IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 20, 2021, accepted September 1, 2021, date of publication September 7, 2021,
date of current version September 20, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3110972

Utilizing Google Search Data With Deep Learning, Machine Learning and Time Series Modeling to Forecast Influenza-Like Illnesses in South Africa

SEUN O. OLUKANMI, FULUFHELO V. NELWAMONDO, (Senior Member, IEEE), AND NNAMDI I. NWULU, (Senior Member, IEEE)

Institute for Intelligent Systems, Department of Electrical and Electronic Engineering Science, University of Johannesburg, Johannesburg 2006, South Africa

Corresponding author: Seun O. Olukanmi (seun_fagbemi@yahoo.com)

The work of Seun O. Olukanmi was supported by the Global Excellence and Stature (GES) Scholarship Fund of the University of Johannesburg.

ABSTRACT Influenza-like illnesses (ILI) result in deaths and hospitalizations across the globe. Traditional surveillance systems rely on data from general medical practitioners. The process is resource-intensive and plagued with delay. Although recent studies have shown the potential utility of free and fast alternatives like web and social media data, the reliability cannot be generalized due to differences in technological culture. Meanwhile, there is a scarcity of studies exploring these free online data for (sub-Saharan) African countries. In this paper, we utilize Google trends (GT) data for ILI forecasting in South Africa. We study models based on deep learning (Long short-term memory (LSTM) and feedforward neural networks (FNN)), machine learning (Multiple linear regression (MLR), elastic net (EN), support vector machine (SVM)), and statistical time series (seasonal autoregressive integrated moving average (SARIMA)) algorithms. The FNN and SVM models using GT data alone, produce forecasts close in accuracy to those fitted to actual ILI data. The algorithms rank differently across various performance measures. Generally, the deep learning techniques perform better than the other algorithms in our study. However, tuning the former is quite intricate. Combining GT and historical ILI data enhances the models. The non-deep-learning algorithms benefit more from this enhancement. Furthermore, we observe that search volume increases proportional to and timeously with reported infection rates, suggesting that South Africans search Google in the week they feel flu symptoms. Thus, monitoring Google search trends is a reliable proxy for monitoring flu spread in South Africa.

INDEX TERMS Deep learning, flu surveillance, Google Trends, ILI reporting, influenza forecasting, machine learning, South Africa.

I. INTRODUCTION

Like the coronavirus disease (COVID-19), influenza (flu) is a severe respiratory infection that can cause complications and death in humans. Flu primarily affects young children, the elderly, and persons with underlying health conditions. It is caused by influenza viruses that spread in all parts of the world, leading to up to 5 million cases of acute illness, and about 290 000 to 650 000 deaths annually [1]. In response to this worldwide epidemic, the world health organization (WHO) monitors flu activity globally to strengthen its prevention and control [1]. In South Africa, where there is high HIV and tuberculosis infection rate, published estimates reveal that between 6734 and 11, 619 influenza deaths occur every year, and 22, 481 out of 47,000 incidences of acute influenza-like illness (ILI) result in hospitalization [2], [3]. Sustainable surveillance, a vital goal of the South African national influenza policy and strategic plan developed by the Department of Health for 2017 to 2021, currently employs traditional laboratory-based monitoring systems. A drawback of these systems is that they are resource-intensive and slow, often with a lag of 1 to 2 weeks before reports are available.

The increased availability of online real-time data streams such as web search volumes, internet forums and social media data, has led to a new field of research called infodemiology or more broadly, digital epidemiology. The field explores the use of these alternative data streams for disease surveillance. While the term infodemiology coined by Eysenbach [4] is older, digital epidemiology was recently defined by Salathe [5] as epidemiology that uses data generated

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Liu.

outside of the public health system. One assumption behind using these alternative data streams, for example search engine data, is that patients commonly use search engines for initial self-diagnosis, providing the health-information seeking trends of a populace over time [6]. Furthermore, patients may prefer to use search engines for information or education about health conditions that may attract societal stigma. Online platforms are also low cost, real-time and give a much more satisfactory spatial resolution of disease surveillance [7]. These advantages have become more important in the wake of the COVID-19 pandemic, as demonstrated by recent studies [8]–[12].

Despite the current evidence base in this field, the results cannot be generalized across countries due to varying cultural, technological, and economic inclinations [7]. This is substantiated by studies in Italy [13] and Turkey [14] where no significant relationship was found between Google search data and national disease records. In Germany, though the Google search data correlated well with reported Lyme disease incidence, it failed to improve forecast accuracy significantly [15]. Tran *et al.* [16] also detailed the low validity of Google search data for behavioural forecasting of national suicide rates in 4 countries. Furthermore, most of the studies focused on the USA, and there remains a need for studies that investigate the feasibility and performance of models built with these alternative data streams for disease surveillance across other countries of the world. This is especially needed in Africa where severe respiratory infections are a leading cause of mortality and where such studies are scarce [17]. Research has also suggested the need for contextualized approaches when using these online data for disease surveillance [17]. In a recent study [18], we had taken the first step of investigating correlation between Google search data for several flu-related queries and the traditional flu surveillance data in South Africa. We established that significant correlation exists for certain flu queries and highlighted such queries.

In this paper, we are concerned with investigating the predictive utility of Google search data for ILI surveillance in South Africa using the query terms (21 of them) compiled in our previous study. To the best of our knowledge, there is no published study on the use of Google search data for forecasting influenza activity in South Africa.

Our specific contributions are as follows: we developed forecasting models based on deep learning (LSTM and FNN), machine learning (Multiple linear regression (MLR), elastic net (EN) and support vector machine (SVM)), and seasonal autoregressive integrated moving average (SARIMA) algorithms. We studied the performance of the models using (i) Google search data alone, (ii) a combination of Google search data and historical ILI data, and (iii) historical ILI data only. Our experimental study reveals the relative strengths of the various algorithms across a variety of performance measures. The performance measures studied include root mean squared error (RMSE), mean absolute error (MAE), Pearson correlation coefficient (PCC), peak weak difference (PWD) and peak magnitude difference (PMD) between the estimated and the true incidence rates. Our results demonstrate the usefulness of Google search data as a reliable proxy for monitoring flu spread in South Africa.

II. RELATED WORKS

Since the early days of the infodemiology research area and over the years, various online data streams have been explored by several studies. Some of these include news articles [19]–[21], data from health-related websites and blogs [22], Wikipedia [23]–[25] and more commonly, search engines [4], [26]–[28] and Twitter data [29]–[33]. Twitter data has the advantage of providing spatio-temporal insights. More importantly, the infection-related tweets are personal testimonies, thus they reflect actual illness more accurately [34]. However, the data collection procedure, preprocessing and classification as required in studies using Twitter data are generally more complicated with the potential of false positives and negatives plaguing such studies. Moreover, Twitter has the limitation of not being commonly used by everyone [35]. On the other hand, search engines are more universal due to increased Internet penetration [35]. In addition, data obtained from search engines are simpler to use in investigating the epidemiological patterns of a specific disease over time. One search engine data source is Google Trends (GT), a free web tool that provides relative volume of Google engine searches for user-specified queries. GT is generic and publicly accessible. This is unlike Google Flu Trends (GFT) which although focused on providing flu estimates, had disadvantages of non-reproducibility since its underlying data and algorithm were non-accessible and was eventually discontinued in 2015 due to concerns about its reliability and accuracy [36], [37].

Though GT data has been explored for the monitoring of multiple communicable and non-communicable diseases including dengue fever [38], [39], tuberculosis [40], [41], multiple sclerosis [42], Type 2 diabetes [43], Lyme disease [6], dementia [44], Ebola [45], pertussis [46], gastroenteritis [47], cancer [48], zika [49] and more recently COVID-19 [50], [51], Aiello *et al.* [52] in their recent review noted that it has mostly been used for ILI monitoring and surveillance. A few of the studies that have utilized GT data for ILI surveillance are [53]–[55]. Some of these studies incorporate climate data [56] while others include traditional disease data to form a hybrid system with improved performance [57]–[59]. Some studies stop at investigating and reporting a correlation between the web data and traditional disease records [12], [46], [60], while others go ahead to nowcast or forecast disease trends using the web or social media data [27], [39], [61]. Nowcasting is a short-term prediction that gives the current disease incidence trends while forecast aims to estimate future disease trends.

The models that have been commonly used for disease incidence forecasting in the digital epidemiology field include the statistical ARIMA/ARIMAX time series methods [49], [56], [39], [62]–[64], traditional machine learning algorithms such linear regression, random forests, elastic net and support

vector machines [17], [41], [65], [66]. The application of deep learning techniques is a recent development. Studies that have applied deep learning methods include [9], [67]–[71]. Deep learning techniques are attractive due to their competitive performance.

In a very recent review of studies that utilized Internet-based user-generated data for public health surveillance by Abad *et al.* [72], 56% of the studies were from the US. Other countries with a significant number of studies were the UK, Australia, Canada, and Italy. In another recent review of studies that used Twitter for public health research by Edo-Osagie *et al.* [73], the top five countries in their breakdown of study activity by country were the US, UK, Canada, India, and China, with more than half of the studies originating in the US. Mogo [7] suggests that the acclaimed results from the developed countries cannot be generalized due to varying cultural, technological, and economic inclinations. The very few recently published works that have utilized digital data streams for nowcasting or forecasting disease trends in sub-Saharan Africa include [17] and [74].

It is evident from the aforementioned reviews that there is a scarcity of studies investigating the use of these online data streams for disease surveillance in Africa even though severe respiratory infections may have significant impact on morbidity and mortality in this part of the world [75]. This paper seeks to address this gap by showing the predictive usefulness of Google search trends for ILI forecasting in South Africa. We develop models to design nowcasting/forecasting systems. The study shows that Google search patterns suffice as features to monitor flu spread and they can also be combined with historical ILI data, yielding even better forecasting performances. Consequently, this study brings us a step closer to achieving sustainable surveillance, a key goal of the South African national influenza policy and strategic plan, developed by the Department of Health for 2017 to 2021 [3].

III. MATERIALS AND METHODS

A. DATA

This study was approved by the ethics committee of the Faculty of Engineering and the Built Environment, University of Johannesburg. The data used in the study covered a period of 459 weeks, comprising of weekly data from the 1st week of 2010 to the 43rd week of 2018. The data sources are described in the following subsections:

1) ILI DATA

We obtained the national-level outpatient ILI data over the study period from the viral watch, a flu surveillance program of the National institute for communicable diseases (NICD) carried out by general practitioners in all the nine provinces of South Africa [76]. The data are anonymized, and they are weekly counts of patients who meet the ILI case definition of a fever of 38° and cough or sore throat with onset $\leq$10 days.

