iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** II **Month of publication:** February 2025

DOI: https://doi.org/10.22214/ijraset.2025.66948

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue II Feb 2025- Available at www.ijraset.com

Predicting Disease Outbreaks Using ML

Ms. Sayed Shifa Mohd Imran[1], Dr. Sweety Garg[2]
[1] *Student, Department of MSc.IT, Nagindas Khandwala College, Mumbai, India,*
[2]*Assistant Professor, Department of Computer and Information Science, Nagindas Khandwala College, Mumbai, India*

Abstract: ***This research focuses on forecasting disease outbreaks in India that occur after natural disasters by utilizing machine learning algorithms. Natural events like cyclones, floods, and earthquakes can lead to subsequent outbreaks of diseases like cholera, malaria, and dengue, exacerbating public health emergencies. In its forecasting, data concerning disasters, including both environmental and social elements such as temperature, rainfall, humidity, population density, and access to healthcare, is considered for predicting disease outbreak occurrences. Two algorithms, namely XGBoost and Long Short-Term Memory (LSTM), have been employed for the prediction of disease epidemics. The dataset includes 5,000 instances of disaster events, each containing variables such as disaster type, geographical region, and pertinent environmental factors.***

The XGBoost model, which utilizes a gradient boosting approach, demonstrated the highest performance, achieving an accuracy rate of 98.40%. It exhibited strong accuracy in forecasting both disease occurrences and non-outbreaks, showing high precision and a balanced performance across both categories. On the other hand, the LSTM model, which reached an accuracy of 92.70%, struggled more significantly with class imbalance, especially in its predictions regarding disease outbreaks. Although it was accurate in predicting non-outbreaks, LSTM faced difficulties in identifying the minority class (disease outbreaks), resulting in low recall and F1 scores for that particular category.

The results highlight the necessity of incorporating machine learning into disaster and public health planning to ensure prompt and efficient forecasting of disease outbreaks following a disaster. This research illustrates that predictive modeling can enhance disaster preparedness, optimize resource utilization, and guide public health strategies. In future research, it would be valuable to incorporate real-time data, develop advanced predictive algorithms, and create platforms that aid in disaster response and mitigation in actual situations, ultimately contributing to saving lives and reducing the impact of health crises in a post-disaster context.

Keywords: LSTM, XGBoost, Disaster, Disease, Prediction, Accuracy

I. INTRODUCTION

Natural disasters represent catastrophic occurrences that yield long-lasting effects on the environment, economy, and public health. In India, a nation particularly vulnerable to various natural disasters such as floods, cyclones, earthquakes, and heat waves, these events pose significant threats to human life and well-being. One of the most severe repercussions of such disasters is the emergence of diseases, which exacerbates the suffering of affected communities. Diseases such as cholera, malaria, dengue, and influenza have a high propensity to spread rapidly in the aftermath of these events, leading to public health crises that can overwhelm a country's healthcare system. Given the complex and unpredictable nature of disease outbreaks following disasters, there is an increasing need to utilize advanced data analysis and machine learning algorithms to forecast and manage health risks associated with these diseases. Researchers are focusing on the interactions between various disaster types, environmental conditions, and disease occurrences to create predictive models that can assist public health officials and policymakers in implementing timely interventions to mitigate the spread of infectious diseases.

This study introduces a deep learning model aimed at predicting the likelihood of disease outbreaks in relation to different types of natural disasters in India. The analysis utilizes a synthetic dataset that simulates real-world disaster scenarios, taking into account factors such as temperature, rainfall, humidity, population density, and healthcare availability. Several machine learning algorithms, including XGBoost—a novel gradient boosting technique—and Long Short-Term Memory (LSTM) networks, have been employed to forecast disease occurrences linked to specific disaster events.

The findings of this research are intended to enhance understanding of the predictive capabilities of social and environmental factors and to develop decision-support tools that can guide actions in response to disasters. By enhancing early warnings and providing room for proactive interventions in public health, such a study's findings could save lives and develop disaster-resilience in high-risk regions in the long term. By taking such a path, such a study confirms that artificial intelligence can address real, high-priority global issues, including protecting public health in a changing environment and increasingly prevalent natural disasters.

