforming nine state-of-theart baseline models across multiple key performance metrics. By harnessing the collective wisdom embedded within tweets pertaining to infectious disease symptoms, our model demonstrated a remarkable capability to detect outbreak onsets even before official reports from health authorities. Furthermore, it exhibited the potential to identify novel diseases that manifest symptoms akin to those of established diseases, thus broadening the scope of disease surveillance and detection.

These findings underscore the transformative potential of social media in revolutionizing existing disease surveillance methodologies. As the user base of social media platforms continues to expand exponentially, so does the wealth of data available for analysis. This exponential growth presents unprecedented opportunities for enhancing disease monitoring and response strategies, paving the way for more effective public health interventions and proactive measures.

In summary, our study serves as a testament to the immense promise held by social media in bolstering disease surveillance efforts. By seamlessly integrating cutting-edge methodologies with the vast troves of data available on platforms like Twitter, we stand poised to usher in a new era of precision public health, where proactive detection and timely response are paramount in safeguarding global health and well-being.

References

[1] K. Al-Surimi, M. Khalifa, S. Bahkali, A. El-Metwally, and M. Househ, "The potential of social media and internet-based data in preventing and fighting infectious diseases: From internet to Twitter," in *Emerging and Re-Emerging Viral Infections*. Berlin, Germany: Springer, 2016, pp. 131–139.

[2] G. Bello-Orgaz, J. Hernandez-Castro, and D. Camacho, "A survey of social web mining applications for disease outbreak detection," in *Intelligent Distributed Computing VIII* (Studies in Computational Intelligence), D. Camacho, L. Braubach, S. Venticinque, and C. Badica, Eds., Cham, Switzerland: Springer, 2015, pp. 345–356.

[3] M. A. Al-garadi, M. S. Khan, K. D. Varathan, G. Mujtaba, and A. M. Al-Kabsi, "Using online social networks to track a pandemic: A systematic review," *J. Biomed. Informat.*, vol. 62, pp. 1–11, Aug. 2016. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1532046416300351

[4] H. Hornmoen and C. McInnes, "Social media communication during disease outbreaks: Findings and recommendations," in *Social Media Use in Crisis and Risk Communication*, H. Hornmoen and K. Backholm, Eds., Bingley, U.K.: Emerald, Oct. 2018, pp. 255–275, doi: 10.1108/978-1-78756-269-120181016.

[5] S. Romano, S. Di Martino, N. Kanhabua, A. Mazzeo, and W. Nejdl, "Challenges in detecting epidemic outbreaks from social networks," in *Proc. 30th Int. Conf. Adv. Inf. Netw. Appl. Workshops (WAINA)*, Mar. 2016, pp. 69–74.

[6] L. E. Charles-Smith et al., "Using social media for actionable disease surveillance and outbreak management: A systematic literature review," *PLoS ONE*, vol. 10, no. 10, Oct. 2015, Art. no. e0139701, doi: 10.1371/journal.pone.0139701.

[7] S. E. Jordan, S. E. Hovet, I. C.-H. Fung, H. Liang, K.-W. Fu, and Z. T. H. Tse, "Using Twitter for public health surveillance from monitoring and prediction to public response," *Data*, vol. 4, no. 1, p. 6, Dec. 2018.

[8] H. Rosenberg, S. Syed, and S. Rezaie, "The Twitter pandemic: The critical role of Twitter in the dissemination of medical information and misinformation during the COVID-19 pandemic," *Can. J. Emergency Med.*, vol. 22, no. 4, pp. 418–421, Jul. 2020.

[9] L. Samaras, E. García-Barriocanal, and M.-A. Sicilia, "Comparing social media and Google to detect and predict severe epidemics," *Sci. Rep.*, vol. 10, no. 1, p. 4747, Mar. 2020. [Online]. Available: http://www.nature.com/articles/s41598-020-61686-9

[10] S. Amin, M. I. Uddin, D. H. Al Saeed, A. Khan, and M. Adnan, "Early detection of seasonal outbreaks from Twitter data using machine learning approaches," *Complexity*, vol. 2021, no. 1, Mar. 2021, Art. no. e5520366. [Online]. Available: https://www.hindawi.com/journals/complexity/2021/5520366/

Authorized licensed use limited to: Ontario Tech University. Downloaded on November 10,2026 at 19:29:46 UTC from IEEE Xplore. Restrictions apply.

