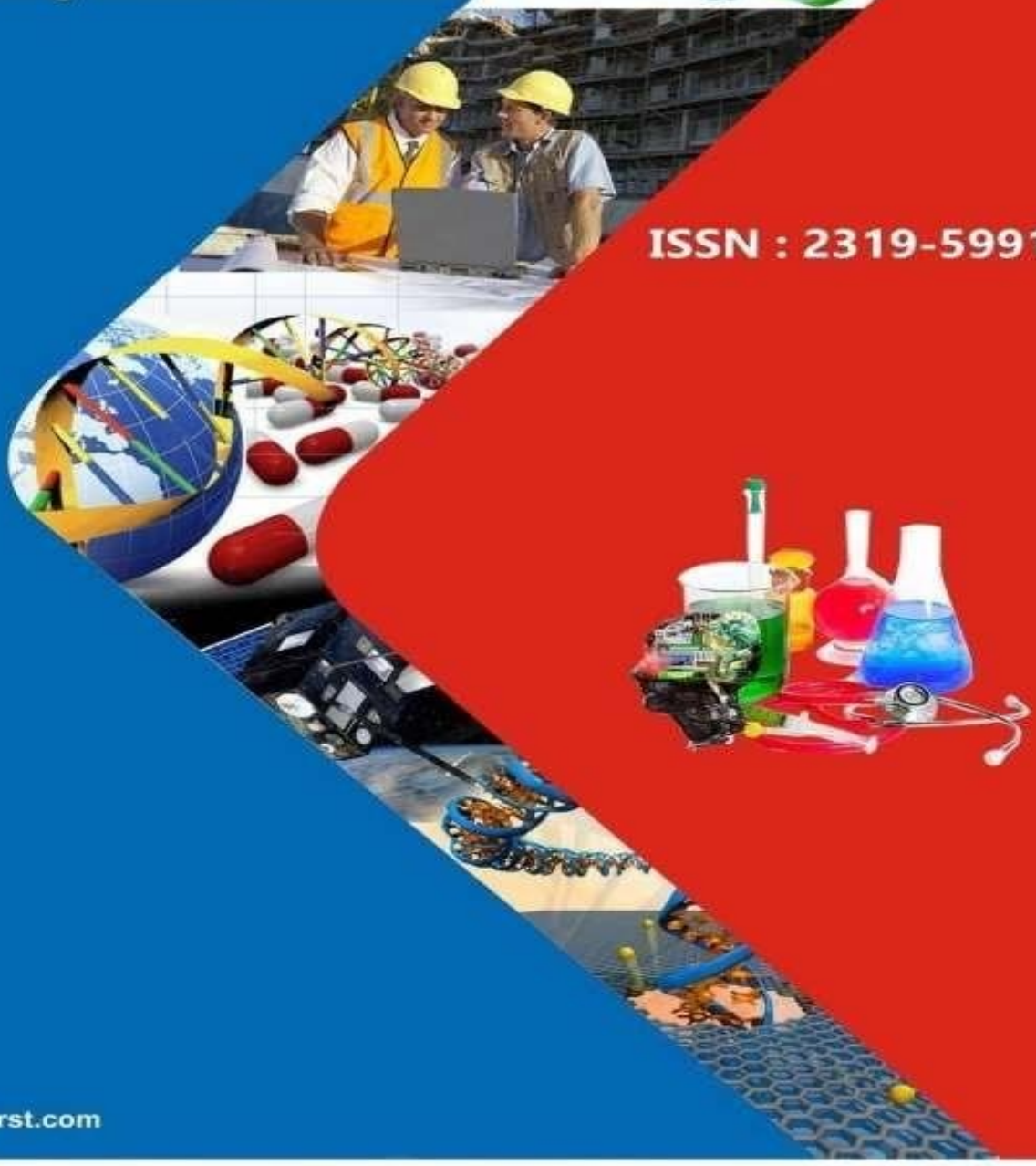
International Journal of Engineering Research and Science & Technology

IJERST



ISSN : 2319-5991

www.ijerst.com

Email: editor@ijerst.com or editor.ijerst@gmail.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 2, 2025

Using Supervised Machine Learning to Forecast Cancer Death Cases

[1] Dr.Farheen Mohammed, [2] Mohammed Majeed Ansari, [3] Syed Mohammed Amash, [4] Aiyan Ahmed
[1]Associate Professor,Department of CSE-AIML, Lords Institute of Engineering & Technology.
,[234] Student Department of CSE-AIML, Lords Institute of Engineering & Technology.

Abstract—

Cancer kills a disproportionate number of people in India compared to other countries. Using supervised machine learning techniques, this study aims to forecast cancer death rates in India. Using data from the Global Burden of Disease Study, we can see the cancer death rates in India from 1990 to 2017 broken down by gender, age group, and location. Following the completion of data pretreatment procedures such as feature engineering and missing value imputation, we use three separate supervised learning algorithms: random forest regression, decision tree regression, and linear regression. We compare the three models' performance using a number of metrics and find that the random forest regression model is the best. In order to aid the health department in their efforts, the study aims to use the best model to offer a long-term forecast of cancer mortality in India. Policymakers and healthcare professionals in India may use our study to influence their efforts to enhance cancer care and decrease cancer incidence.