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue II Feb 2025- Available at www.ijraset.com

II. LITERATURE REVIEW

Stojanovic et al. (2020) identify several such environmental risk factors, including the use of unsanitary or contaminated water; lack of toilets and proper shelter; and crampedness among shelters, together with social ones, including such as poverty, lack of healthcare access, displacement being a significant catalyst for infectious disease epidemiology within humanitarian responses.

Chen et al. (2020) have reviewed machine learning techniques for disease outbreak prediction and indicate that models based on regression analysis, decision trees, and neural networks can predict with a high level of accuracy possibilities of outbreaks from environmental and social data when combined together. The identification of different information inclusions from climate to levels of sanitation, population density is considered for various models that deal with the forecast of diseases.

Mendoza et al. (2020) restrict the discussion to machine learning in order to predict outbreaks of malaria by using climatic, demographic data, and health infrastructure data. This paper demonstrates the contribution of the Random Forest and Support Vector Machines models for in-depth understanding regarding malaria risks in areas where either flood or other natural disasters affect localities. They also go on to elaborate how inclusions of environmental variables, including rainfall and temperature, immensely improve accuracy in such a prediction.

Fang & Lee (2019) explored some machine learning models for epidemic predictions and put more concentration on a boosting tree algorithm known as the XGBoost algorithm. They apply the XGBoost model and find its power in handling big and complicated data estimations, predicting the occurrence of infectious diseases in disaster areas. It thus proves that using XGBoost because of the feature incorporation has proved very wide, such as environmental factors, social data, and real-time input from monitoring systems.

Knight & Goldstein, 2018 discuss disease outbreaks in post-disaster situations and barriers to prediction. The problem with data quality and its availability is highlighted in this paper. Better and timely data from environmental sensors, social media, and health reporting will lead to improvement in the prediction models. Integration of machine learning with real-time monitoring will enhance the public health response and reduce the uncertainty in disease forecasting at times of disasters.

III. RESEARCH OBJECTIVE

1) Comparing the performance of a variety of machine learning algorithms in predicting disease progression
2) Build and train machine learning models to predict disease outbreaks

IV. METHODOLOGY

This study aims at predicting disease epidemics following a natural disaster using a machine learning algorithm. There is a 5000-sample dataset, comprising types of disaster (e.g., cyclone, flood, earthquake), environmental factors (temperature, rain, humidity), social factors (health access, population density), and disease occurrences. XGBoost, a gradient boost algorithm, and LSTM, a deep neural network, have been trained for disease epidemic prediction. Training and testing sets have been prepared for both algorithms, and both have been evaluated for accuracy and classification in terms of the test set. Prediction for disaster types such as cyclone and visualization with plots has been conducted. In this work, one can observe that with a simple and effective algorithm, disease epidemics can be predicted concerning a disaster, and with such predictions, public health planning and disaster management can be eased.

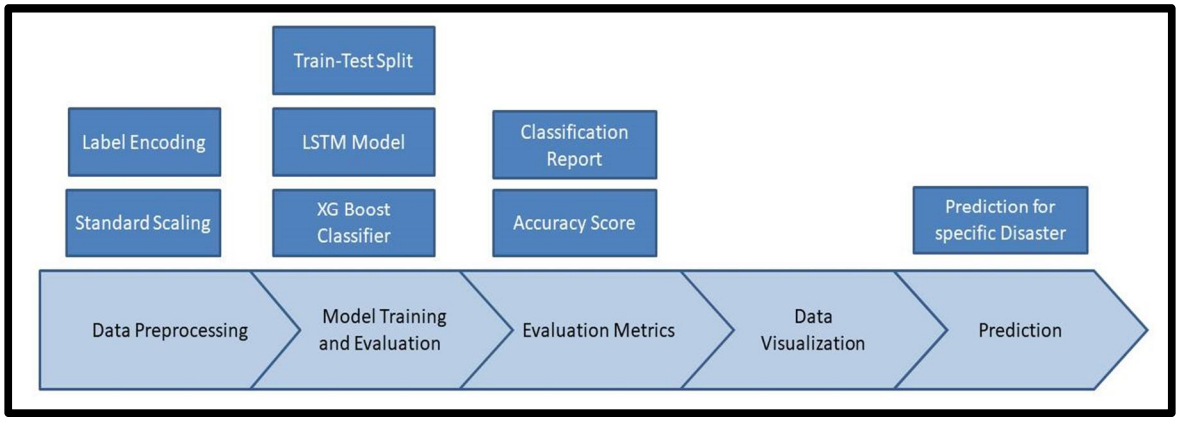


Fig. 1 Methodology

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue II Feb 2025- Available at www.ijraset.com

