An Analysis of Methods for Forecasting Epidemic Disease Outbreaks using Information from Social Media

Disha Sushant Wankhede, Rohan Rajendra Sadawarte, Mahek Ibrahim Mulla, Shreya Rahul Jadhav

Abstract: Predicting the rise or fall of an epidemic or pandemic is an essential part of establishing control over it. Post-World War 1, when there was an outbreak of the "Black Plague" there weren't any means to analyze and predict. Although today we are equipped with tools like Machine Learning and Artificial Intelligence which have certainly enabled us to prevent unnecessary loss of life. It helps prepare the health officials to build the infrastructure and interpret the intensity of preparedness regulation of resources. The aim of this survey is to analyze and shed some light on the various algorithms and methods such as - regression models, neural networks, ARIMA, etc. Before building any model, gathering and processing the data is also essential. Hence our paper also focuses on which social media platforms proved beneficial in comparison to all we found and then made fit to be incorporated into the models. While researching for this paper, we observed that every disease has a different transmission type that leads to an outbreak and is a key factor in constructing a model. The literature evaluation in this work is centered on various prediction algorithms and their strategies for extracting online data from social media sites like Facebook and Twitter, all of which have drawn a lot of interest in early disease diagnosis for public health.

Keywords: Machine Learning, Artificial Intelligence, social media, Pandemic, Epidemic, Outbreak, Covid-19, Influenza, Regression Models, Neural Networks.

I. INTRODUCTION

The Covid-19 pandemic, also known as the Coronavirus, first appeared in Wuhan, China, and then spread throughout the world. Following the unprecedented and rapid breakout of the Covid-19 pandemic, analysis algorithms showed the crests and troughs of the infected people, mortality rate, availability of resources, etc. We were also able to roughly predict the waves of Covid. The next disease our paper covers the spread of is Influenza, commonly known as Flu; seeing that Influenza was a major disease before Covid in 2019. The other diseases we encountered from a technical perspective were Malaria and Thunderstorm Asthma. Malaria and Flu might not be serious in some countries but in some parts of the globe, it has proven to be an epidemic. Breakout of any pandemic or epidemic for that matter ensues unseen chaos owing to loss of life. The spread of disease differs based on various factors like the transmission type, climate conditions, geographic conditions, any existing treatments and resources, and the severity of the virus. Our paper makes note of the transmission type and the severity of the virus as a deciding factor on which model is efficient.

Owing to the rise in the usage of social media data, it provides for a huge volume of data and a variety of information. As per our analysis, social media apps like Twitter and Sina Weibo, and queries from search engines like Google Trends and Baidu Index have been used extensively to extract and study data. As India went through the second wave of Covid-19, the velocity of the generated tweets for resources like hospital beds, oxygen support, ventilators, and other life-necessary essentials increased tremendously helping the government and the citizens to provide on-ground support to each other. Data poured in from other sources like Facebook, Instagram, and news articles were processed by the prediction algorithms.

We hope our paper will be a useful guide for future researchers for reference.

II. LITERATURE SURVEY

Several types of research have been carried out to predict and discover other connected factors in the spread of epidemics and more recently the Covid-19 pandemic. We have observed and studied a few of the models and summarized them below.

Sujin Bae et al [1] implemented adjusted SEIR models by extracting features of covid 19 pandemic statements from various social media platforms and press releases of the Korea Centers for Disease Control and Prevention using the Monte Carlo methodology to predict social media effects on the spread of Covid-19. The adjusted SEIR model forecasts better results than the legacy SEIR model. Elaine O. Nsoesie et al [2] extracted Google search engine data on symptoms of influenza and natural treatments and applied various Combining machine learning and statistics such as RF Regression, LR, SVM, and ARIMA. A higher average R2 was obtained in Random Forest (0.78) and Support Vector Machines (0.88).

Manuscript received on 27 June 2022 | Revised Manuscript received on 02 July 2022 | Manuscript Accepted on 15 July 2022 | Manuscript published on 30 July 2022.

* Correspondence Author

Mrs. Disha Sushant Wankhede*, Assistant Professor, Department of Computer Science, Vishwakarma Institution of Information Technology, Pune (Maharashtra), India. Email: disha.wankhede@viit.ac.in

Rohan Rajendra Sadawarte, Students, Department of Computer Science, Vishwakarma Institution of Information Technology, Pune (Maharashtra), India

Mahek Ibrahim Mulla, Students, Department of Computer Science, Vishwakarma Institution of Information Technology, Pune (Maharashtra), India

Shreya Rahul Jadhav, Students, Department of Computer Science, Vishwakarma Institution of Information Technology, Pune (Maharashtra), India

Mat´ıas N´u˜nez et al [3] trained neural ordinary differential equation (neural ODE) to predict virus outbreaks utilizing data from Daily surveys on Facebook of Covid-19 symptoms. The model learns from the multivariate time series data to effectively predict new cases about 2 months in advance.

Beakcheol Jang et al [4] extracted Keywords related to the illness from online news articles. Word embedding and sorting keywords using the Pearson correlation coefficient has showcased 0.8705 prediction accuracy.

Samira Yousefinaghani et al [5] extract data from Twitter, Google Trends, and Johns Hopkins COVID-19 data repository wherein US states were divided into two categories. In category 1, states that have experienced initial waves were able to predict 100% of first waves 2-6 days earlier and only 6% of second waves. In category 2, states that haven't experienced the initial wave predict 78% of second waves 1-2 weeks in advance.

Samina Amin et al [6] obtain data from Twitter and proposes machine learning models like Support Vector Machine, K-Nearest Neighbor, Random Forest, and Decision Tree to predict and detect flu outbreaks.Results indicate Random forest outperforms other models with 87% accuracy. All machine learning models showcase results ranging from 84% to 88%.

To create a prediction model for the flu, Ali Wahid et al. [7] employed support vector regression, locally weighted smoothing (LOESS) regression, linear regression, and polynomial regression. By analyzing official US CDC data from tweets for 22 weeks, researchers concluded that there is no correlation between official flu data from the CDC and tweets that mention the illness. Additionally, Machine learning algorithms that forecast flu from official CDC data and tweets with mentions of the illness perform better than conventional forecasting models.

Abir EL Azzaoui et al [8] extracted over 10,000 tweets over 2 months with covid-related hashtags, user location, symptoms, subjectivity, and polarity and applied NLP techniques to filter tweets. The prediction considers susceptible, infected, predicted, removed (dead or recovered), and confirmed cases. Using SNS big data analysis method, the proposed framework predicts outbreak 7 days earlier with 0.989 indicator results.

