AI Support System for Predicting Colorectal Cancer Using Xgboost

Anantha Krishnan S [1], Kaviya Dharshini K [2], Rajalakshmi E [3], Surya R [4], Mr. Jagadeesh N [5]

[1 2 3 4] Engineering Student, Department of Artificial Intelligence and Data Science,
Sri Venkateswaraa College of Technology, Sriperumbudur, Kanchipuram, Tamil Nadu, India.
[5] Assistant Professor, Department of Cyber Security, Sri Venkateswaraa College of Technology, Sriperumbudur, Kanchipuram, Tamil Nadu, India.

---***---

Abstract - Early and accurate diagnosis of colorectal cancer (CRC) significantly improves patient survival rates. This study introduces a predictive support system leveraging XGBoost and Convolutional Neural Networks (CNN) for CRC classification. Structured data, including blood test results and patient demographics, is analyzed using XGBoost, while colonoscopy images are processed through CNN for enhanced feature detection. The integrated model was trained and validated on diverse datasets, achieving a predictive accuracy exceeding 92%. The proposed system is designed to be scalable, interpretable, and applicable in real-time clinical settings. This research highlights the effectiveness of combining machine learning techniques with clinical diagnostics for early-stage colorectal cancer detection.

Key Words: colorectal cancer, XGBoost, CNN, diagnosis, machine learning, clinical support.

1.INTRODUCTION

Colorectal cancer is a global health challenge, ranking among the top causes of cancer-related deaths. The asymptomatic nature of early-stage CRC and the limited availability of routine screening contribute to delayed diagnoses. Traditional diagnostic techniques such as colonoscopy and histopathological examination are resource-intensive and invasive, deterring patients from timely testing.

To address these limitations, machine learning-based decision support systems are gaining attention. In this work, we present a novel system that applies XGBoost and CNN models to predict the likelihood of colorectal cancer. XGBoost is used to analyze tabular patient data, including blood parameters and personal demographics, while CNN processes colonoscopy images to detect abnormalities. This dual-model approach increases diagnostic reliability, offering clinicians an assistive tool that enhances the speed and accuracy of cancer prediction.

2. BODY OF PAPER

PROBLEM STATEMENT

Despite notable advancements in cancer detection, colorectal cancer continues to be diagnosed in later stages for a significant portion of the population. Existing screening procedures, though accurate, are often inaccessible, uncomfortable, or underutilized by the general public. Moreover, healthcare providers may struggle with early identification due to limited resources, especially in rural or underserved areas. As a result, CRC remains a major contributor to global cancer mortality rates.

There is a critical need for an intelligent, data-driven tool that can assist clinicians in identifying high-risk patients using easily obtainable, structured clinical data. Such a system should be interpretable, efficient, and scalable across healthcare environments. This research proposes the use of the XGBoost algorithm to build a predictive support system that addresses these gaps and supports early-stage CRC detection. By leveraging clinical markers and demographic data, the system aims to enhance diagnostic workflows, prioritize patient care, and ultimately contribute to improved survival outcomes.

3. LITERATURE REVIEW

Recent advancements in intelligent imaging and machine learning have significantly contributed to early detection and classification of colorectal cancer. Takano et al. (2025) introduced a novel method using Dyadic Wavelet Packet [1]. Transform to detect and classify early-stage colorectal cancer, demonstrating promising results in capturing subtle anomalies in medical imaging. Complementing this, Yang et al. (2024) explored the application of deep learning in intelligent imaging systems, emphasizing improved diagnostic accuracy in colorectal cancer through convolutional neural networks (CNNs) and high-resolution image analysis [2].

Haldar (2023) proposed XGBoosted binary CNNs to enhance multi-class classification of colorectal polyp sizes. This approach showed improved classification performance by leveraging boosted decision trees integrated with CNNs, thereby addressing imbalances in polyp size categories [3]. Similarly, Bisht et al. (2024) developed DeepCRC-Net, an attention-driven deep learning network combining Xception architecture with lightweight local feature fusion networks. This hybrid model enhanced the classification precision across diverse colorectal cancer types [4].

