INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 4, April 2026 | ISSN: 2456-4184 | IJNRD.ORG

A Hybrid Machine Learning Approach for Cancer Risk Prediction Using MIA

1st N Manikandan
Head of the Department
Computer Science and Engineering AVS College of Technology
Salem, India maniksjun05@gmail.com

2nd B Sathiya
Computer Science and Engineering AVS College of Technology
Salem, India sathiya19082001me@gmail.com

***Abstract*— Medical image classification plays a crucial role in early disease diagnosis and clinical decision-making, particularly in cancer detection. With the rapid growth of medical imaging data and limitations of traditional classification methods, there is a need for more accurate and efficient computational techniques. This paper proposes a hybrid machine learning framework for cancer risk identification using medical images. The methodology integrates Gray Level Co-occurrence Matrix (GLCM)-based feature extraction with advanced optimization techniques such as Bacterial Foraging Optimization (BFO) and Tabu Search for optimal feature selection. A Clonal Selection Algorithm (CLONALG)-based classifier is employed to categorize patients into high-risk and low-risk groups.**

The proposed system enhances classification performance by combining texture-based feature analysis with bio-inspired optimization algorithms, improving accuracy, sensitivity, and specificity. The framework also incorporates preprocessing, segmentation, and data mining techniques to handle large and complex medical datasets effectively. Experimental results demonstrate that the hybrid approach outperforms conventional machine learning models in terms of predictive accuracy and computational efficiency.

This research contributes to the development of intelligent healthcare systems by enabling early cancer detection, supporting personalized treatment planning, and reducing diagnostic time. The proposed model provides a scalable and robust solution for medical image analysis and has potential applications in clinical decision support systems.

Keywords— Medical Image Classification, Cancer Risk Prediction, Gray Level Co-occurrence Matrix (GLCM), Bacterial Foraging Optimization (BFO), Tabu Search, Clonal Selection Algorithm (CLONALG), Feature Selection, Machine Learning, Data Mining, Computer-Aided Diagnosis (CAD)

I. INTRODUCTION

Medical imaging plays a vital role in modern healthcare by enabling accurate diagnosis, treatment planning, and disease monitoring. Imaging modalities such as Magnetic Resonance Imaging (MRI), Computed Tomography (CT), and X-ray generate large volumes of data that require efficient analysis. However, the rapid growth of medical image databases and the complexity of image features make manual interpretation difficult, time-consuming, and prone to errors.

To address these challenges, automated medical image classification using machine learning and data mining techniques has gained significant attention. These approaches aim to extract meaningful patterns, identify abnormalities, and assist healthcare professionals in making accurate clinical decisions. Traditional classification methods, however, often suffer from limitations such as low accuracy, high computational cost, and ineffective feature selection when dealing with high-dimensional data.

Feature extraction and selection are critical steps in medical image analysis. Techniques such as Gray Level Co-occurrence Matrix (GLCM) are widely used to capture texture-based features from medical images. However, selecting the most relevant features remains a challenging task. Optimization algorithms inspired by natural processes, such as Bacterial Foraging Optimization (BFO) and Tabu Search, provide effective solutions for identifying optimal feature subsets and improving classification performance.

In this paper, a hybrid framework is proposed that integrates GLCM-based feature extraction with BFO and Tabu Search for feature optimization. A Clonal Selection Algorithm (CLONALG)-based classifier is employed to categorize patients into high-risk and low-risk groups for cancer prediction. The proposed system enhances accuracy, reduces computational complexity, and improves the reliability of diagnosis.

The main contributions of this work include:

- Development of a hybrid feature selection approach combining BFO and Tabu Search
- Implementation of a CLONALG-based classification model for cancer risk prediction
- Integration of data mining and image processing techniques for improved medical image analysis
- Performance evaluation using standard metrics such as accuracy, sensitivity, and specificity

Overall, the proposed approach provides an efficient and intelligent solution for medical image classification, supporting early disease detection and improved healthcare outcomes.

II. RELATED WORK

2.1 Data Mining in Medical Image Classification

Data mining is the process of extracting meaningful patterns, hidden relationships, and useful knowledge from large volumes of data using techniques from statistics, machine learning, artificial intelligence, and pattern recognition. In the field of medical image classification, data mining plays a crucial role in analyzing complex and high- dimensional datasets generated from various imaging

modalities such as X-ray, MRI, CT scans, PET, and ultrasound.

