Leveraging Google Search Data and Artificial Intelligence Methods for Provincial-level Influenza Forecasting: A South African Case Study

https://doi.org/10.3991/ijoe.v18i11.29899

Seun O. Olukanmi(✉), Fulufhelo V. Nelwamondo, Nnamdi I. Nwulu
Department of Electrical and Electronic Engineering Science, University of Johannesburg, Johannesburg, South Africa
seun_fagbemi@yahoo.com

Abstract—This paper investigates the usefulness of Google search patterns with Artificial Intelligence (AI) techniques for timely influenza-like illness (ILI) forecasting for each of the nine South African provinces. Traditional surveillance methods are limited by delay in reporting. Existing digital disease surveillance studies that employ alternative online data have scarcely explored sub-Saharan African countries. In South Africa, Google search data has only been recently studied for ILI surveillance at the national level. Meanwhile, the differences in socio-economic and technological conditions across provinces call for finer spatial investigation. We perform correlation analysis between Google trends (GT) data for 21 ILI-related terms and real-life ILI surveillance data for each province. Next, we develop models to assess the predictive performance of these GT data for forecasting ILI rates, using time series, machine learning, and deep learning methods. We observe sufficient correlation for only two of the nine provinces: Gauteng and Western Cape. Thus, GT data could only be used to forecast ILI in these two provinces. Interestingly, these two provinces are regarded as the most economically developed. In the other seven provinces, LSTM, a deep learning technique, gives more accurate predictions than a baseline autoregressive model when only past ILI data are used for forecasting future ILI trends. The results reveal that, for provinces for which GT data is sufficiently available, it is not only free and fast, but is an effective predictor on its own as well as when added to past ILI data for forecasting future ILI infection rates. The correlation analysis suggests an association between provincial socio-economic development and the use of digital platforms for disease surveillance. Overall, the study established the need for finer scale ILI forecasting which will inform targeted planning for disease surveillance and interventions.

Keywords—influenza forecasting, regional ILI surveillance, artificial intelligence, deep learning, machine learning, Google trends, digital epidemiology, infodemiology, infoveillance

1 Introduction

Influenza (flu), a severe respiratory illness, remains a public health burden worldwide. Every year, it is responsible for millions of deaths and/or hospitalizations globally [1]. In South Africa, it causes more than 10 000 deaths annually [2], [3]. Accurate and timely predictions of disease incidence at national and regional scales can reduce the impact of outbreaks and facilitate targeted and effective public health responses and interventions. Efforts are on the increase by health care systems around the world towards collecting large amounts of quality disease incidence data. For example, the Centers for Disease Control (CDC) in the United States records ILI surveillance and publicly makes available datasets at different geographic scales [4]. In South Africa, the Department of Health set viable surveillance as a basic goal of the national influenza policy and strategic plan put in place for 2017 to 2021. The surveillance presently utilizes reports from hospitals and general practitioners and is anchored by the National Institute for Communicable Diseases (NICD). However, these systems are costly and plagued by a delay of up to two weeks before the surveillance reports are available.

This limitation of delay, together with the increased availability of personal health information shared/collected online in today's age of big data, has given rise to a new research area termed digital epidemiology. Salathe [5] defined digital epidemiology as "epidemiology that uses data generated outside the public health system for disease surveillance". Some of the Internet-based data streams that have been explored in past studies include search engine [6]–[10] and Twitter data [11]–[14], news alerts [15]–[17], Wikipedia [18]–[20] and health-related blogs and websites [21]. Google trends (GT) is a commonly used search queries data source, which gives the relative amount of Google searches for particular queries in a given location. Several studies have demonstrated the usefulness and advantages of these online data for improved surveillance of different diseases such as ILI [22]–[24], tuberculosis, hepatitis [25], [26], Type 2 diabetes [27], zika [28], Ebola [29], AIDS [30], cancer [31], Lyme disease [32], dementia [33] , and the recent COVID-19 pandemic [34]–[36].

Previous studies in the digital epidemiology field have made use of artificial intelligence (AI) methods such as statistical ARIMA/ARIMAX time series models [37]–[40], conventional machine learning techniques including support vector machines (SVM), random forests, linear regression, elastic net [41]–[44], and recently, deep learning models [45]–[48] due to their competitive performance. However, it was noted in [49] that due to differences in economic, technological, and cultural dispositions, the results across countries cannot be generalized. This claim is corroborated by the works of Cervellin et al. [50] and Bilge et al. [51] who found no significant correlation between Google search data and the real-world disease records. Meanwhile, 56% of studies reviewed by Abad et al. [52] originated from the US alone. Only a few studies have focused on African countries [41], [53].

For South Africa, a few studies have emerged in the last year. One recent study investigated and established significant correlation between Google search volume for some ILI-related terms and real-world ILI incidence data [54]. A more recent study demonstrated the performance of Google search data for predicting ILI incidence in South Africa [10]. However, these studies have only focused on monitoring ILI at the

national level. The predictive performance of Google search data for ILI surveillance at the provincial level remains unstudied. National level forecasts provide a high-level overview of disease incidence across the country, offering value to national public health officials but giving only broad information to health officials at the provincial level [55]. Disease surveillance at finer geographic scales presents a potential for targeted interventions and population-specific decision-making that aligns with public health infrastructure available at such levels [55]. This work aims to address this gap.

2 Materials and methods

2.1 Data

The real-world ILI data for each of the nine provinces of South Africa was provided by the National Institute for Communicable Diseases' (NICD) viral watch influenza surveillance program [56]. The case definition of ILI includes a fever (temperature ≥ 38°C) and cough or sore throat with symptoms starting within the last 10 days. The anonymized ILI data are the weekly numbers of patients who meet such case definition from week 1 in 2010 to week 43 in 2018 (459 weeks) in each province.

Over the same study period (week 1 in 2010 to week 43 in 2018), we downloaded the weekly search index for 21 influenza-related terms from Google trends (GT). These 21 terms are the terms that showed significant correlation with the actual national ILI records [54]. GT is a web tool that is freely accessible and returns anonymized, aggregated, and normalized search volume (NSV) for user queries from a particular location or region. If the search amount for a term is too low for a specified time period, GT returns 0 as the NSV [57].

2.2 Provincial correlation analysis

Using the 21 ILI-related terms, we performed Pearson's correlation analysis to test for the association between the weekly Google search data of each of these terms and the weekly provincial records of ILI over each of the epidemiological years (from 2010 to 2018) as defined in [54]. We excluded all rows with missing values from the ILI records of each province before performing the correlation analysis. Furthermore, there were no ILI records for KwaZulu Natal province from 2015 to 2018, so these years were left out of the province's correlation analysis. The analysis was done using the *cor.test* function from the R *stats* package and we set significance at $p < 0.05$.

2.3 Data preparation for the models

Due to the presence of missing instances and outliers in the ILI data for each province, we performed some cleaning on the training and test data in order to enhance the forecasting capabilities of the models. Missing instances and outliers in the training data were replaced using the t*sclean* from the R *forecast* package [58], while missing values in the test set were estimated using *na.interp* also from the *forecast* package.

2.4 Algorithms

We used a range of statistical, machine learning, and deep learning algorithms that have been previously applied for influenza forecasting [10], [41], [43], [44], [46], [47] such as seasonal ARIMA (SARIMA), multiple linear regression (MLR), elastic net (EN), support vector machine regression (SVM), feedforward neural network (FNN), and long short-term memory (LSTM).

The SARIMA technique [59] is an extension of the commonly used ARIMA time series forecasting method. It allows the modeling of the seasonal component of the input data. When external regressors are included, then it is referred to as SARIMAX. The *R auto.arima* function was used to implement the SARIMA models [58].

Multiple linear regression predicts a target response variable using several explanatory variables. The *lm* function in *R* was used for the implementation of the MLR models in this study.

Elastic net regression involves a combination of the penalties of both the Least Absolute Shrinkage and Selection Operator (LASSO) and ridge methods [60]. The *R cv.glmnet* function [61], [62] was used for the implementation of the elastic net models.

The SVM regression technique relies on kernel functions to map the explanatory variables into higher dimensional spaces, making the data linearly solvable [63]. These models were implemented using the *svm* function in *R (e1071* package) [64].

Feedforward neural networks comprise nodes arranged into the input, hidden and output layers. It is termed feedforward because the connections between the nodes do not form a cycle [65].

The LSTM network was designed to mitigate the problem of short-term memory that is encountered with the basic recurrent neural networks (RNNs) [66]. It is a special type of RNN architecture that has cyclic connections linking the nodes, which makes important information to persist LSTMs are found in time series data processing because their architecture is naturally suited to process sequences and lists. We used the *Keras* library with *Tensorflow* backend in Python to implement the FNN and LSTM methods.

2.5 Experimental model(s) per province

We describe the models that were deployed for forecasting future influenza trends for each province in this section. These models stem from the AI algorithms outlined in the previous section and the important input features for each province. For easy referencing and comparison, we maintain the same naming convention for the various provincial models as in the South African national ILI forecasting study [10]. For example, GT-MLR is a MLR model fitted to GT data only, while ILI-SARIMA is a SARIMA model based on past ILI data only. The provincial models fall under three categories and are described as follows:

ILI Trends as a Function of GT data only (Gauteng and Western Cape Provinces). These set of models apply only to the Gauteng and Western Cape provinces. The other seven provinces were excluded due to the sparsity of Google search data and the consequent low correlation with the real-world ILI records. For these models, we supplied only the provincial Google search volume of the 21 ILI-related terms as inputs

(independent variables) in order to predict the ILI incidence rates (dependent variable) for zero to two weeks ahead. This is important for the evaluation of Google search data alone for influenza surveillance at the provincial level in South Africa in the case of lack of the real ILI records. The models were trained on 367 instances (80%) of the data while their performances were evaluated on the remaining 20%. We maintained the same number of training samples for the one and two weeks-ahead forecast models and reduced the test data by 1 and 2 instances respectively. Similar to Ref [10], the four sets of models under this category include GT-EN, GT-FNN, GT-MLR, and GT-SVM. The specific parameters tuning for the GT-FNN models for each of the two provinces concerned are outlined below:

***Gauteng**:* The GT-FNN models for Gauteng have 21 nodes in the input layer which depict the Google search volume of the 21 ILI-related terms. In contrast, the output layer has just one node which is the predicted ILI incidence rate. The optimal hidden layer parameters for the nowcast models (zero week ahead forecasts) were fixed experimentally. Four hidden layers with 1024, 512, 512, and 256 units were selected, each with the *relu* activation function and a dropout rate of 0.2 used after each layer to avoid model overfitting, Similarly, four hidden layers with 512, 512, 256 and 128 units were fixed for the one week ahead models, all with a dropout rate of 0.2 following each layer. The parameter settings for the two weeks ahead forecast models were the same as in the one week ahead models except that the last hidden layer did not require a dropout technique. *Sigmoid* activation function was used in the output layer of all the GT-FNN models while the *adam* optimizer was used for compilation. The Gauteng models were trained for 100 epochs.

Western Cape: Like the Gauteng GT-FNN models, the Western Cape models also have 21 input nodes. The nowcast models (zero week ahead prediction models) have 512, 256, 256, and 128 nodes in the four hidden layers respectively. All the hidden layers had the *relu* activation function but with no dropout layer at all. There are also four hidden layers in the one week ahead forecast models, with 512, 256, 256, and 128 units from the first to the last layer. No dropout technique was applied here as well. The two weeks ahead prediction models have the same parameters as the one week ahead forecasts. All the models have the *relu* activation function in the hidden layers and the *adam* optimizer for compilation. The zero and one week ahead forecast models were trained for 50 epochs while the two-weeks ahead prediction model learned for 40 epochs.