2) GOOGLE SEARCH DATA

Google search data was obtained from Google Trends (GT), a free web tool that gives aggregated search volume for any query submitted by users. The data are anonymized and normalized relative to all searches conducted on the Google search engine based on geolocation, category, and time period. There are 25 categories each with several subcategories to choose from. The search term index can be downloaded as a CSV file at national and regional levels. GT returns 0 search index for a query if its search volume is low for a given period of time [77]. We collected the weekly search index for the 21 flu- related queries highlighted in our previous study [18]. The queries showed moderate to strong Pearson correlation with the NICD ILI data ($r \geq 0.5$; $p < 0.05$) for at least 5 of the 9 years under study. These queries fall under three broad divisions namely, flu nomenclature, flu symptoms, and flu treatment and are presented in Table 1.

B. DATA PRE-PROCESSING

The ILI data from NICD had 21 missing instances from the total of 467 weekly data. To improve the forecasting performance of the models, the t*sclean* from the R *forecast* package [78] was used to estimate missing values and outlier replacements for the training data. To replicate the real-world scenario, we left the test set uncleaned. For the models based on the FNN and LSTM deep learning algorithms, we performed min-max scaling on the data to the range *[0,1]* before giving it as input.

C. ALGORITHMS

The different models employed in this study are briefly described on the following subsections:

1) SEASONAL ARIMA WITH(OUT) EXTERNAL REGRESSORS

The SARIMA technique [79] denoted as ARIMA(p, d, q) (P, D, Q)$_m$, is a time series forecasting method proposed by Box and Jenkins. It models the input data in two components, namely, the non-seasonal ARIMA part and additional seasonal terms. Parameter p denotes the order of the autoregressive (AR) model, d the degree of differencing, and q the order of the moving average (MA) model. P, D, Q are the AR, differencing, and MA terms for the seasonal component, while m is the number of observations in each year. SARIMA can be extended by including external regressors. The extended version is commonly called SARIMAX. In our study, we implemented the SARIMA(X)-based models using the *auto.arima* function in the R *forecast* package [78].

2) MULTIPLE LINEAR REGRESSION (MLR)

MLR is a machine learning algorithm involving more than one explanatory variable being used to predict a response variable by modelling the linear relationship between the explanatory and response variables. The MLR-based models were implemented using the *lm* function in R *stats* package.

TABLE 1. Queries showing moderate to strong correlation (R ≥ 0.5) with ILI data for at least 5 of the 9 years under study [18].

S/N	QUERY/(SUB-CATEGORY)	CATEGORY	2010	2011	2012	2013	2014	2015	2016	2017	2018
1	Common Cold (Disease)	All	0.45	**0.70**	**0.73**	**0.76**	**0.71**	**0.87**	**0.55**	**0.76**	**0.81**
2	Common Cold (Disease)	Health	0.22*	**0.51**	**0.78**	**0.63**	**0.70**	**0.84**	**0.52**	**0.80**	**0.77**
3	Cough (Search term)	All	**0.53**	0.45	**0.52**	**0.65**	**0.74**	**0.66**	**0.74**	**0.81**	**0.70**
4	Cough (Search term)	Health	0.35	0.47	0.47	**0.71**	**0.70**	**0.59**	**0.73**	**0.81**	**0.62**
5	Cough (Topic)	All	**0.55**	0.41	**0.61**	**0.73**	**0.77**	**0.64**	**0.78**	**0.78**	**0.67**
6	Cough (Topic)	Health	0.44	**0.59**	**0.52**	**0.71**	**0.66**	**0.56**	**0.76**	**0.78**	**0.66**
7	Flu (Search term)	Health	0.36	**0.81**	**0.65**	**0.78**	**0.79**	**0.84**	**0.75**	**0.84**	**0.84**
8	Flu (Search term)	All	**0.58**	**0.86**	**0.72**	**0.73**	**0.82**	**0.88**	**0.77**	**0.86**	**0.83**
9	Flu symptoms (Search term)	All	0.15*	**0.71**	**0.69**	**0.65**	**0.75**	**0.70**	**0.72**	**0.75**	**0.86**
10	Flu symptoms (Search term)	Health	0.16*	**0.72**	**0.68**	**0.72**	**0.70**	**0.73**	**0.67**	**0.69**	**0.87**
11	"Flu symptoms" (Search term)	All	0.29	**0.71**	0.42	**0.66**	**0.79**	**0.68**	**0.64**	**0.72**	**0.82**
12	"Flu symptoms" (Search term)	Health	0.29	**0.70**	**0.58**	**0.69**	**0.67**	**0.73**	**0.65**	**0.75**	**0.81**
13	Influenza (Search term)	All	0.05*	**0.60**	**0.50**	0.42	0.20*	**0.66**	**0.80**	**0.80**	**0.58**
14	Influenza (Disease)	All	**0.56**	**0.88**	**0.70**	**0.76**	**0.83**	**0.88**	**0.81**	**0.87**	**0.83**
15	Influenza (Disease)	Health	0.44	**0.86**	**0.75**	**0.79**	**0.83**	**0.86**	**0.78**	**0.85**	**0.84**
16	Oseltamivir (Medication)	All	0.24*	**0.57**	0.28*	**0.60**	**0.52**	**0.58**	**0.68**	**0.71**	**0.77**
17	Oseltamivir (Medication)	Health	0.18*	0.34	**0.50**	**0.63**	**0.50**	**0.69**	**0.64**	**0.70**	**0.68**
18	Symptoms of flu (Search term)	All	0.02*	**0.65**	0.44	0.39	0.47	**0.64**	**0.50**	**0.59**	**0.77**
19	Symptoms of flu (Search term)	Health	0.11*	**0.61**	0.47	0.41	0.44	**0.58**	**0.52**	**0.58**	**0.80**
20	Tamiflu (Search term)	Health	0.20*	0.44	**0.54**	**0.60**	0.31	**0.66**	**0.76**	**0.66**	**0.72**
21	Tamiflu (Search term)	All	0.11*	**0.53**	**0.55**	**0.56**	0.49	**0.73**	**0.55**	**0.66**	**0.75**

3) ELASTIC NET (EN)

Elastic net is a regularized and variable selection regression method that combines the penalties of the ridge and Least Absolute Shrinkage and Selection Operator (LASSO) methods [80]. The EN-based models were implemented using the *cv.glmnet* function from the *glmnet* package in R [81], [82].

4) SUPPORT VECTOR MACHINE REGRESSION (SVM)

SVM regression involves mapping the independent variables into a higher dimensional feature space using the kernel trick [83]. We implemented the SVM models using the function *svm* in the *e1071* R package [84].

5) FEEDFORWARD NEURAL NETWORK (FNN)

A feedforward neural network is an artificial neural network (ANN) comprising nodes arranged into layers. The input layer is the first layer, the output layer is the last layer, while the layers between are referred to as the hidden layers. This type of ANN is called feedforward because information only travels forward from the input through the hidden layers and to the output layer. There are no cycle connections through which the network output can be fed back into the nodes [85].

6) LONG SHORT-TERM MEMORY (LSTM)

LSTMs [86] are a variant of recurrent neural networks (RNN). RNNs, unlike FNNs are artificial neural networks with feedback connections among the nodes, allowing information to be retained. LSTMs were developed to overcome the gradient vanishing and long-term dependency problems which basic RNNs suffer from. LSTMs are well suited to process sequence data and thus are being applied to time series prediction tasks.

The FNNs and LSTMs were implemented using the *Keras* library with *Tensorflow* backend in Python.

D. EXPERIMENTAL STUDY

In this section, we describe different experimental models that are combinations of the various algorithms and data for forecasting future trends of flu in South Africa. The three categories of models that were explored based on input data are described as follows:

1) PREDICTING ILI RATES VIA TIME SERIES MODELLING OF HISTORICAL ILI DATA ONLY (ILI-DATA ONLY MODELS)

Supplying the historical ILI trends as input, we studied two techniques for time series modelling namely:

***ILI-SARIMA*:** This model is based on the ARIMA algorithm (SARIMA) using 80% of the data (367 instances) to train the model and 20% of the data (92 instances) for testing the performance of the model. The training/test set size is as suggested in the popular book of Hyndman and Athanasopoulos [87].

***ILI-LSTM*:** For the second model, the LSTM algorithm was used. The train/test split was maintained as in the ILI-SARIMA model. We used a stack of two LSTM layers for our implementation, with each layer having 200 units. The input data was reshaped as 4 time steps (4 weeks) for ILI prediction at the next time step (the following week). A dropout technique with rate of 0.2 was used after each LSTM layer to prevent overfitting of the model. These optimal parameter values for the model were determined experimentally by evaluating the effect on the model's forecasting performance. The *adam* optimizer was used for the model compilation and the model was trained for 100 epochs.

A simple mathematical representation of these two models based on the input and output data is given as:

$$\boldsymbol{o} = \boldsymbol{f}(\boldsymbol{p}) \tag{1}$$

where $\boldsymbol{o}$ is the predicted ILI rates and $\boldsymbol{p}$ is the past ILI trends.

2) PREDICTING ILI RATES WITH GT DATA ONLY (GT-DATA ONLY MODELS)

These set of models take only the Google search data of the highlighted 21 queries as inputs (explanatory variables) and predict the ILI rates as the outputs (response variable). This helps us ascertain the GT data's predictive ability for flu surveillance in South Africa in the absence of real-life ILI data. We considered zero to two weeks ahead forecasts. For the zero week ahead estimates (nowcast), we used GT data of the queries at week (t) to predict the ILI incidence rates at the end of week (t). For one week ahead forecasts, we used GT data of all the queries at the current week (t) to predict the ILI incidence rates for the next week (t+1), while for two weeks ahead forecasts, GT data of all the queries at the current week (t) were used to predict the ILI incidence rates of the next two weeks (t+2). The same number of training instances as in the ILI data only models (in section C-1) were maintained for the different week-ahead estimates while the test set reduced by 1 and 2 instances for one week and two weeks ahead, respectively. The four model categories with their implementation details are briefly described below:

***GT-MLR*:** These models are based on the multivariate linear regression algorithm.

***GT-EN*:** The three GT-EN models are based on the elastic net algorithm. We performed a 10-fold cross-validation to find the optimal value of the shrinkage parameter, *lambda* over a varying set of *alphas* from 0 to 1. When *alpha* equals 0, we have a ridge regression and when *alpha* equals 1, we have a lasso regression.

***GT-SVM*:** For these models based on the SVM algorithm, four different kernels (linear, radial, polynomial, and sigmoid) were tested experimentally on the data, and the radial kernel was selected as it performed the best. The optimal values for the model parameters, *cost*, *gamma* and *epsilon* were determined experimentally by evaluating the effect on the forecasting performance (in terms of RMSE) of the different possible combinations of the values for the three parameters from a predefined set of values.

***GT-FNN*:** These models adopt the feedforward neural network (FNN) algorithm. The input layers have 21 nodes, representing the Google search data of the 21 queries, while the output layers have just 1 node representing the predicted ILI rate. A dropout technique with rate of 0.2 was used after each hidden layer to prevent overfitting of the models. For the nowcasts, the optimal model parameters were determined experimentally as three hidden layers with 256, 128 and 64 units respectively from the first to the last hidden layer, all with the *relu* activation function. For the one week ahead forecasts, there are four hidden layers with 28, 56, 56 and 128 units respectively. The two weeks ahead forecasts also have four hidden layers with 24, 58, 58 and 128 units from the first to the last hidden layer. The three GT-FNN models used the *adam* optimizer for compilation and 100 epochs for training.

