National College of Ireland

# A study on Influenza Reports: The Impact of Various Influencing Factors and a Predictive Modeling Approach to Forecasting Flu Cases in European Countries

MSc Research Project
Data Analytics

Vasanthi Badami
Student ID: 20186690

School of Computing
National College of Ireland

Supervisor: Prof. Aaloka Anant

**National College of Ireland**
**Project Submission Sheet**
**School of Computing**

National College of Ireland

| | |
|---|---|
| **Student Name:** | Vasanthi Badami |
| **Student ID:** | 20186690 |
| **Programme:** | Data Analytics |
| **Year:** | 2021 |
| **Module:** | MSc Research Project |
| **Supervisor:** | Prof. Aaloka Anant |
| **Submission Due Date:** | 31/01/2022 |
| **Project Title:** | A study on Influenza Reports: The Impact of Various Influencing Factors and a Predictive Modeling Approach to Forecasting Flu Cases in European Countries |
| **Word Count:** | 8024 |
| **Page Count:** | 21 |

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

**ALL** internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

| | |
|---|---|
| **Signature:** | |
| **Date:** | 31st January 2022 |

## PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

| | |
|---|---|
| Attach a completed copy of this sheet to each project (including multiple copies). | ☐ |
| **Attach a Moodle submission receipt of the online project submission**, to each project (including multiple copies). | ☐ |
| **You must ensure that you retain a HARD COPY of the project**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. | ☐ |

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

| **Office Use Only** | |
|---|---|
| Signature: | |
| Date: | |
| Penalty Applied (if applicable): | |

# A study on Influenza Reports: The Impact of Various Influencing Factors and a Predictive Modeling Approach to Forecasting Flu Cases in European Countries

Vasanthi Badami
20186690

**Abstract**

Influenza symptoms are identical to those of COVID 19. As a result, distinguishing between the two disorders might be difficult at times. Similar infections such as influenza should not be ignored in circumstances such as the COVID 19 pandemic. Therefore this study aims to forecast influenza cases in the most regularly visited tourist countries because they are already on high alert to prevent the virus from spreading further. Assisting these countries in their preparation for influenza vaccination campaigns could help to reduce the pandemic's complexity. Other contributing factors such as weather and the population of different age groups in the respective countries were also taken into account in this study. However, they did not reveal a strong link with the number of influenza cases reported. Conversely, youngsters aged 0 to 14 and those aged 100 and above exhibited an improved correlation. As a result, all of these characteristics were taken into account when forecasting influenza cases. To move forward with the model implementation, stand-alone models such as ARIMA, SVR, RF, and attention LSTM are used, followed by an ARIMA-LSTM hybrid model to see how the hybrid model performs on this problem. The attention LSTM model performed exceptionally well, with a $RMSE$ value of 0.08, however, the RF model had a second-lowest $RMSE$ score (84.991). Other models did not appear to offer much optimism in terms of achieving the best fit model for predictions. As a result of this research, it appears that attention LSTM will be the best fit model.

## 1 Introduction

Disease monitoring systems that are both efficient and timely have the potential to assist public health professionals in developing actions to minimize the effects of epidemics. Influenza is one of the world's most fatal diseases, harming public health. Because influenza and COVID-19 have similar symptoms, it can be difficult to distinguish between the two, making it difficult for hospitalized patients to receive timely treatment because health staff must wait for the results. A study highlights the importance of keeping an eye on COVID-19 patients against co-infecting viral infections. Co-infections discovered in COVID-19 cases highlight the importance of flu shot and call for expanded vaccination rates to reduce identified risk factors and hospitalizations ( Alosaimi et al. (2021)). As

a result, there was a strong desire to reduce the impact of influenza reports. Forecasting influenza cases could aid health workers in making early predictions and vaccinating individuals against the virus. This study took into account a variety of other affecting elements, such as weather, demographic mobility, and population density, in addition to influenza reporting data.

It is obvious that the rate of influenza transmission fluctuates depending on the season, and numerous studies have investigated this fact. Influenza reports are highly linked to environmental conditions such as high temperature, low temperature, and rainfall. As a consequence of viral behavior being impacted by the environment, influenza can spread widely (Carter-Templeton et al. (2021)). Based on these analyses, health institutions will have a better understanding of the peak season for viral activity and a more relaxed time to plan vaccine campaigns and prepare for potential consequences. Research has demonstrated that lower temperatures are directly linked to influenza virus transmission. In addition, high temperatures have been proven to lessen the likelihood of virus infection. Months with low temperatures had more occurrences of influenza cases, whereas months with high temperatures had fewer or no infection reports, according to the data collected for this research. Later in this investigation, a correlation heatmap will be used to verify this finding.

Some research suggests that human movement, from short-distance to long-distance commuting, has an impact on how infectious disease outbreaks propagate. In terms of international tourism, Europe is the undisputed leader. Every year, more than 700 million inbound tourists visit the region as per the article statistics and facts, published in statista.com[1]. Europe has numerous tourist attractions that annually attract several visitors from all over the world. Many people also travel to the most well-known European countries in search of better living and career prospects. Such facts provide a large amount of income to the country, but as with any pro, there is a con. Along with great revenue, such countries may also welcome contagious diseases. To evaluate this fact, the study investigated the most popular tourist destinations in Europe (Croatia, Czechia, Finland, France, Germany, Greece, Italy, Netherlands, Spain, Switzerland, and the UK). Even though the results are not as expected, these attributes do indicate some correlation with the frequency of influenza cases. Many researchers have undertaken studies to establish such well-managed health systems in the past, but studying data during the COVID-19 pandemic adds more value because we may gain unique insights from the data since people's mobility is restricted to nearly zero during lock downs.

It is widely known that children and the elderly have weaker respiratory hygiene. Therefore, Pneumonia and influenza cause the most morbidity in children and the elderly Sebastian et al. (2008). However, the methods of transmission, as well as age-related variations in infection severity, are unknown. To conform with this theory, the study looked at various age groups to see age-specific sensitivity to the influenza spread. The dataset comprised age groups such as 0 to 14, 15 to 29, 30 to 44, 45 to 59, 60 plus, and 100 plus. To check both genders' immune performance, the total populations of females and males are considered separately for the research. However, both female and male populations showed the same level of correlation with influenza cases.

[1]Travel and tourism in Europe `https://www.statista.com/topics/3848/travel-and-tourism-in-europe/`

**Research Objective:** Analyze many such factors that influence influenza transmission and forecast future values based on the observations.

Taking into account all these elements, this research focuses on estimating influenza cases based on influencing factors such as weather, tourist arrivals, and various age groups in a country. The accuracy of machine learning models and how they interact with other contributing elements were demonstrated in this study. This research also sought to answer concerns such as how temperature change affected viral dissemination and how demographic shifts affected influenza cases based on recorded reports from the same period. Taking into account various age groups also aided this study's understanding of how viral dissemination affected different age groups. This research also includes knowledge of machine learning models, which were immensely beneficial in forecasting.

