

https://dspace.org/

# Leveraging deeplearning and unconventional data for realtime surveillance forecasting and early warning of respiratory pathogens outbreak

**Z. Movahedi Nia; L. Seyyed-Kalantari; M. Goitom; Bruce Mellado; A. Ahmadi ...**

ELSEVIER SCIENCE BV
https://hdl.handle.net/10539/46008
WOS

Downloaded from a DSpace repository. https://dspace.org/

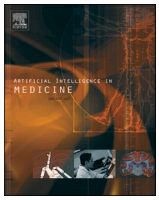
Artificial Intelligence In Medicine 161 (2025) 103076



Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed



Research paper



# Leveraging deep-learning and unconventional data for real-time surveillance, forecasting, and early warning of respiratory pathogens outbreak

Z. Movahedi Nia [a,b,c], L. Seyyed-Kalantari [a,d], M. Goitom [a,b,e], B. Mellado [a,b,f], A. Ahmadi [g], A. Asgary [a,b,g], J. Orbinski [a,b,h], J. Wu [a,c], J.D. Kong [a,b,i,j,k,*]

[a] Africa-Canada Artificial Intelligence and Data Innovation Consortium (ACADIC), Canada
[b] Global South Artificial Intelligence for Pandemic and Epidemic Preparedness and Response Network (AI4PEP), Canada
[c] Laboratory for Industrial and Applied Mathematics (LIAM), York University, Toronto, Canada
[d] Department of Electrical Engineering and Computer Science, York University, Toronto, Canada
[e] School of Social Work, York University, Toronto, ON M3J 1P3, Canada
[f] School of Physics, Institute for Collider Particle Physics, University of the Witwatersrand, Johannesburg, South Africa
[g] Advanced Disaster, Emergency and Rapid-response Simulation (ADERSIM), York University, Toronto, Ontario, Canada
[h] Dahdaleh Institute for Global Health Research, York University, Toronto, Canada
[i] Artificial Intelligence & Mathematical Modelling Lab (AIMM Lab), Dalla Lana School of Public Health, University of Toronto, 155 College St, Toronto, ON M5T 3M7, Canada
[j] Institute of Health Policy, Management and Evaluation (IHPME), University of Toronto, Canada
[k] Department of Mathematics, University of Toronto, Bahen Centre for Information Technology, 40 St. George Street, Toronto, ON M5S 2E4, Canada

ARTICLE INFO

*Keywords:*
Canada
Convolutional neural network (CNN)
COVID-19
Early warning system
Gated recurrent unit (GRU)
Graph neural network (GNN)
Southern African countries

ABSTRACT

*Background:* Controlling re-emerging outbreaks such as COVID-19 is a critical concern to global health. Disease forecasting solutions are extremely beneficial to public health emergency management. This work aims to design and deploy a framework for real-time surveillance, prediction, forecasting, and early warning of respiratory disease. To this end, we selected southern African countries and Canadian provinces, along with COVID-19 and influenza as our case studies.

*Methodology:* Six different datasets were collected for different provinces of Canada: number of influenza cases, number of COVID-19 cases, Google Trends, Reddit posts, satellite air quality data, and weather data. Moreover, five different data sources were collected for southern African countries whose COVID-19 number of cases were significantly correlated with each other: number of COVID-19 infections, Google Trends, Wiki Trends, Google News, and satellite air quality data. For each infectious disease, i.e. COVID-19 and Influenza for Canada and COVID-19 for southern African countries, data was processed, scaled, and fed into the deep learning model which included four layers, namely, a Convolutional Neural Network (CNN), a Graph Neural Network (GNN), a Gated Recurrent Unit (GRU), and a linear Neural Network (NN). Hyperparameters were optimized to provide an accurate 56-day-ahead prediction of the number of cases.

*Result:* The accuracy of our models in real-time surveillance, prediction, forecasting, and early warning of respiratory diseases are evaluated against state-of-the-art models, through Root Mean Square Error (RMSE), coefficient of determination (R2-score), and correlation coefficient. Our model improves R2-score, RMSE, and correlation by up to 55.98 %, 39.71 %, and 44.47 % for 56 days-ahead COVID-19 prediction in Ontario, 34.87 %, 25.52 %, 50.91 % for 8 weeks-ahead influenza prediction in Quebec, and 51.04 %, 32.04 %, and 28.74 % for 56 days-ahead COVID-19 prediction in South Africa, respectively.

* Corresponding author at: Artificial Intelligence & Mathematical Modelling Lab (AIMM Lab), Dalla Lana School of Public Health, University of Toronto, 155 College Street, Toronto, ON M5T 3M7, Canada.
*E-mail address:* jude.kong@utoronto.ca (J.D. Kong).

https://doi.org/10.1016/j.artmed.2025.103076
Received 9 August 2024; Received in revised form 24 January 2025; Accepted 30 January 2025
Available online 1 February 2025
0933-3657/© 2025 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

*Conclusion:* This work presents a framework that automatically collects data from unconventional sources, and builds an early warning system for COVID-19 and influenza outbreaks. The result is extremely helpful to policy-makers and health officials for preparedness and rapid response against future outbreaks.

## 1. Introduction

More than 13 million people lost their lives to infectious diseases in 2019, prior to the COVID-19 pandemic [1]. Containing emerging and re-emerging outbreaks is a critical concern to global health. Infectious diseases are one of the major causes of death in children under 5 years old [2]. Furthermore, epidemics could increase the risk of mental health disorders in both patients and healthcare workers [3]. Other than health and well-being, outbreaks affect the education system, employment, social status, economic growth, agriculture, and other elements of human development such as gender equality and human rights [4–6]. Latest studies show a mutual causation between poverty and the spread of communicable diseases [7]. Surveillance and early alert systems are extremely beneficial to controlling and reducing the burden of infectious diseases.

Early Warning Systems (EWS) for infectious diseases include planning, designing, and implementing methods for prediction or early detection of occurring and re-occurring outbreaks [20]. Despite considerable efforts in developing syndromic surveillance and EWS for infectious diseases, an efficient and practical solution is not yet available. One of the first disease surveillance systems, ProMED-mail, launched in 1994, relies on official subscribers from more than 185 countries worldwide to report suspicious symptoms [8,9]. Developed in 1997, GPHIN is a more automated surveillance system that instead of subscribers, depends on news articles retrieved through extensive search queries to detect disease outbreaks [10]. Successively, some researchers proposed to use health-related data such as emergency department visits [11], telephone triage calls [12], ambulance dispatches [13], hospital records [14], pharmacy records [15,16], and school absenteeism [17] for syndromic surveillance.

Subsequent works suggested that the number of Internet searches or article views is a more convenient source of information for building health surveillance systems, as they are easier to access, more cost-efficient, provide both pre- and post-diagnostic statistics, and most often correlate pretty well with the number of cases, or mortalities [18–21]. A great advantage of search data and article views is that they provide information days to weeks earlier than traditional health-related data [22]. Therefore, contemporary methods mostly focus on building EWS using the number of search queries, and page views [23,24].

A great volume of literature on syndromic surveillance systems concentrates on lower respiratory infections, since they are the fourth leading cause of mortality, worldwide [25–29]. COVID-19 alone with more than 700 million confirmed cases has killed approximately 7 million people, as of December 19, 2023 [30]. The COVID-19 pandemic has highlighted the core necessity of EWS for rapid response to emerging and re-emerging diseases. On December 31st, 2019, the World Health Organization (WHO) Country Office in China was informed of pneumonia-like symptoms from an unknown etiology occurring in Wuhan, Hubei province of China that did not respond well to conventional flu treatments. The 2019 novel coronavirus quickly spread among other cities and provinces of China and eventually, across countries through travelers. Thailand, Japan, and South Korea were the first countries to report cases of COVID-19. WHO declared the outbreak a Public Health Emergency of International Concern (PHEIC), on January 30th, 2020, and a pandemic on March 11, 2020, when more than 118,000 cases had been reported in 114 different countries [31]. Due to the drastic impacts of the COVID-19 pandemic on healthcare services and the global economy [32,33], many researchers urged to build an EWS for COVID-19 waves [34–37].

Although the mentioned works provide rich information on syndromic surveillance systems, they do not propose a forecasting method for imminent outbreaks. Disease prediction enables preparedness, planning, rapid response, and recovery. It is a fundamental tool for the public health emergency management. One of the first forecasting methods for COVID-19 waves was proposed by F. Stevenson, et al. [38], where a multi-variate time series prediction method was implemented using Long-Short Term Memory (LSTM). LSTM is a Recurrent Neural Network (RNN) capable of predicting time series using sequences of multiple variables. LSTM solves the vanishing gradients problem of RNN by enabling the model to learn long data relationships and sequence patterns [39]. To forecast the COVID-19 waves in South Africa, in [38] four different variables are fed into LSTM, namely, Google mobility data, Facebook mobility data, stringency index, and previous daily COVID-19 number of cases. The LSTM model is built to predict the number of cases up to 14 days ahead. Google and Facebook released mobility data during the COVID-19 pandemic to help policymakers and health officials combat COVID-19, however, the datasets are no longer updated and could not be used for future EWS [40,41].

Another source of data that reportedly correlates well with respiratory infection statistics is weather and air quality data [42]. D. Aragão, et al. [43] gathered air quality data from multiple sources and built a multi-variate 1 week-ahead prediction model using stacked LSTM for COVID-19 mortalities in different states and major cities of Brazil. In [44] different parishes of Louisiana state of the USA were clustered using a weighted k-means clustering algorithm based on Xgboost, according to their population density, race, ethnicity, median age, and average daily temperature and humidity. For each cluster, an LSTM model was trained to forecast COVID-19 number of cases, mortalities, and recoveries.

Introduced in 2014, Gated Recurrent Unit (GRU) is an RNN model that also mitigates the vanishing gradients problem while using fewer parameters compared to LSTM, making it faster to train [45]. Some papers have implemented an EWS using both LSTM and GRU to compare their performance. LSTM and GRU are commonly combined with Convolutional Neural Networks (CNN) for better accuracy [46–48]. Batool and Tian [49] used weather parameters such as temperature and humidity to build a forecasting model for COVID-19 cases and mortalities in Pakistan. Their results show that LSTM performs better than GRU. Khennou and Akhloufi [50] used weather parameters, i.e. temperature, humidity, and precipitation, and the number of COVID-19 tests to build a model using LSTM, GRU, and their combination for predicting COVID-19 cases in different provinces of Canada. The results indicate that the models perform differently for different provinces, and no model is superior to the other, in general.

A few papers have considered the location properties to build an EWS. Considering location properties significantly improves the accuracy of the EWS model [59]. Hartono [51] proposed to cluster countries based on their COVID-19 transmission dynamics and using topological autoencoders. Then, for each cluster, a reference country is selected, and an LSTM model is trained based on that to forecast the number of COVID-19 cases. They performed their model on 250 different countries where the transmission dynamics of the countries are determined by the number of COVID-19 cases over time. SI-LSTM and SI-CNN-LSTM [52] use 5 data sources, namely, Google Trends, telecoms mobility, Google mobility, website test requests, and National Health Service (NHS) 119 calls to build an EWS model for COVID-19 cases and fatalities. The 5 data sources of different Local Authority Districts (LAD) of the UK go through layers of LSTM for SI-LSTM and CNN-LSTM for SI-CNN-LSTM and are eventually combined using a dense layer. The results indicate that the SI-LSTM model mostly outperforms the SI-CNN-LSTM model.

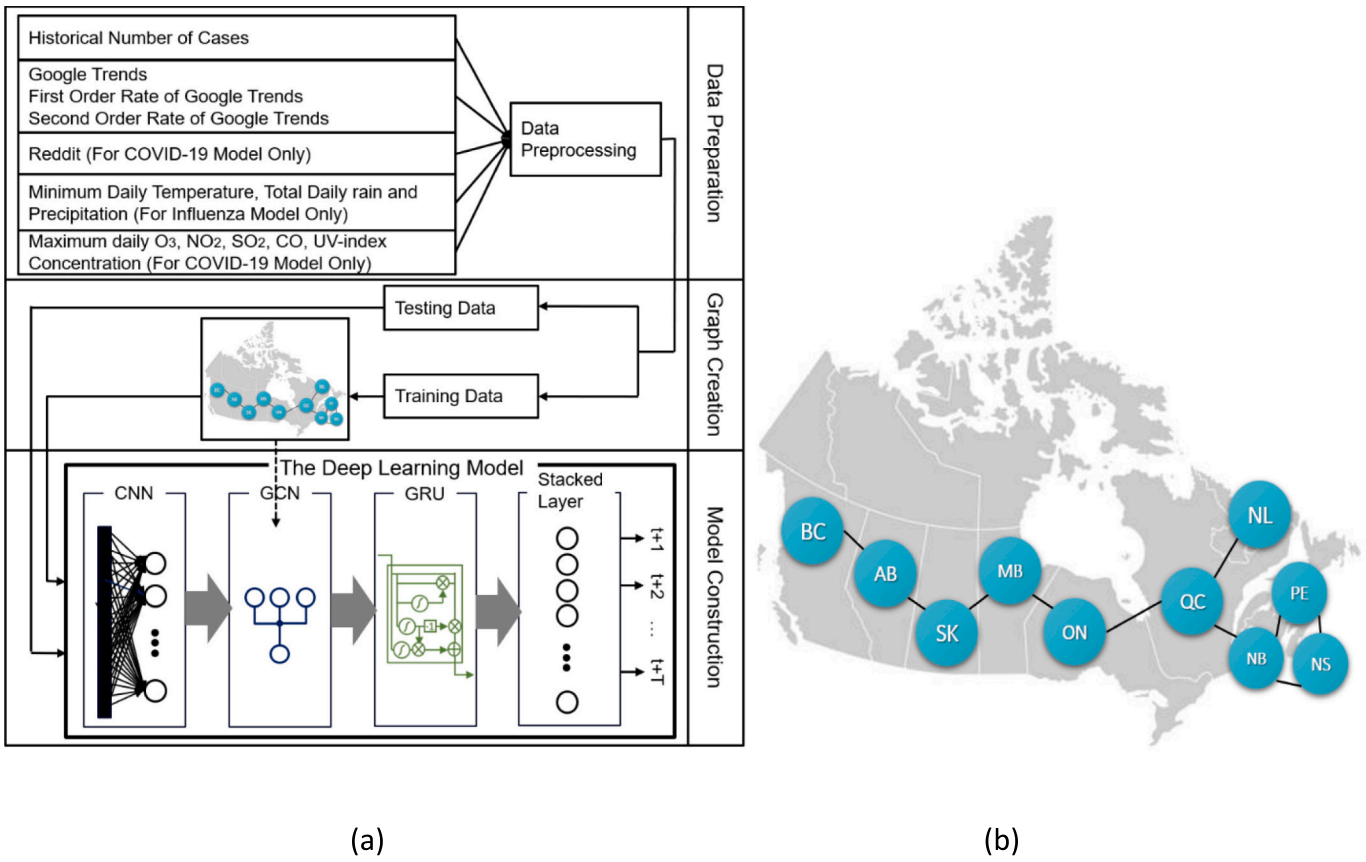



**Fig. 1.** (a) The overall model for Canada, and (b) the graph used for building the GNN layer for Canada.

The methods proposed in [51,52] require training several LSTM models and have high computational and space complexity. A more efficient and location-based method for building an EWS is using Graph Neural Networks (GNN). GNN is a branch of artificial neural networks used to model and process data that can be represented as graphs. The most commonly used form of GNN is Graph Convolutional Networks (GCN) which could be recognized as a generalized CNN for modeling graph-structured data [53]. Authors in [54] derived 3 indicators from Facebook mobility data, namely, staying put, co-location network, and social connectedness network. A fully connected graph was built based on different districts of Germany where the social connectedness indicator was set as weights for the edges. The other two indicators, staying put and co-location network, were fed into the GNN, and then a Generalized Additive Model (GAM) to perform 3 weeks-ahead prediction.

