Iranian Journal of Blood & Cancer

Iran J Blood Cancer, 2024, 16(4), 20-29
https://ijbc.ir

Original Article

Predictive Modeling and Spatial Analysis of Cervix Uteri and Breast Cancer in India using Machine Learning and Big Data Frameworks

Durga pujitha Krotha[1, *], Fathimabi Shaik[2]

[1]Department of Information Technology, Velagapudi Ramakrishna Siddhartha Engineering College, Vijayawada, India.

Scan and read the article online

Citation Krotha DP, Shaik F. Predictive Modeling and Spatial Analysis of Cervix Uteri and Breast Cancer in India using Machine Learning and Big Data Frameworks. Iran J Blood Cancer. 2024 Dec 30;16(4): 20-29.

doi 0

Article info:
Received: 11 Nov 2024
Accepted: 25 Dec 2024
Published: 30 Dec 2024

Keywords:
Cancer
Big data
Machine Learning
Gradient boosting
Geographically weighted Regression

Abstract

Background: Cancer remains a critical public health issue in India, with rising cases of breast cancer and cervical cancer. Accurate predictions and spatial analysis of cancer incidence are essential for shaping prevention strategies and targeting interventions in high-risk regions.

Methods: This study utilized a big data framework employing machine learning techniques from the SparkML library to predict cancer cases and analyze spatial distributions across Indian states from 2016 to 2021. Three machine learning models used Random Forest Regressor, Gradient Boosting Regressor, and Geographically Weighted Regression (GWR) were applied to the dataset. Spatial autocorrelation analysis used Moran's I statistic to identify clustering patterns.

Results: The spatial analysis revealed significant clustering of cancer cases, particularly in 2020, with a z-score of 2.23, a p-value of 0.02, and a Moran's index of 0.15. Among the machine learning models, GWR achieved a predictive accuracy of 98% for both breast cancer and cervical cancer, while the Random Forest Regressor and Gradient Boosting Regressor achieved 95% and 97% accuracy, respectively, over the six-year period. Gradient Boosting outperformed other models in identifying key predictors and ensuring high predictive accuracy.

Conclusions: The findings highlight the efficacy of Gradient Boosting and GWR in predicting cancer incidence and analyzing spatial patterns. These models provide critical insights into cancer clustering and risk factors, supporting the development of targeted prevention strategies and policy interventions for high-risk regions in India. The results emphasize the utility of machine learning techniques in public health research and cancer control.

1. INTRODUCTION

In the era of advanced technology and data-driven insights, mapping cancer incidence in India has become more achievable using big data technologies. By tapping into vast repositories of information from national cancer registries and health databases, researchers can now analyse trends in cancer types, prevalence, and distribution across various regions of the country. By employing data analytics tools,

* Corresponding Author:
Durga pujitha Krotha
Affiliation: Department of Information Technology, Velagapudi Ramakrishna Siddhartha Engineering College, Vijayawada, India.
E-mail: **durgapujitha135@gmail.com**

[DOI: 10.61186/ijbc.16.4.20]

researchers can uncover insights into the prevalence of different cancer types, their geographical distribution, and the demographic characteristics of affected populations.

The role of big data technologies in cancer research and prevention extends beyond mapping incidence to encompass early detection, diagnosis, and personalized treatment. Machine learning algorithms can be leveraged to analyze complex datasets and detect patterns that may indicate the presence of cancer at an early stage. By integrating genetic and clinical data, researchers can develop precision medicine approaches that tailor treatment plans to individual patients based on their unique genetic makeup and disease characteristics. Furthermore, predictive analytics can monitor treatment outcomes, predict cancer recurrence, and optimize patient care pathways, thereby improving survival rates and quality of life for cancer patients.

In the healthcare sector, Big Data Analytics (BDA) will enable the use of new technologies for both health management and patient treatment.[1],[2] On the other hand, big data (BD), or unstructured data, differs from the standard formats used in data processing. Large data sets that are too big to handle, store, or analyze with conventional techniques are referred to as "big data". It is not examined; it is stored. Without a clear schema, this type of data is hard to find and analyze, so making it useful requires a certain set of tools and methodology. Integrating data that is kept in both structured and unstructured formats has many benefits for a business.[3] Big data analytics (BDA) is therefore thought to have promise

Big Data analytics refers to techniques and tools for analyzing vast amounts of data to extract and interpret information [4] Spark ML [5] can be used to forecast future events by utilizing the outcomes of big data analysis.18 Medical reports state that among Indian women of all ages, breast cancer is one of the most prevalent cancers. Breast cancer is one of the leading causes of death for women in this nation. The only way to address this is through early disease detection, which offers the best chance of enhancing treatment and curing the illness [6]. Cervical cancer remains one of the most common cancers in women globally; only breast cancer has a higher incidence than it [7].

It is possible to control, mitigate, and map factors that aid in the detection of the dynamics of the disease and its transmission with the correct information. In addition, information regarding the geographic distribution, trend, and hotspots of the outbreak can be found, as can techniques for estimating the associated risk [3] To determine which regions, the cases are grouped nationwide. To better understand the social context and the spread of the epidemic in India, we examine the spatial distribution of case incidence and its relationship to sociodemographic factors.[8], [9], [3], [10] The Getis-Ord Gi* statistic was used to determine the hotspots for cancer cases. The geographic distribution of the epidemic is a significant aspect that can be investigated using GIS and spatial statistics.[11], [12], [13] Classification and data mining techniques are helpful in data organization. Especially in the medical domain, where methods such as k Nearest Neighbors (k-NN), Support Vector Machine (SVM), Decision Tree, Naive Bayes (NB), and Analysis are commonly employed for diagnosis and decision-making.[7] Our analysis is predicated on cases that have occurred in different Indian states in chronological order. [13] employed techniques for machine learning The Random Forest Regressor, 6radient Boosting Regressor,[4] and Demographic factors have a stronger correlation with the transmission rate of cases [10].