Loukas Samaras et al [9] obtained data for keywords “measles" in regional languages using google trends and official data from ECDC. Pearson Correlation Coefficient is 0.779 indicating a strong relationship between measles cases and predicted cases, the mean standard error was 12.19% for combined results. Jiachen Sun and Peter A [10]. Used Spearman Correlation on Covid data collected from Twitter and Google Trends. The factors they considered were Population size, density, enplanements, and GDP of the study period. Although prediction wasn't the main focus of this paper, they have concluded that the correlation is high but differs for every state. This difference in correlation may be caused due to demographics like the size of the population, density, air traffic & economic development.

Tejas Shinde et al [11] extracted volumes of pandemic-related data from Twitter. Using NLP techniques and classification algorithms like SVM and Naive Bayes. Pearson correlation coefficient is around 0.8 indicating a positive correlation between daily tweets and confirmed cases.

Erhu Du et al [12] implement a framework for agent-based modeling that combines an epidemiological SEIR model, a behavioral adoption model, and a general opinion dynamics model to simulate linked "opinion-behavior-disease" dynamic processes in a hypothetical affected environment by extracting data from global information, neighbor information, and social media.

Cuilian Li et al [14] Weibo social media platform and Google, Baidu search engines are used to analyze the possibility of predicting COVID-19 outbreak. For the keyword coronavirus, the value of r was found to be 0.958, 0.933, 0.944, and the lag period of 9, 12, and 12 days for Google Trends, Baidu Index, and Sina Weibo Index respectively.

Loukas Samaras et al [15], using Twitter as a social media platform and Google as a search engine, compared Google and Twitter data on Influenza in Europe with ECDC data using the ARIMA model and also with self-built-model. The R-value was found to be 0.933, and 0.943. The Mean absolute percentage error was found to be 21.358, and 18.742 for Google and Twitter respectively.

Cuihua Shen et al [16], Classification of the data with sick posts and further classified as grouped, ungrouped, and tagged from famous social media platform Weibo. For classification of sick posts various machine learning algorithms like Decision tree, Extra tree, Extra trees, K nearest neighbors, Multilayer perceptron, Support vector machine, and random forest have been used. The sick posts are also further classified as ingroup sick posts and outgroup sick posts.

Matteo Cinelli et al [17], have done a comparative analysis of 5 different social media platforms during Covid19. They discuss two forms of information on social media: questionable and reliable. Also, there is an analysis of reactions to these two pieces of information.

Yufang Wang et al [18], PDE i.e., Partial Differential Equation model for prediction of Influenza outbreak with the help of Twitter data is used. Twitter filter streaming API is used to extract tweets which are then classified in "Full Tweets" i.e., Tweets containing keywords flu, influenza, cold, cough, and headache. Almost 90% of accuracy was found with this model and Tweet classification

May Oo Lwin et al [19], Facebook and Twitter as social media platforms, and Google as a search engine is used to graphically represent the correlation between data of Zika virus. It can be easily inferred from the graphs presented that social media data can be useful to predict the outbreak.

Zhenghong Peng et al [20], Weibo as a social media platform and Mobile data for location-related information is used to extract "help-seeking" posts related to covid with age and location. With Ordinal Least Square regression and Kernel Density Analysis, the collected data was found to have a high correlation. The study predicts not only the outbreak but also where the medical supply is needed.

U Venkatesh et al [21], Data from Google Trends, YouTube, and News Articles are used to analyze Covid trends in India. The correlation (r) of the keyword “coronavirus” for Google Web search was 0.82 and for YouTube search was 0.82. Google Trends keyword trend with an actual number of cases from officials found to be with an almost accurate 2-3 weeks earlier (lag period) prediction. Jiawei Li et al [22], Weibo Post data and with the help of Linear regression model according to this study, there is a direct association between the quantity of Weibo posts and the number of Chinese instances that are reported. Weibo postings are categorized into one of two categories using qualitative analysis. 2. The outbreak's evolving epidemiological characteristics 3. Public response to epidemic prevention 4. reaction metrics and related subjects. A slight difference was found in the R2 score for Hubei and outside Hubei, R2=0.621 - Hubei, R2=0.652 - other than Hubei (P< 0.01). J. Sooknanan et al [23], discuss all the mathematical models used before for the prediction of a Disease Also, for incorporating social media (Twitter) data, they used two mathematical models, and simply found a strong correlation between the actual data and the predicted data. They have used Twitter data for their analysis and prediction. Lei Qin et al [24], Baidu Search index is used along with one mathematical model to predict the Covid-19 outbreak. Five parameter estimation methods were adopted viz. 1. subset selection 2. forward selection 3. ridge regression 4. lasso regression 5. elastic net with RMSE (Root Mean Square Error) of 51.6671, 70.0168, 415.2922, 519.7440, 510 - 530 and with R-value as 0.9963, 0.9913, 0.6937, 0.4858, 0.48-0.50 respectively. A lag period of 10 days was also found with model prediction and the actual number of cases. Aditya Joshi et al [25], Twitter data is used for early detection of acute disease named Thunderstorm Asthma in Melbourne, Australia. This study was performed in 4 step process, 1. Initial Selection 3. Duplication Removal 4. Time Between Events–based Monitoring Algorithm. They performed 18 experiments out of which three of them were able to spot the thunderstorm asthma outbreak up to nine hours before the period specified in the official report, while five others were able to spot it before the initial news story. Ali Alessa et al [26], Twitter data and data from different articles were collected to predict Influenza outbreaks. They used many machine learning and Artificial Intelligence models such as Neural networks, SVM, Naive Bayes, Agent-Based Modeling, etc. It was an overall review of many ML and AI algorithms for the prediction of the flu and mentioned the Pearson correlation, RMSE, and MAE of each model used. Madhur Verma et al [27], Google trends data is extracted from Haryana and Chandigarh, states of India of outbreaks like Chikungunya, Malaria, Dengue, and Enteric Fever. They have found different r values between google trends and cases of different outbreaks. With the help of plots, they have also studied different lag periods for different outbreaks. Hyekyung Woo et al [28], Daum and Naver's Blog, and Twitter data were used for the prediction of Influenza in Korea. They used Lasso and SVM models and found a very high correlation with r = 0.965