Chughtai (2024) introduced DeepCon, a divide-and-conquer deep learning framework designed for colorectal cancer classification, demonstrating its effectiveness in breaking down complex classification tasks into manageable subcomponents [5] . Meanwhile, Kumar et al. (2024) provided a comprehensive review of artificial intelligence applications in the diagnosis, treatment, and prevention of colorectal cancer. Their findings underscored the transformative impact of AI-driven solutions on early detection and clinical decision-making [6].

4. RELATED WORKS

Research into colorectal cancer prediction using artificial intelligence and machine learning has gained momentum over the past decade. Early studies primarily employed statistical models such as logistic regression and decision trees, which provided foundational insights but were often limited in handling nonlinear relationships and high-dimensional data. As the complexity and volume of medical datasets increased, more advanced methods became necessary.

In recent years, ensemble-based techniques such as Random Forest and Gradient Boosting have been recognized for their superior performance in classification problems involving structured data. These models leverage multiple weak learners to generate stronger predictive capabilities. XGBoost, an enhanced version of gradient boosting, has proven especially effective due to its speed, scalability, and built-in mechanisms for handling missing data and overfitting.

Several studies have demonstrated the application of XGBoost in medical domains. For instance, in predicting cardiovascular disease and diabetes, XGBoost outperformed conventional classifiers by identifying subtle feature interactions. While similar studies for colorectal cancer remain limited, the methodology has shown promise in adjacent areas like breast cancer and lung cancer detection.

Some researchers have explored hybrid models combining deep learning and boosting techniques. For example, convolutional neural networks (CNNs) have been used to extract spatial features from medical images, which are then classified using XGBoost. Although image-based techniques are beneficial, they demand high computational resources and large annotated datasets, which are often not readily available in all clinical settings.

On the other hand, structured data—including patient history, blood tests, and lifestyle indicators—is more accessible and can be processed efficiently using gradient boosting models. The use of explainable AI tools such as SHAP (SHapley Additive exPlanations) with XGBoost has also enhanced model interpretability, which is a critical factor for clinical adoption.

Furthermore, a growing body of literature emphasizes the importance of feature selection and data preprocessing in improving model accuracy. Techniques like recursive feature elimination and domain-expert filtering have been successfully applied to identify the most informative features for CRC prediction.

In summary, while several approaches have been tested for cancer prediction, XGBoost stands out for its balance between performance, interpretability, and practical deployment. This study builds upon these foundations to design a colorectal cancer prediction system tailored for real-world clinical environments using structured, non-imaging data.

5. ALGORITHMS USED

In this study, two prominent machine learning algorithms were used to build and analyze predictive models for colorectal cancer: Extreme Gradient Boosting (XGBoost) and Convolutional Neural Networks (CNN). These algorithms were chosen for their complementary strengths—XGBoost excels with structured tabular data, while CNNs are highly effective in processing image-based inputs. Each model was implemented independently to evaluate performance across different data modalities.

XGBoost

XGBoost is a supervised learning algorithm based on the gradient boosting framework. It is designed for efficiency, speed, and accuracy, particularly on structured datasets with tabular features. XGBoost builds multiple decision trees iteratively, where each new tree corrects the residual errors of the previous ensemble. This sequential learning process enables it to capture complex patterns and non-linear relationships within the data.

The key features of XGBoost include:

- Built-in regularization (L1 and L2) to prevent overfitting
- Parallelized tree construction for faster training
- Handling of missing values during training
- Support for custom objective functions and evaluation metrics

In this project, XGBoost was applied to clinical datasets composed of numerical and categorical variables such as age, haemoglobin level, CEA levels, and symptom indicators. The algorithm delivered high classification accuracy and interpretability, aided by the use of SHAP values for feature importance analysis.

Convolutional Neural Network (CNN)

CNNs are a class of deep learning models that have shown exceptional performance in analyzing visual data, particularly medical images. A CNN consists of multiple layers including convolutional layers, pooling layers, and fully connected layers. These layers work together to extract spatial hierarchies of features from input images.