A. Sample Dataset

The dataset is composed of data about natural disasters and disease that have ensued out of them. There are 5000 samples and 2014 to 2024 information. There is following information in each sample:

1) Date: On which date did the incident occur
2) Disaster_Type: Nature of disaster, for example, a heatwave, tornado, earthquake, cyclone, etc.
3) Region: That region (e.g., North, South, or East) of a country in which a disaster occurred
4) Temperature: Temperature in Celsius at the time of disaster
5) Rainfall: Actual rain that took place during the disaster (in mm).
6) Humidity: Relative humidity at the time of disaster.
7) Population density: How many people live in the impacted zone per square kilometer
8) Healthcare_Access: On a 1 to 10 scale, a variable for ease with which one can access care
9) Disease_Type: The disease type encountered following a disaster (e.g., Cholera, Flu, None).
10) Disease_Reported: A 1 (1=1, 0=0) value for whether a disease outbreak occurred in a disaster

	Date	Disaster_Type	Region	Temperature	Rainfall	Humidity	Population_Density	Healthcare_Access	Disease_Type	Disease_Reported
0	2021-09-22	Flood	South	29.363503	380.285723	80.259667	918.054802	7	Malaria	0
1	2019-07-16	Tornado	East	31.145819	39.989966	65.258689	533.877486	8	None	0
2	2017-11-05	Heatwave	West	37.701814	8.233798	93.345042	1257.041829	6	Malaria	0
3	2020-01-06	Flood	North	20.019469	396.884624	73.961483	936.897083	9	Cholera	0
4	2014-10-09	Flood	Central	27.280729	244.741158	47.672162	473.609740	3	Malaria	0

Fig. 2 Sample Dataset

B. Libraries used

1) Pandas: N-dimensional, multi-dataframe, multi-indexed, high-performance dataframe and analysis tool
2) numpy: Python module for numerical computation
3) LSTM: Excels in sequential prediction tasks
4) Matplotlib.pyplot: Interactive and static plotting module
5) Seaborn: High level plotting tool with a strong base in Matplotlib capable of producing plots and charts
6) Sklearn.preparing: It contains tools for preparing data.
7) sklearn.model_selection: Holds useful utility function for cross-validation, training and testing partitioning, etc.
8) xgboost: A high-performance, efficient gradient boosting platform with a deep background in both classification and regression work
9) tensorflow.keras.models: Part of a deep part of TensorFlow, used for developing neural network model
10) datetime and timedelta: For date and time calculation and manipulation

C. Implemented methods

1) *Preparation of Data & Feature Engineering:* The first part in disease prediction in the case of a natural disaster is preparing the data and choosing significant features. Certain important features in the dataset include disaster (e.g., Cyclone, Flooding), environment (e.g., Temperature, Precipitation, Humidity), and social factors (e.g., Population, Healthcare Facility Coverage). Categorical values (e.g., types of disease and disaster) are represented in numerical form through Label Encoding in an attempt to encode them. Numerical values (e.g., Temperature, Precipitation) are scaled via Standardization in a way that all feature values contribute proportionately towards the prediction model. The target variable, Disease Reported, seeks to obtain information regarding whether a disease outbreak occurred in a post-disaster scenario.
2) *Model training:* To train a classification model with prepared data. Model Selection, two types of machine learning models has been utilized:

- XGBoost: A gradient boosting algorithm that is particularly geared towards working with complex relationships in the data.
- LSTM (Long Short-Term Memory): A neural network with a specific shape for sequential, temporally dependent information.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

On the whole dataset, both of these models are trained, and through them, trends between a disaster's factors (e.g., temperature, rainfall) and disease occurrences are discovered. Cross-validation techniques confirm that a model generalizes to new, unseen data and is not overfit.