[11] N. N. Fakhry, E. Asfoura, and G. Kassam, "Tracking coronavirus pandemic diseases using social media: A machine learning approach," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 10, 2020, doi: 10.14569/ijacsa.2020.0111028.

[12] S. Lakamana, Y.-C. Yang, M. A. Al-Garadi, and A. Sarker, "Tracking the COVID-19 outbreak in India through Twitter: Opportunities for social media based global pandemic surveillance," in *Proc. AMIA Annu. Symp.*, May 2022, pp. 313–322. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9285154/

[13] A. Joshi et al., "Automated monitoring of tweets for early detection of the 2014 Ebola epidemic," *PLoS ONE*, vol. 15, no. 3, Mar. 2020, Art. no. e0230322. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0230322

[14] V. K. Jain and S. Kumar, "An effective approach to track levels of Influenza-A (H1N1) pandemic in India using Twitter," *Proc. Comput. Sci.*, vol. 70, pp. 801–807, Jan. 2015. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1877050915032846

[15] A. Alessa and M. Faezipour, "A review of influenza detection and prediction through social networking sites," *Theor. Biol. Med. Model.*, vol. 15, no. 1, p. 2, Feb. 2018, doi: 10.1186/s12976-017-0074-5.

[16] CDC. (Feb. 2020). *Coronavirus Disease 2019 (COVID-19)*. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/science/about-epidemiology/identifying-source-outbreak.html

[17] J. K. Miller et al., "Statistical outbreak detection by joining medical records and pathogen similarity," *J. Biomed. Informat.*, vol. 91, Mar. 2019, Art. no. 103126.

[18] I. Cooper, A. Mondal, and C. G. Antonopoulos, "A SIR model assumption for the spread of COVID-19 in different communities," *Chaos, Solitons Fractals*, vol. 139, Oct. 2020, Art. no. 110057.

[19] M. O. Alassafi, M. Jarrah, and R. Alotaibi, "Time series predicting of COVID-19 based on deep learning," *Neurocomputing*, vol. 468, pp. 335–344, Jan. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231221015150

[20] F. Shahid, A. Zameer, and M. Muneeb, "Predictions for COVID-19 with deep learning models of LSTM, GRU and bi-LSTM," *Chaos, Solitons Fractals*, vol. 140, Nov. 2020, Art. no. 110212. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960077920306081

[21] A. B. Said, A. Erradi, H. A. Aly, and A. Mohamed, "Predicting COVID-19 cases using bidirectional LSTM on multivariate time series," *Environ. Sci. Pollut. Res.*, vol. 28, no. 40, pp. 56043–56052, Oct. 2021.

[22] E. Afzali, A. Adegoke, Z. Jin, W. Qiu, and L. Wang. (2020). *Hybrid VAR-LSTM Networks Modeling and Forecasting COVID-19 Data in Canada*. [Online]. Available: https://ssc.ca/en/about/sections-regions/biostatistics/covid-19-student-case-study-competition

[23] E. Yang et al., "A simulation-based study on the comparison of statistical and time series forecasting methods for early detection of infectious disease outbreaks," *Int. J. Environ. Res. Public Health*, vol. 15, no. 5, p. 966, May 2018. [Online]. Available: http://www.mdpi.com/1660-4601/15/5/966

[24] H. Abbasimehr, R. Paki, and A. Bahrini, "A novel approach based on combining deep learning models with statistical methods for COVID-19 time series forecasting," *Neural Comput. Appl.*, vol. 34, no. 4, pp. 3135–3149, Feb. 2022, doi: 10.1007/s00521-021-06548-9.