On Twitter and Google data, Soheila Molaei et al. [29] used nonlinear methods such as the number of patients seen by doctors has been predicted using (ARX), (ARMAX), (NARX), (DeepMLP), and (CNN). Among the suggested solutions, the DNN and the entropy-based method generated the best results, lowering the mean average error by as much as 25% and minimizing time complexity. Yuzhou Zhang et al. [30] a multivariate seasonal autoregressive integrated moving average model combining Australian influenza and local search data utilizing data from Google Trends and Baidu Index to track influenza epidemics. Flu outbreaks occurred in Australia, the US, the UK, and China between August and October, whereas they peaked in the US between January and March. Nduwayezu, Maurice et al [31] in their paper about Modeling and Prediction of Nigeria's Malaria Epidemic Using Twitter Data and Precipitation Volume have used various Natural Language Processing toolkits to clean the sentences i.e., the tweets collected through Twitter API. They cleaned and narrowed it down to 12425 tweets ready to be preprocessed. The only limitation found was that their dataset wasn't very comprehensive. Laboratory-confirmed H7N9 cases in humans, as well as data from the Weibo Posting Index (WPI) and Baidu Search Index (BSI) that were relevant to H7N9, were used by Ying Chen et al. [32] and were gathered weekly. Predicting the number of H7N9 cases using seasonal autoregressive integrated moving average (SARIMA) models with BSI and WPI. Anahita Safarishahrbijari et al [33] used a previously made influenza A (H1N1) model pandemic projection. Balsam Alkouz et al [34], designed Tweetfluenza that filters influenza tweets and classifies them into two classes, ones that are to be reported and those that aren’t. They used a linear regression model to predict the influenza-based tweets in the future and studied its feasibility. A new deep learning structure that predictions the real-time influenza-like illness rate (ILI percent) in Guangzhou, China, is proposed by Xiagnlei Zhu et al. [35] using Twitter and Google Correlate of 9 years of data. Long short-term memory (LSTM) neural networks have been used to precisely anticipate accuracy due to the long-term attribute and variety of influenza epidemic data. The model takes into account all the data in the set, focusing on resolving the practical issues of predicting the Guangzhou influenza outbreak. Yuvaana Sundarakrishnan et al [36] made an Epidemic Prediction model by analyzing the spread of the disease through the population. They have made use of artificial neural networks and other ML techniques to predict the spread in particular geographical areas. They have used FluNet (WHO-based influenza surveillance). Aditya Lia Ramadona et al. [37] used Twitter data to create an algorithm to calculate a dynamic mobility-weighted index (MI), which measures the neighborhood's exposure to viral importation. The coefficients and predictiveness of the MI index for lags up to 6 months were determined using a Bayesian Spatio-temporal regression model. The study area's monthly incidence of dengue was seen to steadily grow from December to March of the following year, then decline until the beginning of the rainy season in October. Over the course of the study period, the incidence of dengue was generally declining.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Bijaya Adhikari et al [38], designed a deep neural network with data from US National Centers for Disease Control (CDC), to predict an epidemic through representations of relevant curves in a continuous feature space. They also predicted any future incidences, when will the intensity of the epidemic be at its peak, and the onset of the upcoming wave. EpiDeep learning features Four epidemic forecasting tasks were accomplished by combining historical epidemic seasons with the observed current season. Shoko Wakamiya et al [39], In order to improve the effectiveness and efficiency of the current influenza surveillance and reporting systems, a new tool for predicting and reporting national influenza trends has been developed. By calculating the Pearson correlation coefficients, they were able to provide influenza epidemic results to hospital control practitioners. They abstracted their data from the Taiwan Centers for Disease Control and Prevention (TWCDC). Gaoyang Liu et al [40], made software by the name of SocInf, which makes a mimic model which functions in close relation to a public model. It is trained on machine learning models Xgboost, logistics, and online cloud platforms. Their mimic model was able to achieve an accuracy of 73% and predict whether a given record is present in the training dataset. Wankhede et al [41],[42],[43],[44] Has focused on various Machine learning, Deep learning algorithm used in healthcare domain work. The author mentioned all existing algorithms used in this domain. Gaikwad V. [45][46] Light-Weight Key Establishment Mechanism for Secure Communication Between IoT Devices and Cloud. Predict Socio-Economic Status of an Area from Satellite Image Using Deep Learning Trupti Vasantrao Bhandare et al [47] , [48] proposed and implemented a novel approach to Heart Disease Diagnosis Using Deep Learning Methods.

III. SOCIAL MEDIA AND SEARCH ENGINES

Various Social Media platforms have been used during the research carried out and have majorly varied depending on the basis of the location the research was conducted in. In this paper, we are going to summarize the Social Media platforms that have been used and how or not they were beneficial to the prediction of outbreaks. The most widely used social media platform that has been used is Twitter. More than 25% of the papers we surveyed have predominantly used it as the official source to get data. Not only is the data textual making it easier to extract keywords but can also be easily searched due to a robust searching algorithm. The tweets generated are region-specific helping the model be more efficient in the regional conditions. The trending tweets highlight and bring out any recent happenings going on in the world and the specific location. Any tweet in any location and tweeted at any time that uses a trending hashtag can get further publicly acknowledged. Unlike locally used search engines and region-specific social media platforms, Twitter is for the most part used across the globe. All these characteristics chalk up to Twitter being reliable.

As Google has more than 3 billion searches in just one single day and has a 92% share of the search engine market, we found many research works based on this search engine in our survey. The main feature of this search engine that helps researchers the most is Google Trends. Google trends provide a dashboard that includes all the data visualized not only in the form of tables or graphs but also in the form of interactive maps. Researchers can analyze and get the data with its chief features like keyword searches, regional searches, variation in trends by locality, and even distinction in searching patterns by locality. While numerous search engines like Mozilla, Yahoo, Baidu, etc. are present in the market, Google is the most widely used search engine and because of its advantageous characteristics, researcher's majority prefer Google to extract the data from the search engine.