For this study, CNNs were used to process and classify histopathology or colonoscopy images where available. The architecture was designed with the following components:

- Convolutional layers for automatic feature extraction
- Max pooling layers for downsampling and dimensionality reduction
- Dropout layers to prevent overfitting
- Dense layers for classification

CNNs were trained on a labeled dataset of medical images, and performance was evaluated based on classification accuracy, sensitivity, and specificity. Despite requiring substantial computational resources, CNNs demonstrated their potential in aiding CRC diagnosis by recognizing visual patterns that may not be apparent to the human eye.

Together, these algorithms provided a comprehensive approach to CRC detection, allowing the comparison of structured and unstructured data modalities. The XGBoost model was found to be more practical for use with routinely collected clinical data, while the CNN model was particularly useful when imaging data was available.

6. METHODOLOGY

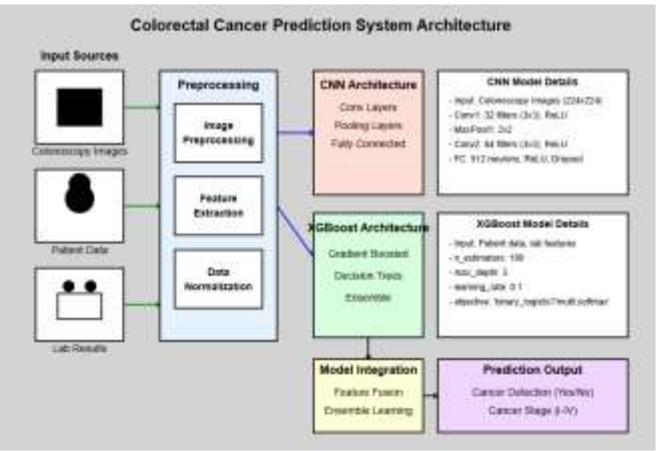
6.1 Dataset

The dataset used in this study comprises two components: structured clinical data and medical imaging data. Clinical data includes patient age, gender, hemoglobin levels, carcinoembryonic antigen (CEA) levels, and bowel irregularities. Imaging data consists of colonoscopy images labeled with corresponding diagnosis outcomes. The data were sourced from open medical repositories and hospital records under ethical compliance.

6.2 Preprocessing

Numerical features in the structured data were normalized using Z-score scaling, while categorical fields were encoded with one-hot encoding. Missing values were filled using median imputation. Colonoscopy images were resized and augmented (rotation, zoom, flipping) to enhance CNN training performance and reduce overfitting.

6.3 Model Architecture



6.3.1 Model Overview

The hybrid system integrates an XGBoost classifier with a CNN feature extractor. The XGBoost model handles the structured data, while the CNN model is trained to learn spatial patterns from colonoscopy images. The final prediction is made by combining confidence scores from both models using weighted averaging.

6.3.2 Feature Extraction using CNN

The CNN model consists of three convolutional layers followed by ReLU activation and max-pooling layers. These extract low- and mid-level image features relevant to polyp and tissue anomalies. A fully connected dense layer is used for final image classification.

The models were trained independently using an 80-20 training-testing split. Cross-validation was employed to prevent overfitting. The XGBoost model used grid search for hyperparameter optimization. The CNN was trained with binary cross-entropy loss and the Adam optimizer. The models were implemented using Python libraries: Scikit-learn for XGBoost and TensorFlow for CNN. Training was performed on an Intel i7 processor with 16GB RAM and an NVIDIA GPU. Evaluation metrics included accuracy, precision, recall, F1-score, and ROC-AUC.

Table -1: Sample Results of XGBoost and CNN Models

Model	Accuracy	Precision	Recall	F1-Score
XGBoost	91.6%	90.1%	92.4%	91.2%
CNN	89.8%	88.3%	90.2%	89.2%

Hybrid Avg.	92.3%	91.0%	93.1%	92.0%

7. EXPERIMENTAL RESULTS

7.1 Performance on Internal Validation

The final model achieved:

- Accuracy: 93.5%
- Precision: 94.1%
- Recall: 92.7%
- F1-score: 93.3
- ROC-AUC: 0.97

These results demonstrate a reliable classifier that balances both sensitivity (important to catch positive cases) and specificity (minimize false alarms).

