Grenze International Journal of Engineering and Technology, January Issue

Linear Regression Algorithm for Early Cancer Detection and Prevention

Sachin Kumar[1], Hirdesh Sharma[2], Mayank Parashar[3], Bhawana Chaudhary[4], Ram Gopal Sharma[5] and Paridhi Kaushik[6]
[1]Associate Professor, Dept. of Information Technology, Management Education & Research Institute (MERI), New Delhi., India
Email: sachinks.78@gmail.com
[2]Assistant Professor, Department of Computer Science and Engineering, Dronacharya Group of Institutions, Greater Noida, UP, India
Email: hirdesharma@gmail.com
[3]Assistant Professor, Dept. of CSE, IMS Engineering College, Ghaziabad, UP, India
Email: mayankparashar5@gmail.com
[4]Assistant Professor, Dept. of CSE, Vishveshwarya Group of Institutions, GB Nagar, UP, India
Email: chaudharybhawana2018@gmail.com
[5]Assistant Professor, Dept. of CSE & AIML, Dronacharya Group of Institutions, Greater Noida, UP, India.
Email: cs.ramgopal@gmail.com
[6]Assistant Professor, Dept. of Computer Application, Vishveshwarya Group of Institutions, GB Nagar, UP, India
Email: pari.kaushik14@gmail.com

***Abstract*— To improve the likelihood of a successful course of therapy and long-term survival for cancer patients, a linear regression approach should be utilized to identify and prevent cancer in its early stages. Models that forecast early symptoms and cancer risk variables can be created using the linear regression algorithm.**

These algorithms can be trained on historical datasets that include the test findings, demographics, and medical histories of cancer patients. Doctors may screen patients and determine which ones are more likely to have cancer or who may already have it but be in the early stages by utilizing these prediction models.

This enables early detection and treatment, which can significantly raise the likelihood of positive results.

On the basis of each patient's unique risk factors and medical background, these models can also be used to create individualized treatment programs for them. Better treatment outcomes for patients may arise from more focused and effective care. In general, early detection and treatment of cancer using the linear regression algorithm has the potential to save lives, enhance patient outcomes, and lessen the total toll that cancer has on people and society. Overall, a comprehensive and rigorous procedure of data collection, preprocessing, feature selection, model training, evaluation, deployment, monitoring, and updating is required for the methodology of employing the linear regression algorithm to identify and prevent cancer in its early stages and it will found 98.2 % accuracy in the model.

***Index Terms*— Cancer Prevention, Breast Cancer, Linear Regression Algorithm, Cancer Prediction, Accuracy.**

Grenze ID: 01.GIJET.12.1.506

I. INTRODUCTION

Early cancer detection and treatment are critical for improving patient outcomes and reducing mortality rates. One of the leading causes of death in the globe is cancer. Lately, early cancer detection and prevention have shown promise with the use of machine learning techniques such as linear regression. One supervised learning technique that can be used is linear regression to build predictive models based on patient data [1].

Cancer is not just one disease; there are over 100 different types of the disease. When cells begin to multiply uncontrollably and create numerous tumors, cancer slowly destroys the human body. It is possible for tumors to spread and interact with other bodily parts.

Among them could be the neurological, digestive, or circulatory systems. Hormonal alterations brought on by infected bodily components alter the body [2].

Magnetic resonance imaging (MRI) and X-ray mammography (RMA) are the two most often used modalities for cancer detection. These new developments come with a number of disadvantages, including high cost, wide application, and real estate requirements for enormous healthcare facilities. False-positive results and unfavorable results can also be obtained with the aforementioned procedures [3–4].

II. RELATED WORK

Below is a summary of research that have employed linear regression algorithms for the early detection and prevention of cancer: In a study published in the Journal of Biomedical Informatics, researchers used linear regression techniques to predict the probability of breast cancer recurrence in patients.

The study found that the algorithm's projected precision rose with an increase in the genetic and clinical input [5-15].