3) *Prediction and evaluation of disease:* The final stage entails disease epidemic prediction and model testing. The model will make predictions of disease occurrences in terms of probability, with consideration for new disaster events' features (e.g., an ongoing impending earthquake with specific environmental factors). Performance is gauged in terms of accuracy, precision, recall, and F1-score regarding the accuracy with which disease prediction is performed by the models. The performance of the XGBoost and LSTM models is compared in an attempt to select the most effective model for disease epidemic prediction in a disaster scenario. Optimum model performance is leveraged for real-time prediction in a quest to allow public health agencies to manage disaster-related health risks effectively.

V. RESULTS

LSTM Accuracy: 92.70%
LSTM Classification Report:

	precision	recall	f1-score	support
0	0.93	1.00	0.96	922
1	0.69	0.12	0.20	78
accuracy			0.93	1000
macro avg	0.81	0.56	0.58	1000
weighted avg	0.91	0.93	0.90	1000

XGBoost Accuracy: 98.40%
XGBoost Classification Report:

	precision	recall	f1-score	support
0	0.99	0.99	0.99	922
1	0.91	0.88	0.90	78
accuracy			0.98	1000
macro avg	0.95	0.94	0.94	1000
weighted avg	0.98	0.98	0.98	1000

Fig. 3 Accuracy of LSTM and XGBoost Model

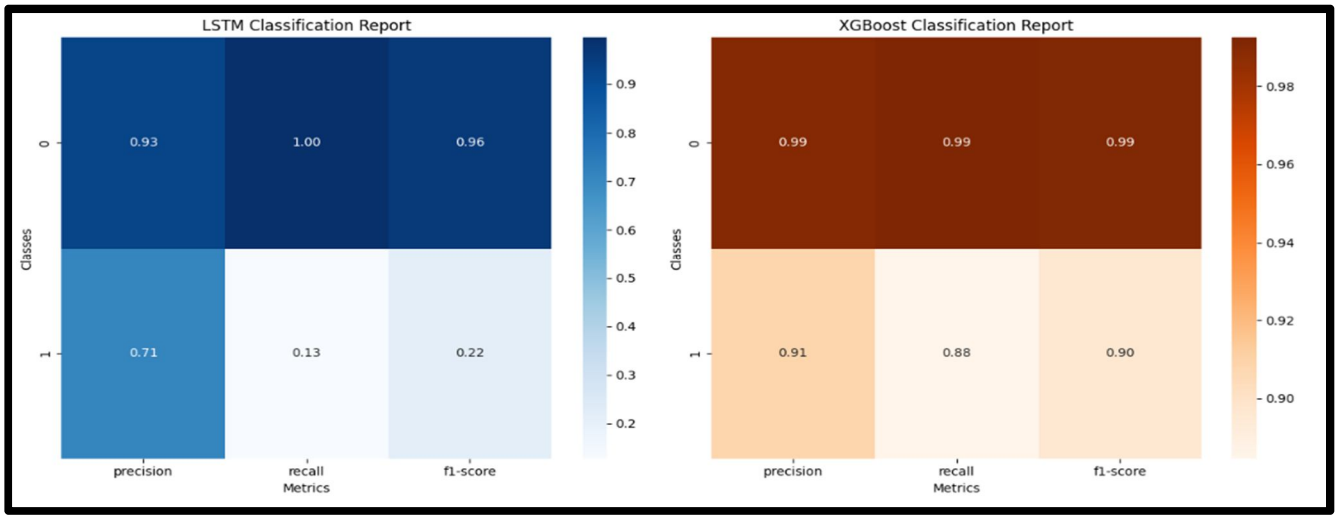


Fig. 4 Classification Report of LSTM and XGBoost Model

In above figures, The LSTM model demonstrated an accuracy of 92.70%, indicating its strong capability in predicting both disease and non-disease outbreaks. Nevertheless, its performance was notably skewed towards the dominant class, exhibiting a perfect recall of 1.00 for class 0 (no disease outbreak), which signifies its successful identification of all instances without disease outbreaks. In contrast, the model's performance for class 1 (disease outbreak) was considerably less effective, with a recall of only 0.12, highlighting its inability to recognize a significant number of actual disease outbreaks. This disparity resulted in a diminished F1 score of 0.20 for class 1, despite the overall high accuracy. The macro and weighted averages reflected this inconsistency, with the weighted average F1 score of 0.90 primarily representing the model's strong performance in class 0.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