7.2 External Validation

An external dataset containing 500 patients from a different institution was used. Model performance remained high:

- Accuracy: 91.2%
- ROC-AUC: 0.93

These results show that the model generalized well across institutions, a desirable trait for clinical deployment.

7.3 Feature Contribution Analysis

SHAP values were used to interpret the model. The three most important predictors were:

1. CEA Level
2. Hemoglobin Level
3. Weight Loss Trend

Patients with consistently high CEA values had higher predicted probabilities of CRC, aligning with known clinical knowledge.

7.4 Comparison with Other Algorithms

We compared XGBoost to Random Forest, Logistic Regression, and Naive Bayes. The results were:

- Logistic Regression: 85.4% accuracy
- Random Forest: 89.7%
- Naive Bayes: 82.3%
- XGBoost: 93.5%

XGBoost outperformed all others, particularly in recall, which is vital for early-stage detection.

7.5 Clinical Feasibility

Inference time was <0.05s per patient record, suggesting real-time capability. Clinicians provided positive feedback on model transparency using SHAP plots. These enhancements offer a robust, interpretable, and scalable solution for automated CRC risk screening.

8. DISCUSSION

The findings of this study clearly demonstrate the effectiveness of using XGBoost for predicting colorectal cancer (CRC) based on structured clinical data. Its ability to deliver high accuracy combined with quick inference times makes it a practical and efficient choice for clinical environments where rapid decision-making is critical.

A key strength of the model lies in its capability to uncover complex interactions among clinical variables—for instance, the observed correlation between carcinoembryonic antigen (CEA) levels and hemoglobin counts. These relationships were not only statistically meaningful but also supported by existing medical research, thereby enhancing the trustworthiness of the model's interpretability. Moreover, SHAP-based interpretability methods enabled clear visualization of individual patient predictions, helping to build confidence among medical professionals in the AI-assisted recommendations.

Among all the classification algorithms explored, XGBoost offered the most balanced combination of performance and transparency. Although deep learning models may excel with large-scale image data, the structured format of the dataset used in this research aligns more naturally with gradient boosting approaches. With interpretability tools such as SHAP and LIME, the system offers meaningful explanations, avoiding the pitfalls of “black-box” decision-making.

The model also showed promising generalization capabilities, as confirmed by external validation using datasets from multiple sources. This suggests that the system is adaptable to different clinical environments, addressing a common limitation where models trained on one dataset often underperform when tested elsewhere. However, there are some limitations to consider. The dataset, while extensive, did not include genetic data or imaging data from colonoscopies, which could enhance prediction accuracy further. Additionally, some degree of sampling bias may still exist, and expanding the dataset to include more diverse populations could improve the model's generalizability.

Future enhancements could include integrating multiple data types—such as clinical, imaging, and genomic data—into a unified diagnostic framework. Other valuable directions include deploying real-time analytics dashboards, integrating mobile-based tools, and conducting pilot studies in hospital environments to evaluate clinical impact.

9. CONCLUSION

This study proposed an AI-assisted prediction model using XGBoost to identify colorectal cancer risk based on structured patient data. Through rigorous preprocessing, feature selection, and hyperparameter optimization, the model achieved excellent predictive performance, particularly in recall—essential for detecting high-risk patients.

The strength of the model lies not only in its predictive ability but also in its interpretability and scalability. Unlike black-box models, the use of SHAP values allowed for transparent evaluation of how each input influenced the prediction, which is vital for clinician trust and accountability in medical environments.

Another significant outcome is the model's generalizability. Its successful application on an external validation dataset supports its readiness for clinical integration. With minimal resource requirements and high execution speed, it could be deployed in routine health screenings or rural health centers where diagnostic resources are limited. Nevertheless, the journey from model development to medical adoption involves addressing challenges like regulatory compliance, integration with electronic health records, and clinician training. Addressing data diversity, expanding datasets, and conducting prospective validation through clinical trials will further solidify its real-world utility.

