Ensemble Model For Epidemic Detection In Maharashtra Using Machine Learning Techniques

Ms. Deepali Y. Kirange
Assistant Professor

Dr. Varsha M. Pathak
Professor

Dr. Yogesh N. Chaudhari
Assistant Professor
KCES's Institute Of Management And Research, Jalgaon

Abstract
This study presents a smarter way to detect epidemics early using a machine-learning ensemble stacking approach. Traditional methods and individual machine learning models often struggle to identify outbreaks quickly because they rely on clinical data, which takes time to collect and analyze. To overcome this, our model uses real-time data from Google Search Trends, which can reflect public health concerns as they arise. Our model can detect early signs of disease outbreaks by analyzing search queries. It combines multiple machine-learning algorithms to improve accuracy and responsiveness. We tested its performance using past epidemic data and related search trends, and the results show that it works significantly better than individual models like Random Forest, SVM, and Gradient Boosting. The model achieves higher precision, recall, and F1 scores, making it a reliable tool for predicting diseases like dengue and malaria, particularly in Maharashtra. This research highlights the power of search trend analysis as a fast, cost-effective, and scalable method for tracking disease outbreaks. With its ability to process large amounts of data in real-time, the stacking-based model offers a promising way to improve epidemic detection and support quicker public health responses.

Date of Submission: 25-02-2025 Date of Acceptance: 05-03-2025

I. Introduction

Epidemic outbreaks have raised major concerns in communities about infectious disease control, prevention, and management to reduce the spread of disease and limit the affected areas[1]. An epidemic refers to the unexpected outbreak of a disease within a particular population, geographic region, or community, occurring at a significantly higher rate than expected. Epidemics can be triggered by infectious agents like viruses, bacteria, fungi, or parasites, and may spread rapidly, causing widespread illness, social disruption, and financial strain. Examples of recent epidemics include Coronavirus (2019- 2022), Dengue (2020), Zika (2015-present day), Ebola (2014-16), Swine Flu (2009-10). The Coronavirus outbreak all over the world, the Ebola outbreak in West Africa, the Zika virus in Latin America, and localized influenza outbreaks, occur annually in various parts of the world. Dengue cases have increased globally over the past two decades. Around 90 countries have seen active dengue transmission in 2024 [2]. Dengue is a mosquito-borne viral infection caused by the Dengue virus (DENV), which belongs to the Flavivirus family. It is one of the most significant viral diseases transmitted by mosquitoes, especially in tropical and subtropical regions of the world. Mosquitoes are considered an important animal vector that can cause several diseases to human beings. Mosquito-borne infectious disease is accepted as important tropical infections and is the focused topic in tropical medicine [3], [4]. Dengue is primarily transmitted by the Aedesaegypti mosquito and to a lesser extent by the Aedes albopictus species. Both mosquito species are also responsible for the transmission of other diseases such as Zika, Chikungunya, and yellow fever, Malaria. These are imminent threats to humanity, in part due to the expansion of their arthropod vectors, in particular, Aedesaegypti and albopictus, into new geographical areas. This is facilitated by anthropogenic changes including climate change. Dengue is one of the most prevalent vector-borne diseases, with four distinct serotypes of dengue flavivirus circulating in more than 110 countries worldwide, mainly in tropical and sub-tropical regions [5].

II. Literature Review

Malaria and Dengue are two common mosquito-borne infections that cause high illness and death rates worldwide. A system that can support existing disease surveillance by providing timely information on outbreaks can help reduce their impact. Recently, an Internet-based surveillance system, which analyzes people's online search behavior, has emerged as a promising tool for detecting infectious disease outbreaks early [6]. In the present decade, teledensity in India is rapidly increasing, and the internet has emerged as an indispensable need of people [7], [8]. A large proportion of internet users go online to search for medical or health-related information [9]. Recent studies have also shown that the Internet is among the primary sources of information for the population actively using the Internet [10], [11], [12], [13]. Data generated from queries fed into search engines has been recorded and can be used for surveillance purposes as it is used for marketing purposes. Targeted sources include Internet search metrics, online news stories, social network data, and blog/microblog data [6].

