Monkey Pox Visualisations, Forecasting, and Analysis.

1stKumkum Geervani *Dept. of Computer Science PES University SRN:PES1UG20CS184*

Bengaluru,India [kumkum1geerv](mailto:kumkum1geervani@gmail.com)[ani@gmail.com](mailto:ani@gmail.com)

2ndK Trisha

*Dept. of Computer Science*

*PES University SRN:PES1UG20CS187*

Bengaluru, India [trishanaidu802@gmail.com](mailto:trishanaidu802@gmail.com)

3rd Diya Anne Tharappan *Dept. of Computer Science PES University SRN:PES1UG20CS137*

Bengaluru,India

[diya.star2002@gmail.com](mailto:diya.star2002@gmail.com)

***Abstract*—The “Monkeypox” dataset, which was published by Edouard Mathieu, Fiona Spooner, Saloni Dattani, Hannah Ritchie, and Max Roser in 2022, is the focus of this study provided online at OurWorldInData.org.virus, which includes the daily total of cases that have been confirmed for every nation where the virus has been detected. In this project, we seek to reveal underlying patterns and structures in the data while also offering summary-level insights about the monkeypox dataset.**

***Index Terms*—virus,disease,vaccine**

1. INTRODUCTION AND BACKGROUND

In 2018, two unrelated travellers from Nigeria were found to have monkeypox in the UK. In the UK at that time, the first confirmed mortal-to-mortal transmission outside of Africa. This individual worked in healthcare and may have acquired the complaint from contaminated bedding. Additionally, cases in travellers to Singapore and Israel were mentioned. More cases were reported in the UK in 2019 and 2021.

In June 2022, the World Health Organization announced that, in keeping with its objective to prevent false links with particular places or species, it will come up with a new name for the complaint. Monkeypox is transmitted similarly to smallpox through a bone, albeit the illness is typically milder with monkeypox.In the DRC, the complaint reappeared in 1996, with 88 percent of cases involving transmission from mortal to mortal. In tropical Central and West Africa, there are frequently small virus outbreaks with mortality rates of around 10 percent and secondary mortal-to-mortal infection rates of almost the same quantity. Until 2003, when a monkeypox outbreak spread to the US, the human complaint was restricted to the rain forests of Western and Central Africa. Champaign tykes from the start contracted the illness and spread it to their owners. There were no fatalities, and the complaint was designed to be minor. The complaint was filed between 1970 and 2019 in 10 African nations, primarily in Central and West Africa. In 2018, two unrelated travellers from Nigeria were found to have monkeypox in the UK. In the UK at that time, the first confirmed mortal-to-mortal transmission outside of Africa. This individual worked in healthcare and may have acquired the complaint from contaminated bedding. Additionally, cases in travellers to Singapore and Israel were mentioned. More cases were reported in the UK in 2019 and 2021.

The virus that causes monkeypox is similar to the one that causes smallpox, although monkeypox sickness is typically less severe. Because it was initially isolated in monkeys, it is known as monkeypox. However, rodents—not monkeys—are the virus’s main carriers. The smallpox vaccine offers some protection from the monkeypox virus, and those who have never had the smallpox vaccine may be more susceptible to developing monkeypox disease. The monkeypox virus, which causes the infectious disease, can affect some animals, including humans. The first signs include a fever, headache, muscle aches, swollen lymph nodes, and fatigue. Monkeypox in humans can be contracted by coming into touch with an infected person’s rash or bodily fluids, such as respiratory droplets. A person can contract an infection by close intimate contact, whether sexual or not.

The first monkeypox death in India was reported on July 31, 2022. The victim was a 22-year-old man who had just returned from the United Arab Emirates. The resident went back to the UK on May 4, which led to the outbreak’s initial incidence there.

Through our project, we would like to discover hidden patterns and correlations that exist in the data through which we can inhibit more cases from occurring.We seek to extract insights from the monkey pox dataset used for this project using understandable charts, graphs, and visualisations.

1. PREVIOUS WORK

The work that has already been done for diverse purposes is summarised in this section, with an emphasis on the aspects that are pertinent to our work. The underlying presumptions, an outline of the methodology, and the outcomes have all been mentioned. Additionally, any restrictions and any lacuna implied by the subsequent approaches have been properly mentioned.