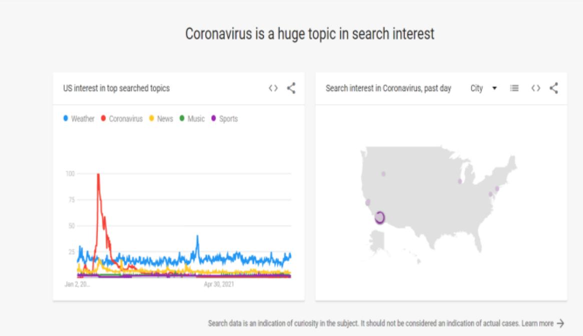
Coronavirus is a huge topic in search interest
US interest in top searched topics
Weather
Coronavirus
News
Music
Sports
100
75
50
25
Jan 2, 20...
Apr 30, 2021
Search interest in Coronavirus, past day
City
Search data is an indication of curiosity in the subject. It should not be considered an indication of actual cases. Learn more →

Fig. 1 - Keyword and Search Analysis from Google Trends

Sina Weibo, also called Facebook and Twitter of China is a microblogging platform that is widely used in China. Weibo offers Weibo API to extract data which has helped many research works based in China. As China is the epicenter for the Covid-19 outbreak, almost all the research works in our survey, which is based on China, have considered Weibo's data to predict outbreaks like covid. Not only China-specific is the limitation for Weibo's data to predict outbreaks, but the Chinese government policy of internet censorship can also affect prediction models. China's leading internet search provider is Baidu. Its features and offerings are comparable to those of Google, but it places a strong emphasis on China, where it dominates the search industry. Baidu was included in this survey owing to the fact that our paper spans 3 years of papers related to epidemics and pandemics - 2019, 2020, and 2021. China being the epicenter of Covid-19, Baidu became used extensively to gather data and track keywords and the latest news regarding the outbreak and other details. Facebook was used in less than 5% of all the papers used for the survey. Facebook isn't used as widely as Twitter or Google Trends to report the latest ongoings. Any data that is collected is a small amount and does not give a good analysis. The data collected from it is mostly in the image and video format which can't be used for extraction of the actual data that efficiently.

Method/Algorithm

The aim of this study is to investigate the Epidemic Outbreak from Social Media Data classifying techniques using peer-reviewed research papers published between 2015 and 2021 in Scopus and Web of Science indexed journals.

The databases that are extensively searched for this survey work were: (1) IEEE Xplore Digital Library, (2) Science Direct, (3) PubMed, (4) Google Scholar, (5) MDPI and (6) ResearchGate. The search criterion includes social media ∧ Pandemic ∧ Epidemic ∧ Outbreak ∧ Covid-19 ∧ Influenza ∧ Regression Models ∧ Machine Learning ∧ Artificial Intelligence. Algorithm 1 clearly reveals the method used for deciding which literature to review, also there are inclusion criteria (IC) and exclusion criteria (EC) for the paper as indicated in Table 2.

Considering Inclusion Criteria (CIC)	Considering Inclusion Criteria (CEC)
CIC1: There must be a peer review of the paper	CEC1: Different databases contain duplicate studies.
CIC2: Scopus or Web of Science indexing is required for journals that publish papers	CEC2: Study that uses different techniques/algorithms other than mentioned social media.
CIC3: The paper should use only Pandemic, Epidemic, Outbreak, Covid-19, Influenza	CEC3: Peer-reviewed paper that is less frequently cited.
	CEC4: MSc and Ph.D. papers.
	CEC5: Case study papers

Algorithm 1 Searching different databases for papers is a good strategy related to the Research Prediction of Epidemic Outbreak from Social Media Data.

```
procedure TOPIC (Research Prediction of Epidemic Outbreak from Social Media Data)
SearchDatabases ← IEEEX explore, Google Scholar, Science Direct, PubMed, MDPI
SearchYear ← 2015 − 2022 ∧ A few old papers are exceptionally valuable to enrich our understanding Section 1
  i ← 1 // Initialize counter
    N ← 5 // N is the number of search databases
        for i ≤ N do //Using for loop to traverse searches
            Keyword ← Social_Media, Pandemic, Epidemic, Outbreak, Covid-19, Influenza,
                Regression Models, Machine Learning, Artificial Intelligence.
                            // Keywords for search
                if Search_Link ∈ Search_Databases and Year ∈ Search_Year then
                    Search (Social_Media ∧ Pandemic ∧ Epidemic ∧ Outbreak ∧
                        Covid-19 ∧ Influenza ∧ Regression Models ∧ Machine
                        Learning ∧ Artificial Intelligence)
                            // If condition matches then
                        searching Started
                end if
        end for
        if Number of Papers/articles ≥ 0 then
            Refine Papers
            ConsideringInclusionCriteria ← IC1, IC2, IC3, IC4
            ConsideringExclusionCriteria ← EC1, EC2, EC3, EC4, EC5, EC6
        end if
end procedure
```

IV. ALGORITHMS

Sr. No.	Algorithm	Data Source and size	Accuracy / Output Extracted
1	Regression Models		
	Linear Regression	Twitter, US CDC official data 5.5 Months Data [7]	There were two types of tweets: self-reported and non-self-reported. RMSE - 1172.2 R squared- 0.81
		Google Trends and European Center of Disease and Prevention - ECDC [13] 60 Months Data [9]	Used "measles" keyword in different languages for different countries R - 0.779 p < 0.01 MSE - 45.2 (12.19%)
		Weibo and Chinese National Health Commission 1 Month Data (January 2020) 115,299 Weibo posts[22]	Conclusions were in two parts: Hubei and other than Hubei P<.001; R2=0.621 - Hubei P<.001; R2=0.652 - other than Hubei
		The Arabic Tweets were - 518558, English Tweets - 571126 [34]	RMSE Number of Tweets: Nov: 0.576 Jan: 0.362 Number of Hospital Visits: Nov: 0.74 Jan: 0.54
	Lasso Regression	Google search queries, Twitter, HealthMap, Pan American Health Organization (PAHO) 9 Months Data [13]	Selection of terms with high predictive power via LASSO Country-wise prediction for 1-, 2- and 3- weeks. Colombia 1-week: RMSE: 617.795, r: 29.888, ρ (Pearson correlation): 0.871