Subjects—Cancer in India, Cancer Deaths, Prediction, ML with Supervision, Linear and Decision Tree Regression, RF and Polynomial Regression, and Measures for Success.

I. INTRODUCTION

Cancer is a major killer in India, just as it is elsewhere. Cancer rates are on the rise in India due to a combination of causes including an aging population, changes in diet and lifestyle, and pollution. The number of new cancer diagnoses in India is projected to reach 1.39 million in 2025, up from an anticipated 1.16 million in 2018, according to the National Cancer Registry Program. Forecasting cancer deaths in India is essential for allocating funds for cancer screenings, prevention, and treatment. In recent years, machine learning has shown encouraging advancements in a number of fields, including cancer prediction and detection. There has been substantial use of supervised machine learning techniques for illness incidence, mortality, and morbidity prediction. We introduce a supervised machine learning approach to cancer mortality prediction in India. The Global Burden of Disease Study provides cancer death statistics in India from 1990 to 2017, broken down by gender, age group, and location. The goal of this study is to compare and contrast three different supervised learning algorithms—linear regression, decision tree regression, and random forest regression—in terms of their ability to forecast cancer mortality rates in India. We analyze each model using a range of measures, including accuracy, precision, recall, and F1 score, to see which one performs the best in predicting cancer death in India over the following five years. Researchers hope these findings would help Indian policymakers and medical professionals find better ways to "prevent and cure" cancer. The "what follows is an organization of the paper's remaining sections: We review the research on medical ML and cancer prognostic prediction in Section 2. The methods utilized is detailed in Section 3. The techniques of polynomial regression that are used in the proposed framework are described in Section 4. In Section 5, we provide the findings of our investigation, including the outcomes of our model assessment and our predictions. At the conclusion of the paper, the authors address the work's limitations and suggest directions for further research.

II. LITERATURE STUDY

Cancer is a big public health problem since the rates of new cases and fatalities from it are predicted to climb in the coming years. The impact of cancer has been the subject of much study, with many studies attempting to predict cancer rates of occurrence and mortality via the use of statistical models and machine learning algorithms. The goal of this literature review is to compile the most recent findings from studies that have addressed the challenge of predicting cancer rates of incidence and mortality across different regions and different scientific perspectives. In their analysis of data from India's National Cancer Registry Programme, K. Sathishkumar et.al. [1] projected cancer incidence rates for the years 2022 and 2025. Based on the study, it is projected that the number of new cases of cancer in India would increase from an estimated 1.4 million in 2022 to 1.7 million in 2025.

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 2, 2025

Romanian researchers C. Tudor et al. came up with a game-changing approach to cancer risk assessment and prediction in [2]. The study's findings, based on Google Trends data, showed a strong correlation between online searches and cancer and mortality rates. From 2015–2025, the model projected an increase in cancer rates and fatalities in Romania. Predicting future car prices was done using regression models in [3] by R. Gupta et.al. While the findings may not be directly relevant to cancer rate estimation, they do demonstrate the potential use of regression models in this setting. With a focus on ovarian cancer, M. Dalmartello et.al. investigated the anticipated cancer mortality rates throughout Europe in 2022 [4]. The need of early detection and adequate treatment is highlighted by the study's forecast of an increase in ovarian cancer mortality in 2022 compared to 2021. Research based on data from a cancer registry indicated that both the incidence and death rate of cancer were projected to climb by 2025, according to [5] B. Trächsel et.al. The importance of cancer preventive programs in Switzerland for reducing the country's cancer incidence was also highlighted by the study. Through the application of a statistical model, K. W. Jung et.al. [6] were able to predict the cancer incidence and mortality rates in Korea up to 2022. The poll found that cancer rates are projected to climb sharply compared to the prior year. The importance of effective cancer control strategies in Korea was also highlighted by the study results. Predicting cancer rates throughout Europe was the goal of B. Sekeroglu, K. Tuncal, and colleagues in [7]. The study used data from the World Health Organization and found that machine learning models were very good in predicting cancer incidence rates. The study's findings suggest that machine learning models might be useful in the fight against and treatment of cancer. In their analysis and prediction of COVID-19, Shaikh et al. [8] used regression models and time series forecasting. Arima and Prophet, two time series forecasting methods, and linear, ridge, and Lasso, three regression models, were used to predict the number of COVID-19 cases. Based on the results, it was clear that the Prophet algorithm produced the most precise forecasts of the bunch. Research by Rahib et al. [9] used data from the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) program to model and forecast cancer incidence and death rates. An increasing tendency was seen in the incidence rates of many cancers, the most frequent of them being breast cancer and lung cancer. The authors Jain et al. Four distinct ML algorithms—support vector machine, logistic regression, decision tree, and K-NN—were used to forecast the occurrence of breast cancer. Among the models evaluated, the results showed that the SVM algorithm produced the most precise predictions. In [11] J & Jakka et.al.'s use of ML models [11]. The study used a number of machine learning models, such as support vector machine (SVM), decision tree (DT), and random forest (RF), to forecast the COVID-19 cases. When compared to the other models, the results showed that the SVM model produced the most accurate predictions. According to Xie et al. and E.-D. J. Med. Discovery, in [12] Using the Box-Jenkins ARIMA model, the study examined and forecasted cancer incidence rates. Using the ARIMA model, the incidence rates of several cancer kinds may be accurately predicted. In their study, Ai et al. [13] used a support vector machine (SVM) model to predict the occurrence of certain malignancies in individuals. The results show that the SVM model was able to accurately forecast the presence of several malignancies. The authors Dong and Taslimitehrani used a pattern-aided regression model in their work [14]. Using a regression model, this study examined disease incidence rates in relation to several demographic variables. The results proved that the regression model could accurately predict the onset of different diseases. The present status of cancer in India was discussed and recommendations for future cancer research were made by Ali et al. in [15]. The current status and future prospects of cancer research in India were the subjects of the study. The authors have put out a number of recommendations that they think may improve cancer research and control initiatives in India. Finally, these studies demonstrate the potential application of several methods, including statistical models and machine learning algorithms, to predict cancer rates and incidences in different regions. A growing cancer burden is suggested by the data in most regions, highlighting the need of effective cancer control and preventive initiatives. The study shows that there is potential for improving the accuracy of cancer predictions by integrating internet questions with machine learning algorithms.