Future ILI Trends as a Function of Past ILI Data Only (Al Provinces). The ILI-data only models for each of the nine provinces take the historical ILI incidence rates as input and forecast the future ILI incidence rates as output. Like in Ref. [10], the two-time series modelling techniques that were studied are the ILI-SARIMA and ILI-LSTM models. The ILI-SARIMA models are centered on the ARIMA algorithm (SARIMA) and we maintained the 80/20% training/evaluation set size as in the GT-data only models. For the ILI-LSTM models, the input data for each province was reshaped as 4 time-steps (4 weeks) for predicting the ILI incidence rate for the following week (the next time-step). The output layer has 1 node representing the predicted ILI rate. The ILI-LSTM models have the same training/evaluation split size as in the ILI-SARIMA

model and the optimal model parameters for each province were determined experimentally by evaluating the effect on the models' forecasting performance. All the models used the *adam* optimizer for compilation. The specific ILI-LSTM model settings for each province are given below:

Eastern Cape: The Eastern Cape model is a simple LSTM layer of 200 units with a dropout technique (rate = 0.2). Due to the wide range of the input data, we performed min-max normalization to rescale the data to the range [0,1], and the model learned for 50 epochs.

Free State: The ILI-LSTM model for the Free State province is a stack of two LSTM layers, each with 100 units and a dropout technique with a rate of 0.1 following each LSTM layer. For this model, no input data scale normalization was necessary, and the model learned for 200 epochs.

Gauteng: The Gauteng ILI-LSTM model is also a stack of 2 LSTM layers. Each layer has 200 units, and a dropout technique (rate = 0.2) followed each layer. Min-max rescaling of the input data was done, and the model learned for 100 epochs.

Kwazulu-Natal: A single LSTM layer with 200 units was used with a dropout rate of 0.2 following the layer. No input data rescaling was done, and the model learned for 50 epochs.

Limpopo: A stack of two LSTM layers with 100 units each was used, and a dropout technique (rate = 0.1) followed each LSTM layer. No data rescaling was necessary, and the model learned for 200 epochs.

Mpumalanga: A stack of two LSTM layers with 100 units each was also used for the Mpumalanga province, with a dropout (rate = 0.2) following each layer. Like the Limpopo ILI-LSTM model, the model was trained for 200 epochs.

Northern Cape: The model parameter settings for Northern Cape province is the same with the Limpopo province.

North West: The Limpopo model parameters also apply to the North West province but here, the model was trained for 150 epochs.

Western Cape: For the Western Cape ILI-LSTM model, a single layer of LSTM (200 units) was used, and a dropout technique (rate = 0.2) followed the layer. The model learned for 100 epochs.

ILI Trends as a Function of GT and Past ILI Data (Gauteng and Western Cape Provinces). These models apply to the Gauteng and Western Cape provinces only because of the scantiness of GT data in the other seven provinces. The input here is a combination of the historical ILI incidence rates and the Google search volume of the 21 terms. The training/evaluation data sizes remained the same as in the ILI-SARIMA model. Similar to [10], the models under this category include ILI-GT-EN, ILI-GT-FNN, ILI-GT-MLR, ILI-GT-LSTM and ILI-GT-SARIMAX. They are grouped into two classes: the first four models (ILI-GT-EN, ILI-GT-FNN, ILI-GT-MLR, and ILI-GT-SVM) belong to the class of machine learning/deep learning regression models that add ILI data from the past one or two weeks as part of the explanatory features, while the last two models (ILI-GT-LSTM and ILI-GT-SARIMAX) belong to another class of statistical/deep learning time series techniques that extend the ILI data with GT data. The suitable parameter settings for the models were selected experimentally. The implementation

details of the ILI-GT-FNN and ILI-GT-LSTM models for the two provinces are described below:

Gauteng ILI-GT-FNN: The optimal parameters for the model that includes the ILI data of the past one week are: four hidden layers with 26, 59, 59, and 160 units, each layer with a *relu* activation function and a dropout technique (rate = 0.2) except the last hidden layer. Likewise, the model incorporating the ILI data of the past two weeks has four hidden layers with 29, 58, 58 and 168 units respectively. Each hidden layer except the last one also has a dropout layer (rate = 0.2) following it. The two Gauteng ILI-GT-FNN models used the *adam* optimizer for compilation and were trained for 100 epochs.

Western Cape ILI-GT-FNN: For the Western Cape province, the model incorporating the past one-week ILI data has four hidden layers with 29, 58, 58, and 136 nodes respectively. All the hidden layers had the *relu* activation function and no dropout layer was applied. Similarly, the model including the past two weeks ILI data has four hidden layers with 26, 58, 58, and 136 units respectively. No dropout technique was applied to this model as well. Both models also used the *adam* optimizer for compilation and were trained for 40 epochs.

Gauteng ILI-GT-LSTM: This model comprises two LSTM layers with 300 and 200 units respectively. The input data was reshaped as in the ILI-LSTM models. A dropout technique (rate = 0.1) was applied after each LSTM layer, and the model used the *adam* optimizer for compilation and learned for 50 epochs.

Western Cape ILI-GT-LSTM: The Western Cape model also has two stacked LSTM layers with 50 units each. A dropout method (rate = 0.1) followed the first LSTM layer, while no dropout layer was used after the second LSTM layer. The model was trained for 40 epochs.

2.6 Evaluation metrics

For the correlation analysis per province, the higher the Pearson's correlation coefficient (PCC), the stronger or better the association between the GT data of a particular term and the true ILI data.

The forecasting performances of the different models were compared using a few other popular evaluation metrics determined on the evaluation data. This includes the root mean squared error (RMSE) and mean absolute error (MAE) of the forecasted and the real ILI occurrence. Lower RMSE and MAE values, and higher PCC values depict better model performance. The capability of the models to accurately predict the peak week of ILI occurrence, and the magnitude of the peak was also evaluated. As in Ref. [10], "peak week difference (PWD) is calculated as the difference between the forecasted and real peak week", while the "peak magnitude difference (PMD) is the difference between the estimated and the true ILI peak height". Lower values of these two metrics indicate better model performance.

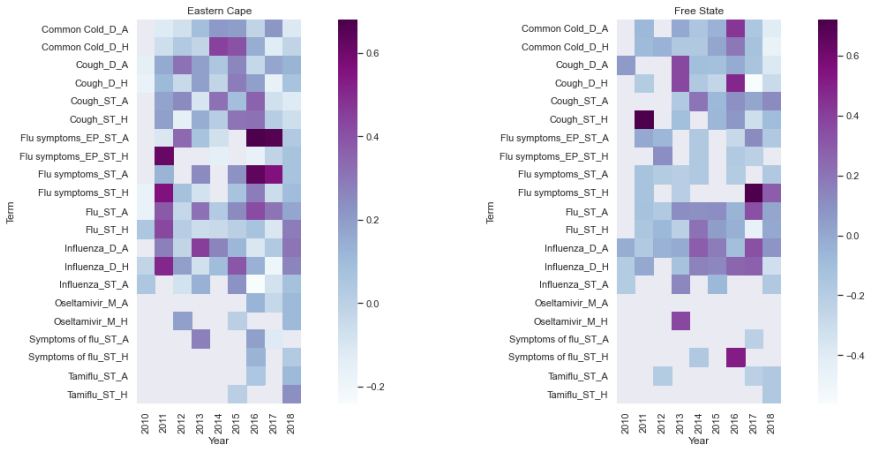
3 Results

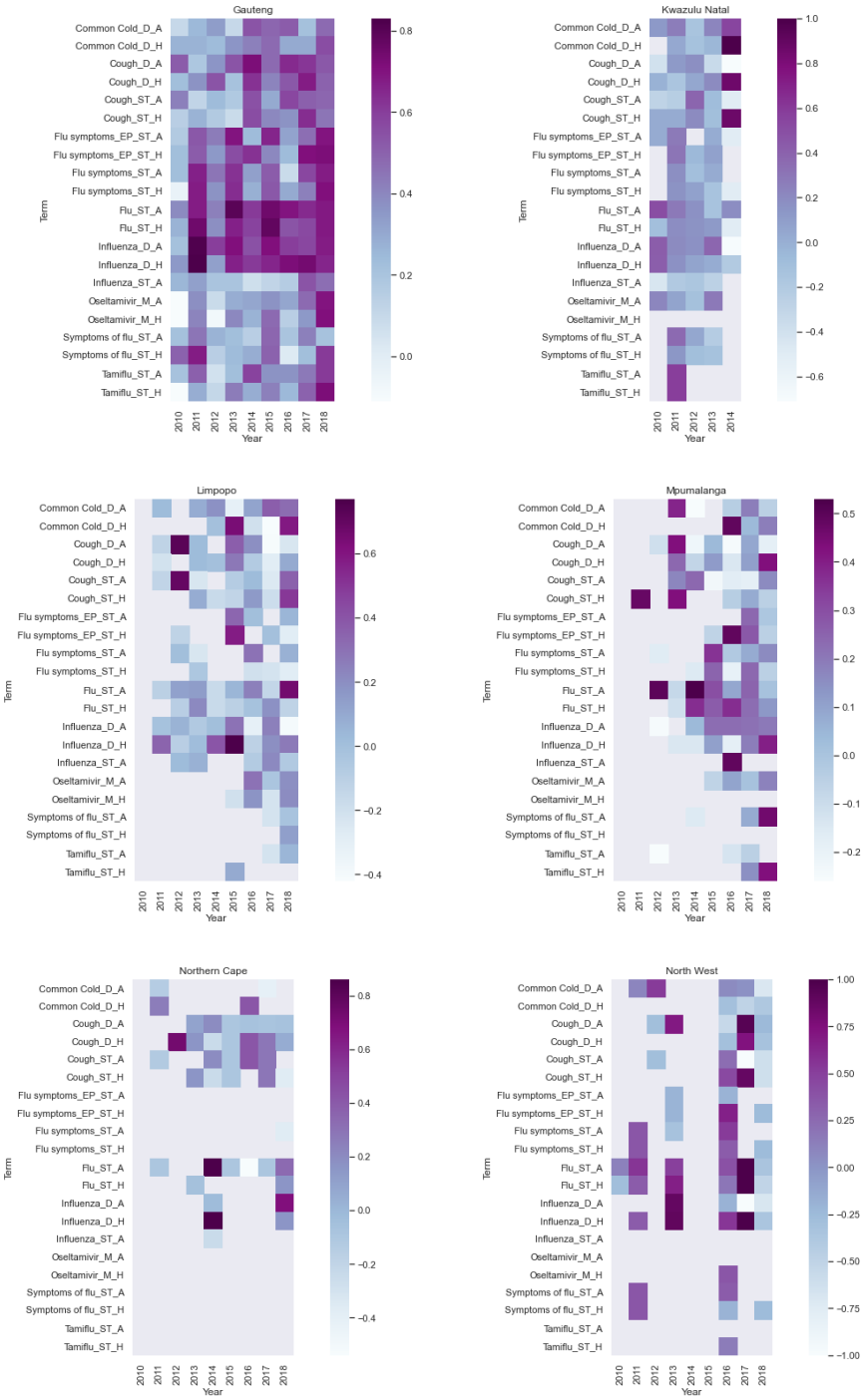
3.1 Correlation coefficients per province

The coefficients of the correlation between GT data of each of the ILI-related terms, and the ILI records of each province can be visualized in the heatmaps presented in Figure 1. The darker the colour of the heatmap cells, the higher the correlation coefficient. We interpret correlation coefficient $r = 0.5 < 0.7$ as moderate correlation and $r \geq 0.7$ as strong correlation. The heatmaps show that the correlation coefficients are significantly higher in Gauteng and Western Cape provinces but are low and sparse in the other seven provinces. There are also many missing values in the results of the correlation analysis across the years in these seven provinces. Due to this, Google search data was excluded from the ILI forecasting models for those provinces.