A Saied et.al [1] explores how the shortage of vaccines during the COVID-19 pandemic in the countries of Africa sparked the birth of the African vaccine industry. It elaborates on the shift of African countries relying heavily on imported drugs and vaccines from other countries to being self sufficient post pandemic on the vaccine front. It also delved deep into the challenges and perspectives of building vaccine production capacities in Africa. Africa is a tropical continent, therefore for our experiment, we were interested in finding out how

temperature influences various elements of the illness. Further- more when we were drawing inferences from our dataset, we found that Africa had one of the lowest ranking percentages of monkeypox cases and were curious to know if the increased fortification and self sufficiency of Africa against the COVID- 19 pandemic had any role to play in it. As vaccines must be stored at low temperatures to preserve their effectiveness, maintaining an uninterrupted cold chain from manufacture to patient is of the utmost importance. According to the paper’s description of how this procedure was acquired, in Sierra Leone,the Arktek Cold Storage Device was used to deliver vaccinations that saved the lives of countless Africans.

The papers Suraweera et.al [2],Murugaiah et.al[3],Fuminari et.al[5] helped us understand the science behind the monkey- pox disease. Specifically Suraweera et.al [2] investigates the structural biology, molecular connections, and precise mecha- nism of action of apoptosis inhibitors encoded in poxviruses and how they affect host-virus interactions to eventually facil- itate viral infection and spread.

Murugaiah et.al[3] focuses on these interactions of com- plement components (particularly C1q, C4b-binding protein, properdin, factor H, Mannose-binding lectin, and Ficolins) with various viruses and their effects.

Fuminari et.al[5] examines the confirmed monkeypox cases that have lately been found in the Netherlands up through May 31, 2022, the incubation period of monkeypox using the stated time of exposure and symptom onset. Monkeypox outbreaks have been documented since the beginning of May 2022 in nations, primarily in Europe, where the monkeypox virus is not indigenous. Monkeypox was categorised as a group A notifiable illness in the Netherlands on May 21, 2022. As of May 31st, 31 cases of monkeypox had been PCR-lab confirmed nationwide. The cases ranged in age from 23 to 64, were all men, and identified as MSM. At the time of data collection, 18 instances had indicated that the date of the first symptoms and the date of exposure were associated to the attendance of an event where exposure was thought to be most likely. In the paper,they used a likelihood-based technique, which allows for exposure to be either a single time point or a time interval, to fit parametric distributions to the reported incubation periods among 18 individuals with symptom onset and exposure histories for monkeypox.The calculation was carried out in R-4.0.5 using the rstan-2.21.2 package. The lognormal, gamma, and weibull distributions were contrasted with one another to see which parametric distribution best fit the data. A lognormal distribution best fit the observed incubation times for monkeypox. The mean incubation period was calculated using this best-fitting distribution to be 8.5 days (95 percent credible intervals (CrI): 6.6-10.9 days), with the 5th percentile being 4.2 days and the 95th percentile being

17.3 days.

Marceline et.al[6] discusses about a male youngster with monkeypox who is under 10 years old and lives in the Netherlands. A potential source of infection was not found despite careful source tracing. There were no subsequent instances found among close contacts.A male child under the

age of 10 who had no pertinent medical history presented to an emergency room for children (ER) in Amsterdam, the Netherlands, at the end of June 2022. He received vaccinations in accordance with the Dutch government’s national immuni- sation programme, and at age 5, he contracted chicken pox. He had a sore throat without a fever three weeks before to his visit, but it went away on its own the next day. He left for Turkey the following day for a weeklong vacation. Two tiny circular skin lesions appeared on his left lower jaw and face when he got back. Under the assumption of a mild dermatomycosis, the general practitioner (GP) started the patient with antifungal cream. The child’s face developed further lesions in the days that followed. A second visit to the doctor was made, and an antibiotic cream was applied for possible impetigo vulgaris. The boy was admitted to the hospital on the clinical suspicion of monkeypox after 20 lone lesions occurred on different body parts. A healthy child with stable vital signs and no fever was seen during a physical checkup at the hospital. According to the paper, thorough source and contact tracing failed to locate any potential sources of the virus. The patient’s sequence was assigned to clade 3 lineage B.1 by whole genome sequencing, although it was not directly linked to any other strains from the Amsterdam area. No probable source could be found, leaving us with an unanswered question about the transmission that we hope to discover when analyzing our dataset.