III. METHODOLOGY

A. Dataset

Data collection including details on fatalities caused by cancer on a global scale from 1990 to 2019. Cancer rates, fatalities, and other disease-related factors are estimated globally using this data set. Anyone interested in studying patterns and trends in cancer throughout the world may use this data, which is sourced from the Worldwide Burden of Disease project. Dataset components contain a number of files that record variables such as risk factors, cancer rates, mortality, years lost to illness, years lost to disability adjusted for disability, and years lived with disability.

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 2, 2025

The 02 total-cancer-deaths-bytype.csv file includes the total number of individuals who died from each type of cancer from 1990 to 2019. Geographical region, gender, and age are the three main classification criteria used to sort the data. The global cancer burden and trends in cancer mortality may be better understood with the use of this dataset. Check out this link: https://www.kaggle.com/datasets/belayethossainds/cancerand-deaths-dataset-19902019-globally. One prominent usage of the popular statistical approach known as regression analysis is to model the relationship between a dependent variable and a collection of independent variables. This is shown in B. Regression Model. Predicting a continuous numerical number, such a product's future sales volume or price, is its intended use. Among the several types of regression models that are accessible are linear, decision tree, random forest, and polynomial regression. When it comes to regression methods, linear regression is by far the most popular and user-friendly. It takes for granted that there is a straight line connecting the two sets of explanatory variables. That is, a change in one variable will have an equivalent effect on another. By simulating the relationship between the two variables along a "straight" line, linear regression seeks to identify the value at which the actual value deviates least from the predicted value. A non-parametric choice tree regression is a regression analysis that uses decision trees. By further subdividing the data according to the independent components, a decision about the dependent variable may be achieved. Iteratively partitioning the data into subgroups based on a set of criteria that aim to make each subset more homogeneous is how the decision tree is built. The ultimate result is shaped like a tree, with each branch standing for a projected value of the dependent variable according to the model. To improve the accuracy of its predictions, random forest regression employs an ensemble approach that combines many decision trees. A large number of decision trees are generated by the method, with each tree randomly using a different subset of the data and independent variables. Once every decision tree has made its prognosis, an average prediction is determined. In polynomial regression, the "link" between the dependent and independent variables is represented as a polynomial of degree n. It comes in handy when the relationship between the variables isn't straight forward. Polynomial regression does not overlook complex data patterns, although linear regression could. This approach increases the likelihood of overfitting, which might provide erroneous predictions when used with new data. C. Criteria for Assessment By modifying the regression parameters, one may determine how well a regression model performs. Here are a few of the most common regression parameters and what they mean: • "R2 Score (Coefficient of Determination): The R2 score measures how well the data fits into the regression model. It reveals to what extent the independent variable (or variables) can explain the variance in the dependent variable. When the R2 score is perfect, it means that the model fits the data perfectly. A variable's "Explained Variance Score," or "EVS" (s), is the proportion of its fluctuation that can be explained by the independent variable. Also in the range of 0 to 1, where 1 signifies that the model adequately accounts for all variations in the dependent variable. You may see how far your estimations are from the target numbers by looking at the Mean Squared Error (MSE). It seems that the model is more in line with the data when the MSE is lower. • The MAE is a statistical metric that indicates the degree to which the actual value of a variable deviates from its predicted value. Being less prone to outliers than MSE, it provides a respectable indicator of the accuracy of the model. • Since RMSE is equal to the square root of MSE, it is equivalent to the dependent variable in units of Root Mean Squared (RMSE). "Calculating the root-mean-squared errors (RMSEs)" is a typical way to evaluate multiple regression models. Section IV: Indian Cancer Screening Procedures The following is a detailed description of each "block diagram component" for forecasting the number of cancer-related fatalities in India using supervised machine learning. In addition to conventional approaches, this article also use polynomial regression, which is both easier and faster.