A simple mathematical representation of these models based on the input and output data is given as:

$$\boldsymbol{o} = \boldsymbol{f}(\boldsymbol{q_i}, \boldsymbol{q_2}, \boldsymbol{q_3}, \ldots, \boldsymbol{q_{21}}) \tag{2}$$

where $\boldsymbol{o}$ is the predicted ILI rates and $(\boldsymbol{q_i}, \boldsymbol{q_2}, \boldsymbol{q_3}, \ldots, \boldsymbol{q_{21}})$ are the 21 Google queries.

3) PREDICTING ILI RATES WITH GT AND HISTORICAL ILI DATA (ILI-GT-DATA MODELS)

Hybrid models comprising online data and traditional disease data have been shown to have improved performance [58], [59]. For this set of models, the past ILI trends data and the Google search data of the 21 queries were fed as input. We were able to determine the improvement in the forecasting performance of the models that had used only the past ILI or GT data through these models. The training/test data sizes are the same as in the *ILI-SARIMA* model. The models are described below along with their implementation details:

***ILI-GT-SARIMAX*:** Here, the ILI-SARIMA model was enhanced with the GT data as external regressors. As in the GT-data only models we considered zero to two weeks ahead estimates, resulting in three models. For the zero week ahead estimates (nowcasts), GT data of the queries at current week (t) were used as external regressors to predict ILI rate of the same week (t). For one-week ahead estimates, GT data of the queries at the current week (t) were used as regressors to predict the ILI incidence rate of the next week (t+1), while for two weeks ahead estimates, GT data of the queries at the current week (t) were used as regressors to predict the ILI incidence rate of the next two weeks (t+2).

***ILI-GT-LSTM*:** This model is made up of a single LSTM layer with 200 units, and a dropout technique with rate 0.2 immediately following the layer. Similar to the ILI_LSTM model, we reshape the input of past

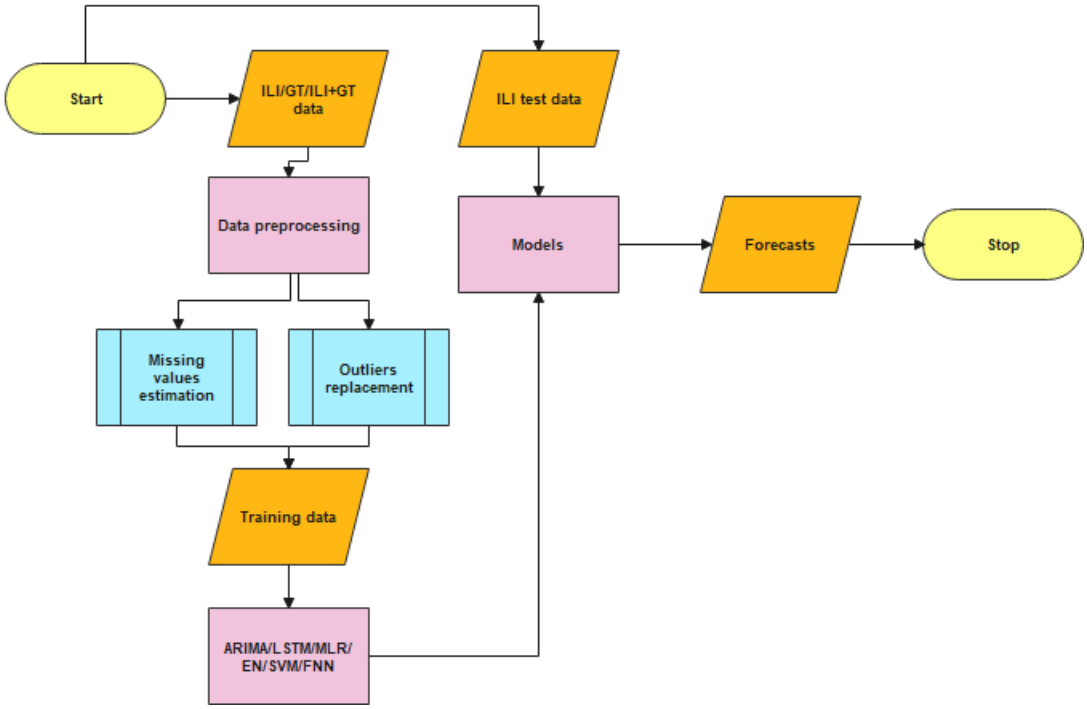

FIGURE 1. ILI forecasting framework.

ILI and GT data as 4 time steps (4 weeks) for the prediction of ILI at the next time step (the following week). The optimal parameter values for the model were also determined experimentally by evaluating the effect on the model's forecasting performance. We also used the *adam* optimizer and 100 epochs for compiling and training this model.

***ILI-GT-MLR*:** These MLR models take as explanatory variables, the GT data of all the 21 queries as well as historical ILI data from the past one or two weeks, in order to predict the current week ILI incidence rate.

***ILI-GT-EN*:** These models are based on the elastic net algorithm and the optimal parameter values selection process are the same as in the GT-EN models. The input data are the same as in the ILI-GT-MLR models.

***ILI-GT-SVM*:** For these SVM models, the radial kernel was also selected as it performed the best and the optimal values for the parameters were also determined experimentally as in the GT-SVM models. The input data are also the same as in the ILI-GT-MLR models.

***ILI-GT-FNN*:** The input layers of these FNN-based models have 22 nodes, representing the Google search data of the 21 queries and the historical ILI data of the past one or two weeks. As in the GT-FNN models, the output layers have just 1 node representing the predicted ILI rate. For the model incorporating the ILI data of the past one week, the optimal model parameters were determined experimentally as four hidden layers with 27, 54, 54 and 128 units respectively from the first to the last hidden layer, all with the *relu* activation function. A dropout technique with rate of 0.2 was used after each hidden layer of this model to prevent overfitting. The model using the ILI data of the past two weeks also have four hidden layers with 27, 54, 54 and 128 units respectively. This second model used a dropout technique with rate 0.2 after each of the first two hidden layers only. Both models used the *adam* optimizer for compilation and 40 epochs for training.

A simple mathematical representation of these models based on the input and output data is given as:

$$\boldsymbol{o} = \boldsymbol{f}(\boldsymbol{p}, \boldsymbol{q_i}, \boldsymbol{q_2}, \boldsymbol{q_3}, \ldots, \boldsymbol{q_{21}}) \tag{3}$$

where $\boldsymbol{o}$ is the predicted ILI rates, $\boldsymbol{p}$ is the ILI data of the past one or two weeks and $(\boldsymbol{q_i}, \boldsymbol{q_2}, \boldsymbol{q_3}, \ldots, \boldsymbol{q_{21}})$ are the 21 Google queries.

Our overall forecasting framework is summarized in Figure 1.

E. PERFORMANCE METRICS

We compared the predictive performance of the different models by calculating different measures on the test set. The root mean squared error (RMSE), mean absolute error (MAE) and Pearson correlation coefficient (PCC) of the estimated and the true incidence rates. The lower the RMSE and MAE, the better the performance of the models, while the higher the PCC, the better the model performance.

We also estimated the ability of each model to predict the week of the peak ILI incidence and the height of the peak. Peak week difference (PWD) is estimated as the difference between the true and estimated peak week. In contrast, peak magnitude difference (PMD) corresponds to the difference

between the true and the forecasted peak height. The lower the PWD and PMD, the better the model performance.

IV. RESULTS

A. ILI-DATA ONLY MODELS (ILI-SARIMA AND ILI-LSTM)

Table 2 shows the performances of the ILI-data only models, ILI-SARIMA and ILI-LSTM. The recorded accuracy metrics include the RMSE, MAE, PCC, PWD and PMD. Each of them provides a way to judge the accuracy of the models. There are two peaks reflecting the two influenza seasons of 2017 and 2018 in the test period. The deep learning time series model ILI-LSTM performed better than its statistical counterpart (ILI-SARIMA) on almost all the metrics except the PWD where the ILI-SARIMA predicted the first peak week correctly. Figure 2 presents a visualization of the true ILI data over the study period (red) while showing the predicted signal (blue) over the test period, for comparison.

TABLE 2. Performance of the ILI-Data only models.

Model	RMSE	MAE	PCC	PWD	PMD
ILI-SARIMA (1,0,2)(2,1,0){52}	20.7140	14.7060	0.8450	Peak 1: 0 Peak 2: +1	Peak: 37 Peak 2: 67
ILI-LSTM	11.7309	7.8912	0.9244	Peak 1: +1 Peak 2: +1	Peak: 5 Peak 2: 10

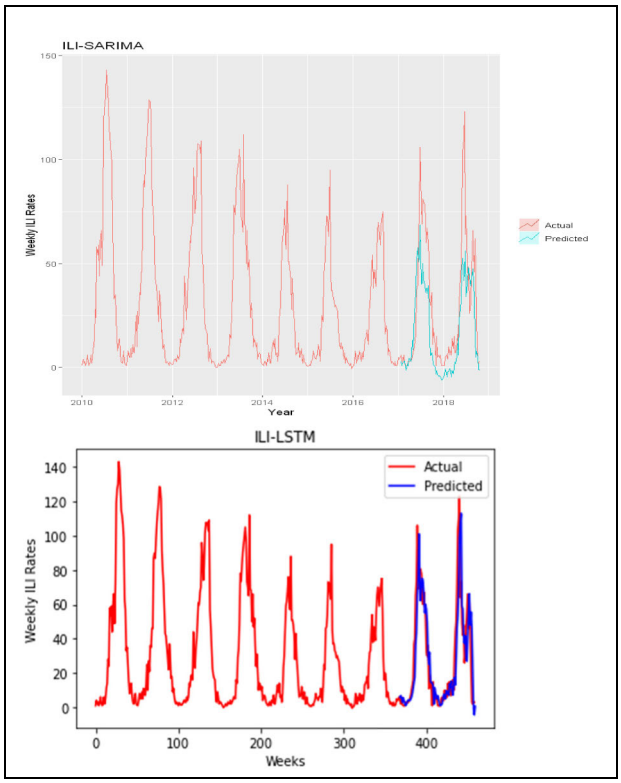

FIGURE 2. Comparison of actual and predicted weekly ILI counts from the ILI-SARIMA and ILI-LSTM models for the two flu seasons in the test period.

B. GT-DATA ONLY MODELS (GT-MLR, GT-EN,GT-SVM AND GT-FNN)

In Table 3, we present the performances of the GT-data only regression models. The table records the accuracy metrics for three prediction scenarios: nowcasting (predicting ILI using search data of same week at each time step in the test period), forecasting ILI one week ahead and forecasting ILI two weeks ahead at each time step in the test period. The predicted signals for each case are visualized in Figure 3.

