**Disease Outbreak Prediction using Machine Learning**

A Project Report submitted in partial fulfillment of the requirements for the award of the degree of

**Bachelor of Technology**

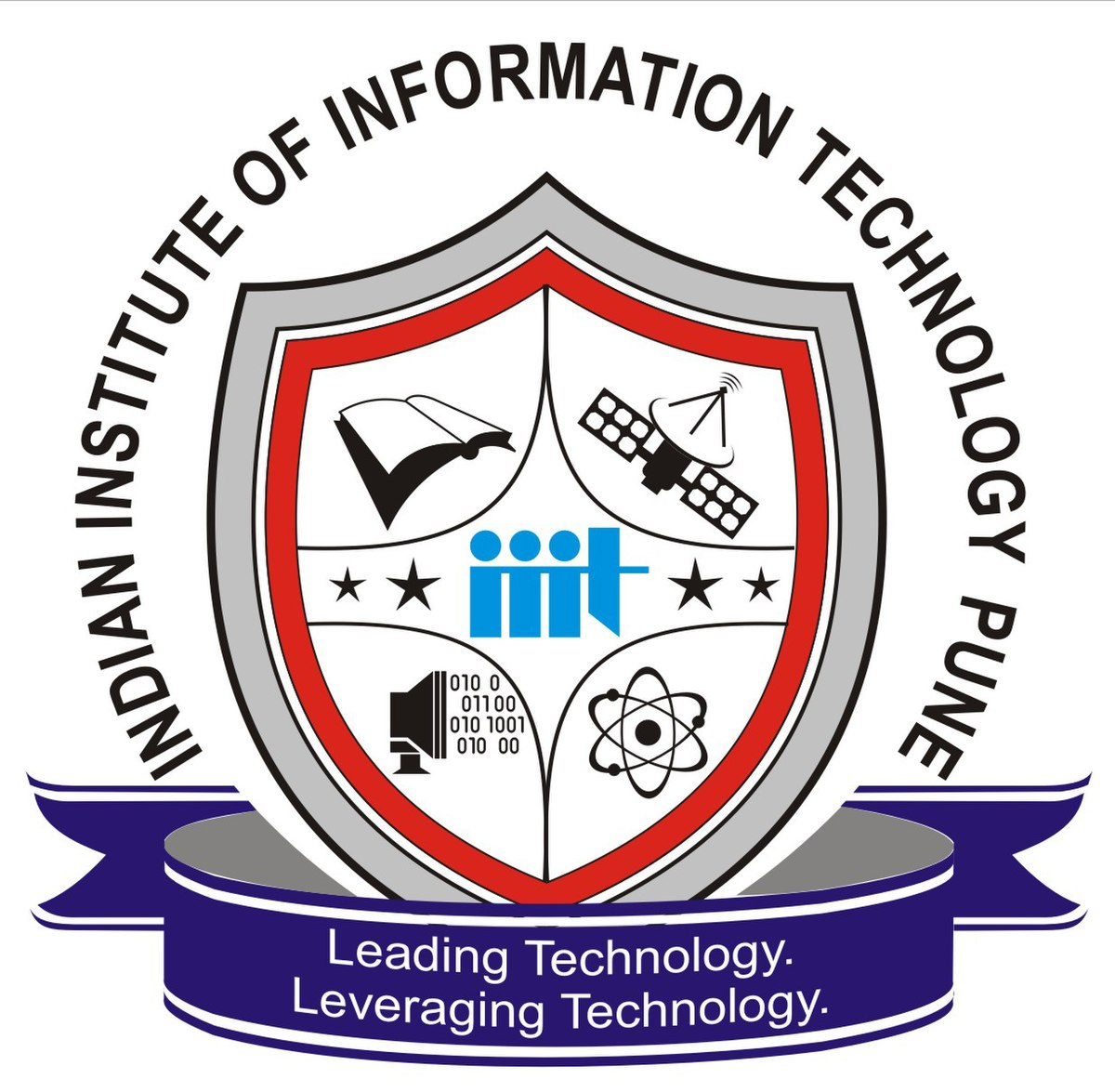
### in

**Computer Science and Engineering**

By

### ANSHUL (112115022)

### Semester: IV

****

#### Name of Department: Department of Computer Science and Engineering

#### 

#### Indian Institute of Information Technology, Pune

**(An Institute of National Importance by an Act of Parliament)**

#### MAY 2023

**BONAFIDE CERTIFICATE**

This is to certify that the project report entitled **“Disease Outbreak Prediction using Machine Learning”** submitted by **Anshul** bearing the **MIS No: 112115022**, in completion of his/her project work under the guidance of **Supervisor’s Name** is accepted for the project report submission in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology** in the **Department of Computer Science and Engineering**, Indian Institute of Information Technology, Pune (IIIT Pune), during the academic year **2022-23**.

