International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 04 | April - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Age-Specific Comparative Analysis of ML Algorithms for Colon Cancer Classification Using Gene Expression Data

Ch.Harini Devi[1], Dr Vanitha Kakollu[2]
[1] **UG** Student, GITAM Deemed to be University, Visakhapatnam
[2] Assistant professor, GITAM Deemed to be University, Visakhapatnam

ABSTRACT: The field of machine learning (ML) has made significant contributions to the medical domain, notably in the automatic classification of cancer types such as colon cancer. This type of cancer, while generally more common among the elderly, can affect people at any age. The incorporation of ML techniques has minimized the reliance on manual examination and enhanced the capacity to process large datasets in medical research. Utilizing advanced gene technology and analysis of gene expression data, this study addresses the challenges posed by the high dimensionality and limited size of such datasets. The findings reveal that most of the ML models employed here have surpassed expected accuracy levels, highlighting their effectiveness in the precise classification and diagnosis of colon cancer. Consequently, these developments hold promise for improving patient care and the overall efficiency of healthcare services in dealing with colon cancer.

Keywords: Machine Learning, Colon Cancer, Support Vector Machine, Random Forest, Decision Tree, K-Nearest Neighbour.

1. INTRODUCTION

Colon cancer ranks among the top causes of cancer-related deaths worldwide. Its early detection significantly improves survival rates, underscoring the necessity for precise and efficient diagnostic tools. This research aims to evaluate and compare the effectiveness of machine learning (ML) models for colon cancer classification. Models such as Logistic Regression, Support Vector Machines (SVM), Random Forests, Gradient Boosting (XGBoost, LightGBM), and Neural Networks will be applied to curated datasets, including histopathological and clinical data. The study uses preprocessing techniques, feature selection, and hyperparameter optimization for robust training. Metrics like accuracy, precision, recall, F1-score, and ROC-AUC are employed for evaluation. Results indicate ensemble models and neural networks outperform simpler algorithms, contributing to advanced AI tools for precision oncology.

KNeighborsClassifier: The KNeighborsClassifier is a supervised machine learning algorithm used for classification tasks. It belongs to the family of instance-based or lazy learning algorithms. Instead of learning explicit models from the training data, it memorizes instances of training data and classifies new instances based on their similarity to existing examples. The algorithm works by calculating the distance between the input data point and its k nearest neighbors in the feature space, where k is a user-defined hyperparameter. Classification is then performed based on the majority class among the k neighbors, typically using a simple majority voting scheme. KNeighborsClassifier is versatile, intuitive, and easy to implement, making it suitable for various classification tasks, including medical diagnosis, pattern recognition, and recommendation systems. However, its performance heavily depends on the choice of the distance metric and the value of k, and it may suffer from high computational costs, especially with large datasets.

Naïve bayes: Naive Bayes is a probabilistic classification algorithm based on Bayes' theorem, which assumes independence among features. It

II. MACHINE LEARNING ALGORITHMS CLASSIFICATION

Random forest: The Random Forest classifier is a powerful ensemble learning method used for classification tasks. It operates by constructing multiple decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees. Each decision tree is trained on a random subset of the training data and a random subset of features, promoting diversity among the trees.

DecisionTree: The DecisionTree is a machine learning algorithm used for classification tasks. It constructs a tree-like structure by recursively partitioning the feature space into smaller subsets based on the values of input features. At each node of the tree, the algorithm selects the feature that best splits the data, aiming to maximize the purity of the resulting subsets in terms of class labels.

Support Vector Machine (SVM) : Support Vector Machine (SVM) Classifier is a powerful supervised learning algorithm used for classification and regression tasks. It works by finding the hyperplane that best separates different classes in the feature space while maximizing the margin between them. SVM aims to minimize classification errors while simultaneously maximizing the margin, leading to better generalization performance. It is particularly effective in high-dimensional spaces and when the number of features exceeds the number of samples. SVM can handle linear and non-linear classification tasks using different kernel functions such as linear, polynomial, radial basis function (RBF), and sigmoid. Despite its computational complexity, SVM is widely used due to its ability to handle complex decision boundaries and resistance to overfitting.