1) RMSE

The RMSE for these set of models ranged from 13.44 to 25.48 with different algorithms and at different forecast horizons. The deep learning model (GT-FNN) had the lowest forecast error, followed by GT-SVM and GT-EN, while GT-MLR showed the worst performance based on this metric. The RMSE of the four models increases as the forecast horizon increases from zero (nowcasts) to two weeks ahead.

2) MAE

Like the RMSE, GT-FNN had the lowest MAE, followed by GT-SVM and GT-EN, while GT-MLR had the highest forecast error. The values ranged from 9.87 (GT-FNN, lag 0) to 21.70 (GT-MLR, lag 2). Like the RMSE, the MAE increased for the models as we move from nowcasts to two weeks ahead forecasts.

3) PCC

The PCC for the models generally decreased as the forecast horizon increased from zero to two weeks ahead. For the nowcasts (zero week ahead), GT-SVM had the highest PCC of 0.9068. GT-MLR had the highest PCC of 0.8764 at one week ahead forecasts while GT-FNN had the highest PCC of 0.8372 at two weeks ahead forecasts.

4) PWD

There are two peaks reflecting the two influenza seasons of 2017 and 2018 in the test period. For the nowcasts, all the models predicted the first peak week accurately, GT-MLR still predicted the second peak week accurately while the rest of the models peaked one week earlier than the true peak. At the one week ahead estimates, GT-MLR and GT-EN predicted the first flu season peak to be one week later than the true peak, while the GT-SVM and GT-FNN models predicted the peak to be two weeks later than the true peak. For the second flu season, GT-MLR, GT-EN and GT-SVM had accurate peak predictions, while GT-FNN predicted the peak to be three weeks later than the true peak. At the two weeks ahead forecasts, GT-MLR and GT-EN predicted the peak of the first flu season to be two weeks later while GT-SVM and GT-FNN predicted the peak as three weeks later than the true peak. For the second season in the test period, GT-MLR and GT-SVM got accurate peak week prediction while GT-EN and GT-SVM were one week later in their peak week prediction.

5) PMD

While GT-MLR had the highest forecast error in terms of RMSE and MAE, it showed the best performance in

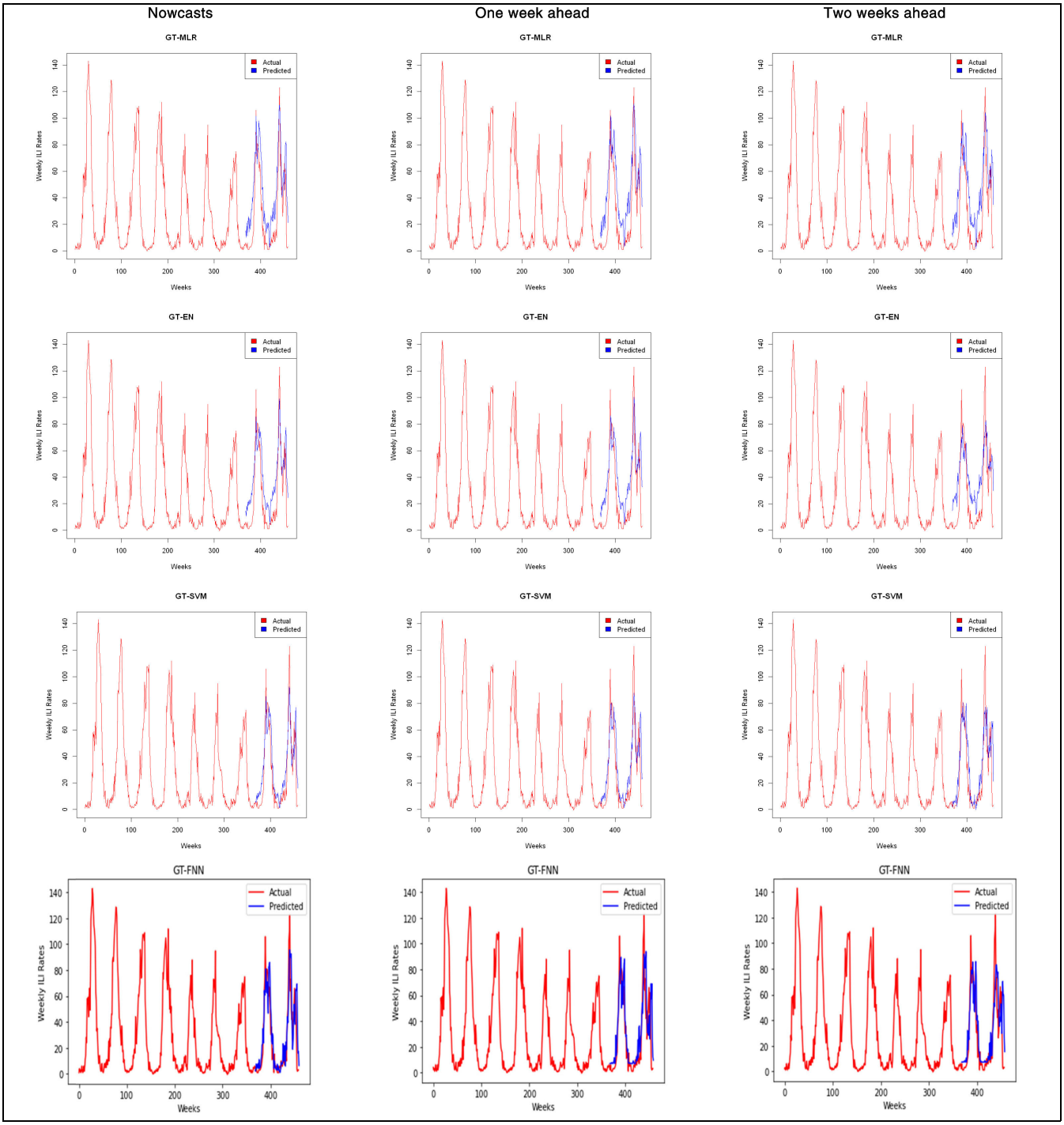

FIGURE 3. **Comparison of actual and predicted weekly ILI counts from the GT-MLR, GT-EN, GT-SVM AND GT-FNN models for same week, one week ahead and two weeks ahead for the two flu seasons in the test period.**

predicting the magnitude of the peaks in the 2 seasons in the test period at all the forecast horizons (zero to two weeks ahead), as shown in Figure 3. The other three models (GT-EN, GT-SVM and GT-FNN) had relatively comparable performance in predicting the PMD at one week and two weeks ahead forecasts.

C. ILI-GT-DATA MODELS

In this section of our study, we show the performances of the models incorporating both ILI and GT data. These models perform the best among the three categories of models. Table 4 presents the performance of the ILI-GT-SARIMAX model, Table 5 shows the performance of the ILI-GT-LSTM

TABLE 3. Performance of the GT-Data only models: nowcasts, one week ahead and two weeks ahead estimates.

	Model	RMSE	MAE	PCC	PWD	PMD
Nowcasts	GT-MLR	21.7947	18.9286	0.9046	Peak 1: 0 Peak 2: 0	Peak 1: 4 Peak 2: 13
	GT-EN (*alpha = 0*)	18.8756	16.6586	0.9024	Peak 1: 0 Peak 2: -1	Peak 1: 21 Peak 2: 24
	GT-SVM (*cost = 0.07, gamma = 0.015, epsilon = 0.01*)	14.3519	11.6196	0.9068	Peak 1: 0 Peak 2: -1	Peak 1: 21 Peak 2: 31
	GT-FNN	13.4447	9.8693	0.9031	Peak 1: 0 Peak 2: -1	Peak 1: 25 Peak 2: 27
One week ahead	GT-MLR	24.5443	21.3728	0.8764	Peak 1: +1 Peak 2: 0	Peak 1: 4 Peak 2: 12
	GT-EN (alpha = 0)	21.1270	18.1429	0.8528	Peak 1: +1 Peak 2: 0	Peak 1: 20 Peak 2: 23
	GT-SVM (*cost = 0.2, gamma = 0.01, epsilon = 0.01*)	16.7801	12.4337	0.8580	Peak 1: +2 Peak 2: 0	Peak 1: 26 Peak 2: 35
	GT-FNN	15.6713	11.2970	0.8722	Peak 1: +2 Peak 2: +3	Peak 1: 17 Peak 2: 29
Two weeks ahead	GT-MLR	25.4801	21.6988	0.8340	Peak 1: +2 Peak 2: 0	Peak 1: 10 Peak 2: 19
	GT-EN (alpha = 0.5)	22.5198	19.9732	0.8357	Peak 1: +2 Peak 2: +1	Peak 1: 27 Peak 2: 40
	GT-SVM (*cost = 0.3, gamma = 0.02, epsilon = 0.05*)	17.9880	12.4457	0.8269	Peak 1: +3 Peak 2: 0	Peak 1: 25 Peak 2: 46
	GT-FNN	17.3423	11.8596	0.8372	Peak 1: +3 Peak 2: +1	Peak 1: 21 Peak 2: 40

model while Table 6 presents the performances of the ILI-GT-MLR, ILI-GT-EN, ILI-GT-SVM and ILI-GT-FNN models. The visualization of the performance of the models can be seen in Figures 4, 5 and 6.

1) RMSE

The deep learning models (ILI-GT-FNN) had the lowest RMSE value of 10.54 which is comparable with the RMSE of the ILI-GT-LSTM and ILI-GT-SVM models which had

TABLE 4. Performance of the ILI-GT-SARIMAX models: nowcasts, one week ahead and two weeks ahead estimates.

	Model	RMSE	MAE	PCC	PWD	PMD
Nowcasts	ILI-GT-SARIMAX (3,1,0) (1,1,1) {52}	14.6380	10.6329	0.8824	Peak 1: 0 Peak 2: +1	Peak 1: 17 Peak 2: 44
One week ahead	ILI-GT-SARIMAX (0,1,1) (1,1,1) {52}	15.8359	11.0241	0.8635	Peak 1: 0 Peak 2: +1	Peak 1: 25 Peak 2: 50
Two weeks ahead	ILI-GT-SARIMAX (3,1,0) (1,1,1) {52}	16.3721	11.6043	0.8514	Peak 1: 0 Peak 2: +1	Peak 1: 16 Peak 2: 47

TABLE 5. Performance of ILI-GT-LSTM model.

Model	RMSE	MAE	PCC	PWD	PMD
ILI-GT-LSTM	10.7078	7.5335	0.9370	Peak 1: +1 Peak 2: 0	Peak 1: 5 Peak 2: 4

values of 10.71 and 10.96 respectively. The other machine learning models (ILI-GT-MLR and ILI-GT-EN) had performance comparable to one another with increasing RMSE values as the added ILI data was lagged from the previous week (t−1) to previous two weeks (t−2). The out-of-sample RMSE of the ILI-GT-SARIMAX nowcast model (14.64) is comparable to the RMSE obtained with GT-SVM nowcast model (14.35).