Gauteng: The highest correlation coefficient ($r = 0.83; p < 0.05$) recorded was for term *influenza (sub-category: Disease, Category: Health)* in 2011, followed closely by the terms*: influenza (Disease, All, 0.82)* also in 2011, and *flu (Search term, All, 0.80)* in 2013.

Western Cape: For Western Cape, the highest correlation coefficient ($r = 0.68; p < 0.05$) was obtained for a similar term *influenza* (*Disease, All*) in 2011, followed by *cough* (*Search Term, Health, 0.67*) in 2017, *influenza (Disease, Health*) in 2015, flu (Search Term, Health, 0.66), and flu (Search Term, All, 0.65) both in 2017.




Gauteng
Kwazulu Natal
Limpopo
Mpumalanga
Northern Cape
North West
Common Cold_D_A
Common Cold_D_H
Cough_D_A
Cough_D_H
Cough_ST_A
Cough_ST_H
Flu symptoms_EP_ST_A
Flu symptoms_EP_ST_H
Flu symptoms_ST_A
Flu symptoms_ST_H
Flu_ST_A
Flu_ST_H
Influenza_D_A
Influenza_D_H
Influenza_ST_A
Oseltamivir_M_A
Oseltamivir_M_H
Symptoms of flu_ST_A
Symptoms of flu_ST_H
Tamiflu_ST_A
Tamiflu_ST_H
Term
Year

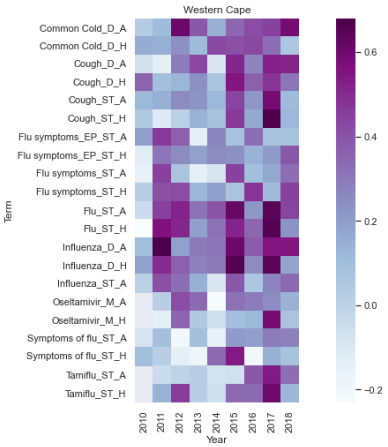


Fig. 1. The Correlation between 21 ILI-related search terms and ILI incidence rates (per year from 2010 – 2018) for each province. The darker the colour of the heatmap sells, the higher the correlation coefficient

3.2 Performance of the GT-Data only models (Gauteng and Western Cape Provinces)

In this section, we present the performance of the models that incorporate GT data only for Gauteng and Western Cape provinces.

RMSE and MAE

Gauteng: The RMSE values for the Gauteng GT-data only models ranged from 10.32 to 13.30, with SVM and the elastic net (EN) techniques showing the best and worst performance respectively. Similar to the national study [10], the forecast error grows as the forecast horizon increases. The mean absolute error (MAE) values for the Gauteng models ranged from 7.84 (GT-SVM) to 11.27 (GT-EN). The deep learning model (GT-FNN) had comparable performance with the SVM-based counterpart.

Western Cape: The range of RMSE values for Western Cape is from 6.93 to 8.18 for the nowcasting scenario. Similar to the Gauteng models, GT-SVM models performed the best while GT-EN models had the poorest performance. GT-SVM (4.85) had the lowest MAE value for the Western Cape province, followed by GT-FNN (5.74) and GT-MLR (5.97), while the highest value is from GT-EN (6.34).

A visualization of the RMSE and MAE values for the nowcasting scenario for both provinces can be seen in Figures 2 and 3 respectively.

PCC. For both provinces, the Pearson's correlation coefficient values reduced as the forecast horizon increases. In Gauteng, GT-SVM had the highest value of 0.8050 for the same week predictions, followed by GT-EN (0.8005), GT-FNN (0.7565) and GT-MLR (0.7266) respectively. However, for one and two weeks ahead forecasting, the GT-EN models gave the best values.

For the Western Cape province, the EN-based models had the highest values (0.7567) for same week predictions, with GT-SVM having a comparable value of

0.7246. At the one and two weeks ahead onwards, the GT-SVM model had the highest PCC values.

PWD and PMD. For Gauteng, the GT-MLR model estimated the week of the two peak influenza seasons in the evaluation period accurately. The GT-EN and GT-SVM model had similar predictions for the first peak week (5 weeks earlier) while the second peak week was estimated correctly by all the models.

For Western Cape, the nowcast models all had accurate predictions for the first peak week except the GT-SVM model. The second peak week was predicted as two and three weeks after the real peak week by the GT-MLR and GT-EN models respectively. On the other hand, GT-SVM and GT-FNN estimated the second peak week as two weeks prior to the real peak week. For the one-week onward estimates, GT-SVM and GT-FNN forecasted the first flu peak week accurately while GT-MLR and GT-EN predicted it as one week later. For the second peak week, all the models forecasted it as one week sooner. Only GT-SVM and GT-FNN predicted the two peaks correctly at the two weeks onwards estimates.

Similar to the national study [10], GT-MLR performed the best in estimating the size of the two peaks at all the forecast horizons, while the GT-EN showed the worst performance.

Figure 4 is a visualization of the true versus predicted ILI rates over the test period for Gauteng while Figure 5 is the visualization for the Western Cape province.

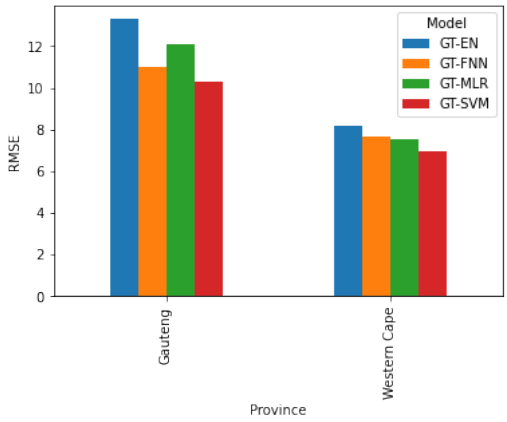


Fig. 2. RMSE of the GT-EN, GT-FNN, GT-MLR, and GT-SVM models for Gauteng and Western Cape provinces

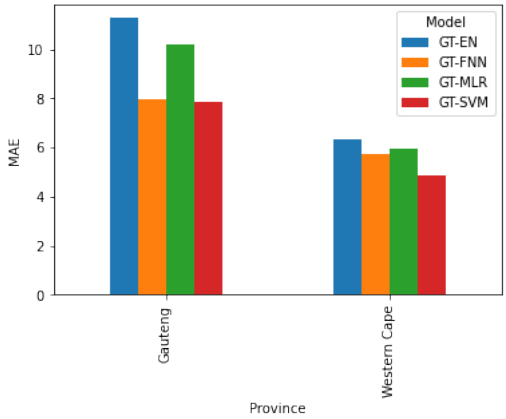
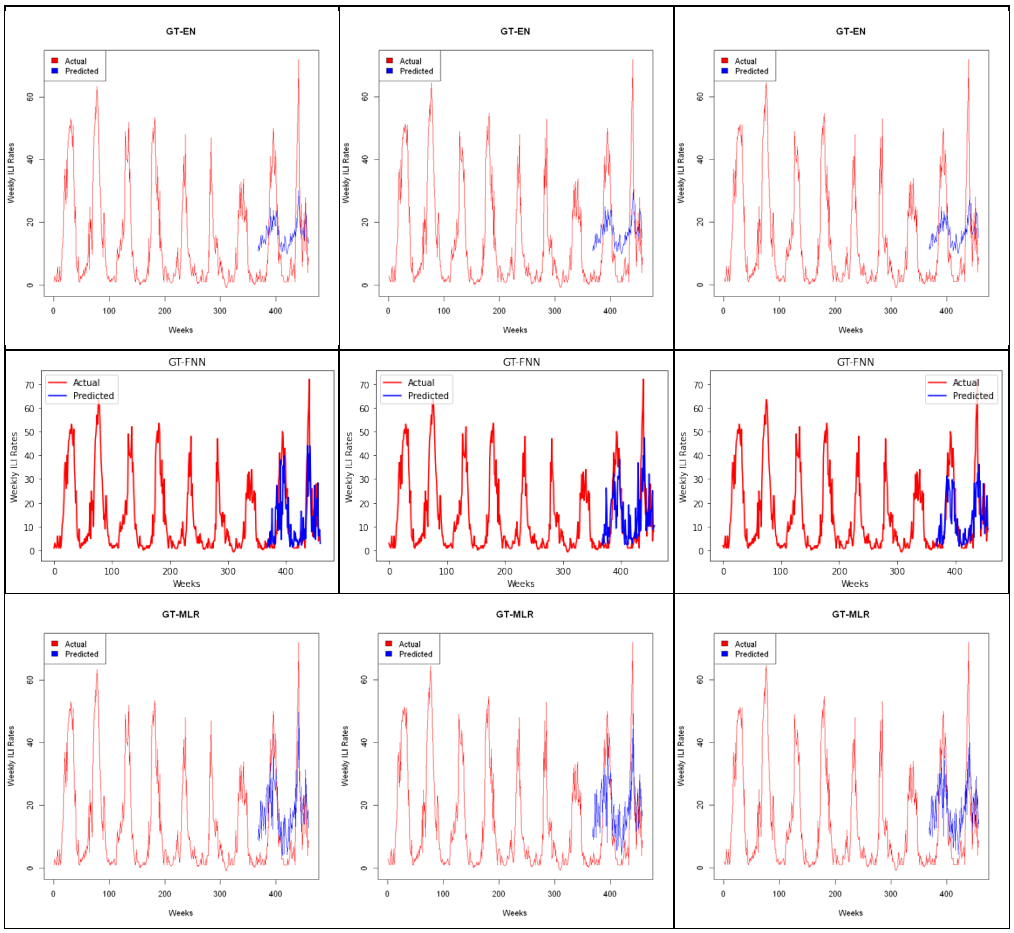


Fig. 3. MAE of the GT-EN, GT-FNN, GT-MLR, and GT-SVM models for Gauteng and Western Cape provinces



