Available online at http://www.journalcra.com

International Journal of Current Research

Vol. 17, Issue, 01, pp.31172-31175, January, 2025
DOI: https://doi.org/10.24941/ijcr.48385.01.2025

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

ISSN: 0975-833X

RESEARCH ARTICLE

ENHANCING THE PREDICTIVE ACCURACY OF PROSTATE CANCER OUTCOMES VIA A COMPARATIVE STUDY OF K-NEAREST NEIGHBOR AND GRADIENT BOOSTING ALGORITHMS

[1]Balaji and [2]Raja, S. R.

[1] Research Scholar, Department of Computer Applications, Centre for Open and Digital Education, Hindustan Institute of Technology and Science, India; [2]Associate Professor, Master of Computer Applications, Center for Open and Digital Education, Hindustan Institute of Technology and Science, Chennai, India

ARTICLE INFO

Article History:
Received 20th October, 2024
Received in revised form
17th November, 2024
Accepted 24th December, 2024
Published online 24th January, 2025

Key Words:
Prostate cancer, K Nearest Neighbor, Gradient Boost, SPSS Tool, Standard deviation, Algorithms, Health, Prediction.

Corresponding author:
Dr. Raja, S.R.

ABSTRACT

In this research paper is to relate effectiveness of KNN algorithm and the Gradient Boost algorithm for editing prostate cancer, in order to determine which one is more efficient. Materials and methods: This study aimed to relate K Nearest Neighbor and Gradient Boost machine learning algorithms for predicting prostate cancer. Each algorithm was run more than ten times, and the top five performing models were recorded for each. The analysis was performed on a sample size of 20, divided into two groups of N=10. Our approach achieved an accuracy rate of over 81%, suggesting potential for developing an effective prostate cancer diagnostic tool. Results and discussion: The suggested machine learning methods have the potential to improve prostate cancer diagnosis and could have a significant impact on patient outcomes. The significant value is p=0.01 which is less than the 0.05. So there is a significant variance between the two sets. Conclusion: The study highlights the significance of accurate prostate cancer prediction for early detection and effective treatment. The research results indicated that the Gradient Boost model achieved superior accuracy of 81% in comparison to KNN, which achieved an accuracy of 66%.

Citation: Balaji and Raja, S. R. 2025. "Enhancing the Predictive Accuracy of Prostate Cancer Outcomes via a Comparative Study of K-Nearest Neighbor and Gradient Boosting Algorithms". *International Journal of Current Research*, 17, (01), 31172-31175.

INTRODUCTION

Today all around the world, prostate cancer is a very famous cancer. We can say that most people between 70 to 80 years of age are affected by this cancer. It is a highly harmful disease. In this paper, we can predict the presence of cancer using two algorithms, K Nearest Neighbor and Gradient Boost. With these algorithms, we can easily determine whether prostate cancer is present or not. Prostate cancer affects a maximum of 300,000 people per year, and it is more prevalent in the US. India ranks third in the world in terms of the number of cases of this cancer. Prostate cancer affects more men than women. The prediction tools were categorized based on two factors: the patient's clinical condition and the specific outcome being predicted. The main dangerous issues for prostate cancer include family history, age, obesity, and other environmental factors. However, family inheritance is one of the main reasons for prostate cancer (1). Earlier, prostate cancer was considered one of the most challenging issues in all of medicine. The prediction model for prostate cancer can be used to detect patients who are at a high risk of increasing a violent form of the disease, and to guide decisions about screening, biopsy, and treatment options, Prostate-Specific Antigen (PSA) testing is controversial, as it can lead to overdiagnosis and overtreatment of slow-growing, non-aggressive cancers. A biopsy is a process that involves removing a small sample of prostate tissue and examining (2) it under a microscope to look for cancerous cells. Biopsies can be done using a needle that is injected into the prostate through the rectum or the skin between the scrotum and anus (Transperineal biopsy). During a process, a local anesthetic is used to numb the area. A prostate cancer prediction model can also be used for detecting cancer patients who are unlikely to benefit from aggressive treatments and who may be better served by active surveillance or watchful waiting (3). There are a total of 17,500 articles in Google Scholar, 87 on IEEE Xplore, and 444 on PubMed about Prostate Cancer. We have collected articles from these three sources because they are widely used and highly cited. For the past five years, we have collected articles, but we can find more if we search further. These sources provide a way to sort the best articles and make them easier to find. These sources are popular among authors and researchers. There are also other sources, such as Science Direct, Web of Science, and Elsevier, which can be useful for finding additional articles. In 2021 Adamaki and Zoumpoulis referred Prostate cancer biomarkers are detected from diagnosis to prognosis and using therapeutics of precision-guided (4). Artificial neural network (ANN) has the best algorithm for predicting survivability of prostate cancer patients because it can learn from data and identify complex patterns that are difficult to detect using traditional machine learning techniques(5). ANNs are composed of multiple layers of interconnected nodes that process information and learn from it. They can be trained to recognize patterns in data and make predictions based on those patterns. The existing algorithm for prostate cancer prediction has shown a low accuracy rate, indicating a need for improvement in this area. Our goal is to present a more accurate and effective solution in our article. The current lack of accuracy has motivated us to pursue this research, as prostate cancer affects an important portion of the population. It is also one of the most harmful and dangerous types of cancer in the world. The main objective is to identify more accurate methods for predicting prostate cancer and to protect individuals from developing this form of cancer. The main aim in regards to prostate cancer is to detect it at the earliest stage. It produces an appropriate treatment to increase the chances of a positive outcome (6). This may involve regular screening, such as through prostate-specific antigen (PSA) tests changes in lifestyle to reduce the risk issues connected with the development of prostate cancer.

