DOI: https://doi.org/10.54393/fbt.v5i1.131

FUTURISTIC BIOTECHNOLOGY
https://fbtjournal.com/index.php/fbt
ISSN (E): 2959-0981, (P): 2959-0973
Vol 05 Issue 01, (Jan-Mar, 2025)

Original Article

OPEN ACCESS

Investigating the Usage of Random Forest Method on Next-Generation Sequencing Data to Predict MSH2 and MSH6 Associated Mutations

Obaid Ullah[1], Muzamal Hussain[2], Nazia Kanwal[2*], Aamir Amin[2], Ahmar Saeed[1], Mudasir Zaheer[1] and Sana Fatima[2]

[1]Department of Computer Science, University of Agriculture, Faisalabad, Pakistan
[2]Department of Biological Sciences, The Superior University, Lahore, Pakistan

ARTICLE INFO

Keywords:
Colorectal cancer (CRC), Random Forest (RF), Machine Learning (ML), MSH2, MSH6, DNA, Diagnosis, NGS, Mutation

How to Cite:
Ullah, O., Hussain, M., Kanwal, N., Amin, A., Saeed, A., Zaheer, M., & Fatima, S. (2025). Investigating the Usage of Random Forest Method on Next-Generation Sequencing Data to Predict MSH2and MSH6 Associated Mutations: Random Forest Method to Predict MSH2and MSH6 Associated Mutations. Futuristic Biotechnology, 5(1), 26-31. https://doi.org/10.54393/fbt.v5i1.131

***Corresponding Author:**
Nazia Kanwal
Department of Biological Sciences, The Superior University, Lahore, Pakistan.
kanwal.n@superior.edu.pk

Received date: 31st August, 2024
Acceptance date: 17th March, 2025
Published date: 31st March, 2025

ABSTRACT

Colorectal cancer (CRC) is one of the most prevalent cancers and the second leading cause of cancer-related deaths globally. Germline mutations in CRC are associated with the MSH2 and MSH6 genes, which prevent infection for the DNA MMR pathway. **Objectives:** To enhance CRC-related prediction of mutations using the Random Forest algorithm on NGS data of MSH2 and MSH6 gene. Given the tremendous amount of genetic information obtained from NGS, a model for the early diagnosis and individual treatment of CRC is necessary. **Methods:** The raw sequencing data of MSH2 and MSH6 genes were meticulously downloaded from the NCBI's SRA database. The three datasets of 1000, 2000, and 3000 sequences were carefully analyzed to assess genomic features, including ORF count, nucleotide content, AT/CG ratio, G-quadruplex signal, and mutation rates, to understand their correlation with colorectal cancer. The data were then divided into a training set (80%) and a test set (20%) for model training and testing in Python, employing the Biopython package for mutation analysis and feature extraction. The model was rigorously evaluated using accuracy, confusion matrix, and classification report, instilling confidence in the research process for accurate CRC mutation prediction. **Results:** The Random Forest model yielded high accuracy of 96.25%, 98.37%, and 99. 5% for the datasets of 1000, 2000, and 3000 sequences, respectively. The confusion matrix showed that the model was very accurate in identifying true negatives, especially in the large data set. **Conclusions:** The study successfully applied the Random Forest algorithm to predict CRC using NGS data of MSH2 and MSH6 gene mutations. The model's potential to revolutionize CRC research is both exciting and optimistic.

INTRODUCTION

Colorectal cancer (also known as CRC) is one of the most widespread cancers globally and has incidence rates that are on the rise in both developed and developing nations. Colorectal cancer is the second most common cause of death, and a rapid increase has been observed from the year 2000 to 2019 [1]. Colorectal cancer rates in several studies from different regions of Pakistan varied between 4 and 6. 8% [2]. It is believed that genetic mutations, family history, and inherited syndromes all contribute to its hereditary nature. However, lifestyle factors such as food high in red and processed meats, laziness, obesity, excessive alcohol consumption, and smoking increase the risk of developing this condition [3]. Complications associated with colorectal cancer include bowel obstruction, haemorrhage, metastasis to other organs, such as the liver and lungs, and systemic manifestations, such as anaemia and fatigue [4]. MSH6 and MSH2 genes are necessary for the DNA mismatch repair (MMR) system. This system maintains genomic stability by identifying and repairing base mismatches during DNA replication [5]. Microsatellite instability, or MSI, is distinguished by the accumulation of insertion or deletion errors at microsatellite DNA sequences. These errors are brought on by mutations in the genes that control microsatellites.