#### Supervisor’s Name HOD Name

Project Guide Head of the Department

Designation of the Supervisor Assistant Professor

Department of the Supervisor Department of CSE/ECE

IIIT Pune IIIT Pune

Project Viva-voce held on

**Undertaking for Plagiarism**

I **Anshul** solemnly declare that research work presented in the **report** titled **“Disease Outbreak Prediction using Machine Learning”** is solely **my** research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete report has been written by **me**. I understand the zero tolerance policy of **Indian Institute of Information Technology Pune** towards plagiarism. Therefore, **I** declare that no portion of my **report** has been plagiarized and any material used as reference is properly referred/cited. I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of the degree, the Institute reserves the rights to withdraw/revoke my **B.Tech** degree.

**Students’/ Student’s Name and Signature with Date**

**Conflict of Interest**

**Manuscript title: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

The authors whose names are listed immediately below certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers’ bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

**Students’/ Student’s Name and Signature with Date**

## ACKNOWLEDGEMENT

This project would not have been possible without the help and cooperation of many. I would like to thank the people who helped me directly and indirectly in the completion of this project work.

First and foremost, I would like to express my gratitude to our honorable Director, **Prof. O.G. Kakde**, for providing his kind support in various aspects. I would like to express my gratitude to my project guide **Supervisor’s Name**, **Department of CSE**, for providing excellent guidance, encouragement, inspiration, constant and timely support throughout this **B.Tech Project**. I would like to express my gratitude to the **Head of Department (Name)**, **Department of CSE**, for providing his kind support in various aspects. I would also like to thank all the faculty members in the **Department of CSE/ECE** and my classmates for their steadfast and strong support and engagement with this project.

**(Note: Students may add or change the Acknowledgement as per their preference, but it must not exceed one page.)**

## Abstract

Disease outbreak prediction has become increasingly relevant in today's world due to the ongoing global health challenges. The rapid spread of infectious diseases, such as COVID-19, has highlighted the critical need for accurate and timely prediction of disease outbreaks to inform public health interventions and prevent further spread. Intelligent models, including machine learning and data analysis techniques, have emerged as powerful tools for predicting and managing disease outbreaks.

This report is a summary of how Machine learning Algorithms are being used today in prediction of diseases outbreaks today.

This report discusses how we have been predicting the outbreaks of various diseases today through Classification and Regression model in Machine Learning. Models have been applied to disease outbreak prediction, including supervised, unsupervised learning approaches. We describe how these models utilize different algorithms, such as decision trees, support vector machines, KNN, etc to analyze diverse datasets, including epidemiological data, environmental data, and social media data. To measure the accuracy of model, various methods have been used as F1 score, AUC curve. This report also highlights the advantages and limitations of these models in terms of accuracy, interpretability, scalability, and generalizability.

Furthermore, the challenges and ethical considerations associated with using machine learning models for disease outbreak prediction, such as data privacy, bias, interpretability, and accountability is discussed. We emphasize the importance of addressing these challenges to ensure the responsible and effective use of machine learning in disease outbreak prediction.

Finally, it concludes with future directions and recommendations for further research in the field of disease outbreak prediction using machine learning models. We highlight the need for developing robust, interpretable, and scalable models that can integrate diverse data sources and provide accurate and timely predictions.