2) MAE

The ILI-GT-FNN deep learning model had the lowest MAE value of 7.33, followed by the ILI-GT-LSTM, ILI-GT-MLR and ILI-GT-SVM models with comparable values of 7.53, 7.57 and 7.98 respectively. Similar to the other performance metrics, the MAE value increases as the forecast horizon increases.

3) PCC

The ILI-GT-FNN deep learning model also had the highest Pearson correlation coefficient of 0.9397, followed by its recurrent network counterpart (ILI-GT-LSTM) with comparable PCC value of 0.9370. The ILI-GT-SARIMAX models generally had slightly lower PCC values.

4) PWD

The ILI-GT-MLR, ILI-GT-EN, ILI-GT-SVM, and ILI-GT-FNN models all predicted the peak week as one week later than the true peak week when the ILI data of the past one week were part of the explanatory variables, while they predicted the peak week as two weeks later than the true

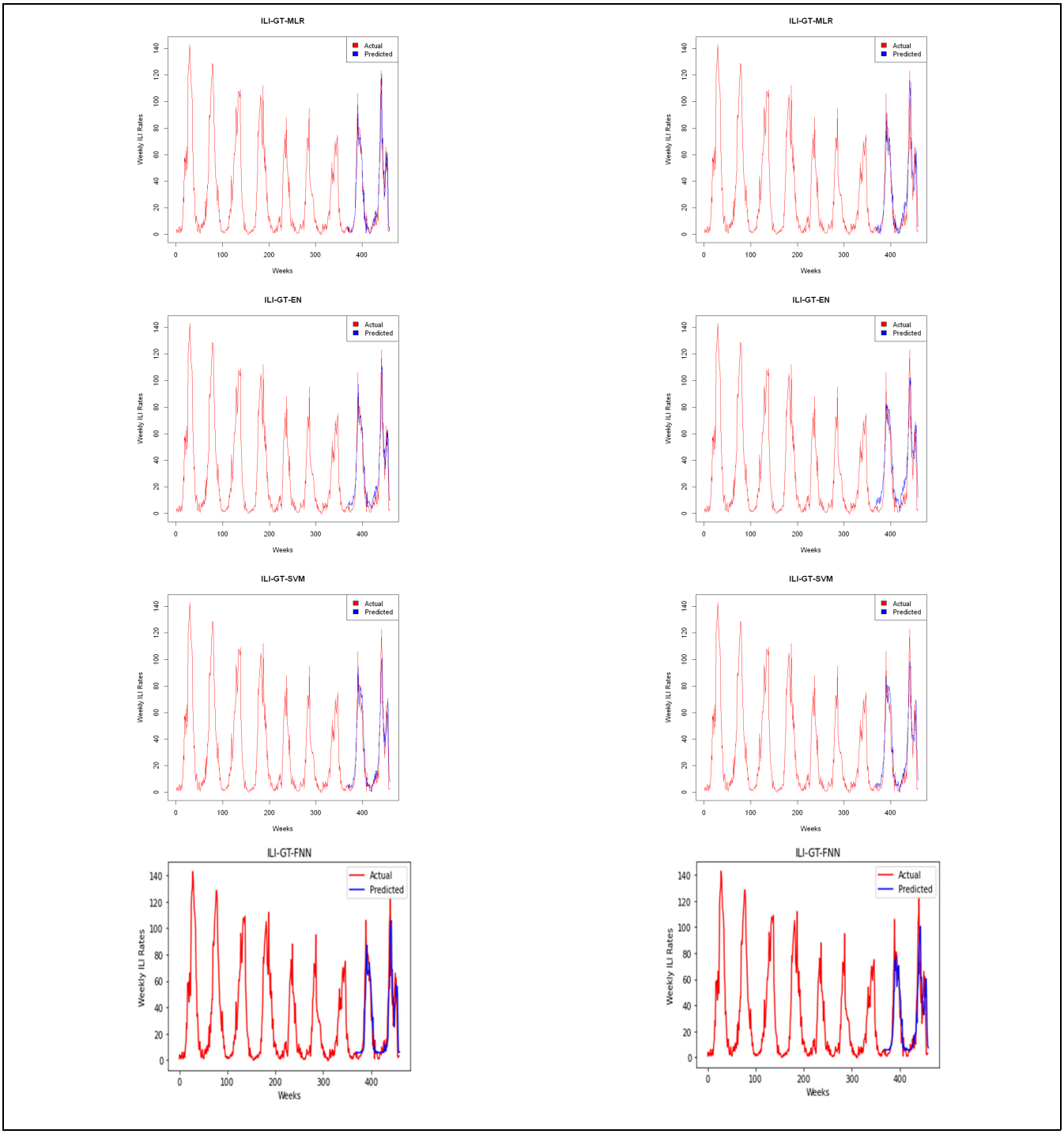

FIGURE 4. Comparison of actual and predicted weekly ILI counts from the ILI-GT-MLR, ILI-GT-EN, ILI-GT-SVM AND ILI-GT-FNN models for the two flu seasons in the test period. Col 1: The weekly ILI estimates for week t were produced given search data at week t + ILI data at week (t−1). Col 2: : The weekly ILI estimates for week t were produced given search data at week t + ILI data at week (t−2).

peak week when the ILI data of the past two weeks were part of the explanatory variables to forecast the current week ILI rates. The performance of the ILI-GT-SARIMAX models were the same with the increase in forecast horizon: the first peak week was predicted correctly while the second peak week was predicted as one week later than the true peak week.

5) PMD

The ILI-GT-LSTM model had the lowest PMD values, followed by the ILI-GT-MLR model.

V. DISCUSSION

The results of the GT-data only models (Table 3) show that Google search data alone can be used to forecast ILI

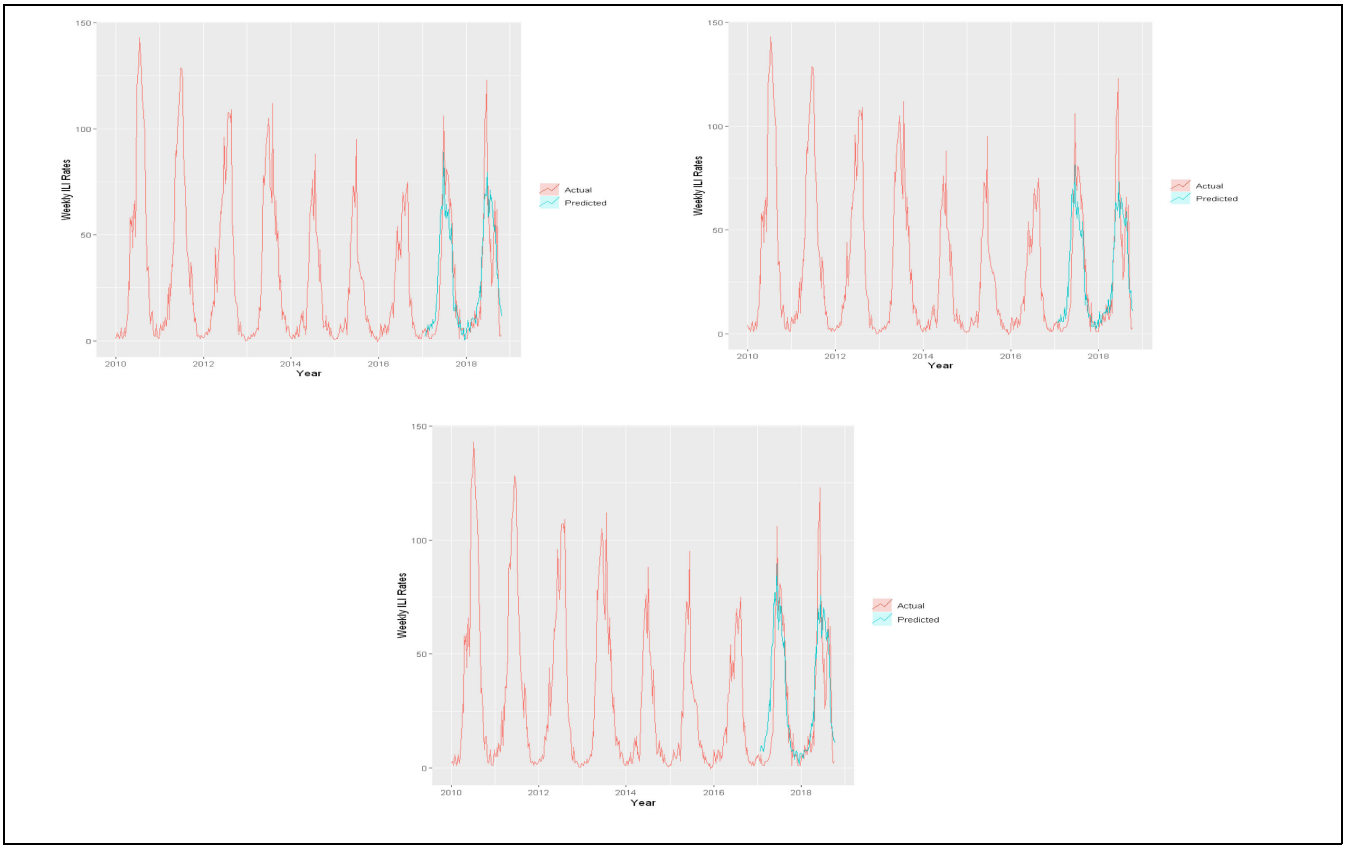

FIGURE 5. Comparison of actual and predicted weekly ILI counts from the ILI-GT-SARIMAX models using GT data of the same week (t) (1st row, left), past one week (t−1) (1st row, right) and past two weeks (t−2) (2nd row) as external regressors.

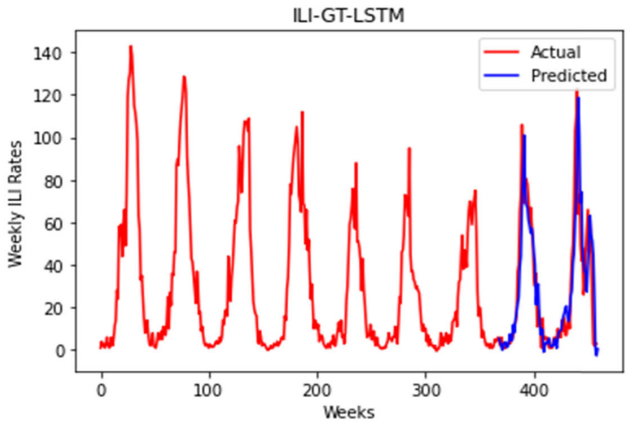

FIGURE 6. Comparison of actual and predicted weekly ILI counts from ILI-GT-LSTM model.

incidence rates with accuracy close to what would be obtained using the real ILI data from general practitioners. The real ILI data collection process is plagued with delay, and it is costly. GT data on the other hand is free and available without delay. This shows that GT data may be used as a reliable proxy for monitoring flu spread in South Africa in the absence or delay of the real ILI data. This affords more rapid response to mitigate the effects of any epidemic outbreak.