DOI: https://doi.org/10.54393/fbt.v5i1.131

Due to the genome's instability, an environment conducive to neoplastic transformation is created [6]. Hereditary nonpolyposis colorectal cancer (HNPCC), also known as Lynch syndrome, is more likely to occur in people who have inherited mutations in either the MSH6 or MSH2 gene [7]. Next-generation sequencing (NGS) development was a key find in molecular biology, allowing knowledgeable scientists to understand broader aspects of the human genome quickly. This high-speed method can produce clear and in-depth images, decoding millions of DNA sequences simultaneously [8]. Machine learning and AI are taking on increased importance in the ever-changing healthcare industry, which is possible due to their flexibility in function [9]. Significantly, the prospects of ML with next-generation sequencing (NGS) for cancer prediction have advanced their way into human consciousness [10]. Random Forest (RF) is an ensemble learning approach that ensures dependable predictions from big genetic datasets by providing high accuracy and well-handling overfitting [11]. Its principal function is to build several trees during the training period. For classification problems, it uses the mode of the classes, and for regression problems, it uses mean perdition [12]. Applying the Random Forest model to sequence data from next-generation sequencing (NGS) on colorectal cancer may significantly raise prediction accuracy. The method advances the development of personalized treatment procedures by using NGS's massive genetic data to make diagnoses that are much more determined [13].

This research aimed to predict CRC related mutations prediction based on different DNA features in MSH2 and MSH6 genes.

METHODS

The current study focused on the Next Generation Sequences of Homo Sapiens based on the MSH2 (Accession Number: SRR25243226 and SRR25243227) and MSH6 genes (Accession Number: SRR1518357) from the NGS reads from NCBI's SRA repository in FASTA format [14]. These sequences were grouped into three sets with thousand, two thousand, and three thousand sequences, respectively. Reference sequences for MSH2 (Accession Number: NG_007110.2) and MSH6 (Accession Number: NG_007111.1) normal genes were retrieved from the NCBI database that was used to determine the mutation rates in NGS reads. Features including ORF count, average nucleotide (Cytosine, Guanine, Thymine, and Adenine), AT/CG ratio and its content, presence of G-quadruplex, and mutation rates were extracted from the collected sequences in this phase by comparing NGS reads with normal sequences.These eleven features were then stored in a .csv file. Matrices selection was applied to input the construction of the classification process. All the analyses were done on VS Code, but Google Colabs was used to compute the mutation rate. The total mutation rate was calculated per NGS read through Python library-assisted comparisons between NGS data features and reference sequences. The methodology of feature extraction was based on previously published work by Kurian and Jyothi. A random forest classifier was used to evaluate the dataset's classification performance. Initially, the necessary libraries, such as the Random Forest Classifier and train_test_split, were imported from Scikit-learn for selection and functions for accuracy. The confusion matrix and classification report from Scikit-learn were exported along with pandas for data operation. The dataset was loaded and split into features and target variables, with irrelevant columns removed. Python and biopython package were used for feature extraction and computing the mutation rate in this study. The data were then divided into training and testing sets with an 80:20 split using train_test_split. A Random Forest model with two estimators was built and fitted to the training set. The test set was used as input, and the model's predictions and accuracy were determined. A confusion matrix and classification report were also generated to analyze the model's results and give a clear view of the model's performance. The data of mutation from the NGS reads related to MSH2 and MSH6 genes were analyzed and correlated with colorectal cancer prediction, employing the Random Forest model to assess the statistical significance of these mutations (Figure 1).