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

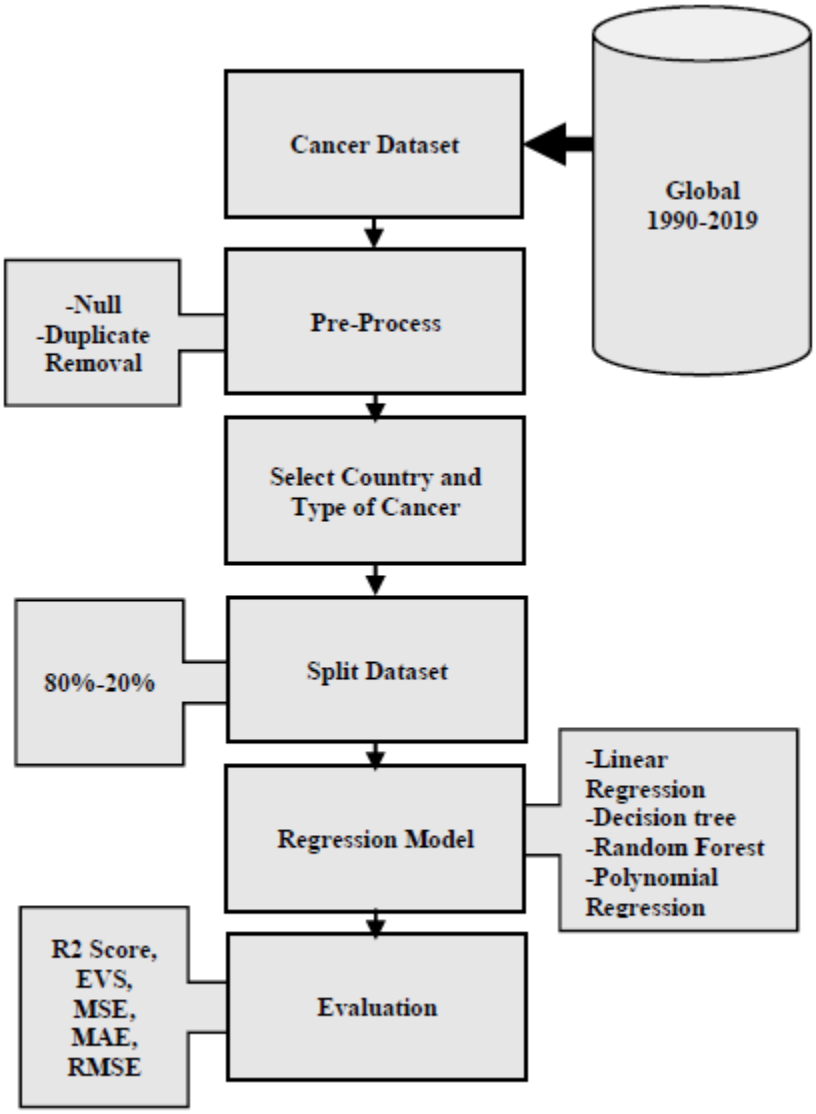


Fig. 1. Lung Sound Recognization Proposed Block Diagram

In order to predict future cancer rates, the authors suggest a regression model. The first step is to get the global cancer dataset from 1990 to 2019 that is accessible on Kaggle. In Step 2, we replace null or duplicate entries with new ones as part of the pre-processing of the dataset. The dropna() and drop_duplicates() functions in the pandas package are used for this purpose. The third step is to split the dataset in half, creating a training set and a testing set, after limiting it to a certain country and cancer kind. In the fourth phase, we'll utilize the training data to train four different regression models: linear regression, decision tree, random forest, and polynomial regression. Step 5: To evaluate the regression models, five separate metrics are used: R2 score, EVS, MSE, MAE, and RMSE to stand for root mean squared error and explained variance score, respectively. By comparing each regression model using these metrics, we can choose the one that produces the most reliable predictions.

V. RESULTS AND EVALUATION

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 2, 2025

We use a variety of indicators to evaluate our models, and they include linear regression, decision tree regression, random forest regression, and polynomial regression. We use a publicly available dataset from the Global Burden of Disease Study to forecast cancer mortality in India.

```
import pandas as pd
df=pd.read_csv("/content/02 total-cancer-deaths-by-type.csv")
pd.set_option('display.max_columns', None)
df.head()
```

	Entity	Code	Year	Deaths - Liver cancer - Sex: Both - Age: All Ages (Number)	Deaths - Kidney cancer - Sex: Both - Age: All Ages (Number)	Deaths - Lip and oral cavity cancer - Sex: Both - Age: All Ages (Number)	Deaths - Tracheal, bronchus, and lung cancer - Sex: Both - Age: All Ages (Number)	Deaths - Larynx cancer - Sex: Both - Age: All Ages (Number)	Deaths - Gallbladder and biliary tract cancer - Sex: Both - Age: All Ages (Number)	Deaths - Malignant skin melanoma - Sex: Both - Age: All Ages (Number)	Deaths - Leukemia - Sex: Both - Age: All Ages (Number)	Deaths - Hodgkin lymphoma - Sex: Both - Age: All Ages (Number)	Deat Mult myel - Bo Age: (Num
0	Afghanistan	AFG	1990	851	66	89	983	260	180	47	1055	102	
1	Afghanistan	AFG	1991	866	68	89	982	263	182	48	1089	108	
2	Afghanistan	AFG	1992	890	68	91	989	268	185	51	1171	118	
3	Afghanistan	AFG	1993	914	70	93	998	275	189	53	1252	126	
4	Afghanistan	AFG	1994	933	71	94	996	282	193	54	1296	130	