Furthermore, the results from the ILI-GT-data models in Tables 4, 5 and 6 show that better influenza forecasting models can be developed for South Africa by incorporating both past ILI and GT data. These set of models outperform every other category of models that we considered.

The results also show that the choice of algorithm is equally crucial. The models rank differently across different performance measures. On the overall, across the categories of models considered, the models based on deep learning methods consistently outperform the traditional statistical/machine learning methods on most of the metrics. The ILI-GT-LSTM model proves especially good at predicting the peak magnitudes. This accuracy, however, comes at a cost of detailed tuning of several model parameters which can be complicated. Meanwhile, each of the SVM based models (GT-SVM and ILI-GT-SVM) yields performance that is comparable to its deep learning based counterpart (GT-FNN, ILI-GT-FNN and ILI-GT-LSTM) in terms of RMSE, MAE, PCC and PWD. This shows that SVM is also a highly effective machine learning method for the purpose as supported by the findings of Nsoesie *et al.* [17]

Also, we observe that the performance gaps between the traditional linear models on the one hand and the deep

TABLE 6. Performance of the ILI-GT-MLR, ILI-GT-EN, ILI-GT-SVM AND ILI-GT-FNN models.

	Model	RMSE	MAE	PCC	PWD	PMD
GT data of week(t) + ILI data of week(t-1)	ILI-GT-MLR	11.5944	7.5747	0.9262	Peak 1: +1 Peak 2: +1	Peak 1: 9 Peak 2: 2
	ILI-GT-EN (alpha = 0.2)	11.6645	8.9105	0.9339	Peak 1: +1 Peak 2: +1	Peak 1: 9 Peak 2: 12
	ILI-GT-SVM (*cost = 0.6, Gamma = 0.005, epsilon = 0.01*)	10.9619	7.9146	0.9377	Peak 1: +1 Peak 2: +1	Peak 1: 12 Peak 2: 22
	ILI-GT-FNN	10.5426	7.3329	0.9397	Peak 1: +1 Peak 2: +1	Peak 1: 19 Peak 2: 17
GT data of week(t) + ILI data of week(t-2)	ILI-GT-MLR	14.5611	10.2385	0.8821	Peak 1: +2 Peak 2: +2	Peak 1: 15 Peak 2: 8
	ILI-GT-EN (alpha = 0)	14.3501	11.6684	0.9112	Peak 1: +2 Peak 2: +2	Peak 1: 24 Peak 2: 22
	ILI-GT-SVM (*cost = 0.5 gamma=0.005, epsilon = 0.05*)	12.8062	9.6631	0.9168	Peak 1: 0 Peak 2: 0	Peak 1: 18 Peak 2: 25
	ILI-GT-FNN	12.8253	8.7767	0.9099	Peak 1: +2 Peak 2: +2	Peak 1: 29 Peak 2: 22

learning models (Table 3) on the hand are significantly reduced when historical ILI data are combined with GT data to predict current/future ILI rates (Table 6). By implication, where historical ILI data of up to the previous two weeks is available (which is the case in South Africa), these models that are reasonably accurate affords simplicity (requires no complicated tuning) which is an attractive property in practice. The simple GT-data only MLR (GT-MLR) also predicts the peak weeks accurately and can be used if the peak timing is the desired performance metric.

We find that, regardless of the modelling method, the prediction errors were lowest for the same week forecasts (nowcasts) and the models' performances generally decreases on all the accuracy metrics as the forecast horizon increases. This may suggest that in South Africa, people search within or around the week when they have started experiencing symptoms and/or go for treatment at health facilities in the same week.

VI. CONCLUSION

This paper addresses the scarcity of studies exploring the utility of web and social media data for ILI surveillance in South Africa. Specifically, we established the predictive utility of Google Search data for ILI rate forecasting. We explored models based on deep learning techniques (LSTM and FNN), machine learning algorithms (Multiple linear regression (MLR), elastic net (EN) and support vector machine (SVM)), and seasonal autoregressive integrated moving average (SARIMA) models. The choice of algorithm plays a significant role in model performance. The studied models rank differently across various criteria, though deep learning techniques are optimal overall with appropriate tuning. Notably also, SVM-based models compete closely with the deep learning techniques. The findings also show that reasonable forecasts can be made a few weeks ahead using the proposed models. We observe that search volume increases proportional to and timeously with reported infection rates. This may suggest that South Africans tend to search Google to confirm their symptoms or for common flu home remedies around the week they feel flu symptoms. The implication is that monitoring Google search data is a reliable proxy for monitoring flu spread. The study established that models based on Google search data alone produce forecasts comparable in accuracy to those fitted to real-life ILI data; and that the models that incorporate GT and historical ILI data have enhanced by forecasting capability. Thus, Google search data, although free and readily available without delay can be utilized to effectively address problems associated with traditional systems such as resource-intensiveness and delay. This will, in turn, allow for better epidemic preparedness, moving us closer to achieving sustainable surveillance, a key goal of the South African national influenza policy and strategic plan, developed by the Department of Health for 2017 to 2021.

ACKNOWLEDGMENT

The authors gratefully acknowledge Prof. Cheryl Cohen and Jo Mcanerney of the South African National Institute for Communicable Diseases (NICD) for providing the ILI surveillance data.

REFERENCES

[1] World Health Organization, "Influenza (seasonal)," *Bull. World Health Org.*, Nov. 2018. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs211/en/

[2] L. Blumberg, C. Cohen, H. Dawood, O. Hellferscee, A. Karstaedt, K. McCarthy, S. Madhi, M. McMorrow, J. Moyes, J. Nel, A. Puren, E. Variava, W. Ramkrishna, G. Reubenson, S. Tempia, F. Treurnicht, S. Walaza, and H. Zar. (2017). *Influenza NICD Recommendations for the Diagnosis, Prevention, Management and Public Health Response*. Accessed: Mar. 13, 2018. [Online]. Available: http://www.nicd.ac.za/wp-content/uploads/2017/03/Influenzaguidelines-final_24_05_2017.pdf

[3] Department of Health. (2017). *National Influenza Policy and Strategic Plan: 2017 to 2021*. [Online]. Available: http://www.health.gov.za/index.php/component/phocadownload/category/339

[4] G. Eysenbach, "Infodemiology: Tracking flu-related searches on the web for syndromic surveillance," in *Proc AMIA Annu. Symp.*, 2006, pp. 244–248.

[5] M. Salathé, "Digital epidemiology: What is it, and where is it going?" *Life Sci., Soc. Policy*, vol. 14, no. 1, pp. 1–5, Dec. 2018, doi: 10.1186/s40504-017-0065-7.

[6] A. Seifter, A. Schwarzwalder, K. Geis, and J. Aucott, "The utility of 'Google trends' for epidemiological research: Lyme disease as an example," *Geospatial Health*, vol. 4, no. 2, p. 135, May 2010, doi: 10.4081/gh.2010.195.

[7] E. Mogo. *Social Media as a Public Health Surveillance Tool: Evidence and Prospects*. Accessed: Nov. 2, 2018. [Online]. Available: http://www.enterprise.sickweather.com/downloads/SW-SocialMedia_WhitePaper.pdf

[8] M. Sulyok, T. Ferenci, and M. Walker, "Google trends data and COVID-19 in Europe: Correlations and model enhancement are European wide," *Transboundary Emerg. Diseases*, vol. 68, pp. 2610–2615, Jul. 2021, doi: 10.1111/tbed.13887.