In another context, the primary application of machine learning has been in the identification and diagnosis of cancer.[14] However a prognosis for a disease can only be determined following a medical professional's diagnosis, and a prognostic prediction needs to consider more than just the diagnosis. [15] In fact, several researchers from various fields usually employ different subsets of biomarkers and clinical factors, such as the patient's age and general health, the type and location of the cancer, and the grade and size of the tumor, to determine a cancer patient's prognosis.[16], [17], [18] The attending physician usually needs to carefully integrate data from the following sources: clinical (patient-based), histology (cell-based), and demographic (population-based) to arrive at a reasonable prognosis. Even for the most seasoned medical professional, it is difficult to finish. Predicting cancer susceptibility and preventing cancer deliver similar difficulties for patients as well as doctors. One can predict one's risk of developing cancer based on a range of factors, including age, weight (obesity), high-risk behaviours (smoking, heavy drinking), family history, diet, and exposure to environmental cancer-causing agents (asbestos, radon, UV radiation, and PCBs [19], [20], [21], [22]. However, accurate prognoses and predictions are rarely possible with the limited information available from these conventional "macro-scale" behavioural, environmental, and clinical parameters. Ideally, very specific molecular information about the patient's genetic makeup or the tumor is required [23].

Geographic information systems and spatial mapping are therefore growing in popularity across the globe. Researchers will find the study's conclusions useful when mapping any infectious disease. Autocorrelation and spatial data are necessary for geographic modeling. Since then, a lot

[DOI: 10.61186/ijbc.16.4.20]

of work has been done on developing methods and techniques for assessing spatial autocorrelation, and many geographers have been eager to use Moran's I statistic.

By combining numerous datasets into a single dataset, a significant amount of information is gathered, which facilitates the use of machine learning techniques to predict cancer cases for the following year based on six years of data. It is essential to keep in mind that the clinical process cannot be realized without the knowledge of both medicine and nursing. Additionally, map out the geographic distribution of cancer cases across the country, using methods such as spatial analysis techniques that identify higher-risk cases based on strong correlations from a variety of factors. This study's goal is to describe how cancer cases are predicted for India using machine learning techniques and spatial analysis methods based on strong correlation to identify areas at risk all around the country.

2. MATERIALS AND METHODS

2.1. Dataset Description

The data used in this paper, which cover the years 2016–2021, are from the Open Government Data (OGD) Platform India (available at: https://data.gov.in/) and India shape file from diva-gis (available at: https://data.gov.in/). A year's supply of relatively small-scale datasets is comprised of 35 data samples, each of which has four unique features. In contrast, the latter dataset comprises 216 data samples, each of which is characterized by ten distinct features. This paper uses this large-scale dataset as a prediction shown in **Table 1**.

2.2. Proposed Work

The purpose of this project is to analyze the geographical distribution of different cancer cases in India, using advanced spatial analysis and machine learning techniques. Local Moran's I analysis combined with spatial autocorrelation will be employed to identify disease clusters and hotspots. The integration of big data sources, including comprehensive cancer data and India Shape file, will enhance the richness of the analysis and provide a comprehensive view of the disease landscape. The proposed work will begin with the collection and processing of cancer-related data and the India Shapefile shown in **Figure 1**, Using Spatial autocorrelation techniques, we will measure the degree of dispersion or clustering of cancer incidence rates across different regions. Local Moran's I analysis will be used to identify significant spatial clusters with similar cancer incidence rates.

Predictions of cancer incidence will be generated using machine learning models, specifically the Random Forest Regressor and Gradient Boosting algorithms. These models will be implemented using a Spark session to leverage distributed computing for efficient processing of large datasets. Additionally, the spatial distribution and future incidence rates will be predicted using Geographically Weighted Regression (GWR) to account for spatial variability in the data. The models will be evaluated using metrics such as Root Mean Squared Error (RMSE), R-squared (R^2), and Mean Squared Error (MSE), GWR will also provide spatial predictions of cancer incidence rates, highlighting regions with higher or lower predicted rates. This systematic approach aims to provide detailed insights into the spatial distribution of cancer cases in India, supporting public health planning and intervention strategies with accurate predictive models and comprehensive spatial analysis.

2.2.1. Spatial Autocorrelation

A popular statistic for evaluating spatial autocorrelation is Moran's I, which gauges how much a variable is dispersed or clustered spatially within a given region. It assesses whether similar and dissimilar values are randomly distributed or if they tend to occur close to one another (positive spatial autocorrelation). For a variable x in a study area with n spatial units (e.g., regions or points), the formula for Moran's I is as follows (1):

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \times \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\cdot(x_i-\bar{x})\cdot(x_j-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

where:

I is the Moran's I value for the variable x

x_i and x_j are the values of the variable x at locations i and j, respectively.