XGBoost demonstrated superior performance, achieving an accuracy of 98.40%. It exhibited a well-balanced capability in predicting both classes, attaining a high recall value of 0.99 for class 0 and 0.88 for class 1. This indicates that XGBoost accurately identified the presence and absence of disease outbreaks with a significantly greater margin compared to LSTM. Furthermore, the precision and F1-score for both classes were notably higher in XGBoost, particularly for class 1, which had a recall value of 0.88 and an F1-score of 0.90, contrasting sharply with the considerably lower performance of LSTM. In summary, XGBoost's effectiveness in managing both majority and minority classes renders it a more dependable model for predicting disease outbreaks in disaster scenarios.

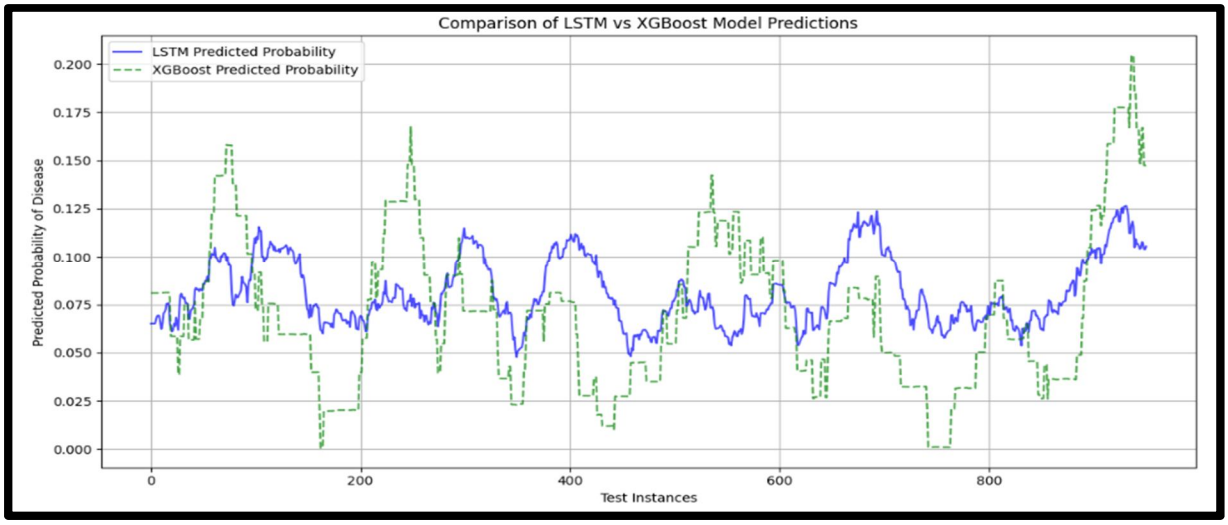


Fig.5 Comparison between LSTM and XGBoost Model

In fig. 5, the code evaluates two predictive models, XGBoost and LSTM, in the context of forecasting disease outbreaks using disaster-related data. Initially, it processes the data by encoding categorical variables and applying feature scaling, followed by dividing the dataset into training and testing subsets. XGBoost is trained using the original feature values, while LSTM is trained on values that have been reshaped for time-series analysis. To reduce noise, the predictions from both models are smoothed using a moving average, and the results are displayed graphically to facilitate a comparison of the predicted probabilities of disease outbreaks. This visual representation aids in assessing the performance of both models, highlighting their reliability and variability in predictions. XGBoost achieves a performance level of 98.40%, surpassing LSTM, which has a performance level of 92.70%. Consequently, XGBoost is deemed the more suitable model for this predictive task.

```
The most likely disease after an Earthquake is: Cholera
```

```
The most likely disease after a Cyclone is: Flu
```

Fig. 6 Prediction of Disease after a Disaster using XGBoost model

In fig. 6, this code is designed to develop and train an XGBoost model aimed at predicting the most probable disease that may arise during a disaster. The model incorporates various explanatory factors, such as the type of disaster, temperature, rainfall, humidity, population density, and the availability of medical care. Initially, the data is loaded, and categorical variables, including Disaster_Type and Disease_Type, are transformed into numerical values using LabelEncoder. Subsequently, the dataset is divided into training and testing subsets. The training set is utilized to train the XGBoost model, while the testing set is employed for making predictions. The accuracy of the model is assessed using accuracy_score, and the result is expressed as a percentage, indicating the model's effectiveness in predicting disease types.