		Baidu, National Health Commission (NHC) 1.3 Months Data [24]	RMSE: 519.7440 MAE (mean absolute error): 358.0978 0.1032 is the mean absolute percentage error (MAPE). 0.9597 for the Pearson correlation
2	Monte Carlo	Korea Centers for Disease Control and Prevention, BigKinds 1.5 Months Data [1]	Position(verified) and negative(non-verified/fake) function, two motives of social media. The adjusted SEIR model has double the accuracy as the legacy model MSE: 359670.9
3	SEIR	BigKinds, a search site for news databases, provided the social media data, press releases Korea Centers for Disease Control and Prevention [1], Around 2000 posts for a period of 1.5 months	Adjusted SEIR model had double accuracy than the legacy model
		Twitter, Instagram, YouTube, Reddit, Gab and Google Trends [17]	E as the average number of reactions to a post; ER (for reliable posts) EU (for unreliable posts) α=EU/ER Twitter α~1
4	ARIMA	Search Data | Google [2],	There was no accuracy since the paper predicted and compared the spread of the disease region wise.
		Google Trends, European Center of Disease and Prevention - ECDC [9]	correlation coefficient R= 0.779 in two-tailed significance p< 0.01 [13]
		It is coordinated by the European Center for Disease Control and Prevention (ECDC).23 weeks data from December 13, 2018, until May 20, 2019.[15]	Spearman Correlation = 0.933 [23].
		Baidu Search Index (BSI) and Weibo Posting Index (WPI) data, Data collected from 2013 to 2017, [32]	BSI (β = 0.008, p < 0.001) WPI (β = 0.002, p = 0.036)
6	Neural Networks	Twitter Data of UK 7 Months [37]	Accuracy: 0.9532
		Twitter and Google Data 5 years data: Comparison between successive years [43]	CNN (Convolutions Neural Network) (2009-10) RMSE: 3.69 MAE 18.06 (2011-14) RMSE: 4.05 MAE 6.04
		Data from The US national Centers for Disease Control and Prevention (CDC) and taken over the span of 7 years Data (2010-2017) [57]	LSTM - Future Incidence Prediction for the Year 2016 - 2017 RMSE: 0.98 MAPE: 0.31
7	SVM	Google Search Query 72 months Data [4]	R2: 0.877 RMSE: 1.078
		Twitter - 6000 sample set 24 Months Data [9]	Positive tweets - symptoms Negative tweets - related information but not symptoms. Precision: 0.87 Recall: 0.85 F1- score: 0.84 Accuracy: 0.85
		Twitter - 5000 tweets [15]	Precision: 91.3621 Recall: 96.042 F1- score: 93.64 Accuracy: 91.66
		Twitter, Internet Articles [37]	Pearson correlation (r): 0.93
		[40]	Optimal feature selection by LASSO r=.956; P<.001
		Twitter 6 Months Data [48]	Relationship: 0.75 Exactness: 0.985 Score: 0.9737 Recall: 0.963 F1- Reliability: 0.96511
8	Decision Tree	Data Set | Twitter 6000 tweets [9]	Accuracy - 0.84 (test)

		China CDC, Weibo User Pool [24]	Accuracy - 0.830
9	Random Forest	Twitter Data 6000 tweets [9]	Tweets were classified as positive and negative Positive: about symptoms. Negative: related information but not about symptoms Accuracy: 0.84
		Weibo User pool, CDC data China 250 million users [24]	Sick Posts: Posts that report any symptoms or diagnoses that are likely related to COVID-19 1. ingroup sick posts - user's own or immediate family members 2. outgroup sick posts - people not in the user's immediate family F1 Score: 0.835 Precision: 0.840 Accuracy: 0.830 Recall: 0.830
10	Naive Bayes	Twitter Data 5000 Tweets [15]	Organizing keywords into five categories: 1. Only tweets relating to coronavirus 2. arbitrary tweets 3. Tweets about symptoms 4. Tweets about new cases 5. Individual tweets relating to COVID-19 Correlation: 0.8
		Tweeter 1 month data (Oct 2015 - Nov 2015) [37]	The accuracy of the used method is determined by the sentiment polarity (Naive Bayes polarity is 70%)

Statistics from Above Survey

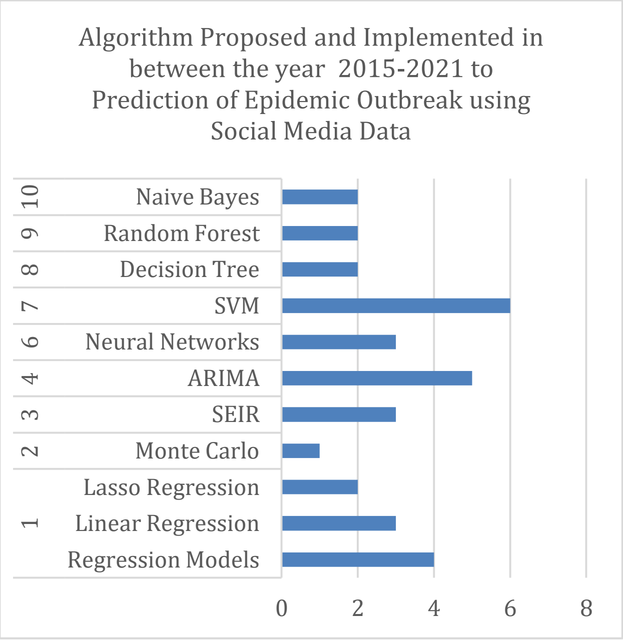

Fig 2 - Algorithm Used in Various Surveys

V. DISEASES AND OUTBREAK FEATURES

The most widespread and deadly pandemic after the Spanish Influenza happens to be the novel Coronavirus (Covid-19) pandemic. Throughout the papers that we surveyed, many researchers across the globe have developed and implemented various algorithms and models to track the outbreak characteristics of the pandemic. The Coronavirus Disease (COVID-19) is caused by the SARS-CoV-2 virus.

Most people suffering from Covid-19 will only experience mild to moderate respiratory symptoms and won't need medical attention. According to research, the symptoms begin to manifest 14 days into the incubation phase. However, some people will experience a serious illness and need medical attention. The elderly and those with underlying medical conditions including cancer, diabetes, chronic lung disease, or cardiovascular disease are more likely to experience serious illness. Any person can become unwell from COVID-19 and may do so at any age, becoming seriously ill or passing away. Reportedly, 3.4% of the people that contacted Covid-19, have died. There have been 2-3 variants that followed the first wave and are still being mutated into existence. The majority of the research conducted and algorithms made revolved around keyword extraction of words like "Coronavirus", "Covid-19", and "Pandemic". There have also been social media posts aimed at seeking help and resources, such as when India went through the second wave of the pandemic, there was a shortage in resources and such posts helped get the necessary items to the masses. Taking note of the next epidemic that we surveyed, which was influenza we observed that, in the last 140 years, there have been six significant influenza outbreaks, the most severe of which was the 1918 flu pandemic, which is estimated to have killed 50–100 million people. The most recent pandemic, the 2009 swine flu pandemic, killed less than 300,000 people and was considered mild. Malaria is a fever illness caused by Plasmodium parasites that are transmitted to humans by mosquito bites from infected female Anopheles mosquitoes. Human malaria is caused by five parasitic species, two of which – P. falciparum and P. vivax – are the most dangerous. The malaria parasite Plasmodium falciparum is the deadliest and most common on the African continent. In most places outside of Sub-Saharan Africa, P. vivax is the most common malaria parasite. The first signs of malaria, such as fever, headache, and chills, come 10–15 days after the infective mosquito bite and might be mild and difficult to distinguish from other illnesses. P. falciparum malaria can escalate to severe sickness and death in as little as 24 hours if left untreated.