Keywords: **Machine Learning, Supervised Learning, Unsupervised Learning, Classification, F1 Score, AUC Curve, Regression, KNN, Support Machine vector, Decision Tree, epidemiology**

Abstract 1

## TABLE OF CONTENTS

**Abstract i**

**(i)** [**List of Figures/Symbols/Nomenclature iv**](#_heading=h.gjdgxs)

**(ii)** [**List of Tables v**](#_heading=h.30j0zll)

1. **Introduction 1**
2. **Motivation of Work**
3. **Literature Review 5**

**3.1 Epidemic Prediction using Machine Learning Model**

**3.2 Coronavirus Outbreak Prediction using Machine Learning**

**3.3 Infectious disease outbreak prediction using Media articles with Machine Learning Model**

1. **Conclusion and Future Scope 10**

**5 References**

Table of content 2

# 

# List of Figures [/ Symbols/ Nomenclature](#_heading=h.gjdgxs)

List of Figures 3

# List of Table

List of tables 4

**Introduction**

In recent years, the world has witnessed alarming outbreaks of zoonotic diseases, such as COVID-19, Zika virus, Ebola, and others, which have posed significant threats to human health and well-being. Many countries whose economies have been most negatively impacted by the illness pandemic, have experienced economic disruption. These diseases, which are transmitted to humans, have caused widespread morbidity, mortality, and socioeconomic disruptions, underscoring the need for early prediction and detection of unknown diseases that may cause harm to humanity. Machine learning, a powerful subset of artificial intelligence (AI), holds great potential in addressing this challenge by developing predictive models that can aid in the timely identification and containment of disease outbreaks.

COVID-19, caused by the novel coronavirus SARS-CoV-2, has ravaged the globe since its emergence in late 2019, leading to millions of infections and deaths, overwhelming healthcare systems, and disrupting economies and societies worldwide. The Zika virus outbreak in 2015-2016 raised concerns due to its association with birth defects and neurological complications. The Ebola virus disease outbreaks, recurring in Africa since the 1970s, have caused significant mortality and posed challenges in containment and control. These outbreaks have exposed the vulnerabilities in our ability to detect and respond to emerging diseases, highlighting the critical need for improved early prediction and detection mechanisms.

In conclusion, these diseases and other outbreaks have proven to be formidable challenges for humanity. So we need a machine learning model that can predict an outbreak of these diseases so that we can take the necessary precautions to stop its transmission in order to be better prepared for future outbreaks.

The early prediction of unknown diseases is a complex task that requires a multidisciplinary approach, involving expertise from fields such as epidemiology, virology, data science, and public health. Machine learning models have shown promise in this regard, as they can analyze large datasets and identify patterns and risk factors that may signal the emergence of a disease outbreak before it escalates into an epidemic or pandemic.

By leveraging historical and real-time data daily from newspaper articles and social media on disease transmission, environmental factors, genetic information, and social determinants of health, machine learning models can provide valuable insights into the likelihood, severity, and spread of unknown diseases, enabling timely and targeted interventions.

Additionally, machine learning models will continuously learn from incoming data and adapt, enhancing their accuracy and prognostication skills over time. These models can be improved and updated as more data becomes available resulting in more precise and timely predictions, increasing their usefulness in disease outbreak prediction and control.

Introduction 5

**Motivation of Work**

The global spread of diseases like COVID-19 and Zika has demonstrated how interconnected and vulnerable our world is to emerging infectious diseases. Rapid transmission through global travel and interconnected supply chains has made containment and control of outbreaks challenging. These diseases have resulted in widespread illness, mortality, economic disruptions, and social upheaval, underscoring the devastating consequences of unpreparedness.

One of the main motivations for undertaking this topic is the availability of vast amount of data on news articles and social media posts related to disease outbreaks. News articles from reputable sources provide valuable information on disease outbreaks, including geographical locations, reported cases, mortality rates, and response measures. Social media platforms, on the other hand, capture real-time conversations, public opinions, and user-generated content related to disease outbreaks. This wealth of data can be leveraged to train machine learning algorithms and build predictive models that can identify patterns, trends, and early warning signals associated with disease outbreaks.

Furthermore, newspaper articles and social media play a key role in capturing the public's perception and response to disease outbreaks. Public opinions, concerns, and rumors related to disease outbreaks are often shared and discussed on social media platforms. News articles reflect the media's coverage of disease outbreaks, including the tone, sentiment, and level of urgency conveyed to the public. Analyzing this information can provide valuable insights into the public's behavior, adherence to preventive measures, and their impact on disease spread. Incorporating such social and media-related factors into machine learning models can enhance the accuracy and effectiveness of disease outbreak predictions.