Answering these questions will assist in preventing further disease spread. It is not just the government's (primary stakeholder) job to maintain public health during difficult times like COVID-19, but it is also everyone's responsibility because public is the secondary stakeholder of this research. This study's findings not only help health agencies but also raise public awareness about precautionary measures individuals can take to protect their families and communities. To do this, all communities can help to reduce the pandemic's complexity by finding solutions to problems such as the spread of other catastrophic co-infections such as Influenza-Like Illness (ILI) or influenza. The findings on traveler arrivals and their link to influenza cases should help the public understand the risks they face after traveling. Furthermore, research on different age groups can inform the public about who they should be concerned about during epidemics and pandemics. Tourist locations in Europe were chosen to conduct these experiments because they receive a large number of visitors from all over the world. In addition, some of these countries were already sensitive to climate change, which harmed public health, particularly among the younger and elderly age groups. Some of the current changes in human disease patterns are associated with the onset of climate change on a global scale (Danzon (2000)). This research is conducted on the total incidence of influenza cases recorded and published to regional WHO databases (World Health Organization). Weather data from the Climate Change Knowledge Portal and Power Access Viewer, tourist arrival statistics from CEIC, and world population data from the UN Department of Economic and Social Affairs were used to support this study. Rather than looking at the data of each region of a country, this study will concentrate on data at the country level. The climatological collected data will not be targeted on any unfavorable weather incidents that occurred during this time frame. The study concentrates only on tourist arrival data to see how population movement influences the rise in influenza occurrences.

# 2 Related Work

Many studies have been undertaken to test the idea of influenza transmission, and certain other researchers have concentrated on forecasting influenza cases in various parts of the world. However, there is still a desire to concentrate on accurate estimates of influenza reports. As a result, the researchers took into account influencing aspects in their research to develop a highly accurate algorithm and forecast high accuracy results. This study focuses on three major contributing elements by considering and drawing inspiration from a variety of similar studies. Before beginning this study, the following research materials were gathered and carefully observed to construct better prediction models. This literature review will focus on studies that looked at various machine learning models for highly accurate predictions, as well as studies that looked at the elements that influence influenza spread. This literature review will concentrate on three main study categories:

- Research on the factors that influence the spread of influenza.
- Studies on the machine learning models for evaluating and forecasting influenza and influenza-like illness.
- Studies on forecasting machine learning models

## 2.1 Factors that influence influenza transmission are being studied in the following publications

The National Health Service (NHS) in the United Kingdom (UK) has a difficult time throughout the winter because of the cold weather as well as the increased risk of developing respiratory diseases, particularly influenza Hussain et al. (2005). The author of this article uses an appropriate statistical time series technique to model and analyze weekly hospitalizations in the West Midlands of the United Kingdom from week 15 of 1990 to week 14 of 1999. This study included attributes such as minimum temperature, general practitioner consultants, and total hospital admissions. Furthermore, there is a strong inter-dependency between the hospital admission summer and winter residuals, which could be regarded as hidden patterns throughout the recent ten years interval of time.

The study by Birrell et al. (2019), constituted the first attempt in the UK to create consistent short-term predictions of seasonal influenza epidemic to guide wintertime health service preparation. Precise short-term forecasts of the number of patients in primary and secondary care, especially at the local scale, can help health facility planners make the most use of limited capacity. Other prerequisites suggested by this study to establish such an effective forecasting system include thorough serological data, the inclusion of vaccine coverage statistics, and adding greater flexibility to models to allow for less precise projections because of the occurrence of certain biases. The research paper Poirier et al. (2021) presents a machine-learning modeling approach that generates real-time influenza activity estimates and short-term predictions for France's twelve continental zones. This study uses a variety of data sources, including Google search activity, actual and local meteorological data, flu-related Twitter microblogs, e-healthcare data, and past illness occurrence synchronicities throughout regions. The findings suggest that all data sources help improve influenza surveillance, and algorithms include all these data sources

to produce precise and timely forecasts. As a consequence, considering other factors that influence influenza infections could bring more value to this study.

Over the last few decades, scientists have spent a lot of time researching how people's movement affects disease transmission. Unfortunately, the mechanism by which human mobility influences disease propagation remains a mystery. The influence of human movement on contagious disease propagation was investigated in the work Changruenngam et al. (2020) by constructing the individual-based SEIR (Susceptible Exposed Infectious Removed) model, which included a model of human mobility. The researchers discovered that their method will provide a geo-temporal epidemic spreading trend that is not represented by a typical homogenous epidemiological model. The infection has the potential to expand to densely populated metropolitan regions before spreading to more rural regions, and then to all nearby places. The moment of the infection's first entrance, which is related to the landscape of mobility determined by relative appeal, can capture the infection's diversified dissemination. These studies can help us better comprehend and help to estimate how the disease will spread.

Human movement is a major factor in the transmission of contagious diseases. Existing data, on the other hand, is limited in terms of coverage, availability, timeliness, and granularity. The authors of the study Venkatramanan et al. (2021) examine the utility of a machine-learned AMM (Anonymized Mobility Map) based on millions of smartphones in forecasting disease outbreaks. This study incorporated an anonymized mobility map into a metapopulation framework to forecast influenza in the United States and Australia retrospectively. The advantage of considering such meta population data showed the improved model's capacity to forecast disease propagation even across state borders is demonstrated in this study. This research helped to provide timely infectious epidemic forecasting on a worldwide scale employing human mobility statistics, broadening their usefulness in contagious disease research. In the time and spatial spread of infectious illnesses, human movement is critical.

It is believed that particular age groups are more susceptible to influenza infections, which could raise the risk of disease spread. As a result, it is critical to do research on different age groups and to protect them from disease exposures. To verify this fact the study Sebastian et al. (2008) worked to determine which age groups are most likely to be affected by seasonal influenza and who might have the biggest indirect impact on the population. Using publicly available health care and vital statistics sources, the study was able to determine the peak pneumonia and influenza health-care stress by age and analyse possible indirect consequences at various population levels. The findings show that pneumonia and influenza hospitalization and mortality adversely impact the very old and the very young. This study revealed that young children are more likely to visit the hospital. The young and elderly had the highest number of hospitalizations for pneumonia and influenza disease, as well as the highest number of recorded mortality rates. At its peak, the strain on young school-aged children and working adults was moderate. The lowest peak rates were seen in older school children. Flu activity showed modest changes in the age-specific time frame of community outbreaks in the investigation by Stockmann et al. (2014). The number of medically-attended visits for laboratory-confirmed influenza infection peaked among older children aged 12 to 18 years old for twelve consecutive influenza seasons in Utah, according to this study. As a result of this research, it is

recommended that older children be vaccinated against influenza transmission to other age groups. Along with children elderly individuals aged 65 and up are more susceptible to influenza viruses because their immune systems deteriorate with age. Because of their weakened immune systems, they are more susceptible to infection, and thus a majority of influenza-related loss of life occurs in the elderly Monto et al. (2009).

As a result, we will be able to advance immunization initiatives by focusing on susceptible groups. Such immunization initiatives should pay special attention to vulnerable age groups. To do so, influenza forecasting among various age groups is important, as it will better assist health professionals in combating the transmission rate.

## 2.2 Studies on standalone and Hybrid models

In the research, Volkova et al. (2017), the author created and analyzed the predictive capacity of neural network architectures constructed on LSTMs (Long Short Term Memory) units effective for nowcasting and forecasting ILI patterns in the influenza seasons from 2011 to 2014. Authors combined information from people who post on social media, such as subjects, stylistic tendencies, embeddings, and communication behavior utilizing mentions and hashtags, to create these models. The study then uses a variety of evaluation measures to statistically assess the predictive capacity of various social media data and compare the effectiveness of state of the art regression modeling techniques with neural networks.