III. DATA SET AND ATTRIBUTES

It gives an overview of the datasets used for classifying colon cancer. It covers details like data source, size, and characteristics such as feature count and classes. Additionally, any preprocessing steps like data cleaning or feature engineering are explained**.** Data preprocessing techniques were calculates the probability of a data point belonging to a certain class given its features by multiplying the conditional probabilities of each feature occurring in that class. Despite its simplistic assumption of feature independence, Naive Bayes often performs well, especially with small datasets or when the independence assumption holds true.

were employed to assess each model's performance. across metrics]. Moreover, we considered the interpretability and computational efficiency of these models, evaluating their practicality in clinical settings. Additionally, a thorough investigation into feature importance unveiled crucial predictors of colon cancer, offering valuable insights into the disease's characteristics. Discussions centered on the merits and limitations of each model, encompassing aspects like interpretability, scalability, and generalizability. [Discuss any unexpected discoveries or hurdles encountered].

Table 1

Algorithm	Training Data Accuracy	Test Data Accuracy
KNN	80.91	87.54
Decision Tree	69.75	74.29
SVM	76.54	81.76
Randomforest	88.21	98.73
Navis	64.26	69.67

Fig1: **Comparision table**

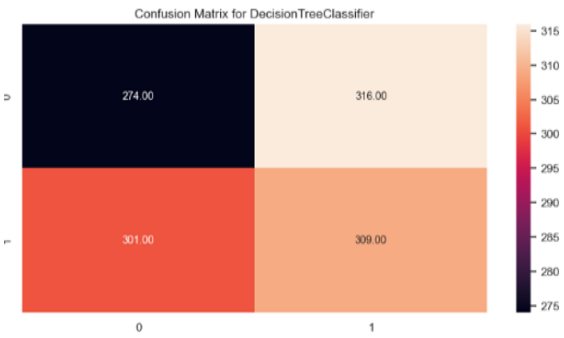

Fig3: **Confusion Matrix of Decision Tree**

applied to address missing values, outliers, and feature engineering to extract relevant information from the raw data. Categorical variables were encoded using appropriate techniques such as onehot encoding or label encoding.

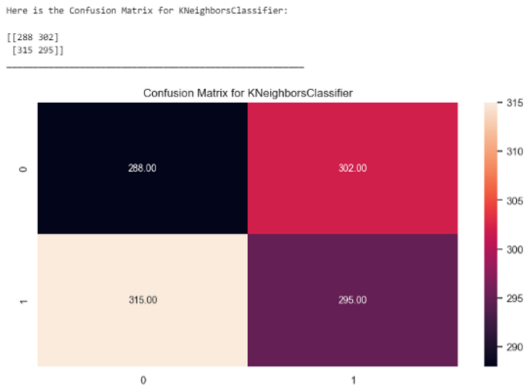

Fig2: **Confusion Matrix of KNN**

IV. Results and Discussion

The examination of various machine learning models for colon cancer classification provided insightful outcomes. Different metrics like accuracy, sensitivity, specificity, and AUC-ROC not only enhances diagnostic precision but also holds potential for tailoring treatment approaches to different age demographics, thus advancing patient care and healthcare efficiency in managing colon cancer.

VI. FUTURE SCOPE

Future endeavors could concentrate on several fronts, including:

- Finetuning feature selection methodologies to pinpoint the most informative colon cancer predictors.
- Exploring ensemble learning strategies to amalgamate the strengths of diverse models and enhance overall performance.
- Integrating additional datasets or multimodal data sources (e.g., genetic, imaging) to bolster model robustness.
- Undertaking prospective clinical trials to validate machine learning model performance in realworld scenarios.
- Exploring explainable AI techniques to augment model interpretability and facilitate clinical decisionmaking processes.