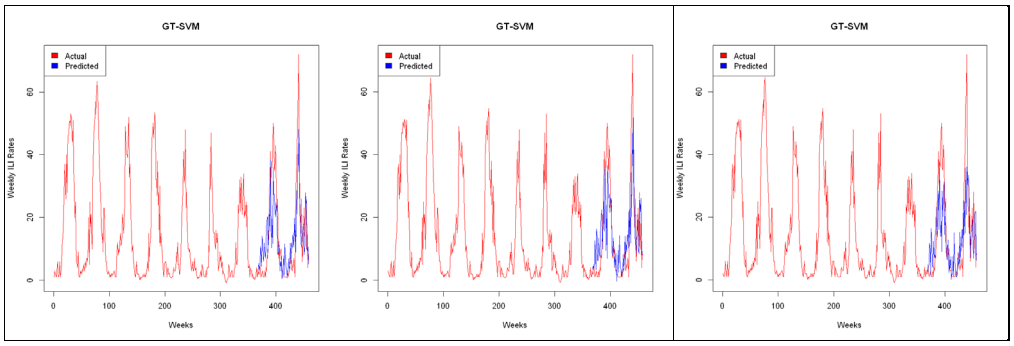
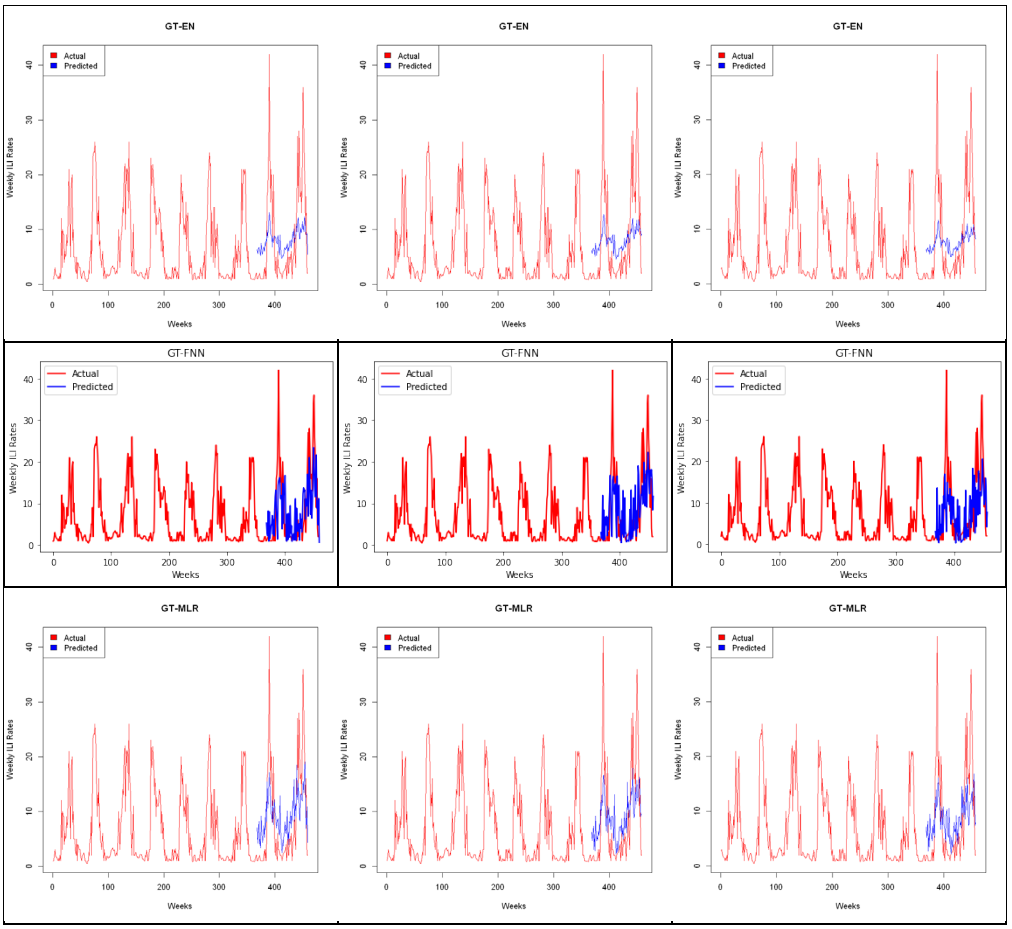


Fig. 4. True versus estimated weekly rates of ILI occurrence produced by the GT-EN, GT-FNN, GT-MLR, and GT-SVM models for Gauteng nowcasts (first column), one week onward (second column) and two weeks onward (third column) over the evaluation period



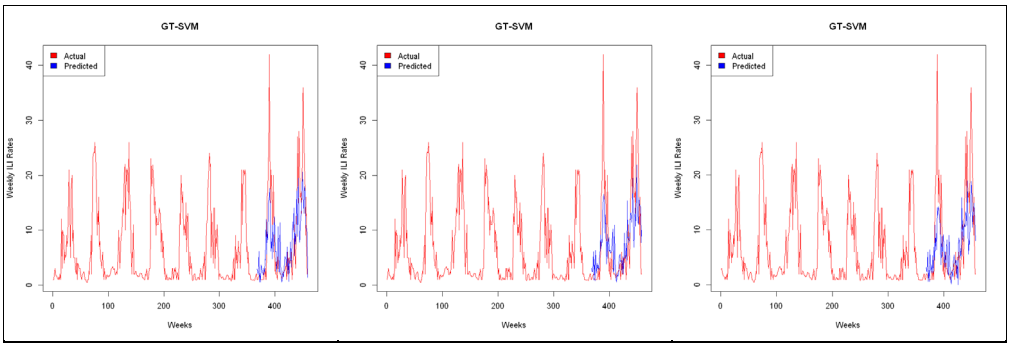


Fig. 5. True versus forecasted weekly ILI incidence rates produced by the GT-EN, GT-FNN, GT-MLR, and GT-SVM models for Western Cape nowcasts (first column), one week onward (second column) and two weeks onward (third column) over the evaluation period

3.3 Performance of the ILI-data only models (All provinces)

This section describes the performances of the models that incorporated past ILI data only for all the nine provinces. These are the ILI-SARIMA and ILI-LSTM models. We evaluate the performance of the models using the same metrics as in the GT-data models. Figures 6 and 7 gives a visualization of the RMSE and MAE values for all the provinces, while Figures 8 and 9 shows the real versus the predicted ILI incidence rates over the test period (indicating the models' performance in predicting the peak week and magnitude) for all the provinces. The figures reveal that ILI-LSTM, the deep learning times series model performs significantly better than the statistical ARIMA counterpart on all metrics.

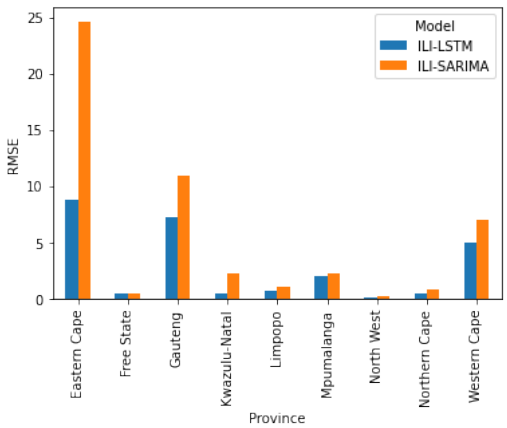


Fig. 6. RMSE values of the ILI-LSTM and ILI-SARIMA models for all provinces

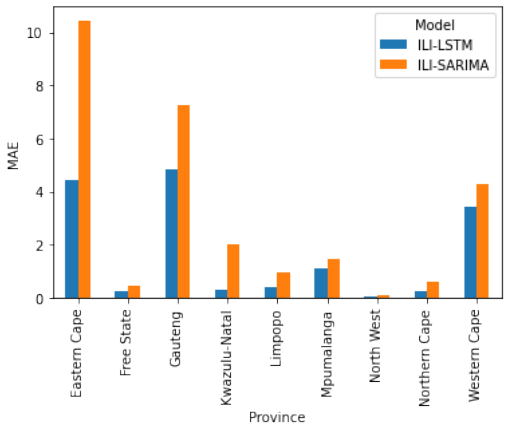
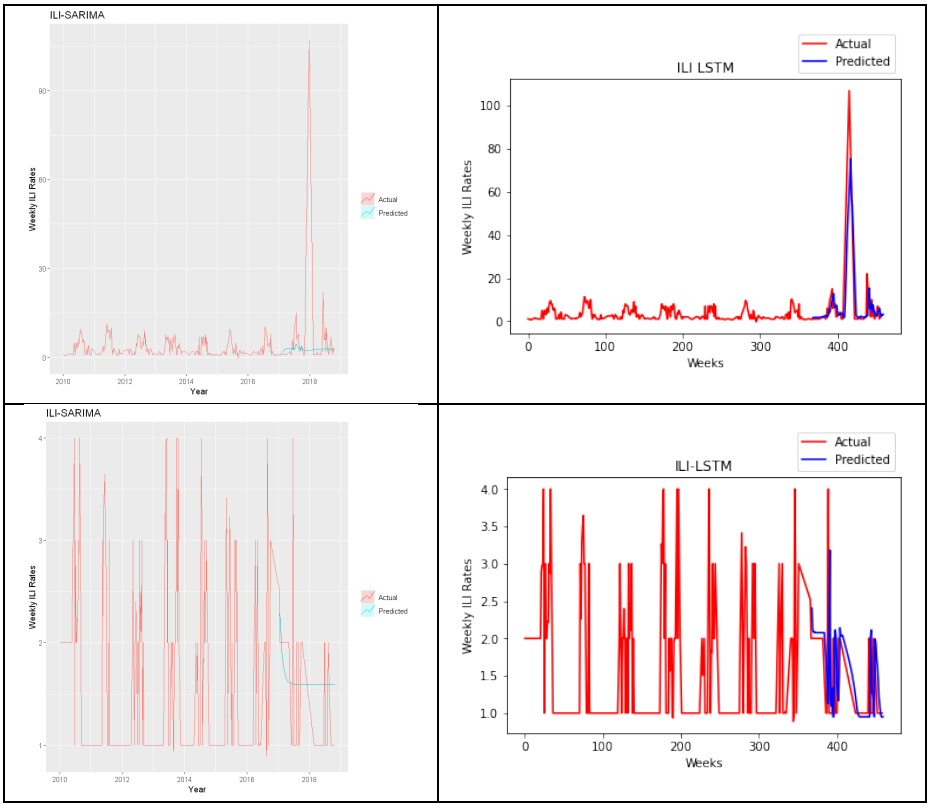