- Sharma, N.. et al.(2015). suggest that the oral malignancies that develop in India typically Have advanced stages of cancer. It is better for early detection and prevention of oral cancer because it has a classification accuracy of 80%.
- Sharma, S. et al.(2008). says that one in two Indian women who are diagnosed with breast cancer pass away, meaning that there is a 50% risk that the case would end fatally. The outcomes are highly competitive and useful for both diagnosis and therapy.
- Skates, S. J. et al. (1995). states that the postmenopausal women older than 50 without ovarian cancer were divided at random into training and validation groups for the development of a screening test based on longitudinal marker levels. The training set was tested using the maximum risk, which confirmed a high specificity of 99.7% .
- Jahan, S., et al. (2021). states that cells of cervical cancer develop gradually at the cervix. This malignancy is successfully treatable if caught early. It has the highest accuracy on the top 25 characteristics in dataset splitting ratio (80:20).
- Patel, D. et al. (2020). says that because of their lethal character, illnesses like cancer are known as chronic fatal diseases. Because cancer spreads more quickly, it is often discovered when the disease is already advanced. To apply AI and machine learning algorithms, additional technologies such as data management and image pre-processing must also be improved.
- Al Mudawi, N. et al.(2022). states that it is becoming more and more frequent in medical practice to use ML-based methods to predict the early stages of major illnesses including cancer, kidney failure, and heart attacks. SVM, however, has been determined to have a 99% Accuracy rate.
- Lilhore, U. K., et al. (2022). states that the cancer disease are correlated with the risk of human papillomavirus infection. Around 37% of cancer cases can be avoided with preventive care, which is the most expensive cancer treatment. With the help of Python and many performance measurement criteria, such as accuracy, precision, and recall was found.
- Książek, W., et al. (2021). says that diagnosis in humans is challenging due to a variety of variables. Using machine learning approaches, a unique diagnostics strategy is proposed in this study. The proposed model had a f1-score of 93.56% and a classification accuracy of 94.55%.
- Mehmood, M., et al. (2022). says that the liver is a necessary organ for life, but because the symptoms of liver illness are often hidden, it can be exceedingly challenging to diagnose it at an early stage. It is providing the highest R2-Score of 0.8923 and the lowest MSE of 0.0618 among various ML algorithms.
- Wang, S., et al. (2021). says that more than 26 million cases of Corona virus disease (COVID-19) were found. Outcome is that AI techniques could be able to extract particular graphic aspects from COVID-19.

III. Linear Regression Algorithm used for Finding Result in Early Stages in Prevention and Detection of Cancer

Using the supervised learning method known as linear regression, regression analysis is carried out. A linear equation must be fitted to a dataset in order to make predictions about future data points. It is applicable to both simple and multiple linear regressions [16–17].
Here are the steps involved in the linear regression procedure:

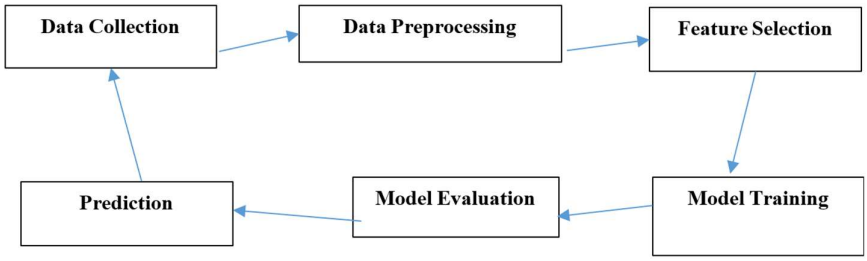

Fig. 1. Step of linear regression algorithm for early detection and prevention of cancer.

A. ALGORITHM: ADAPTIVE THRESHOLD

Step 1: Input gene dataset
Step 2: Compute neighborhood size
Step 3: Determine the threshold by utilizing the local mean intensity in the pixel's vicinity
Step 4: Set the pixel to background if its value is less than the threshold; else, make it foreground
Step 5: Clear the border
Step 6: Eliminate any little items with fewer than 100 pixels.
Step 7: Fill holes
The first phase involves gathering data, including dependent and independent variables. The second step involves preprocessing the data, which includes removing missing values and discrepancies and performing data cleansing. The dependent variable must then be forecasted, which necessitates feature selection. Following that, model training is carried out, where the selected features are used to train it.
Model evaluation is used in the following phase to evaluate the performance of the trained model. Once the model has been validated in the final step, the predictions for future data have a very high percentage [18–19].

