A Hybrid Outbreak Detection using Ontology-based Data Collection from Social Media

Ghazaleh Babanejaddehaki
Department of Electrical Engineering and Computer Science
York University
Toronto, Canada
ghazalba@yorku.ca

Aijun An
Department of Electrical Engineering and Computer Science
York University
Toronto, Canada
aan@yorku.ca

Heidar Davoudi
Faculty of Science
Ontario Tech University
Toronto, Canada
heidar.davoudi@ontariotechu.ca

2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3503-3748-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/BIBM58861.2023.10385809

***Abstract*—Rapidly spreading diseases pose a significant threat, leading to substantial loss of life and economic devastation, as seen in the global COVID-19 outbreaks. Developing disease prediction models is crucial for preemptive pandemic control and minimizing their impact. As internet accessibility grows through computers and mobile devices, social media platforms provide a direct conduit to disseminate vital health information to the public. Unlike traditional methods that rely on bureaucratic channels, these platforms offer accurate and timely information distribution. We propose a framework that employs ontology to identify these symptoms and gather relevant tweets. Subsequently, the XGBoost-BiLSTM hybrid model harnesses this data to predict the count of infected cases. This hybrid model capitalizes on XGBoost's prowess in handling limited dataset sizes, a prevalent challenge during outbreaks with insufficient time series data. Moreover, it enriches data for BiLSTM, amplifying its efficacy in predicting and monitoring outbreaks. To construct our dataset, we extracted tweets discussing symptoms from six distinct infectious disease outbreaks (Ebola, Zika, MERS, H1N1, Chikungunya, COVID-19) spanning from 2012 to 2021. Our results demonstrate that the proposed hybrid model outperforms nine cutting-edge and baseline models. This advancement can significantly assist health authorities in minimizing fatalities and preparing preemptively for potential outbreaks.**

***Index Terms*—Outbreak detection, Machine learning, Deep learning, Time series, Ontology, Social media**

I. Introduction

Social media plays a crucial role in disease monitoring and prediction, enabling the integration of low-cost, real-time health information sources with conventional surveillance systems [1]. These platforms provide quick and affordable channels for unreported health data and health-related opinions. Among them, Twitter, due to its real-time nature, descriptive posts, and diverse user base, has garnered attention in various research fields, including the analysis of public responses during outbreaks [2]. Research has indicated that integrating social media data with conventional surveillance information can significantly improve outbreak detection and prevention [2], [3]. However, most existing methods have primarily relied on disease names or common symptoms as keywords for data collection. The emergence of COVID-19 underscored the importance of capturing a more comprehensive array of symptoms. In light of this, we have introduced a framework that exclusively relies on symptom-related tweets, allowing for seamless adaptation to evolving situations without being constrained by predetermined disease names. Our approach leverages an *ontology knowledge base* that provides an extensive range of symptom names, thereby enhancing the accuracy of outbreak detection by encompassing a diverse semantic spectrum of symptom descriptions. The challenge of outbreak detection has been explored from various angles. Time Series Analysis (TSA) stands as a prevalent approach in outbreak detection and prediction, utilizing historical data to anticipate current and future scenarios [4], [5]. Both univariate and multivariate temporal analyses employ statistical [4], deep [5], and shallow [2] models to forecast new cases and fatalities. Another commonly used statistical technique, regression analysis, is often employed to model the spread of diseases. While deep learning approaches can capture intricate dynamics of infectious diseases and surpass shallow models [6], they are hindered by the requirement for extensive training datasets, which might be lacking for emerging diseases. In this paper, we introduce a hybrid model that leverages the advantages of boosting models, specifically XGBoost, renowned for handling small training datasets [7], for training a deep learning model.