$\bar{x}$ is the mean value of the variable across all locations.

w_{ij} represents the spatial weight between locations i and j. It indicates the strength of the spatial relationship between the two locations. Commonly used spatial weights include binary contiguity weights or inverse distance weights, among others.

n is the total number of spatial units in the study area. Moran's value ranges from -1 to 1.

2.2.2. Hotspot analysis

Hotspot analysis locates spatial clusters of high or low values within a dataset using a statistical method called local Moran's I analysis. By computing local measures of spatial autocorrelation for each distinct location within a study area, it expands on the Global Moran's I statistic. With the

[DOI: 10.61186/ijbc.16.4.20]

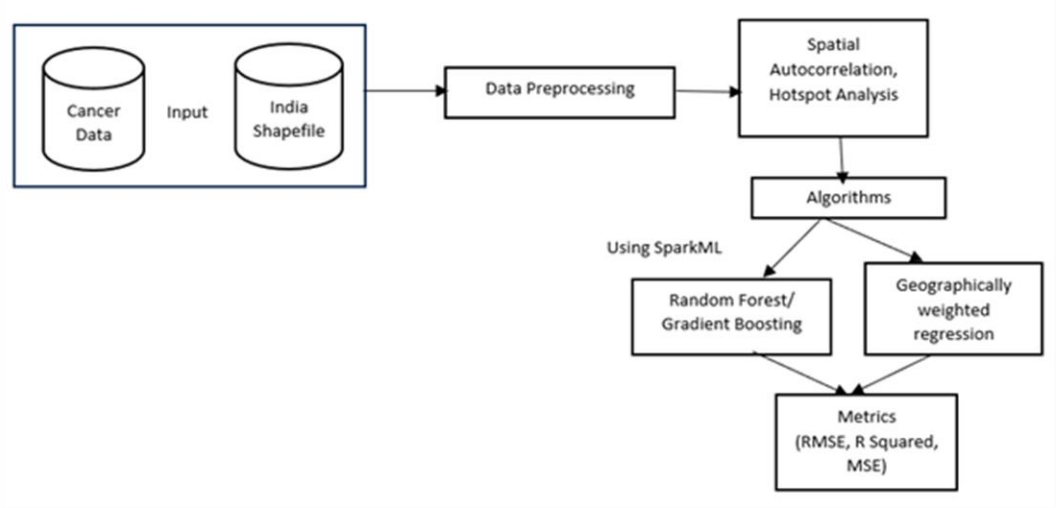

Figure 1. Proposed Work Methodology

Table 1. Cancer Incidence and Mortality Datasets: For Single-Year and Multi-Year Analysis

Small-scale Dataset	
State	The state or region name
Year	The year of data collection
Active cases	The Total number of active cases in each state
Death cases	The Total number of death cases in each state
Large-scale Dataset	
State	The state or region name
dteday	The specific date of data collection
Breast cancer	The number of reported cases of breast cancer
Cervix Uteri	The number of reported cases of cervix uteri
Active cases	The total number of active cases reported
Death cases	The total number of cancer-related deaths reported
Total cases	Total cases of cancers, active and death cases
Year	The year of data collection
Longitude	The geographic longitude coordinate of the location
Latitude	The geographic latitude coordinate of the location

use of this analysis, particular regions exhibiting notable spatial dispersion or clustering of an interest variable can be located. For local Moran's I, the formula is (2):

$$I_i = \frac{(x_i - \bar{x})}{s^2} \sum_{j=1}^{n} w_{ij} \cdot (x_j - \bar{x})$$

where:

I_i is the Local Moran's I value for location i.

x_i is the value of the variable of interest (e.g., disease incidence) at location i.

$\bar{x}$ is the mean value of the variable across all locations.

s^2 is the variance of the variable.

w_{ij} represents the spatial weight between location i and location j. It indicates the strength of the spatial relationship between the two locations. n is the total number of locations in the study area.

2.2.3. Prediction models

2.2.3.1. Random Forest Regressor

A Random Forest Regressor is an ensemble learning method that operates by constructing many decision trees at training time and outputting the average of the individual tree predictions. This approach improves the predictive accuracy and controls overfitting, making it a robust and versatile model for regression tasks, particularly when dealing with non-linear relationships and complex datasets[24].

In the context of Spark ML, the Random Forest Regressor is implemented within the 'pyspark.ml' library, which provides tools for large-scale machine learning. Here's how it operates:

- The input dataset is split into training and testing sets. Features are assembled into a vector using 'VectorAssembler', a critical step in Spark ML pipelines.

[DOI: 10.61186/ijbc.16.4.20]

- During training, multiple decision trees are constructed. Each tree is trained on a random subset of the training data (bootstrap aggregating or bagging). Additionally, at each split in the tree, a random subset of features is considered for splitting, which introduces further randomness and diversity among the trees.
- For regression tasks, the final prediction of the Random forest model is the average of the predictions from all individual trees.

The Mathematical formulation of a Random forest Regressor involves several key concepts:

Given a training dataset D with n samples, multiple subsets D_i are generated by sampling with replacement. Each subset D_i is used to train an individual decision tree T_i. For each tree, at each node, a random subset of features is elected to determine the best split. Each tree T_i produces a prediction $h_i(x)$ for an input x. The final prediction of the random forest regressor is the average of the predictions from all trees (3):

$$\hat{y} = \frac{1}{T}\sum_{i=1}^{T} h_i(x)$$

Where T is the total number of trees, and $h_i(x)$ is the prediction from the i-th tree.