M.M.H.et.al[4]explains how data from the CovIdentify re- search (6765 participants) and the MyPHD study (8580 partic- ipants), including smartwatch data from 1265 people of whom

126 tested positive for COVID-19, were used to construct an Intelligent Testing Allocation (ITA) technique.We looked at the paper’s thorough model and parameter search, which revealed the best intervals and overall metrics for monitoring continuous digital biomarkers to raise the COVID-19 diag- nostic test’s positivity rate. These insights helped us with the analysis of the monkeypox dataset.

Sitaula et.al[7] The study examines 13 different deep learn- ing (DL) models that have already been trained to detect the monkeypox virus. Four widely recognised metrics—Precision, Recall, F1-score, and Accuracy—are used for analysis and fine-tuning, respectively. After determining which DL models perform the best, ensembling is done to boost overall perfor- mance using a majority vote over the resulting probabilistic outputs. The suggested ensemble approach was tested on a publically available dataset, and the average Precision, Recall, F1-score, and Accuracy were 85.44 percent, 85.47 percent,

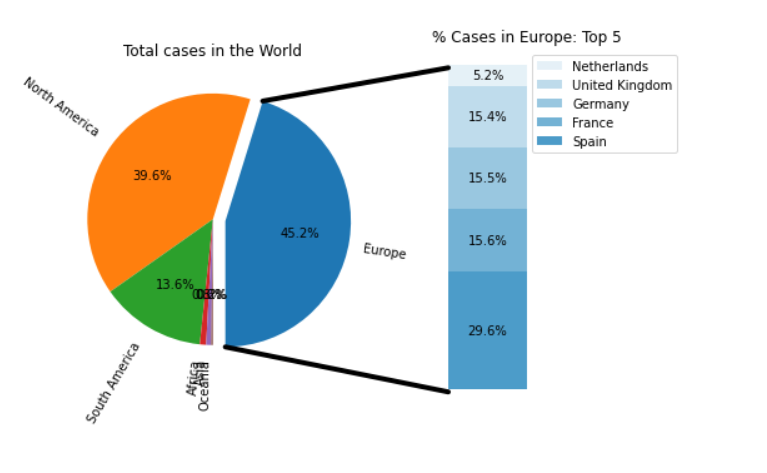
85.40 percent, and 87.13 percent, respectively.This paper’s main contributions include suggesting that all 13 pre-trained DL models for MonkeyPox detection be used with a single, fine-tuned architecture and comparing them; pick the top- performing DL models for group learning using an ablative study; Compare the proposed strategy to cutting-edge tech- niques; and The best-performing DL model’s explainability should be demonstrated using Grad-CAM and LIME. This paper greatly helped us unearth the key characteristics in the image classification of monkeypox images and broadened our knowledge on the disease.

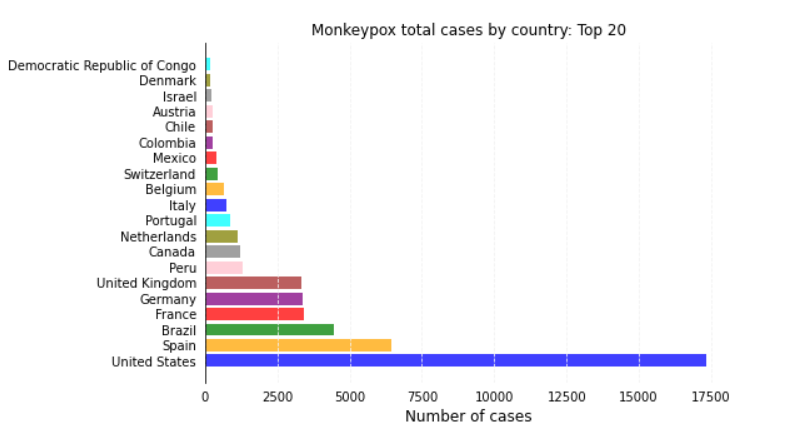
1. PROBLEM STATEMENT

Our project primarily consists of three stages: Data preprocessing and visualizations, forecasting and analysis.