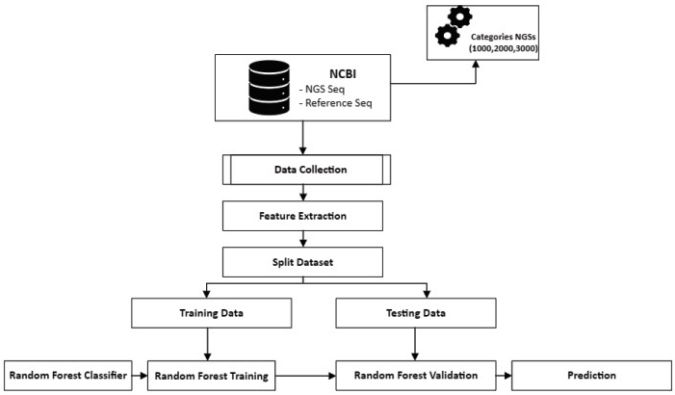

Figure 1: Architectural Diagram of Methodology for CRC Prognosis.

RESULTS

The study adopted the Random Forest (RF) ML algorithm for analysis. Random Forest (RF) is one of the most used techniques of the ensemble learning methodology, offering reliable outcomes from large genetic datasets as it offers high precision and manages overfitting. Thus, the

DOI: https://doi.org/10.54393/fbt.v5i1.131

presented software tool can help obtain accurate and fast results from next-generation sequencing data analysis. Total datasets were divided into training sets, and testing sets with a ratio of 80:20 for the progression of ML classification. Different tendencies surfaced in this study that compared ML models trained on colorectal cancer NGS reads based on datasets of varying sizes. The same datasets are used for all four models. Confusion matrix is characterized as 2 × 2 matrix which signifies the total number of classes to be classified. The sum of sets for testing for each class was derived by using row summation from the confusion matrix. Figure 2 indicates the confusion matrix results of the Random Forest algorithm. Random Forest (RF) successfully identified genuine negatives with high accuracy, especially in more enormous datasets, but struggled to classify real positives, especially in smaller datasets. The mutation rates in the MSH2 and MSH6 genes were found to be significantly higher in colorectal cancer (CRC)-positive samples compared to normal sequences.

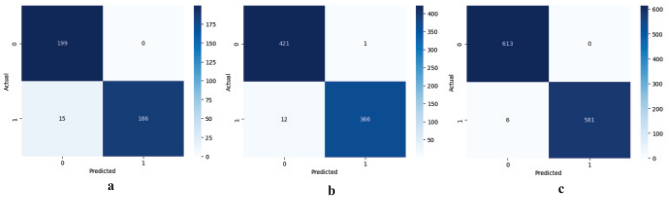

Figure 2: Graphical Explanation of Confusion Matrix generated by VS Code after classification. (a): On the first dataset of 1000 reads, the model correctly classified 199 positive instances and 186 negative instances. The model was misclassified 15 times, predicting negative while it predicted positive, and there were no misclassifications where the model predicted positive while it predicted negative, (b): On the second dataset of 2000 reads, the model appropriately classified 421 positive instances and 366 negative instances. The model was misclassified 12 times, predicting negative while it predicted positive, and there was only 1 misclassifications where the model predicted positive while it predicted negative, and (c): On the third dataset of 3000 reads, the model correctly classified 613 positive instances and 581 negative instances.The model was misclassified 6 times, predicting negative while it predicted positive, and there were no misclassifications where the model predicted positive while it predicted negative.

One way to get the classification accuracy rate is to multiply the total testing data size by 100 and divide the result by the number of adequately indicated classes. Table 1 presents the results of classification accuracy rates of RF. Machine learning models showed clear accuracy patterns when evaluated on varying datasets of NGS reads. With a modest improvement with increasing dataset size, despite dealing with datasets of varied sizes, Random Forest attained excellent accuracy, ranging from 96.25 per cent for the smallest dataset to 99.5 per cent for the biggest.

Table 1: Classification Accuracy Rate

ML Model	Dataset Size		
	1000	2000	3000
Random Forest (RF)	96.25	98.37	99.5

Random Forest performs well on smaller datasets (1000 samples) with average recall, accuracy, and F1-score values close to 0.96. RF performs admirably, and in the most extensive dataset, it achieves practically perfect results. Table 2 indicates the System Generated Classification report of Random Forest. Precision refers to the proportion of accurate positive predictions among all positive predictions. The proportion of true positive predictions among all actual positive is called recall. F1-score is used as the harmonic average of precision and recall, and thus, a balance between precision and recall. The support is the number of the actual occurrences of the class in the dataset. Table 2 shows the accuracy of the Random Forest model in classifying MSH2 and MSH6 gene mutations in CRC for different dataset sizes. The model's performance increases with the increase in the dataset size from 1000 to 3000 sequences, and the precision, recall, and F1-score for both MSH2 and MSH6 are closer to the maximum values. In the largest dataset of 3000 sequence reads, the precision and recall for MSH2 were 0.99 and 1.00, respectively, and MSH6 had a perfect score of 1.00 in all the measures.