To predict location-specific ILI trends, scientists used social media signals and neural network models. LSTM methods outperform formerly used machine learning methods like SVM and ADABoost, according to studies. This research has also shown that gathering numerous attributes for the investigation will aid machine learning models in performing effectively. For multi-step influenza epidemic forecasting challenges, the study Kara (2021) presented a hybrid approach that integrates an LSTM neural network and a genetic algorithm (GA). In forecasting influenza cases, the LSTM approach is utilized to alleviate complexity and nonlinearity concerns. The genetic algorithm is being used to acquire the epoch capacity of the network to improve the effectiveness and productivity of the neural network. According to the results of the studies, the given hybrid model beat several well-known machine learning techniques, a fully connected neural network, and a statistical model for peak times. In the future, researchers want to use convolutional LSTM neural networks to study how multivariable time-series datasets influence influenza forecast accuracy.

In the research Norrulashikin et al. (2021), a hybrid ARIMA-SVR technique is used to forecast monthly influenza occurrences in Malaysia. By using Box-Jenkins approach and the Support Vector Regression model, the ARIMA model is employed to capture linear and nonlinear variables in monthly influenza cases. The study's findings revealed that three models, ARIMA, ARIMA-SVR, and SVR, were effective in forecasting monthly instances. The investigation then discovered that these three models outperformed the Naive model. SVR demonstrated great accuracy when compared to ARIMA SVR and ARIMA models, according to the authors, and they suggested SVR as the best fit model for influenza case predicting. Future studies could include other machine learning approaches for hybrid models, such as the installation of a Genetic Algorithm for model improvement and the use of alternative lags for the Super Vector Regression model com-

ponent, and evaluate its improvements over present methods, according to the author. To estimate the trends of reported Influenza-like illnesses (ILI) in Cameroon from 2012 to 2018, researchers in the study Nsoesie et al. (2021) used and analyzed a variety of machine learning and strong statistical models, such as support vector machines regression, random forest (RF) regression, ARIMA, multivariable linear regression models. Most of the algorithms had statistically identical R-squared and Root Mean Squared Error (RMSE), however, SVM and RF had the greatest average R-squared for forecasting ILI for every 100,000 people at the national level. Future research, according to the author, will focus on building ensemble techniques for real-time predictions that involve the majority of various algorithms. The study by researchers Novaes de Amorim et al. (2021) created a stacked ensemble method that aggregates the forecasts from multiple competing approaches in the present ILI forecasting frontier. Using real-time hospital datasets on weekly ILI visitor numbers from the 2012 to 2018 influenza season at the Alberta Children's Hospital in Canada, this study evaluated the precision and reliability of the model with one to four weeks ahead forecast goals. According to the findings, the stacked ensemble's forecasting accuracy surpasses that of the standalone models across all forecast objectives.

## 2.3 Studies on forecasting machine learning models

The ARMA (Autoregressive Moving Average) time series model is a traditional stochastic model that can be found in a variety of domains, including foreign exchange, flue forecasting, biological science, and rainfall forecasting. Identifying the model orders, determination of model coefficients, and predictions are the three elements of using the ARMA model for time series analysis. The study used convolutional neural networks to solve the challenge of model selection in ARMA time series forecasting, where the system trained the networks using generated time series containing known ground truth rather than actual data. When compared to the efficacy of these networks to that of classic approaches, such as the Bayesian and Akaike information Criteria, a study by Tang and Rllin (2018) observed that trained networks outperform these approaches in terms of speed and accuracy by the degree of magnitude.

The trending forecasting models are the ARIMA models. ARIMA exhibited a considerable favorable response in time series forecasting in a variety of applications. ARIMA has recently been employed in several studies to forecast COVID-19 cases, as well as a variety of other stand-alone models to achieve excellent accuracy. The article Perone (2020) looked at numerous time series forecasting approaches for predicting the spread of COVID-19 in Italy during the second wave of the pandemic. The ARIMA, ETS (Exponential Smoothing State-space model), NNAR (Neural Network AutoRegression model), and ETS-NNAR, ARIMA-ETS, ARIMA-ETS-NNAR, and ARIMA-NNAR, models were used in the study. The study claims that apart from ARIMA-ETS, hybrid models outperform single models in terms of capturing linear and nonlinear epidemic trends. ARIMA is the most successful single model, and ARIMA-NNAR outperformed all other hybrid models. This research also revealed that when models are combined to handle both linear and nonlinear patterns, they perform well. Finally, in the short to mid-term, the combining ETS, ARIMA, and NNAR have demonstrated to be an adequately accurate hybrid model. Another study Temr and Yldz (2021) created a hybrid model employing ARIMA and LSTM to produce a monthly sales quantity budget based on an organization's

historical revenue data. Forecasting time-series of sales using a linear ARIMA model, a nonlinear LSTM model, and a HYBRID (LSTM-ARIMA) model created to increase the performance of the system over a single model. As a consequence of the research, performance evaluation metrics from each of the application's methods were compared, and a monthly overall sales budget for 2017 was created. When the $MSE$ and $MAPE$ of each of these methods were compared, the study found that hybrid models had a lower error and other models also produced realistic results using historical data. In some circumstances, we will be unable to obtain sufficient data. Models do not perform well when there is inadequate data, so we need a solution to this problem to attain prediction efficiency with limited data. Insufficient data, according to the study Ma (2021), can lead to model inefficiency. The study concentrates on the LSTM model believing that the LSTM model's prediction error would increase substantially with insufficient data, and it will be prone to large bias for the mid-term and long-term predictions. To address this issue, the study suggested an LSTM-Markov model, that employs the Markov model to minimize the LSTM model's prediction error. The training losses of the LSTM were calculated, and the probability transition matrix for the Markov model was created using the errors. Finally, the prediction results were produced by integrating the LSTM model output data with the Markov Model prediction errors. The results demonstrate that employing the LSTM Markov model reduced prediction error. As a result, using such stochastic models is recommended to improve model efficiency and, in turn, reduce error.

# 3 Methodology

The Cross-Industry Standard Process for Data Mining (CRISP-DM) approach was employed in this study to carry out planned tasks. The six stages of the data science life cycle will be presented in this methodology. This methodology, as shown in figure 1, represents six stages of project planning that enhance machine learning projects.

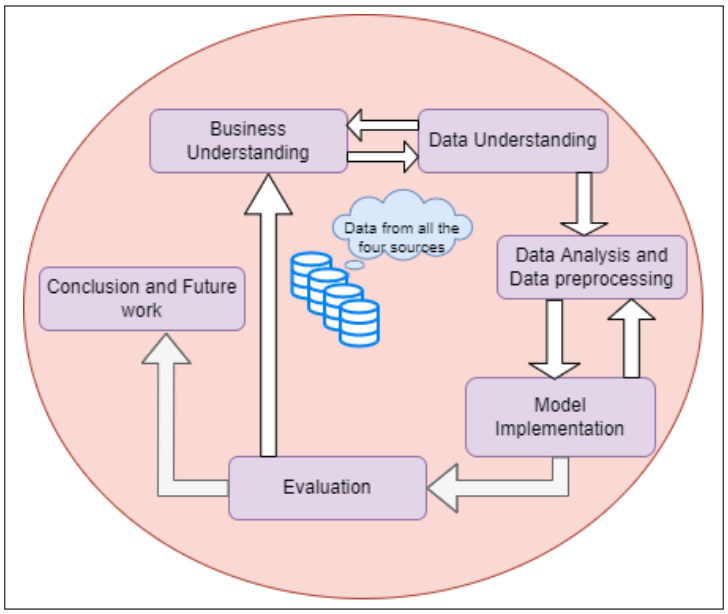



Figure 1: Customized CRISP - DM

## 3.1 Business understanding

Tourism is a major source of revenue for many European countries. As a result, it's no surprise that such countries draw an increasing number of tourists each year with their appealing tourism packages. As everyone knows, people's mobility can cause major problems for the hosting country's public health, because such activities could also facilitate the spread of infectious diseases. As a consequence, this research is being carried out to assess the situation by focusing on the contagious disease influenza. As a result, health officials are the key stakeholders in this study, with the general public serving as secondary stakeholders. This study aims at forecasting influenza cases at least for the next one weeks. The main aim of this study is to help health officials plan their workloads based on expected case counts.