[25] V. Papastefanopoulos, P. Linardatos, and S. Kotsiantis, "COVID-19: A comparison of time series methods to forecast percentage of active cases per population," *Appl. Sci.*, vol. 10, no. 11, p. 3880, Jun. 2020. [Online]. Available: https://www.mdpi.com/2076-3417/10/11/3880

[26] V. K. R. Chimmula and L. Zhang, "Time series forecasting of COVID-19 transmission in Canada using LSTM networks," *Chaos, Solitons Fractals*, vol. 135, Jun. 2020, Art. no. 109864. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0960077920302642

[27] Y. Zheng et al., "A learning-based model to evaluate hospitalization priority in COVID-19 pandemics," *Patterns*, vol. 1, no. 6, Sep. 2020, Art. no. 100092. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666389920301203

[28] C.-A. Hu et al., "Using a machine learning approach to predict mortality in critically ill influenza patients: A cross-sectional retrospective multicentre study in Taiwan," *BMJ Open*, vol. 10, no. 2, Feb. 2020, Art. no. e033898, doi: 10.1136/bmjopen-2019-033898.

[29] Z.-G. Fang, S.-Q. Yang, C.-X. Lv, S.-Y. An, and W. Wu, "Application of a data-driven XGBoost model for the prediction of COVID-19 in the USA: A time-series study," *BMJ Open*, vol. 12, no. 7, Jul. 2022, Art. no. e056685. [Online]. Available: https://bmjopen.bmj.com/content/12/7/e056685

[30] K. Guo et al., "Traffic data-empowered XGBoost-LSTM framework for infectious disease prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 1, pp. 774–785, Jan. 2024.

[31] J. Luo, Z. Zhang, Y. Fu, and F. Rao, "Time series prediction of COVID-19 transmission in America using LSTM and XGBoost algorithms," *Results Phys.*, vol. 27, Aug. 2021, Art. no. 104462. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2211379721005775

[32] Z. Hu, Q. Ge, S. Li, L. Jin, and M. Xiong, "Artificial intelligence forecasting of COVID-19 in China," 2020, *arXiv:2002.07112*.

[33] T. Zeng, Y. Zhang, Z. Li, X. Liu, and B. Qiu, "Predictions of 2019-nCoV transmission ending via comprehensive methods," 2020, *arXiv:2002.04945*.

[34] S. J. Fong, G. Li, N. Dey, R. Gonzalez-Crespo, and E. Herrera-Viedma, "Finding an accurate early forecasting model from small dataset: A case of 2019-nCoV novel coronavirus outbreak," *Int. J. Interact. Multimedia Artif. Intell.*, vol. 6, no. 1, p. 132, 2020.

[35] J. Li, Q. Dai, and R. Ye, "A novel double incremental learning algorithm for time series prediction," *Neural Comput. Appl.*, vol. 31, no. 10, pp. 6055–6077, Oct. 2019, doi: 10.1007/s00521-018-3434-0.

[36] J. Zheng, X. Fu, and G. Zhang, "Research on exchange rate forecasting based on deep belief network," *Neural Comput. Appl.*, vol. 31, no. 1, pp. 573–582, Jan. 2019, doi: 10.1007/s00521-017-3039-z.

[37] W. Zou and Y. Xia, "Back propagation bidirectional extreme learning machine for traffic flow time series prediction," *Neural Comput. Appl.*, vol. 31, no. 11, pp. 7401–7414, Nov. 2019, doi: 10.1007/s00521-018-3578-y.

[38] D. M. Giustini, S. M. Ali, M. Fraser, and M. N. K. Boulos, "Effective uses of social media in public health and medicine: A systematic review of systematic reviews," *Online J. Public Health Informat.*, vol. 10, no. 2, Sep. 2018. [Online]. Available: https://ojphi.org/ojs/index.php/ojphi/article/view/8270

[39] L. A. Ellis, P. Collin, T. A. Davenport, P. J. Hurley, J. M. Burns, and I. B. Hickie, "Young men, mental health, and technology: Implications for service design and delivery in the digital age," *J. Med. Internet Res.*, vol. 14, no. 6, p. 160, Nov. 2012. [Online]. Available: https://www.jmir.org/2012/6/e160

[40] Y. T. Yang, M. Horneffer, and N. DiLisio, "Mining social media and web searches for disease detection," *J. Public Health Res.*, vol. 2, no. 1, Mar. 2013, Art. no. jphr.2013, doi: 10.4081/jphr.2013.e4.