Specifically, we harness Twitter data extracted through an ontology of disease symptoms to establish a multivariate time series dataset for the number of disease case predictions. Subsequently, we introduce a hybrid model that seamlessly integrates the power of XGBoost [7] and Bi-LSTM [8]. The initial phase involves the utilization of XGBoost for data augmentation, succeeded by the subsequent phase that employs Bi-LSTM to further elevate prediction precision and identify the early stages of outbreaks. The main idea is to augment training data with the XGboot model prediction to facilitate BiLSTM model training. The main contributions of this study are as follows:

- By utilizing a disease ontology, we identify symptoms of infectious diseases and extract corresponding data from Twitter.
- Our novel hybrid model predicts the number of infected

Authorized licensed use limited to: Ontario Tech University. Downloaded on November 10,2026 at 20:17:49 UTC from IEEE Xplore. Restrictions apply.
979-8-3503-3748-8/23/$31.00 ©2023 IEEE

cases using tweets related to infectious disease symptoms, providing earlier insights than official health authorities' reports.
- We compile, refine, and preprocess a dataset to serve as a foundation for future research endeavors.
- Our approach surpasses the performance of nine alternative machine learning methods employed for predicting infectious disease cases.

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 provides a brief overview of research related to outbreak detection and the utilization of machine-learning algorithms for predicting infectious disease cases and deaths. Social media is increasingly being explored by researchers and medical professionals for various purposes, including professional development, networking, health campaigns, and disease monitoring [9]. Review articles [9], [10] have examined social media's role in disease surveillance and highlighted its advantages in terms of speed, accuracy, and cost-effectiveness. These studies recommend integrating social media with existing surveillance systems and discussing its impact on health protection, public health messaging, and lessons learned from health campaigns. In our study, we compare the performance of our proposed method with several baselines and state-of-the-art models. We are going to mention some of the recent works that have used these models. The vector autoregressive (VAR) [11] model has proven to be especially useful for forecasting and describing the dynamic behavior of economic and financial time series. To predict positive COVID-19 cases, Wang et al. [12] proposed a validated Vector Autoregression (VAR) time series model.

XGBoost (eXtreme Gradient Boosting) [7] is commonly used for feature selection. Lv et al. [13] 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. [14] 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. [15] and Fang et al. [16] 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. [17] 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 is an extension of the previously described LSTM models that apply two LSTMs to the input data [18]. For predicting the number of Covid-19 infected cases there are research works e.g. in India, Arura et al. [19] employed recurrent neural network (RNN)-based variants such as deep LSTM, convolutional LSTM, and Bi-LSTM. Wang et al. [20] used LSTM networks and rolling updating mechanisms to incorporate new forecasting outcomes into model training for the subsequent iteration of different countries.

Temporal Fusion Transformer (TFT) [21] is an attention-based, performance- and interpretability-optimized Deep Neural Network. Caldas et al. [22] utilized TFT to forecast the volume of non-urgent patients in Portuguese public hospitals by Health Regional Areas (HRA). Zhou et al. [23] 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. The approach makes use of a simple time-series model with three key elements: trend, seasonality, and holidays. It is also suitable for single-variable time series. Prophet uses a piecewise linear model and a saturating growth model to predict trends. Recent studies by Sah et al. [24] 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 [25]. There are some recent works using DeppAR to predict the active cases [25], [26]. N-Beats model functions generally in a similar manner to conventional decomposition methods, like the seasonality-trend-level approach [27] and the recent works that used N-Beats for Covid19 outbreak detection [25], [28].

III. Preliminaries

In this section, we present brief definitions of the two models that are used throughout the proposed hybrid method.

A. Extreme Gradient Boosting(XGBoost):

XGBoost [7], an advancement of the gradient boosting decision tree (GBDT) algorithm, excels in speed, accuracy, and unique information processing. It outperforms traditional statistical models in predicting nonlinear data [15]. As a boosting algorithm, XGBoost sequentially combines multiple

Authorized licensed use limited to: Ontario Tech University. Downloaded on November 10,2026 at 20:17:49 UTC from IEEE Xplore. Restrictions apply.

weak learners, such as decision trees, to form a robust predictive model. Unlike conventional GBDT, XGBoost employs second-order Taylor expansion for the loss function, includes regularization for better generalization, and allows custom loss functions [29]. Finally, XGBoost borrows from random forest concepts, introducing row and column sampling to control computation and overfitting. These features make XBoost appropriate for the proposed hybris architect, handling small training data. Refering to the phase 1 of the Figure 1, XGBoost aims to reduce the regularized objective in the following manner:

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

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 Long Short-Term Memory (Bi-LSTM):