In [55] a graph is created from the provinces/regions of Italy, and the weights are set based on the mobility flows between the locations. The number of swab tests used in each location, mobility in each location, and demographic information such as the population density of each region are used as features. The data goes through a GCN layer and then an LSTM layer to predict the number of COVID-19cases in each location. In [56] a fully connected graph was designed for different departments of France, where the weights of the edges were set based on mobility between departments. Four features were considered for nodes, number of COVID-19 cases in previous days, population of the department, vaccination rate of the department, and mobility in the department. The features of the nodes were fed to a GNN layer and a Message Passing Neural Network (MPNN) was implemented that performed 14 day-ahead prediction for the number of COVID-19 cases and hospitalizations.

Authors in [57] build a fully connected graph from different states of the USA, where the edges are the correlation between the number of COVID-19 cases in the two states. They used GCN to perform a 1-day ahead prediction for forecasting the COVID-19 effective reproduction number of different states. In [58] a fully connected graph is built for different regions of four different European countries, namely, Italy, Spain, France, and England, where the edges are set based on the Facebook mobility between the regions. The number of COVID-19 cases in different regions is used to first train a GNN and then fine-tune a Model-Agnostic Meta-Learning (MAML) model to provide a 14-day-ahead prediction model for the number of COVID-19 cases.

In [60] a fully connected graph is made from different counties in the USA, where the weights of the edges are set using the mobility between the counties. Then, the next-day number of COVID-19 cases and deaths in different counties is predicted using a GNN. Authors in [61] build a graph for 193 counties of the USA, where neighboring nodes are connected and their weight is determined by the population size of the nodes. The nodes are explained by four features, number of active cases, hospitalizations, ICU stays, and total number of cases. The features go through a Graph Attention Network (GAT), and then a GRU layer to perform 10 days-ahead prediction for COVID-19 number of cases. Cola-GNN [62] proposes a model consisting of three layers for forecasting the influenza number of cases in different prefectures of Japan, states of the USA, and regions of the USA. The first layer is a simple RNN network with an attention layer on top of it. The second layer is a CNN, and the third layer is a GNN, where the weights of the graph are set using the cross-location attention matrix. The result is a 20-week-ahead influenza cases prediction.

This paper focuses on building and deploying technology for real-time surveillance, prediction, forecasting, and early warning of respiratory diseases in Canadian provinces and southern African countries. To improve the accuracy further using GNN, two graphs are formed, one on provinces of Canada and the other on southern African countries. To make the model practical for prognosticating future outbreaks, only

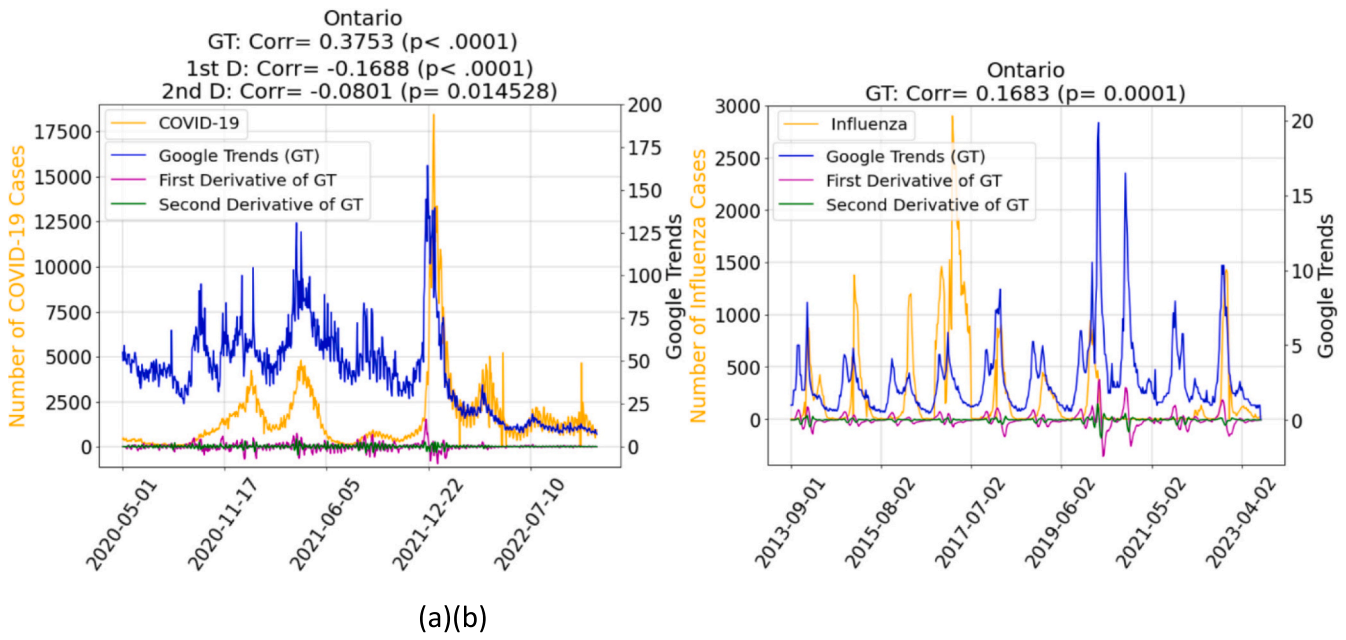



**Fig. 2.** Google Trends, and the first, and second derivatives of Google Trends, and the number of (a) COVID-19 and (b) influenza cases.

datasets that are freely available on websites or through RESTful APIs are used. We apply our model to forecast COVID-19 and influenza cases 56 days ahead; however, our model applies to any respiratory infectious disease. Our contribution to this domain is three-fold:

- To provide a framework for forecasting future respiratory infections, particularly COVID-19 in southern African countries which has not been studied previously, and COVID-19 and influenza in different provinces of Canada.
- To build a model that could automatically extract data and utilize it for COVID-19, influenza, and other respiratory infectious disease syndromic surveillance.
- To improve the accuracy of state-of-the-art EWS models.

All our datasets and codes are available from [87]. In the following, section 2 describes, the methodology, section 3 presents the results, and sections 4 and 5 provide the discussion and conclusion of our manuscript.

## 2. Methodology

### 2.1. Canada

For each of the COVID-19 and influenza infectious diseases, a different EWS model is built. Fig. 1(a) shows the overall design of an EWS model, which is composed of three parts. The first and second parts, data preparation and graph creation, process the data to format it for training the deep learning model. The graph used in the graph creation part was the same for both models, i.e. COVID-19 and influenza, and is shown in Fig. 1(b). The 3 territorial provinces, i.e. Northwest Territories, Yukon, and Nunavut were excluded because not enough surveillance data was available for them. The third part of the design (Fig. 1(a)), model construction includes training and testing the model using the prepared data to forecast the number of cases. Each part of the model is explained in the following.

**Table 1**
Features that are significantly correlated with the number of cases in different provinces.

| | ON | QC | BC | AB | MB | SK | NB | NS | NL | PE |
|---|---|---|---|---|---|---|---|---|---|---|
| Correlation with COVID-19 Number of Cases | | | | | | | | | | |
| GT | 0.375** | 0.455** | 0.469** | 0.390** | 0.299** | 0.325** | 0.455** | 0.469** | 0.199** | 0.171** |
| 1D GT | −0.168** | −0.109** | – | – | −0.067* | – | – | – | −0.065* | – |
| 2D GT | – | – | – | – | 0.112** | – | 0.0778* | – | 0.115** | 0.091** |
| Reddit | 0.635** | 0.618** | 0.364** | 0.481** | 0.196** | 0.569** | 0.336** | 0.0728* | 0.236** | – |
| O3 | 0.115** | – | 0.318** | 0.221** | 0.157** | 0.267** | – | – | – | – |
| NO2 | 0.180** | 0.213** | 0.095** | 0.112** | – | – | 0.0684* | – | – | – |
| SO2 | −0.255** | – | −0.293** | −0.139** | −0.200** | −0.191** | – | – | – | – |
| CO | −0.150** | −0.099** | −0.103** | −0.122** | – | – | – | – | – | – |
| Correlation with Influenza Number of Cases | | | | | | | | | | |
| GT | 0.248** | 0.367** | 0.392** | 0.342** | 0.303** | 0.194** | 0.455** | 0.379** | 0.288** | 0.214** |
| 1D GT | – | – | 0.1087* | – | – | – | – | – | 0.134** | 0.136** |
| 2D GT | – | – | – | – | – | – | −0.094* | – | – | – |
| Min Temp. | −0.382** | −0.491** | −0.416** | −0.418** | −0.353** | −0.331** | −0.376** | −0.206** | −0.366** | −0.423** |
| Total Rain | −0.219** | −0.39** | −0.218** | −0.320** | −0.245** | −0.246** | −0.113* | – | −0.254** | −0.138** |
| Total Prec. | −0.139** | −0.308** | – | −0.210** | −0.164** | −0.174** | – | – | −0.155** | – |

In the table above,
* means that p-value $< .05$ and
** means that $p$-value $< .01$.

#### 2.1.1. Data preparation

*2.1.1.1. Datasets.* Data retrieved from websites, or by RESTful APIs such as the number of searches, page views, and news articles are superior to conventional health-related data, e.g. clinical visits, telephone triage, and hospital records, for two main reasons, (1) they could be retrieved automatically, and (2) they provide information days to weeks earlier [22]. Therefore, four different sources of data, namely, Google Trends, Reddit posts, weather data, and air quality data are used for this project. The number of COVID-19 and influenza cases are retrieved using the COVID19Tracker API, and the official Government of Canada website, respectively [64,72]. To have enough data to train a deep learning model, information is collected from May 1st, 2020 to July 31st, 2022 on a daily basis for COVID-19 and from September 1st, 2013 to September 1st, 2023 on a weekly basis for influenza.

Google Trends: provides a normalized number of Google searches on a particular topic, or search term [66]. It classifies data based on category, keyword, and country/region. For this project, the volume of Google searches for all categories of COVID-19 and influenza disease topics were gathered for various provinces of Canada on a daily and weekly basis, respectively. When cases of a particular disease rise and its symptoms become widely noticed, people tend to search for it online. Therefore, Google Trends on the disease topic are most often significantly correlated with the number of cases and can be used as an indicator for building a forecasting model. In Google Trends API, each topic has an ID; and it is equal to "/g/11j2cc_qll" and "/m/0cycc" for COVID-19 and influenza disease topics, respectively. By searching the topic ID, the normalized number of searches in a certain period on a daily basis (using the dailydata package of pytrends) is obtained and used as a predictor in the forecasting model. The first derivative of Google Trends shows whether the number of searches is increasing or decreasing, and the second derivative specifies how fast the volume of searches is changing. Fig. 2 shows the number of COVID-19 and influenza cases, Google Trends, and the first and second derivatives of Google Trends for Ontario. Since the first and second derivatives of Google Trends could also be used as indicators for forecasting outbreaks, they were fed into the models as different features, as well.

In Fig. 2(b), the number of influenza cases (orange line) from May 31, 2020, to January 23, 2022, during the COVID-19 pandemic, is very low and mostly zero. By removing this period from the datasets, the accuracy of the forecasting model significantly increased. Table 1 shows the correlation between the number of influenza cases and other indicators after removing the COVID-19 pandemic period, and the correlation between the number of COVID-19 cases and other indicators, for Canadian provinces.

Reddit posts: are retrievable using the PullPush API [67]. The PullPush API accepts the name of the subreddit, a keyword, and a time period as input and returns the posts along with their exact submission date and time. Each Canadian province has a subreddit, enabling the collection of posts containing the keyword "covid" from May 1st, 2020 onwards in the respective provincial subreddits. As shown in Table 1, the volume of the posts for most of the provinces is significantly correlated with the number of COVID-19 cases in that province. The reason is that when a certain disease becomes prevalent in a particular region, people of that region discuss it on social media, and the number of posts related to that disease increases. Therefore, the number of Reddit posts was used as a feature to feed into the deep-learning model for forecasting COVID-19 waves. However, it was not possible to gather a Reddit dataset for forecasting influenza waves because the PullPush API did not return posts dated before 2015.

Weather data: was retrieved from Open-Meteo historical weather API [68,73–76]. The Open-Meteo historical weather API returns weather parameters such as minimum temperature, total rain, and total precipitation on a daily basis for a certain coordinate. We collected data for different coordinates in Canada at a granular resolution (higher than 333 Km resolution). We then minimized the temperature values over a certain province, and summed the values for rain and precipitation. Weather parameters, i.e. minimum temperature, total rain, and total precipitation are correlated with the number of influenza cases of each province (after removing the COVID-19 period from the data), as demonstrated in Table 1, but not with the number of COVID-19 cases. Adding weather parameters to the COVID-19 model degraded the final accuracy. Therefore, weather parameters, i.e. minimum temperature, total rain, and total precipitation were fed as additional features to the deep-learning model only for forecasting influenza waves.