Fig. 7. MAE values of the ILI-LSTM and ILI-SARIMA models for all provinces



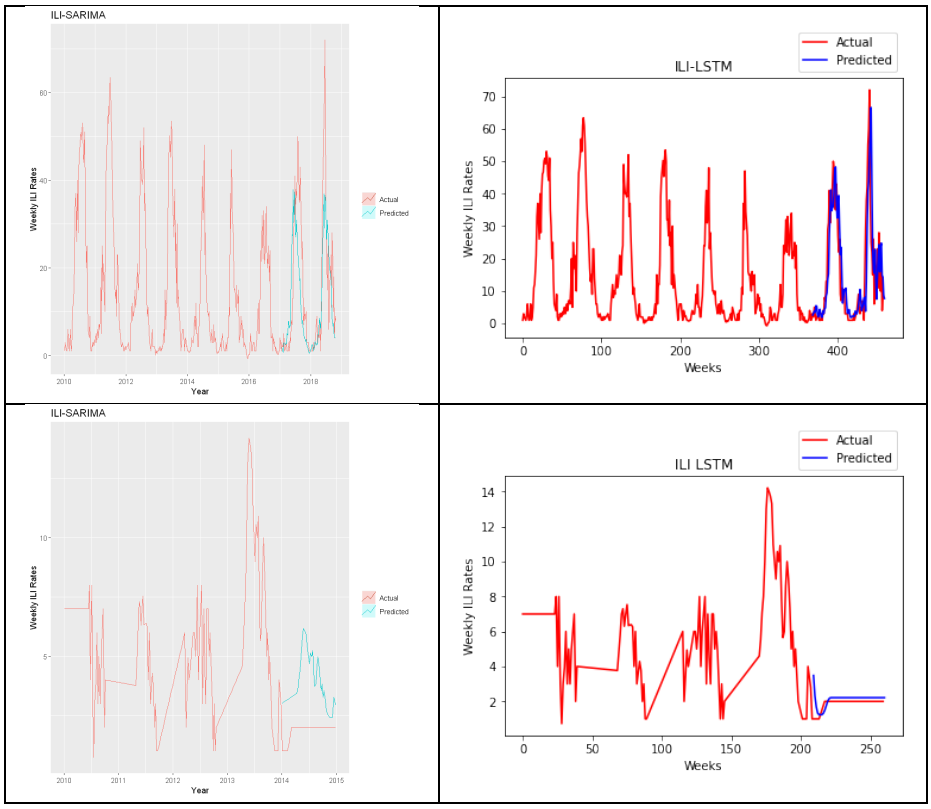
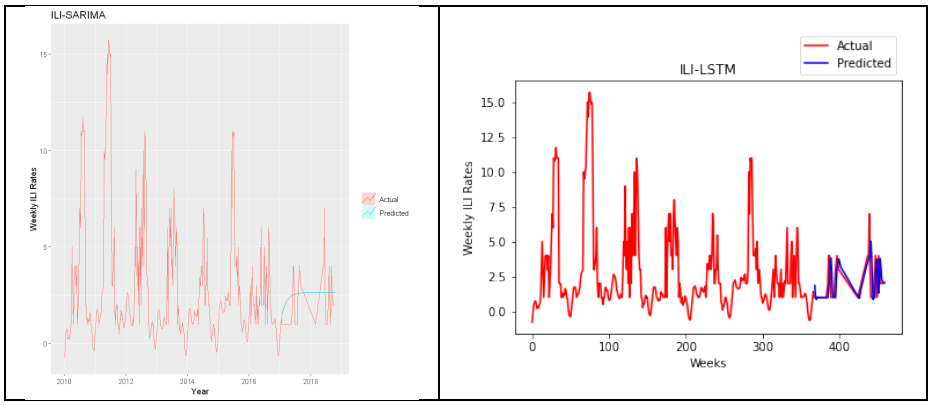
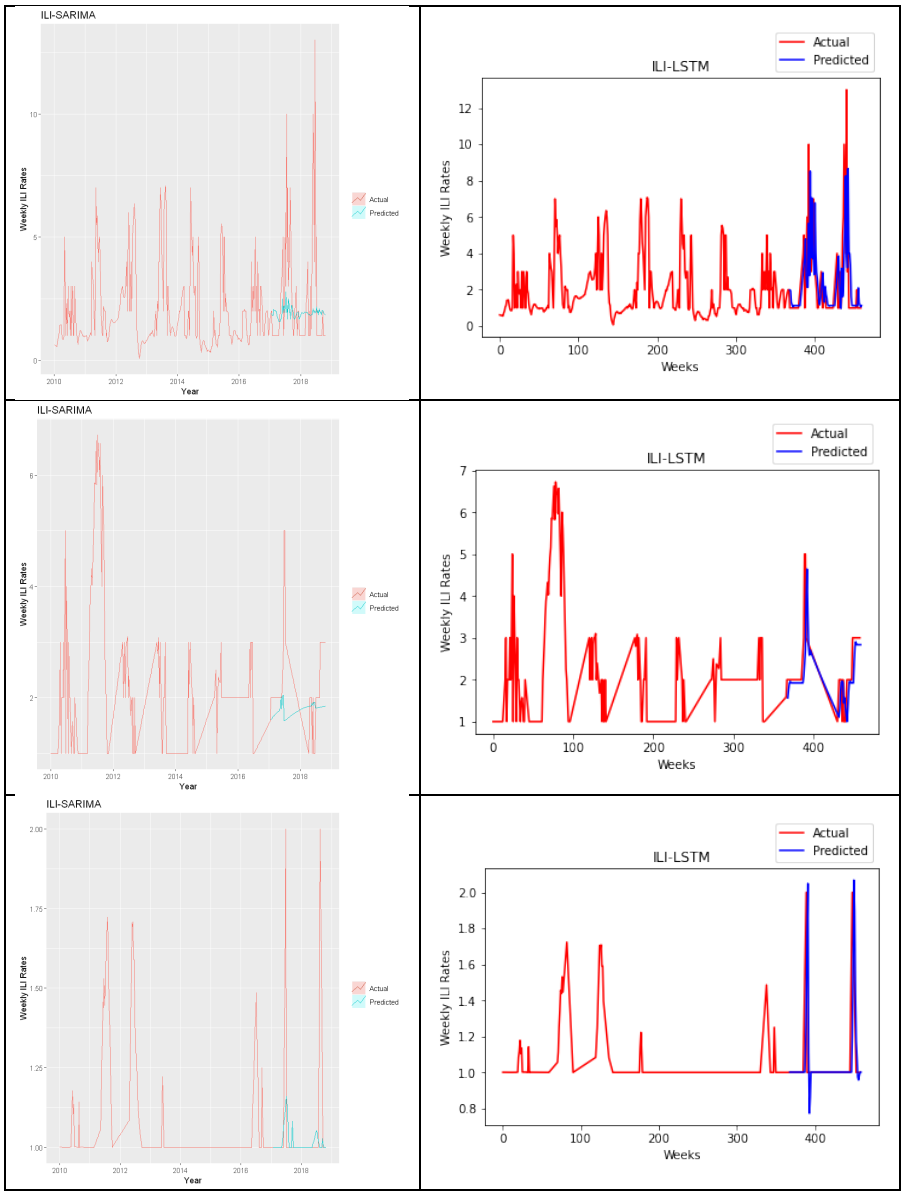


Fig. 8. True versus estimated weekly rates of ILI incidence from the ILI-SARIMA and ILI-LSTM models for Eastern Cape, Free State, Gauteng, and KwaZulu Natal provinces over the evaluation period




ILI-SARIMA
Weekly ILI Rates
Year
Actual
Predicted
ILI-LSTM
Weeks

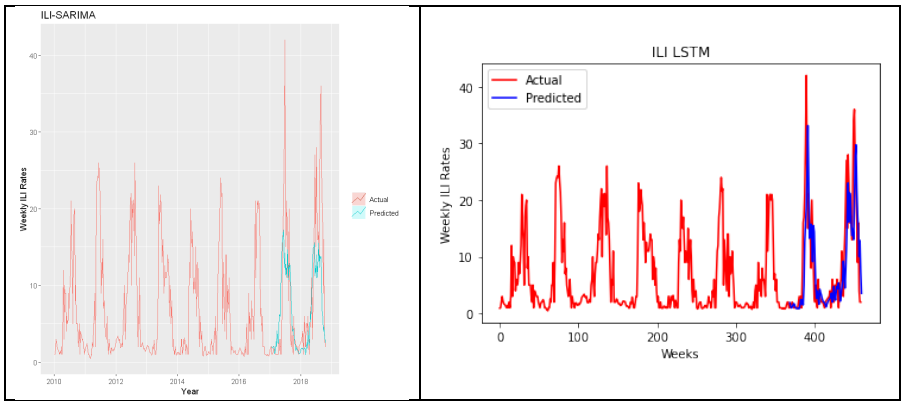


Fig. 9. True versus forecasted weekly rates of ILI incidence from the ILI-SARIMA and ILI-LSTM models for Limpopo, Mpumalanga, Northern Cape, North West, and Western Cape provinces over the evaluation period

3.4 Performance of the ILI-GT-data models (Gauteng and Western Cape provinces)

Here, we present the performances of the two classes of models that include both ILI and GT data. The first class is the regression models (ILI-GT-EN, ILI-GT-FNN, ILI-GT-MLR, and ILI-GT-SVM) while the second class is the statistical/deep learning time series techniques (ILI-GT-LSTM and ILI-GT-SARIMAX). The RMSE and MAE of the first class of models for Gauteng and Western Cape provinces can be visualized in Figures 10 and 11 respectively. Similarly, Figures 12 and 13 show the RMSE and MAE bar plots for the second class of models for both provinces.

RMSE and MAE

Gauteng: As can be seen in Figure 10, the performance of the regression models is comparable in terms of RMSE, with ILI-GT-SVM having the lowest value of 6.94 and ILI-GT-MLR having the highest value of 7.23 when ILI data of the prior week was incorporated. The values were slightly higher when ILI data of the previous two weeks were used. For the time series models (Figure 12), the ILI-GT-SARIMAX nowcast model had RMSE value of 10.21, comparable to 10.32 obtained from the GT-SVM nowcast model. The ILI-GT-LSTM model had RMSE value of 7.04. From Figure 11, the smallest MAE value of 4.82 was obtained from the ILI-GT-FNN model, while the highest value of 5.45 was obtained from ILI-GT-EN model when ILI data from the prior week was added as part of the features. The other two models in the regression models class (ILI-GT-MLR and ILI-GT-SVM) had similar MAE values of 4.91 and 4.95 respectively. From the second class (Figure 13), the ILI-GT-LSTM model had MAE value of 5.17 while ILI-GT-SARIMAX had a value of 7.17 when the ILI data was extended by GT data of the same week.

Western cape: Similarly for Western Cape province, the regression models performed comparably in terms of RMSE. The ILI-GT-FNN had the lowest RMSE value of 4.65 and ILI-GT-EN had the highest value of 4.97 when ILI data of the past one

week was added. The time series counterparts had values ranging from 4.96 (ILI-GT-LSTM) to 6.66 (ILI-GT-SARIMAX) when GT data of the same week was added as an extension to the ILI data.

For the MAE values, there were similar performances, with the ILI-GT-FNN having the smallest value of 3.34 (when the previous one-week ILI data was used), while the ILI-GT-SARIMAX had the highest value of 4.20 (Figure 13).

PCC. For Gauteng, the ILI-GT-LSTM deep learning method had the biggest PCC value of 0.9064, followed closely by ILI-GT-SVM (0.9063). Similarly for Western Cape, ILI-GT-FN had the biggest PCC value of 0.8746. The lowest PCC values are from the ILI-GT-SARIMAX time-series models.

PWD and PMD

Gauteng: All the regression models (the first class of models) estimated the first peak to be two weeks in advance of the real peak week, except the ILI-GT-FNN model that predicted the peak accurately when the ILI data of the previous week were added as a feature. The peak week estimation of the ILI-GT-LSTM model was one week late, while all the ILI-GT-SARIMAX models were less accurate at predicting the peak weeks, with all of them predicting the first peak week as five weeks earlier. The ILI-GT-MLR models gave the smallest peak magnitude difference for Gauteng.

Western cape: When we added the ILI data of the previous week, the peak estimation of all the regression models was one week late. However, they were all accurate in the estimation of the second peak week. The ILI-GT-LSTM predicted both peaks as one week later than the true peak weeks. The ILI-GT-MLR/ILI-GT-SVM models gave the lowest PMD values for the Western Cape province.

Figures 14 and 16 show the predicted versus true ILI incidence rates for Gauteng while Figures 15 and 17 show the same for Western Cape province. These figures reveal the performance of the models in terms of peak week difference and peak magnitude difference.