According to Rand Obeidat et. al [14], the search queries and keyword trends can be truly reliable to be used for the prediction of skin disease outbreaks. The search-term surveillance can provide an additional tool for infectious disease surveillance.

Seun O. Olukanmi et al [15] says that the South Africans tend to search Google to confirm their symptoms or for common flu home remedies around the week, they feel flu symptoms. Monitoring Google search data is a reliable proxy for monitoring flu spread. Google search data alone produces forecasts of Influenza-like illnesses.

According to Nirmalya Thakur et al [16], the temporal patterns of query rates related to Disease X, along with their geographical distribution and key search themes, can serve as a valuable indicator of public interest and the demand for information about Disease X.

Dong-Her Shih et. al [17] says that influenza-like illnesses are accurately detected by combining Google Trends with temperature data. It provides more accurate predictions than humidity data, suggesting that temperature is a more reliable monitoring indicator.

III. Methodology

Google Search Trends is a useful tool for analyzing the popularity of search queries over time. It allows users to explore various topics, compare search terms, and track how interest varies across regions and times. It can provide valuable insights into public interest and behavior related to disease outbreaks or emerging health issues like Dengue, Malaria, etc.

Google Trends presents the frequency at which a specific search term is considered as input into Google's search engine relative to the overall search volume on the site during a specific timeframe. Mathematically, if n (q, l, t) represents the number of searches for the query q in the location l during the period t, the relative popularity (RP) of the query is computed as shown in Equation (1). In Equation (1), Q (l, t) is a set of all the queries made from location l at time t, Π (n (q, l, t) > τ) is a dummy term with a value of 1 when n (q, l, t) > τ (query is popular) and 0 otherwise. The resulting numbers are scaled within the range of 0 to 100 based on the proportion of the topic relative to the total number of search topics. This defines the Google Trends Index (GTI), as shown in Equation (2) [14].

$RP\ (q,\ l,\ t) = n\ (q,\ l,\ t)\ /\ \Sigma q \epsilon Q\ (l,\ t)\ n\ (q,\ l,\ t) \times \Pi\ (n\ (q,\ l,\ t) > \tau)$ ....(1)

$GTI\ (q,\ l,\ t) = RP\ (q,\ l,\ t)\ /\ max\ \{RP\ (q,\ l,\ t)\ t \epsilon 1,\ 2\ldots,\ T\} \times 100$............. (2)

Prompt outbreak detection plays a crucial role in controlling and preventing infectious diseases. India's Integrated Disease Surveillance Programme (IDSP), launched in November 2004, serves as a vital tool for disease monitoring and outbreak response. Now a nationwide program, IDSP covers almost 97% of districts, tracking 22 epidemic-prone notifiable diseases through its one-stop portal. This platform facilitates the surveillance of disease trends and manages outbreaks via trained rapid response teams (RRTs) [18] . Disease data, collected through S (syndromic), P (presumptive), and L (laboratory) forms, is reported from the community level to state and central authorities. However, the current reporting process takes 7 to 10 days for the central surveillance unit to detect outbreaks. As a result, systems that can complement existing mechanisms and provide timely intelligence on infectious diseases could significantly mitigate the impact of outbreaks.

General Setting

Under the IDSP, three types of forms, 'P', and 'L'—are required to be submitted for disease surveillance. The 'S' form is used for reporting suspected cases through syndromic surveillance conducted by health workers at sub-centers, covering a population of 3,000 to 5,000. Medical officers in various health facilities, ranging from primary health centers to tertiary hospitals, including private practitioners, complete the 'P' form, or presumptive form. It reports on approximately more than 29 diseases based on clinical examinations. The 'L' form, or laboratory form, is used by laboratories (both public and private) to report 12 types of laboratory-confirmed cases. Data is collected from Monday to Sunday and reported the following Monday. Reporting units forward the data to the next level each Monday, and after verification and compilation, it reaches the District Surveillance Units by Wednesday. From there, it is transmitted to the State Surveillance Units (SSU) at the State/UT level and finally sent to the Central Surveillance Unit (CSU) in New Delhi.