Moreover, the timeliness of news articles and social media posts can greatly benefit disease outbreak prediction models. Traditional data sources for disease surveillance, such as official reports or medical records, may suffer from delays in data collection, aggregation, and analysis. In contrast, news articles and social media posts are often updated in real-time, providing a more current and dynamic picture of disease outbreaks. By leveraging the real-time nature of news and social media data, machine learning models can detect early warning signals and capture emerging trends, enabling timely interventions and response measures.

In conclusion, the role of newspaper articles and social media in disease outbreak prediction cannot be overstated. The availability of vast amounts of data, the capturing of public perceptions and behaviors, and the real-time nature of news and social media make them invaluable sources of information for building accurate and timely machine learning models for disease outbreak prediction. Incorporating these data sources into predictive models can enhance their accuracy, timeliness, and effectiveness, contributing to the development of robust and proactive approaches for disease outbreak prediction and response.

Motivation of work 6

**Literature Review**

Effective outbreak prediction models are needed to learn more about the anticipated spread and infectious disease effects, insights from other legislative bodies and government. The prediction models are for suggestion of new strategies and assess effectiveness of those that have already been put in place. (A & G., 2020) [1].

1. **“Epidemic Outbreak Prediction Using Machine Learning Model”** -by Soham Shinde, Seema Yadav, Ashelesha Somvanshi (Information Technology, K.J Somaiya Institute of Engineering and Information Technology, Mumbai, India)

The paper published in the IEEE Xplore digital library presents a comprehensive literature review on the use of machine learning models for predicting epidemic outbreaks. In this literature review, I will summarize and discuss the key findings and contributions of the paper, as well as highlight the relevant literature that has been previously done in this area.

**1.1 Outline of their work**

Their work is an outline of existing research on applying machine learning models for predicting epidemic outbreaks. It addresses a number of research that used machine learning methods for predicting the occurrence and outbreak of disease like Zika in various parts of the world, using decision trees, support vector machines, and deep learning algorithms[2]. The authors have discussed different types of data used in epidemic prediction, including clinical data, environmental data, social media data, and mobility data, as well as the challenges associated with data collection, preprocessing, and integration. They highlight the importance of feature selection and feature engineering in improving the efficiency of machine learning models for epidemic prediction and also highlights the various factors that are used as predictors in prediction models, such as climate variables, socio-economic factors, demographic factors, and mobility patterns.

For prediction they have built two kinds of model, Classification model for probability prediction of any case and Regression model to predict the probable number of cases.

In their research they have collected three major datasets (Zika Data Repository from CDC), Historical weather data collected online through API, population density dataset of the virus struck zones to do the prediction for the outbreak of Zika virus.

Literature Review 7

**1.2. Model Built**

For all Machine Learning activities, they have primarily employed Python programming language along with Pandas, NumPy, Scikit-Learn package to build models.

ADA Boost model performed pretty well as compared to other baseline classification models.

7 models were tested for prediction out of which Linear Regression model gave standard results.