Fig. 2. Data Reading

```
TYPES_OF_CANCER = df.columns[3:]
print(TYPES_OF_CANCER)
```

```
Index(['Deaths - Liver cancer - Sex: Both - Age: All Ages (Number)',
       'Deaths - Kidney cancer - Sex: Both - Age: All Ages (Number)',
       'Deaths - Lip and oral cavity cancer - Sex: Both - Age: All Ages (Number)',
       'Deaths - Tracheal, bronchus, and lung cancer - Sex: Both - Age: All Ages (Number)',
       'Deaths - Larynx cancer - Sex: Both - Age: All Ages (Number)',
       'Deaths - Gallbladder and biliary tract cancer - Sex: Both - Age: All Ages (Number)',
       'Deaths - Malignant skin melanoma - Sex: Both - Age: All Ages (Number)',
       'Deaths - Leukemia - Sex: Both - Age: All Ages (Number)',
       'Deaths - Hodgkin lymphoma - Sex: Both - Age: All Ages (Number)',
       'Deaths - Multiple myeloma - Sex: Both - Age: All Ages (Number)',
       'Deaths - Other neoplasms - Sex: Both - Age: All Ages (Number)',
       'Deaths - Breast cancer - Sex: Both - Age: All Ages (Number)',
       'Deaths - Prostate cancer - Sex: Both - Age: All Ages (Number)',
       'Deaths - Thyroid cancer - Sex: Both - Age: All Ages (Number)',
       'Deaths - Stomach cancer - Sex: Both - Age: All Ages (Number)',
       'Deaths - Bladder cancer - Sex: Both - Age: All Ages (Number)',
       'Deaths - Uterine cancer - Sex: Both - Age: All Ages (Number)',
       'Deaths - Ovarian cancer - Sex: Both - Age: All Ages (Number)',
       'Deaths - Cervical cancer - Sex: Both - Age: All Ages (Number)',
       'Deaths - Brain and central nervous system cancer - Sex: Both - Age: All Ages (Number)',
       'Deaths - Non-Hodgkin lymphoma - Sex: Both - Age: All Ages (Number)',
       'Deaths - Pancreatic cancer - Sex: Both - Age: All Ages (Number)',
       'Deaths - Esophageal cancer - Sex: Both - Age: All Ages (Number)',
       'Deaths - Testicular cancer - Sex: Both - Age: All Ages (Number)',
       'Deaths - Nasopharynx cancer - Sex: Both - Age: All Ages (Number)',
       'Deaths - Other pharynx cancer - Sex: Both - Age: All Ages (Number)',
       'Deaths - Colon and rectum cancer - Sex: Both - Age: All Ages (Number)',
       'Deaths - Non-melanoma skin cancer - Sex: Both - Age: All Ages (Number)',
       'Deaths - Mesothelioma - Sex: Both - Age: All Ages (Number)'],
      dtype='object')
```

Fig. 3. Types of Cancer

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 2, 2025

```
Country = df['Entity'].unique()
print(Country)
```

```
['Afghanistan' 'African Region (WHO)' 'Albania' 'Algeria' 'American Samoa'
 'Andorra' 'Angola' 'Antigua and Barbuda' 'Argentina' 'Armenia'
 'Australia' 'Austria' 'Azerbaijan' 'Bahamas' 'Bahrain' 'Bangladesh'
 'Barbados' 'Belarus' 'Belgium' 'Belize' 'Benin' 'Bermuda' 'Bhutan'
 'Bolivia' 'Bosnia and Herzegovina' 'Botswana' 'Brazil' 'Brunei'
 'Bulgaria' 'Burkina Faso' 'Burundi' 'Cambodia' 'Cameroon' 'Canada'
 'Cape Verde' 'Central African Republic' 'Chad' 'Chile' 'China' 'Colombia'
 'Comoros' 'Congo' 'Cook Islands' 'Costa Rica' "Cote d'Ivoire" 'Croatia'
 'Cuba' 'Cyprus' 'Czechia' 'Democratic Republic of Congo' 'Denmark'
 'Djibouti' 'Dominica' 'Dominican Republic' 'East Asia & Pacific (WB)'
 'Eastern Mediterranean Region (WHO)' 'Ecuador' 'Egypt' 'El Salvador'
 'England' 'Equatorial Guinea' 'Eritrea' 'Estonia' 'Eswatini' 'Ethiopia'
 'Europe & Central Asia (WB)' 'European Region (WHO)' 'Fiji' 'Finland'
 'France' 'G20' 'Gabon' 'Gambia' 'Georgia' 'Germany' 'Ghana' 'Greece'
 'Greenland' 'Grenada' 'Guam' 'Guatemala' 'Guinea' 'Guinea-Bissau'
 'Guyana' 'Haiti' 'Honduras' 'Hungary' 'Iceland' 'India' 'Indonesia'
 'Iran' 'Iraq' 'Ireland' 'Israel' 'Italy' 'Jamaica' 'Japan' 'Jordan'
 'Kazakhstan' 'Kenya' 'Kiribati' 'Kuwait' 'Kyrgyzstan' 'Laos'
 'Latin America & Caribbean (WB)' 'Latvia' 'Lebanon' 'Lesotho' 'Liberia'
 'Libya' 'Lithuania' 'Luxembourg' 'Madagascar' 'Malawi' 'Malaysia'
 'Maldives' 'Mali' 'Malta' 'Marshall Islands' 'Mauritania' 'Mauritius'
 'Mexico' 'Micronesia (country)' 'Middle East & North Africa (WB)'
 'Moldova' 'Monaco' 'Mongolia' 'Montenegro' 'Morocco' 'Mozambique'
 'Myanmar' 'Namibia' 'Nauru' 'Nepal' 'Netherlands' 'New Zealand'
 'Nicaragua' 'Niger' 'Nigeria' 'Niue' 'North America (WB)' 'North Korea'
 'North Macedonia' 'Northern Ireland' 'Northern Mariana Islands' 'Norway'
 'OECD Countries' 'Oman' 'Pakistan' 'Palau' 'Palestine' 'Panama'
 'Papua New Guinea' 'Paraguay' 'Peru' 'Philippines' 'Poland' 'Portugal'
 'Puerto Rico' 'Qatar' 'Region of the Americas (WHO)' 'Romania' 'Russia'
 'Rwanda' 'Saint Kitts and Nevis' 'Saint Lucia'
 'Saint Vincent and the Grenadines' 'Samoa' 'San Marino'
 'Sao Tome and Principe' 'Saudi Arabia' 'Scotland' 'Senegal' 'Serbia'
 'Seychelles' 'Sierra Leone' 'Singapore' 'Slovakia' 'Slovenia'
 'Solomon Islands' 'Somalia' 'South Africa' 'South Asia (WB)'
 'South Korea' 'South Sudan' 'South-East Asia Region (WHO)' 'Spain'
 'Sri Lanka' 'Sub-Saharan Africa (WB)' 'Sudan' 'Suriname' 'Sweden'
 'Switzerland' 'Syria' 'Taiwan' 'Tajikistan' 'Tanzania' 'Thailand' 'Timor'
 'Togo' 'Tokelau' 'Tonga' 'Trinidad and Tobago' 'Tunisia' 'Turkey'
 'Turkmenistan' 'Tuvalu' 'Uganda' 'Ukraine' 'United Arab Emirates'
 'United Kingdom' 'United States' 'United States Virgin Islands' 'Uruguay'
 'Uzbekistan' 'Vanuatu' 'Venezuela' 'Vietnam' 'Wales'
 'Western Pacific Region (WHO)' 'World' 'World Bank High Income'
 'World Bank Low Income' 'World Bank Lower Middle Income'
 'World Bank Upper Middle Income' 'Yemen' 'Zambia' 'Zimbabwe']
```

```
years_title = df.Year.unique()
print(years_title)
```

```
[1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003
 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017
 2018 2019]
```