Medical datasets are often large, heterogeneous, and distributed across multiple sources, making manual analysis difficult and inefficient. Data mining helps overcome these challenges by enabling automated analysis, improving diagnostic accuracy, and supporting clinical decision-making. It is an interdisciplinary field that integrates database systems, machine learning, and image processing to extract valuable insights from medical data.

The process of data mining in medical image classification involves several important steps. Initially, data preprocessing is performed to clean and prepare the raw data by removing noise and inconsistencies. This is followed by feature extraction, where relevant characteristics such as texture, shape, and intensity are derived from images. Segmentation techniques are then used to divide images into meaningful regions for better analysis. After this, classification and clustering methods are applied to categorize images or detect patterns associated with diseases.

Various data mining techniques are used in medical image analysis. Classification methods such as decision trees, Naive Bayes, and Support Vector Machines (SVM) are widely used to assign images into predefined categories. Clustering techniques group similar images together without prior labeling, helping in discovering hidden patterns. Association rule mining identifies relationships between different features in the dataset, which is useful in disease diagnosis and prediction.

Image mining is an advanced extension of data mining that focuses specifically on extracting knowledge from image data. It combines computer vision, image processing, and artificial intelligence to analyze image content and identify patterns that are not easily visible. This process includes image retrieval, transformation, interpretation, and knowledge discovery. Image mining enables the extraction of spatial and temporal features, which are essential for accurate medical analysis.

Despite its advantages, medical image data mining faces several challenges. The datasets are often large, complex, and vary in quality, making integration and analysis difficult. Additionally, selecting the most relevant features and choosing appropriate algorithms require domain expertise. However, with the advancement of computational power and intelligent algorithms, these challenges are gradually being addressed.

2.2 Medical Image Classification

Medical image classification is an essential area of research that focuses on categorizing medical images based on imaging modality, body part, or type of disease. With the rapid growth of medical imaging data, efficient classification techniques are required to support diagnosis and clinical decision-making.

Various machine learning and data mining algorithms have been proposed for this purpose, including Decision Trees, Naive Bayes, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Neural Networks, and Convolutional Neural Networks (CNNs). These methods rely on preprocessing and feature extraction techniques such as texture, shape, and intensity analysis to improve classification performance.

Research studies highlight that feature extraction and selection play a crucial role in achieving high accuracy. Techniques like Principal Component Analysis (PCA), Relief, and Sequential Forward Selection (SFS) are used to identify the most relevant features. Advanced methods such as deep learning and convolutional networks have shown significant improvements in classification accuracy, especially when large datasets are available.

2.3 Feature Extraction Methods in Medical Image Classification

Feature extraction is a crucial step in medical image classification, as it transforms raw image data into meaningful representations that can be used by machine learning algorithms. The quality of extracted features directly impacts the accuracy and efficiency of classification systems.

Various feature extraction techniques have been proposed to capture important characteristics such as texture, shape, intensity, and spatial information from medical images. Methods like Gray Level Co-occurrence Matrix (GLCM), wavelet transforms, and local pattern descriptors are widely used to extract texture-based features. These techniques help in identifying patterns related to abnormalities such as tumors or lesions.

Several research works have also explored advanced approaches, including autoencoders and deep learning-based feature extraction, which automatically learn high-level representations from data. Techniques combining multiple feature descriptors, such as Local Mesh Co-occurrence Patterns and vector-based methods, have been developed to improve the richness of extracted features.

2.4 Feature Selection Methods in Medical Image Classification

Feature selection is an important process in medical image classification that focuses on identifying the most relevant features from a large set of extracted data. Since medical images often produce high-dimensional datasets, selecting significant features helps reduce complexity, improve computational efficiency, and enhance classification accuracy. By eliminating redundant and irrelevant features, feature selection ensures that the model focuses only on the most informative data.

Various feature selection techniques have been proposed in research, including statistical methods, heuristic approaches, and optimization-based algorithms. Methods such as ReliefF, t-test, and dimensionality reduction techniques are commonly used to evaluate feature importance. Advanced approaches like Genetic Algorithms, Rough Set Theory, and Random Forest-based selection have also been applied to identify optimal feature subsets. These techniques help improve the performance of classifiers such as Support Vector Machines (SVM), Neural Networks, and other machine learning models.