## 3.2 Data understanding

In every data mining and machine learning investigation, data is critical. Researchers need a realistic and authentic dataset to do legitimate research. Before beginning any research, researchers should have a thorough understanding of the data. Moreover, collecting and analyzing the data is a pivotal time in any research. Data for this study was gathered from four different sources. This information was gathered over a two-decade period, from 2000 to 2021. Flunet[2] dataset from the WHO regional databases was used to construct the total number of influenza reports for the selected regions. WHO collects Flunet data from GISRS (Global Influenza Surveillance and Response System) and from the laboratories which are actively participating with GISRS. Figure 7 shows the map of various transmission zones, these Influenza Propagation Zones are specific geographical grouping of countries, territories, or areas having influenza transmission trends that are similar. Flunet data is updated weekly and provided on a country-by-country basis. The GISRS data in this collection has been verified in the laboratory.

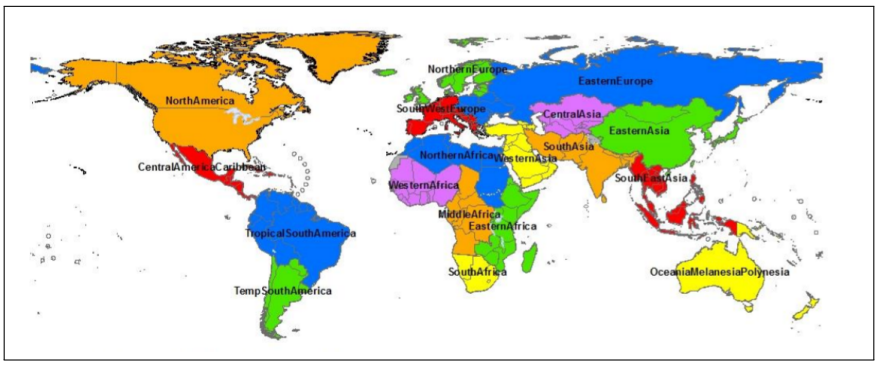



Figure 2: Influenza transmission zones

This study contains weather, tourist arrival, and demographic data to support the conspiracy of contributing elements for the influenza transmission rate. Weather data is collected from two sources. Yearly weather Climate Change Knowledge Portal (CCKP)[3] provides past and future data. We can get the mean, minimum, and maximum temperature data from this portal. Aggregated yearly or monthly time series are also available.

[2]WHO FluNet Dataset `https://www.who.int/tools/flunet`

[3]Climate Change Knowledge Portal `https://climateknowledgeportal.worldbank.org/download-data`

For this study, we gathered monthly aggregated data from 1991 to 2020, which will give us temperature variation mean values for all 12 months of the year. The mean temperature of the months helps to distinguish between the coldest and warmest months of the year, making it easier to track influenza spread during these months. Because the weather data did not indicate a strong link with the weekly influenza reports, the study retrieved more weather data to test the hypothesis discussed in the literature review. Second weather data collected from Data Access Viewer NASA POWER[4]. Data Access Viewer is a tool that allows us to view solar and meteorological data from NASA studies to help you with various research. This portal provides daily meteorological data based on country latitude longitudinal details. Temperature, humidity, precipitation, wind, and pressure are among the weather parameters provided in this dataset. The only temperature at a 2-meter range is evaluated in this study to check for disease propagation at a 2-meter range. Daily data was turned into weekly data to see whether there was any association, but this study still found no good correlation with meteorological data.

Data on tourist arrivals is gathered from the CEIC data global database[5]. Since 1992, this database has always been the gold standard for understanding advanced and emerging markets all across the world. International tourism statistics available in this portal put the spotlight on various tourism aspects. For this investigation, a considerable proportion of tourist arrival data on a monthly and yearly average was collected. In addition, the data did not demonstrate the anticipated link with influenza data. As a result, the population data which is also an important factor that could influence influenza transmission rate has been included in this investigation.

Population data is collected from United Nations (UN), Department of Economic and Social Affairs[6]. This website encompasses diverse types of population data, as well as different age groups and genders. The population data of various age groups and genders are covered in this study.

# 4 Data pre-processing and Data Analysis

## 4.1 Data pre-processing

The key phase in any data analytic project is data analysis, pre-processing, and then preparation. These processes will answer the majority of questions before we start applying any model. Therefore, data analysis is critical in moving forward with any assumptions. To achieve that this study integrated All of the collected datasets to obtain helpful insights. Many adjustments were made to the collected data because they were in different formats. The daily weather data from Data Access Viewer is transformed into a weekly format to coordinate with weekly updated influenza data. The monthly average temperature from the Climate Change Knowledge portal was integrated with influenza data to see how colder and warmer months affect influenza transmission.

CEIC Data comprises yearly and monthly tourist arrival numbers, which are then merged into a single CSV by taking the monthly average, and for some countries, the yearly average, to see how increased visitor numbers affect the number of influenza reports

[4]Power Data access viewer `https://power.larc.nasa.gov/data-access-viewer/`

[5]CEIC data global database `https://www.ceicdata.com/en/belgium/tourism-statistics/be-international-tourism-number-of-arrivals`

[6]Department of Economic and Social Affairs `https://population.un.org/wpp/Download/Standard/Population/`

registered across respective countries. Statistics from the Union Nations' population were collected, and the yearly average of each age group's data was calculated.

Dataset is converted into time step data where time steps allow us to examine and analyze the data over a set of time intervals. The time steps attribute is a unique integer assigned to each row that starts at zero and indicates the year, month, and week. TimeSteps are a means of breaking down each moment of time or event in a data set into smaller chunks. To generate time series of the data, the Timestep attribute is being constructed. The pandas.interpolation() method is used to check for NaN values. Rather than hard-coding the value, it uses a variety of interpolation algorithms to fill in the gaps. Numerical values are preferred for fitting data into a machine learning system. Therefore, categorical variables will be transformed to numerical after careful understanding (Figure 3).