Fig. 4. Total Country and Years

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 2, 2025

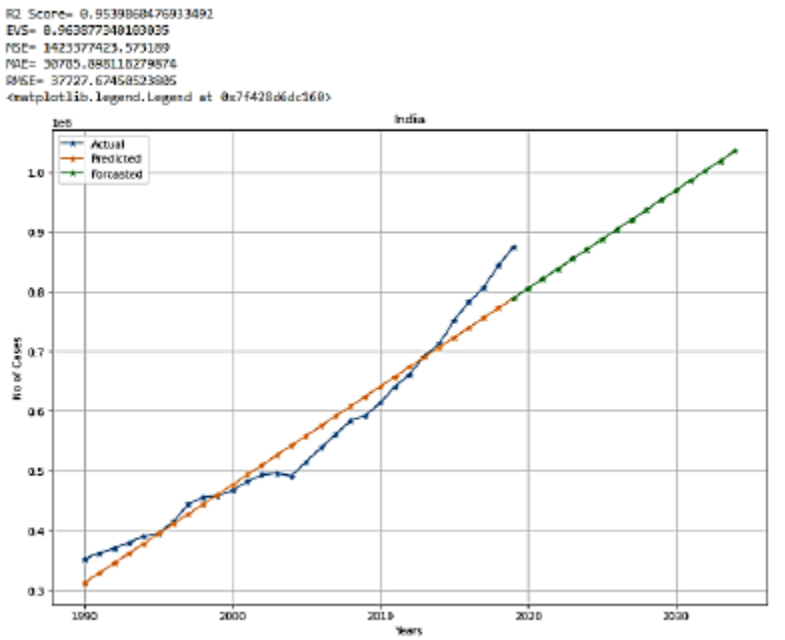


Fig. 5. LR Regression

Figure 6 displays the number of cancer cases in India on the y-axis and the years on the x-axis. That linear regression model is trained using data from 1990 to 2019, and then it is used to predict future data up to 2035.