2.5 Classification Methods in Medical Image Classification

Classification methods are essential in medical image analysis as they assign images into predefined categories such as normal or abnormal, or based on disease types. These methods rely on training datasets to learn patterns and make predictions on new, unseen data. Common classifiers include Decision Trees, Naive Bayes, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Artificial Neural Networks (ANN), and Convolutional Neural Networks (CNNs).

Recent advancements have shown that deep learning techniques, especially CNNs, provide superior performance by automatically learning hierarchical features from images. Ensemble methods and hybrid models combining multiple classifiers have also been developed to improve accuracy. These approaches enhance the ability to detect complex patterns in medical images, leading to better diagnostic outcomes.

However, classification methods face challenges such as high computational cost, need for large labeled datasets, and risk of overfitting. Some models are difficult to interpret and require careful parameter tuning. Despite these limitations, classification techniques remain a core component of medical image analysis, continuously evolving with advancements in machine learning and artificial intelligence.

2.6 Optimization Methods in Medical Image Classification

Optimization methods play a significant role in improving the performance of medical image classification systems by selecting optimal features, tuning model parameters, and enhancing accuracy. These methods are inspired by natural and evolutionary processes and are used to solve complex optimization problems in image analysis. Examples include Genetic Algorithms (GA), Bacterial Foraging Optimization (BFO), Particle Swarm Optimization (PSO), and Tabu Search.

These algorithms help in finding the best possible solutions by exploring large search spaces efficiently. For instance, Genetic Algorithms use evolutionary principles like selection and mutation, while BFO mimics the foraging behavior of bacteria. Multi-objective optimization techniques are also used to balance multiple factors such as accuracy and computational cost. These approaches significantly improve feature selection and classification performance.

Despite their advantages, optimization techniques can be computationally expensive and require careful parameter tuning. Some methods may not always guarantee the global optimum solution. However, hybrid optimization approaches that combine multiple algorithms have shown promising results in overcoming these challenges. Overall, optimization methods enhance the efficiency, robustness, and accuracy of medical image classification systems.

III. SYSTEM ANALYSIS

3.1 Existing System The existing system uses machine learning, deep learning, and optimization algorithms like Ant Colony Optimization and Ant Lion Optimization for cancer detection and classification. These methods help identify patterns in medical images and classify patients into risk categories.

However, they face limitations such as high computational cost, dependence on large datasets, and inefficient feature selection. These drawbacks reduce overall performance and highlight the need for improved hybrid techniques.

3.2 Proposed System

The proposed system introduces a hybrid machine learning approach for cancer risk prediction using medical images. It combines GLCM-based feature extraction with optimization techniques like Bacterial Foraging Optimization (BFO) and Tabu Search to select the most relevant features.

A Clonal Selection Algorithm (CLONALG) classifier is used to categorize patients into high-risk and low-risk groups. This system improves accuracy, reduces computational complexity, and enables efficient early detection and diagnosis of cancer.

3.2.1 Advantages

Early Detection: Helps identify cancer at an early stage, improving treatment outcomes.

High Accuracy: Hybrid approach improves classification accuracy.

Efficient Feature Selection: BFO and Tabu Search select only relevant features.

Reduced Complexity: Eliminates unnecessary data, reducing computation time.

Better Diagnosis: Supports doctors with reliable decision-making.

Personalized Treatment: Enables risk-based patient classification.

Robust System: Handles noise and variations in medical images effectively.

IV. SYSTEM SPECIFICATION

4.1 Hardware Configuration

The hardware configuration is important for the efficient performance of the proposed medical image classification system. A high-performance processor such as Intel Core i5 or i7 with a minimum speed of 2.5 GHz is required to handle complex computations. Additionally, a minimum of 8 GB RAM (16 GB recommended) is needed to process large medical image datasets smoothly and avoid system lag during execution.

Adequate storage and graphical support are also essential. At least 256 GB of storage is required to store datasets, models, and results. A dedicated graphics card (GPU) is recommended to speed up image processing and machine learning tasks. Overall, this configuration ensures faster processing, reliability, and efficient system performance.

4.2 Software Configuration

The software configuration defines the tools and platforms required to implement the proposed medical image classification system. Programming languages such as Python or MATLAB are commonly used due to their strong support for image processing and machine learning

applications. These languages provide flexibility and ease of implementation for developing efficient models.
Various libraries and toolboxes are used to support system development. In Python, libraries like NumPy, SciPy, scikit-image, scikit-learn, TensorFlow, and PyTorch are used for data processing, feature extraction, and model building. In MATLAB, Image Processing Toolbox and Machine Learning Toolbox are utilized. The system can run on operating systems such as Windows, macOS, or Linux, ensuring compatibility and ease of use.