Recurrent Neural Networks (RNNs) demonstrate exceptional proficiency in handling sequential time series and temporal dependencies. Long Short-Term Memory (LSTM) networks addressing challenges posed by long-term dependencies by employing memory cells - concealed layer units introduced by Hochreiter and Schmidhuber [18]. These memory cells incorporate input, output, and forget gates, enabling the management of network state and temporal aspects [30]. The input and output gates regulate the flow of data into and out of memory cells, with the forget gate facilitating selective data retention. Activated data finds its storage within memory cells, contingent upon the activation of input or output units. While standard RNNs lack foresight into future data, Bidirectional LSTM (Bi-LSTM) rectifies this by considering both past and future information. Through its dual forward and backward states, Bi-LSTM captures a comprehensive panorama of temporal details over time.

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

$$h_i = LSTM(h_i, \leftarrow) \oplus LSTM(h_i, \rightarrow) \qquad (2)$$

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.

IV. Proposed Method

The proposed methodology comprises three parts: *data collection*, *data preparation*, and the *hybrid machine-learning model*. Figure 1 illustrates the steps for each component in detecting six infectious diseases. Each part is elaborated on in the subsequent subsections.

A. Data Collection:

This research presents a novel approach by utilizing daily tweet frequencies of infectious disease symptoms for outbreak detection, instead of relying solely on disease names. This approach proves particularly advantageous for emerging diseases like COVID-19, where early stages often lack well-defined names. Additionally, we incorporate an ontology, a structured representation capturing disease symptom relationships, which enhances data management, retrieval, and reasoning. The ontology offers access to a comprehensive symptom list, featuring semantic names like "fatigue" and "tiredness". For this study, we concentrate on a disease ontology [31], using SPARQL Query to extract symptoms subclasses linked to diseases through the *has_symptom* relationship as depicted in Figure 2. Within the timeframe of 2012–2021, we analyze six major infectious diseases (Ebola, Zika, MERS, H1N1, Chikungunya, COVID-19), each associated with 65 identified symptoms. These symptoms serve as keywords for tweet collection, retrieved from Twitter, with a requirement of containing at least one specified keyword. Concurrently, daily case counts for each disease are obtained from health authorities or CDC [32], [33]. We evaluate the proposed model's efficacy through experiments involving three sets of datasets. The first set employs the *ontology* to extract tweets referencing infectious disease symptoms (65 keywords). We also compile two Twitter datasets: one extacting tweets based on *disease names* and *common symptoms* cited by health authorities (18 keywords), and another solely based on *disease names* (6 keywords).

B. Data Preparation:

After obtaining the tweets, initial data preparation is critical. In the primary phase, we employ natural language processing techniques [34] to eliminate stop words, sparse keywords, and non-informative terms. Punctuation and diacritics are removed, and tweet text is tokenized accordingly. The subsequent phase focuses on filtering out irrelevant and noisy tweets related to illnesses. Strategies outlined in [35] are adopted to sift through tweets unrelated to the disease and to detect sarcasm (as we collected data using specific keywords like "vomit," we determine the relevance of a tweet to a disease based on these keywords. However, we also need to exclude tweets containing these keywords that may be sarcastic, as they could skew the model's accuracy. Therefore, we identify and remove such sarcastic tweets from the dataset.

The final step involves converting cleaned disease-related tweets into a time series dataset, recording daily tweet counts mentioning keywords by country. The initial dataset may contain synonyms for symptoms (e.g., "fatigue," "tiredness," or "exhaustion" from the ontology). These synonyms are aggregated for simplicity, and reducing complexity. For instance, daily tweets about "fatigue," "tiredness," and "exhaustion" are

Authorized licensed use limited to: Ontario Tech University. Downloaded on November 10,2026 at 20:17:49 UTC from IEEE Xplore. Restrictions apply.