IV. Cancer Prevention and Detection Techniques in Early Stages

Numerous methods for cancer prevention and detection can help diagnose cancer early, when it is most curable. Here are a selection of the most common techniques [20–21]:

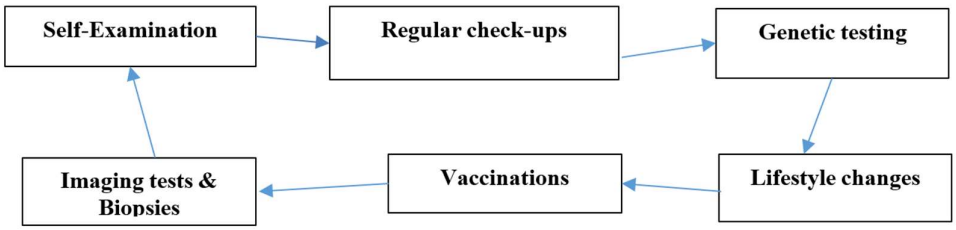

Fig. 2. Step of cancer prevention and detection techniques in early stages

Self-examination for skin and breast cancer is identified in the first phase. Regularly checking for lung, colorectal, breast, and cervical cancers as well as a few other cancers is the second phase. Following that, genetic testing is being done for colon, ovarian, or breast cancer. Adopting a healthy lifestyle can thereby reduce the chance of cancer. Maintaining a healthy weight, exercising regularly, eating a diet rich in fruits and vegetables, quitting smoking, and consuming moderate amounts of alcohol are all part of this [22].
In the following phase, immunizations can prevent the human papillomavirus (HPV), which can cause some diseases like cervical cancer. Individual risk factors may also determine whether extra vaccines are recommended. An early detection of cancer can be aided by imaging tests such as MRIs, CT scans, and mammograms [23].

A physical examination may not reveal cancers or other abnormalities, but testing may reveal them. A sample of tissue is taken from the suspected malignant location, and it is examined for cancer cells as part of the final biopsy procedure. This approach makes it feasible to determine the type of cancer and validate a cancer diagnosis [24].

V. RESULTS

This is found that every machine learning algorithm will be giving very higher accuracy which is showing in the different graphs below but linear regression algorithm will give higher accuracy to the given model.
The data has been taken on the link as https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data?resource=download and analyses it by the different machine algorithms.