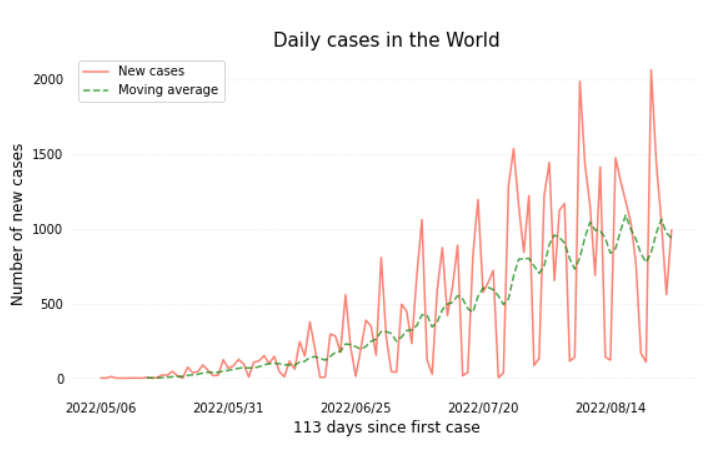
*A. Data preprocessing and visualizations*

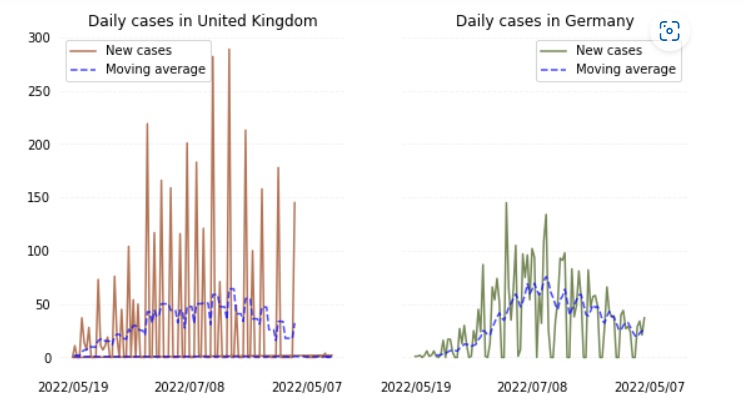
We wanted to structure the unstructured data in our dataset so we imported the monkeypox dataset and performed data cleaning activities like removing null values, and finding and eliminating duplicate values. To determine the actual number of nations, the value “World” in the location field of our data frame was removed from the data frame because it can be regarded as the total of all the countries that were provided. Additionally, we changed the names of some columns, such as changing “Congo” to the “Democratic Republic of Congo”. We then acquired the total cases by continent and the top twenty worst-hit countries followed by the top five badly affected countries in each continent, followed by visualizations for the same.





The null numbers for new deaths and total deaths are considered to be attributable to the absence of these, therefore they were changed to 0. We calculated the number of deaths-the deaths by continent and deaths by country, followed by graphing daily cases of monkeypox in the world. We also moved on to find the moving averages of the emerging new cases in each country, namely, France, Brazil, Spain and the UK.





*B. Auto regressive Forecasting*

Based on information about past behavior, an autoregressive (AR) model predicts future behavior.

When there is a relationship between the time series values and their foregoing and following values, this type of analysis is employed. In a linear model, the variable's previous values are used to predict its future value.

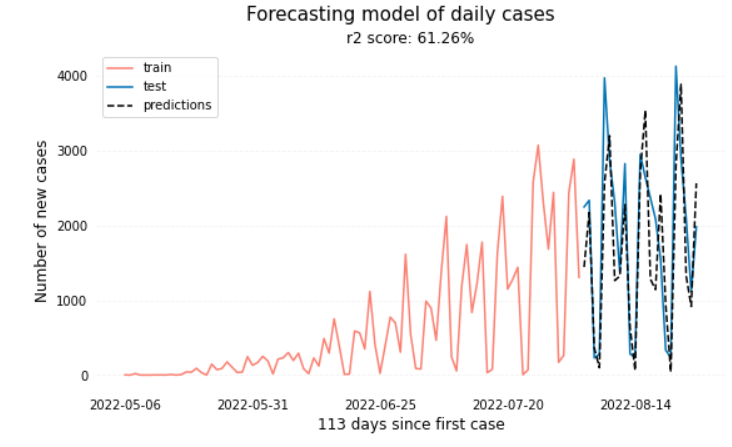
We have incorporated the 'new cases' feature into our model. The model is further tested and predictions are made using these values.