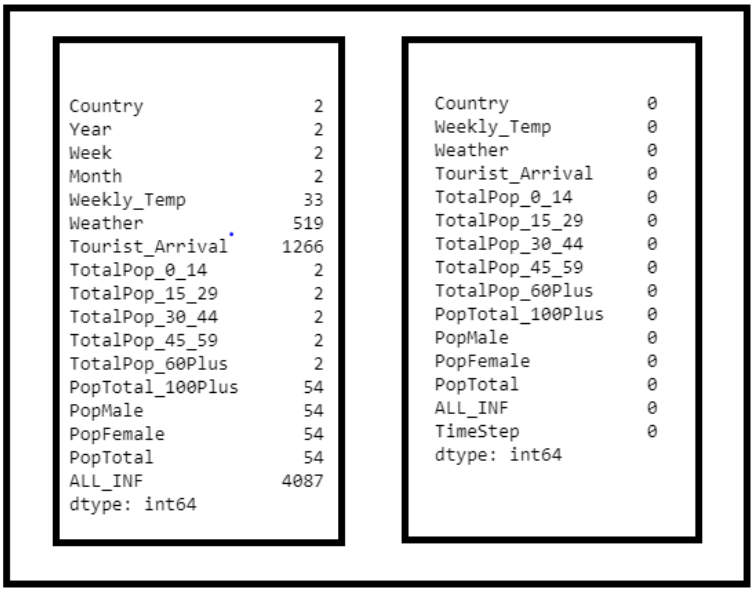



Figure 3: Data Before and after pre-processing

## 4.2 Data Analysis

It is essential to visualize and work on exploratory data analysis(EDA) to have a comprehensive understanding of the dataset. Exploratory data analysis was carried out in this study using various visualizations and pandas data frame features. As shown in figure 4 the most influenza cases were reported in France and the United Kingdom. Also, we can see that the number of cases increases from year to year in most nations, with the greatest number of influenza cases recorded in 2019 for most of the individual countries. The total population of each country is visualized in figure 5, it is observed that France, Germany, Italy, Spain, and the United Kingdom all have significant populations. However, only France and the United Kingdom recorded a high number of influenza cases. It will demonstrate how successfully other countries may have handled the transmission rate. We can't conclude because there could be other reasons why these countries have a low number of influenza instances.

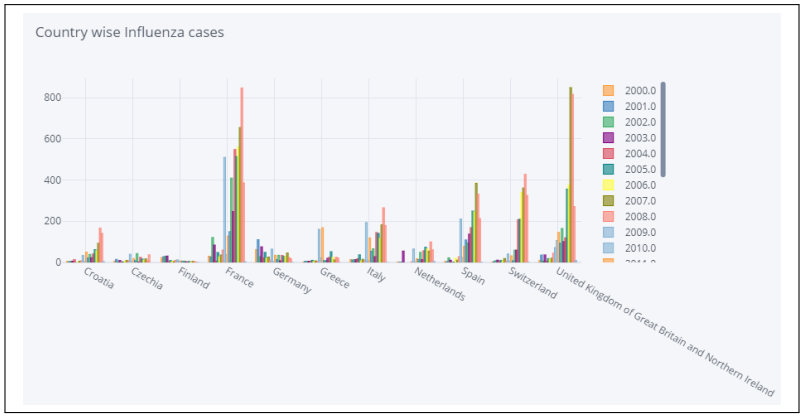



Figure 4: Total influenza cases by country

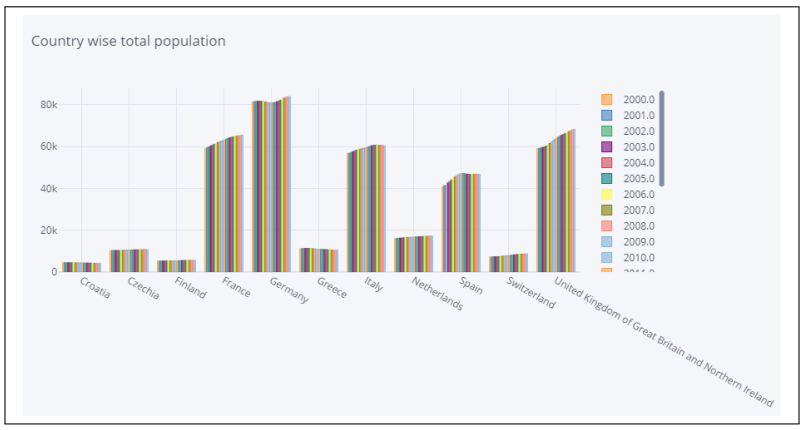



Figure 5: Total population of individual countries

As a consequence, other contributing elements such as visitor arrival, weather, and demographic age groups were investigated in this study. In the figure 6 It has been observed that France and Czechia have received a greater number of tourists. The theory that tourism is one of the causes for viral infection transmission could not be supported by this representation.

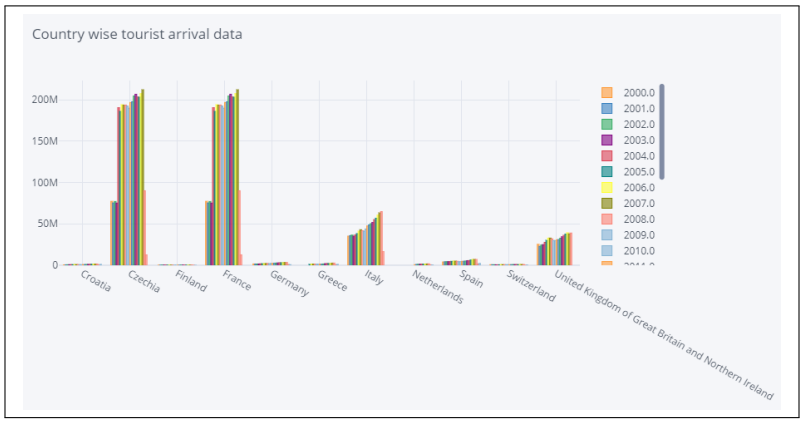



Figure 6: Tourist Arrival

This study subsequently developed a correlation heat map, as shown in figure 7, to verify the correlation of these parameters. The number of cases documented for chosen countries is stored in the ALL INF(ALL the Influenza cases recorded) variable. The ALL INF variable did not show a significant link with any other parameters in the map, although it did show a better correlation with the population age group 0 to 14 and 100 and up. Despite the fact that meteorological data were analyzed on a monthly and weekly basis, there was no significant association. As a result, all age groups in the population will be included for future research.

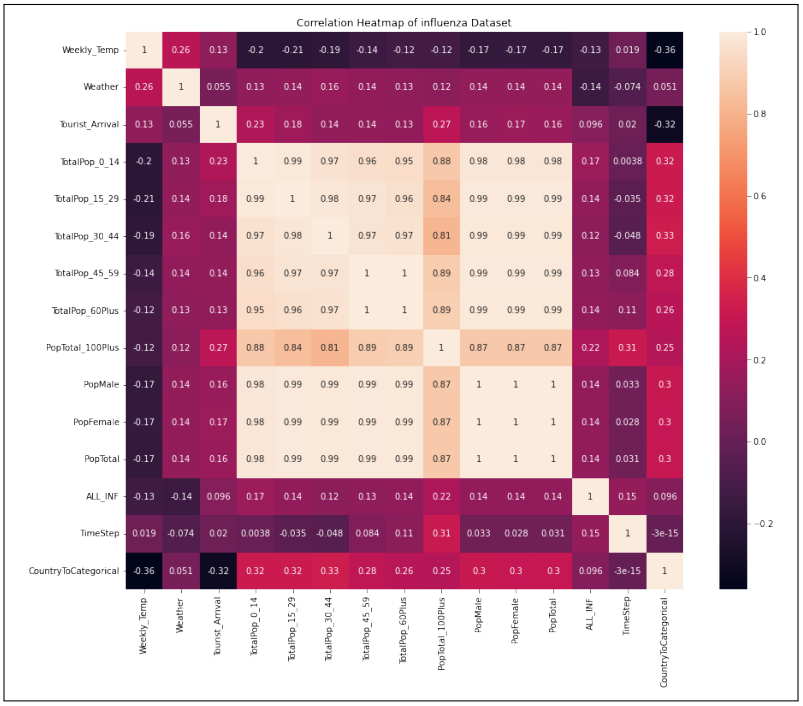



Figure 7: Correlation heatmap

# 5 Model Implementation

This section will provide detailed descriptions of the model used in this study : ARIMA, Random Forest Regressor, SVM, Attention LSTM model, and Hybrid model (ARIMA+ Attention LSTM Model).