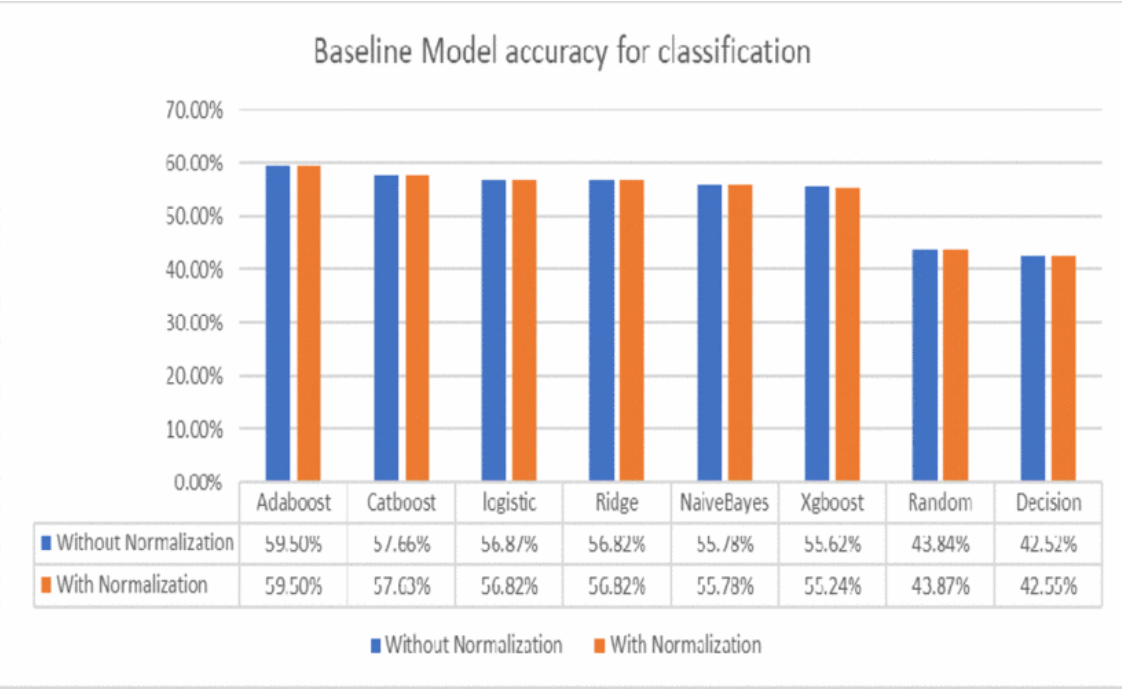
****

Fig1: Performance evaluation of classification model[3]

But after hyperparameter tuning, the performance of Random Forest Regressor is better as compared to other models.

Literature Review 8

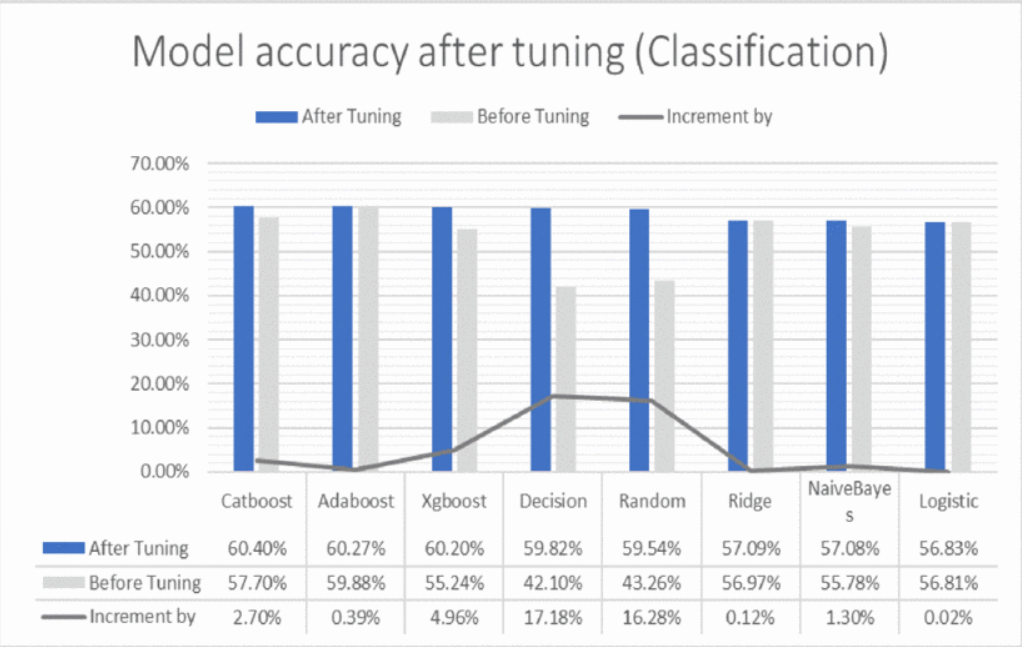


Fig 2: Performance evaluation after tuning[3]

**1.3 Results**

For Classification: Auto Machine Learning method used: TPOT, Best Model: XG Boost Classifier, Performance Evaluation: 60.37%

For Regression: Auto Machine Learning method used: Auto Sklearn, Best Model: Random Forest Regressor, Performance Evaluation: Mean Square Error: 441754.63

**2. “Coronavirus Outbreak prediction using Machine Learning”-** by Aman Kumar, Vivek Kumar and Vaishali Gupta (School of Computer Science, Galgotias University, Greater Noida)

**2.1. Outline of their work**

A comprehensive experimental setup was created in this study in order to examine the efficacy and effectiveness of machine learning models for predicting COVID-19 positive cases. The public data source Kaggle provided the dataset for this study, which was made up of 11,435 records with 20 symptoms and one label. The dataset included an abundant amount of information for the analysis and prediction of COVID-19 cases, including details on sickness characteristics at different stages, pathology data by location, disease distribution, and fatality rates.