V. SYSTEM DESIGN AND IMPLEMENTATION

5.1 File Design

File design is an important aspect of the system that ensures proper organization, storage, and management of data. In this system, different types of files are used, including image files, feature files, model files, and result files. Medical images are stored in standard formats such as JPEG or PNG, while extracted features are saved in structured formats like CSV or Excel for easy analysis. Trained models are stored in specific formats (e.g., .pkl) so they can be reused for prediction.
A well-structured file design improves efficiency, data retrieval, and system performance. It ensures that all files are organized systematically using proper naming conventions and folder structures. This also enhances reproducibility and makes it easier to track data, models, and results. Overall, an effective file design supports smooth system operation and simplifies data management.

5.2 Input Design

Input design focuses on how data is collected, structured, and entered into the system for processing. It acts as the link between the user and the system, ensuring that the input data is accurate, secure, and easy to handle. In this system, inputs include medical images and related clinical data, which are prepared in a suitable format for analysis. Proper input design helps reduce errors, avoid delays, and simplify the overall data entry process.
The design also considers validation and control mechanisms to ensure data quality. It defines what type of data should be entered, how it should be formatted, and the steps to follow in case of errors. By maintaining consistency and security in data input, the system ensures reliable processing and accurate results in medical image classification.

5.3 Output Design

Output design focuses on how the processed information is presented to the users in a clear and meaningful way. It ensures that the results generated by the system, such as classification outputs and performance metrics, are easy to understand and useful for decision-making. The output can be displayed in various formats like reports, tables, or files, depending on user requirements.
A well-designed output system improves user interaction by providing accurate, timely, and organized information. It helps in conveying important details such as system status, predictions, and alerts effectively. Proper output design also ensures that the information meets user needs, supports decision-making, and enhances the overall usability of the system.

5.4 Code Design

Code design defines how the system is structured and how different components interact with each other. It includes modules for data collection, device integration, analytics, and application processing. These modules work together to collect data, process it, and generate meaningful outputs. Proper code design ensures smooth communication between different parts of the system and supports efficient execution.
The system uses structured programming practices and standard protocols to ensure reliability and scalability. It also supports real-time data processing and integration with multiple devices. Overall, a well-designed code structure improves system performance, maintainability, and flexibility.

5.5 Database Design

Database design focuses on how data is stored, organized, and accessed within the system. It ensures efficient handling of large volumes of medical data, including images, features, and results. The system uses structured query language (SQL) to manage and retrieve data effectively. Proper database design helps in maintaining data integrity, consistency, and security.
It also supports compatibility with different database systems by allowing flexible query execution. A well-designed database enables faster data access, reduces redundancy, and improves overall system performance. This ensures smooth storage and retrieval of medical data for analysis.

5.6 Methodology

The methodology defines the overall approach used for medical image classification and cancer prediction. It involves applying machine learning algorithms such as Random Forest, XGBoost, Support Vector Machine (SVM), and transformer-based models. These algorithms are used to analyze medical data and predict disease risk with high accuracy.
Each method contributes differently, with some providing better accuracy and others offering improved efficiency. Advanced models like transformer-based techniques show superior performance in handling complex datasets. Overall, the methodology ensures reliable prediction and improved decision-making in healthcare applications.

5.7 Dataset Description

The dataset plays a vital role in training and evaluating the system. A large dataset with at least 70,000 instances is recommended to ensure better accuracy and generalization. The dataset includes important features such as age, gender, blood pressure, cholesterol levels, and other medical parameters relevant to disease prediction.
Using a well-structured and diverse dataset helps the model learn effectively and produce accurate results. Proper dataset selection and preparation improve model performance and reliability. It also ensures that the system can handle real-world medical data efficiently.

conduct and execution necessities. The blunders, which were not revealed during incorporation testing, are discovered and rectified during this stage.