## 5.1 ARIMA

AutoRegressive Integrated Moving Average (ARIMA) is an autoregressive model that uses historical influenza infection data to predict future values. ARIMA models are highly effective in forecasting in a variety of investigations. As a result, the ARIMA model was utilized to evaluate the fact and predict influenza cases in this study. ARIMA is comprised of two components autoregression (AR) and Moving Average (MA). AR denotes that the parameter is regressed on its previous lagged values as it evolves. The linear composite of error factors whose values occurred concurrently and at various times in the past is denoted by the MA section of the equation. The ARIMA family of models allows you

to express specific phenomena across time in a concise manner and estimate the future outcomes of certain parameters with the confidence interval around the predictions.

ARIMA models have different mathematical formulas depending on the developer. The one used in this study is the one implemented in the package Python "statsmodels".

If Xt is the time serie, and  is its mean then $ARIMA(p, d, q)$
$Yt = (1B)d(1Bs)DXt - ; \Phi(B)(Bs))Yt = \theta(B)\Theta(Bs)Zt, ZtN(0, \sigma)$
$\Phi(z) = 1pi = 1izi, (z) = 1pi = 1izi; (z) = 1 + qi = 1izi, (z) = 1 + qi = 1izi$

p is the order of the autoregressive model.
q is the mean of mobile part of the model.
d is the order of the differentiation of the model.
D is the differentiation part of the sustainability of the model.
s is the period of the model.

To perform ARIMA modelling, we must first determine whether the time series is stationary.

```
CountryToCategorical : P-Value = 0.0 => Stationary.
TimeStep : P-Value = 0.0 => Stationary.
PopTotal : P-Value = 0.0 => Stationary.
PopTotal_100Plus : P-Value = 0.0 => Stationary.
Weather : P-Value = 0.0 => Stationary.
Tourist_Arrival : P-Value = 0.0 => Stationary.
ALL_INF : P-Value = 0.0 => Stationary.
```

Here we can see all the series in the data set are stationary. After this, we can perform the Johansen cointegration test in multivariate time series to see if the series is correlated to each other or not.

The Johansen co-integration test results are as follows:

```
Column Name   >  Test Stat > C(95%)    =>   Signif
 ----------------------------------------
PopTotal >  6114.14   > 60.0627   =>   True
PopTotal_100Plus >  4487.6    > 40.1749   =>   True
Weather >  2946.63   > 24.2761   =>   True
Tourist_Arrival >  1487.34   > 12.3212   =>   True
ALL_INF >  71.19     > 4.1296    =>   True
```

Here we can see that the multivariate time series we are using are correlated. Now we can apply the Auto ARIMA model. Which will tell us the order of p and q for our VARMA model.

| | p | q | RMSE ALL_INF | MAE | MAPE |
|---|---|---|---|---|---|
| 0 | 7.0 | 0.0 | 94835.320633 | 82119.638459 | 3.413856e+06 |
| 1 | 7.0 | 0.0 | 94835.320633 | 82119.638459 | 3.413856e+06 |
| 2 | 5.0 | 0.0 | 94672.089486 | 81986.423497 | 3.408348e+06 |
| 3 | 6.0 | 0.0 | 94642.319022 | 81962.232175 | 3.407348e+06 |
| 4 | 4.0 | 0.0 | 94809.663737 | 82105.876690 | 3.413312e+06 |

By analyzing the $RMSE$ score, the optimal orders are p = 7 and q = 0 will give the best score and forecasting values for this multivariate time series.

## 5.2 SVR

Support Vector Regression (SVR) is a supervised learning technique for forecasting discrete floating values. The purpose of this type of SVM is to find the best line for a given situation. The hyperplane with the greatest number of points in SVR is the right fit line. Unlike other regression models, the SVR seeks to fit the best line within a threshold value rather than minimising the difference between the actual and predicted values. The threshold value is the distance between the hyperplane and the boundary line. This study implements an SVM regressor with a linear kernel. The grid search technique for the parameters C and gamma was used in this study to fine-tune the results. The set tested is 'C': [1,5,10,100,1000,10000], 'gamma' : ('auto','scale').

The optimal parameters are :

```
tuned hpyerparameters :(best parameters)  {'C': 1000, 'gamma': 'auto'}
accuracy : 0.31947297805681163
```

## 5.3 Random Forest Regressor

Random forest is a bagging technique-based Supervised Learning algorithm. The bagging technique is developed to make machine learning algorithms more stable and accurate. Random forest employs averaging to increase predicted accuracy and control overfitting by fitting multiple classification decision trees on diverse subsamples of the dataset. It works by creating a large number of decision trees during training and then producing the class that would be the mode of the individual trees' mean prediction (regression). The topmost decision node in a tree is the root node, which corresponds to the best predictor. Decision trees can handle both numerical and categorical data.

In this investigation, all datasets were converted to categorical representation. The parameters max depth and max features were fine-tuned in this investigation employing the grid search technique. The set tested is ['max depth':[100,1000,10000], 'max features':['auto', 'sqrt', 'log2']].

The optimal parameters are :

```
tuned hpyerparameters :(best parameters)  {'max_depth': 100, 'max_features': 'sqrt'}
accuracy : 0.8918699545093635
```

## 5.4 LSTM

The LSTM (Long Short-Term Memory) architecture is a Recurrent Neural Network (RNN)-based design used in time series forecasting and as well as natural language processing. When dealing with Time-Series the LSTM Model is indeed very effective. Attention technique can be utilized in conjunction with LSTM to increase model performance. In natural language processing, the attention mechanism was presented as an upgrade over the encoder decoder-based neural machine translation system (NLP). As a result, in this work, an LSTM with an attention mechanism was used to assess performance. The model that was employed in this investigation is shown below :

| Layer (type) | Output Shape | Param # | Connected to |
|---|---|---|---|
| encoder_input (InputLayer) | [(None, 4, 5)] | 0 | [] |
| permute (Permute) | (None, 5, 4) | 0 | ['encoder_input[0][0]'] |
| dense (Dense) | (None, 5, 4) | 20 | ['permute[0][0]'] |
| permute_1 (Permute) | (None, 4, 5) | 0 | ['dense[0][0]'] |
| multiply (Multiply) | (None, 4, 5) | 0 | ['encoder_input[0][0]', 'permute_1[0][0]'] |
| lstm (LSTM) | [(None, 4, 64), (None, 64), (None, 64)] | 17920 | ['multiply[0][0]'] |
| repeat_vector (RepeatVector) | (None, 1, 64) | 0 | ['lstm[0][1]'] |
| dense_1 (Dense) | (None, 1, 128) | 8320 | ['repeat_vector[0][0]'] |
| dropout_dec_input (Dropout) | (None, 1, 128) | 0 | ['dense_1[0][0]'] |
| dense_2 (Dense) | (None, 1, 1) | 129 | ['dropout_dec_input[0][0]'] |
| multiply_1 (Multiply) | (None, 1, 128) | 0 | ['dropout_dec_input[0][0]', 'dense_2[0][0]'] |
| lstm_1 (LSTM) | (None, 64) | 49408 | ['multiply_1[0][0]', 'lstm[0][1]', 'lstm[0][2]'] |
| decoder_dense (Dense) | (None, 1) | 65 | ['lstm_1[0][0]'] |

```
Total params: 75,862
Trainable params: 75,862
Non-trainable params: 0
```

Model is trained until 55 epochs and then recorded loss value of 0.0074 which is the lowest score. The loss function is recorded and the evolution of loss function $MSE$ for the training and validation set is shown in the graph 8 below.