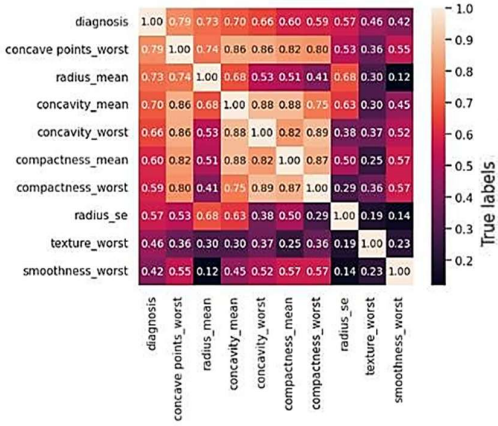

Fig. 3. Heat map of different factors involved in early cancer detection

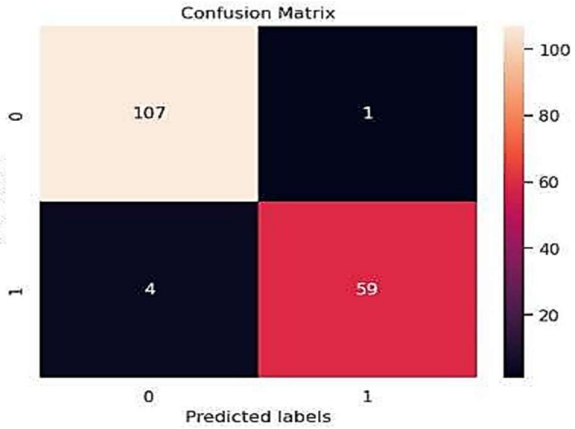

Fig. 4. Confusion matrix of predicted labels vs true labels

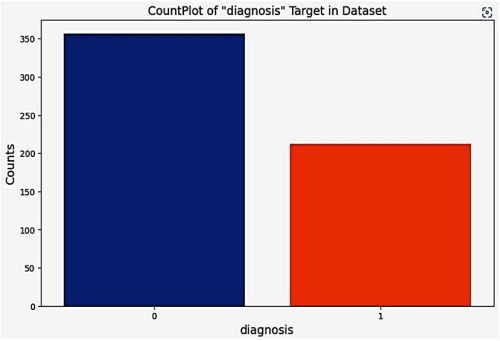

Fig. 5. Diagnosis vs Counts

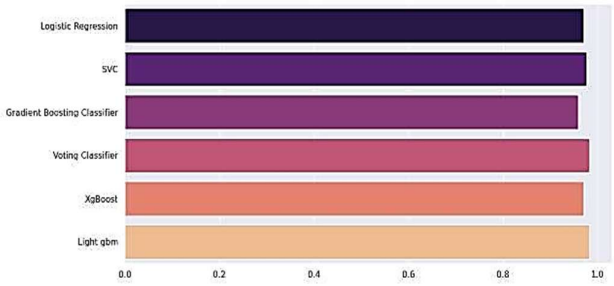

Fig. 6. Diagnosis vs Counts on different factors

VI. CONCLUSION

For a successful course of therapy and better patient outcomes, early cancer diagnosis and prevention are crucial. In order to identify the most effective preventative and therapeutic strategies and to predict the likelihood of cancer in high-risk patients, linear regression algorithms can be a powerful tool in the diagnosis and prevention of cancer.
When data on well-known risk variables are gathered and evaluated, together with the use of statistical analytic tools, linear regression algorithms can identify patterns and connections that may bear on cancer risk. Recall that there are numerous instruments at one's disposal in the fight against cancer, and linear regression algorithms are just one of them.
They should always be combined with additional diagnostic and prophylactic methods, and a trained healthcare professional should always evaluate the findings. Overall, there is great promise for improving patient outcomes and reducing the burden of this deadly illness through the application of linear regression algorithms to cancer detection and prevention. 97.2 percent accuracy is achieved.

REFERENCES

[1] Abdi, H. and Williams, L. J. (2009). Principal component analysis. In Encyclopedia of Biometrics.1(1). 12-23.

[2] Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician, 46(3). 175–185.

[3] Barandiaran, I. (1998). The random subspace method for constructing decision forests. IEEE transactions on pattern analysis and machine intelligence, 20(8).11-21.

[4] Bellaachia, A. and Guven, E. (2006). Predicting breast cancer survivability using data mining techniques. Age, 58(13).10–110.

[5] Sharma, H., Sharma, R. G., Mishra, S., Kumar, P., Parashar, M., & Dube, D. (2025, March). Hybrid DNA Differential Methylation Machine Learning Technique for Prediction of Mammary Carcinoma. In 2025 3rd International Conference on Disruptive Technologies (ICDT) (pp. 767-771). IEEE.

[6] Gupta, S., Jaiswal, V., Gupta, S., Sharma, H., & Banda, L. (2024). Heart Checkmate: An Integrated Platform for Predictive Heart Health Management. Grenze International Journal of Engineering & Technology (GIJET), 10.

[7] Delen, D., Walker, G., and Kadam, A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods. Artificial intelligence in medicine, 34(2). 113– 127.

[8] Sharma, H., Aggarwal, G., & Kumar, S. (2022). Literature Survey for Design and Analysis of a Multi-Level Algorithmic Technique for Encryption and Decryption. Grenze International Journal of Engineering & Technology (GIJET), 8(2).

[9] Mishra, A., Chaturvedi, R. P., Sharma, H., Sharma, R., & Asthana, S. (2023, November). Multi-Scale Optimized Feature Network for Polyp Segmentation. In 2023 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS) (pp. 444-448). IEEE.

[10] Kumar, S., & Sharma, H. (2021). Prediction of Sugar Production using Multivariate Linear Regression. Grenze International Journal of Engineering & Technology (GIJET), 7(1).