[41] O. Edo-Osagie, B. De La Iglesia, I. Lake, and O. Edeghere, "A scoping review of the use of Twitter for public health research," *Comput. Biol. Med.*, vol. 122, Jul. 2020, Art. no. 103770. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0010482520301426

[42] J. Kim and I. Ahn, "Infectious disease outbreak prediction using media articles with machine learning models," *Sci. Rep.*, vol. 11, no. 1, p. 4413, Feb. 2021. [Online]. Available: https://www.nature.com/articles/s41598-021-83926-2

[43] Z. M. Zain and N. M. Alturki, "COVID-19 pandemic forecasting using CNN-LSTM: A hybrid approach," *J. Control Sci. Eng.*, vol. 2021, pp. 1–23, Jul. 2021. [Online]. Available: https://www.hindawi.com/journals/jcse/2021/8785636/

[44] S. Sah, B. Surendiran, R. Dhanalakshmi, S. N. Mohanty, F. Alenezi, and K. Polat, "Forecasting COVID-19 pandemic using prophet, ARIMA, and hybrid stacked LSTM-GRU models in India," *Comput. Math. Methods Med.*, vol. 2022, pp. 1–19, May 2022. [Online]. Available: https://www.hindawi.com/journals/cmmm/2022/1556025/

[45] C. A. Sims, "Macroeconomics and reality," *Econometrica*, vol. 48, no. 1, pp. 1–48, Jan. 1980. [Online]. Available: https://www.jstor.org/stable/1912017

[46] K. Goto, B. Kumarendran, S. Mettananda, D. Gunasekara, Y. Fujii, and S. Kaneko, "Analysis of effects of meteorological factors on dengue incidence in Sri Lanka using time series data," *PLoS ONE*, vol. 8, no. 5, May 2013, Art. no. e63717. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0063717

[47] Q. Wang, Y. Zhou, and X. Chen, "A vector autoregression prediction model for COVID-19 outbreak," 2021, *arXiv:2102.04843*.

[48] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*. New York, NY, USA: Assoc. for Comput. Machinery, Aug. 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[49] P. Wang, X. Zheng, G. Ai, D. Liu, and B. Zhu, "Time series prediction for the epidemic trends of COVID-19 using the improved LSTM deep learning method: Case studies in Russia, Peru and Iran," *Chaos, Solitons Fractals*, vol. 140, Nov. 2020, Art. no. 110214. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S096007792030610X

Authorized licensed use limited to: Ontario Tech University. Downloaded on November 10,2026 at 19:29:46 UTC from IEEE Xplore. Restrictions apply.

[50] C.-X. Lv, S.-Y. An, B.-J. Qiao, and W. Wu, "Time series analysis of hemorrhagic fever with renal syndrome in Mainland China by using an XGBoost forecasting model," *BMC Infectious Diseases*, vol. 21, pp. 1–13, Dec. 2021, doi: 10.1186/s12879-021-06503-y.

[51] M. Alim, G.-H. Ye, P. Guan, D.-S. Huang, B.-S. Zhou, and W. Wu, "Comparison of ARIMA model and XGBoost model for prediction of human brucellosis in Mainland China: A time-series study," *BMJ Open*, vol. 10, no. 12, Dec. 2020, Art. no. e039676. [Online]. Available: https://bmjopen.bmj.com/content/10/12/e039676

[52] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

[53] P. Baldi, S. Brunak, P. Frasconi, G. Soda, and G. Pollastri, "Exploiting the past and the future in protein secondary structure prediction," *Bioinformatics*, vol. 15, no. 11, pp. 937–946, Nov. 1999, doi: 10.1093/bioinformatics/15.11.937.