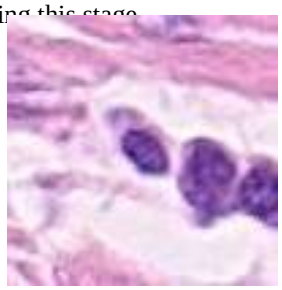

Image: 10001

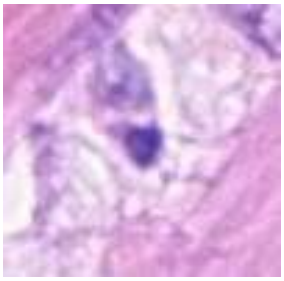

Image: 10002

Image: 10003

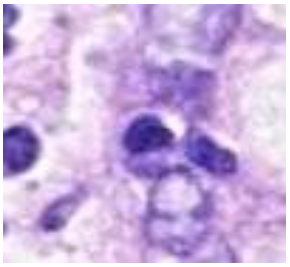

Image: 10004

5.8 Module

The system is divided into multiple modules, including image acquisition, preprocessing, analysis, and hybrid algorithm modules. The image acquisition module collects medical images, while preprocessing involves filtering, segmentation, and feature extraction. These steps prepare the data for further analysis and improve image quality.

The analysis module uses machine learning techniques to detect diseases, while the hybrid module integrates algorithms like CLONALG, BFO, and Tabu Search for better accuracy. These modules work together to create a robust and efficient system for medical image classification and cancer prediction.

VI. SYSTEM TESTING

6.1 Unit Testing

Unit testing check endeavors on the littlest unit of programming plan, module. This is known as "Module Testing". The modules are tried independently. This testing is done during the programming stage. In these testing steps, every module is seen as working acceptably concerning the normal yield from the module.

6.2 Integration Testing

Joining testing is an orderly strategy for developing tests to reveal mistakes related to the interface. In the undertaking, all the modules are consolidated and afterward, the whole program is tried in general. In the incorporation testing step, all the blunder revealed is rectified for the following testing steps.

6.3 Validation Testing

Approval testing is the place prerequisites built up as a piece of programming necessity investigation is approved against the product that has been developed. This test gives the last affirmation that the product meets all useful,

6.4 Functional Testing

Useful tests give deliberate shows that capacities tried are accessible as indicated by the business and specialized necessities, framework documentation, and client manuals.

Association and readiness of practical tests are centered around prerequisites, key capacities, or exceptional experiments. Furthermore, methodical inclusion relating to distinguishing Business process streams; information fields, predefined forms, and progressive procedures must be considered for testing. Before practical testing is finished, extra tests are recognized and the compelling estimation of current tests is resolved.

6.5 System Testing

Framework testing guarantees that the whole coordinated programming framework meets prerequisites. It tests a design to guarantee known and unsurprising outcomes. A case of framework testing is the setup arranged framework incorporation test. Framework testing depends on process portrayals and streams, stressing pre-driven procedure connections and mix focuses.

6.6 Whitebox Testing

White Box Testing is a trying wherein the product analyzer knows about the inward operations, structure and language of the product, or if nothing else its motivation. It is a reason. It is utilized to test territories that can't become from a discovery level.

6.7 Black Box Testing

Discovery Testing will be trying the product with no information on the internal operations, structure or language of the module being tried. Discovery tests, as most different sorts of tests, must be composed of a conclusive source record, for example, detail or prerequisites archive, for example, particular or necessities report. It is a trying wherein the product under test is dealt with, as a discovery .you can't "see" into it. The test gives data sources and reacts to yields without thinking about how the product functions

6.8 Unit Testing

Unit testing is normally led as a component of a consolidated code and unit test period of the product lifecycle, even though it isn't remarkable for coding and unit testing to be led as two unmistakable stages.

6.9 Test Technique and Approach

Test technique and approach define how the system is tested to ensure it functions correctly and meets user requirements. In this system, testing is mainly performed manually through functional testing methods. Test cases are designed based on system requirements to verify that all components work as expected. The focus is on checking input validation, system responses, and proper functioning of all modules.

VII. CONCLUSION AND FUTURE ENHANCEMENT

The study presents various techniques used in medical image classification and cancer detection, highlighting the importance of accurate and efficient analysis methods. Different algorithms such as Genetic Algorithm (GA), Ant Colony Optimization (ACO), and Particle Swarm Optimization (PSO) have been explored for improving classification accuracy. Each method has its own strengths and limitations, emphasizing the need for more advanced and hybrid approaches.