The Random Forest Regressor in Spark ML is a powerful tool for regression tasks, capable of handling large-scale data efficiently using distributed computing. Its ability to produce accurate predictions, manage non-linear relationships, and prevent overfitting makes it a versatile choice for various regression problems. Implementing it is PySpark leverages the scalability and speed of the Spark framework, making it suitable for big data applications.

2.2.3.2. Gradient Boosting Regressor

Gradient Boosting Regressor (GBR) is a powerful and flexible machine learning algorithm used for regression tasks. It belongs to the family of ensemble methods and aims to improve predictive performance by combining the strengths of multiple weak learners, typically decision trees. The algorithm iteratively builds an ensemble by adding models that correct the errors of the combined model [25]. This approach effectively minimizes bias error, making GBR highly accurate and effective for various complex datasets.

The primary idea behind Gradient Boosting is to create a strong predictive model by sequentially adding weak models (typically shallow decision trees) in a stage-wise manner. Each new model is trained to correct the errors made by the previous models, thereby improving the overall accuracy. The key steps in Gradient Boosting involve initialization, residual calculation, model fitting, and model updating. The process starts with an initial model, typically a simple model that makes a constant prediction. The initial prediction is usually the mean of the target values for regression tasks (4).

$$F_0(x) = \arg min \sum_{i=1}^{n} L(y_i, \gamma)$$

Where L is the loss function, y_i are the actual values, and γ is a constant prediction.

For each subsequent iteration m, the algorithm computes the residuals (errors) of the current model $F_{m-1}(x)$. These residuals represent the difference between the actual target values and the predicted values (5).

$$r_{im} = -\left[\frac{\partial L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)}\right]$$

Where r_{im} is the residual for the i-th observation in the m-th iteration.

A new weak learner (decision tree) is trained to predict the residuals. The new model aims to capture the patterns in the residuals that were not captured by the previous ensemble of models (6).

$$h_m(x) = \arg min \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + h(x_i))$$

Where $h_m(x)$ is the new weak learner trained to fit the residuals.

The predictions of the new model are scaled by a learning rate n and added to the current ensemble model to update it (7).

$$F_m(x) = F_{m-1}(x) + nh_m(x)$$

Where $n(0 < n \leq 1)$ is the learning rate, controlling the contribution of each weak learner to the final model.

This process is repeated for a specified number of iterations or until the model performance converges. Each iteration adds a new model to the ensemble, gradually improving the overall prediction accuracy by reducing bias.

Gradient boosting Regressor is a highly effective algorithm for regression tasks, combining multiple weak learners to create a strong predictive model. Its ability to iteratively reduce bias error and its flexibility in handling different types of data and loss functions make it a valuable tool in the machine learning toolkit. When implemented in environments like Spark ML, it leverages distributed computing to handle large-scale datasets efficiently, making it suitable for big data applications.

2.2.3.3. Geographically Weighted Regression (GWR) Analysis

GWR evaluates a local model of the variable or process you are trying to understand or predict by fitting a regression equation to every feature in the dataset. GWR constructs

[DOI: 10.61186/ijbc.16.4.20]

these separate equations by incorporating the dependent and explanatory variables of the features falling within the neighborhood of each target feature. The shape and extent of each neighborhood analyzed is based on the Neighborhood type and Neighborhood Selection method parameters. Adaptive bisquare is a specific spatial kernel function used in Geographically Weighted Regression (GWR) to assign weights to neighboring observations when estimating local regression coefficients. The kernel function and the choice of its parameters, such as bandwidth, play a crucial role in how GWR models capture spatial relationships and variations. The bisquare kernel function is a type of weighting function used to emphasize observations that are closer to the center and down weight those that are farther away. This is important in spatial analysis because nearby observations are more likely to have similar characteristics and be more relevant in modeling spatial relationships.

The bisquare kernel function is defined as follows (8):

$$w(d) = \begin{cases} \left(1 - \left(\frac{d}{b}\right)^2\right)^2 & if\ 0 \le \frac{d}{b} < 1 \\ 0 & if\ \frac{d}{b} \ge 1 \end{cases}$$

Where:

w(d) is the weight assigned to an observation at a distance d from the center.

b is the bandwidth parameter of the kernel function.

GWR is particularly useful for analyzing cancer cases due to its ability to capture and model spatial heterogeneity. Cancer incidence rates often vary significantly across different regions due to factors such as environmental exposures, socioeconomic conditions, and access to healthcare. GWR helps in understanding how these factors influence cancer rates in different locations. By analyzing spatial clusters, GWR can identify hotspots with unusually high or low cancer incidence rates. This information is crucial for public health officials to focus resources and interventions in areas with higher needs. With GWR, health authorities can tailor public health interventions based on local factors. For instance, areas identified with high cancer rates due to environmental factors can be targeted for specific environmental health initiatives. GWR enhances the accuracy of predictive models for cancer incidence by incorporating spatial variability. This leads to better forecasts and more reliable identification of at-risk areas. Insights from GWR analyses inform policy decisions and resource allocation, ensuring that interventions are effectively distributed based on the specific needs of different regions.

3. RESULTS AND DISCUSSION

Moran's I is a measure of spatial autocorrelation that helps determine whether a spatial pattern is clustered, dispersed, or random. It assesses the degree to which similar values of a variable are geographically clustered or dispersed. In the context of cancer case analysis, Moran's I can identify patterns in the distribution of cancer incidences, helping to reveal whether high or low values of cancer rates are geographically clustered.