Table 2: System Generated Classification Report of Random Forest

Variables	Dataset Size	1000				2000				3000			
Machine Learning Model	Class Label	Precision Value	Recall Value	F1- Score Value	Support Value	Precision Value	Recall Value	F1- Score Value	Support Value	Precision Value	Recall Value	F1- Score Value	Support Value
Random Forest (RF)	MSH2	0.93	1.00	0.96	199	0.97	1.00	0.98	422	0.99	1.00	1.00	613
	MSH6	1.00	0.93	0.96	201	1.00	0.97	0.98	378	1.00	0.99	0.99	587
	Avg/Total	0.97	0.96	0.96	400	0.98	0.98	0.98	800	1.00	0.99	0.99	1200

Precision is the ratio of true positive predictions over all predicted positive predictions. The recall is the fraction of all actual positives that are true positives. F1-score is the harmonic mean of precision and recall, thus balancing out the two. This is referred to as support and denotes the number of real instances of each class in the dataset, or how many instances of each class have occurred in the dataset.

DOI: https://doi.org/10.54393/fbt.v5i1.131

DISCUSSION

The current study shows that it is possible to use machine learning, specifically Random Forest algorithm, to predict CRC based on the genomic changes in the MSH2 and MSH6 genes. The current research approach based on the NGS data is a reliable method for finding mutations that may cause CRC and thus, makes a significant contribution to developing the concept of precision medicine. Combining NGS with machine learning can help in early diagnosis, essential for enhancing patient prognosis and developing individualized treatment plans. The results of the current study suggested that the Random Forest model achieved classification accuracy rates of 96.25%, 98.37%, and 99.5% for the datasets of 1000, 2000, and 3000 sequences, respectively. These results suggest that Random Forest is a very efficient method in terms of relevant mutation detection, and the accuracy of the method increases with the growth of data, which underlines the significance of large datasets in improving CRC prediction.These results prove that the proposed model is effective and can be applied to analyze big genomic data, which can be used for the early diagnosis and personalized treatment of colorectal cancer because accuracy more than 90% is considered best as mentioned in previous studies [16, 17]. Another study focusses on using Random Forest (RF) algorithms to analyze NGS data for prediction of cancer-associated genomic alterations [18]. Although current study aims to predict colorectal cancer using MSH2 and MSH6 genes, the former literature investigated multiple tumor types, such as colorectal, melanoma, lung, and glioma, based on a 27-gene panel. Both studies use RF because of its ability to handle large data and minimize overfitting. The classification accuracy rates of our RF model were 96. 25%, 98.37%, and 99.5% for the datasets of 1000, 2000, and 3000 sequences, respectively, which proves that the proposed method's performance increases with the dataset's size proposed method's performance increases with the dataset size. It was claimed that the accuracy of the RF model was 99.77% with an ROC-AUC of 0. 99 indicates good predictability across a wider range of genomic changes [18]. Although the two studies have different emphases and dataset compositions, both show that RF is a useful tool for genomic data analysis and is highly accurate in predicting cancer-related mutations. Another study on the Lung Cancer Research obtained an accuracy of 87% in biomarkers for NSCLC and SCLC using RNA-Seq data. The biomarkers for NSCLC include BRAF, KRAS, NRAS, and EGFR, while those for SCLC include ATF6, ATF3, PGDFA, PGDFD, PGDFC, and PIP5K1C. On the other hand, the present study on CRC using MSH2 and MSH6 gene mutations and NGS data showed accuracy rates of 96%. 25%, 98. 37%, and 99. 5% for the datasets of 1000, 2000, and 3000 sequences, respectively. Accuracy in the Random Forest method refers to the percentage of correct predictions (both true positives and true negatives) made by the model relative to the total predictions made. An accuracy of 96.25% means that 96.25% of the total predictions for the dataset were correct [19]. This implies that the Random Forest model in our study was more efficient, probably because of the genomic interest and the feature engineering from NGS data as opposed to the general and intricate RNA-Seq data used in the lung cancer study. The Random Forest algorithm is a very efficient and reliable tool in genomic data analysis for disease prognosis and personalized medicine. It can handle big and intricate data, does not overtrain, is suitable for genetic data, and has high prediction power [20]. The high accuracy of the Random Forest model in classifying mutations in MSH2 and MSH6 genes in colorectal cancer can be helpful in early diagnosis and treatment planning. Random Forest consists of multiple decision trees that enhance the overall prediction, making it a very efficient method for predicting the relationships existing in genomic data. Consequently, it holds an essential position in forming precision medicine and individualized therapy [21]. The current study's research suggests some directions for enhancing future research in the academic field. Including other omics data, like transcriptomics and proteomics, could complement the current knowledge of CRC and enhance the prediction models. Further, the methodology could be used to predict other types of cancer by analyzing other genetic mutations and thus increase the applicability of this study. Clinical confirmation through performing clinical trials is crucial to confirm the model's predictions in clinical scenarios, which is significant for the application of research outcomes in CRC diagnosis. In addition, creating systems that can sequence and predict in real-time may bring about a meaningful change in the diagnostic process since accurate interventions can be made at the right time, enhancing the patient's quality of life. One limitation of this study is the diversity of the datasets used. While the study utilized datasets of different sizes, the diversity of the sequences might still be limited. Additionally, the study primarily focuses on mutations in the MSH2 and MSH6 genes, leaving out other genetic and epigenetic factors contributing to CRC, which might limit the model's comprehensiveness. Furthermore, the computational demands of processing and analyzing large NGS datasets can be significant, posing a challenge in optimizing the model for resource efficiency without compromising accuracy.