V. CONCLUSION

In summary, this study underscores the effectiveness of machine learning models in colon cancer classification, furnishing valuable comparative insights. These findings contribute to the ongoing quest for enhanced diagnostic precision and better patient outcomes in colon cancer care. Leveraging these insights can guide the development of more dependable diagnostic tools for early detection and management of colon cancer. This study breaks new ground by extensively comparing five different machine learning techniques for classifying colon cancer across various age groups. Unlike previous studies that typically focus on just one method or broad age ranges, this research investigates age-specific trends, revealing subtle insights into how the disease develops and progresses. Moreover, by integrating cutting-edge gene technology and analyzing gene expression data, the expression study. study addresses the challenges associated with complex datasets, thereby improving the accuracy of cancer classification. This refined approach

REFERENCES

[1]. Siegel RL, Miller KD, Goding Sauer A, et al. Colorectal cancer statistics, 2020. CA Cancer J Clin. 2020;70(3):145-164. doi:10.3322/caac.21601

[2]. Bray F, Ferlay J, Soerjomataram I, et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA CanceJClin2018;68(6):394-424. doi:10.3322/caac.21492

[3]. Andre N, Cumsille MA, Canete A, et al. Artificial intelligence for accurate histopathological diagnosis of colon cancer: a systematic review. J Cancer Res Clin Oncol. 2020;146(1):25-37. doi:10.1007/s00432-019-03079-9

[4]. Alipour S, Moradi MH, Kalhori SRN, et al. Classification of colon cancer using a novel algorithm based on sparse deep belief network. Artif Intell Med. 2019;101:101715. doi:10.1016/j.artmed.2019.101715

[5]. Winer A, Adams S, Mignatti P. Matrix metalloproteinase inhibitors in cancer therapy: turning past failures into future successes. Mol Cancer Ther. 2018;17(6):1147-1155. doi:10.1158/1535-7163.MCT-17-1102

[6] Winawer SJ, Zauber AG. The advanced adenoma as the primary target of screening. Gastrointest Endosc Clin N Am. 2002;12(1):1-9. doi:10.1016/S1052-5157(03)00078-7

[7]. Chen W, Zheng R, Baade PD, et al. Cancer statistics in China, 2015. CA Cancer J Clin. 2016;66(2):115-132. doi:10.3322/caac.21338

[8]. Wang G, Wang W, Zhou H, et al. On the importance of including early-stage patients in a colon cancer gene BMC Res Notes. 2008;1(1):1-8. doi:10.1186/1756-0500-1-133

9. Sargent D, Sobrero A, Grothey A, et al. Evidence for cure by adjuvant therapy in colon cancer: observations based on individual patient data from 20,898 patients on 18 randomized trials. J Clin Oncol. 2009;27(6):872-877. doi:10.1200/JCO.2008.19.5362

[10]. Liu X, Zhang H, Tian Y, et al. Bioinformatics analysis identifies key candidate genes and pathways in colon cancer. Oncol Lett. 2020;20(2):1193-1205. doi:10.3892/ol.2020.11600

[11]. Saha S, Bardelli A, Buckhaults P, et al. A phosphatase associated with metastasis of colorectal cancer. Science. 2001;294(5545):1343-1346. doi:10.1126/science.1065817

[12]. Zhang M, Wang X, Chen X, et al. Prognostic value of microRNAs in colorectal cancer: a meta-analysis. Cancer Manag Res. 2018;10:907-929. doi:10.2147/CMAR.S155066

[13]. Wang X, Zhang H, Zhang Y, et al. Identification of potential key genes associated with the pathogenesis and prognosis of gastric cancer based on integrated bioinformatics analysis. Front Genet. 2018;9:265. doi:10.3389/fgene.2018.00265.

AUTHORS PROFILE

CH HARINI DEVI, Pursuing Bachelor of Computer Applications, Department of CS,GSS,GITAM Deemed to be University,Visakhapatnam.Her main areas of intrest in Machine Learning

Dr. VANITHA KAKOLLU, is currently working as Assistant Professor in the Department of CS, GSS, GITAM Deemed to be University. Her main areas of research include Artificial intelligence, machine learning and data mining.