3.1. Spatial Pattern Detection

Positive Spatial Autocorrelation (+ve) indicates that similar values (e.g., high or low cancer incidence rates) are clustered together. This can reveal hotspots of cancer incidence and Negative Spatial Autocorrelation (-ve) suggests that dissimilar values are adjacent to each other, indicating a dispersed pattern.

To understand the temporal changes in the spatial distribution of cancer cases, Moran's I is computed independently for each year. This year-wise approach allows for tracking how the spatial patterns evolve over time. By performing a year-wise analysis, the project examines the spatial distribution of cancer cases over time. For instance, the calculated Moran's I values for recent years might be 2019, 2020, and 2021 are 0.10 (decreased clustering), 0.15 (peak clustering), 0.11 (reduced clustering compared to 2020). Moran's I is a valuable tool for analyzing the spatial autocorrelation of cancer cases, providing insights into whether and how cancer incidences are clustered or dispersed across geographic regions. Year-wise analysis of Moran's I enable tracking of spatial patterns over time, identifying significant trends and aiding in the development of targeted public health interventions. The peak clustering observed in 2020 underscores the importance of spatial analysis in understanding and addressing cancer incidence patterns in **Figure 2**.

Applying Local Moran's I to cervix uteri and breast cancer cases enables the identification of spatial clusters (areas with similar high or low values) and outliers (areas where the value significantly differs from surrounding areas). This local analysis is crucial for understanding the spatial patterns of cancer incidences and identifying potential risk factors associated with geographic location. In **Figure 3**, We identified High-High Clusters where high cancer incidence rates are surrounded by similarly high rates. These clusters indicate hotspots of the disease or Low-Low Clusters where

[DOI: 10.61186/ijbc.16.4.20]

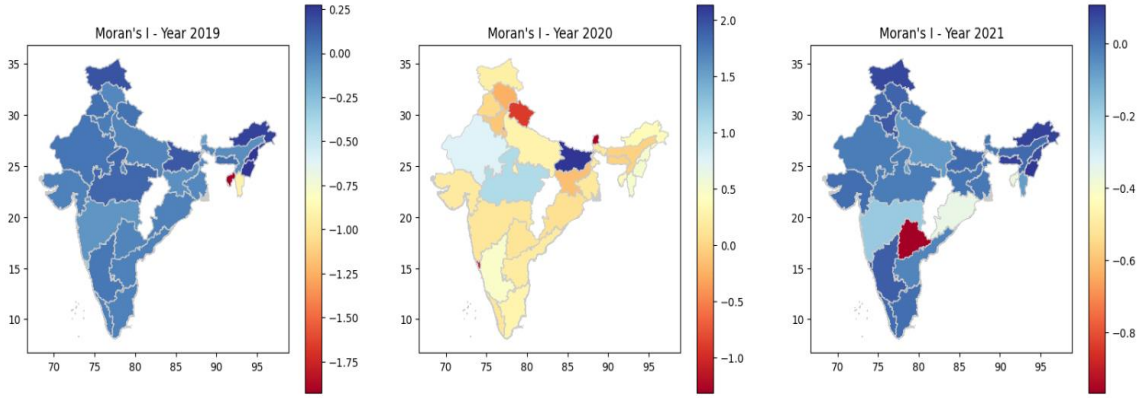

Figure 2. Spatial autocorrelation performed year-wise

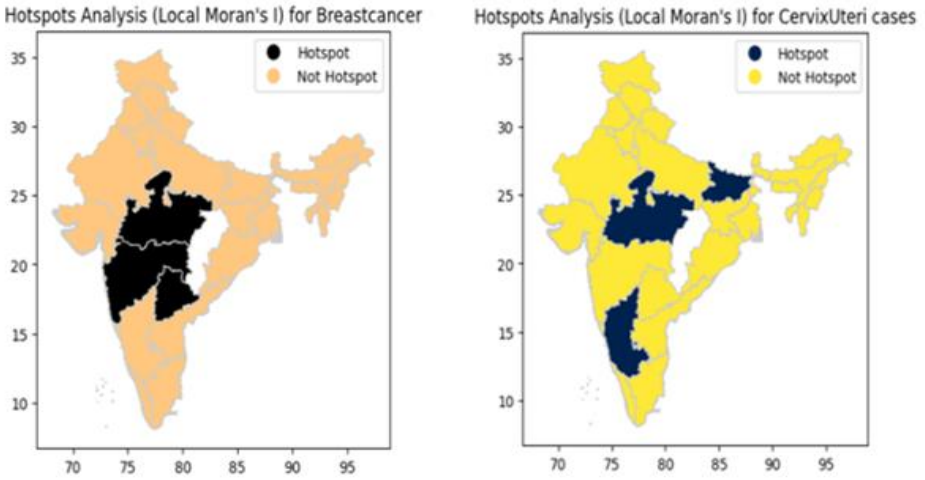

Figure 3. Hotspot analysis on (a) Breast cancer cases (b) Cervix Uteri cases

Table 2. Accuracy of cancer cases using RF and GB models.