Nearly half of the world's population would be in danger of malaria by 2020. Infants, young children, pregnant women, persons with HIV/AIDS, and those traveling to regions where malaria transmission is high, such as migrant workers, mobile populations, and travelers, are at a notably increased risk of catching malaria and suffering severe illness. In this survey, a paper by Nduwayezu, Maurice, et al [31], used an SVM algorithm to collect live tweets from Nigerian Twitter about precipitation and then used it to define a correlation between precipitation and the spread of malaria. Another disease that caused an epidemic was the Zika virus. Zika fever spread from Brazil to other American countries from April 2015 to November 2016. Between October 2015 and January 2016, the Zika virus infected 1.5 million people. Other portions of South and North America, as well as some Pacific islands, were afflicted by the outbreak. In 2013 or 2014, the Zika virus traveled from Oceania to Brazil. In March 2015, Brazil notified the WHO of a skin rash-like condition, and Zika was established as the cause in May 2015. As evidence accumulated that Zika might cause birth abnormalities and neurological issues, the WHO labeled the outbreak a Public Health Emergency of International Concern in February 2016. The virus can be passed from a pregnant woman to her unborn child, causing microcephaly and other severe brain abnormalities. Adults who contract Zika can develop Guillain–Barré syndrome. In about one out of every five occurrences, the Zika virus causes Zika fever, a mild sickness characterized by fever as well as a rash. Prior to the outbreak, Zika was thought to be a mild infection because most infections remain asymptomatic, making precise estimates of the number of patients impossible. In a paper by Sarah F. McGough et al [13], an attempt has been made to make predictions about the number of cases using Pearson's correlation and LASSO regression on data from Google Trends, etc. The quality of prediction was not good in the situation of fewer cases, but will nevertheless help in making better and informed decisions and can be used to make better predictions in the future, Fig 3 Shows sample architecture for Health Predictive Analysis

VI. PROPOSED ARCHITECTURE /BLOCK DIAGRAM

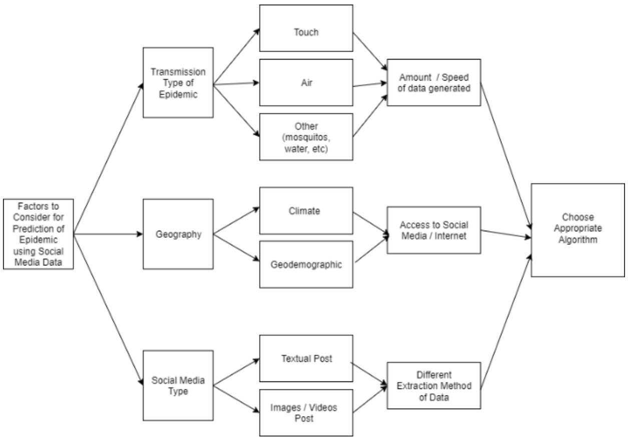
Touch
Transmission Type of Epidemic
Air
Amount / Speed of data generated
Other (mosquitos, water, etc)
Factors to Consider for Prediction of Epidemic using Social Media Data
Geography
Climate
Geodemographic
Access to Social Media / Internet
Choose Appropriate Algorithm
Social Media Type
Textual Post
Images / Videos Post
Different Extraction Method of Data

Fig 3 – Prediction of an Epidemic through Social Media Data Architecture

Figure 3 displays the entire structure mentioned above for the final Prediction of Epidemic Outbreak using Social Media Data. It is clear that the overall framework is divided into four distinct analytical modules. The data set has been taken into consideration for social media data, and after that, a descriptive analysis is carried out, which entails the examination of general data regarding various features of the data set. Utilizing data visualization techniques and algorithms, the diagnostic analysis examines the relationships between various parameters.

VII. CONCLUSION

This article provides a discussion of methods for predicting epidemic disease outbreaks that depend on web-based social info from the media. Writing that has been analyzed generally be justified in arguing that various forecasting models can anticipate the outbreaks of infection, however, these models rely merely on the diseases that can be found in social networks. Different mining techniques can be used to tap into the channel. paired with speedier and more often occurring input to produce precise identification of the disease outbreaks. In order to construct a reliable forecast for the risk of epidemic or infection outbreak, a few factors might be identified as key components. Prediction accuracy, as well as its land scale and duration, are crucial among these. Based on the survey, we've come to the conclusion that, when compared to traditional methods, machine learning performs better at predicting the onset of epidemics. Because of this, these areas need additional research and planning.

REFERENCES

1. Sujin Bae, Eunyoung (Christine) Sung & Ohbyung Kwon (2021), Accounting for social media effects to improve the accuracy of infection models: combatting the COVID-19 pandemic and infodemic, European Journal of Information Systems
2. Elaine O. Nsoesie1*, Olubusola Oladeji1, Aristide S.AbahAbah 2 & Martial L. Ndefo-Mbah(2021), Forecasting influenza-like illness trends in Cameroon using Google Search Data, Scientific Reports [CrossRef]
3. Mat´ıas N´u˜nez,1, 2, 3 Nadia L. Barreiro,4 Rafael A. Barrio,5 and Christopher Rackauckas (2021), Forecasting virus outbreaks with social media data via neural ordinary differential equations, medRxiv
4. Beakcheol Jang, Ph.D.; Inhwan Kim, BSc; Jong Wook Kim2, Ph.D. (2021), Effective Training Data Extraction Method to Improve Influenza Outbreak Prediction from Online News Articles: Deep Learning Model Study, Jmir Medical Informatics
5. Samira Yousefinaghani, Rozita Dara, Samira Mubareka and Shayan Sharif 3 (2021), Prediction of COVID-19 Waves Using social media and Google Search: A Case Study of the US and Canada, Frontiers [CrossRef]
6. Samina Amin, Muhammad Irfan Uddin, Duaa H. alSaeed, Atif Khan, and Muhammad Adnan (2021), Early Detection of Seasonal Outbreaks from Twitter Data Using Machine Learning Approaches, Hindawi" [CrossRef]
7. Ali Wahid, Steven H. Munkeby, Samuel Sambasivam (2021), MACHINE LEARNING-BASED FLU FORECASTING STUDY USING THE OFFICIAL DATA FROM THE CENTERS FOR DISEASE CONTROL AND PREVENTION AND TWITTER DATA, Issues in Informing Science+Information Technology" [CrossRef]
8. Abir EL Azzaoui, Sushil Kumar Singh, Jong Hyuk Park (2021), SNS Big Data Analysis Framework for COVID-19 Outbreak Prediction in Smart Healthy City, Elsevier – ScienceDirect [CrossRef]
9. Loukas Samaras*, Miguel-Angel Sicilia and Elena García-Barriocanal(2021), Predicting epidemics using search engine data: a comparative study on measles in the largest countries of Europe, BMC Public health [CrossRef]