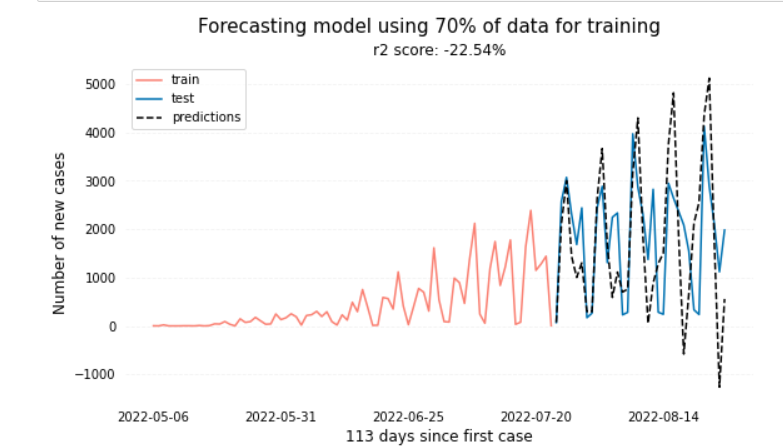
Eighty percent of the data are used for training, and the remaining twenty percent are used for testing purposes.

The AutoReg module was then imported with the aid of statsmodels.tsa.ar model to fit the data from new cases into this model with a log value of 9, making predictions for values within a specified range. We can see the number of new cases after 113 days since the initial occurrence by plotting the graph.

There is unquestionably a sharp rise in the number of new instances. We obtain an r2 score of 61.26% with 80% of the training data and an r2 score of -22.54% with 70% of the training data.

A regression model would not always be viable if the r2 score was low, indicating a low amount of correlation. Therefore, we may conclude that this model performs better after being trained with more data.





*C. ARIMA Forecasting*

Arima(Autoregressive Integrated Moving Average model is a class of statistical models for analyzing and forecasting time series data. Hence we have used Arima model in to predict our data.

Arima works with ACF(Auto Correlation Functions) and PACF(Partial Autocorrelation Function) plots. Based on the ACF and PACF plots, we choose the p and q values which will be used in our ARIMA model to predict the data. ARIMA has 3 components: Autoregressive components with p lags AR(p), Integration component(d), moving average with q lags MA(q). ACF plot gives us the Moving Average(MA) (q) term and PACF plot gives us the AR term(p). d here represents the number of differentiations.

In our model, we are predicting the values of ‘total\_deaths’ and ‘total\_cases’. We plotted ACF and PACF plots using plot\_acf and plot\_pacf functions available in statsmodel. Since the ACF plots in both the above variables showed a gradual decrease, we differentiated once. We then chose p and q values of 1 and 1 for total\_cases and total\_deaths. Hence our model was ARIMA(1,1,1) for ‘total\_cases’ and ARIMA(1,1,1) for ‘total\_deaths’. Any null values present after differntiation were dropped accordingly. We have also used auto arima model to predict the p,d,q values. Auto arima suggested (1,1,1) for ‘total\_cases’ and (3,1,4) for ‘total\_deaths’, but we chose (1,1,1) for our model.We then trained our ARIMA model using the ARIMA functionality from statsmodel library followed by fitting our model. We have created two different ARIMA models for ‘total\_cases’ and ‘total\_deaths’ respectively.