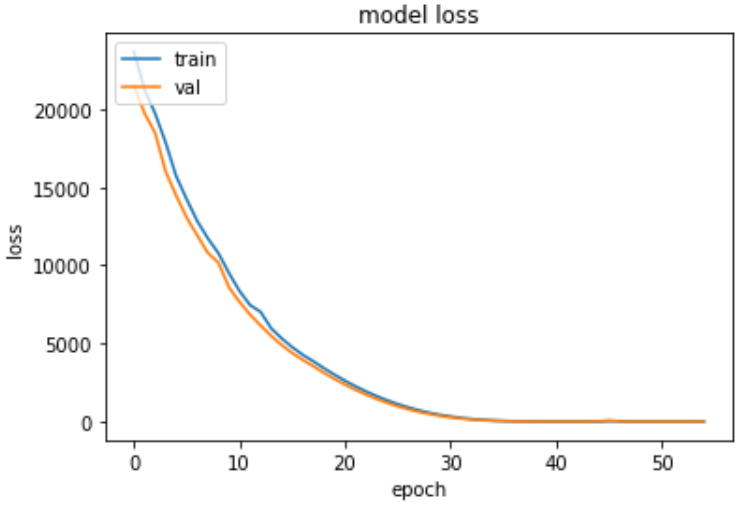



Figure 8: Evolution of model loss

### 5.5 Hybrid Model : Arima+LSTM

In many circumstances, a hybrid model is a blend of one or more models that have been demonstrated to provide desired results.ARIMA and attention LSTM models are combined in this study to create a hybrid model. The Vector Autoregression Moving-Average with Exogenous Regressors (VARMAX) is utilized in this study which models exogenous variables. Exogenous variables, often known as covariates, are parallel sequence input data with observations occurring at the identical time steps as the source series. Therefore, VARMAX is a method that applies to multivariate time series with exogenous variables that do not have a trend. In this study, the attention LSTM model uses the result of this VARMAX as an input parameter. The model did not perform as intended, but it did open up the possibility of thinking about how two models may be integrated and worked on to produce the desired results.

For VARMAX, the optimum previously identified p and q values (p= 7 and q = 0) have been used, and above that, attention LSTM is employed to forecast the future values.

## 6 Evaluation

The dataset was separated into a training and test subset, and following the training phase, each model was assessed on the test data. The five models: ARIMA, Random Forest Regressor, SVM, Attention LSTM model, and ARIMA+ Attention LSTM Model were evaluated based on the performance metrics $RMSE$, $MAE$ and $MAPE$.

| | ARIMA | SVR | Random Forest Regressor | LSTM | ARIMA+LSTM |
|---|---|---|---|---|---|
| $RMSE$ | 94835.32 | 216.845 | 84.991 | 0.086 | 341.76 |
| $MAE$ | 82119.63 | 74.033 | 18.619 | 0.054 | 341.76 |
| $MAPE$ | 3413856 | 128.694 | 101.906 | 0.124 | 100.0 |

- $RMSE$**:** Root Mean Squared Error ($RMSE$) is calculated by taking square root of mean squared error ($MSE$) where $MSE$ stands for the difference between the actual and predicted results, which is calculated by squaring the average difference throughout the data set.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$$

Where,

$\hat{y}$ $-$ *predicted value of* $y$

$\bar{y}$ $-$ *mean value of* $y$

- $RMSE$ calculated for each model, as shown in the table. ARIMA has a high $RMSE$ value (94835.32) and LSTM has a low $RMSE$ value(0.086), As a result, the LSTM model outperformed the competition in this case. The second best model for this problem is the Random Forest Regressor model which has the second lowest $RMSE$ of 84.991.

- **MAE:** Mean Absolute Error(MAE) calculates the absolute average distance between both the actual and predicted data.

$$\text{MAE} = \frac{|(y_i - y_p)|}{n}$$

$y_i$ = actual value

$y_p$ = predicted value

$n$ = number of observations/rows

- ARIMA has a high MAE of 82119.63, while LSTM received the lowest MAE of 0.054. The RF model has the second-lowest MAE (18.619). The hybrid model received a score of 341.76 and an SVR of 74.033. These figures show that LSTM outperformed other models once again.

- $MAPE$**:** Mean Absolute Percentage Error ($MAPE$) is one of the performance evaluation metric to measure forecasting accuracy. Expressed as,

$$M = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

$M$ = mean absolute percentage error

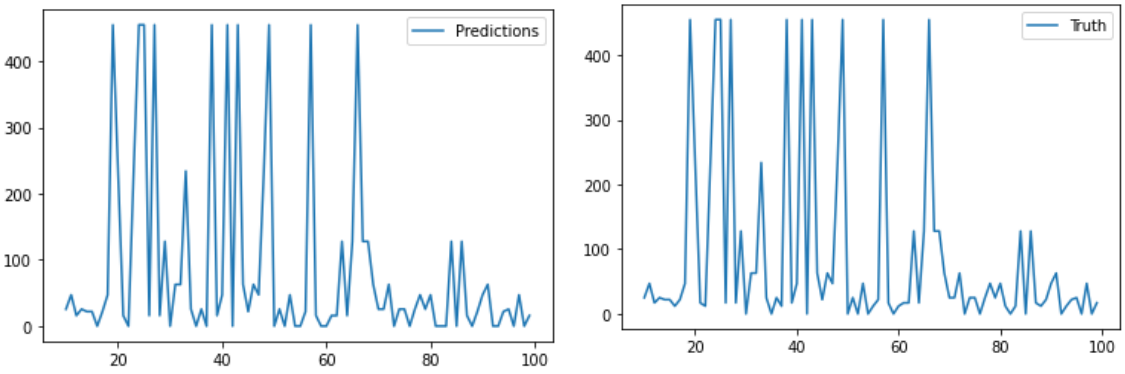
$n$ = number of times the summation iteration happens

$A_t$ = actual value

$F_t$ = forecast value

- Similarly, with a $MAPE$ score of 0.124, the LSTM outperformed the other models, while the hybrid model and Random Forest Regressor performed well compared to SVR and ARIMA with $MAPE$ values of 100 and 101.906, respectively.

As shown in the table above, the LSTM model performed excellently as the best fit model for this challenge of forecasting influenza cases, with $RMSE$ values of 0.086, $MAE$ values of 0.054, and $MAPE$ scores of 0.124. The Predictions and Truth value compared in the following figures.




