**Area wise Diseases analysis using machine learning.**

**Akshat Raj, Kanhaiya Mahto, Satyam Singh, Aman Raj, Nitish Kumar, Shaurya Kumar Sinha**

Department of BCA, JIS College of Engineering,

Kalyani, JIS Road, Nadia 741235, West Bengal

**Abstract:** In recent years, the utilization of machine learning techniques in healthcare has garnered significant attention due to its potential to revolutionize disease prediction and prevention strategies. This project focuses on the development of an area-wise disease prediction model leveraging machine learning algorithms. The primary objective of this research is to predict the occurrence of diseases within specific geographic regions, enabling targeted intervention and resource allocation by healthcare authorities. To achieve this, the project integrates diverse data sources including demographic information, environmental factors, socio-economic indicators, and historical health records. The methodology involves preprocessing and feature engineering to extract relevant information from the heterogeneous datasets. Subsequently, a machine learning pipeline is implemented, encompassing model selection, training, validation, and evaluation stages. Various algorithms such as logistic regression, decision trees, random forests, KNN, Naïve Bayes, Support Vector. This approach aims to provide a robust framework for disease prediction, ultimately aiding in the improvement of public health planning and response efforts.

**Key words:**

Machine Learning · Geographic Health Data · Demographic Information · Environmental Factors · Socio-economic Indicators · Historical Health Records · Data Integration · Preprocessing · Feature Engineering · Predictive Modelling · Logistic Regression · Decision Trees · Random Forests · Neural Networks · Public Health Planning · Resource Allocation · Targeted Intervention · Healthcare Strategy

**INTRODUCTION**

In this era of the modern world, our population is increasing, and urbanization carries enormous general, financial, and environmental challenges, presenting issues in urban management such as traffic resource planning, environmental quality, and public policy and safety services. Addressing health concerns in urban areas has become one of the most important social issues in large metropolitan areas as it affects people’s health, the growth of youngsters, and the socio-economic status of individuals. Disease prediction is a scheme that uses different algorithms to determine the likelihood of disease occurrence based on prior information. For our daily purposes, we need to consider health risks in various places we go every day.

In general, we often use Google Maps to find routes to our destinations. Google Maps can show multiple ways to get to a location, but we may not understand the health-related conditions of those paths. Is it safe from a health perspective? This research introduces the design and execution of a strategy based on past health data and analyses the disease rates in various areas at different times. For this work, we use primary data collected from people based on their previous health issues. In our data collection, we used different algorithms to figure out the highest precision between the machine learning algorithms that provide the greatest accuracy.

In this paper, we use different models and tables to show the prediction rates of various diseases, mostly working with data from the last three years and showing the level of disease prediction in different issues described.

**Literature Review**

For this paper, we have studied the relationship between disease prevalence and various features in the epidemiology literature. Reducing disease incidence and detecting potential outbreaks early are critical objectives, and researchers have used different techniques to achieve these goals have been utilized to predict disease occurrences effectively. Appropriate disease pattern identification and statistical analysis of hidden links using detection algorithms are crucial.

The K-Means Clustering algorithm has been used for unsupervised learning to determine disease patterns. The model was analysed, pre-processed, and implemented to test the dataset and train the algorithm, achieving clusters. Other techniques include the application of the, random forests, and naïve Bayes to predict disease hotspots and reduce health risks. These methods prepare data frames to train models for image recognition, information preprocessing, and hotspot detection.

Machine learning methods, including regression and classification, have been used to predict disease rates. Multi-linear regression has been applied to find links between dependent and independent variables. K-Nearest Neighbour’s (KNN) has been used for classification tasks involving single and multi-class variables

Classification is a unique method of information mining used to categorize each object in a dataset into predefined classes or groups. Data mining generates classification models by observing classified data and discovering predictive patterns.