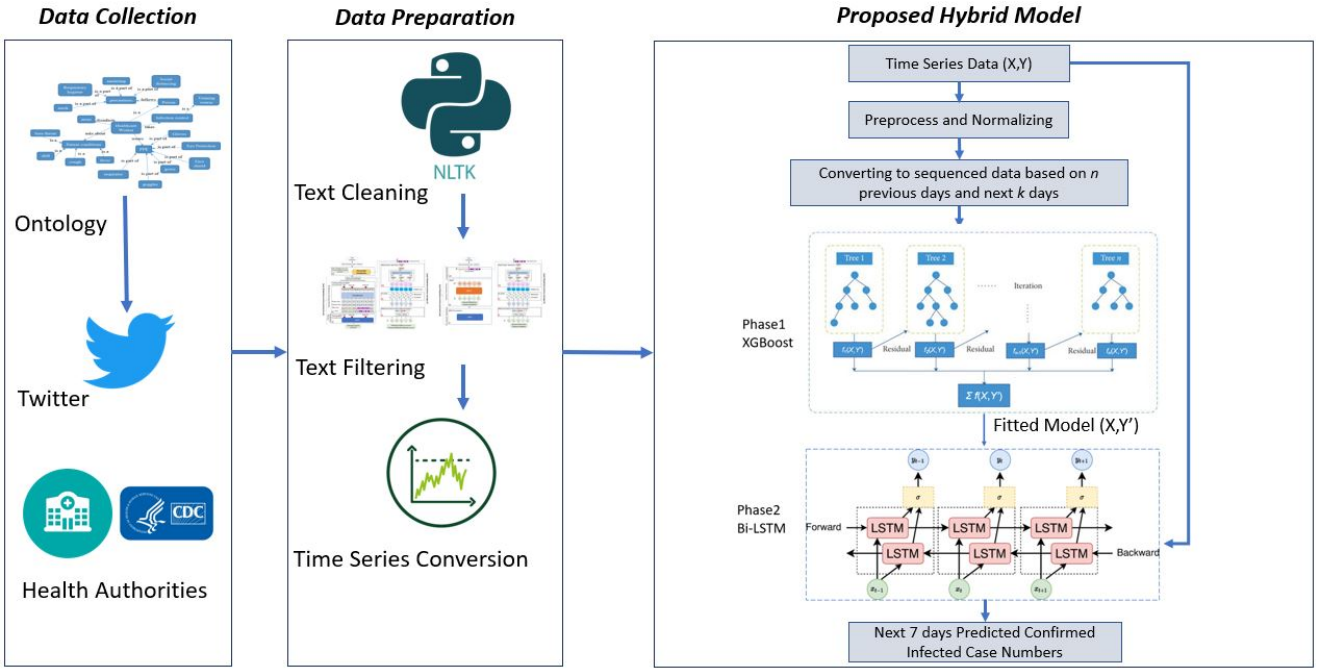


Fig. 1. Proposed outbreak detection method steps

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

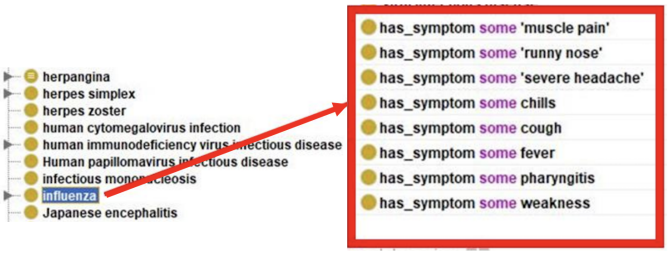


Fig. 2. Viral infectious disease Ontology

grouped as "fatigue." The dataset comprises a time series of daily symptom-related tweet counts as independent variables, and confirmed infected case numbers as the dependent variable.

C. Proposed Hybrid Model:

We propose a hybrid model that integrates both XGBoost and Bidirectional LSTM (Bi-LSTM) to predict confirmed infected cases. XGBoost plays a pivotal role in the second phase by providing a data augmentation mechanism for training the Bi-LSTM model. This approach addresses the challenge of limited data during the training process. Specifically, we enhance the original small dataset by incorporating XGBoost predictions, thereby facilitating the training of the Bi-LSTM model. The subsequent section provides the details of the data augmentation process.

1) Hybrid XGBoost–BiLSTM Proposed Method: In this section, we describe the operation of our proposed hybrid model, taking into account the characteristics of XGBoost and Bi-LSTM explained in previous sections. Figure 1 illustrates the architecture of the proposed framework, which consists of two phases.