Specific Setting

Maharashtra, located in western India, is one of the wealthiest and most industrialized states in the country, with the highest GDP among all Indian states. It has a tele density (number of telephone connections per 100 individuals) of around 91.67. Mumbai, the state's capital, is also the financial capital of India, and has a tele density of approximately 165.49. Internet access is available to around 69% of the population in Maharashtra [19]. The state regularly reports diseases using all three forms of the IDSP. In this study, two major febrile illnesses—dengue and malaria reported in the 'P' form were considered for analysis.

Google search trends

Google is the most widely used search engine, handling an enormous volume of queries daily, with a current market share of approximately 91.62% among all search engines [20]. Google tracks and records these search queries, compiling the data to automatically display trends. These weekly trends are accessible through Google Trends, a publicly available platform (https://trends.google.com/trends/).

Google Trends provides a sample of both real-time data (from the last seven days) and non-real-time data (from 2004 up to 36 hours before the search). After removing personal information, the data is categorized and tagged by topic. The total number of searches in each geographic area divides each data point over a set period, allowing comparison of relative search popularity. Google Trends displays search frequency as a normalized data series, with values scaled from 0 to 100. The score reflects a term's popularity relative to its highest point on the graph. A value of 100 indicates peak popularity, while 50 shows the term was half as popular at its peak, and a score of 0 means it was less than 1% as popular. This data can be downloaded in 'CSV' format for further analysis. Queries made by a very small number of users, duplicate searches, and searches involving special characters are excluded from Google Trends data. Figure 3.1 shows the interface of Google search trends.

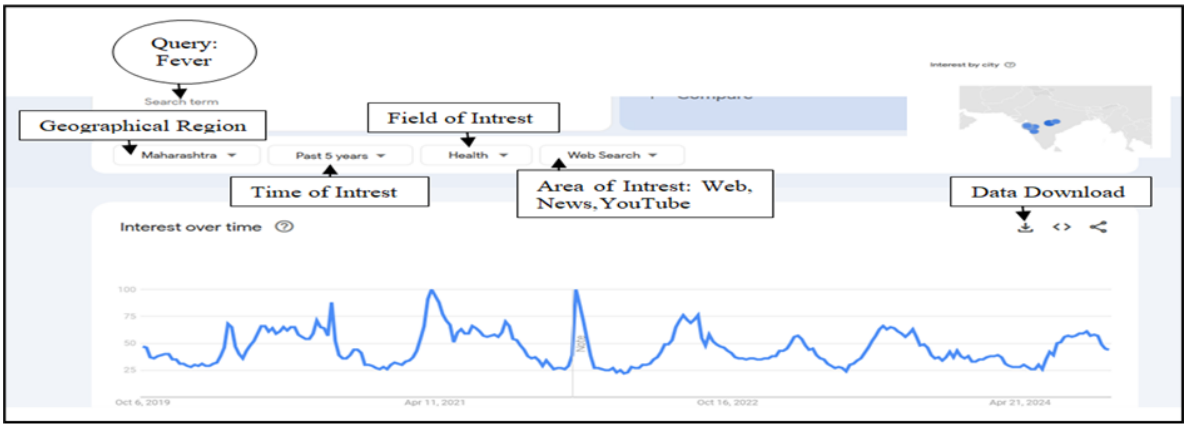

Fig. 3.1 Interface of Google Search Trend

Identified search terms

In our study, to track public interest in disease-related symptoms by monitoring search patterns across various platforms, including Web, YouTube, News, and Shopping searches had been used. Specifically, analyze whether an increase in search activity for terms related to the symptoms of dengue and malaria correlates with a potential epidemic outbreak in a particular region from IDSP. The following are the keywords focus on symptoms associated with these diseases, as recommended by healthcare professionals, WHO, and IDSP.

- **Dengue**: "fever", "weakness", "headache", "body aches", "nausea", "fatigue"
- **Malaria**: "high fever", "weakness", "chills", "unconsciousness", "breathlessness"
- **Other Terms**: "thermometer", "pulse oximeter"

Study Period and Population

The study focuses on the Maharashtra region, using Google Trends to filter searches by state and district (treated as cities for analysis purposes). The districts are Satara, Gadchiroli, Chandrapur, Thane, Pune, Kolhapur, Osmanabad, and Raigad…..N. The past five years' data in weekly format collected from disease outbreak reports, available on the official IDSP government portal, is correlated with Google Trends data to identify patterns and potential outbreaks.