**2.1.A. Preprocessing of Datasets**

The explicit features were converted to numeric values using the label encoder technique. To enhance the performance of the models, non-significant and extraneous options were eliminated, and the dataset's dimensionality was decreased.

**2.1.B. Machine Learning Models**

Three different machine learning models were used in this study: Regression, Decision Tree Regressor and Random Forest method. These models were chosen based on their potential to handle regression tasks and

Literature Review 9

their suitability for forecasting COVID-19 positive cases.

Regression Model: The regression model was trained using the training set, and different regression techniques, such as Polynomial Regression, were applied to find the best fit for the dataset. The accuracy of the model was evaluated using metrics such as accuracy rate, root mean squared error, and F1 score.

Decision Tree Regressor: The Decision Tree Regressor model was used to create a decision tree-based model for predicting COVID-19 positive cases. The model was trained using the training set, and hyperparameter tuning was performed to optimize the model's performance. The accuracy of the model was evaluated using various metrics, including accuracy rate, root mean squared error, and F1 score.

Random Forest Method: The Random Forest method, with the Ada Boost method as a boost, was applied to create an ensemble-based model for predicting the severity of COVID-19 positive cases. The model was trained using the training set, and hyperparameter tuning was performed to optimize the model's performance. The accuracy of the model was evaluated using metrics such as accuracy rate and F1 score.

**2.1.C. Evaluation Metrics**

To assess the performance of the machine learning models, various evaluation metrics were used, including accuracy rate, root mean squared error, and F1 score. The accuracy rate measures the percentage of correctly predicted cases, while the root mean squared error measures the average squared difference between the predicted and actual values.

**2.2 Results**

The experimental results of the study revealed that the machine learning models, particularly the Polynomial Regression model, showed promising accuracy rates in predicting COVID-19 positive cases. The model achieved a 90% accuracy rate, which is higher than the algorithms previously used in similar studies. However, it was noted that 90% accuracy is still insufficient for medical data processing, indicating that further improvements are needed.

The Decision Tree Regressor model and the Random Forest method with the Ada Boost method as a boost also showed good performance in predicting the severity of COVID-19 positive cases, with accuracy rates of 95% and 94%, respectively. The F1 score of the Random Forest method was also high, with a value of 0.86, indicating good accuracy and precision in predicting the severity of cases.

**2.3 Conclusion**

Machine learning algorithms such as KNN, SVM, Decision Tree, and Random Forest have been utilized in this study to construct supervised classification models for analyzing COVID-19 data. Among these models, KNN outperformed others in terms of accuracy, recall, and F1 score, while SVM showed high precision but performed poorly in other criteria. Their study highlights the importance of using machine learning for understanding and predicting disease spread, aiding in real-time preparations for pandemic diseases

Literature Review 10

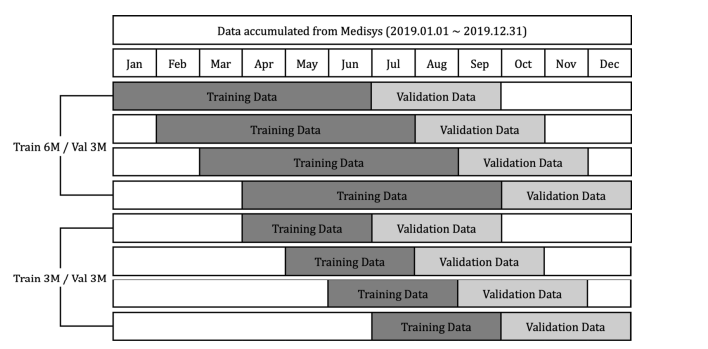
**3**. **“Infectious disease outbreak prediction using Media Articles with Machine Learning Model”**

By Juhyeon Kim and Insung Ahn

**3.1 Methods**

The methodology of their work is based on collecting and analyzing articles and reports related to infectious diseases from Medisys, a real-time Internet media source that provides news articles and reports on

infectious diseases published worldwide. The authors collected data from January to December 2019, consisting of 115,279 articles published in 237 different countries. The data was then normalized between 0 and 1 by each country to adjust for differences in population sizes and opportunities to publish digital data.