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 data set 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

Authorized licensed use limited to: Ontario Tech University. Downloaded on November 10,2026 at 20:17:49 UTC from IEEE Xplore. Restrictions apply.

dependent variable indicating confirmed cases from time t to $t+k-1$. This data set 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.

In the subsequent phase, the trained XGBoost model utilizes the original independent variables (X) to predict the number of confirmed cases in the next k days, denoted by (Y'). Subsequently, we employ the results obtained from the XGBoost model (X, Y') to feed and train the Bi-LSTM model to predict (Y), enabling multi-step ahead forecasts. This process enhances the hybrid model's ability to predict future numbers of infected cases and detect the onset of outbreaks earlier. By combining these two phases in our hybrid model, we leverage the strengths of XGBoost and Bi-LSTM to enhance the accuracy of forecasting and the early detection of outbreaks for infected cases. While XGBoost learns the intrinsic behavior of our multivariate dataset and handles the small dataset size, Bi-LSTM captures the temporal dependencies among independent variables.

V. Experimental Evaluation

This study introduces a novel method for predicting positive cases using disease-related tweets. Implementation utilized Python, Twitter API, and Google Colaboratory. Experiments employed Python 3.6, PyTorch 1.6, NVIDIA Corporation TU104GL GPU, and open-source libraries like Tensorflow, Pandas, Numpy, and Keras. Protege aided in querying and ontology extraction.

A. Data:

As explained in Section (IV.A) the data will be collected from Twitter. We analyze six attributes: source, username, date, tweet, language, and location. Initial cleaning filters bot-generated tweets using a 'source' attribute. Li et al. (2021) identified human-generated tweets. NLP methods [35] remove irrelevant and sarcastic tweets via Affective and Contextual Embeddings (ACE 1 and ACE 2). Geo-referencing uses geotags, user profiles, and content-mentioned places. The final step converts cleaned tweets into time series data, yielding daily disease-related tweet counts by country. Symptoms and disease names are independent variables in our multivariate dataset. Confirmed cases, our dependent variable, comes from health authorities [32], [33]. Table I summarizes disease names and time spans for model training and testing.

B. 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 $\hat{y}_i - y_i$ 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.

C. Experiments

For all baselines, including state-of-the-art methods and our hybrid model, the preprocessed dataset was divided into an 80% training and 20% test set. Time series data was normalized using min-max normalization to maintain the original distribution shape within the [0, 1] range. After predicting on test data, this normalization was reversed to align predictions with the original time series. Notably, some models like DeepAR have automatic scaling, eliminating manual adjustment. To address non-stationarity, the VAR model required the Augmented Dickey–Fuller (ADF) test for confirmation, followed by Box-Cox transformation. This reduced volatility and improved stationarity, which was particularly important for outbreak datasets with trends, as shown in the Covid-19 dataset for the USA in Figure 4 based on Table I's duration. The proposed approach is compared to nine baselines listed in Table IV. Optimal hyperparameter selection is vital for accurate forecasting models. Given the numerous hyperparameters in deep learning, Bayesian optimization fine-tunes them during training. Unlike grid search, Bayesian optimization streamlines parameter exploration, saving time and resources. Leveraging Gaussian processes, Bayesian optimization predicts optimal parameters. This technique is consistently applied for hyperparameter tuning in this study, following related work [36]. Model suitability is assessed through validation data error. Table II showcases the hyperparameter domains across models. The window size, essential for transforming input time series, notably affects minimal forecasting error. Our models also feature key hyperparameters greatly impacting forecasting accuracy.

D. Results

In this section, we present our analysis and results from contributions 1 to 4. Due to the result similarity across the six diseases, we visually represent a single disease in some experiments for simplicity and space. Initial step analysis, as seen in Figure 3, displays a heatmap illustrating the correlation between daily tweets and confirmed COVID-19 cases in the first dataset. A clear correlation emerges between confirmed cases and Covid-19 symptoms like fever (0.89), fatigue (0.87), and difficulty breathing (0.77). Less common symptoms of

Authorized licensed use limited to: Ontario Tech University. Downloaded on November 10,2026 at 20:17:49 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]