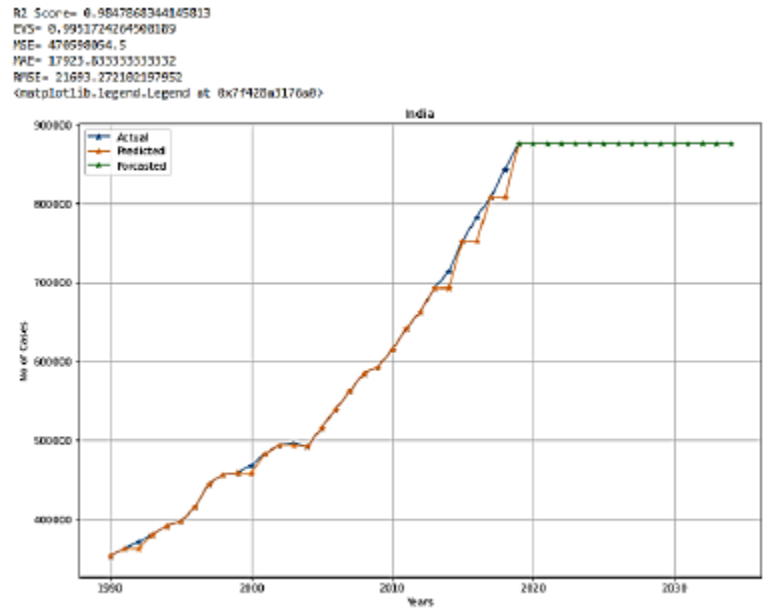


Fig. 6. DT Regression

Figure 6 displays the number of cancer cases in India on the y-axis and the years on the x-axis. The decision tree regression model is trained using data from 1990 to 2019, and then updated with data from forecasts made up to the year 2035.

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 2, 2025

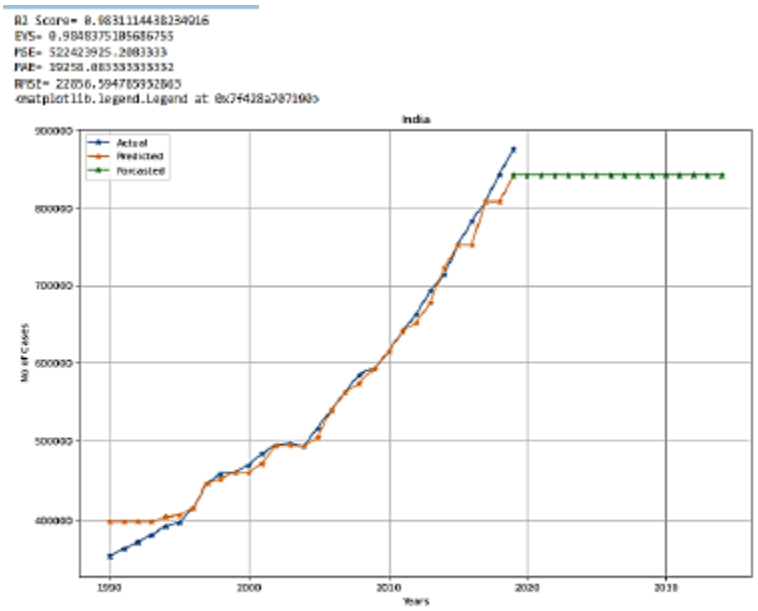


Fig. 7. RF Regression

Figure 7 depicts the x-axis representing years and the y-axis representing the number of cancer cases in India. A random forest regression model is trained using data from 1990 to 2019 and then used for future data predictions up to the year 2035.

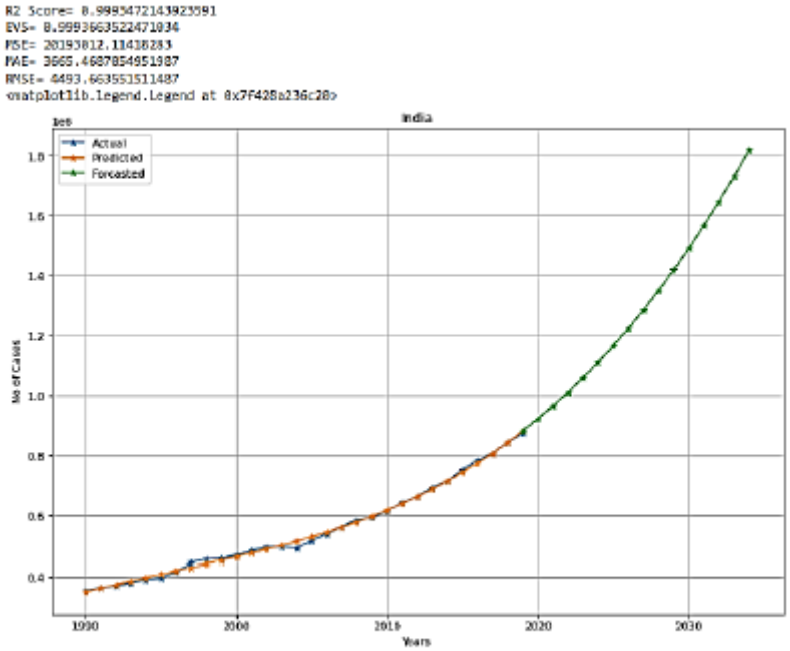


Fig. 8. Polynomial Regression

Figure 8 displays the number of cancer cases in India on the y-axis and the years on the x-axis. The data used to train the polynomial regression model spans the years 1990–2019, with further projections going all the way up to 2035.