IV. Results And Discussion

Data collection

The study uses 254 weeks of data, covering the period from the 37th week of 2019 to the 29th week of 2024. National-level outpatient data was obtained from IDSP (Integrated Disease Surveillance Program) under the National Center for Disease Control (NCDC), Directorate General of Health Services. This data is anonymized and consists of weekly counts of patients showing symptoms that match dengue and malaria case definitions, such as fever, headache, and vomiting. The weekly number of dengue and malaria cases in Maharashtra was compiled, and the process started by collecting Google Trends data for relevant search terms. Parameters such as region, period, and category were set to retrieve the data using the Google Trends API via the PyTrends library. The retrieved search data, containing disease-related keywords over a specific timeframe, was then merged into a single dataset. The mean values were calculated to unify data for the same search terms across different search criteria. Next, data association rules were applied to combine the columns effectively. These rules help identify relationships between search terms, allowing the analysis of search behavior patterns related to disease outbreaks. By applying these rules, the study aims to predict search trends and their correlation with actual outbreak data. The details of the rules are:

A) **Rule 1**: If there is a search for "fever" (A), across web, YouTube, news, or shopping searches, all counts are combined into one by calculating their mean (B).
B) **Rule 2**: If there is a search for "thermometer" (A), there is also likely a spike in searches for "fever" (B).
C) **Rule 3**: if there is a search for "pulse oximeter" (A), there is also likely an increase in searches for "breathlessness" (B).

Data Preprocessing

Google Trends gives relative interest values, which are normalized based on the highest search activity during the selected period. To make the data consistent, Z-score normalization is used. This method adjusts the data by centering it around the average and scaling it based on the standard deviation. This ensures that search trends can be compared accurately across different regions and periods.

$Z = (X - \mu) / \sigma$,

Where,

μ is the mean, and σ is the standard deviation.

Feature extraction helps convert raw data into a format that machine learning algorithms can easily process. In this study, there are two categorical features:

A) **Disease** – A categorical variable with three distinct values (0, 1, 2), representing different diseases.
B) **City** – Another categorical variable with N unique values (0, 1, 2... N), representing different cities.

This transformation makes the data suitable for machine learning models to analyze and detect patterns effectively.

The following is an algorithm for google trend data set preparation.

Google Trend Dataset Preparation Algorithm

Algorithm 1: Google Trend Dataset Preparation
Input: Raw data keywords, time, region
Output: Pre-processed Google Trend dataset
Begin:

1. Import the necessary libraries.
2. Set the following parameters.
2.1. Set the keywords for fetch google trend.
2.2. Set the period for the last 5 years.
2.3. Set the list of Cities to gather data on regional variations of disease trends.
2.4. Set the region to Maharashtra.
2.5. Set all categories.
3. Export the data for each search criterion.
4. Merge all exported datasets (columns of search term data) into a single file.
5. Apply the mean to unify data for the same search term across the different search criteria.
6. Use data association rules to combine the columns effectively.
7. Segment villages into their corresponding cities.
8. Apply preprocessing on dataset.
8.1. Fill missing values in city data with "Unknown."
8.2. Remove duplicate records to ensure data consistency.
8.3. Apply clipping to limit the range of data values, reducing outliers.
9. Map actual patient case counts from the IDSP website to corresponding records.

10. Apply Z-Score Normalization to standardize the dataset.
11. Annotate the dataset.

11.1. Annotate the diseases: [0,1,2] as [None, Dengue, Malaria]

11.2. Annotate the cities: [0,1,2,3,4,5,6,7,……,N] as [Satara, Gadchiroli, Chandrapur, Thane, Pune, Kolhapur, Osmanabad, Raigad…..N]

12. Return the fully pre-processed Google Trend dataset.

End

Machine Learning Algorithms

In this section, we describe different experimental models that are combinations of the various algorithms and data for epidemic detection in Maharashtra.