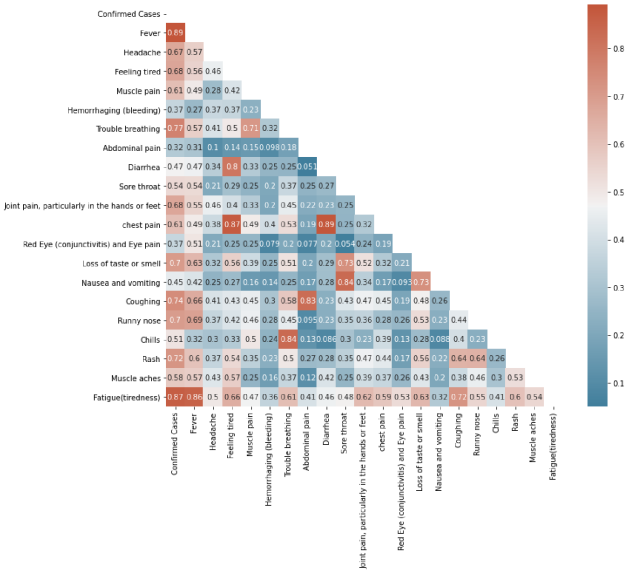


Fig. 3. Covid-19 First Dataset Heatmap

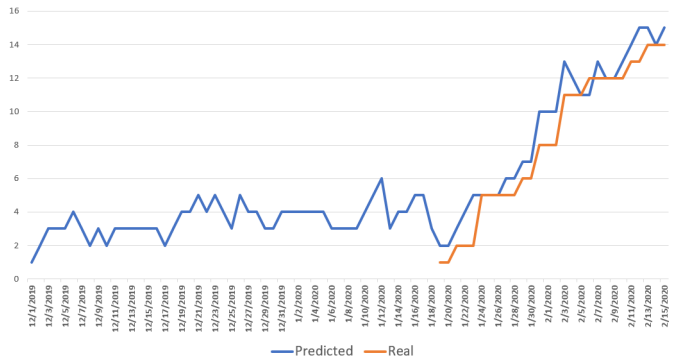


Fig. 4. Predicted and the real number of positive cases for Covid-19 USA

Covid-19, like abdominal pain (0.32) and red eyes (0.37), show a lower correlation. This pattern holds across all six diseases, demonstrating a strong link between specific symptoms and daily case numbers. In our first contribution, we utilized three datasets collected through various methods, as outlined in Section IV. Evaluation using two metrics in Table III shows the highest performance achieved by the first dataset. The ontology likely contributes to this result, enhancing keyword coverage and relevant tweet extraction for improved accuracy. In our second contribution, Figure 4 depicts our hybrid model's predictions compared to actual Covid-19 positive cases reported by US health authorities. We focused on a concise one-month timeframe for a clear comparison. As shown, our hybrid model closely aligns with actual numbers, detecting infections four to six weeks earlier through symptom-related tweets than reported by health authorities. Our third contribution involves identifying outbreaks using historical data. Training on the H1N1 outbreak in India (2015) and leveraging tweets from 2012 to 2021, we identified an outbreak in India in 2017 (Figure 5), matching actual positive case numbers reported by India NCDC [37]. The fourth contribution and analysis examine our approach's forecasting accuracy versus baselines and state-of-the-art models (Table IV). Employing data from six diseases, experiments were conducted ten times, reporting average RMSE and MAPE metrics. Our hybrid model consistently outperforms others in both metrics. XGBoost excels with small datasets like Covid-19, while our model adapts to varying dataset sizes by leveraging XGBoost's capabilities

Authorized licensed use limited to: Ontario Tech University. Downloaded on November 10,2026 at 20:17:49 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	**0.06**	**0.09**	**0.07**	**0.09**	**0.08**	**1.07**
	RMSE	**242.12**	**256.31**	**395.48**	**364.29**	**388.19**	**2419.03**
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	**0.04**	**0.07**	**0.05**	**0.06**	**0.07**	**1.03**
	RMSE	**248.23**	**227.83**	**282.48**	**283.33**	**382.66**	**2396.42**