**Research Methodology**

**Dataset**

The disease dataset is extracted from primary data collection based on fieldwork. This dataset consists of approximately 150000 entries with 10 key attributes. The key features include Name, Years, Months, Disease Type, Affected Areas, Patient Genders, Patient Ages, Patient Locations, and Months of Diagnosis. These features are selected as the system's input variables. The characteristics such as Patient Ages, Patient Genders, and Disease Severity are selected as the system's target variables (Table 1).

**Preprocessing**

We cleaned up our data by removing any incomplete or unknown information. We also organized dates into categories like Year, weekdays, and unknown to make it easier to analyse.

**Predictive Variables**

We're trying to predict Three things: where diseases are likely to happen, which age groups are most affected, and how severe the disease is.

**System Workflow**

We started by gathering data from different sources. Then, we cleaned it up and organized it. Next, we used machine learning techniques to predict our target variables. We tested different models to see which one gave us the most accurate predictions.

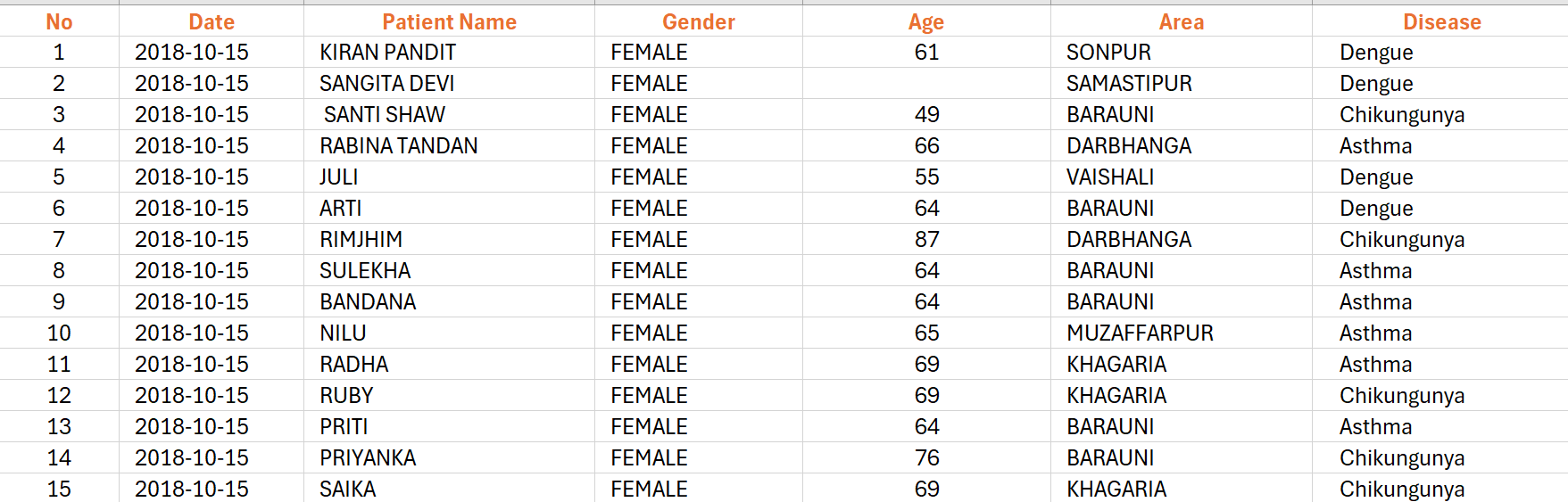
Fig. 1 **Workflow diagram**

ii. The gender is estimated using the classification of K- Nearest Neighbours.

iii. The gender of the perpetrator is estimated using the classification of K-Nearest.

neighbours.

iv. The Final prediction rate for years based on age, gender, time, and year using K-NN algorithm.