A) **Support Vector Machine (SVM):** It uses Google search trends to detect disease outbreaks in real time. The model classifies Google search patterns related to diseases (e.g., dengue, malaria, none) as either indicative of an outbreak or normal behavior.

B) **Random Forest:** Random Forest is an ensemble learning technique that creates multiple decision trees and aggregates their predictions to improve accuracy and reduce overfitting. It works well for this use case because it handles non-linearity, and feature importance, and reduces overfitting.

C) **Gradient Boosting:** Gradient Boosting is an algorithm for classification and regression tasks. It builds an ensemble of decision trees, where each new tree corrects the mistakes of the previous trees, resulting in a robust predictive model. This model can provide insights into which search terms or lagged features are most predictive of outbreaks. It handles imbalanced and non-linear data.

D) **Logistic Regression:** Logistic Regression is ideal for binary classification problems where the goal is epidemic detection (whether an outbreak will occur or not) based on Google Search Trends data. It estimates the probability that an event epidemic will occur given the input features like search trends.

E) **Ensemble Model Using Stacking:** Stacking is a machine learning ensemble technique where multiple models, called base models, are trained on the same dataset. Their predictions are then used as input features for a final model, called a meta-model, which learns how to best combine the base models' predictions. The goal of stacking is to leverage the complementary strengths of different models to produce a more accurate prediction.

An ensemble model using stacking is a powerful approach that combines multiple machine-learning algorithms to enhance prediction performance. For epidemic detection using Google Search Trends, a stacking model leverages the strengths of different base models and combines them with a meta-model to improve accuracy and robustness in predicting disease outbreaks like dengue or malaria. We have used Random Forest, XGBoost, and SVM as base models and Logistic Regression as the meta-model in our stacking approach.

Table 4.1 shows how well these machine learning models perform by measuring accuracy, precision, recall, and F1-score. These metrics help us understand how effective each model is at detecting epidemics. By looking at the results in Table 4.1, we can compare the performance of individual models with the final ensemble model and see how stacking improves overall predictive performance.

Table 4.1 Performance Measure for machine learning techniques

Machine Learning Technique	Accuracy	Precision	Recall	F1 Score
SVM	0.80	0.53	0.62	0.57
Random Forest	0.83	0.74	0.75	0.74
Gradient Boosting	0.93	0.86	0.90	0.88
Logistic Regression	0.73	0.50	0.56	0.53
Ensemble Model Using Stacking	**0.94**	**0.90**	**0.88**	**0.89**

Table 4.2 presents the precision, recall, and F1-score for each disease category. diseases are classified into different categories: Dengue is labeled as Class 1, Malaria as Class 2, and if neither disease is present, it is labeled as Class 0 (None).

Table 4.2 Performance Measure according to disease or class

Disease Name/Class	Accuracy	Precision	Recall	F1 Score
0-None	**0.98**	**0.94**	**0.96**	**0.95**
1-Denuge	**0.96**	**0.75**	**0.67**	**0.71**
2-Malaria	**1.00**	**1.00**	**1.00**	**1.00**

The ROC (Receiver Operating Characteristic) curve is used metric to evaluate the performance of a classification model. It measures the model's ability to distinguish between classes. The x-axis represents the false positive rate and y-axis represents the true positive rate) figure 4.1 shows the ROC Curve for ensemble model.