Air Quality: data was retrieved from the Google earth engine, SENTINEL-5P mission instrument, launched by the European Commission (EC) as part of the European Earth Observation Program (Copernicus) [69]. Using the shapefile of Canada, the concentration of a certain air quality parameter was maximized over the coordinates of a certain province (1113.2 m resolution) to acquire the maximum concentration of that parameter for each day. The maximum daily concentration of air pollutants, i.e. O3, SO2, CO, and NO2 were correlated with the number of COVID-19 cases for most of the provinces, as illustrated in Table 1, and therefore, used as additional features for improving the accuracy of the COVID-19 model. However, since the SENTINEL-5P mission was initiated in 2018, it was not possible to use it in the influenza model.

It is clear from Table 1 that there are 9 and 7 features for each province in the COVID-19 and influenza models, respectively. These features for COVID-19 include the historical COVID-19 number of cases, Google Trends, first and second derivatives of Google Trends, number of Reddit posts, maximum daily concentration of O3, NO2, SO2, and CO, and for influenza include historical influenza number of cases, Google Trends, first and second derivatives of Google Trends, minimum daily temperature, total daily rain, and precipitation, namely.

### 2.2. Data preprocessing

In the data preprocessing block (Fig. 1(a)), missing values are filled by averaging the preceding and succeeding records. All the features that are negatively correlated with the number of cases (Table 1) are normalized and subtracted from 1 so that they have a positive correlation. The data of different provinces are then integrated into different dimensions of a NumPy array. Thus, the final array has three dimensions, (1) data features, (2) time series sequence, and (3) different provinces of Canada. The final array is scaled using the min-max scaling algorithm. Then, the observations are formed. To apply T step-ahead prediction, the observation for time t includes a sequence of all features with length $\tau + 1$, (t-$\tau$, …, t-2, t-1, t) as the input and a sequence of the number of cases only with length T, ($t + 1$, $t + 2$, …, $t + T$) as the label. The input sequence length, which is determined through hyperparameter optimization, is equal to 150 days for COVID-19 and 15 weeks for Influenza models. Hyperparameter optimization which is an important phase in training deep-learning models, greatly affects the final accuracy [84–86]. The output sequence length, T, is 56 days for COVID-19 and 8 weeks (which is also 56 days) for influenza. The observations are then converted into tensors, and fed into the second block, graph creation.

#### 2.2.1. Graph creation

In the graph creation stage, the train and test datasets are separated. The train dataset is used to set the weights of the GNN graph. In a GNN model, nodes that are more related to each other share their data for making a better decision. Therefore, to improve the accuracy of the EWS model, Canadian provinces that have a significant correlation ($p < .05$) in terms of the number of COVID-19 and influenza cases are selected for forming the GNN graph [65]. As displayed in Fig. 1(b), the territorial provinces, i.e. Yukon, Nunavut, and Northwest Territories did not have sufficient data; therefore, their case numbers were not correlated with the rest of the provinces, and were excluded from the model design, and the GNN graph. Our investigations indicate that rather than building a

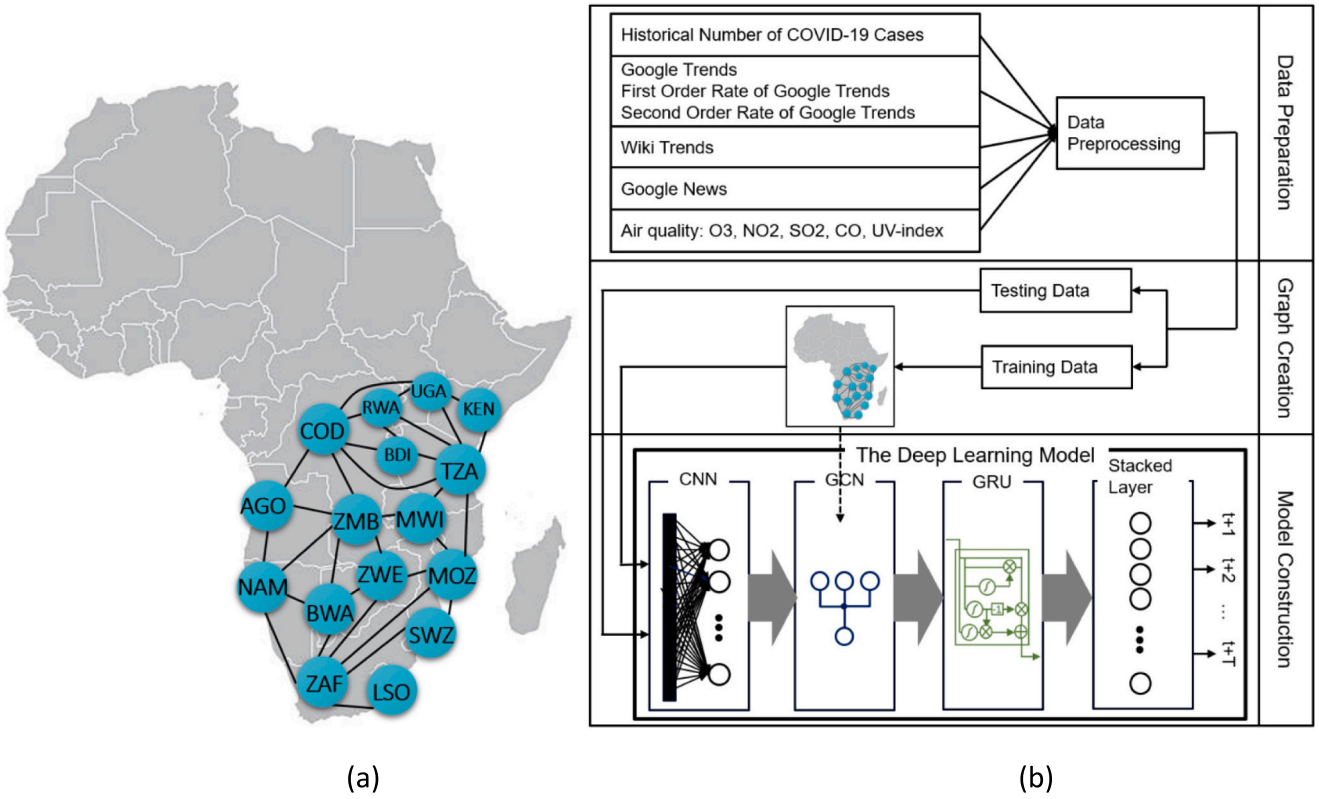



**Fig. 3.** (a) The overall model for southern African countries, and (b) the graph used for building the GNN layer for southern African countries.

fully connected graph, linking only provinces that have a common border with each other increases the accuracy of the model. The edge index matrix of the graph is defined using its topology.

Similar to [57] the weights of the graph are set using the correlation between the number of cases of the two provinces. Because Google and Facebook mobility data are not updated anymore, they cannot be used for designing and training future models [40,41]. Moreover, other sources of human mobility are not available on the provincial-level [70]. Since, mobility data is most often correlated with the number of infections [58], in this work, the correlation between the number of disease cases in the training dataset is used as the weights of the graph.

The edge index and edge weight matrices are fed to the GCN layer in the model construction block to define and form the graph. The training and testing datasets are fed into the model construction block.

#### 2.2.2. Model construction

The model construction block takes the training and testing data to build the deep learning model, which is composed of four main layers, CNN, GCN, GRU, and a linear neural network layer. The first layer is a two-dimensional CNN that includes 9 and 7 input channels for the COVID-19 and influenza models respectively, corresponding to their input features, and one output channel. Each input channel accepts a two-dimensional tensor, i.e. a sequence of the feature for each province. So, the CNN combines all the data feature sequences of a province and provides one sequence for each province to the GCN layer. In other words, the CNN layer turns the three-dimensional input tensor into a two-dimensional tensor and feds it into the GCN. The GCN which is the spatial layer of the model, applies the aggregation and update procedures on the provinces and feeds the result to the GRU which is the temporal layer of the model. The GRU which is capable of remembering relationships in data sequences, provides its result to the linear layer, which has T outputs for 1- to T-steps ahead prediction. A dropout component is placed between each layer for regularization. Each layer has its own hyperparameters, which are optimized prior to training and testing the model. The hyperparameters of the models include sequence length, batch size, dropout probabilities, learning rate and its decay value, activation functions, kernel of CNN, number of layers of GRU, and the hidden layer size of GRU.

### 2.3. Southern African countries

Fig. 3(a) shows the GNN graph formed on southern African countries that are under study to perform 56 days-ahead prediction for COVID-19 number of cases. The input sequence length, determined through hyperparameter optimization is equal to 150 days for the forecasting model of COVID-19 in southern African countries. The designed model which is depicted in Fig. 3(b), is different from the model of Canada (Fig. 1(a)) in two ways: (1) the data sources, and (2) the GCN graph, which are explained in the following.

#### 2.3.1. The data sources

To automate the data gathering stage, datasets are collected from five different online sources that are regularly updated. The number of COVID-19 cases released by the official website of WHO on a daily basis is used as the historical number of cases from June 1st, 2020 to July 31st, 2022 [77]. Google Trends on the COVID-19 disease topic (ID = "/g/11j2cc_qll") which is significantly correlated with the number of COVID-19 cases for most of the countries (for South Africa, corr = 0.5137, $p <$ .000001), is collected on a daily basis for different countries and used alongside its first and second derivatives as additional features to forecast COVID-19 waves.

Wiki Trends is an API that returns the number of views for different Wikipedia pages on a daily basis [78]. When the number of cases of a particular disease increases in a certain region, people of that region tend to look for more information regarding that disease online; as a result, the number of Wikipedia views on that disease rises. Since there exists a separate Wikipedia page on COVID-19 for each country, it is possible to use this resource to build an EWS for COVID-19 waves. The

number of views is significantly correlated with the number of COVID-19 cases for many of the African countries (for South Africa, corr = 0.1878, p < .000001).

Google News API enables search for news pages that have been released in a particular country, and include a specific keyword [63]. The API returns the title, link, and the exact release date of the article. When the number of cases of a certain disease increases in a region, more news articles regarding that disease, its symptoms, and remedy are released; therefore, the number of released Google News articles related to a particular disease is most often significantly correlated with the number of cases of that disease. We found that the number of Google News published with the keyword "covid" is significantly correlated with the number of COVID-19 cases for many of the African countries under study (for South Africa, corr = 0.1797, $p = .0148$).

Air quality data which was gathered using Google Earth Engine API for all countries under study, includes the maximum daily concentration of five different air pollutants, namely, SO2, NO2, CO, O3, and UV-index. Using the shapefile of Africa, all the daily concentration values of the air pollutants over a particular country (1113.2 m resolution) were gathered and the maximum for the whole country was used. Overall, from the five different sources of data, eleven features are extracted; therefore, the CNN layer in the model construction block, includes eleven input channels and one output channel. The CNN layer combines all the features into one feature for each province and provides it to the GCN layer. The GCN uses this data along with the edge index and edge weights of the GCN graph to analyze the data on a spatial level.

### 2.3.2. The GCN graph

As shown in Fig. 3(a), the GCN graph is constructed from 16 different southern African countries. The reason for this choice is that the number of COVID-19 cases in these countries are significantly ($p < .05$) correlated with each other. In a GNN model, nodes that are more related to each other share their data for making a better decision. Therefore, forming the GNN graph using correlated nodes will improve the final accuracy. Our investigations indicate that rather than building a fully connected graph, linking only countries that have a common border with each other increases the accuracy of the model. The weights of the graph are set equal to the correlation of the number of COVID-19 cases in the training dataset between the two countries. Based on the topology of the graph and the weights of the links the edge index and edge weight matrices are created and passed to the GCN layers. The GCN layer also accepts data from the CNN at its input, performs the aggregate and update procedures on spatial dimension and passes the results to the GRU. The GRU which is capable of memorizing sequences of data analyzes the data on temporal dimension and provides the result to the final stacked NN layer which includes T outputs for T step-ahead prediction. Before using the train and test datasets to build the model, hyper-parameters are optimized to achieve the best results.

## 3. Results

The accuracy of the model is evaluated by three different metrics, Root Mean Square Error (RMSE), coefficient of determination (R2-score), and correlation, which are presented in eqs. 1–3, respectively.

$$RMSE = \sqrt{\frac{1}{n}\sum_i (a_i - p_i)^2} \tag{1}$$

$$R^2 = 1 - \frac{\sum_i (a_i - p_i)^2}{\sum_i (a_i - \bar{a})^2} \tag{2}$$

**Table 2**
Evaluation of the accuracy of the proposed model against other models for different number of days in advance for COVID-19 in Ontario.

| Evaluating COVID-19 Model for Ontario | | | | |
|---|---|---|---|---|
| | Model | RMSE | R2-score | Correlation |
| 1 day-ahead | GRU | 1443.1537 | 0.7357 | 0.8192* |
| | Cola-GNN | 1411.5913 | 0.7447 | 0.8344* |
| | CNN-GCN-LSTM | 988.4723 | 0.8211 | 0.9217* |
| | Informer | 1254.1963 | 0.8042 | 0.8242* |
| | Autoformer | 1161.6517 | 0.8144 | 0.8344* |
| | CNN-GCN-GRU | 866.8081 | 0.8691 | 0.9412* |
| 14 days-ahead | GRU | 1259.006 | 0.6546 | 0.7672* |
| | Cola-GNN | 1148.0051 | 0.6814 | 0.7936* |
| | CNN-GCN-LSTM | 980.1725 | 0.7928 | 0.9078* |
| | Informer | 1311.9353 | 0.7391 | 0.7557* |
| | Autoformer | 1259.4979 | 0.7463 | 0.763* |
| | CNN-GCN-GRU | 904.3313 | 0.8497 | 0.9296* |
| 28 days-ahead | GRU | 1531.7383 | 0.5952 | 0.7185* |
| | Cola-GNN | 1329.7801 | 0.6383 | 0.7559* |
| | CNN-GCN-LSTM | 1046.2955 | 0.7809 | 0.8922* |
| | Informer | 1664.3532 | 0.6939 | 0.6993* |
| | Autoformer | 1538.0796 | 0.7094 | 0.7149* |
| | CNN-GCN-GRU | 938.345 | 0.8317 | 0.9096* |
| 42 days-ahead | GRU | 1841.1215 | 0.5374 | 0.6612* |
| | Cola-GNN | 1761.1983 | 0.5515 | 0.6807* |
| | CNN-GCN-LSTM | 1197.0164 | 0.7501 | 0.8664* |
| | Informer | 1769.5534 | 0.6815 | 0.6679* |
| | Autoformer | 1704.6728 | 0.6891 | 0.6755* |
| | CNN-GCN-GRU | 1196.2316 | 0.7546 | 0.8698* |
| 56 days-ahead | GRU | 2497.7258 | 0.4405 | 0.5651* |
| | Cola-GNN | 2065.2009 | 0.5009 | 0.619* |
| | CNN-GCN-LSTM | 1505.3791 | 0.6849 | 0.8139* |
| | Informer | 2387.8943 | 0.6165 | 0.576* |
| | Autoformer | 2272.732 | 0.6277 | 0.5872* |
| | CNN-GCN-GRU | 1505.1116 | 0.6871 | 0.8164* |