After our model building and training, we were then focused on finding the errors of our model. Hence, we split the data into training and testing data as ‘train’ and ‘test’ respectively. We then trained our model on the train data for both ‘total\_cases’ and ‘total\_deaths’. We calculated Mean Absolute Error(MAE), Mean Squared Error(MSE) and Root Mean Squared Error(RMSE) for our models. For ‘total\_cases’ we got MAE around 2008 and RMSE around 6569. This suggests that ARIMA is not the best model for our dataset. Our inference from these error metrics is that there are other exogeneous variables in our dataset which have a greater impact on our prediction data. Hence it was indicative to use other models for our data. We got MSE around 0.26 and RMSE around 1.5 for ‘total\_deaths’. We got such low errors because, the original dataset we used had null data and we imputed the data using Median. Hence we got such low errors.

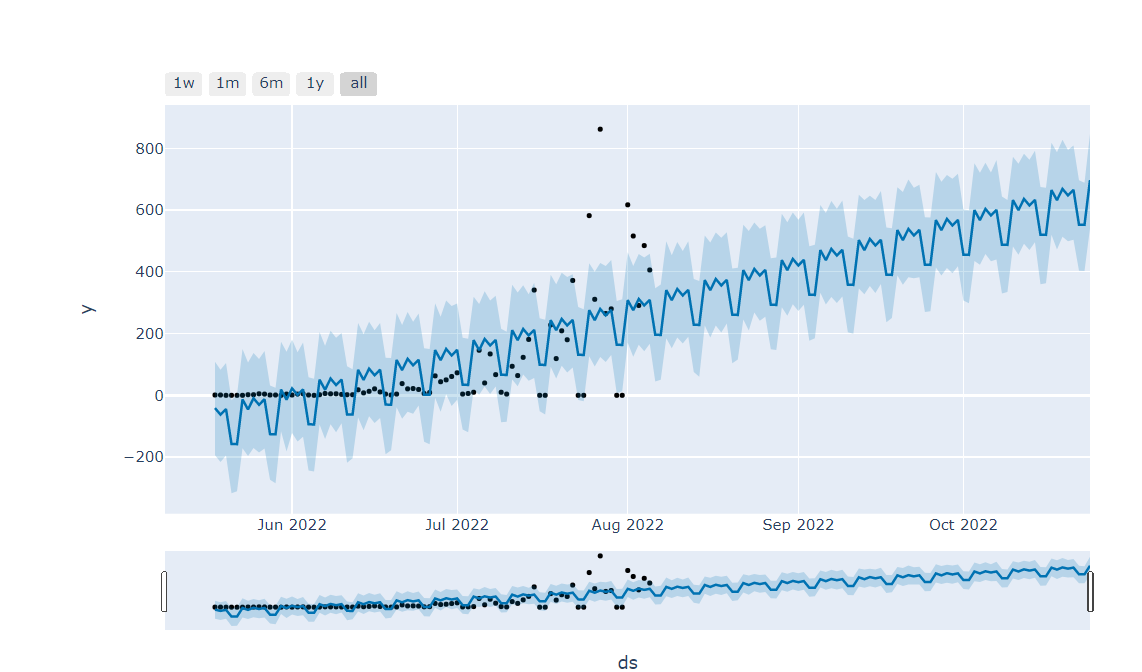
Based on huge errors that we got, we can conclude that ARIMA, which takes into account only the single response variable and not any other exogeneous variables, ARIMA is not suitable for our dataset. As we can see from our dataset, we have other variables like ‘location’, ‘new\_cases\_per\_million’ and many more which have a greater impact on ‘total\_deaths’. Hence choosing another model which considers all these variables is a good option.