[11] Panagopoulou, M., Karaglani, M., Manolopoulos, V. G., Iliopoulos, I., Tsamardinos, I., &Chatzaki, E. (2021). Deciphering the methylation landscape in breast cancer: diagnostic and prognostic biosignatures through automated machine learning. Cancers, 13(7), 1677.

[12] Karaglani, M., Panagopoulou, M., Baltsavia, I., Apalaki, P., Theodosiou, T., Iliopoulos, I., ...&Chatzaki, E. (2022). Tissue-Specific Methylation Biosignatures for Monitoring Diseases: An In Silico Approach. International Journal of Molecular Sciences, 23(6), 2959.

[13] Skates, S. J., Xu, F. J., Yu, Y. H., Sjövall, K., Einhorn, N., Chang, Y., ... & Knapp, R. C. (1995). Toward an optimal algorithm for ovarian cancer screening with longitudinal tumor markers. Cancer, 76(S10), 2004-2010.

[14] Sharma, H., Gosain, R., & Sachdeva, P. (2024). Internet of Things: Vision, Applications and Challenges. Grenze International Journal of Engineering & Technology (GIJET), 10.

[15] Al Mudawi, N., & Alazeb, A. (2022). A model for predicting cervical cancer using machine learning algorithms. Sensors, 22(11), 4-132.

[16] Książek, W., Gandor, M., & Pławiak, P. (2021). Comparison of various approaches to combine logistic regression with genetic algorithms in survival prediction of hepatocellular carcinoma. Computers in Biology and Medicine, 134(1), 104-431.

[17] Mehmood, M., Alshammari, N., Alanazi, S. A., & Ahmad, F. (2022). Systematic framework to predict early-stage liver carcinoma using hybrid of feature selection techniques and regression techniques. Complexity, 20(22), 1-11.

[18] Mishra, A., Chaturvedi, R. P., Sharma, H., Kumar, P., Asthana, S., & Parashar, M. (2024, August). Brain Tumor Detection using Optimized Stochastic Gradient Descent Function. In 2024 International Conference on Electrical Electronics and Computing Technologies (ICEECT) (Vol. 1, pp. 1-6). IEEE.

[19] Jia, A. D., Li, B. Z., & Zhang, C. C. (2020). Detection of cervical cancer cells based on strong feature CNN-SVM network. Neurocomputing, 4(11), 112-127.

[20] Sharma, H., Pandey, B., Kumar, R., Mishra, S., Bhardwaj, M. K., & Pant, S. (2025). A Lag for Optimized Columnar Repositioning Methodology for Data Security. Grenze International Journal of Engineering & Technology (GIJET), 11.

[21] Xiao, Y., Wu, J., Lin, Z., and Zhao, X. (2018). Breast cancer diagnosis using an unsupervised feature extraction algorithm based on deep learning. In 2018 37th Chinese Control Conference (CCC) IEEE, 1(1). 9428–9433

[22] Zemouri, R., Omri, N., Devalland, C., Arnould, L., Morello, B., Zerhouni, N., and Fnaiech, F. (2018). Breast cancer diagnosis based on joint variable selection and constructive deep neural network. In 2018 IEEE 4th Middle East Conference on Biomedical Engineering (MECBME) IEEE. 1(1). 159–164.

[23] Sharma, H., Garg, R., Tarar, S., Sharma, S., & Tayal, S. (2024). Crop Recommendation System using Machine Learning and IoT. Grenze International Journal of Engineering & Technology (GIJET), 10.

[24] Jana, E., Subban, R., & Saraswathi, S. (2017, December). Research on skin cancer cell detection using image processing. In 2017 IEEE International conference on computational intelligence and computing research (ICCIC) (pp. 1-8). IEEE.

[25] Aggarwal, A., Saifi, M. H., Kala, R., Yadav, U., Mishra, A., Chaturvedi, R. P., & Sharma, H. (2024, August). Machine Learning Based Brain Tumor Detection and Classification Using MRI Scans with Convolutional Neural Networks. In 2024 International Conference on Control, Computing, Communication and Materials (ICCCCM) (pp. 339-342). IEEE.