DOI: https://doi.org/10.54393/fbt.v5i1.131

CONCLUSION

The study effectively proves that using Random Forest algorithms for the NGS data of MSH2 and MSH6 genes can accurately predict mutations due to an accuracy of more than 90% in three different types of datasets. The Random Forest model obtained classification accuracy rates of 96. 25%, 98.37%, and 99.5% for the datasets of 1000, 2000, and 3000 sequences, respectively. These results demonstrate the efficiency and applicability of the model in dealing with big genomic data and thus can be used for early diagnosis and individualized treatment of colorectal cancer.

Authors Contribution

Conceptualization: OU, MH
Methodology: OU, MH, AA
Formal analysis: NK
Writing review and editing: OU, MH, NK, SF, AS, MZ
All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

All the authors declare no conflict of interest.

Source of Funding

The authors received no financial support for the research, authorship and/or publication of this article.

REFERENCES

[1] Pan H, Zhao Z, Deng Y, Zheng Z, Huang Y, Huang S, *et al.*The Global, Regional, And National Early-Onset Colorectal Cancer Burden and Trends From 1990 To 2019: Results from The Global Burden of Disease Study 2019. Bmc Public Health.2022;22(1):1896.doi:10.1186/s12889-022-14274-7.

[2] Idrees R, Fatima S, Abdul-Ghafar J, Raheem A, Ahmad Z.Cancer Prevalence in Pakistan: Meta-Analysis of Various Published Studies to Determine Variation in Cancer Figures Resulting from Marked Population Heterogeneity In Different Parts Of The Country. World Journal of Surgical Oncology.2018; 16: 1-11. doi:10.1186/s12957-018-1429-z.

[3] Kastrinos F, Samadder NJ, Burt RW. Use Of Family History and Genetic Testing to Determine Risk of Colorectal Cancer.Gastroenterology.2020;158(2):389-403. doi:10.1053/j.gastro.2019.11.029.

[4] Biller LH and Schrag D. Diagnosis and Treatment of Metastatic Colorectal Cancer: A Review. Jama.2021; 325(7): 669-85. doi:10.1001/jama.2021.0106.

[5] Pećina-Šlaus N, Kafka A, Salamon I, Bukovac A. Mismatch Repair Pathway, Genome Stability and Cancer. Frontiers In Molecular Biosciences.2020; 7: 122. doi:10.3389/fmolb.2020.00122.

[6] Li K, Luo H, Huang L, Luo H, Zhu X. Microsatellite Instability: A Review of What the Oncologist Should Know. Cancer Cell International.2020;20:1-13.doi :10.1186/s12935-019-1091-8.