| Evaluating Influenza Model for Quebec | | | | |
|---|---|---|---|---|
| | Model | RMSE | R2-score | Correlation |
| 1 week-ahead | GRU | 281.9151 | 0.7532 | 0.7021* |
| | Cola-GNN | 254.878 | 0.82 | 0.7967* |
| | CNN-GCN-LSTM | 232.1351 | 0.8861 | 0.8407* |
| | Informer | 291.5814 | 0.7319 | 0.7279* |
| | Autoformer | 267.0295 | 0.7886 | 0.7657* |
| | CNN-GCN-GRU | 90.1274 | 0.9835 | 0.9926* |
| 2 weeks-ahead | GRU | 273.7503 | 0.7722 | 0.7148* |
| | Cola-GNN | 269.2848 | 0.783 | 0.772* |
| | CNN-GCN-LSTM | 235.3229 | 0.8762 | 0.8341* |
| | Informer | 301.3555 | 0.7354 | 0.7303* |
| | Autoformer | 268.4916 | 0.785 | 0.7633* |
| | CNN-GCN-GRU | 117.7067 | 0.9719 | 0.9873* |
| 4 weeks-ahead | GRU | 323.5803 | 0.6692 | 0.6461* |
| | Cola-GNN | 259.0234 | 0.809 | 0.7893* |
| | CNN-GCN-LSTM | 259.3676 | 0.8081 | 0.7887* |
| | Informer | 329.7142 | 0.6584 | 0.6789* |
| | Autoformer | 292.7803 | 0.7293 | 0.7262* |
| | CNN-GCN-GRU | 157.4489 | 0.9498 | 0.9803* |
| 6 weeks-ahead | GRU | 337.4767 | 0.6452 | 0.6301* |
| | Cola-GNN | 262.8911 | 0.799 | 0.7827* |
| | CNN-GCN-LSTM | 304.4590 | 0.7053 | 0.7202* |
| | Informer | 339.8939 | 0.6412 | 0.6675* |
| | Autoformer | 323.726 | 0.6689 | 0.6859* |
| | CNN-GCN-GRU | 157.339 | 0.9498 | 0.9845* |
| 8 weeks-ahead | GRU | 336.2793 | 0.6472 | 0.6315* |
| | Cola-GNN | 272.1649 | 0.776 | 0.7673* |
| | CNN-GCN-LSTM | 299.0645 | 0.7162 | 0.7275* |
| | Informer | 348.8513 | 0.6268 | 0.6579* |
| | Autoformer | 333.4062 | 0.6521 | 0.6747* |
| | CNN-GCN-GRU | 250.4763 | 0.8729 | 0.953* |

In the table above,
* means that the p-value < .000001.

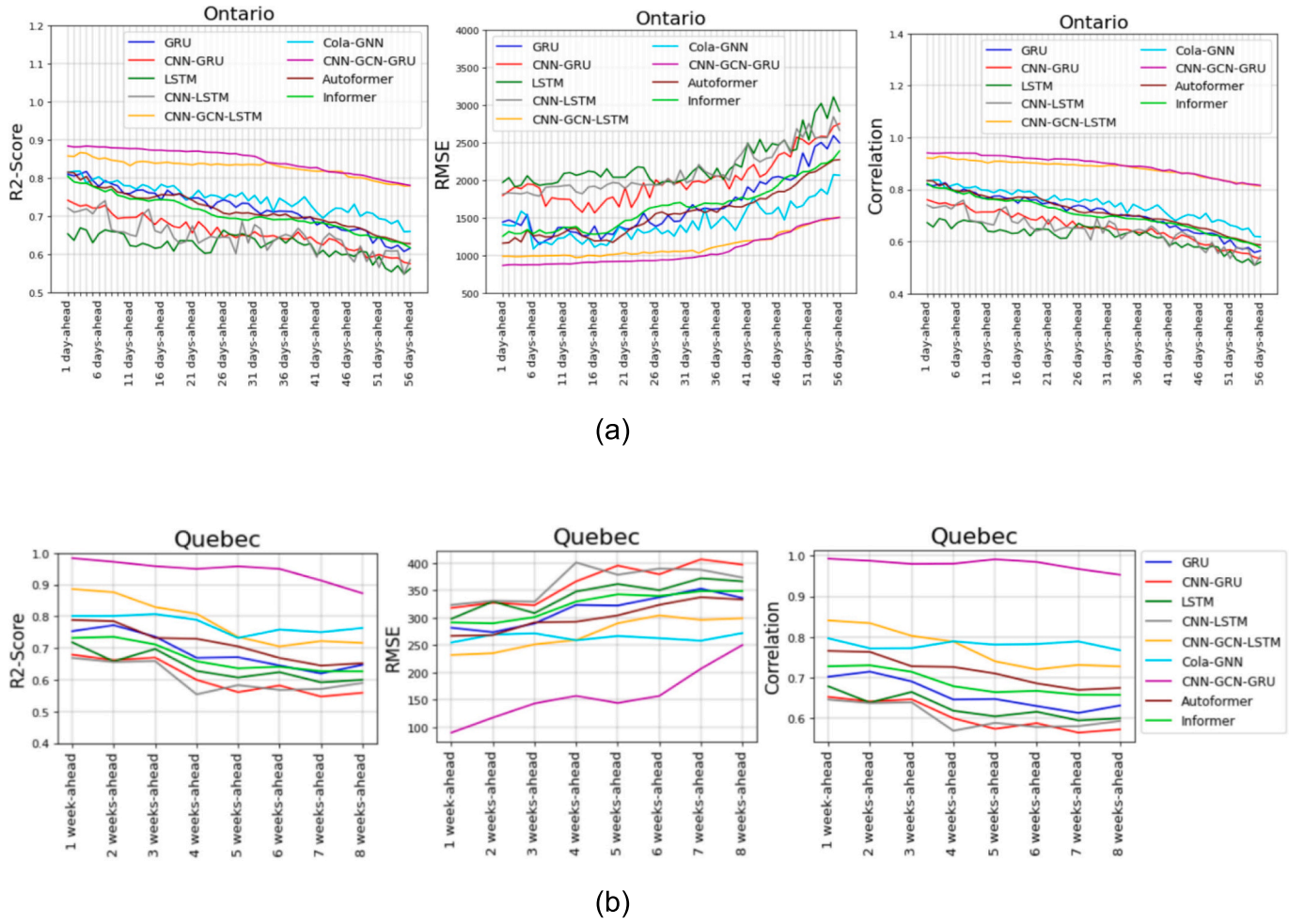



**Fig. 4.** Comparing the deep-learning model with state-of-the-art models in terms of R2-score, RMSE, and correlation for (a) COVID-19 in Ontario, and (b) influenza in Quebec.

$$Corr = \frac{n\sum_i a_i p_i - \sum_i a_i \sum_i p_i}{\sqrt{n\sum_i a_i^2 - \left(\sum_i a_i\right)^2}\sqrt{n\sum_i p_i^2 - \left(\sum_i p_i\right)^2}} \quad (3)$$

Where ai and pi are the *i*th actual and predicted values, respectively, a signifies the mean actual values, and n is the total number of predicted values. We have compared our model (CNN-GCN-GRU) with state-of-the-art models, including LSTM, GRU, CNN-LSTM, CNN-GRU, CNN-GCN-LSTM, Cola-GNN [62], and two transformer models designed for time-series prediction, namely, Informer [79] and Autoformer [80].

### 3.1. Canada

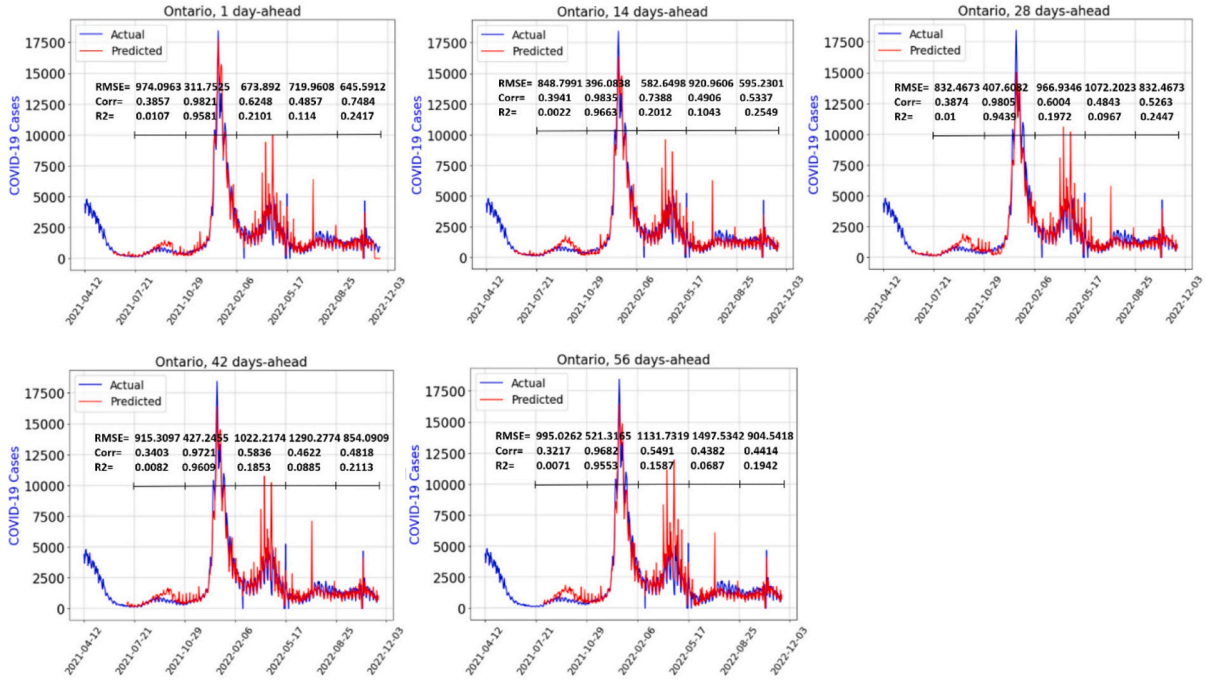
The best results for COVID-19 and influenza were obtained for Ontario and Quebec, respectively. Table 2 compares the performance of our model for COVID-19 in Ontario and Influenza in Quebec, with other methods in terms of RMSE, R2-score, and correlation. Table 2 shows that our model improved R2-score, RMSE, and correlation by up to 55.98 %, 39.71 %, and 44.47 % for 56 days-ahead COVID-19 prediction in Ontario and 34.87 %, 25.52 %, 50.91 % for 8 weeks-ahead influenza prediction in Quebec, respectively.

Fig. 4(a) shows the different metrics for state-of-the-art models for COVID-19 in Ontario and Fig. 4(b) shows that for influenza, in Quebec.

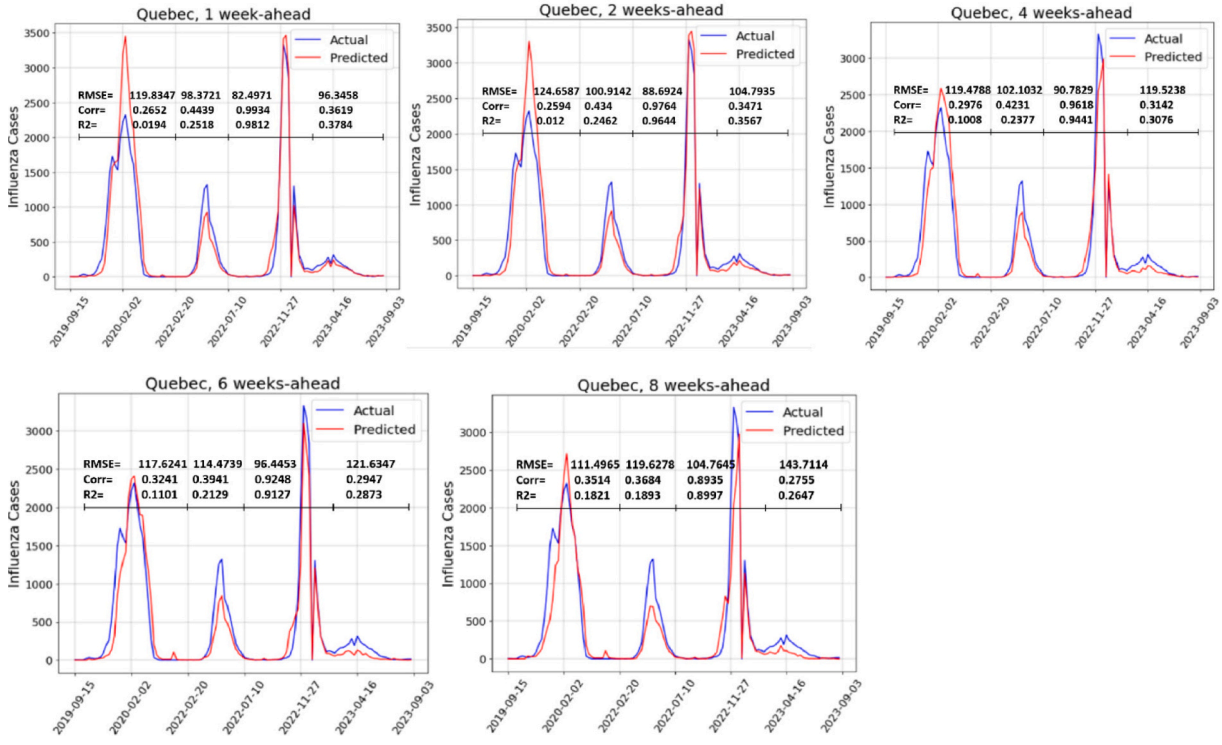
Figs. 5(a and b) compare the predicted and actual values of COVID-19 in Ontario and influenza in Quebec, respectively. For influenza prediction (Fig. 5(b)), the COVID-19 period, from May 31, 2020 to January 23, 2022 was removed from the dataset, since the number of cases was very low and almost zero in that period. According to Fig. 5 the predicted values accurately follow the actual values for different steps in advance.