****

**Workflow diagram :**

**Fig. 1**

Analysis result

Using Python Library, matplotlib

Collect raw Data

Process the data

Area

Date

Age

**Algorithm:**

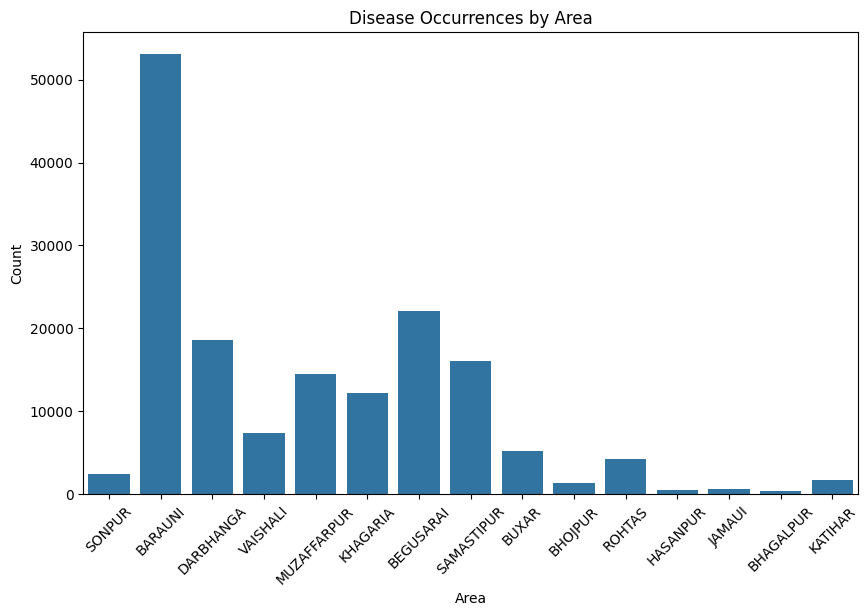
The domain contains many clustering algorithms. There is widespread use and accep-

tance of the K-means partitioning method. Apart from the K-means strategy, the

Random Forest Algorithm is the one we used because it enables consumers to

determine the number of clusters based on those values ,K-NN is also pretended

good result but above two are provides the best accuracy.

** Fig 02:**

**Fig : 03**

**K-Nearest Neighbours:**

**A graph of different colored lines

Description automatically generated** K-nearest neighbours is used when the target variable must be classified in more than two classes. In this dataset, there are three classes of target variable perpetrator gender: male, female, and unknown. Similarly, three categories of young, old, and kid are defined in age. To classify these target variables, K-nearest neighbours Classifier is used.

### Pseudo Code:

### At first KNN Classifier (Data Entry)

### Assign the Number of Cluster in K

### Choose a set of K instances to be cluster canters.

### Data points for each output

### Calculate the Euclidean range.

### Assign next to the data point the cluster.

### Perpetually Calculate centroids and reassign the variables of the cluster.

### Repeat until you reach an appropriate cluster Result, Give back the clusters and

### their values.

**Results and Discussion**

This section summarizes the study and highlights the potential for predicting disease occurrences using various machine learning algorithms and a comprehensive dataset. We analyse disease rates across different segments, including age-based, gender-based, area-based, and monthly disease rates. The data sources and methods used to guide forecasting include various disease statistics, surveys of the general population, literature reviews, and statistical models that extrapolate disease trends into the future.

Algorithmic models that describe the behaviour of observed past values can be used to forecast future disease trends by projecting a time series analysis of disease trends. Any predictive model aims to show a relationship between certain predictors and a dependent variable. To ensure greater accuracy, these models must identify and predict the scope and nature of several factors that will influence disease occurrences and their spread in the future.

**Conclusion:**

The sparsity of disease occurrences in many areas complicates the application of area-specific prediction models. In this work, we used machine learning algorithms to create and test age, gender, year, and month-based predictions of disease occurrences. We employed six types of machine learning algorithms: Random Forest, Logistic Regression, Naïve Bayes, and K-nearest neighbour, Support vector system, Decision Tree. Among these, we observed varying precision across different instances. While K-Nearest Neighbour performed well in some cases, overall, Random Forest Algorithm provided the most appreciated accuracy. Therefore, we utilized Random Forest Algorithm for our disease prediction model. By using these predictive systems, we can achieve stronger precision in the future and identify hot zone regions for disease occurrences. In future work, we aim to use the CNN algorithm to analyse image data and integrate the Google API for visualizing hot zones.