(a) (b)

# 7 Conclusion and Future Work

The subject of this study is to investigate possible machine learning techniques to forecast influenza cases. This study first started to gather public records data for 11 countries between 2000 and 2021: Croatia, Czechia, Finland, France, Germany, Greece, Italy, Netherlands, Spain, Switzerland, and the UK. And then stored them in a CSV File. Furthermore, the number of different age group people in the population data, as well as the number of tourists arriving, were discussed along with weather data. Despite the lack of a strong correlation between these parameters and the number of influenza cases recorded, all of these influencing factors were combined to model, since they showed a nearly greater correlation with the influenza pattern. Five machine learning models were investigated in this study: ARIMA, SVM Regression, Random Forest Regression, LSTM with attention model, and a hybrid model ARIMA+LSTM. This study concluded that LSTM performed very well with the lowest $RMSE$ error rate (0.08) after evaluating these five models based on $RMSE$. This study discovered that the population age groups of 0 to 14 and 100 and over had better correlations, indicating that younger and elderly persons are more vulnerable to influenza infection. Further contributing factors, such as socio-demographic factors, could be collected in the future to assess the fact that discourse about demographic shifts accelerates the transmission rate. In addition, the study can be conducted to evaluate how these diseases spread during less busy times of the year, such as festivals and other public holidays. In the case of COVID-19, we may apply the Attention LSTM model to assist governments and public groups in implementing strategic decisions to rescue humanity in this pandemic emergency.

# 8 Acknowledgement

Prof. Aaloka Anant, my mentor, deserves special thanks for his compassion and support throughout the course term. He supplied the necessary guidelines and strategies for conducting extremely beneficial research that aids society in pandemics such as Covid 19. I would also want to thank my friends and family for their inspiration and support throughout my studies.

# References

Alosaimi, B., Naeem, A., Hamed, M., Alkadi, H., Alanazi, T., Rehily, S., Almutairi, A. and Zafar, A. (2021). Influenza co-infection associated with severity and mortality in covid-19 patients, *Virology Journal* **18**.
**URL:** *https://www.authorea.com/users/356231/articles/479174-influenza-co-infection-associated-with-severity-and-mortality-in-covid-19-patients*

Birrell, P., Zhang, X.-s., Corbella, A., Leeuwen, E., Panagiotopoulos, N., Hoschler, K., Elliot, A., McGee, M., de Lusignan, S., Presanis, A., Baguelin, M., Zambon, M., Charlett, A., Pebody, R. and Angelis, D. (2019). Forecasting the 2017/2018 seasonal influenza epidemic in england using multiple dynamic transmission models: a case study.

Carter-Templeton, H., Templeton, G. F., Nicoll, L. H., Maxie, L., Kittle, T. S., Jasko, S. A., Carpenter, E. E. and Monsen, K. A. (2021). Associations between weather-related data and influenza reports: A pilot study and related policy implications, *Applied Nursing Research* p. 151413.
**URL:** *https://www.sciencedirect.com/science/article/pii/S0897189721000197*

Changruenngam, S., Bicout, D. and Modchang, C. (2020). How the individual human mobility spatio-temporally shapes the disease transmission dynamics, *Scientific Reports* **10**.

Danzon, M. (2000). Climate change and stratospheric ozone depletion early effects on our health in europe, *World Health Organization Regional Publications - European Series* pp. vii–95.

Hussain, S., Harrison, R., Ayres, J., Walter, S., Hawker, J., Wilson, R. and Shukur, G. (2005). Estimation and forecasting hospital admissions due to influenza: Planning for winter pressure. the case of the west midlands, uk, *Journal of Applied Statistics* **32**: 191–205.

Kara, A. (2021). Multi-step influenza outbreak forecasting using deep lstm network and genetic algorithm, *Expert Systems with Applications* **180**: 115153.
**URL:** *https://www.sciencedirect.com/science/article/pii/S0957417421005947*

Ma, R. e. a. (2021). The prediction and analysis of covid-19 epidemic trend by combining lstm and markov method, *Scientific reports vol* pp. 15–46.

Monto, A. S., Ansaldi, F., Aspinall, R., McElhaney, J. E., Montaño, L. F., Nichol, K. L., Puig-Barberà, J., Schmitt, J. and Stephenson, I. (2009). Influenza control in the 21st

century: Optimizing protection of older adults, *Vaccine* **27**(37): 5043–5053.
**URL:** *https://doi.org/10.1016%2Fj.vaccine.2009.06.032*

Norrulashikin, M. A., Yusof, F., Hanafiah, N. H. M. and Norrulashikin, S. M. (2021). Modelling monthly influenza cases in malaysia, *PLOS ONE* **16**(7): 1–12.
**URL:** *https://doi.org/10.1371/journal.pone.0254137*

Novaes de Amorim, A., Deardon, R. and Saini, V. (2021). A stacked ensemble method for forecasting influenza-like illness visit volumes at emergency departments, *PLOS ONE* **16**(3): 1–15.
**URL:** *https://doi.org/10.1371/journal.pone.0241725*

Nsoesie, E., Oladeji, O., Abah, A. and Mbah, M. (2021). Forecasting influenza-like illness trends in cameroon using google search data, *Scientific Reports* **11**.

Perone, G. (2020). Comparison of arima, ets, nnar, tbats and hybrid models to forecast the second wave of covid-19 hospitalizations in italy (hedg-wp 20/18 - now published in the european journal of health economics), *SSRN Electronic Journal* .

Poirier, C., Hswen, Y., Bouzill, G., Cuggia, M., Lavenu, A., Brownstein, J. S., Brewer, T. and Santillana, M. (2021). Influenza forecasting for french regions combining ehr, web and climatic data sources with a machine learning ensemble approach, *PLOS ONE* **16**(5): 1–26.
**URL:** *https://doi.org/10.1371/journal.pone.0250890*

Sebastian, R., Skowronski, D. M., Chong, M., Dhaliwal, J. S. and Brownstein, J. S. (2008). Age-related trends in the timeliness and prediction of medical visits, hospitalizations and deaths due to pneumonia and influenza, british columbia, canada, 1998-2004., *Vaccine* **26 10**.

Stockmann, C., Pavia, A., Hersh, A., Spigarelli, M., Castle, B., Korgenski, K., Byington, C. and Ampofo, K. (2014). Age-specific patterns of influenza activity in utah: Do older school age children drive the epidemic?, *Journal of the Pediatric Infectious Diseases Society* **3**: 163–167.

Tang, W. and Rllin, A. (2018). Model identification for arma time series through convolutional neural networks, *Decision Support Systems* **146**.

Temr, A. and Yldz, . (2021). Comparison of forecasting performance of arima lstm and hybrid models for the sales volume budget of a manufacturing enterprise, *Istanbul Business Research* **50**: 15–46.

Venkatramanan, S., Sadilek, A., Fadikar, A., Barrett, C., Biggerstaff, M., Chen, J., Dotiwalla, X., Eastham, P., Gipson, B., Higdon, D., Kucuktunc, O., Lieber, A., Lewis, B., Reynolds, Z., Vullikanti, A. K., Wang, L. and Marathe, M. (2021). Forecasting influenza activity using machine-learned mobility map, *Nature Communications* **12**.

Volkova, S., Ayton, E., Porterfield, K. and Corley, C. (2017). Forecasting influenza-like illness dynamics for military populations using neural networks and social media, *PLOS ONE* **12**: e0188941.