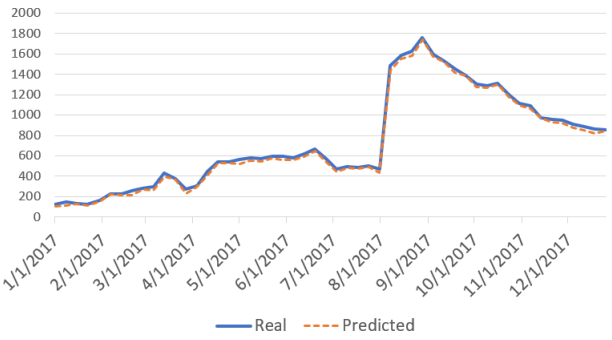


Fig. 5. Predicted and real positive cases for H1N1 India 2017 using proposed hybrid model

and Bi-LSTM's predictive power. N-BEATS, DeepAR, and TFT require ample data for reliable generalization, posing overfitting risks with small time series.

VI. CONCLUSION

Twitter stands as a valuable repository of insights into epidemics and contagious diseases, gleaned from online dialogues. This vast reservoir of information presents the potential to predict future pandemics and gain deeper insights into ongoing outbreaks. Established surveillance strategies, including those centered around testing center data, hospital admissions, and contact tracing, have emerged from research pinpointing correlations between these variables. In this study, we harnessed Twitter data encompassing infectious disease symptoms between 2012 and 2021 to drive a hybrid model for predicting infection numbers and forthcoming outbreaks. Leveraging an ontology that encompasses a wide spectrum of keywords linked to infectious diseases, we identified relevant symptoms. Notably, our proposed hybrid model surpasses nine state-of-the-art baselines across two key metrics. Uniquely leveraging tweets concerning infectious disease symptoms, our model exhibits the capacity to detect the emergence of outbreaks prior to health authorities' official reporting. Moreover, it holds the potential to identify hitherto unknown

Authorized licensed use limited to: Ontario Tech University. Downloaded on November 10,2026 at 20:17:49 UTC from IEEE Xplore. Restrictions apply.

diseases characterized by symptoms closely mirroring those of known ailments. These findings underscore the potential of social media in significantly elevating current surveillance methods, with even greater possibilities as the social media user base continues to expand.

ACKNOWLEDGMENT

We would like to acknowledge the financial support from the Natural Sciences and Engineering Research Council of Canada (NSERC) via an Emerging Infectious Diseases Modelling Initiative (EIDM) grant.

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*. Springer, 2016, pp. 131–139.

[2] S. Amin, M. I. Uddin, D. H. alSaeed, A. Khan, and M. Adnan, "Early Detection of Seasonal Outbreaks from Twitter Data Using Machine Learning Approaches," *Complexity*, vol. 2021, p. e5520366, Mar. 2021, publisher: Hindawi.

[3] 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," *AMIA Annu Symp Proc*, vol. 2022, pp. 313–322, May 2022.

[4] J. K. Miller, J. Chen, A. Sundermann, J. W. Marsh, M. I. Saul, K. A. Shutt, M. Pacey, M. M. Mustapha, L. H. Harrison, and A. Dubrawski, "Statistical outbreak detection by joining medical records and pathogen similarity," *Journal of biomedical informatics*, vol. 91, p. 103126, 2019.

[5] H. Abbasimehr and R. Paki, "Improving time series forecasting using LSTM and attention models," *J Ambient Intell Human Comput*, vol. 13, no. 1, pp. 673–691, Jan. 2022.

[6] 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 & Applic*, vol. 34, no. 4, pp. 3135–3149, Feb. 2022.

[7] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, Aug. 2016, pp. 785–794.

[8] M. Schuster and K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997, conference Name: IEEE Transactions on Signal Processing.

[9] 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 Journal of Public Health Informatics*, vol. 10, no. 2, Sep. 2018, number: 2.

[10] O. Edo-Osagie, B. De La Iglesia, I. Lake, and O. Edeghere, "A scoping review of the use of Twitter for public health research," *Computers in Biology and Medicine*, vol. 122, p. 103770, Jul. 2020.

[11] C. A. Sims, "Macroeconomics and Reality," *Econometrica*, vol. 48, no. 1, pp. 1–48, 1980, publisher: [Wiley, Econometric Society].