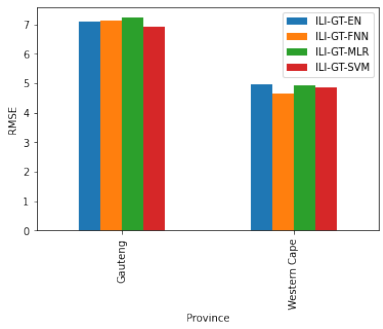


Fig. 10. RMSE of the regression models (ILI-GT-EN, ILI-GT-FNN, ILI-GT-MLR, and ILI-GT-SVM) for Gauteng and Western Cape provinces

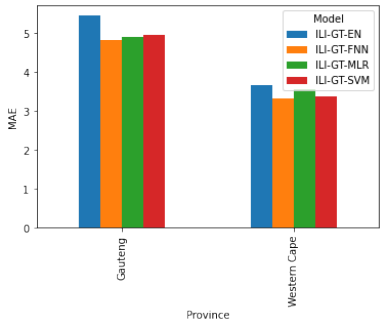


Fig. 11. MAE of the regression models (ILI-GT-EN, ILI-GT-FNN, ILI-GT-MLR, and ILI-GT-SVM) for Gauteng and Western Cape provinces

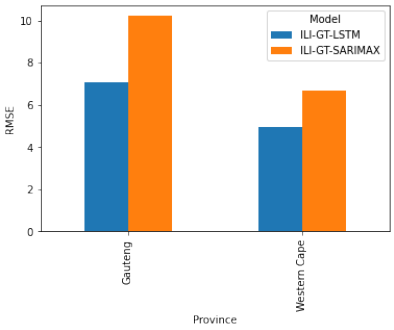


Fig. 12. RMSE of the time series (ILI-GT-LSTM and ILI-GT-SARIMAX) models for Gauteng and Western Cape provinces

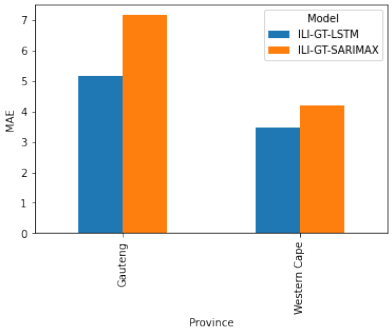


Fig. 13. MAE of the time series (ILI-GT-LSTM and ILI-GT-SARIMAX) models for Gauteng and Western Cape provinces

4 Discussion

In this paper, we first studied the correlation between 21 influenza-related terms and the ILI records in each province. This step is crucial as it shows which province(s) can employ the Google search volume of queries as a valid substitute for influenza forecasting.

The correlation analysis was organized per epidemiological year [54] in order to reveal any correlation trends over the years. The heatmaps (Figure 1) reveal that the GT data of the 21 ILI-related terms are only significantly correlated with provincial ILI records in Gauteng and Western Cape provinces. This may be explained by the fact that Internet access are highest in these two provinces as confirmed by the latest general household survey publications from Statistics South Africa (stats SA) [67], [68]. Another report from stats SA [69] shows that Gauteng and Western Cape are the top two provinces in terms of gross domestic product (GDP) per capita. The same report confirms Gauteng as South Africa's economic powerhouse, contributing 34% to the national economy in 2017. This suggests a relationship between a province's socio-economic profile and the use of digital platforms for disease surveillance in such a province.

The presence of many light-colored cells in the heatmaps for the other seven provinces show sparsity of Google searches for ILI-related terms in those provinces. The high number of missing values in the ILI records of these provinces (which were excluded in the analysis) may also have contributed to the low correlation values. This suggests the need for quality ILI data reporting by the public health authorities in all provinces and not just in the developed ones.

The performance of the GT-data-only models (over different metrics) for Gauteng and Western Cape as shown in Figures 2, 3, 4, and 5 show that the free and real-time Google search data can be used alone as a proxy to estimate ILI rates in those provinces with precision close to that of the best ILI-data only model. The lowest RMSE from GT-SVM is even slightly better than the ILI-only SARIMA model. Similar performance was reported in the national-scale study of [10].

Figures 6, 7, 8, and 9 show the behavior of models that used only historical ILI data to forecast future ILI incidence rates. These models were developed for all nine provinces, and we compare the performance of the statistical time series ILI-SARIMA model to the deep learning ILI-LSTM counterpart. One major issue with the provincial ILI data in those seven provinces excluded from the GT data models is the large number of missing values. Eastern Cape ILI test data also contains outlier We estimated the missing values but did not remove the outlier(s) in the test data, in order to reflect the scenario in real life in which future ILI rates cannot be predetermined. Figures 8 and 9 show that ILI-LSTM performed significantly better than the ILI-SARIMA models, even with the poor quality of the ILI data. For instance, ILI-LSTM was able to predict the future trend of ILI rates in Eastern Cape in the presence of the outlier. This demonstrates the strengths of advanced AI techniques for accurately forecasting ILI rates in the provinces where the complementary GT data is sparse.

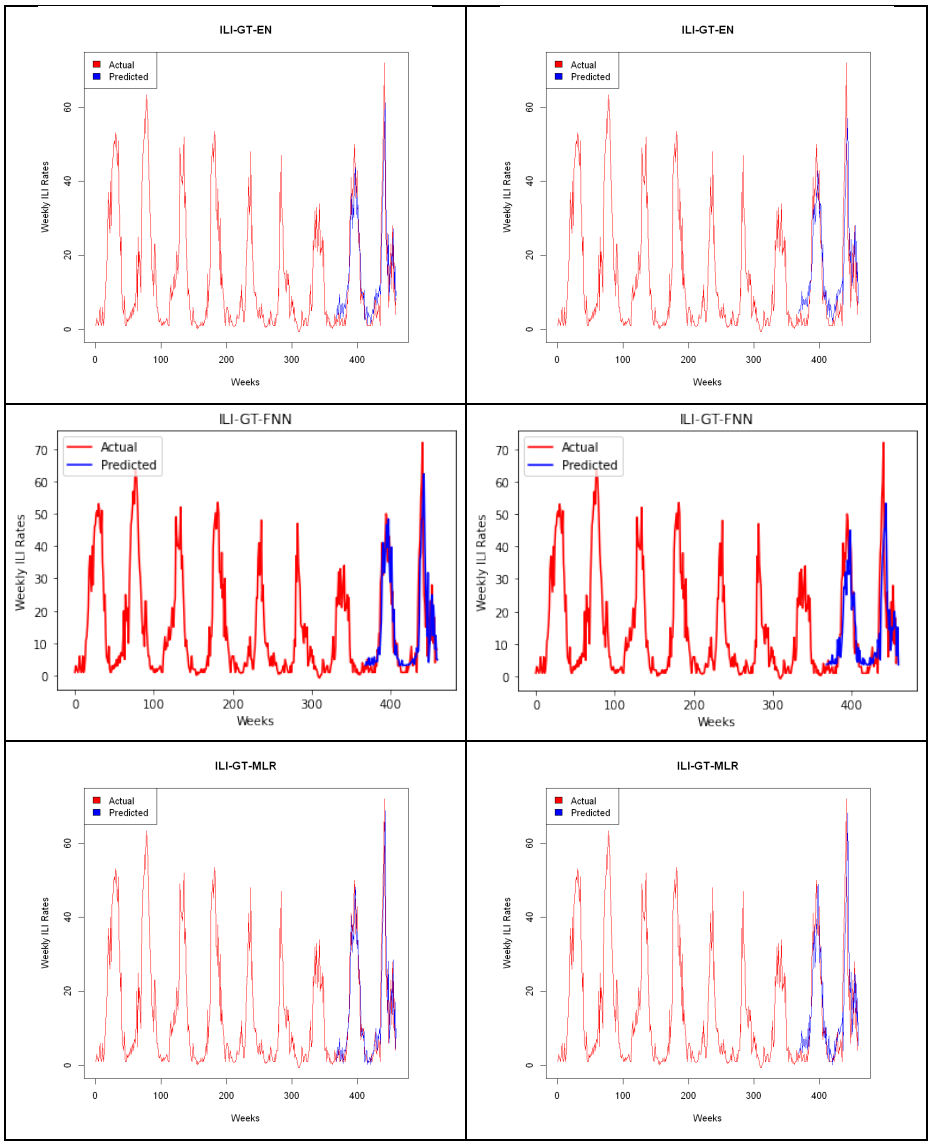
In addition, visualizations of the performance of the ILI-GT-data models for Gauteng and Western Cape provinces in Figures 10 – 17 show that integrating ILI with GT data yields better model performances, confirming previous findings in [10].

We also observe that the machine learning SVM-based models performed comparably to the deep learning FNN-based models. The SVM-based models performed slightly better than the FNN models in some cases, while the reverse is the case in some other instances. This validates the effectiveness of the SVM technique for the purpose. SVM also has the advantage of easier parameter tuning.

5 Conclusion

This study assessed the use of Google Trends data for predicting influenza rates at the provincial level in South Africa. First, we determined the relationship between digital Google search data and the true ILI records in each province by performing correlation analysis. The correlation study shows which province can employ Google search volumes as substitute for influenza surveillance. The outcomes of the analysis show that only the two most developed provinces (Gauteng and Western Cape) had significant correlation, suggesting an association between provincial socio-economic development and the use of digital platforms for disease surveillance. We therefore exclude GT data from the ILI forecasting models of the other seven provinces. For Gauteng and Western Cape, GT data was used as a standalone predictor as well as an enhancing predictor to predict ILI rates. The results for these two provinces show that Google search volume can be employed to successfully tackle the problems of delay and cost associated with the traditional surveillance systems. The potential of online search data for ILI surveillance is expected to increase further as Internet penetration continues to grow in each province. We also show the benefits of advanced AI time series methods (LSTM) in forecasting ILI accurately (even with low-quality data) as compared to traditional ARIMA methods. This is useful in the seven provinces where Google search data is sparse.

Overall, the study established the need for finer scale ILI forecasting which will inform targeted planning for disease surveillance and interventions. Although recent studies show the "predictive utility of Google search data for ILI forecasting" at the national level, this study reveals that Google search data can only be used effectively in the two most developed provinces of South Africa. The study also points to the need for quality reporting of ILI incidence in each province as this significantly impacts forecasting models developed for each province. Provinces with sparse digital surveillance data can harness advanced AI techniques to obtain accurate future ILI rates. All of these allow for better pandemic preparedness and aims to achieve the goal of sustainable surveillance set by the South African Department of Health.