Overall, the methodology described in their report highlights the importance of Internet media as a valuable source of real-time information for studying infectious diseases. The use of Medisys as a data source, along with the normalization of data and reorganization for analysis, provides a systematic approach for researchers to study the occurrence of infectious diseases and make predictions based on the collected data.

**3.2 Model Built and Experimentation**

**3.2.A Building Model**

Three different machine learning models were adapted and compared for their performance in this task. The models used were:

Support Vector Machine (SVM): SVM is known for its consistent performance across various fields. The model parameters of SVM were searched over a specified range.

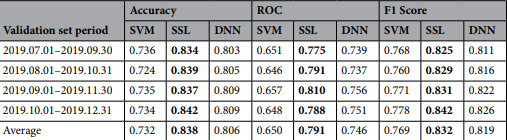
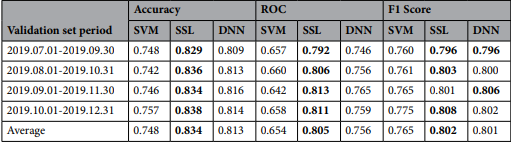
Semi-Supervised Learning (SSL): SSL is known for its good performance when dealing with imbalanced data sets. This could be helpful in the context of disease outbreak detection where data may be imbalanced. The model parameters of SSL were also searched over a specified range.

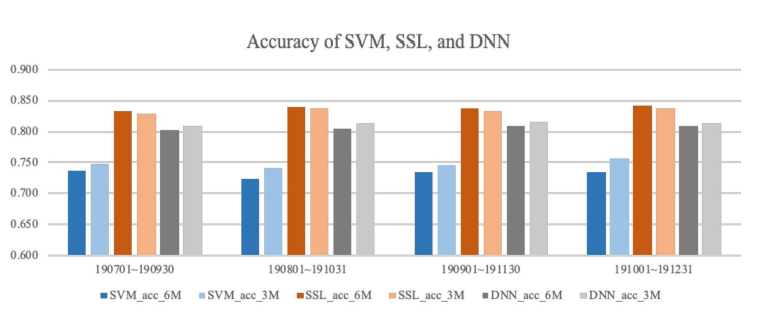
Literature Review 11

**3.2.B Experiments**

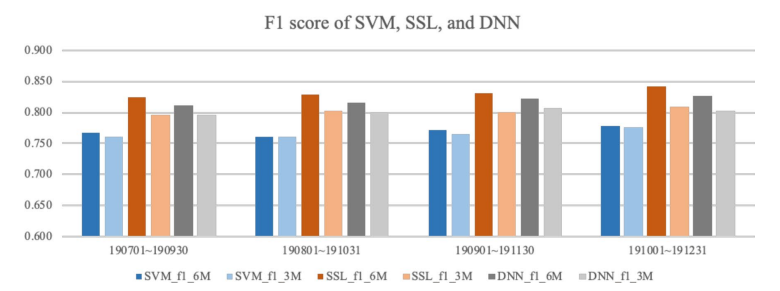
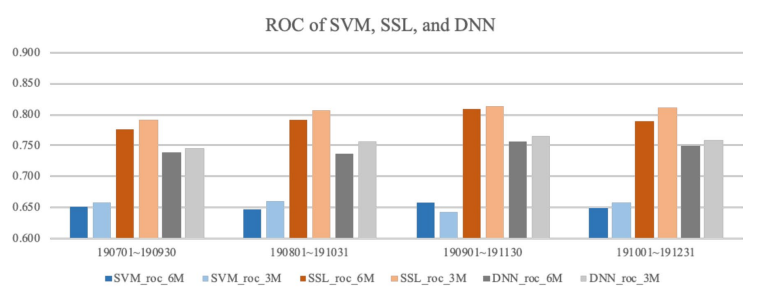
To evaluate the performance of the prediction models, the researchers used three metrics: AUC (Area Under the Curve), Accuracy, and F1 score. AUC is a threshold-independent measure of model performance based on the receiver operating characteristic curve, which assesses the overall value of a classifier. Accuracy measures the total number of correct predictions when the threshold is set to 0. F1 score is a weighted average of precision and recall, where a score of 1 represents the best performance and 0 represents the worst.