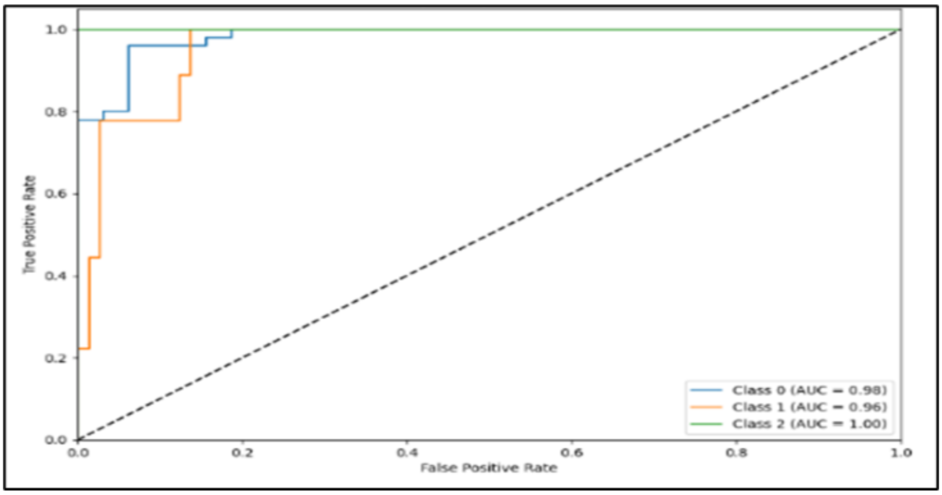

Fig. 4.1 ROC Curve for Ensemble model.

The proposed model has also compared the actual vs. predicted cases in Gadchiroli. Figure 4.2 displays a graph showing this comparison. In the graph, the x-axis represents the weeks, while the y-axis represents the number of cases.

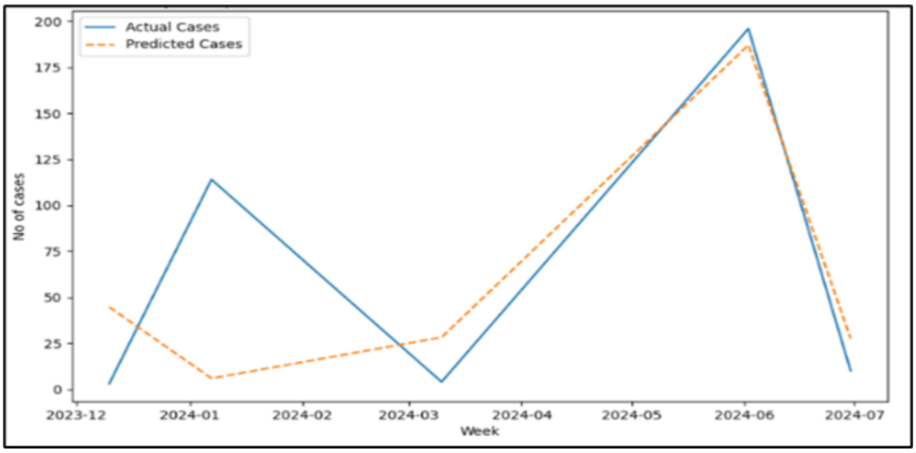

Fig. 4.2 Actual vs Predicted cases of Gadchiroli using Ensemble model.

V. Result And Discussion

Performance Analysis

When comparing proposed techniques with existing techniques for epidemic detection, it is essential to focus on the different approaches, their accuracy, diseases and techniques. Epidemic detection techniques rely on various data sources and analytical methods, including traditional epidemiology, statistical modeling, machine learning, and network science. Here is an analysis of different existing techniques and the proposed improvements often brought forward to enhance epidemic detection. Table 4.3 gives the comparison of ensemble model with existing techniques.

Table 4.3. Performance Analysis of proposed techniques with existing techniques.

Author Name	Title of Paper	Disease	Technique which gives better accuracy	Accuracy (%)
N.Rajathi, S.Kanagaraj[21]	Early Detection of Dengue Using Machine Learning Algorithms (2018)	Dengue	Random Forest	83.3%
C Koushik, Ritwika Bhattacharjee[22]	Symptoms based Early Clinical Diagnosis of COVID-19 Cases using Hybrid and Ensemble Machine Learning Techniques (2021)	COVID-19	Hybrid Model using Gradient Boosting Classifier (GBC)	87.17% (GBC)
C Koushik, Ritwika Bhattacharjee[22]	Symptoms based Early Clinical Diagnosis of COVID-19 Cases using Hybrid and Ensemble	COVID-19	Random Forest Classifier (RFC)	87.24% (RFC)