[54] P. Arora, H. Kumar, and B. K. Panigrahi, "Prediction and analysis of COVID-19 positive cases using deep learning models: A descriptive case study of India," *Chaos, Solitons Fractals*, vol. 139, Oct. 2020, Art. no. 110017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S096007792030415X

[55] N. Hasan, "A methodological approach for predicting COVID-19 epidemic using EEMD-ANN hybrid model," *Internet Things*, vol. 11, Sep. 2020, Art. no. 100228. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2542660520300615

[56] B. Lim, S. Ö. Arik, N. Loeff, and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," *Int. J. Forecasting*, vol. 37, no. 4, pp. 1748–1764, Oct. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169207021000637

[57] F. M. Caldas and C. Soares, "A temporal fusion transformer for long-term explainable prediction of emergency department overcrowding," 2022, *arXiv:2207.00610*.

[58] B. Zhou, G. Yang, Z. Shi, and S. Ma, "Interpretable temporal attention network for COVID-19 forecasting," *Appl. Soft Comput.*, vol. 120, May 2022, Art. no. 108691. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1568494622001545

[59] S. J. Taylor and B. Letham, "Forecasting at scale," *Amer. Statistician*, vol. 72, no. 1, pp. 37–45, Jan. 2018, doi: 10.1080/00031305.2017.1380080.

[60] A. C. Harvey and S. Peters, "Estimation procedures for structural time series models," *J. Forecasting*, vol. 9, no. 2, pp. 89–108, Mar. 1990, doi: 10.1002/for.3980090203.

[61] G. E. Hutchinson, *Introduction to Population Ecology*. New Haven, CT, USA: Yale Univ. Press, 1978.

[62] A. C. Harvey and N. G. Shephard. (1993). *Estimation and Testing of Stochastic Variance Models*. STICERD—Econometrics Paper Series. [Online]. Available: https://ideas.repec.org/p/cep/stiecm/268.html

[63] D. Salinas, V. Flunkert, J. Gasthaus, and T. Januschowski, "DeepAR: Probabilistic forecasting with autoregressive recurrent networks," *Int. J. Forecasting*, vol. 36, no. 3, pp. 1181–1191, Jul. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169207019301888

[64] M. A. Shaik and D. Verma, "Deep learning time series to forecast COVID-19 active cases in INDIA: A comparative study," *IOP Conf. Ser., Mater. Sci. Eng.*, vol. 981, no. 2, Dec. 2020, Art. no. 022041.

[65] P. Arunkumar and L. K. Ramasamy, "Time-series forecasting and analysis of COVID-19 outbreak in highly populated countries: A data-driven approach," *Int. J. E-Health Med. Commun. (IJEHMC)*, vol. 13, no. 2, pp. 1–17, 2021.

[66] R. B. Cleveland, W. S. Cleveland, J. E. McRae, and I. Terpenning, "STL: A seasonal-trend decomposition," *J. Off. Statistic*, vol. 6, no. 1, pp. 3–73, 1990.

[67] B. N. Oreshkin, D. Carpov, N. Chapados, and Y. Bengio, "N-BEATS: Neural basis expansion analysis for interpretable time series forecasting," 2019, *arXiv:1905.10437*.

[68] A. Sulasikin, Y. Nugraha, J. Kanggrawan, and A. L. Suherman, "Forecasting for a data-driven policy using time series methods in handling COVID-19 pandemic in Jakarta," in *Proc. IEEE Int. Smart Cities Conf. (ISC)*, Sep. 2020, pp. 1–6.

[69] S. O. Folorunso, J. B. Awotunde, O. O. Banjo, E. A. Ogundepo, and O. Adeboye, "Comparison of active COVID-19 cases per population using time-series models," *Int. J. E-Health Med. Commun. (IJEHMC)*, vol. 13, no. 2, pp. 1–21, 2021.

[70] D. Cao et al., "Spectral temporal graph neural network for multivariate time-series forecasting," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 17766–17778.