Models	Parameters	Types of cancer	Metrics		
			RMSE	R-Squared	MSE
Gradient Boosting Regressor	GBTRegressor(labelCol='inputfeature',featuresCol='features',maxDepth=5)	Breast cancer	3669.23	0.96	134633
		Cervix uteri	646.28	0.97	417690
Random Forest Regressor	RandomForestRegressor(labelCol='inputfeature',featuresCol='features',maxDepth=5)	Breast cancer	5545.76	0.91	3075547
		Cervix uteri	449.78	0.95	202302

low cancer incidence rates are surrounded by similarly low rates. The regions may represent areas with better health care access and identified outliers, High-Low Outliers where areas have high cancer incidence rates is surrounded by lower rates. These outliers could indicate localized risk factors or Low-High Outliers where areas have low cancer incidence rate is surrounded by higher rates. These might represent regions with effective prevention. The analysis of cervix uteri and breast cancer cases using Local Moran's I reveal distinct spatial patterns. Identifying clusters and outliers helps in understanding potential, social, and healthcare-related risk factors.

3.2 Regressor Analysis

The comparison between Random Forest Regressor (RF) and Gradient Boosting Regressor (GB) using machine

[DOI: 10.61186/ijbc.16.4.20]

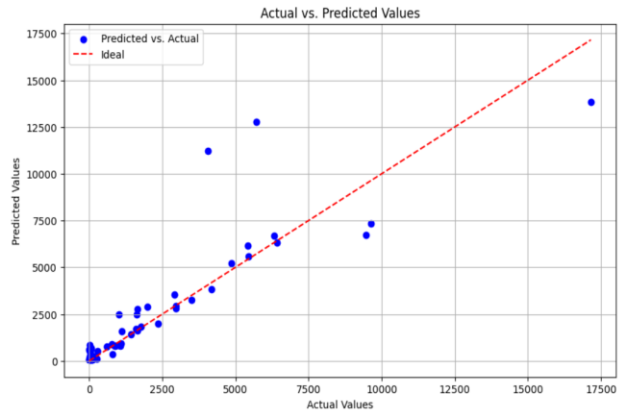

Figure 4. Actual vs Predicted of Gradient boosting Regressor for cervix uteri

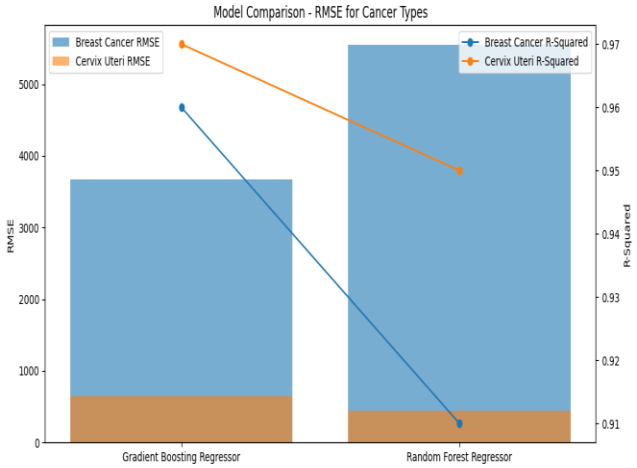

Figure 5. Model comparison across different cancer types.

learning models on cancer datasets from 2016-2021 offers valuable insights into predictive analytics in the realm of healthcare. These models aim to forecast cancer cases, a critical task for healthcare practitioners, policymakers, and researchers alike. In the analysis, both RF and GB algorithms are employed to predict breast cancer and cervix uteri cases based on six years of historical data. **Table 2** represents the results of this comparison, showcasing the predictions made by GB specifically for 2022. Notably, GB achieves a remarkable accuracy rate of 97% for cervix uteri cases, signifying its efficiency in forecasting this specific type of cancer in the Indian context. This high accuracy rate underscores the potential of machine learning algorithms, particularly GB, in aiding early detection, intervention, and resource allocation for cervical cancer, a significant public health concern.

Metrics used for Regression models: Root Mean Squared Error (RMSE) measures the average magnitude of errors in predictions. A lower RMSE indicates better predictive accuracy, R-squared explains how well the model fits the data. A value closer to 1 indicates better predictive power and Mean Squared Error (MSE) quantifies the average squared differences between predicted and actual values. A lower MSE indicates that the predicted values are close to the actual values. **Table 2** also includes parameters for each model.

A perfect model would have predictions closely aligned with the actual values, showing a straight line. It compares predicted values with actual observed values to visualize prediction performance and The bar plot is comparing RMSE values across different models for breast cancer and cervix uteri in which A line plot is used to show how well each model fits the data for both breast cancer and cervix uteri shown in **Figure 4 and 5**.

3.3. GWR Analysis

Geographically weighted regression (GWR) is a spatial analysis technique used to explore how relationships between a dependent variable, such as the "foreign-born" population, and various predictor variables differ across geographical locations. Unlike traditional regression models, which assume uniform relationships across all areas, GWR allows for localized variations by incorporating different bandwidth values in the analysis. This approach adjusts the regression parameters according to the spatial context, providing a more nuanced understanding of spatial heterogeneity. By analyzing significance levels, coefficients, and p-values for the correlations within each localized area, GWR offers insights into how factors influencing the "foreign-born" population vary across different states or regions.

In this study, GWR is employed to forecast cases of Cervix Uteri, with the results for 2020 demonstrating a strong correlation through autocorrelation techniques. The forecasted cases of Cervix Uteri are visualized in **Figure 6**, which projects these predictions onto a map of India. This map highlights the geographic distribution of Cervix Uteri cases, showing areas with higher or lower predicted incidences. Such spatial visualizations are crucial for identifying regions that may require more healthcare interventions and resources.