Furthermore, the model is utilized to forecast the most likely disease that could occur in the event of a specific disaster type. Coded representations of disaster types, along with average values for temperature, rainfall, humidity, population density, and medical care availability, are utilized for making predictions. The predicted disease type is then converted back from its coded value to its actual designation, facilitating the identification of potential disease outbreaks during disasters by considering relevant factors and applying machine learning algorithms.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

VI. CONCLUSION

In summary, the XGBoost classifier effectively forecasts the probabilities of disease outbreaks associated with various disaster events. By incorporating critical factors such as the type of disaster, temperature, rainfall, humidity, population density, and access to healthcare, the model identifies patterns that facilitate the prediction of disease occurrences in real-world disaster situations. The accuracy measured through a testing dataset ensures a dependable assessment of the model's applicability to new data. This research highlights the importance of integrating machine learning with disaster management, enabling proactive and informed interventions to address health risks in the aftermath of disasters.

Moreover, the model's capability to predict the most likely diseases associated with specific disaster types underscores its practical significance in real-world applications. By considering environmental and societal factors in its predictions, this approach can be employed to minimize health risks related to diseases and optimize resource allocation for affected areas. As the model continues to evolve and expand, it holds the potential to significantly enhance decision-making in disaster response efforts, ultimately contributing to saving lives and reducing the impact of diseases following natural disasters.

VII. FUTURE SCOPE

1) Inclusion of Real-Time Data: Integration of the current environmental data of the present temperature, levels of humidity, and even the rainfall adds to increase the success rate for timeliness in predicting chances of outbreaks or the development of diseased conditions.

2) Advanced Models: Using advanced machine learning models, such as deep learning algorithms, can be a promising approach to improving the system's capability in detecting complex patterns and thus improving predictive accuracy, particularly in areas with diverse disaster impacts.

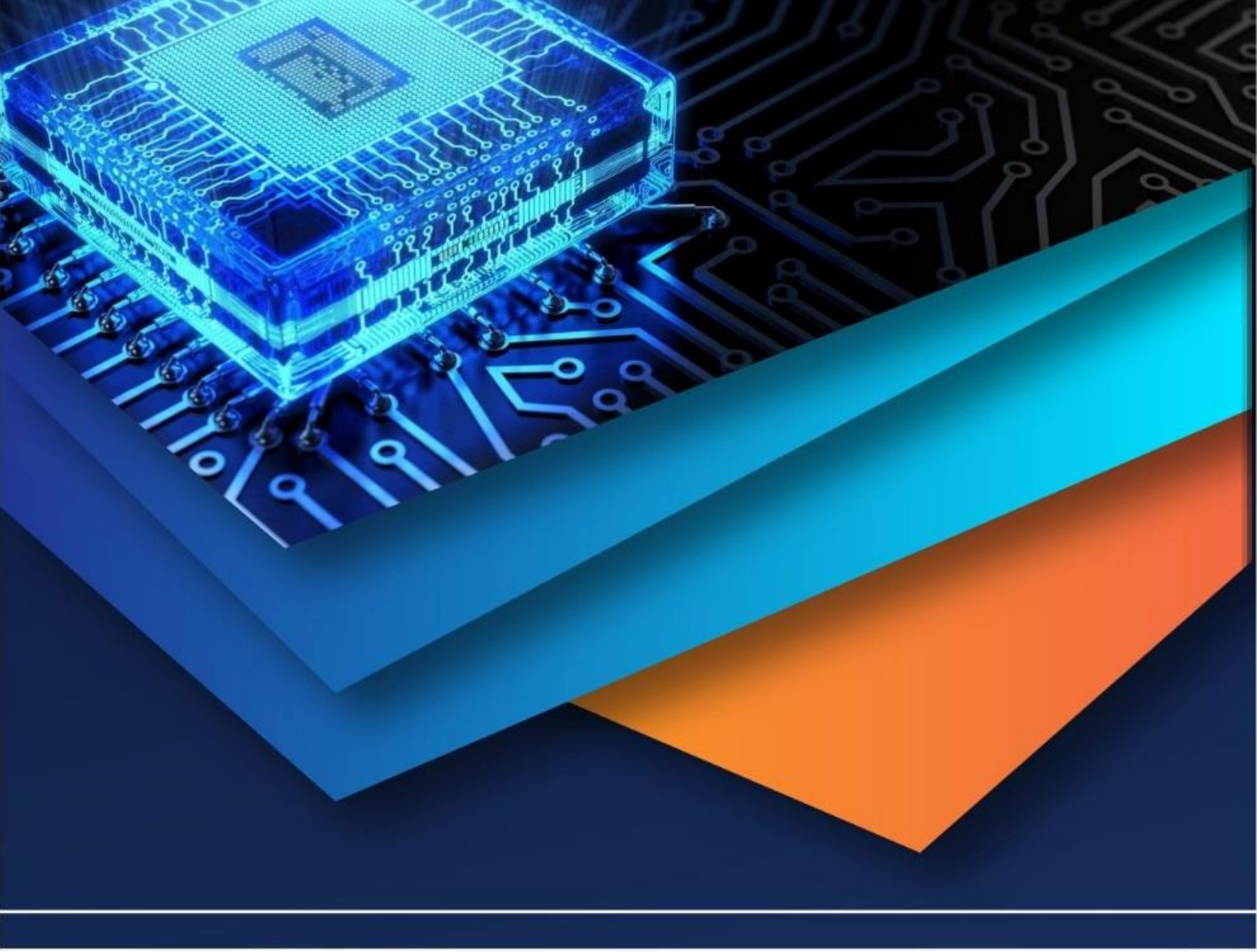
3) Broader and Region-Specific Data: Additional factors include socioeconomic variables, healthcare access, and geographical variations that can further enhance the application of the prediction models to various regions and types of disasters.