[12] Q. Wang, Y. Zhou, and X. Chen, "A Vector Autoregression Prediction Model for COVID-19 Outbreak," Feb. 2021, arXiv:2102.04843 [stat].

[13] 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 Infect Dis*, vol. 21, no. 1, p. 839, Dec. 2021.

[14] J. Luo, Z. Zhang, Y. Fu, and F. Rao, "Time series prediction of COVID-19 transmission in America using LSTM and XGBoost algorithms," *Results in Physics*, vol. 27, p. 104462, Aug. 2021.

[15] 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, p. e039676, Dec. 2020, publisher: British Medical Journal Publishing Group Section: Epidemiology.

[16] 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, p. e056685, Jul. 2022, publisher: British Medical Journal Publishing Group Section: Epidemiology.

[17] K. Guo, C. Shen, X. Zhou, S. Ren, M. Hu, M. Shen, X. Chen, and H. Guo, "Traffic Data-Empowered XGBoost-LSTM Framework for Infectious Disease Prediction," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–12, 2022, conference Name: IEEE Transactions on Intelligent Transportation Systems.

[18] 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.

[19] 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, p. 110017, Oct. 2020.

[20] 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, p. 110214, Nov. 2020.

[21] B. Lim, S. o. Arık, N. Loeff, and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," *International Journal of Forecasting*, vol. 37, no. 4, pp. 1748–1764, 2021.

[22] F. M. Caldas and C. Soares, "A Temporal Fusion Transformer for Long-term Explainable Prediction of Emergency Department Overcrowding," Jul. 2022, arXiv:2207.00610 [cs].

[23] B. Zhou, G. Yang, Z. Shi, and S. Ma, "Interpretable Temporal Attention Network for COVID-19 forecasting," *Applied Soft Computing*, vol. 120, p. 108691, May 2022.

[24] 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," *Computational and Mathematical Methods in Medicine*, vol. 2022, pp. 1–19, May 2022.

[25] V. Papastefanopoulos, P. Linardatos, and S. Kotsiantis, "COVID-19: A Comparison of Time Series Methods to Forecast Percentage of Active Cases per Population," *Applied Sciences*, vol. 10, no. 11, p. 3880, Jan. 2020, number: 11 Publisher: Multidisciplinary Digital Publishing Institute.

[26] P. Arunkumar, L. K. Ramasamy *et al.*, "Time-series forecasting and analysis of covid-19 outbreak in highly populated countries: A data-driven approach," *International Journal of E-Health and Medical Communications (IJEHMC)*, vol. 13, no. 2, pp. 1–17, 2021.

[27] B. N. Oreshkin, D. Carpov, N. Chapados, and Y. Bengio, "N-BEATS: neural basis expansion analysis for interpretable time series forecasting," *CoRR*, vol. abs/1905.10437, 2019.

[28] S. O. Folorunso, J. B. Awotunde, O. O. Banjo, E. A. Ogundepo, and N. O. Adeboye, "Comparison of active covid-19 cases per population using time-series models," *International Journal of E-Health and Medical Communications (IJEHMC)*, vol. 13, no. 2, pp. 1–21, 2021.

[29] 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, 2021.

[30] W. Zaremba, I. Sutskever, and O. Vinyals, "Recurrent Neural Network Regularization," Feb. 2015, arXiv:1409.2329 [cs].

[31] D. Ontology, "Disease Ontology - Institute for Genome Sciences @ University of Maryland."

[32] C. W. 24/7, "CDC Works 24/7," Nov. 2022. [Online]. Available: https://www.cdc.gov/index.htm

[33] WHO, "World Health Organization (WHO)." [Online]. Available: https://www.who.int

[34] P. R. Text, "Processing Raw Text." [Online]. Available: https://www.nltk.org/book/ch03.html

[35] N. Babanejad, H. Davoudi, A. An, and M. Papagelis, "Affective and Contextual Embedding for Sarcasm Detection," in *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 225–243.

[36] 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," *Journal of Electronic Science and Technology*, vol. 17, no. 1, pp. 26–40, Mar. 2019.

[37] 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 20:17:49 UTC from IEEE Xplore. Restrictions apply.