Analyzing GWR results involves examining local coefficients, standard errors, t-values, p-values, and other diagnostics to understand spatial variability in relationships between variables. The high R^2 and adjusted R^2 values indicate a strong fit of the model, suggesting that the predictions explain a large portion of the variability in the dependent variable. **Table 3** and **Table 4** would display these spatial patterns, showing how the relationships

[DOI: 10.61186/ijbc.16.4.20]

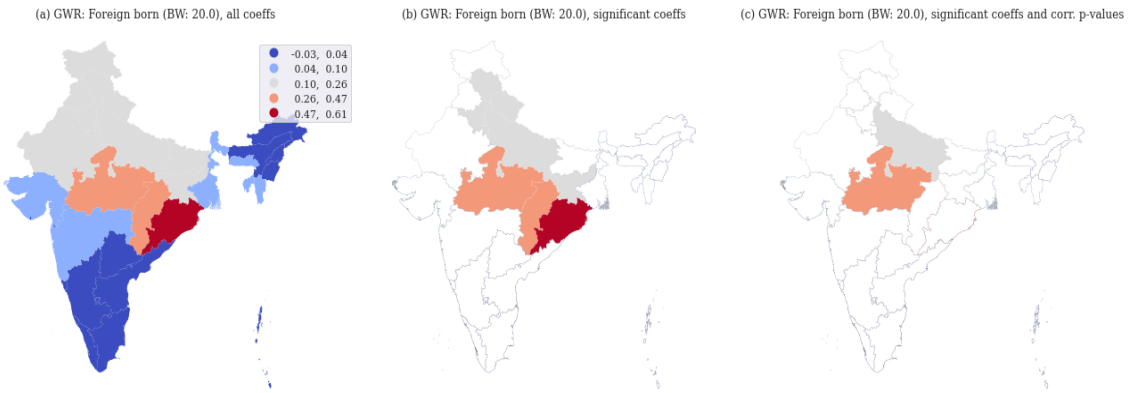

Figure 6. Three subplots (a), (b), and (c) for using GWR to visualize geographic data about individuals who were born elsewhere in the country, particularly about cervix uteri cancer. It seems to be a component of a greater spatial analysis or data visualization.

Table 3. GWR Results with Diagnostic Information

Metric	Value
Spatial kernel	Adaptive bisquare
Bandwidth used	20
Residual sum of squares	3083398.173
Effective number of parameters (trace(S))	14.236
Degree of freedom (n-trace(S))	21.764
Sigma estimate	376.393
Log-likelihood	-255.526
AIC	541.524
AICc	566.554
BIC	565.649
R2	0.986
Adjusted R2	0.977
Adj. alpha (95%)	0.014
Adj. critical t value (95%)	2.585

Table 4. Summary Statistics for GWR Parameter Estimates

Variable	Mean	STD	Min	Median	Max
X0	88.024	166.030	-213.521	96.346	312.299
X1	-0.121	0.202	-0.646	-0.022	0.088
X2	0.223	0.370	-0.165	0.043	1.77
X3	2.089	0.284	1.592	2.229	2.413

between variables change across locations. This visualization helps in pinpointing areas with stronger or weaker correlations, aiding in targeted policy-making and resource allocation based on localized needs.

4. LIMITATIONS

A lot of research depends on the quantity and quality of data, which can be biased, inconsistent, or incomplete. Variations in data sources, data cleaning procedures, and data collection techniques amongst studies may add uncertainty and impair the precision and dependability of the models or analyses carried out. Some research may narrow the scope of their recommendations or implications for more general healthcare policies or interventions by concentrating on particular diseases or datasets. This may limit the findings' applicability in other healthcare environments or geographical areas.

5. CONCLUSION

We used big data analytics to investigate the spatial distribution and prediction of cancer diseases, with a focus on hotspot analysis and spatial autocorrelation. We were able to locate important spatial clusters and hotspots of disease incidence throughout India by using statistical techniques like Local Moran's I. We were able to locate regions with comparable cancer incidence rates that tend to cluster together (positive spatial autocorrelation) by using spatial autocorrelation techniques like Moran's I. Our research also included spatial prediction in addition to spatial description. We were able to precisely predict the incidence of cancer cases over a given time frame, including cases of cervix uteri and breast cancer, across the whole research region by using Random Forest Regressor and Gradient Boosting Regressor models. In this case, the Random Forest Regressor's accuracy is lower than the Gradient Boosting Regressor's, which is 97% accurate for the Cervix Uteri. With a 98% accuracy rate, the Geographically Weighted Regression (GWR) model is used to forecast the spatial distribution of cancer cases. By applying deep learning techniques, big data technologies, and spatial techniques, the model can be extended to

[DOI: 10.61186/ijbc.16.4.20]

identify cases on the map of India. By extending the application of these techniques to other diseases, prompt responses to new health threats, early detection, and proactive disease surveillance are made possible.

Acknowledgment

We sincerely appreciate the insightful feedback and constructive suggestions from the peer reviewers and editors, which have significantly enhanced the clarity and quality of this manuscript.

Conflict of interest

The authors declare that there is no conflict of interests.