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 2, 2025

TABLE I. COMPARE MODEL WITH EVALUATION PARAMETERS

Model	R2 Score	EVS	MSE	MAE	RMSE
Linear Regression	0.95	0.96	1423377423.57	30785.89	37727.67
Decision Tree	0.98	0.99	470598054.5	17923.83	21693.27
Random Forest	0.98	0.98	522423925.20	19258.08	22856.59
Polynomial Regression	0.99	0.99	20193012.11	3665.46	4493.66

CONCLUSION

According to these findings, supervised machine learning methods may be used to predict cancer deaths in India. Healthcare practitioners and lawmakers may be able to use this study's results to better combat cancer. Nevertheless, our results may not be applicable to a broader context due to our limited dataset and attribute selection. We used four separate supervised learning algorithms to tackle the issue of cancer death prediction in India using a publicly available dataset from Kaggle. Multiplenomial, decision tree, random forest, and linear regression were among them. When compared to other models, the polynomial regression one performs better. With values of 0.93 for R2 Score, 0.93 for EVS, 0.93 for MSE, 0.93 for MAE, and 0.93 for RMSE, the polynomial regression model is statistically sound. Future research may look at using more diverse datasets and advanced machine learning algorithms to improve the accuracy of cancer prediction in India. Adding other characteristics, such environmental, socioeconomic, and demographic factors, may enhance the forecasting model's accuracy and value. Using machine learning for cancer prediction could help with cancer control and prevention in India and other countries facing similar problems.

REFERENCES

[1] K. Sathishkumar, M. Chaturvedi, P. Das, S. Stephen, and P. Mathur, “Cancer incidence estimates for 2022 & projection for 2025: Result from National Cancer Registry Programme, India.,” The Indian journal of medical research, vol. 156, no. 4&5, pp. 598–607, 2022, doi: 10.4103/ijmr.ijmr_1821_22.

[2] C. Tudor, “A Novel Approach to Modeling and Forecasting Cancer Incidence and Mortality Rates through Web Queries and Automated Forecasting Algorithms: Evidence from Romania.,” Biology, vol. 11, no. 6, Jun. 2022, doi: 10.3390/biology11060857.

[3] R. Gupta, A. Sharma, V. Anand, and S. Gupta, “Automobile Price Prediction using Regression Models,” in 2022 International Conference on Inventive Computation Technologies (ICICT), 2022, pp. 410–416. doi: 10.1109/ICICT54344.2022.9850657.

[4] M. Dalmartello et al., “European cancer mortality predictions for the year 2022 with focus on ovarian cancer,” Annals of Oncology, vol. 33, no. 3, pp. 330–339, 2022, doi: https://doi.org/10.1016/j.annonc.2021.12.007.

[5] B. Trächsel, E. Rapiti, A. Feller, V. Rousson, I. Locatelli, and J.-L. Bulliard, “Predicting the burden of cancer in Switzerland up to 2025,” PLOS Global Public Health, vol. 2, no. 10, p. e0001112, Oct. 2022, [Online]. Available: https://doi.org/10.1371/journal.pgph.0001112

[6] K. W. Jung, Y. J. Won, M. J. Kang, H. J. Kong, J. S. Im, and H. G. Seo, “Prediction of Cancer Incidence and Mortality in Korea, 2022,” Cancer Research and Treatment, vol. 54, no. 2, pp. 345–351, 2022, doi: 10.4143/crt.2022.179.

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 2, 2025

[7] B. Sekeroglu and K. Tuncal, "Prediction of cancer incidence rates for the European continent using machine learning models," Health Informatics Journal, vol. 27, no. 1, p. 1460458220983878, Jan. 2021, doi: 10.1177/1460458220983878.
[8] S. Shaikh, J. Gala, A. Jain, S. Advani, S. Jaidhara, and M. R. Edinburgh, "Analysis and Prediction of COVID-19 using Regression Models and Time Series Forecasting," in 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2021, pp. 989–995. doi: 10.1109/Confluence51648.2021.9377137.
[9] L. Rahib, M. R. Wehner, L. M. Matrisian, and K. T. Nead, "Estimated Projection of US Cancer Incidence and Death to 2040," JAMA Network Open, vol. 4, no. 4, pp. e214708–e214708, Apr. 2021, doi: 10.1001/jamanetworkopen.2021.4708.
[10] T. Jain, V. K. Verma, M. Agarwal, A. Yadav, and A. Jain, "Supervised Machine Learning Approach For The Prediction of Breast Cancer," in 2020 International Conference on System, Computation, Automation and Networking (ICSCAN), 2020, pp. 1–6. doi: 10.1109/ICSCAN49426.2020.9262403.
[11] V. R. J and A. Jakka, "Forecasting COVID-19 cases in India Using Machine Learning Models," in 2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE), 2020, pp. 466–471. doi: 10.1109/ICSTCEE49637.2020.9276852.
[12] E.-D. J Med Discov and L. Xie, "Time Series Analysis and Prediction on Cancer Incidence Rates," Journal of Medical Discovery, vol. 2, p. jmd17030, Sep. 2017, doi: 10.24262/jmd.2.3.17030.
[13] X. X. Ai, H. Jia, and L. Xin, "SVM-based Cancer Incidence Forecasting of Patients," in 2016 9th International Symposium on Computational Intelligence and Design (ISCID), 2016, vol. 2, pp. 281– 284. doi: 10.1109/ISCID.2016.2074.
[14] G. Dong and V. Taslimitehrani, "Pattern-aided regression modeling and prediction model analysis," in 2016 IEEE 32nd International Conference on Data Engineering (ICDE), 2016, pp. 1508–1509. doi: 10.1109/ICDE.2016.7498398.
[15] I. Ali, W. A. Wani, and K. Saleem, "Cancer scenario in India with future perspectives," Cancer Therapy, vol. 8, no. ISSUE A, pp. 56–70, 2011.