Figs. 5(a and b) also show the accuracy of the model for different waves of COVID-19 in Ontario, and influenza in Quebec, respectively, and demonstrate that the model performs particularly well for one of the waves. The reason is that the hyperparameters which were optimized using the WandB package, were fine-tuned specifically for that wave. Although, we optimized the hyperparameters to maximize overall accuracy (R2-score), the algorithm prioritized optimizing them for that particular wave. Recalibrating the hyperparameters for each wave could increase the accuracy of the model in that time period.

Fig. 6(a and b) which have been created using ArcGIS Pro, further illustrates the accuracy of the COVID-19 and influenza models for different provinces of Canada, respectively. As shown in Fig. 6, the accuracy of the model is acceptable for all the provinces under study;




(a)




(b)

**Fig. 5.** comparing predicted and actual values for different steps in advance for (a) COVID-19 in Ontario and (b) influenza in Quebec.

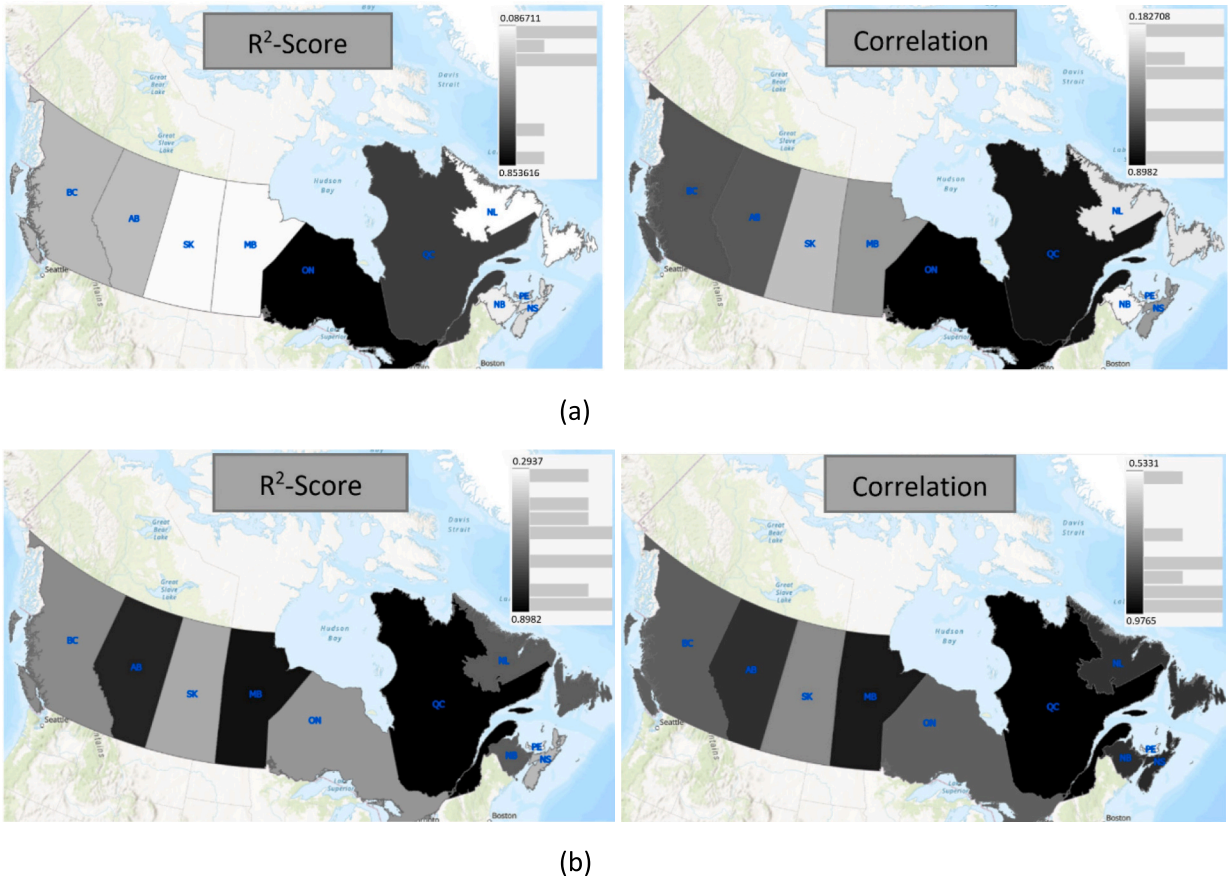



**Fig. 6.** Accuracy of the (a) COVID-19 and (b) influenza forecasting models for different provinces of Canada.

however, higher accuracy can be achieved for a particular province by fine-tuning the hyperparameters specifically for that province. Complete results of COVID-19 and influenza prediction for different days and weeks in advance for all provinces of Canada is available in supplementary file 1. In the following the results for southern African countries is provided.

### 3.2. Southern African countries

Table 3 compares the accuracy of our model with other state-of-the-art methods for South Africa. According to Tables 3, our model improves the RMSE, R2-score, and correlation by up to 52.74 %, 66.89 %, and 33.67 % for 1 day-ahead, 49.88 %, 66.77 %, and 49.17 % for 14 days-ahead, 44.31 %, 64.8 %, and 41.58 % for 28 days-ahead, 48.85 %, 61.44 %, and 48.3 % for 42 days ahead, and 32.04 %, 51.04 %, and 28.74 % for 56 days-ahead prediction, respectively. Fig. 7 evaluates the metrics of our model for different numbers of days in advance prediction against other methods, and displays its superiority.

Fig. 8 compares the predictions with the actual values for three different countries. It is expected for the accuracy to degrade, as the in-advance number of days increases. Nonetheless, our model works fine and could alert of upcoming waves by up to 56 days-ahead prediction. It is evident from Fig. 8 that the peaks and upsurges are mostly predicted without delay. Therefore, our model could efficiently help with preparedness for and rapid response to future COVID-19 waves.

Fig. 8 also provides the accuracy of the model for different COVID-19 waves in South Africa. The statistics show that the hyperparameters have been fine-tuned to work best for one of the waves. Setting new hyperparameters for each period could significantly increase the accuracy of the model for that time.

While Fig. 8 evaluates our model for only three of the southern African countries, our model delivers great results for all the countries under study. Fig. 9 compares the accuracy of the model for different countries under study. By fine-tuning the hyperparameters specifically for a certain country, a higher accuracy for that country will be obtained. A complete evaluation of our model is available in supplementary file 1.

### 3.3. Feature importance and attention map

Feature importance was studied using the permutation importance method, in which each feature is permuted and the degradation in accuracy is calculated. The features that have a higher degradation contribute more to building the final accuracy of the model [81]. In this work, for each feature i.e. data source, 10 different permutations in the input sequence and for all the provinces/countries under study were used. For each permutation, accuracy metrics i.e. RMSE, R2-Score, and correlation, were averaged over 1 to 56 days-ahead (8 weeks-ahead) prediction. Finally, degradations of different features were compared using the Mann-Whitney $U$ test $p$-values. Fig. 10(a, b, and c) shows the feature importance results for COVID-19 in Ontario, influenza in Quebec, and COVID-19 in South Africa, respectively. According to Fig. 10(a), Reddit plays the most important role in building the COVID-19 forecasting model in Ontario. Moreover, Reddit, Google Trends, and

**Table 3**
Evaluation of the accuracy of the proposed model against other models for different number of days in advance for South Africa.

| | Model | RMSE | R2-score | Correlation |
|---|---|---|---|---|
| 1 day-ahead | LSTM | 905.7236 | 0.7449 | 0.8643* |
| | GRU | 959.2577 | 0.7099 | 0.8521* |
| | CNN-LSTM | 1463.3280 | 0.2697 | 0.6855* |
| | CNN-GRU | 1094.8315 | 0.6112 | 0.7922* |
| | CNN-GCN-LSTM | 734.0415 | 0.818 | 0.8872* |
| | Cola-GNN | 737.0097 | 0.7795 | 0.8795* |
| | Informer | 877.828 | 0.7866 | 0.8867* |
| | Autoformer | 875.6571 | 0.7884 | 0.8884* |
| | CNN-GCN-GRU | 691.4755 | 0.8145 | 0.9163* |
| 14 days-ahead | LSTM | 1169.4908 | 0.5578 | 0.7474* |
| | GRU | 1266.6047 | 0.4738 | 0.6924* |
| | CNN-LSTM | 1484.1531 | 0.2593 | 0.6847* |
| | CNN-GRU | 1383.0138 | 0.3636 | 0.6099* |
| | CNN-GCN-LSTM | 806.0130 | 0.7124 | 0.829* |
| | Cola-GNN | 821.0267 | 0.7283 | 0.8543* |
| | Informer | 985.3697 | 0.6835 | 0.8095* |
| | Autoformer | 948.1624 | 0.7086 | 0.8346* |
| | CNN-GCN-GRU | 743.793 | 0.7803 | 0.9098* |
| 28 days-ahead | LSTM | 1379.9717 | 0.3839 | 0.6543* |
| | GRU | 1277.7921 | 0.4784 | 0.6904* |
| | CNN-LSTM | 1513.0831 | 0.2491 | 0.717* |
| | CNN-GRU | 1416.8820 | 0.3479 | 0.6066* |
| | CNN-GCN-LSTM | 977.8331 | 0.6792 | 0.8243* |
| | Cola-GNN | 855.3383 | 0.6916 | 0.8456* |
| | Informer | 1083.904 | 0.6273 | 0.7813* |
| | Autoformer | 1090.6204 | 0.6232 | 0.7772* |
| | CNN-GCN-GRU | 842.5194 | 0.7860 | 0.8967* |
| 42 days-ahead | LSTM | 1341.1658 | 0.4054 | 0.6643* |
| | GRU | 1252.1205 | 0.4877 | 0.7007* |
| | CNN-LSTM | 1444.6724 | 0.3021 | 0.6058* |
| | CNN-GRU | 1442.3221 | 0.3045 | 0.5814* |
| | CNN-GCN-LSTM | 1187.4599 | 0.5325 | 0.7005* |
| | Cola-GNN | 913.7837 | 0.6349 | 0.8169* |
| | Informer | 1135.8344 | 0.5849 | 0.7669* |
| | Autoformer | 1105.2435 | 0.6034 | 0.7854* |
| | CNN-GCN-GRU | 738.963 | 0.7836 | 0.8984* |
| 56 days-ahead | LSTM | 1152.0485 | 0.482 | 0.6962* |
| | GRU | 1163.5759 | 0.4706 | 0.6911* |
| | CNN-LSTM | 1269.0783 | 0.3603 | 0.6624* |
| | CNN-GRU | 1294.9145 | 0.3317 | 0.6441* |
| | CNN-GCN-LSTM | 1158.3298 | 0.5205 | 0.7094* |
| | Cola-GNN | 1081.0475 | 0.5161 | 0.7261* |
| | Informer | 1160.0878 | 0.4992 | 0.7092* |
| | Autoformer | 1132.4348 | 0.5134 | 0.7234* |
| | CNN-GCN-GRU | 880.0286 | 0.6775 | 0.8292* |

In the table above,
* means that the p-value $< .000001$.

historical data are the most important data sources for the COVID-19 model in Ontario. Fig. 10(b) shows that temperature plays the most important role in building the influenza forecasting model in Quebec. Then historical data, and then Google Trends. Fig. 10(c) shows that Google Trends is the most important data source for building the accuracy of the COVID-19 forecasting model in South Africa; and in general, 4 data sources, namely, Google Trends, historical data, Google News, and Wiki Trends are the most important features used for this model. All the *p*-values are significant ($p < .05$), except for the p-value between Google News and Wiki Trends in Fig. 10(c), COVID-19 forecasting model for South Africa.

Attention maps are used to understand which parts of the data contribute more to producing the final decision of the model. In this work, the saliency method was used to depict the attention maps of our models. In the saliency method, the data points that have a higher absolute value of the gradients are considered more important [82,83]. Therefore, the mean value of the gradients over the testing dataset was computed. Figs. 11(a, b, and c) show the attention maps for the forecasting model of COVID-19 in Ontario, Influenza in Quebec, and COVID-19 in South Africa, respectively. Fig. 11(a) verifies that Reddit, historical data, and Google Trends are the most important data sources used. Additionally, it demonstrates that Ontario mostly relies on data derived from Quebec and Atlantic provinces to make its final decision. Furthermore, samples of the sequence that are closer to the predicting sample play a more important role in building accuracy. Fig. 11(b) shows that temperature and historical data greatly help build the final accuracy of the influenza forecasting model in Quebec. Moreover, three provinces, namely, Ontario, Quebec, and New Brunswick play an important role in the final results. Finally, the three samples of the sequence that are closer to the predicting sample contribute the most in providing the prediction. Fig. 11(c) shows that various data sources including Google Trends, historical data, Goolge News, Wiki Trends, first and second derivatives of Google Trends, and NO2 play an important role in providing the final decision of the COVID-19 forecasting model in South Africa. All the countries under study have an equal contribution to building the final accuracy, and samples of the input sequence that are further from the predicting sample are more important to the model.

## 4. Discussion

Scientists and health authorities have long been concerned with building efficient EWS for infectious diseases. Early efforts mostly used health-related data such as clinical records, hospital visits, and telephone triages. Recently, researchers have proposed to use web-based information and data accessible through RESTful APIs, because they are easier and faster to collect, become available weeks to days earlier than conventional health records, and could be gathered automatically through computer programs. Most previous works use datasets, such as Google Mobility and Facebook mobility data, that are no longer updated and could not be employed for future foundations. In this work, datasets

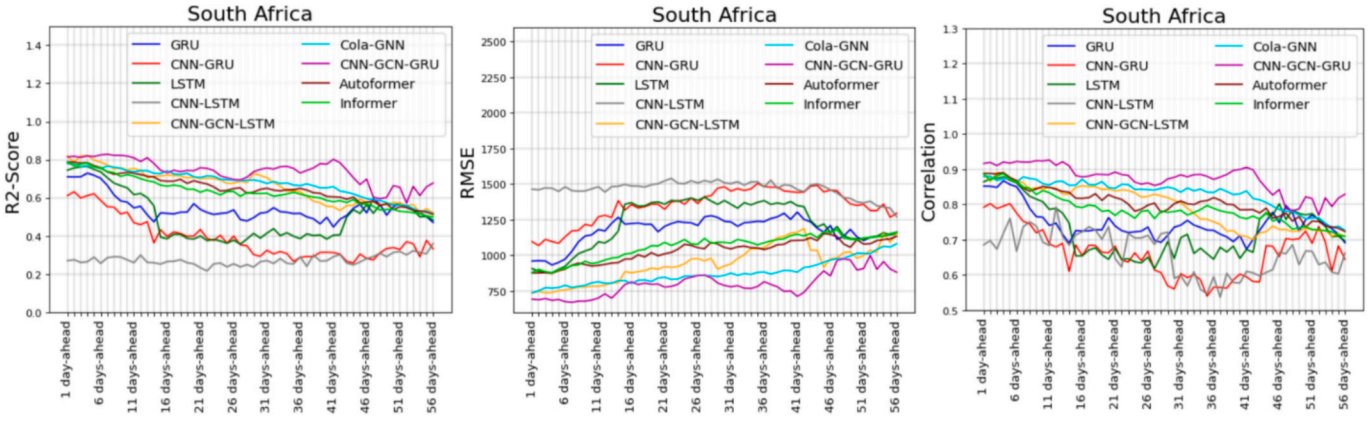



**Fig. 7.** Evaluation of the proposed model against other models for different number of days in advance for South Africa.