References

1. Senthilkumar SA, R.B., Meshram AA, Gunasekaran A, Chandrakumarmangalam S Big data in healthcare management: a review of literature. American Journal of Theoretical and Applied Business, 2018 4(2): p. 57-69.

2. Dash S, S.S., Sharma M, Kaushik S, Big data in healthcare: management, analysis and future prospects. Journal of big data, 2019. 6(1): p. 1-25.

3. Haider MS, S.S., Hassan S, Taniwall NJ, Moazzam MF, Lee BG, Spatial distribution and mapping of COVID-19 pandemic in Afghanistan using GIS technique. SN Social Sciences, 2022. 2(5): p. 59.

4. Shailaja K, S.B., Jabbar MA, Prediction of breast cancer using big data analytics. Int J Eng Technol, 2018. 7(46): p. 223.

5. Daghistani T, A.H., Alshammari R, AlHazme RH, Predictors of outpatients' no-show: big data analytics using apache spark. Journal of Big Data, 2020. 7: p. 1-5.

6. Bhatla N, A.D., Sharma DN, Sankaranarayanan R, Cancer of the cervix uteri: 2021 update. International Journal of Gynecology & Obstetrics, 2021. 155: p. 28-44.

7. Asri H, M.H., Al Moatassime H, Noel T, Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Computer Science, 2016. 83: p. 1064-9.

8. Kebede Kassaw AA, M.Y.T., Sebastian Y, Yeneneh Birhanu A, Sharew Melaku M, Surur Jemal S, Spatial distribution and machine learning prediction of sexually transmitted infections and associated factors among sexually active men and women in Ethiopia, evidence from EDHS 2016. BMC Infectious Diseases, 2023. 23(1): p. 49.

9. Batko K, Ś.A., The use of Big Data Analytics in healthcare. Journal of big Data, 2022 9(1): p. 3.

10. Ozyilmaz A, B.Y., Toprak M, Isik E, Guloglu T, Aydin S, Olgun MF, Younis M, Socio-economic, demographic and health determinants of the COVID-19 outbreak. Healthcare, 2022. 10(4): p. 748.

11. Jenila VM, V.P., Rajasekar SJ Geospatial mapping, Epidemiological modelling, Statistical correlation and analysis of COVID-19 with Forest cover and Population in the districts of Tamil Nadu, India, in 2020 IEEE International Conference on Advent Trends in Multidisciplinary Research and Innovation (ICATMRI) 2020, IEEE: Buldhana, India. p. 1-7.

12. Raymundo CE, O.M., Eleuterio TD, André SR, da Silva MG, Queiroz ER, Medronho RD Spatial analysis of COVID-19 incidence and the sociodemographic context in Brazil. Plos one, 2021. 16(3): p. e0247794.

13. P, G., Spatiotemporal Analysis of COVID-19 Pandemic and Predictive Models based on Artificial Intelligence for different States of India. Journal of The Institution of Engineers (India): Series B, 2021. 102(6): p. 1265-74.

14. Mccarthy JF, M.K., Hoffman PE, Gee AG, O'neil P, Ujwal ML, Hotchkiss J Applications of machine learning and high-dimensional visualization in cancer detection, diagnosis, and management. Annals of the New York Academy of Sciences, 2004. 1020(1): p. 239-62.

15. Colozza M, C.F., Sotiriou C, Larsimont D, Piccart MJ Bringing molecular prognosis and prediction to the clinic. Clinical breast cancer, 2005. 6(1): p. 61-76.

16. Burke HB, B.D., Meiers I, Montironi R Prostate cancer outcome: epidemiology and biostatistics. 2005, Analytical and quantitative cytology and histology: https://europepmc.org/article/med/16220832. p. 211-7.

17. Cochran, A., et al., Prediction of outcome for patients with cutaneous melanoma. Current Diagnostic Pathology, 2003. 9(5): p. 302-312.

18. Fielding LP, F.P.C., Freedman LS, The future of prognostic factors in outcome prediction for patients with cancer. Cancer, 1992. 70(9): p. 2367-77.

19. Leenhouts, H., Radon-induced lung cancer in smokers and non-smokers: risk implications using a two-mutation carcinogenesis model. Radiation and environmental biophysics, 1999. 38(1): p. 57-71.

20. Bach, P.B., et al., Variations in lung cancer risk among smokers. Journal of the National Cancer Institute, 2003. 95(6): p. 470-478.

21. Gasco F, V.M., Martos R, Zafra M, Morales R, Castano MA, Childhood obesity and hormonal abnormalities associated with cancer risk. European journal of cancer prevention, 2004. 13(3): p. 193-7.

22. Domchek SM, E.A., Calzone K, Stopfer J, Blackwood A, Weber BL, Application of breast cancer risk prediction models in clinical practice. Journal of Clinical Oncology, 2003. 21(4): p. 593-601.

23. Colozza, M., et al., Bringing molecular prognosis and prediction to the clinic. Clinical breast cancer, 2005. 6(1): p. 61-76.

24. Dai B, C.R., Zhu SZ, Zhang WW, Using random forest algorithm for breast cancer diagnosis, in 2018 International symposium on computer, consumer and control (IS3C). 2018, IEEE: Taichung, Taiwan. p. 449-452.

25. Al Mudawi N, A.A., A model for predicting cervical cancer using machine learning algorithms. Sensors, 2022. 22(11): p. 4132.

[DOI: 10.61186/ijbc.16.4.20]