[71] X. Zhang et al., "Novel soft computing model for predicting blast-induced ground vibration in open-pit mines based on particle swarm optimization and XGBoost," *Natural Resour. Res.*, vol. 29, no. 2, pp. 711–721, Apr. 2020.

[72] W. Li, Y. Yin, X. Quan, and H. Zhang, "Gene expression value prediction based on XGBoost algorithm," *Frontiers Genet.*, vol. 10, p. 1077, Nov. 2019.

[73] L. Breiman, *Classification and Regression Trees*. Evanston, IL, USA: Routledge, 2017.

[74] R. Yu et al., "LSTM-EFG for wind power forecasting based on sequential correlation features," *Future Gener. Comput. Syst.*, vol. 93, pp. 33–42, Apr. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167739X18314420

[75] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[76] W. Zaremba, I. Sutskever, and O. Vinyals, "Recurrent neural network regularization," 2014, *arXiv:1409.2329*.

[77] D. Ontology. *Disease Ontology—Institute for Genome Sciences*. University of Maryland. [Online]. Available: https://disease-ontology.org/

[78] L. Mhadhbi and J. Akaichi, "DS-ontology: A disease-symptom ontology for general diagnosis enhancement," in *Proc. Int. Conf. Inf. Syst. Data Mining*, Apr. 2017, pp. 99–102.

[79] C. W. 24/7. (Nov. 2022). *CDC Works 24/7*. [Online]. Available: https://www.cdc.gov/index.htm

[80] *World Health Organization (WHO)*. [Online]. Available: https://www.who.int

[81] P. R. Text. *Processing Raw Text*. [Online]. Available: https://www.nltk.org/book/ch03.html

[82] N. Babanejad, H. Davoudi, A. An, and M. Papagelis, "Affective and contextual embedding for sarcasm detection," in *Proc. 28th Int. Conf. Comput. Linguistics*, Dec. 2020, pp. 225–243. [Online]. Available: https://aclanthology.org/2020.coling-main.20

[83] N. Babanejad, A. Agrawal, A. An, and M. Papagelis, "A comprehensive analysis of preprocessing for word representation learning in affective tasks," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 5799–5810. [Online]. Available: https://aclanthology.org/2020.acl-main.514

[84] Z. Li et al., "Measuring global multi-scale place connectivity using geotagged social media data," *Sci. Rep.*, vol. 11, no. 1, p. 14694, Jul. 2021. [Online]. Available: https://www.nature.com/articles/s41598-021-94300-7

[85] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[86] J. Wu, X.-Y. Chen, H. Zhang, L.-D. Xiong, H. Lei, and S.-H. Deng, "Hyperparameter optimization for machine learning models based on Bayesian optimizationb," *J. Electron. Sci. Technol.*, vol. 17, no. 1, pp. 26–40, Mar. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1674862X19300047

[87] E. Brochu, V. M. Cora, and N. de Freitas, "A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning," 2010, *arXiv:1012.2599*.

[88] R. Calandra, A. Seyfarth, J. Peters, and M. P. Deisenroth, "Bayesian optimization for learning gaits under uncertainty," *Ann. Math. Artif. Intell.*, vol. 76, nos. 1–2, pp. 5–23, Feb. 2016, doi: 10.1007/s10472-015-9463-9.

[89] H. Abbasimehr and R. Paki, "Improving time series forecasting using LSTM and attention models," *J. Ambient Intell. Humanized Comput.*, vol. 13, no. 1, pp. 673–691, Jan. 2022, doi: 10.1007/s12652-020-02761-x.

[90] K. Bandara, C. Bergmeir, and S. Smyl, "Forecasting across time series databases using recurrent neural networks on groups of similar series: A clustering approach," *Exp. Syst. Appl.*, vol. 140, Feb. 2020, Art. no. 112896. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417419306128

[91] NCDC. *Home: National Centre for Disease Control (NCDC)*. [Online]. Available: https://ncdc.gov.in/

Authorized licensed use limited to: Ontario Tech University. Downloaded on November 10,2026 at 19:29:46 UTC from IEEE Xplore. Restrictions apply.