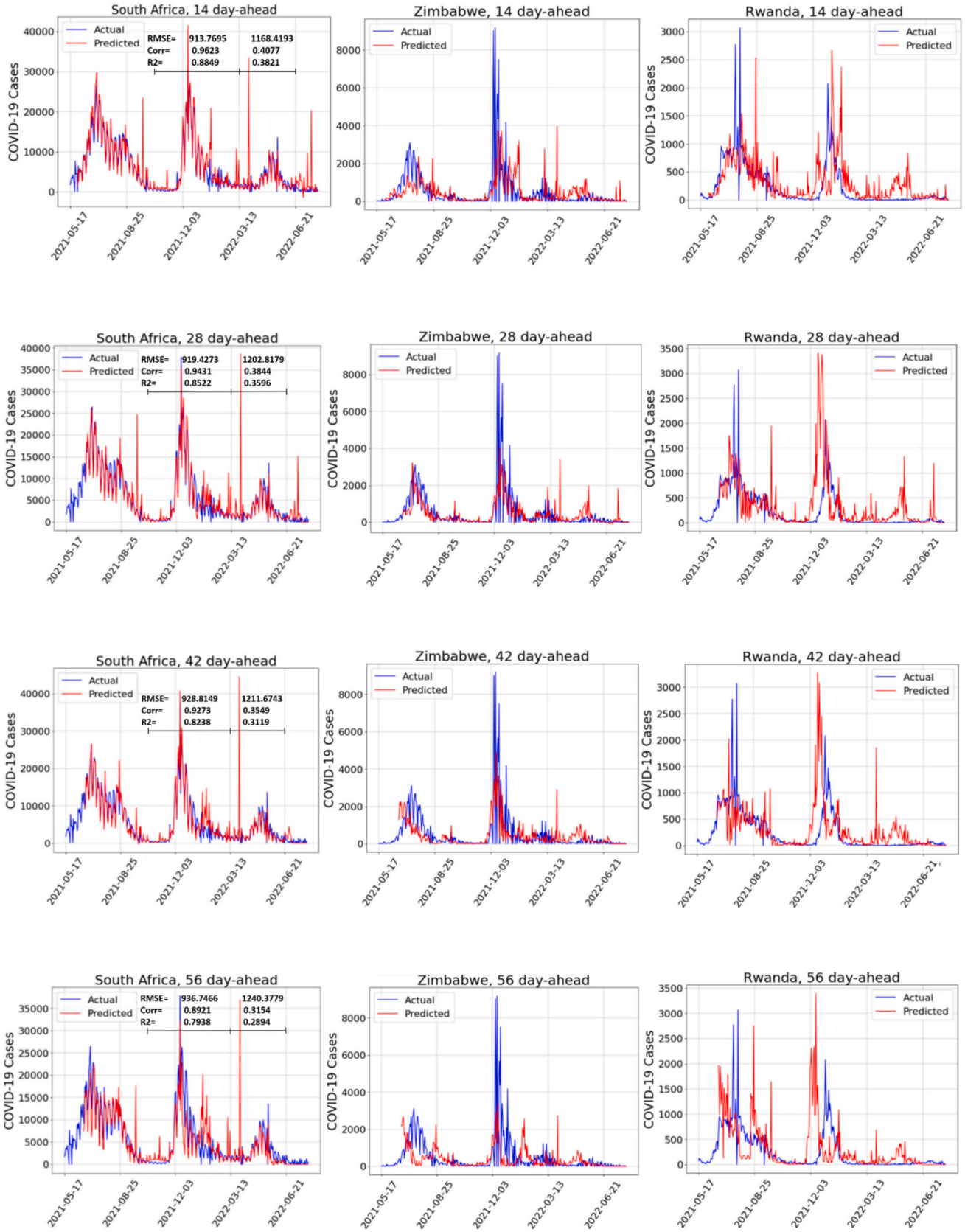



**Fig. 8.** Comparison of the predicted values with the actual values for different number of days in advance for South Africa, Zimbabwe, and Rwanda.

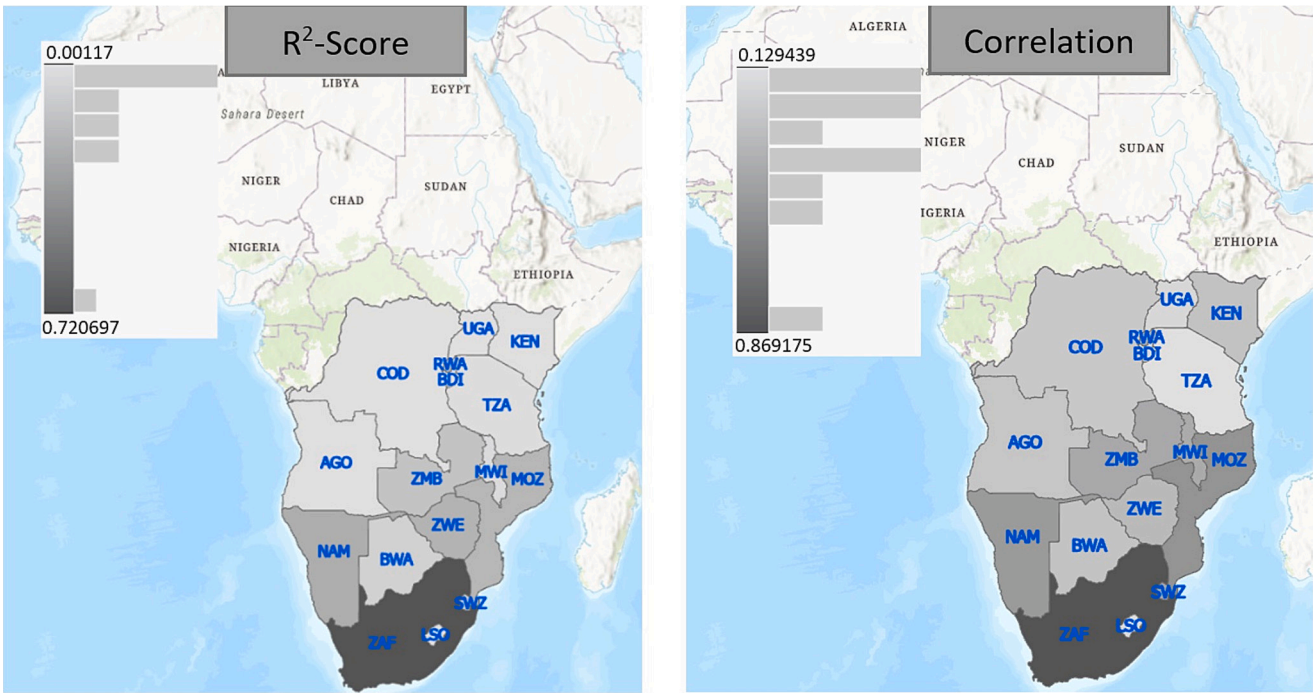



**Fig. 9.** Accuracy of the COVID-19 forecasting model for different southern African countries.

that are publicly available from websites or by RESTful APIs, and are regularly updated have been utilized to build an accurate EWS model for COVID-19 and influenza in provinces of Canada and COVID-19 in southern African countries. Thus, our framework could easily be implemented to automatically gather the required dataset, and forecast future COVID-19 waves.

In addition to the merits mentioned above, our model is capable of accurately predicting COVID-19 upsurges and peaks by up to 56 days in advance. This outstanding performance is observable for all the countries and provinces under study (please refer to supplementary file 1).

Our proposed EWS model is extendable to other continents and nations and for other respiratory infections, as we have implemented it to forecast different respiratory infections (i.e. COVID-19 and Influenza) and in different regions and countries around the globe. The best outcome for forecasting COVID-19 and Influenza in Canada was obtained for Ontario and Quebec, respectively, and for COVID-19 in southern African countries was acquired for South Africa. The RMSE, R2-score, and correlation were improved by up to 55.98 %, 39.71 %, and 44.47 % for 56 days-ahead COVID-19 prediction in Ontario, 34.87 %, 25.52 %, 50.91 % for 8 weeks-ahead influenza prediction in Quebec, and 51.04 %, 32.04 %, and 28.74 % for 56 days-ahead COVID-19 prediction in South Africa, respectively. These outstanding results alongside our fascinating outcomes for other provinces of Canada and countries in southern Africa indicate the strength of our model for building EWSs for respiratory infections (please refer to supplementary file 1 for more information).

There are some limitations to this work. Firstly, a model could be implemented for a certain country only when enough data is accessible for that country. Unfortunately, we were not able to build our framework for Influenza in southern African countries, since the number of Influenza cases is not available for many of the countries in that area [71]. Surveillance strategies need to be established in all nations around the globe, especially developing countries since infectious diseases are more prevalent in low-income regions. Secondly, sometimes datasets are not provided in developer-friendly formats. For example, the number of influenza cases in Canada is provided as PDF files, which could in turn cause difficulties for automatic retrieval through programming languages. Datasets provided in CSV, JSON, excel, or through RESTful APIs are very helpful and beneficial to automatic data retrieval. Thirdly, to implement a real-time forecasting model, capable of predicting future outbreaks, significant computational resources and infrastructure are required. Given the complexity of the model, the hyperparameters need to be optimized and the model needs to be trained and evaluated on Graphics Processing Units (GPUs) to ensure time efficiency. The datasets must be stored in fast and reliable storage systems to enable efficient access and processing during both training and inference. Additionally, the data sources (e.g. Google Trends, Reddit, Wiki Trends, Google News, weather, and air quality) need to be updated beforehand, which increases the time complexity of the overall system. Only after the model is trained and evaluated on a strong infrastructure, the results could be transmitted to and consumed by resource-constrained environments such as smartphones and edge devices. Fourthly, we only focused on respiratory infections in this work. Future research could explore extending the framework to other infectious diseases, e.g. vector-borne diseases, water- and food-borne diseases, or zoonotic diseases, which depend on other indicators such as climate, vegetation, and animal life. Incorporating such indicators would be essential for building a broader EWS. Furthermore, it is also possible to develop a unified EWS capable of forecasting outbreaks across multiple diseases, simultaneously. This would require multi-task learning approaches and modular, scalable frameworks that can accommodate the varying characteristics of different diseases. Fifthly, integrating adaptive algorithms for dynamic data streams is another promising direction. Techniques like online learning and real-time data assimilation could enable the model to update continuously as new data becomes available, reducing the need for periodic retraining. Furthermore, adaptive methods could allow the model to recognize shifts in disease dynamics, such as seasonality or mutations, and adjust predictions accordingly. Moreover, robust methods for handling noisy or incomplete data in dynamic settings, such as Bayesian techniques and uncertainty quantification, could further improve the system's reliability and applicability to real-world scenarios. Sixthly, in addition to forecasting re-emerging outbreaks, it is of high value and importance to concentrate on building EWS for emerging diseases, especially emerging respiratory infectious diseases that have frequently and ubiquitously occurred throughout history, e.g. SARS, MERS, and COVID-19.

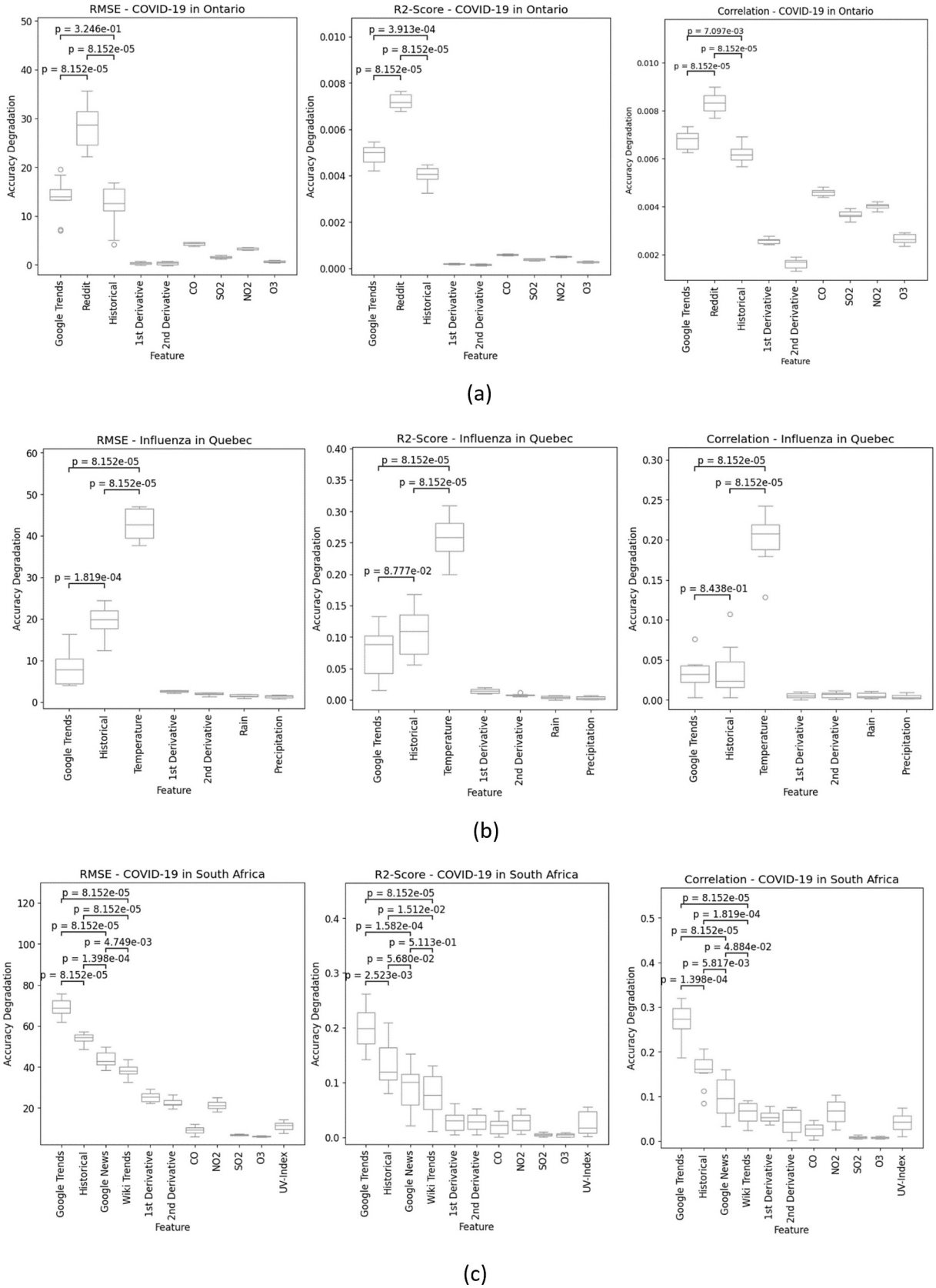



**Fig. 10.** Feature importance for (a) COVID-19 forecast in Ontario, (b) influenza forecast in Quebec, and (c) COVID-19 forecast in South Africa.