Author Name	Title of Paper	Disease	Technique which gives better accuracy	Accuracy (%)
	Machine Learning Techniques (2021)			
N. ThirupathiRaoa , Debnath Bhattacharyya [23]	Prediction of Swine Flu using a Hybrid Voting Algorithm (2021)	Swine Flu	Random Forest	78%
Thoma´s Tabosa de Oliveira, SebastiãoRoge´rio da Silva Neto[24]	A Comparative Study of Machine Learning Techniques for Multi-Class Classification of Arboviral Diseases (2022)	Dengue, Chikunguny a	Gradient Boosting Machines (GBM)	60.15%
Martina Mariki, Elizabeth Mkoba[25]	Combining Clinical Symptoms and Patient Features for Malaria Diagnosis: Machine Learning Approach (2022)	Malaria	Random Forest	82%
	Ensemble Model	**Dengue, Malaria**	**Ensemble Model using Stacking**	**94%**

Table 4.3 presents a performance comparison between the proposed ensemble model and existing machine-learning techniques for epidemic detection. The results indicate that the ensemble model using stacking applied to both Dengue and Malaria, achieves the highest accuracy of 94%, significantly outperforming previous approaches. Among existing techniques, Random Forest has been widely used for various diseases, achieving 83.3% accuracy for Dengue, 82% for Malaria, and 78% for Swine Flu. For COVID-19 diagnosis, the Random Forest Classifier (RFC) and Gradient Boosting Classifier (GBC) demonstrated comparable performance, with accuracies of 87.24% and 87.17%, respectively. However, Gradient Boosting Machines (GBM) showed relatively lower accuracy (60.15%) for multi-class classification of arboviral diseases like Dengue and Chikungunya. The findings highlight that the stacking-based ensemble model not only improves predictive performance but also offers greater adaptability across different diseases. With its ability to process large-scale, real-time data, the ensemble model proves to be the most effective technique, providing a robust solution for early epidemic detection.

VI. Conclusion

This research introduces a new approach to epidemic detection by using a stacking-based ensemble model with data from Google Search Trends. By combining multiple machine learning algorithms, the model performed better than individual methods in predicting outbreaks. The use of real-time search trends helped detect early signs of epidemics, making the system faster and more proactive. The stacking ensemble model outperformed traditional methods like ARIMA and individual machine learning models such as Random Forest and SVM, achieving higher precision and recall, especially in detecting early outbreak signals. By leveraging Google Search Trends, the model captured public health concerns in real-time, providing early warnings even before official case reports were available. Its stacking framework allowed it to integrate different base models and adapt to changing search trends and outbreak patterns, making it effective in identifying new outbreaks or sudden infection surges. Furthermore, using search query data from a large and diverse population ensured that the model was scalable and applicable across different regions and diseases with minimal manual adjustments. Driven by real-time data, this ensemble model has significant potential as an early epidemic detection tool, providing valuable insights to policymakers and healthcare systems to help control the spread of infectious diseases.

References

[1] N. Ibrahim, N. S. M. Akhir, And F. H. Hassan, “Predictive Analysis Effectiveness In Determining The Epidemic Disease Infected Area,” Aip Conf Proc, Vol. 1891, No. October 2017, 2017, Doi: 10.1063/1.5005397.

[2] “World Health Organization.” [Online]. Available: Https://Www.Who.Int/

[3] Jeff Stanley, “Malaria: The Global Resurgence Of Disease”.

[4] A. M. Campuzano And M. N. Restrepo B, “Caracterización Clínica De Los Casos De Dengue Hospitalizados En La,” 2006.

[5] L. A. Reperant And A. D. M. E. Osterhaus, “Aids, Avian Flu, Sars, Mers, Ebola, Zika… What Next?,” Aug. 16, 2017, Elsevier Ltd. Doi: 10.1016/J.Vaccine.2017.04.082.

[6] G. J. Milinovich, G. M. Williams, A. C. A. Clements, And W. Hu, “Internet-Based Surveillance Systems For Monitoring Emerging Infectious Diseases,” Feb. 2014. Doi: 10.1016/S1473-3099(13)70244-5.

[7] T. P. N. A. A. V. Donna L. Hoffman, “Has The Internet Become Indispensable?”.

[8] “Telecom Regulatory Authority Of India,Mahanagar Doorsanchar Bhawan,Delhi”.

[9] G. O. Hellawell, K. J. Turner, K. J. Le Monnier, And S. F. Brewster, “Urology And The Internet: An Evaluation Of Internet Use By Urology Patients And Of Information Available On Urological Topics.”