[9] S. Prasanth, U. Singh, A. Kumar, V. A. Tikkiwal, and P. H. J. Chong, "Forecasting spread of COVID-19 using Google trends: A hybrid GWO-deep learning approach," *Chaos, Solitons Fractals*, vol. 142, Jan. 2021, Art. no. 110336, doi: 10.1016/j.chaos.2020.110336.
[10] D. Fantazzini, "Short-term forecasting of the COVID-19 pandemic using Google trends data: Evidence from 158 countries," *Appl. Econometrics Forthcoming*, Aug. 2020, doi: 10.2139/ssrn.3671005.
[11] A. Mavragani and K. Gkillas, "COVID-19 predictability in the United States using Google trends time series," *Sci. Rep.*, vol. 10, no. 1, pp. 1–12, Dec. 2020, doi: 10.1038/s41598-020-77275-9.
[12] A. Mavragani, "Tracking COVID-19 in Europe: Infodemiology approach," *JMIR Public Health Surveill.*, vol. 6, no. 2, Apr. 2020, Art. no. e18941, doi: 10.2196/18941.
[13] G. Cervellin, I. Comelli, and G. Lippi, "Is Google trends a reliable tool for digital epidemiology? Insights from different clinical settings," *J. Epidemiol. Global Health*, vol. 7, no. 3, p. 185, 2017, doi: 10.1016/j.jegh.2017.06.001.
[14] U. Bilge, S. Bozkurt, B. Yolcular, and D. Ozel, "Can social web help to detect influenza related illnesses in Turkey?" in *Large Scale Projects in eHealth*, B. Blobel, R. Engelbrecht, and M. A. Shifrin, Eds. Amsterdam, The Netherlands: IOS Press BV, 2012, pp. 100–104.
[15] M. Kapitány-Fövény, T. Ferenci, Z. Sulyok, J. Kegele, H. Richter, I. Vályi-Nagy, and M. Sulyok, "Can Google trends data improve forecasting of lyme disease incidence?" *Zoonoses Public Health*, vol. 66, no. 1, pp. 101–107, Feb. 2019, doi: 10.1111/zph.12539.
[16] U. S. Tran, R. Andel, T. Niederkrotenthaler, B. Till, V. Ajdacic-Gross, and M. Voracek, "Low validity of Google trends for behavioral forecasting of national suicide rates," *PLoS ONE*, vol. 12, no. 8, Aug. 2017, Art. no. e0183149, doi: 10.1371/journal.pone.0183149.
[17] E. O. Nsoesie, O. Oladeji, A. S. A. Abah, and M. L. Ndeffo-Mbah, "Forecasting influenza-like illness trends in Cameroon using Google search data," *Sci. Rep.*, vol. 11, no. 1, p. 6713, Dec. 2021, doi: 10.1038/s41598-021-85987-9.
[18] S. Olukanmi and F. Nelwamondo, "Digital influenza surveillance: The prospects of Google trends data for South Africa," in *Proc. icABCD*, Aug. 2020, pp. 397–402, doi: 10.1109/icABCD49160.2020.9183882.
[19] R. Grishman, S. Huttunen, and R. Yangarber, "Information extraction for enhanced access to disease outbreak reports," *J. Biomed. Inform.*, vol. 35, no. 4, pp. 236–246, 2002, doi: 10.1016/S1532-0464(03)00013-3.
[20] M. Abla and M. Blench, "Global public health intelligence network (GPHIN)," in *Proc. 7th Conf. Assoc. Mach. Transl. Amer.*, 2006, pp. 8–12. Accessed: Mar. 9, 2018. [Online]. Available: https://pdfs.semanticscholar.org/7d88/e623aa6ca78510e0093e17e2e00db39bdad5.pdf
[21] A. R. Reilly, E. A. Iarocci, C. M. Jung, D. M. Hartley, and N. P. Nelson, "Indications and warning of pandemic influenza compared to seasonal influenza," *Inf. Syst.*, vol. 9, no. 8, p. 2008, 2008.
[22] A. Hulth, G. Rydevik, and A. Linde, "Web queries as a source for syndromic surveillance," *PLoS ONE*, vol. 4, no. 2, p. e4378, Feb. 2009, doi: 10.1371/journal.pone.0004378.
[23] D. Mciver and J. S. Brownstein. *Wikipedia Usage Estimates Prevalence of Influenza-Like Illness in Near Real-Time*. Accessed: Jun. 21, 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4512362/pdf/ojphi-07-e40.pdf
[24] B. Bardak and M. Tan, "Prediction of influenza outbreaks by integrating Wikipedia article access logs and Google flu trend data," in *Proc. IEEE 15th BIBE*, Nov. 2015, pp. 1–6, doi: 10.1109/BIBE.2015.7367640.
[25] K. S. Hickmann, G. Fairchild, R. Priedhorsky, N. Generous, J. M. Hyman, A. Deshpande, and S. Y. D. Valle, "Forecasting the 2013–2014 influenza season using Wikipedia," *PLOS Comput. Biol.*, vol. 11, no. 5, May 2015, Art. no. e1004239, doi: 10.1371/journal.pcbi.1004239.
[26] P. M. Polgreen, Y. Chen, D. M. Pennock, F. D. Nelson, and R. A. Weinstein, "Using internet searches for influenza surveillance," *Clin. Infectious Diseases*, vol. 47, no. 11, pp. 1443–1448, Dec. 2008, doi: 10.1086/593098.
[27] R. Moss, A. Zarebski, P. Dawson, and J. M. McCaw, "Forecasting influenza outbreak dynamics in Melbourne from internet search query surveillance data," *Influenza Other Respiratory Viruses*, vol. 10, no. 4, pp. 314–323, Jul. 2016, doi: 10.1111/irv.12376.
[28] L. Clemente, F. Lu, and M. Santillana, "Improved real-time influenza surveillance: Using internet search data in eight Latin American countries," *JMIR Public Health Surveill.*, vol. 5, no. 2, Apr. 2019, Art. no. e12214, doi: 10.2196/12214.
[29] E. De Quincey and P. Kostkova, "Early warning and outbreak detection using social networking websites: The potential of Twitter," in *Proc. Int. Conf. Electron. Healthcare*. Berlin, Germany: Springer, 2009, pp. 21–24.
[30] H. Achrekar, A. Gandhe, R. Lazarus, S.-H. Yu, and B. Liu, "Predicting flu trends using Twitter data," in *Proc. IEEE INFOCOM WKSHPS*, Apr. 2011, pp. 702–707.
[31] A. Culotta, "Towards detecting influenza epidemics by analyzing Twitter messages," in *Proc. SOMA*, 2010, pp. 115–122, doi: 10.1145/1964858.1964874.
[32] A. Lamb, M. J. Paul, and M. Dredze, "Separating fact from fear: Tracking flu infections on Twitter," in *Proc. NAACL-HLT*, 2013, pp. 779–789.
[33] S. Yousefinaghani, R. Dara, Z. Poljak, T. M. Bernardo, and S. Sharif, "The assessment of Twitter's potential for outbreak detection: Avian influenza case study," *Sci. Rep.*, vol. 9, no. 1, pp. 1–17, Dec. 2019, doi: 10.1038/s41598-019-54388-4.
[34] R. Nagar, Q. Yuan, C. C. Freifeld, M. Santillana, A. Nojima, R. Chunara, and J. S. Brownstein, "A case study of the New York City 2012–2013 influenza season with daily geocoded Twitter data from temporal and spatiotemporal perspectives," *J. Med. Internet Res.*, vol. 16, no. 10, p. e236, Oct. 2014, doi: 10.2196/jmir.3416.
[35] A. Mavragani, "Infodemiology and infoveillance: Scoping review," *J. Med. Internet Res.*, vol. 22, no. 4, Apr. 2020, Art. no. e16206, doi: 10.2196/16206.
[36] D. Butler, "When Google got flu wrong," *Nature*, vol. 494, no. 7436, pp. 155–156, Feb. 2013, doi: 10.1038/494155a.
[37] D. R. Olson, K. J. Konty, M. Paladini, C. Viboud, and L. Simonsen, "Reassessing Google flu trends data for detection of seasonal and pandemic influenza: A comparative epidemiological study at three geographic scales," *PLoS Comput. Biol.*, vol. 9, no. 10, Oct. 2013, Art. no. e1003256, doi: 10.1371/journal.pcbi.1003256.
[38] R. A. Strauss, J. S. Castro, R. Reintjes, and J. R. Torres, "Google dengue trends: An indicator of epidemic behavior. The Venezuelan case," *Int. J. Med. Informat.*, vol. 104, pp. 26–30, Aug. 2017, doi: 10.1016/j.ijmedinf.2017.05.003.
[39] W. Anggraeni and L. Aristiani, "Using Google trend data in forecasting number of dengue fever cases with ARIMAX method case study: Surabaya, Indonesia," in *Proc. ICTS*, 2016, pp. 114–118, doi: 10.1109/ICTS.2016.7910283.
[40] X. Zhou, J. Ye, and Y. Feng, "Tuberculosis surveillance by analyzing Google trends," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 8, pp. 2247–2254, Aug. 2011, doi: 10.1109/TBME.2011.2132132.
[41] A. Mavragani and G. Ochoa, "Infoveillance of infectious diseases in USA: STDs, tuberculosis, and hepatitis," *J. Big Data*, vol. 5, no. 1, pp. 1–23, Dec. 2018, doi: 10.1186/s40537-018-0140-9.
[42] M. Moccia, R. Palladino, A. Falco, F. Saccà, R. Lanzillo, and V. B. Morra, "Google trends: New evidence for seasonality of multiple sclerosis," *J. Neurol., Neurosurg. Psychiatry*, vol. 87, no. 9, pp. 1028–1029, Sep. 2016, doi: 10.1136/jnnp-2016-313260.
[43] N. Tkachenko, S. Chotvijit, N. Gupta, E. Bradley, C. Gilks, W. Guo, H. Crosby, E. Shore, M. Thiarai, R. Procter, and S. Jarvis, "Google trends can improve surveillance of type 2 diabetes," *Sci. Rep.*, vol. 7, no. 1, pp. 1–10, Dec. 2017, doi: 10.1038/s41598-017-05091-9.
[44] H.-W. Wang, D.-R. Chen, H.-W. Yu, and Y.-M. Chen, "Forecasting the incidence of dementia and dementia-related outpatient visits with Google trends: Evidence from Taiwan," *J. Med. Internet Res.*, vol. 17, no. 11, p. e264, Nov. 2015, doi: 10.2196/jmir.4516.
[45] C. Alicino, N. L. Bragazzi, V. Faccio, D. Amicizia, D. Panatto, R. Gasparini, G. Icardi, and A. Orsi, "Assessing Ebola-related web search behaviour: Insights and implications from an analytical study of Google trends-based query volumes," *Infectious Diseases Poverty*, vol. 4, no. 1, pp. 1–13, Dec. 2015, doi: 10.1186/s40249-015-0090-9.
[46] V. Gianfredi, N. L. Bragazzi, M. Mahamid, B. Bisharat, N. Mahroum, H. Amital, and M. Adawi, "Monitoring public interest toward pertussis outbreaks: An extensive Google trends–based analysis," *Public Health*, vol. 165, pp. 9–15, Dec. 2018, doi: 10.1016/j.puhe.2018.09.001.
[47] C. Pelat, C. Turbelin, A. Bar-Hen, A. Flahault, and A. J. Valleron, "More diseases tracked by using Google trends," *Emerg. Infectious Diseases*, vol. 15, no. 8, pp. 1327–1328, 2009, doi: 10.3201/eid1508.090299.
[48] M. Schootman, A. Toor, P. Cavazos-Rehg, D. B. Jeffe, A. McQueen, J. Eberth, and N. O. Davidson, "The utility of Google trends data to examine interest in cancer screening," *BMJ Open*, vol. 5, no. 6, Jun. 2015, Art. no. e006678, doi: 10.1136/bmjopen-2014-006678.

[49] Y. Teng, D. Bi, G. Xie, Y. Jin, Y. Huang, B. Lin, X. An, D. Feng, and Y. Tong, "Dynamic forecasting of Zika epidemics using Google trends," *PLoS ONE*, vol. 12, no. 1, Jan. 2017, Art. no. e0165085, doi: 10.1371/journal.pone.0165085.

[50] M. Effenberger, A. Kronbichler, J. I. Shin, G. Mayer, H. Tilg, and P. Perco, "Association of the COVID-19 pandemic with internet search volumes: A Google Trends^{TM} analysis," *Int. J. Infectious Diseases*, vol. 95, pp. 192–197, Jun. 2020, doi: 10.1016/j.ijid.2020.04.033.

[51] C. Li, L. J. Chen, X. Chen, M. Zhang, C. P. Pang, and H. Chen, "Retrospective analysis of the possibility of predicting the COVID-19 outbreak from internet searches and social media data, China, 2020," *Eurosurveillance*, vol. 25, no. 10, Mar. 2020, Art. no. 2000199, doi: 10.2807/1560-7917.ES.2020.25.10.2000199.

[52] A. E. Aiello, A. Renson, and P. N. Zivich, "Social media- and internet-based disease surveillance for public health," *Annu. Rev. Public Health*, vol. 41, no. 1, pp. 101–118, Apr. 2020, doi: 10.1146/annurev-publhealth-040119-094402.

[53] S. B. Choi and I. Ahn, "Forecasting seasonal influenza-like illness in South Korea after 2 and 30 weeks using Google trends and influenza data from Argentina," *PLoS ONE*, vol. 15, no. 7, Jul. 2020, Art. no. e0233855, doi: 10.1371/journal.pone.0233855.

[54] S. Cho, C. H. Sohn, M. W. Jo, S.-Y. Shin, J. H. Lee, S. M. Ryoo, W. Y. Kim, and D.-W. Seo, "Correlation between national influenza surveillance data and Google trends in South Korea," *PLoS ONE*, vol. 8, no. 12, Dec. 2013, Art. no. e81422, doi: 10.1371/journal.pone.0081422.

[55] M. Kang, H. Zhong, J. He, S. Rutherford, and F. Yang, "Using Google trends for influenza surveillance in south China," *PLoS ONE*, vol. 8, no. 1, Jan. 2013, Art. no. e55205, doi: 10.1371/journal.pone.0055205.