MATERIALS AND METHODOS

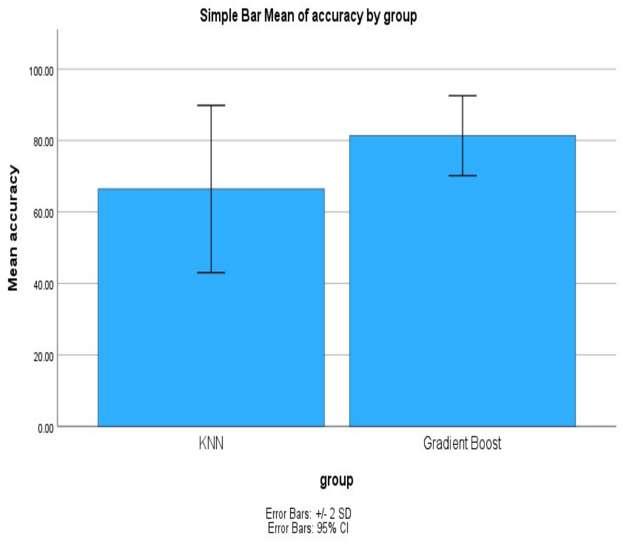
The analysis was performed with a whole sample size of 20, divided into two groups of N=10 (Collins *et al*. 2021). Ethical approval is not required in this study. An alpha value of 0.05 was used as the significance level, which is a common practice in the medical literature (7). The power of the study was set to 80%, indicating that the study was designed to detect a difference between the two groups if one exists. The study's power was calculated as 1-beta, where beta represents the probability of failing to detect a difference between the two groups when one truly exists. Both groups had an environmental ratio of 1, indicating that the study environment was the same for both groups. The study used an equal enrollment ratio between the two groups, set at 1 to ensure fairness and minimize potential biases. Clinicalc.com, a popular website providing sample size calculators, was used to find the appropriate sample size for the study. The accuracy reported in the research papers for the KNN algorithm is quite low, and we need to improve it. To achieve higher accuracy, we will use a different model that is better suited to our needs. By doing so, we hope to achieve more accurate results than the KNN algorithm. During the group preparation, we compared the accuracy of gradient boost and KNN for healthcare purposes. After analyzing the results, we found that gradient boost provided better accuracy than KNN (8). Therefore, we decided to use this algorithm for our prostate cancer prediction project in our research paper. We propose using gradient boost as our primary algorithm. Our research project heavily relied on the use of SPSS, a popular statistical analysis software program used by researchers to interpret complex data sets. Utilizing Google Colab, a cloud-based platform with sample resources, we began our study by preprocessing the data set, including removing null or missing values and outliers. We then split the data set into two groups, with 81% being used to train the model and 19% for testing its performance. We trained separate models for each group, with Group 1 using Logistic Regression algorithm and Group 2 using Gradient Boost algorithm. To evaluate the accuracy of prediction, precision, F1 score, and recall score, we conducted the same testing procedure for both groups to compare their performance with both algorithms used for predicting heart disease in the patient dataset. Our study would not have been possible without SPSS, allowing us to perform complex data analyses and generate meaningful results.