10. Jiachen Sun and Peter A. Gloor1(2021), Assessing the Predictive Power of Online social media to Analyze COVID-19 Outbreaks in the 50 U.S. States, Preprints
11. Tejas Shinde; Parikshit Thatte; Sachin Sachdev; Vidya Pujari (2021), Monitoring of Epidemic Outbreaks Using Social Media Data, IEEE [CrossRef]
12. Erhu Du, Eddie Chen, Ji Liub, Chunmiao Zheng (2021), How do social media and individual behaviors affect epidemic transmission and contro, Elsevier – ScienceDirect
13. Sarah F. McGough1,2,3*, John S. Brownstein2,3,4, Jared B. Hawkins2,3,4, Mauricio Santillana (2017), Forecasting Zika Incidence in the 2016 Latin America Outbreak Combining Traditional Disease Surveillance with Search, social media, and News Report Data, Plos: neglected tropical diseases
14. Cuilian Li1, Li Jia Chen2, Xueyu Chen1, Mingzhi Zhang1, Chi Pui Pang1,2, Haoyu Chen (2020), Retrospective analysis of the possibility of predicting the COVID-19 outbreak from Internet searches and social media data, China, 2020, Eurosurveillance
15. Loukas Samaras, Elena García-Barriocanal, Miguel-Angel Sicilia (2020), Comparing social media and Google to detect and predict severe epidemics, nature - Scientific Reports [CrossRef]
16. Cuihua Shen, Anfan Chen, Chen Luo, Jingwen Zhang, Bo Feng, Wang Liao (2020), Using Reports of Symptoms and Diagnoses on social media to Predict COVID-19 Case Counts in Mainland China: Observational Infoveillance Study, JMIR Publications
17. Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo Antonio Scala (2020), The COVID-19 social media infodemic, nature - Scientific Reports [CrossRef]
18. Yufang Wang, Kuai Xu, Yun Kang, Haiyan Wang, Feng Wang, Adrian Avram (2020), Regional Influenza Prediction with Sampling Twitter Data and PDE Model, MDPI - International Journal of Environmental Research and Public Health [CrossRef]
19. May Oo Lwin, Jiahui Lu, Anita Sheldenkar, Ysa Marie Cayabyab, Andrew Zi Han Yee, Helen Elizabeth Smith (2020), Temporal and textual analysis of social media on collective discourses during the Zika virus pandemic, Springer
20. Zhenghong Peng, Ru Wang, Lingbo Liu, Hao Wu (2020), Exploring Urban Spatial Features of COVID-19 Transmission in Wuhan Based on Social Media Data, MDPI - International Journal of Environmental Research and Public Health" [CrossRef]
21. U Venkatesh, Periyasamy Aravind Gandhi (2020), Prediction of COVID-19 Outbreaks Using Google Trends in India: A Retrospective Analysis, Synapse" [CrossRef]
22. Jiawei Li, Qing Xu, Raphael Cuomo, Vidya Purushothaman, Tim Mackey (2020), Data Mining and Content Analysis of the Chinese Social Media Platform Weibo During the Early COVID-19 Outbreak: Retrospective Observational Infoveillance Study, JMIR Publications"
23. J. Sooknanan, D. M. G. Comissiong(2020), Trending on Social Media: Integrating Social Media into Infectious Disease Dynamics, Springer [CrossRef]
24. Lei Qin, Qiang Sun, Yidan Wang, Ke-Fei Wu, Mingchih Chen, Ben-Chang Shia, Szu-Yuan Wu (2020), Prediction of Number of Cases of 2019 Novel Coronavirus (COVID-19) Using Social Media Search Index, MDPI - International Journal of Environmental Research and Public Health"
25. Aditya Joshi, Ross Sparks, James McHugh, Sarvnaz Karimi, Cecile Paris, Raina MacIntyre (2020), Harnessing Tweets for Early Detection of an Acute Disease Event, NCBI" [CrossRef]
26. Ali Alessa, Miad Faezipour (2018), A review of influenza detection and prediction through social networking sites, BMC [CrossRef]
27. Madhur Verma, Kamal Kishore, Mukesh Kumar, Aparajita Ravi Sondh, Gaurav Aggarwal, Soundappan Kathirvel(2018), Google Search Trends Predicting Disease Outbreaks: An Analysis from India, HIR - Healthcare Informatics Research" [CrossRef]
28. Hyekyung Woo, Youngtae Cho, Eunyoung Shim, Jong-Koo Lee, Chang-Gun Lee, Seong Hwan Kim (2016), Estimating Influenza Outbreaks Using Both Search Engine Query Data and Social Media Data in South Korea, JMIR Publications"
29. Soheila Molaei, Mohammad Khansari, Hadi Veisi & Mostafa Salehi (2019), Predicting the spread of influenza epidemics by analyzing Twitter messages, Springer [CrossRef]
30. Yuzhou Zhang, Laith Yakob, Michael B. Bonsall & Wenbiao Hu (2019), Predicting seasonal influenza epidemics using cross-hemisphere influenza surveillance data and local internet query data, Scientific Reports [CrossRef]
31. Nduwayezu, Maurice; Satyabrata, Aicha; Han Suk-Young; Kim Jung Eon; Kim Hoon; Park Junseok; Hwang Won-Joo (2019), Malaria Epidemic Prediction Model by Using Twitter Data and Precipitation Volume in Nigeria, Journal of Korea Multimedia Society
32. Ying Chen, Yuzhou Zhang, Zhiwei Xu, Xuanzhuo Wang, Jiahai Lu & Wenbiao Hu (2019), Avian Influenza A (H7N9) and related Internet search query data in China, Scientific Reports [CrossRef]
33. Anahita Safarishahrbijari; Nathaniel D Osgood (2019), Social Media Surveillance for Outbreak Projection via Transmission Models: Longitudinal Observational Study, JMIR Publications [CrossRef]
34. Balsam Alkouz; Zaher Al Aghbari; Jemal Hussien Abawajy (2019), Tweetfluenza: Predicting flu trends from Twitter data, IEEE Explore [CrossRef]
35. Xianglei Zhu, Bofeng Fu, Yaodong Yang, Yu Ma, Jianye Hao, Siqi Chen, Shuang Liu, Tiegang Li, Sen Liu, Weiming Guo & Zhenyu Liao (2019), Attention-based recurrent neural network for influenza epidemic prediction, BMC Bioinformatics
36. Yuvaana Sundarakrishnan, Akshay George Koshy, K.P. Vijayakumar (2019), Epidemic Prediction, International Research Journal of Engineering and Technology (IRJET)
37. Aditya Lia Ramadona, Yesim Tozan, Lutfan Lazuardi, Joacim Rocklöv (2019), A combination of incidence data and mobility proxies from social media predict the intra-urban spread of dengue in Yogyakarta, Indonesia, PLOS [CrossRef]
38. Bijaya Adhikari, Xinfeng Xu, Naren Ramakrishnan and B. Aditya Prakash (2019), EpiDeep: Exploiting Embeddings for Epidemic Forecasting, ACM Digital Library [CrossRef]
39. Shoko Wakamiya, Mizuki Morita, Yoshinobu Kano, Tomoko Ohkuma, Eiji Aramaki(2019), Tweet Classification Toward Twitter-Based Disease Surveillance: New Data, Methods, and Evaluations, JMIR Publications [CrossRef]
40. Gaoyang Liu, Chen Wang, Kai Peng, Haojun Huang, Yutong Li, and Wenqing Cheng (2019), SocInf: Membership Inference Attacks on Social Media Health Data with Machine Learning, IEEE Explore
41. Mrs. Disha Sushant Wankhede, Dr. Selvarani Rangasamy,"REVIEW ON DEEP LEARNING APPROACH FOR BRAIN TUMOR GLIOMA ANALYSIS" Journal of Information Technology in Industry, VOL. 9 NO. 1 (2021) pp. 395 - 408, DOI: https://doi.org/10.17762/itii.v9i1.144 [CrossRef]
42. Disha Sushant Wankhede, R. Selvarani, Dynamic architecture based deep learning approach for glioblastoma brain tumor survival prediction, Neuroscience Informatics, Volume 2, Issue 4, 2022, 100062, ISSN 2772-5286, https://doi.org/10.1016/j.neuri.2022.100062. (https://www.sciencedirect.com/science/article/pii/S2772528622000243) [CrossRef]
43. Wankhede, D.S., Pandit, S., Metangale, N., Patre, R., Kulkarni, S., Minaj, K.A. (2022). Survey on Analyzing Tongue Images to Predict the Organ Affected. Hybrid Intelligent Systems. HIS 2021. Lecture Notes in Networks and Systems, vol 420. Springer, Cham. https://doi.org/10.1007/978-3-030-96305-7_56 [CrossRef]
44. "Artificial Intelligence and its Subsets: Machine Learning and its Algorithms, Deep Learning, and their Future Trends ", International Journal of Emerging Technologies and Innovative Research (www.jetir.org | UGC and ISSN Approved), ISSN:2349-5162, Vol.9, Issue 5, page no. ppi112-i117, May-2022, Available at: http://www.jetir.org/papers/JETIR2205914.pdf
45. Shetty, A. Thorat, R. Singru, M. Shigawan, and V. Gaikwad, "Predict Socio-Economic Status of an Area from Satellite Image Using Deep Learning," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 2020, pp. 177-182, doi: 10.1109/ICESC48915.2020.9155696. [CrossRef]
46. Gudapati S.P., Gaikwad V. (2021) Light-Weight Key Establishment Mechanism for Secure Communication Between IoT Devices and Cloud. In: Satapathy S., Bhateja V., Janakiramaiah B., Chen YW. (eds) Intelligent System Design. Advances in Intelligent Systems and Computing, vol 1171. Springer, Singapore. https://doi.org/10.1007/978-981-15-5400-1_55 [CrossRef]
47. Trupti Vasantrao Bhandare, & Selvarani Rangasamy. (2021). Review on Heart Disease Diagnosis Using Deep Learning Methods. International Journal of Next-Generation Computing, 12(2), 91–102. https://doi.org/10.47164/ijngc.v12i2.206 [CrossRef]
48. Vasantrao, BhandareTrupti & Rangasamy, Selvarani. (2021). Weighted Clustering for Deep Learning Approach in Heart Disease Diagnosis. International Journal of Advanced Computer Science and Applications. 12. 10.14569/IJACSA.2021.0120944. [CrossRef]