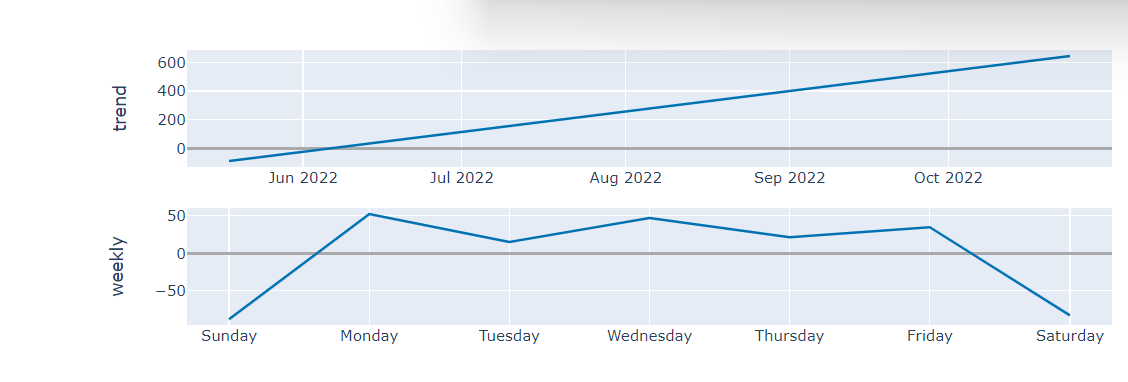
*D. Prophet Forecasting*

An open-source library called Prophet is made specifically for forecasting univariate time series datasets. In an effort to produce accurate forecasts for data with trends and seasonal structure by default, it is simple to use and built to automatically discover a decent set of hyperparameters for the model. In our project, we make use of the prophet forecasting model.

However, its limitations are that since it is an emerging virus and without the same data for a specific time, it will not provide us a complete view of what is about to happen, leaving Time Series a limited solution. Nonetheless, we have still tried to implement the same.

In our project we have tried to forecast new cases for the countries –‘India’ and the ‘USA’. We have created a data frame for the USA named usa\_data and a data frame for India labeled India\_data. Each dataset was then split into two partitions-train and test. The train had 80 percentage of the data and the test had 20 percentage of the data. Using prophet’s time series forecasting model we were able to forecast data as per the training data for each country. We also used interactive widgets and visualizations to effectively comprehend the forecasting done by the model. The forecasting was shown in such a way that the user could choose to view the forecasting one in a week, a month, six months, and even a year. We also used Prophet’s built-in functions to uncover metrics such as the trend and seasonality of the new cases emerging in each country.





The RMSE value obtained after evaluating this model was the lowest among all other models we tried, namely the Autoregressive and the ARIMA model, giving us an RMSE of 456.24. We were not satisfied with this metric but were pleased that it was not too huge as the errors predicted by other models.

1. ACKNOWLEDGMENT

The first and foremost person we would like to thank is our Data Analytics professor- Professor Anand MS for sharing his guidance and insights whenever we asked him queries or showcased our work to him. We would also like to extend our gratitude to all the teaching assistants and Data Analytics faculty for curating this course and project.

REFERENCES

[1] AbdulRahman A Saied,Asmaa A Metwally,Manish Dhawan,Om Prakash Choudhary,Hani Aiash, ”Strengthening vaccines and medicines manu- facturing capabilities in Africa: challenges and perspectives”

[2] Chathura D. Suraweera,Mark G. Hinds,Marc Kvansakul,”Poxviral Strategies to Overcome Host Cell Apoptosis”

[3] Valarmathy Murugaiah,Praveen M. Varghese,Nazar Beirag,Syreeta De Cordova,Robert B. Sim,Uday Kishore,”Complement Proteins as Soluble Pattern Recognition Receptors for Pathogenic Viruses” Accepted: 17 December 2020 / Published: 23 December 2020

[4] Shandhi, M.M.H., Cho, P.J., Roghanizad, A.R., A method for intelligent allocation of diagnostic testing by leveraging data from commercial wearable devices: a case study on COVID-19. npj Digit. Med. 5, 130 (2022).

[5] Miura Fuminari, van Ewijk Catharina Else, Backer Jantien A., Xiridou Maria, Franz Eelco, Op de Coul Eline, Brandwagt Diederik, van Cleef Brigitte, van Rijckevorsel Gini, Swaan Corien, van den Hof Susan, Wallinga Jacco. Estimated incubation period for monkeypox cases confirmed in the Netherlands, May 2022. Euro Surveill.

[6] Tutu van Furth A Marceline, van der Kuip Martijn, van Els Anne L, Fievez Lydia CR, van Rijckevorsel Gini GC, van den Ouden Anton, Jonges Marcel, Welkers Matthijs RA. Paediatric monkeypox patient with unknown source of infection, the Netherlands, June 2022. Euro Surveill.

[7] Chiranjibi Sitaula, Tej Bahadur Shahi, “Monkeypox virus detection using pre-trained deep learning-based approaches”.