Pseudo code

Group 1: K-Nearest Neighbor

KNN is the most important machine learning algorithm and also mainly used for classification and regression.

```
1. Import the "pandas", "numpy", "train_test_split",
   "Standard Scaler", “K-Neighbors Classifier"
   "Classification_report", "score of the accuracy, f1, recall,
   precision and model" from the "sklearn" library.
2. Load the dataset using Pandas and assign it to a variable.
3. Create a label encoder and use it to encode the
   diagnosis_result column.
4. Preprocess the data by dropping any rows with missing
   values.
5. Create empty lists to store the performancemetrics.
6. For each run in a loop of 10 runs:
   a. Divide the dataset into two groups. One is training and
   another for testing sets using "train_test_split".
   b. Standardize the features using "StandardScaler".
   c. Create a KNN model using "K-Neighbors Classifier".
   d. using KNN model on the training set for fit
   e. Predict the labels for the testing set using "predict".
   f. Compute the performance metrics using "score of the
   accuracy, f1, recall, precision and model"
   g. To add the performance metrics to the respective lists
   h. Print the performance metrics.
7. Compute the mean metrics across all runs using
   "np.mean".
8. Print the mean metrics.
9. Print the classification report using
   "classification_report".
```



Group 2. Gradient Boost Algorithm

Gradient Boost is one of the machine learning algorithms. It is mainly used for regression and classification.

Table 1. Prostate Cancer disease prediction with GB and KNN

	Levene's test for equality of variance					Significance				95% confidence Interval of Difference	
		F	Sig.	t	df	One Sided P	Two Sided P	Mean Difference	Std. Error difference	Lower	Upper
Accuracy	Equal variances assumed	4.3	.051	-3.640	18	<.001	.002	-14.92	4.10	23.54	-6.311
	Equal variances not assumed			-3.640	12.9	.002	.003	-14.92	4.10	-23.78	-6.062

Table 2. Classification report for KNN and Gradient Boost consists of performance equivalence like accuracy, and score of the f1, recall and precisions

	Group	N	Mean	Std Deviation	Std Mean Error
Accuracy	KNN	10	66.44%	11.696 %	3.698
	Gradient Boost	10	81.37%	5.594%	1.769

Table 3. Independent samples comparing Gradient Boost for prostate cancer with KNN algorithm, 95% is the confidence interval where p = 0.051 (p>0.05) significance value. There is no difference between the two sets

Algorithm	Accuracy	f1_score	Recall score	Precision score
KNN	83%	87%	93%	83%
Gradient boost	87%	92%	92%	92%

It involves combining multiple weak models to create a strong model by iteratively minimizing the loss function through gradient descent.

1. Import necessary libraries: pandas, numpy, train_test_split, StandardScaler, gradient Boosting Classifier, classification_report, Label Encoder.
2. Store the dataset from a CSV file using pandas read_csv function and load it in a variable called "data".
3. Create a LabelEncoder object and use it to encode the "diagnosis_result" column in the dataset to numerical values.
4. Preprocess the data by dropping missing values from the dataset using the function.
5. Divide the data into 2 sets. One is for training purpose and remaining for testing purpose. Using the train_test_split function from k learn.model_selection. Use a test size of 0.5 and a random state of 9.
6. Create a Standard Scaler object and use it to standardize the training set.
7. Create a Gradient Boosting Classifier object and store it in a variable called gradient_boost".
8. Create a list of test sizes ranging from 0.1 to 0.9 with a step of 0.05.
9. Create an empty list called "accuracies".
10. Iterate over the test_sizes list and for each test Size:
    a. Divide the data into two sets one for training and another for testing Sets. Using train_test_split function with the current test size and a random state of 9.
    b. Standardize the two data sets using the Standard Scaler object created earlier.
    c. Fit the Gradient Boosting Classifier model for training set.
    d. Using trained model to predict the labels for testing set.
    e. Compute the model accuracy by relating. The labels predicted to the original labels and calculating the mean.
    f. Add the accuracy to the accuracies List.
    g. Print the test size and accuracy for each iteration.
11. Compute the mean accuracy across all runs by using the np.mean function on the accuracy list.
12. Print the mean accuracy.
13. Calculate the model performance using the classification_report function from sklearn.metrics.

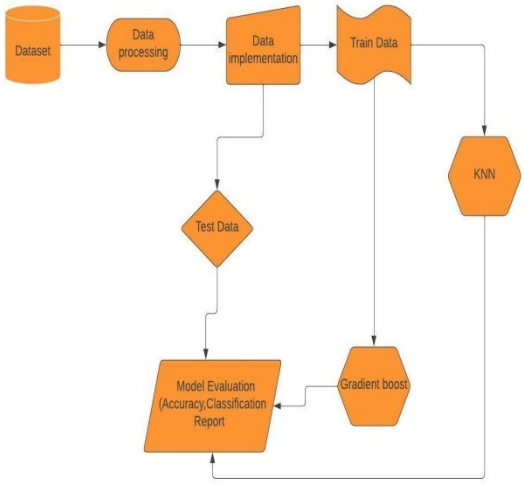


Fig. 1. Flow chart dispatched the methodology adopted in the study

Statistical Analysis

SPSS software version 29 used the SPSS tool for our statistical software. The groups K Nearest Neighbor and Gradient Boost, for calculating the mean, Std-deviation, Std-error, and t-test to compare the performance of two algorithms. We will choose the test with an independent sample t-test and then select our variables.