The study is in line with previous research that utilizes machine learning techniques to analyze large datasets for disease prediction. It builds on existing literature by focusing on media articles and reports as a potential data source for disease prediction, and by considering the geographic location of diseases by country. The use of multiple evaluation metrics provides a comprehensive assessment of model performance.





Literature Review 12



**3.3 Discussion**

The future scope of disease outbreak prediction using news articles and social media data in conjunction with machine learning algorithms is promising. The integration of these data sources can enhance the accuracy and timeliness of disease outbreak predictions, enabling effective public health interventions and resource allocation. Further research and development in this area can significantly contribute to improving global health security and mitigating the impact of infectious diseases.

Literature Review 13

**Conclusion and Future Scope**

**Conclusion**

The integration of news articles and social media data into disease outbreak prediction algorithms can have several potential benefits. Firstly, it can enable early detection of outbreaks, allowing public health agencies to implement timely and targeted interventions to prevent the spread of diseases. Secondly, it can facilitate real-time monitoring of outbreaks, providing up-to-date information on disease dynamics, affected areas, and at-risk populations. Thirdly, it can aid in resource allocation and preparedness planning, allowing for efficient deployment of healthcare resources and interventions where they are most needed.

However, there are also challenges that need to be addressed in utilizing news articles and social media data for disease outbreak prediction. One challenge is the need for accurate data collection and preprocessing, as news articles and social media data can be noisy and may contain false or misleading information. Ensuring data accuracy and reliability is essential to avoid erroneous predictions and interventions. Another challenge is the need for data privacy and ethical considerations, as news articles and social media data may contain sensitive information about individuals. Proper data anonymization and protection measures must be in place to ensure privacy and compliance with relevant regulations.

**Future Scope**

The future scope of machine learning algorithms in disease outbreak management is vast and promising.

As technology continues to evolve and data availability improves, machine learning algorithms can be further harnessed to revolutionize how outbreaks are detected, predicted, and managed, leading to more effective and timely responses to disease outbreaks. Integrating reliable data from media articles and social media can also contribute to better dataset quality and decision-making. Integration of these data sources can enhance the accuracy and timeliness of disease outbreak predictions, enabling effective public health interventions and resource allocation. Further research and development in this area can significantly contribute to improving global health security and mitigating the impact of infectious diseases Continued research, innovation, and collaboration between stakeholders are essential for unlocking the full potential of machine learning algorithms in managing disease outbreaks in the future.

Conclusion and Future Scope 14

# References

# 1. A Remuzzi and G. Remuzzi, "COVID-19 and Italy: what next?", Lancet, vol. 395, no. 10231, pp. 1225-1228, 2020.

2. Vaishali Gupta and Sanjeev Prasad, "Prediction of Epidemic Disease Outbreaks Using Machine Learning".

3. S. Shinde, S. Yadav and A. Somvanshi, "Epidemic Outbreak Prediction Using Machine Learning Model," 2022 5th International Conference on Advances in Science and Technology (ICAST), Mumbai, India, 2022, pp. 127-132, doi: 10.1109/ICAST55766.2022.10039594.

4.S.V. Scarpino and G. Petri, "On the predictability of infectious disease outbreaks", Nat Communication, vol. 10, pp. 898, 2019.

5. A. Kumar, V. Kumar and V. Gupta, "Corona Virus Outbreak Prediction Using Machine Learning," 2022 3rd International Conference on Intelligent Engineering and Management (ICIEM), London, United Kingdom, 2022, pp. 632-635, doi: 10.1109/ICIEM54221.2022.9853077

6. Kim, J.; Ahn, I. Infectious disease outbreak prediction using media articles with machine learning models. Sci. Rep. 2021, 11, 4413.

7.

References 15