[10] T. Tonsaker, M. Gillian, B. Phd, And C. Trpkov, “Health Information On The Internet Gold Mine Or Minefield?” [Online]. Available: Www.Statcan.Gc.Ca/

[11] M. D. A. F. C. O. M. B. ,Ch. B. S. A. Ph. D. David Cunningham, “Surveys Of Physicians And Electronic Health Information”.

[12] J. A. Diaz, R. A. Griffith, J. J. Ng, S. E. Reinert, P. D. Friedmann, And A. W. Moulton, "Patients' Use Of The Internet For Medical Information."

[13] F. A. Moretti, V. E. De Oliveira, And E. M. K. Da Silva, "Access To Health Information On The Internet: A Public Health Issue?," Rev Assoc Med Bras, Vol. 58, No. 6, Pp. 650–658, 2012, Doi: 10.1590/S0104-42302012000600008.

[14] R. Obeidat, I. Alsmadi, Q. Bani Bakr, And L. Obeidat, "Can Users Search Trends Predict People Scares Or Disease Breakout? An Examination Of Infectious Skin Diseases In The United States," Infectious Diseases: Research And Treatment, Vol. 13, P. 117863372092835, 2020, Doi: 10.1177/1178633720928356.

[15] S. O. Olukanmi, F. V. Nelwamondo, And N. I. Nwulu, "Utilizing Google Search Data With Deep Learning, Machine Learning And Time Series Modeling To Forecast Influenza-Like Illnesses In South Africa," Ieee Access, Vol. 9, Pp. 126822–126836, 2021, Doi: 10.1109/Access.2021.3110972.

[16] N. Thakur, S. Cui, K. A. Patel, I. Hall, And Y. N. Duggal, "A Large-Scale Dataset Of Search Interests Related To Disease X Originating From Different Geographic Regions," Data (Basel), Vol. 8, No. 11, 2023, Doi: 10.3390/Data8110163.

[17] D. H. Shih, Y. H. Wu, T. W. Wu, S. C. Chang, And M. H. Shih, "Infodemiology Of Influenza-Like Illness: Utilizing Google Trends' Big Data For Epidemic Surveillance," J Clin Med, Vol. 13, No. 7, 2024, Doi: 10.3390/Jcm13071946.

[18] "Integrated Disease Surveillance Programme." [Online]. Available: Https://Idsp.Mohfw.Gov.In/

[19] Kantar, "Internet In India 2023," Pp. 1–20, 2023.

[20] "Search Engine Market Share 2023-2024".

[21] N. Rajathi, S. Kanagaraj, R. Brahmanambika, And K. Manjubarkavi, "Early Detection Of Dengue Using Machine Learning Algorithms." [Online]. Available: Http://Www.Ijpam.Eu

[22] C. Koushik, R. Bhattacharjee, And C. S. Hemalatha, "Symptoms Based Early Clinical Diagnosis Of Covid-19 Cases Using Hybrid And Ensemble Machine Learning Techniques," 2021 5th International Conference On Computer, Communication, And Signal Processing, Icccsp 2021, Pp. 59–64, 2021, Doi: 10.1109/Icccsp52374.2021.9465494.

[23] N. T. Rao, D. Bhattacharyya, E. Stephen, N. Joshua, And C. V Satyanarayana, "Prediction Of Swine Flu Using A Hybrid Voting Algorithm," Vol. 12, No. 10, Pp. 1169–1177, 2021.

[24] T. Tabosa De Oliveira Et Al., "A Comparative Study Of Machine Learning Techniques For Multi-Class Classification Of Arboviral Diseases," Frontiers In Tropical Diseases, Vol. 2, No. February, Pp. 1–10, 2021, Doi: 10.3389/Fitd.2021.769968.

[25] M. Mariki, E. Mkoba, And N. Mduma, "Combining Clinical Symptoms And Patient Features For Malaria Diagnosis: Machine Learning Approach," Applied Artificial Intelligence, Vol. 36, No. 1, 2022, Doi: 10.1080/08839514.2022.2031826.