This will show the mean accuracy value for our variables. Furthermore, to determine the precision of our findings, we calculated a 95% confidence interval.

RESULTS

Table 1. The mean accuracy of Gradient Boost for Prostate Cancer was 81.37, which was significantly higher than the mean accuracy of KNN Algorithm, which was 66.44. The standard deviation of Gradient Boost was 5.594, which was significantly lower than the standard deviation of KNN Algorithm, which was 11.696. The error value for KNN Algorithm was 3.698, and the error value for Gradient Boost was 1.769.

DISCUSSION

A study was shown to compare the presentation of KNN and Gradient Boost algorithms in predicting prostate cancer. The results indicated that Gradient Boost outperformed KNN in terms of accuracy score (9). However, the study's determination should be taken with caution due to the relatively minimum model size and the violation of the assumption of equal variance between the two groups. To confirm these results, future research with a larger sample size and alternative statistical methods is necessary. In summary, KNN and Gradient Boost algorithms can enhance the prostate cancer accuracy value prediction models. KNN is an established algorithm that has been successful in various applications (10), whereas Gradient Boost is a newer algorithm that has shown potential in surpassing KNN in certain contexts. Choosing the appropriate algorithm depends on the specific data characteristics and prediction model objectives (11). The current study used two different models (KNN and Gradient Boost) to predict prostate cancer. In the future research could explore the use of other statistical methods, such as logistic regression, decision trees, or neural networks, to improve the accuracy of these models. The current study used age, diagnostic results, and radius, texture, area, and symmetry results as predictor variables for prostate cancer prediction. In the future research could explore the inclusion of other variables like race/ethnicity, PSA, and lifestyle factors, to further improve the accuracy of these models(12).

CONCLUSION

In conclusion, the Gradient Boost model had a higher accurate rate of 81% related to the KNN model which had an accurate rate of 66%. This indicates that the Gradient Boost model may be a more effective tool for predicting prostate cancer.

REFERENCES

Huang *et al*, 2020, Psychometric Properties of the parental Boarding Instrument in a Sample of Canadians Children", child Psychiatry and human Development 51(5):754-68

Buuhigas *et al* 2022, "The Architecture of Clonal Expansions in Morphologically Normal Tissue from Cancerous and non-Cancerous Prostate Molecular Cancer".

Heidenreich *et al* 2008, "EAU Guidelines on prostate cancer ",European Urology 53(1):68-80

Adamaki *et al* 2021 "Prostate Cancer Biomarkers: From Diagnosis to Prognosis and Precision Guided therapeutics", Pharmacology & therapeutics 228: 107932

Selvi, H., Saravanan, M.S. "Using machine learning techniques for predicting and diagnosing dyslexia in school age children", 2019, Journal of Advanced Research in Dynamical and Control Systems, Volume:11, Issue: 4, Page:753-760.

Fleshner et. al 2017, "Prostate Cancer Prevention: Past, Present and Future", cancer 110(9):1889-99

Collins *et al* 2021, "Sample Sizes of prediction Model studies in prostate cancer were rarely justified and often insufficient, " Journals of clinical Epidemiology 133(May): 53-60

Rebbeck *et al* 2017, "Prostate cancer genetics: variations by Race, Ethnicity and Geography", Seminars in Radiation oncology 27(1):3-10

M.Rhifky Wayahdi *et al* 2022, "KNN and XGBoost Algorithm for lung cancer prediction", Journal of Science Technologies 4(1):179-86

Ritch *et al* 2018, "Recent Trends in the management of Advanced prostate cancer", F1000Research 7 (September)

Selvi, H., Saravanan, M.S. "A Study of dyslexia using different machine learning algorithm with data mining techniques", 2018, International Journal of Engineering and Technology(UAE), Volume:7, Issue: 4, Page: 3406-3411, DOI:10.14419/ijet.v7i4.14545

Nagpal *et al*, 2020, "Development and Validation of a Deep learning algorithm for Gleason Grading of prostate cancer from biopsy specimens", JAMA oncology 6(9): 1372-80.

****