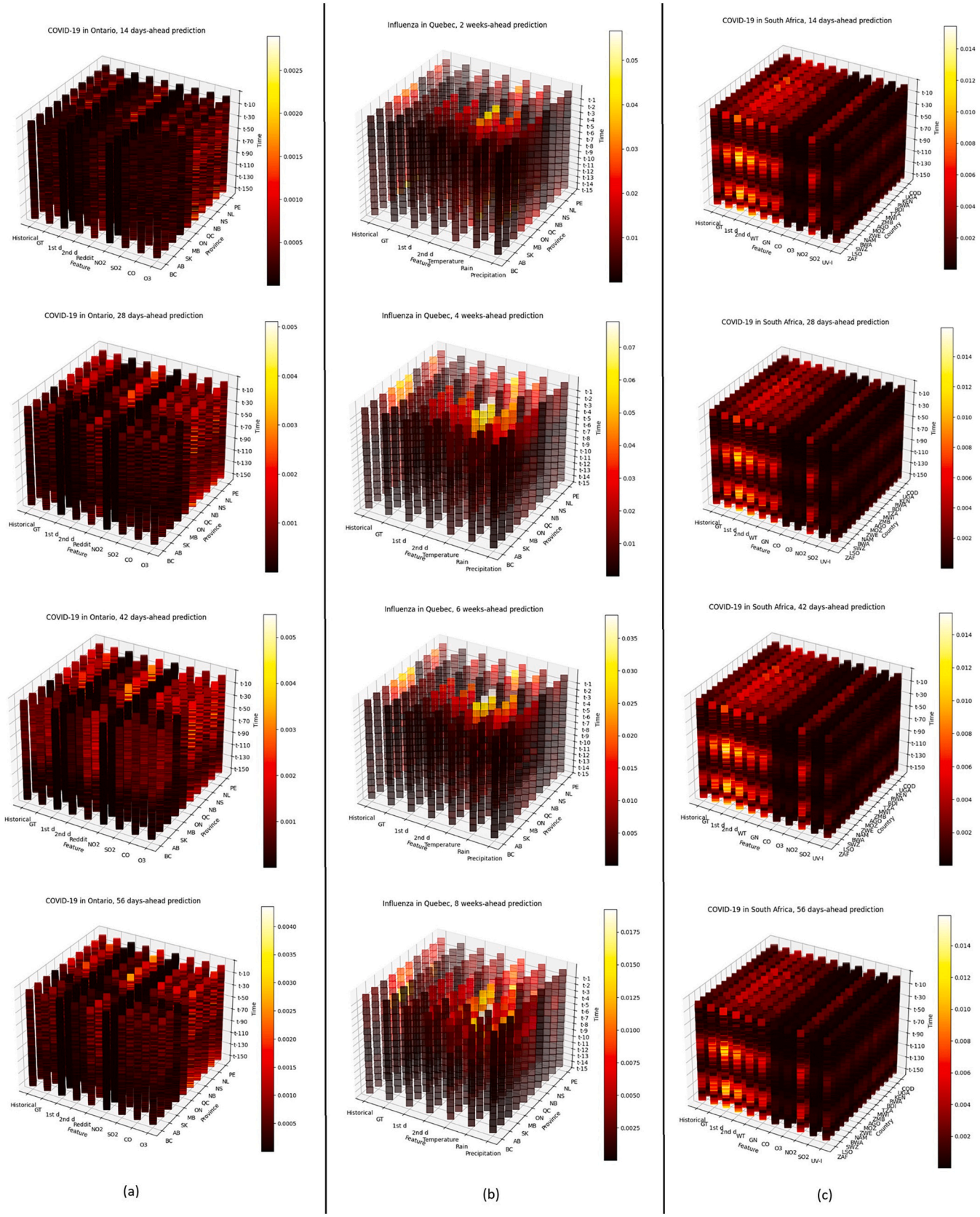



**Fig. 11.** Attention map for (a) COVID-19 forecast in Ontario, (b) influenza forecast in Quebec, and (c) COVID-19 forecast in South Africa.

## 5. Conclusion

Disease prediction models are fundamental tools for public health emergency management, as they enable preparedness, planning, and rapid response and recovery. The COVID-19 pandemic has revealed the core necessity of EWS for infectious diseases. Due to the grave impacts of the COVID-19 pandemic on healthcare services, and the global economy, many researchers have urged to build an EWS for COVID-19 waves. This work presents an efficient EWS based on GNNs for COVID-19 in Canadian provinces and southern African countries and influenza in different provinces of Canada. Four neural network layers, CNN, GCN, GRU, and a linear NN layer are placed sequentially to combine the input feature, apply spatial processing, perform temporal analysis, and enable T days-ahead prediction, respectively.

The results indicate that the best accuracy for COVID-19 is acquired in Ontario, which for 56 days-ahead prediction has a value of 1505.1116, 0.6871, and 0.8164 for RMSE, R2-score, and correlation, and for influenza is obtained in Quebec which for 8 weeks-ahead prediction sets at a value of 250.4763, 0.8729, and 0.953 for that, respectively. Moreover, the best results for southern African countries were obtained for South Africa which for RMSE, R2-score, and correlation has a value of 0.05914, 0.8145, and 0.9163 for 1 day-ahead, 0.06437, 0.7803, and 0.9098 for 14 days-ahead, 0.07425, 0.7077, and 0.8588 for 28 days-ahead, 0.06389, 0.7835, and 0.8984 for 42 days-ahead, and 0.078, 0.6775, and 0.8292 for 56 days-ahead prediction, respectively.

## CRediT authorship contribution statement

**Z. Movahedi Nia:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **L. Seyyed-Kalantari:** Writing – review & editing, Validation, Resources, Formal analysis. **M. Goitom:** Writing – review & editing, Validation, Resources, Formal analysis. **B. Mellado:** Writing – review & editing, Validation, Resources, Methodology, Investigation, Formal analysis, Conceptualization. **A. Ahmadi:** Writing – review & editing, Validation, Resources, Formal analysis. **A. Asgary:** Writing – review & editing, Validation, Supervision, Software, Resources, Funding acquisition, Formal analysis. **J. Orbinski:** Writing – review & editing, Supervision, Resources, Funding acquisition, Formal analysis. **J. Wu:** Writing – review & editing, Supervision, Resources, Funding acquisition, Formal analysis. **J.D. Kong:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

## Funding sources

This research is funded by Canada's International Development Research Centre (IDRC) (Grant No. 109981), IDRC and UK's Foreign, Commonwealth and Development Office (FCDO) (Grant No. 110554-001), SSHRC New Frontier in Research Fund- Exploratory (Grant No. NFRFE-2021-00879) and NSERC Discovery Grant (Grant No. RGPIN-2022-04559) and NSERC Discovery Launch Supplement (Grant No: DGECR-2022-00454).

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.artmed.2025.103076.

## Data availability

All our datasets and codes are available from https://github.com/Zahra1221/EWSF/tree/main [87].

## References

[1] Gray A, Sharara F. Global and regional sepsis and infectious syndrome mortality in 2019: a systematic analysis, the lancet. Global Health 2022;10(S2). https://doi.org/10.1016/S2214-109X(22)00131-0.
[2] World Health Organization, Children: improving survival and well-being. https://www.who.int/news-room/fact-sheets/detail/children-reducing-mortality. [Accessed 18 December 2023].
[3] Yuan K, Gong YM, Liu L. Prevalence of posttraumatic stress disorder after infectious disease pandemics in the twenty-first century, including COVID-19: a meta-analysis and systematic review. Mol Psychiatry 2021;26:4982–98. https://doi.org/10.1038/s41380-021-01036-x.
[4] Smith KM, Machalaba CC, Seifman R, Feferhotz Y, Karesh WB. Infectious disease and economics: the case for considering multi-sectoral impacts, Elsevier, one. Health 2019;7. https://doi.org/10.1016/j.onehlt.2018.100080.
[5] Stemple L, Karegeya P, Gruskin S. Human rights. Gender, and Infectious Disease: From HIV/AIDS to Ebola, Human Rights Quarterly, John Hopkins University Press 2016;28(4):993–1021. https://doi.org/10.1353/hrq.2016.0054.
[6] Boutayeb A, The impact of infectious diseases on the development of Africa, springer, handbook of disease burden and quality of life measures, New York, NY, 2010:1171–1188, PMCID: PMC7120372, doi: https://doi.org/10.1007/978-0-387-78665-0_66.
[7] Sahasranaman A, Jensen HJ. Poverty in the time of epidemic: a modelling perspective. PLoS ONE 2020;15(11). https://doi.org/10.1371/journal.pone.0242042.
[8] ProMED-mail, About ProMED-mail. https://web.archive.org/web/20150601102650/http://www.promedmail.org/aboutus/. [Accessed 19 December 2023].
[9] Madoff LC, Woodall JP. The internet and the global monitoring of emerging diseases: lessons from the first 10 years of ProMED-mail. Elsevier, Archives of Medical Research 2005;36(6):724–30. https://doi.org/10.1016/j.arcmed.2005.06.005.
[10] Mykhalovskiy E. The global public health intelligence network and early warning outbreak detection: a Canadian contribution to global public health. Can J Public Health 2006;97(1):42–4. https://doi.org/10.1007/BF03405213.
[11] Hope KG, Merritt TD, Durrheim DN, Massey PD, Kohlhagen JK, Todd KW, et al. Evaluating the utility of emergency department syndromic surveillance for a regional Pblic health service. Communicable Disease Intelligence Quarterly 2010;34(3):310–8.
[12] Caudle JM, Dijk A, Rolland E, Moore KM. Telehealth Ontario detection of gastrointestinal illness outbreaks. Springer, Canadian J of Public Health 2009;100(4):253–7. https://doi.org/10.1007/BF03403942.
[13] Schrell S, Ziemann A, Garcia-Castrillo Riesgo L, Rosenkotter N, Llorca J, Popa D, et al. Local implementation of a syndromic influenza surveillance system using emergency department data in Santander. Spain, Oxford Academic, J of Public Health 2013;35(3):397–403. https://doi.org/10.1093/pubmed/fdt043.
[14] Dembek Z F, Carley K, Siniscalachi A, Hadler J, Hospital admissions syndromic surveillance – Connecticut, September 2001-November 2003, centers for Disease Control & Prevention (CDC), MMWR, 2004;53:50–52, available from: https://www.jstor.org/stable/23315689.
[15] Hogan WR, Tsui F-C, Ivanov O, Gesteland PH, Grannis S, Overhage M, et al. Detection of pediatric respiratory and diarrheal outbreaks from sales of over-the-counter electrolyte products. JAMIA 2003;10(6):555–62. https://doi.org/10.1197/jamia.M1377.
[16] Davies GR, Finch RG. Sales of over-the-counter remedies as an early warning system for winter bed crises. Elsevier, Clinical Microbiology and Infection 2003;9(8):858–63. https://doi.org/10.1046/j.1469-0691.2003.00693.x.
[17] Yan W, Nie S, Xu B, Dong H, Palm L, Diwan VK. Establishing a web-based integrated surveillance system for early detection of infectious disease epidemic in rural China: a field experimental study. BMC Med Inform Decis Mak 2012;12(4). https://doi.org/10.1186/1472-6947-12-4.
[18] Johnson HA, Wagner MM, Hogan WR, Chapman W, Olszewski RT, Dowling J, et al. Analysis of web access logs for surveillance of influenza. IOS, Studies in Health Technology and Informatics 2004;107:1202–6. https://doi.org/10.3233/978-1-60750-949-3-1202.
[19] Brownstein JS, Freifeld CC, Madoff LC. Digital disease detection – harnessing the web for public health surveillance. N Engl J Med 2010;360(21):2153–7. https://doi.org/10.1056/NEJMp0900702.
[20] Milinovich GJ, Avril SMR, Clement ACA, Brownstein JS, Tong S, Hu W. Using internet search queries for infectious disease surveillance: screening diseases for suitability. BMC Infectious Disease 2014;14(690). https://doi.org/10.1186/s12879-014-0690-1.
[21] Santillana M, Ma WZ, Althouse BM, Ayers JW. What can digital disease detection learn from (an external revision to) Google flu trends? Elsevier, American Journal of Preventive Medicine 2014;47(3):341–7. https://doi.org/10.1016/j.amepre.2014.05.020.
[22] Polgreen PM, Chen Y, Pennock DM, Nelson FD, Weinstein RA. Using internet searches for influenza surveillance. Oxford Academic, Clinical Infectious Diseases 2008;47(11):1443–8. https://doi.org/10.1086/593098.

[23] Bragazzi NL, Alicino C, Trucchi C, Paganino C, Barberis I, Martini M, Sticchi L, Trinka E, Brigo F, Ansaldi F, Icardi G, Orsi A. Global reaction to the recent outbreak of Zika virus: insights from Big Data analysis. PLoS ONE 2017;12(9). https://doi.org/10.1371/journal.pone.0185263.

[24] Seo D-W, Shin S-Y. Methods using social media and search queries to predict infectious disease outbreaks. Healthc Inform Res 2017;23(4):343–8. https://doi.org/10.4258/hir.2017.23.4.343.

[25] World Health Organization, The top 10 causes of death. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death. [Accessed 22 December 2023].