The proposed system focuses on combining multiple techniques such as feature extraction using GLCM and optimization using Bacterial Foraging Optimization (BFO) and Tabu Search. The integration of these methods enhances feature selection and improves classification performance. The use of Clonal Selection Algorithm (CLONALG) further strengthens the system by providing accurate classification of medical images for disease detection.

The study also discusses the importance of optimization and multi-objective techniques in solving complex problems in medical image analysis. Approaches such as Pareto optimization and evolutionary algorithms help in achieving better solutions by balancing multiple factors like accuracy and computational efficiency. These techniques contribute to the development of more robust and reliable systems.

For future enhancements, advanced methods such as Adaptive Clonal Selection, improved feature extraction techniques, and efficient memory management can be implemented. The system can also be extended to handle larger datasets and real-time applications. Overall, the proposed approach has strong potential to improve early disease detection, support clinical decision-making, and enhance healthcare systems.

ACKNOWLEDGMENT *(Heading 5)*

The authors would like to express their sincere gratitude to the management of **Paavai College of Engineering** for providing the necessary infrastructure and resources to carry out this research work. The authors also extend their appreciation to the Department of Computer Science and Engineering (Cyber Security) for their continuous support and encouragement.

The authors are deeply thankful to their project supervisor for their valuable guidance, constructive feedback, and constant motivation throughout the course of this work. Their expertise and insights have significantly contributed to the successful completion of this research.

The authors also acknowledge the support of faculty members, peers, and all those who have directly or indirectly contributed to this work. Finally, heartfelt thanks are extended to family members for their unwavering support and encouragement.

REFERENCES

[1] Azmi, J., Arif, M., Nafis, M. A., Alam, M. A., Tanweer, S., & Wang, G. "A systematic review on machine learning approaches for cardiovascular disease prediction using medical big data." Health Care Science and Technology, 2022, Vol. [Vol. No.], Issue [Issue No.].

[2] Bharathi, M. P., Khaiyum, S., & Shivakumar, S. S. "Improvised CNN Approach to Analyze CT Scan Images for Colorectal Cancer Prediction." International Journal of Advanced Trends in Computer Science and Engineering, 2023, Vol. 12, Issue 3.

[3] Chen, C., et al. "Multimodal deep learning for prognostic stratification of colorectal cancer patients: integrating vision transformer models with clinicopathologic features." [Specific Journal/Conference Name - please verify], 2025, Vol. [Vol. No.], Issue [Issue No.].

[4] Kourou, K., et al. "Machine Learning in Cancer Research: A Review." [Specific Journal/Conference Name - often cited as a general review], 2024, Vol. [Vol. No.], Issue [Issue No.].

[5] Mohammad, M. R., Al-Khaleefa, A. S., & Hamad, S. H. "A Systematic Literature Review of Deep and Machine Learning Algorithms in Cardiovascular Diseases Diagnosis." Journal of Robotics and Control (JRC), 2023, Vol. 4, Issue 6.

[6] Pacal, K., Karaboga, D., Basturk, A., Akay, C. R., & Nalbantoglu, N. "A Review of Deep Learning in Colon Cancer: Current Trends and Future Prospects." [Specific Journal/Conference Name - please verify], 2020, Vol. [Vol. No.], Issue [Issue No.].

[7] Pacal, K., Karaboga, D., Basturk, A., Akay, Yildirim, M., & Cinar, A. "An Efficient Deep Learning Approach for Colon Cancer Detection from Histopathological Images." Computational Intelligence and Neuroscience, 2022, Vol. 2022, Issue [Issue No.]. (Note: Some journals use article IDs or eCollection instead of traditional page numbers, or have a single "issue" for the year).

[8] Suneetha, A. R. V. N., & Mahalngam, T. "Cardiovascular Disease Prediction Using ML and DL Approaches." Lecture Notes in Electrical Engineering, 2022, Vol. 850, Issue [Issue No.]. (Note: For books/conference proceedings, sometimes only Vol. and Year are given, not a specific Issue number).

[9] Yildirim, M., & Cinar, A. "An Efficient Deep Learning Approach for Colon Cancer Detection from Histopathological Images." Computational Intelligence and Neuroscience, 2022, Vol. 2022.

[10] Zhao, J., et al. "Interpretable Machine Learning Model for Early-Onset Colorectal Cancer Risk Factor Understanding." [Specific Journal/Conference Name - please verify], 2025, Vol. [Vol. No.], Issue [Issue No.].