ILI-GT-EN
Actual
Predicted
Weekly ILI Rates
Weeks
ILI-GT-EN
Actual
Predicted
Weekly ILI Rates
Weeks
ILI-GT-FNN
Actual
Predicted
Weekly ILI Rates
Weeks
ILI-GT-FNN
Actual
Predicted
Weekly ILI Rates
Weeks
ILI-GT-MLR
Actual
Predicted
Weekly ILI Rates
Weeks
ILI-GT-MLR
Actual
Predicted
Weekly ILI Rates
Weeks

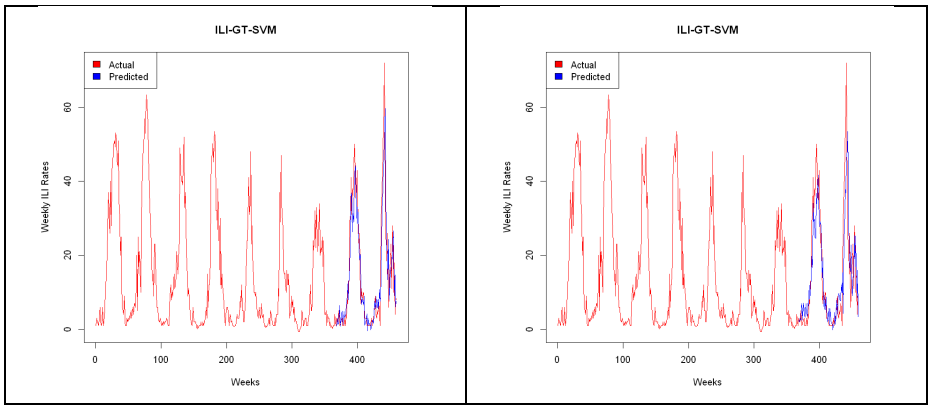
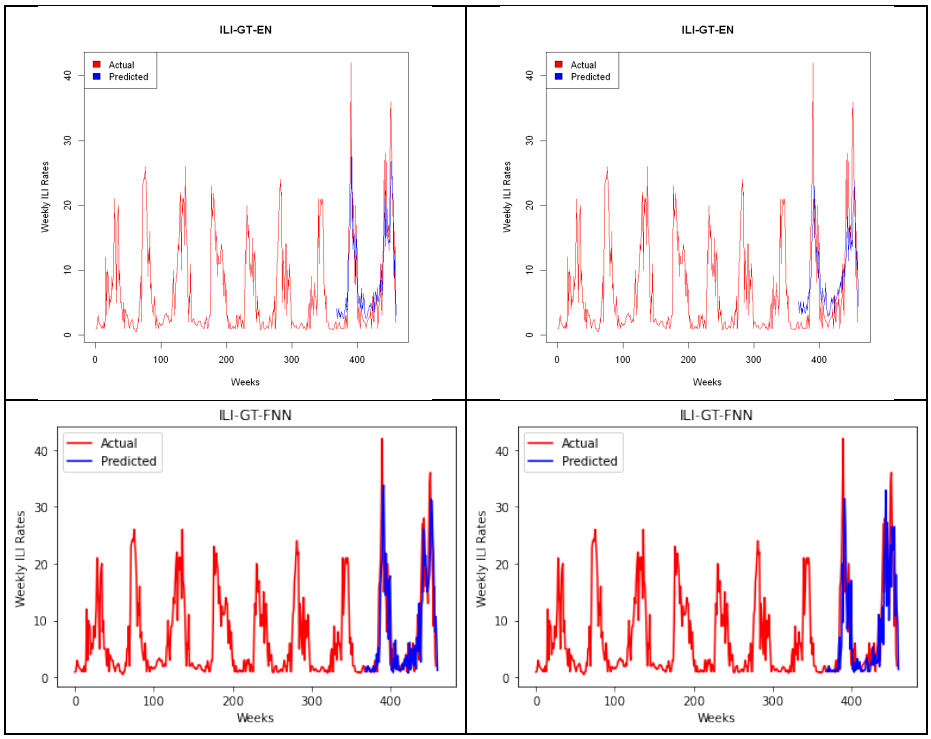


Fig. 14. True versus estimated weekly rates of ILI incidence produced by the regression models (ILI-GT-EN, ILI-GT-FNN, ILI-GT-MLR, and ILI-GT-SVM) for Gauteng. First column: ILI forecasts for current week were produced from Google search volume of same week and ILI data of previous week. Second column: ILI forecasts for current week were produced from Google search volume of same week and ILI data of the previous two weeks



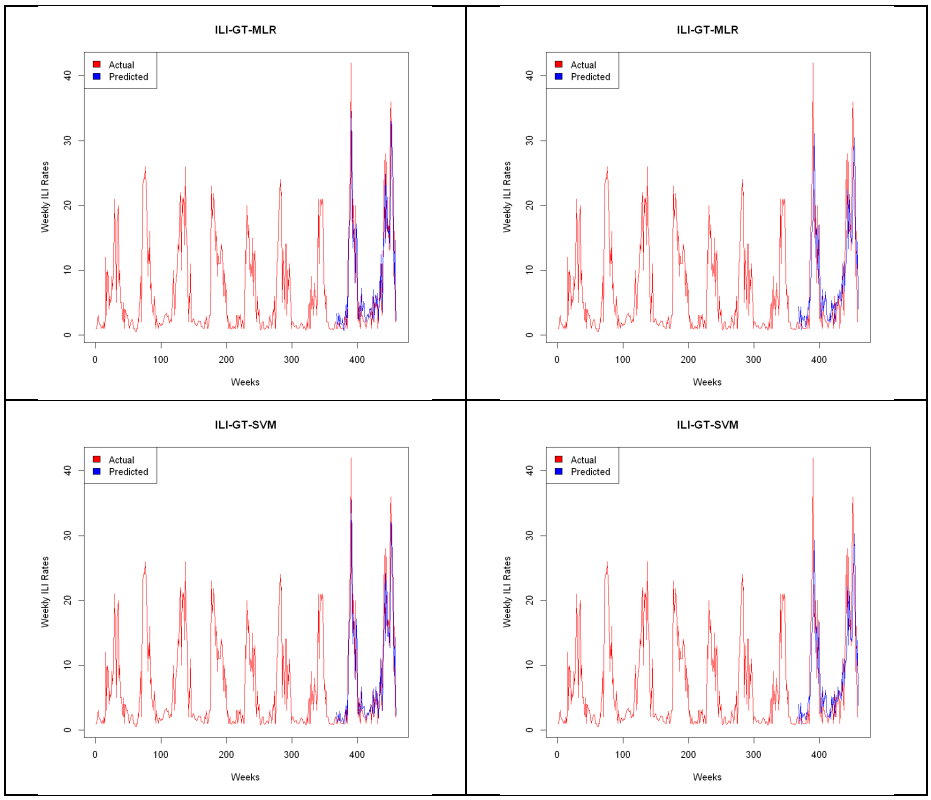


Fig. 15. True versus forecasted weekly rates of ILI incidence produced by the regression models (ILI-GT-EN, ILI-GT-FNN, ILI-GT-MLR, and ILI-GT-SVM) for Western Cape. First column: ILI forecasts for current week were produced from Google search volume of same week and ILI data of the previous week. Second column: ILI forecasts for current week were produced from Google search volume of same week and ILI data of the previous two weeks

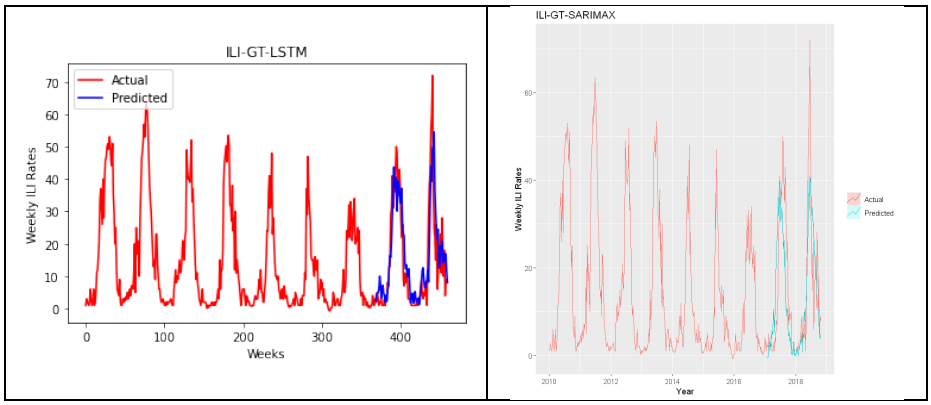


Fig. 16. True versus estimated weekly rates of ILI incidence by the time series models (ILI-GT-LSTM and ILI-GT-SARIMAX) for Gauteng (adding Google search volume of the same week)

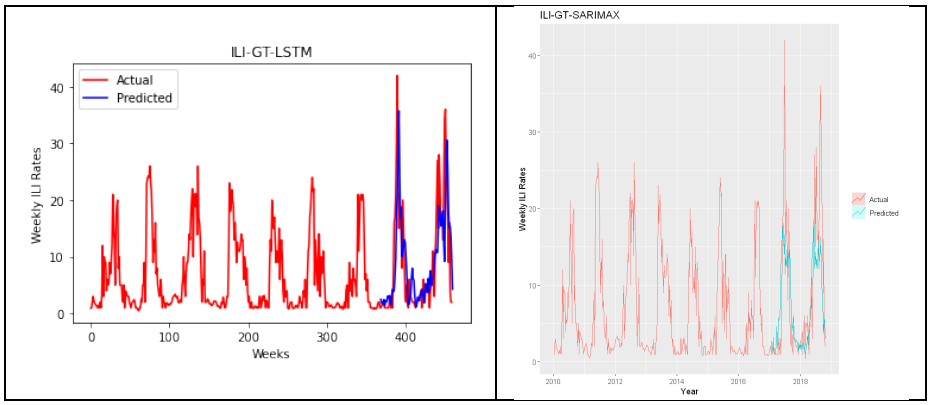


Fig. 17. True versus forecasted weekly rates of ILI incidence by the time series models (ILI-GT-LSTM and ILI-GT-SARIMAX) for Western Cape (adding Google search volume of the same week)

6 Acknowledgment

We are grateful to Prof. Cheryl Cohen and Jo Mcanerney of the NICD in South Africa, for supplying the ILI data.

7 References

[1] World Health Organization (WHO), "Influenza (Seasonal)," *Bulletin of the World Health Organization*, 2014. 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, "Influenza NICD Recommendations for the diagnosis, prevention, management and public health response," 2017. [Online]. Available: http://www.nicd.ac.za/wp-content/uploads/2017/03/Influenza-guidelines-final_24_05_2017.pdf. Accessed: Mar. 13, 2018.

[3] Department of Health, "National Influenza Policy and Strategic Plan: 2017 to 2021," [Online]. Available: http://www.health.gov.za/index.php/component/phocadownload/category/339. Accessed: Mar. 13, 2018.

[4] 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, p. eabb1237, Jun. 2021. https://doi.org/10.1126/sciadv.abb1237

[5] M. Salathé, "Digital epidemiology: what is it, and where is it going?," *Life Sci. Soc. Policy*, vol. 14, no. 1, 2018. https://doi.org/10.1126/sciadv.abb1237

[6] G. Eysenbach, "Infodemiology: tracking flu-related searches on the web for syndromic surveillance.," in *Proc. AMIA Annual Symposium*, 2006, pp. 244–8, doi: PMC1839505.

[7] P. M. Polgreen, Y. Chen, D. M. Pennock, F. D. Nelson, and R. A. Weinstein, "Using Internet Searches for Influenza Surveillance," *Clin. Infect. Dis.*, vol. 47, no. 11, pp. 1443–1448, Dec. 2008. https://doi.org/10.1086/593098

[8] R. Moss, A. Zarebski, P. Dawson, and J. M. McCaw, "Forecasting influenza outbreak dynamics in Melbourne from Internet search query surveillance data," *Influenza Other Respi. Viruses*, vol. 10, no. 4, 2016. https://doi.org/10.1111/irv.12376

[9] L. Clemente, F. Lu, and M. Santillana, "Improved real-time influenza surveillance: Using internet search data in eight Latin American countries," *J. Med. Internet Res.*, vol. 21, no. 4, 2019. https://doi.org/10.2196/12214

[10] S. O. Olukanmi, F. V. Nelwamondo, and N. I. Nwulu, "Utilizing Google Search Data with Deep Learning, Machine Learning and Time Series Modeling to Forecast Influenza-Like Illnesses in South Africa," *IEEE Access*, vol. 9, pp. 126822–126836, 2021. https://doi.org/10.1109/ACCESS.2021.3110972

[11] E. De Quincey and P. Kostkova, "Early Warning and Outbreak Detection Using Social Networking Websites: The Potential of Twitter," in *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, 2009, vol. 27, pp. 21–24. https://doi.org/10.1007/978-3-642-11745-9_4

[12] H. Achrekar, A. Gandhe, R. Lazarus, S.-H. Yu, and B. Liu, "TWITTER IMPROVES SEASONAL INFLUENZA PREDICTION," 2012. [Online]. Available: http://www.cs.uml.edu/~bliu/pub/healthinf_2012.pdf. Accessed: Jun. 21, 2017.