AUTHORS PROFILE

Mrs. Disha S. Wankhede, Assistant Professor VIIT Pune, Enthusiastic ML engineer eager to contribute to team success through hard work, attention to detail, and excellent organizational skills. Motivated to learn, grow, and excel in machine learning and data analysis. Strong education professional with a Master's Degree focused on Computer Science from RGPV University and Bachelor's Degree in Computer Science and Engg. from SGBAU Amravati. Currently pursuing PhD in Artificial intelligence from Alliance University Bangalore. The research area includes Deep Learning and Image Processing in Healthcare and Agriculture Domain.

Rohan Rajendra Sadawarte, B. Tech (CSE) from Vishwakarma Institute of Information Technology, Pune. I am currently working as a Software Developer Engineer (SDE) in renowned digital marketing company named Merkle Sokrati, Pune. Moreover, I have 6 months of working experience in the same company as an intern. I am passionate, innovative, a problem solver, a hard worker and most importantly an eager learner! I have worked in many development technologies like React.js, Node.js, Django, etc. I am AI-ML enthusiast as well. I always push myself to learn new things and keep myself well prepared for any situation life will put me in, not only professionally but in practical world as well. I believe every situation makes you a better person, so always keep a positive approach.

Mahek Ibrahim Mulla, B. Tech (CSE) from Vishwakarma Institute of Information Technology, Pune. I am currently onboard with Nutanix technologies, a pioneer in Hyper Converged Infrastructure, as an Associate Resident Consultant. Moreover, I have 6 months of working experience in the same company as a Remote Resident Expert intern. I am a Cloud enthusiast and have explored public cloud platforms such as AWS cloud and GCP. I have also dived deeper into the concepts of networking and completed my CCNA certification. I possess the inquisitiveness to learn new technologies emerging in the computer science field. Alongside my technical skills, I regularly volunteer at an NGO because I believe it's important to give back to the world. I enjoy volunteering at various events as an event manager and an anchor. I even have a blog page with my friends wherein I have uploaded a few of my writings.

Shreya Rahul Jadhav, B. Tech (CSE) from Vishwakarma Institute of Information Technology. I will be starting my Masters in Computer Science this Fall at Rochester University of Technology, Rochester, NY. Recently published a dataset on Indian Sign Language. AI enthusiast, recently made a software to identify continuous sign language gestures. I an always open to knowledge, in a field as vast as technology, there's always a lot to learn.