[56] Y. Zhang, H. Bambrick, K. Mengersen, S. Tong, and W. Hu, "Using Google trends and ambient temperature to predict seasonal influenza outbreaks," *Environ. Int.*, vol. 117, pp. 284–291, Aug. 2018, doi: 10.1016/j.envint.2018.05.016.

[57] S. Yang, M. Santillana, J. S. Brownstein, J. Gray, S. Richardson, and S. C. Kou, "Using electronic health records and internet search information for accurate influenza forecasting," *BMC Infectious Diseases*, vol. 17, no. 1, pp. 1–9, Dec. 2017, doi: 10.1186/s12879-017-2424-7.

[58] S. Yang, M. Santillana, and S. C. Kou, "Accurate estimation of influenza epidemics using Google search data via ARGO," *Proc. Nat. Acad. Sci. USA*, vol. 112, no. 47, pp. 14473–14478, 2015, doi: 10.1073/pnas.1515373112.

[59] M. W. Davidson, D. A. Haim, and J. M. Radin, "Using networks to combine 'big data' and traditional surveillance to improve influenza predictions," *Sci. Rep.*, vol. 5, no. 1, pp. 1–5, Jul. 2015, doi: 10.1038/srep08154.

[60] A. Husnayain, A. Fuad, and L. Lazuardi, "Correlation between Google trends on dengue fever and national surveillance report in Indonesia," *Global Health Action*, vol. 12, no. 1, Jan. 2019, Art. no. 1552652, doi: 10.1080/16549716.2018.1552652.

[61] F. S. Lu, S. Hou, K. Baltrusaitis, M. Shah, J. Leskovec, R. Sosic, J. Hawkins, J. Brownstein, G. Conidi, J. Gunn, J. Gray, A. Zink, and M. Santillana, "Accurate influenza monitoring and forecasting using novel internet data streams: A case study in the Boston metropolis," *JMIR Public Health Surveill.*, vol. 4, no. 1, p. e4, Jan. 2018, doi: 10.2196/publichealth.8950.

[62] S. Chadsuthi, S. Iamsirithaworn, W. Triampo, and C. Modchang, "Modeling seasonal influenza transmission and its association with climate factors in Thailand using time-series and ARIMAX analyses," *Comput. Math. Methods Med.*, vol. 2015, pp. 1–8, Nov. 2015, doi: 10.1155/2015/436495.

[63] Q. Mao, K. Zhang, W. Yan, and C. Cheng, "Forecasting the incidence of tuberculosis in China using the seasonal auto-regressive integrated moving average (SARIMA) model," *J. Infection Public Health*, vol. 11, no. 5, pp. 707–712, Sep. 2018, doi: 10.1016/j.jiph.2018.04.009.

[64] T. U. Xiao-Qing, Z. H. Zhan-Lin, G. O. Zheng, Y. E. Mahan, H. U. Bing-Xue, T. I. Tian, A. B. Ainiwaer, C. H. Zhen, G. U. Hailili, F. A. Xu-Cheng, and D. A. Jiang-Hong, "Forecasting influenza like illness in Urumqi based on ARIMAX model," *Chin. J. Disease Control Prevention*, vol. 22, no. 6, pp. 590–593, 2018, doi: 10.16462/J.CNKI.ZHJBKZ.2018.06.012.

[65] C. Poirier, A. Lavenu, V. Bertaud, B. Campillo-Gimenez, E. Chazard, M. Cuggia, and G. Bouzillé, "Real time influenza monitoring using hospital big data in combination with machine learning methods: Comparison study," *JMIR Public Health Surveill.*, vol. 4, no. 4, Dec. 2018, Art. no. e11361, doi: 10.2196/11361.

[66] M. Santillana, A. T. Nguyen, M. Dredze, M. J. Paul, E. O. Nsoesie, and J. S. Brownstein, "Combining search, social media, and traditional data sources to improve influenza surveillance," *PLOS Comput. Biol.*, vol. 11, no. 10, Oct. 2015, Art. no. e1004513, doi: 10.1371/journal.pcbi.1004513.

[67] Q. Xu, Y. R. Gel, L. L. R. Ramirez, K. Nezafati, Q. Zhang, and K.-L. Tsui, "Forecasting influenza in Hong Kong with Google search queries and statistical model fusion," *PLoS ONE*, vol. 12, no. 5, May 2017, Art. no. e0176690, doi: 10.1371/JOURNAL.PONE.0176690.

[68] S. Chae, S. Kwon, and D. Lee, "Predicting infectious disease using deep learning and big data," *Int. J. Environ. Res. Public Health*, vol. 15, no. 8, p. 1596, Jul. 2018, doi: 10.3390/IJERPH15081596.

[69] E. L. Aiken, A. T. Nguyen, C. Viboud, and M. Santillana, "Toward the use of neural networks for influenza prediction at multiple spatial resolutions," *Sci. Adv.*, vol. 7, no. 25, Jun. 2021, Art. no. eabb1237, doi: 10.1126/SCIADV.ABB1237.

[70] S. Volkova, E. Ayton, K. Porterfield, and C. D. Corley, "Forecasting influenza-like illness dynamics for military populations using neural networks and social media," *PLoS ONE*, vol. 12, no. 12, Dec. 2017, Art. no. e0188941, doi: 10.1371/journal.pone.0188941.

[71] C.-T. Yang, Y.-A. Chen, Y.-W. Chan, C.-L. Lee, Y.-T. Tsan, W.-C. Chan, and P.-Y. Liu, "Influenza-like illness prediction using a long short-term memory deep learning model with multiple open data sources," *J. Supercomput.*, vol. 76, no. 12, pp. 9303–9329, Dec. 2020.

[72] Z. Shakeri Hossein Abad, A. Kline, M. Sultana, M. Noaeen, E. Nurmambetova, F. Lucini, M. Al-Jefri, and J. Lee, "Digital public health surveillance: A systematic scoping review," *NPJ Digit. Med.*, vol. 4, no. 1, pp. 1–13, Dec. 2021, doi: 10.1038/s41746-021-00407-6.

[73] 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, doi: 10.1016/j.compbiomed.2020.103770.

[74] N. L. Bragazzi and N. Mahroum, "Google trends predicts present and future plague cases during the plague outbreak in madagascar: Infodemiological study," *JMIR Public Health Surveill.*, vol. 5, no. 1, Mar. 2019, Art. no. e13142, doi: 10.2196/13142.

[75] M. Yazdanbakhsh and P. G. Kremsner, "Influenza in Africa," *PLoS Med.*, vol. 6, no. 12, Dec. 2009, Art. no. e1000182, doi: 10.1371/journal.pmed.1000182.

[76] National Institute for Communicable Diseases. *Weekly Influenza and Respiratory Syncytial Virus Surveillance Report Week 19, 2019*. Accessed: Jul. 30, 2019. [Online]. Available: http://www.nicd.ac.za/wpcontent/uploads/2019/05/WeeklyRespiratory-pathogens-surveillancereport-Flu-RSV-week19-of2019FINAL.pdf

[77] *Google Trends: Understanding the Data*. Accessed: Aug. 1, 2019. [Online]. Available: https://storage.googleapis.com/gweb-news-initiativetraining.appspot.com/upload/GO802_NewsInitiativeLessons_Fundamentals-L04-GoogleTrends_1saYVCP.pdf

[78] R. J. Hyndman and Y. Khandakar, "Automatic time series forecasting: The forecast package for R," *J. Stat. Softw.*, vol. 27, no. 1, pp. 1–22, 2008, doi: 10.18637/jss.v000.i00.

[79] G. E. P. Box and G. M. Jenkins, *Time Series Analysis, Forecasting and Control*. London, U.K.: Holden-Day, 1970.

[80] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Roy. Stat. Soc. B, Stat. Methodol.*, vol. 67, no. 2, pp. 301–320, 2005, doi: 10.1111/j.1467-9868.2005.00503.x.

[81] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *J. Statist. Softw.*, vol. 33, no. 1, pp. 1–22, 2010, doi: 10.18637/jss.v033.i01.

[82] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for Cox's proportional hazards model via coordinate descent," *J. Stat. Softw.*, vol. 39, no. 5, pp. 1–13, 2011, doi: 10.18637/jss.v039.i05.

[83] H. Drucker, C. J. Burges, L. Kaufman, A. J. Smola, and V. Vapnik, "Support vector regression machines," in *Proc. Neural Inf. Process. Syst.*, vol. 9, 1997, pp. 155–161. Accessed: Apr. 7, 2021. [Online]. Available: http://ci.nii.ac.jp/naid/10018343800/en/

[84] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, F. Leisch, C. C. Chang, C. C. Lin, and M. D. Meyer, "Package 'e1071,'" *R J.*, 2019.

[85] D. Svozil, V. Kvasnička, and J. Pospíchal, "Introduction to multi-layer feed-forward neural networks," *Chemometrics Intell. Lab. Syst.*, vol. 39, no. 1, pp. 43–62, Nov. 1997, doi: 10.1016/S0169-7439(97)00061-0.

[86] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/NECO.1997.9.8.1735.

[87] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 2nd ed. Melbourne, VIC, Australia: OTexts, 2018.

SEUN O. OLUKANMI received the bachelor's degree in computer science from Ladoke Akintola University of Technology, Nigeria, in 2010, and the M.Sc. degree in computer science from the University of KwaZulu-Natal, South Africa, in 2016. She is currently pursuing the Ph.D. degree with the Institute for Intelligent Systems, University of Johannesburg, South Africa. Her current research interests include artificial intelligence and data science for the social good.

FULUFHELO V. NELWAMONDO (Senior Member, IEEE) received the B.Sc. and Ph.D. degrees in electrical engineering (computational intelligence) from the University of the Witwatersrand, Johannesburg, South Africa.

He is currently a Visiting Professor of electrical engineering with the Institute for Intelligent Systems, University of Johannesburg. He has published over 140 academic articles in artificial intelligence. He is a member of the South African Institute of Electrical Engineers and a Senior Member of the Association of Computing Machinery (ACM). He became the Youngest Recipient of the Harvard–South African Fellowship Program, in 2008, and was awarded the Silver Order of Mapungubwe by the President of South Africa, in 2017. He is a Registered Professional Engineer (Pr. Eng) with the Engineering Council of South Africa.

NNAMDI I. NWULU (Senior Member, IEEE) is currently a Full Professor with the Department of Electrical and Electronic Engineering Science, University of Johannesburg, and the Director of the Centre for Cyber Physical Food, Energy and Water Systems (CCP-FEWS). His research interests include the application of digital technologies, mathematical optimization techniques, and machine learning algorithms in food, energy, and water systems. He is a Senior Member of the South African Institute of Electrical Engineers (SMSAIEE). He is a Professional Engineer registered with the Engineering Council of South Africa (ECSA) and a Y-rated Researcher with the National Research Foundation, South Africa. He is the Editor-in-Chief of the *Journal of Digital Food Energy and Water Systems* (JDFEWS) and an Associate Editor of the *IET Renewable Power Generation* (IET-RPG) and the *African Journal of Science, Technology, Innovation and Development* (AJSTID).