[13] A. Lamb, M. J. Paul, and M. Dredze, "Separating Fact from Fear : Tracking Flu Infections on Twitter," *Proc. NAACL-HLT 2013*, no. June, pp. 789–795, 2013.

[14] 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. https://doi.org/10.1038/s41598-019-54388-4

[15] 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. https://doi.org/10.1016/S1532-0464(03)00013-3

[16] M. Abla and M. Blench, "Global Public Health Intelligence Network (GPHIN)," in *7th Conference of the Association for Machine Translation in the Americas*, 2006, pp. 8–12. [Online]. Available: https://pdfs.semanticscholar.org/7d88/e623aa6ca78510e0093e17e2e00db39bdad5.pdf. Accessed: Mar. 09, 2018.

[17] 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.

[18] D. J. Mciver, J. S. Brownstein, and M. Salathé, "Wikipedia Usage Estimates Prevalence of Influenza-Like Illness in the United States in Near Real-Time," *PLoS Comput Biol*, vol. 10, no. 4, 2014. https://doi.org/10.1371/journal.pcbi.1003581

[19] B. Bardak and M. Tan, "Prediction of influenza outbreaks by integrating Wikipedia article access logs and Google flu trend data," 2015. https://doi.org/10.1109/BIBE.2015.7367640

[20] K. S. Hickmann, G. Fairchild, R. Priedhorsky, N. Generous, J. M. Hyman, A. Deshpande, S. Y. Del Valle, "Forecasting the 2013–2014 Influenza Season Using Wikipedia," *PLoS Comput. Biol.*, vol. 11, no. 5, 2015. https://doi.org/10.1371/journal.pcbi.1004239

[21] A. Hulth, G. Rydevik, and A. Linde, "Web Queries as a Source for Syndromic Surveillance," *PLoS One*, vol. 4, no. 2, 2009. https://doi.org/10.1371/journal.pone.0004378

[22] 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 July, 2020. https://doi.org/10.1371/journal.pone.0233855

[23] 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*, 2013. https://doi.org/10.1371/journal.pone.0081422

[24] 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, p. e55205, Jan. 2013. https://doi.org/10.1371/journal.pone.0055205

[25] 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. https://doi.org/10.1186/s40537-018-0140-9

[26] X. Zhou, J. Ye, and Y. Feng, "Tuberculosis surveillance by analyzing google trends," *IEEE Trans. Biomed. Eng.*, 2011. https://doi.org/10.1109/TBME.2011.2132132

[27] 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.*, 2017. https://doi.org/10.1038/s41598-017-05091-9

[28] 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*, 2017. https://doi.org/10.1371/journal.pone.0165085

[29] 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," *Infect. Dis. Poverty*, vol. 4, no. 1, 2015. https://doi.org/10.1186/s40249-015-0090-9

[30] A. Mavragani and G. Ochoa, "Forecasting AIDS prevalence in the United States using online search traffic data," *J. Big Data*, vol. 5, no. 1, Dec. 2018. https://doi.org/10.1186/s40537-018-0126-7

[31] 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, 2015. https://doi.org/10.1136/bmjopen-2014-006678

[32] A. Seifter, A. Schwarzwalder, K. Geis, and J. Aucott, "The utility of 'Google Trends' for epidemiological research: Lyme disease as an example," *Geospat. Health*, 2010. https://doi.org/10.4081/gh.2010.195

[33] 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.*, 2015. https://doi.org/10.2196/jmir.4516

[34] M. Effenberger, A. Kronbichler, J. Il Shin, G. Mayer, H. Tilg, and P. Perco, "Association of the COVID-19 pandemic with Internet Search Volumes: A Google TrendsTM Analysis," *International Journal of Infectious Diseases*, vol. 95. Elsevier B.V., pp. 192–197, Jun. 01, 2020. https://doi.org/10.1016/j.ijid.2020.04.033

[35] S. Mohammad, S. M. Ayyoubzadeh, H. Zahedi, M. Ahmadi, and S. R. N. Kalhori, "Predicting COVID-19 Incidence Through Analysis of Google Trends Data in Iran: Data Mining and Deep Learning Pilot Study," *JMIR Public Heal. Surveill 2020;6(2)e18828 https//publichealth.jmir.org/2020/2/e18828*, vol. 6, no. 2, p. e18828, Apr. 2020. https://doi.org/10.2196/18828

[36] 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," *Euro Surveill.*, vol. 25, no. 10, 2020. https://doi.org/10.2807/1560-7917.ES.2020.25.10.2000199

[37] 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, 2015. https://doi.org/10.1155/2015/436495

[38] 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. Infect. Public Health*, vol. 11, no. 5, pp. 707–712, Sep. 2018. https://doi.org/10.1016/j.jiph.2018.04.009

[39] T. Xiao-qing, Z. Zhan-lin, G. Zheng, Y. Mahan, H. Bing-xue, T. Tian, A. Ainiwaer, C. Zhen, G. Hailili, F. Xu-cheng, and D. Jiang-hong, "Forecasting influenza like illness in Urumqi based on ARIMAX model," *CHINESE J. Dis. Control Prev. 2018, Vol. 22, Issue 6, Pages 590-593*, vol. 22, no. 6, pp. 590–593. https://doi.org/10.16462/j.cnki.zhjbkz.2018.06.012

[40] W. Anggraeni and L. Aristiani, "Using Google Trend data in forecasting number of dengue fever cases with ARIMAX method case study: Surabaya, Indonesia," in *Proceedings of 2016 International Conference on Information and Communication Technology and Systems, ICTS 2016*, Apr. 2017, pp. 114–118. https://doi.org/10.1109/ICTS.2016.7910283

[41] 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. https://doi.org/10.1038/s41598-021-85987-9

[42] A. Mavragani and G. Ochoa, "Infoveillance of infectious diseases in USA: STDs, tuberculosis, and hepatitis," *J. Big Data*, vol. 5, no. 1, Dec. 2018. https://doi.org/10.1186/s40537-018-0140-9

[43] 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 Heal. Surveill.*, vol. 4, no. 4, p. e11361, Oct. 2018. https://doi.org/10.2196/11361

[44] 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, 2015. https://doi.org/10.1371/journal.pcbi.1004513

[45] 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 and Fractals*, vol. 142, Jan. 2021. https://doi.org/10.1016/j.chaos.2020.110336

[46] 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*, 2017. https://doi.org/10.1371/journal.pone.0188941

[47] 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, 2020. https://doi.org/10.1007/s11227-020-03182-5

[48] 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, Aug. 2018. https://doi.org/10.3390/ijerph15081596

[49] E. Mogo, "Social Media As A Public Health Surveillance Tool: Evidence And Prospects," Baltimore MD. [Online]. Available: enterprise.sickweather.com/downloads/SW-SocialMedia_WhitePaper.pdf.

[50] G. Cervellin, I. Comelli, and G. Lippi, "Is Google Trends a reliable tool for digital epidemiology? Insights from different clinical settings," *J. Epidemiol. Glob. Health*, vol. 7, no. 3, 2017. https://doi.org/10.1016/j.jegh.2017.06.001

[51] 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: IOS Press BV, 2012, pp. 100–104.

[52] Z. S. H. 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 Digital Medicine*, vol. 4, no. 1. Nature Research, pp. 1–13, Dec. 01, 2021. https://doi.org/10.1038/s41746-021-00407-6

[53] N. L. Bragazzi and N. Mahroum, "Google trends predicts present and future plague cases during the plague outbreak in Madagascar: Infodemiological study," *J. Med. Internet Res.*, vol. 21, no. 3, 2019. https://doi.org/10.2196/13142

[54] S. Olukanmi and F. Nelwamondo, "Digital influenza surveillance: The prospects of Google trends data for South Africa," in *Proc. icABCD 2020*, Aug. 2020. https://doi.org/10.1109/icABCD49160.2020.9183882

[55] D. Osthus and K. R. Moran, "Multiscale influenza forecasting," *Nat. Commun. 2021 121*, vol. 12, no. 1, pp. 1–11, May 2021. https://doi.org/10.1038/s41467-021-23234-5

[56] "Weekly Influenza and Respiratory Syncytial Virus Surveillance Report Week 19, 2019," 2019. [Online]. Available: http://cran.r-project.org/web/package=mem. Accessed: Jul. 30, 2019.

[57] "Google Trends: Understanding the data.". [Online]. Available: https://storage.googleapis.com/gweb-news-initiative-training.appspot.com/upload/GO802_NewsInitiativeLessons_Fundamentals-L04-GoogleTrends_1saYVCP.pdf. Accessed: Aug. 01, 2019

[58] R. J. Hyndman and Y. Khandakar, "Journal of Statistical Software Automatic Time Series Forecasting: the forecast Package for R," vol. 27, 2008. https://doi.org/10.18637/jss.v027.i03

[59] G. E. P. Box and G. M. Jenkins, *Time Series Analysis: Forecasting & Control*. Holden-Day, 1970.

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

[61] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *J. Stat. Softw.*, vol. 33, no. 1, pp. 1–22, 2010. https://doi.org/10.18637/jss.v033.i01
[62] 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. https://doi.org/10.18637/jss.v039.i05
[63] H. Drucker, C. J. Burges, L. Kaufman, A. J. Smola, and V. Vapnik, "Support vector regression machines," *Neural Inf. Process. Syst.*, vol. 9, pp. 155–161, 1997, Accessed: Apr. 07, 2021. [Online]. Available: http://ci.nii.ac.jp/naid/10018343800/en/
[64] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, F. Leisch, C. C. Chang, C. C. Lin, and M. D. Meyer, "Package 'e1071,'" *The R Journal*, 2019.
[65] D. Svozil, V. Kvasnička, and J. Pospíchal, "Introduction to multi-layer feed-forward neural networks," *Chemom. Intell. Lab. Syst.*, vol. 39, no. 1, pp. 43–62, Nov. 1997. https://doi.org/10.1016/S0169-7439(97)00061-0
[66] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997. https://doi.org/10.1162/neco.1997.9.8.1735
[67] "Statistical release (General Household Survey)," 2016. [Online]. Available: www.statssa.gov.za. Accessed: Nov. 16, 2021
[68] "Statistical Release (General Household Survey)," 2019. [Online]. Available: www.statssa.gov.za. Accessed: Nov. 16, 2021.
[69] "Four facts about our provincial economies | Statistics South Africa." http://www.statssa.gov.za/?p=12056&gclid=Cj0KCQiAys2MBhDOARIAFf1D1eojlyn1El-jWBNFAdDHViUj2FL5iFieg8WqCIuHP56iNgVseZjo5iUaAjQVEALw_wcB. Accessed Nov. 16, 2021.

8 Authors

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 completing her Ph.D. degree with the Institute for Intelligent Systems, department of Electrical and Electronic Engineering Science, University of Johannesburg, South Africa. Her current research interests include artificial intelligence and data science for social impact.

Fulufhelo V. Nelwamondo received the B.Sc. and Ph.D. degrees in electrical engineering (computational intelligence) from the University of Witwatersrand, 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 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).

Article submitted 2022-02-01. Resubmitted 2022-04-29. Final acceptance 2022-05-08. Final version published as submitted by the authors.