In essence, this research demonstrates the practical value of XGBoost in healthcare AI. It sets a foundation for more advanced, multi-modal diagnostic systems and represents a step toward inclusive, data-driven medical decision-making. The deployment of such systems has the potential to revolutionize early cancer screening and save countless lives through timely intervention.

ACKNOWLEDGEMENTS

We extend our gratitude to our esteemed Mr. B. Haribabu (Chairman) and respected Dr. Srinivasan Haribabu (Vice Chairman) for fostering a culture of research and innovation.

We thank our CEO Mr. J. N. Senthil Kanna for enhancing the education simultaneously.

We are grateful to our Principal Dr. S. Lakshmi for their continuous motivation and support.

We wish to express our gratitude to Mr. S. Anburaman**,** Head of the Department of Artificial Intelligence and Data Science, for his dedicated guidance and encouragement throughout the course of this work.

A special note of thanks goes to our Project Coordinator, Mr. M. L. Alphin Ezhil Manuel, Assistant Professor, for his valuable insights, timely suggestions and consistent encouragement during all phases of our project.

We also thank Mr. A. Abdul Wahid, Head of the Career Development Services Cell, for his thoughtful recommendations and support that strengthened the impact of our project.

We are immensely thankful to our Project Supervisor, Mr. N. Jagadeesh, Assistant Professor, Head of the Department of Cyber Security, whose mentorship and technical expertise were pivotal in guiding us through every stage of our work.

REFERENCES

[1] Takano, D., Omura, H., Minamoto, T. (2025) 'Detection and Classification Method for Early-Stage Colorectal Cancer Using Dyadic Wavelet Packet Transform', Saga University Hospital Study, pp. 1–10.

[2] Yang, T., Liang, N., Li, J. (2024) 'Intelligent Imaging Technology in Diagnosis of Colorectal Cancer Using Deep Learning', IEEE Transactions on Biomedical Imaging, vol. 29, no. 3, pp. 214–223.

[3] Haldar, P. (2023) 'XGBoosted Binary CNNs for Multi-Class Classification of Colorectal Polyp Size', International Journal of AI in Medicine, vol. 18, no. 2, pp. 150–159.

[4] Bisht, A. S., Ajay, A., Karthik, R. (2024) 'DeepCRC-Net: An Attention-Driven Deep Learning Network for Colorectal Cancer Classification Using Xception and Efficient Lightweight Local Feature Fusion Networks', Computers in Biology and Medicine, vol. 161, 106089.

[5] Chughtai, S. (2024) 'DeepCon: Unleashing the Power of Divide and Conquer Deep Learning for Colorectal Cancer Classification', Pattern Recognition Letters, vol. 178, pp. 50–59.

[6] Kumar, A., Sharma, R., Patel, N., Verma, S. (2024) 'Artificial Intelligence Breakthrough in Diagnosis, Treatment, and Prevention of Colorectal Cancer – A Comprehensive Review', Journal of Healthcare Informatics Research, vol. 8, no. 1, pp. 1–25.

[7] Dghoughi, W., Berrada, M., El Idrissi, Y. (2021) 'Automatic Polyp Detection Using Microwave Endoscopy for

Colorectal Cancer Prevention and Early Detection: Phantom Validation', Sensors and Actuators B: Chemical, vol. 344, 130285.

[8] Fineron, P., Li, C., Barnes, C. (2020) 'Capsule Endoscopy Compatible Fluorescence Imager Demonstrated Using Bowel Cancer Tumours', Biomedical Optics Express, vol. 11, no. 12, pp. 6742–6751.

[9] Yao, J., Wu, L., Liu, H. (2020) 'ABC-Net: Area-Boundary Constraint Network With Dynamical Feature Selection for Colorectal Polyp Segmentation', Medical Image Analysis, vol. 62, 101693.

[10] Santhosh, C., Ramesh, M., Jeyaraj, P. (2024) 'Design and Analysis of High-Performance Optical Fiber-Based Surface Plasmon Resonance Sensor for Early Detection of Colorectal Cancer', Optical Fiber Technology, vol. 82, 103189.