[7] Nolano A, Medugno A, Trombetti S, Liccardo R, De Rosa M, Izzo P, *et al.* Hereditary Colorectal Cancer: State of The Art in Lynch Syndrome. Cancers.2022; 15(1): 75. doi:10.3390/cancers15010075.

[8] Ahuja SK, Shrimankar DD, Durge AR. A Study and Analysis of Disease Identification Using Genomic Sequence Processing Models: An Empirical Review. Current Genomics.2023;24(4):207.doi:10.2174/0113892029695232311010514 55.

[9] Qayyum MU, Sherani AMK, Khan M, Hussain HK. Revolutionizing Healthcare:The Transformative Impact of Artificial Intelligence in Medicine.Bin: Bulletin Of Informatics. 2023; 1(2): 71-83.

[10] Iqbal MJ, Javed Z, Sadia H, Qureshi IA, Irshad A, Ahmed R, *et al.* Clinical Applications of Artificial Intelligence and Machine Learning in Cancer Diagnosis: Looking into The Future.Cancer Cell International.2021; 21(1): 270. doi:10.1186/s12935-021-01981-1.

[11] Pachouly J, Ahirrao S, Kotecha K, Selvachandran G, Abraham A. A Systematic Literature Review on Software Defect Prediction Using Artificial Intelligence: Datasets, Data Validation Methods, Approaches, And Tools. Engineering Applications for Artificial Intelligence.2022; 111: 104773. doi:10.10 16/j.engappai.2022.104773.

[12] Singh R and Pal S. Machine Learning Algorithms and Ensemble Technique to Improve Prediction of Students Performance. International Journal of Advanced Trends in Computer Science and Engineering.2020;9(3).doi:10.30534/ijatcse/2020/221932020.

[13] Rubio-Mangas D, García-Arranz M, Torres-Rodriguez Y, León-Arellano M, Suela J, García-Olmo D. Differential Presence of Exons (DPE): Sequencing Liquid Biopsy by Nges. A New Method for Clustering Colorectal Cancer Patients. BMC Cancer.2023;23(1): 1-14. doi:10.1186/s12885-022-10459-w.

[14] Eldem V and Balcı MA. Mining NCBI Sequence Read Archive Database: An Untapped Source of Organelle Genomes for Taxonomic and Comparative Genomics Research.Diversity.2024;16(2):104.doi:10.3390/d16020104.

[15] Kurian B and Jyothi V. Breast Cancer Prediction Using an Optimal Machine Learning Technique for Next Generation Sequences. Concurrent Engineering.2021;29(1):49-57. doi:10.1177/1063293X21991808.

[16] Risal S, Zhu W, Guillen P, Sun L. Improving Phase Prediction Accuracy for High Entropy Alloys with

DOI: https://doi.org/10.54393/fbt.v5i1.131

Machine Learning. Computational Materials Science. 2021;192:110389.doi:10.1016/j.commatsci.2021.110389.

[17] Ibrahim I and Abdulazeez A.The Role of Machine Learning Algorithms for Diagnosing Diseases. Journal Of Applied Science and Technology Trends. 2021; 2(01): 10-9. doi:10.38094/jastt20179.

[18] Pellegrino E, Jacques C, Beaufils N, Nanni I, Carlioz A, Metellus P, *et al.* Machine Learning Random Forest for Predicting Oncosomatic Variant Ngs Analysis. Scientific Reports.2021;11(1):21820.doi:10.1038/s41598-021-01253-y.

[19] Lavanya C, Pooja S, Kashyap AH, Rahaman A, Niranjan S, Niranjan V. Novel Biomarker Prediction for Lung Cancer Using Random Forest Classifiers.Cancer Informatics.2023;22.doi:10.1177/11769351231167992.

[20] Qi Y. Random Forest for Bioinformatics.Ensemble Machine Learning: Methods And Applications.2012 307-23. doi:10.1007/978-1-4419-9326-7_11.

[21] Sun Z, Wang G, Li P, Wang H, Zhang M, Liang X. An Improved Random Forest Based on The Classification Accuracy and Correlation Measurement of Decision Trees. Expert Systems with Applications.2024;237:121549.doi:10.1016/j.eswa.2023.121549.