[26] Corley C D, Cook D J, Mikler A R, Singh, K P, using web and social Media for Influenza Surveillance, springer, advances in computational biology, advances in experimental medicine and biology, 2010;680:59-564, New York, NY, doi: https://doi.org/10.1007/978-1-4419-5913-3_61.

[27] Ginsberg J, Mohebbi MH, Oatel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature 2009;457:1012–4. https://doi.org/10.1038/nature07634.

[28] Corley CD, Cook DJ, Mikler AR, Singh KP. Text and structural data Mining of Influenza Mentions in web and social media. Int J of Environ Res and Public Health 2010;7(2):596–615. https://doi.org/10.3390/ijerph7020596.

[29] Yih W K, Teates K S, Abrams A, Kleinman K, Kulldorff M, Pinner R, Harmon R, Wang S, Platt R, Telephone Triage Service Data for Detection of Influenza-Like Illness, PLoS ONE 4(4): e5260, doi: https://doi.org/10.1371/journal.pone.0005260.

[30] World Health Organization, WHO Coronavirus (COVID-19) Dashboard. https://covid19.who.int/. [Accessed 20 December 2023].

[31] Centers for Disease Control and Prevention, CDC Museum COVID-19 Timeline. https://www.cdc.gov/museum/timeline/covid19.html. [Accessed 20 December 2023].

[32] Movahedi Nia Z, Ahmadi A, Bragazzi NL, Woldegerima WA, Mellado B, Wu J, Orbinski J, Asgary A, Kong JD. A cross-country analysis of macroeconomic responses to COVID-19 pandemic using Twitter sentiments. PLoS ONE 2022;17(8). https://doi.org/10.1371/journal.pone.0272208.

[33] Movahedi Nia Z, Prescod C, Westin M, Patricia Perkins, Goitom M, Fevrier K, et al. A cross-sectional study to assess the impact of the COVID-19 pandemic on healthcare services and clinical admissions using statistical analysis and discovering hotpots in three regions of the greater Toronto area. BMJ Open 2024;14(3):1–12. doi: bmjopen-2023-082114.

[34] Marty R, Ramos-Maqueda M, Khan N, Reichert A. The evolution of the COVID-19 pandemic through the lens of google searches. Sci Rep 2023;13(19843). https://doi.org/10.1038/s41598-023-41675-4.

[35] Kodera S, Hikita K, Rashed EA, Hirata A. The effects of time window-averaged mobility on effective Reporduction number of COVID-19 viral variants in urban cities. J Urban Health 2023;100:29–39. https://doi.org/10.1007/s11524-022-00697-5.

[36] Nindera RD, Sari NP, Aryandono T. Validation: the use of Google trends as an alternative data source for COVID-19 surveillance in Indonesia, sage, Asia Pacific J of. Public Health 2020:. https://doi.org/10.1177/10105395209408.

[37] Ortiz-Martinez Y, Garcia-Robledo E, Vasquez-Castaneda DL, Bonilla-Aldana DK, Rodriguez-Marales AJ. Can Google® trends predict COVID-19 incidence and help preparedness? The situation in Colombia, Elsevier. Travel Med Infect Dis 2020;37. https://doi.org/10.1016/j.tmaid.2020.101703.

[38] Stevenson F, Hayasi K, Bragazzi NL, Kong JD, Asgary A, Lieberman B, et al. Development of an early alert system for an additional wave of COVID-19 cases using a recurrent neural network with Long short-term memory. Int J Environ Res Public Health 2021;18(14). https://doi.org/10.3390/ijerph18147376.

[39] Yu Y, Si X, Hu C, Zhang J. A review of recurrent neural networks: LSTM cells and network architecture. MIT Press Direct, Neural Computation 2019;31(7):1235–70. https://doi.org/10.1162/neco_a_01199.

[40] Google, See how your community moved differently due to COVID-19. https://www.google.com/covid19/mobility/. [Accessed 23 December 2023].

[41] Meta, Movement Range Maps. https://dataforgood.facebook.com/dfg/tools/movement-range-maps. [Accessed 23 December 2023].

[42] Khalis M, Toure A B, Badisy I E, Khomsi K, Najmi H, Bouaddi O, Marfak A, Al-Delaimy W K, Berraho M, Nejjari C, Relationship between meteorological and air quality Parmeters and COVID-19 in Casablanca region, Morocco, 2022;vol. 19(9) doi: https://doi.org/10.3390/ijerph19094989.

[43] Aragão DP, Oliveira EV, Bezerra AA, Satos D, Silva Junior A, Pereira IG, et al. Multivariate data driven prediction of COVID-19 dynamics: towards new results with temperature, humidity and air quality data, Elsevier. Environ Res 2022;204. https://doi.org/10.1016/j.envres.2021.112348.

[44] Vadyala SR, Betgeri SN, Sherer EA, Amritphale A. Prediction of the number of COVID-19 confirmed cases based on K-means-LSTMvol. 11. Array: Elsevier; 2021. https://doi.org/10.1016/j.array.2021.100085.

[45] Cho K, Merrieboer B, Bahdanau D, Bengio Y. On the properties of neural machine translation: encoder-decoder approaches. arXiv, [Preprint] 2014. https://doi.org/10.48550/arXiv.1409.1259.

[46] Dairi A, Harrou F, Zeroual A, Hittawe MM, Sun Y. Comperative study of machine learning methods for COVID-19 transmission forecasting. Elsevier, J of Biomedical Informatics 2021;118. https://doi.org/10.1016/j.jbi.2021.103791.

[47] Shahid F, Zameer A, Muneeb M. Predictions for COVID-19 with deep learning models of LSTM, GRU and bi-LSTM, Elsevier. Chaos, Solitons Fractals 2020;140. https://doi.org/10.1016/j.chaos.2020.110212.

[48] Ayoobi N, Sharifrazi D, Alizadehsani R, Shoeibi A, Gorriz JM, Moosaei H, et al. Time series forecasting of new cases and new deaths rate for COVID-19 using deep learning methods, Elsevier. Results in Physics 2021:27. https://doi.org/10.1016/j.rinp.2021.104495.

[49] Batool H, Tian L. Correlation determination between COVID-19 and weather parameters using time series forecasting: a case study in Pakistan, Hindawi. Math Probl Eng 2021;2021(9953283). https://doi.org/10.1155/2021/9953283.

[50] Khennou F, Akhloufi MA. Forecasting COVID-19 spreading in Canada using deep learning, MedRXiv. Prepring 2021. https://doi.org/10.1101/2021.05.01.21256447.

[51] Hartono P. Similarity maps and pairwise predictions for transmission dynamics of COVId-19 with neural networks, Elsevier. Informatics in Medicine Unlocied 2020;20. https://doi.org/10.1016/j.imu.2020.100386.

[52] Ward T, Johnsen A, Ng S, Chollet F. Forecasting SARS-COV-2 transmission and clinical risk at small spatial scales by the application of machine learning architectures to syndromic surveillance data, nature machine intelligence 2022;4:814–27. https://doi.org/10.1038/s42256-022-00538-9.

[53] Scarselli F, Gori M, Tsoi AC, Hagenbuchner M, Monfardini G. The graph neural network model, IEEE, transactions on. Neural Netw 2009;20(1):61–80. https://doi.org/10.1109/TNN.2008.2005605.

[54] Fritz C, Dorigatti E, Rügamer D. Combining graph neural networks and spatio-temporal disease models to improve the prediction of weekly COVID-19 cases in Germany. Sci Rep 2022;12(3930). https://doi.org/10.1038/s41598-022-07757-5.

[55] Gatta V, Moscato V, Postiglione M, Sperli G. An epidemiological neural network exploiting dynamic graph structured data applied to the COVID-19 outbreak. IEEE Trans Big Data 2021;7(1):45–55. https://doi.org/10.1109/TBDATA.2020.3032755.

[56] Skianis K, Nikolentzos G, Gallix B, Thiebaut R, Exarchakis G. Predicting COVID-19 positivity and hospitalization with multi-scale graph neural networks. Sci Rep 2023;13(5235). https://doi.org/10.1038/s41598-023-31222-6.

[57] Davahli MR, Fiok K, Karwowski W, Aljuaid AM, Taiar R. Predicting the dynamics of the COVID-19 pandemic in the United States using graph theory-based neural networks, J. Environ. Res. Public Health 2021;18(7). https://doi.org/10.3390/ijerph18073834.

[58] Panagopoulos G, Nikolentzos G, Vazirgiannis M. Transfer graph neural networks for pandemic forecasting. arXiv, [Preprint] 2009. https://doi.org/10.48550/arXiv.2009.08388.

[59] Davidson MW, Haim DA, Radin JM. Using networks to combine “big data” and traditional surveillance to improve influenza prediction. Sci Rep 2015;5(8154). https://doi.org/10.1038/srep08154.

[60] Kapoor A, Ben X, Liu L, Perozzi B, Barnes M, Blais M, O'Banion S. Examining COVID-19 forecasting using spatio-temporal graph neural networks. arXiv, [Preprint] 2007. https://doi.org/10.48550/arXiv.2007.03113.

[61] Gao J, Sharma R, Qian C, Glass LM, Spaeder J, Robmerg J, et al. STAN: spatio-temporal attention network for pandemic prediction using real-world evidence. AMIA 2021;28(4). https://doi.org/10.1093/jamia/ocaa322.

[62] Deng S, Wang S, Rangwala H, Wang L, Ning Y. Graph message passing with cross-location attentions for lon-term ILI prediction. arXiv 2019. https://doi.org/10.48550/arXiv.1912.10202.

[63] PyPI. GoogleNews. https://pypi.org/project/GoogleNews/. [Accessed 11 January 2024].

[64] COVID19Tracker, Overview. https://api.covid19tracker.ca/docs/1.0/overview. [Accessed 26 June 2024].

[65] Jafari M, Ansari-Pour N. Why, When and how to adjust your P values?vol. 20(4): PP. National Library of Medicine; 2019. p. 604–7. https://doi.org/10.22074/cellj.2019.5992.

[66] PyPI, pytrends. https://pypi.org/project/pytrends/. [Accessed 11 January 2024].

[67] PullPush, PullPush Reddit API Documentation. https://pullpush.io. [Accessed 26 June 2024].

[68] Open-Meteo, Historical Weather API. https://open-meteo.com/en/docs/historical-weather-api. [Accessed 27 June 2024].

[69] Catalog Earth Engine Data. Sentinel-5P. https://developers.google.com/earth-engine/datasets/catalog/sentinel-5p. [Accessed 11 January 2024].

[70] Hu T, Wang S, She B, Zhang M, Huang X, Cui Y, et al. Hu Yaxin, Fu X, Wang X, Wang P, Zhu X, Bao S, guan W, Li Z, human mobility data in the COVID-19 pandemic: characteristics, applications, and challenges, Taylor & Francis. Int J of Digital Earth 2021;14(9):1126–47. https://doi.org/10.1080/17538947.2021.1952324.

[71] World Health Organization, Global Influenza Programme. https://www.who.int/tools/flunet. [Accessed 11 February 2024].

[72] Government of Canada, Respiratory Virus Detections in Canada – Archives. https://www.canada.ca/en/public-health/services/surveillance/respiratory-virus-detections-canada.html. [Accessed 26 June 2024].

[73] Zippenfenig P, Open-Meteo.com Weather API. Zenodo [Computer software] 2023. https://doi.org/10.5281/ZENODO.7970649.

[74] Hersbach H, Bell B, Berrisford P, Biavati G, Horányi A, Muñoz Sabater J, et al. ERA5 hourly data on single levels from 1940 to present. ECMWF, [Data set] 2023. https://doi.org/10.24381/cds.adbb2d47.

[75] Muñoz Sabater J. ERA5-Land hourly data from 2001 to present. ECMWF, [Data set] 2019. https://doi.org/10.24381/CDS.E2161BAC.

[76] Schimanke S, Ridal M, Le Moigne P, Berggren L, Undén P, Randriamampianina R, et al. CERRA sub-daily regional reanalysis data for Europe on single levels from 1984 to present. ECMWF, [Data set] 2021. https://doi.org/10.24381/CDS.622A565A.

[77] World Health Organization, WHO COVID dashboard. https://data.who.int/dashboards/covid19/data?n=c. [Accessed 11 January 2024].

[78] Wikimedia, Wikimedia REST API. https://wikimedia.org/api/rest_v1/#/Pageviews%20data/get_metrics_pageviews_per_article_project__access__agent__article__granularity__start__end . [Accessed 11 January 2024].
[79] Zhou H, Zhang S, Peng J, Zhang S, Li J, Xiong H, et al. Informer: beyond efficient transfomer for long sequence time-series forecasting. ArXiv, [Preprint] 2021. https://doi.org/10.48550/arXiv.2012.07436.
[80] Wu H, Xu J, Wang J, Long M. Autoformer: decomposition transformers with auto-correlation for long-term series forecasting. ArXiv, [Preprint] 2022. https://doi.org/10.48550/arXiv.2106.13008.
[81] Altmann A, Tolosi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. Wiley, Bioinformatics 2010;26(10):1340–7. https://doi.org/10.1093/bioinformatics/btq134.
[82] Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classification models and saliency maps. ArXiv, [Preprint] 2014. https://doi.org/10.48550/arXiv.1312.6034.
[83] Karpathy A, Johnson J, Fei-Fei L. Visualizing and understanding recurrent networks. ArXiv, [Preprint] 2015. https://doi.org/10.48550/arXiv.1506.02078.
[84] El-kenawy EM, Khodadadi N, Mirjalili S, Abdelhamid AA, Eid MM, Ibrahim A. Greylag Goose Optimization: nature-inspired optimization algorithm. In: Expert systems with applications. 238. Elsevier; 2024. https://doi.org/10.1016/j.eswa.2023.122147. 122147. (E).
[85] El-kenawy EM, Rizk FH, Zaki AM, Elshabrawy M, Ibrahim A, Abdelhamid AA, et al. iHow optimization algorithm: a human-inspired metaheuristic approach for complex problem solving and feature selection. Journal of Artificial Intelligence in Engineering Practice 2024;1(2):36–53. https://doi.org/10.21608/jaiep.2024.386694.
[86] El-Kenawy EM, Rizk FH, Zaki AM, Mohamed ME, Ibrahim A, Abdelhamid AA, et al. Football optimization algorithm (FbOA): a novel metaheuristic inspired by team strategy dynamics. Journal of Artificial Intelligence and Metaheuristics 2024;8(1):21–38. https://doi.org/10.54216/JAIM.080103.
[87] EWSF. A deep-learning model for building an early warning system for respiratory infections. GitHub 2025. https://github.com/Zahra1221/EWSF/tree/main. [Accessed 24 January 2025].