4) Implementation of Predictive Systems: Realistic development of a prediction system in real time, which the disaster management authorities can use, will help in planning health care in advance, resource mobilization, and issuing early warnings to avoid or reduce the chances of disease outbreaks.

REFERENCES

[1] Stojanovic, V. B., et al. (2020). Environmental and Social Determinants of Health in the Context of Disaster Response. International Journal of Environmental Research and Public Health, 17(22), 8497.
[2] Chen, S. J., et al. (2020). Predicting Disease Outbreaks Using Machine Learning: A Review. Journal of Medical Systems, 44(10), 184.
[3] Mendoza, A. K., et al. (2020). Machine Learning for Disease Prediction and Outbreak Response: An Application to Zika Virus. Computers in Biology and Medicine, 125, 103999.
[4] Fang, R. H., & Lee, Y. S. (2019). A Survey of Machine Learning Models for Epidemic Prediction and Forecasting. International Journal of Environmental Research and Public Health, 16(6), 957.
[5] Knight, L. J., & Goldstein, D. A. (2018). Social Vulnerabilities and Disease Transmission in Post-Disaster Settings: A Global Perspective. Journal of Disaster Research, 13(4), 781-792.
[6] Bhaduri, B., & Chou, L. (2021). Real-time prediction of health risks during and after natural disasters using machine learning. Science Advances, 7(3), eabd3456.
[7] Chakraborty, P., & Saha, S. (2019). Data-driven models for predicting the spread of infectious diseases in disaster-prone regions. Journal of Disaster Research, 14(2), 201-210.
[8] Ravindra, P., & Sahoo, B. (2017). Machine learning approaches in disaster management: A survey. International Journal of Disaster Risk Reduction, 22, 204-217.
[9] Zhang, H., et al. (2017). Real-time epidemic forecasting and monitoring in post-disaster settings using machine learning models. Nature Communications, 8, 1301.
[10] Dey, S., & Gupta, A. (2019). Machine Learning for Disaster Response: Predicting Disease Outbreaks. Computational Biology and Chemistry, 81, 73-84.
[11] Singh, A., & Kumar, A. (2020). Forecasting Disease Outbreaks Following Natural Disasters in India Using Machine Learning Algorithms. Journal of Environmental Health, 82(6), 25-34.
[12] Soni, N., & Sharma, V. (2021). Predictive Modeling of Disease Outbreaks Using Environmental and Socioeconomic Data in the Aftermath of Natural Disasters. Journal of Disaster Studies, 16(4), 319-334.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)