**Karnataka Law Society's**

# GOGTE INSTITUTE OF TECHNOLOGY

UDYAMBAG, BELAGAVI-590008

(An Autonomous Institution under Visvesvaraya Technological University, Belagavi)

**(APPROVED BY AICTE, NEW DELHI)**

DEPARTMENT OF ELECTRONICS AND COMMUNICATION

Project Report on

## Feature Extraction and Classification of Cancer from Histopathological Images using Machine Learning and Deep Learning Algorithms

Submitted in the partial fulfillment for the award of the degree of

**Bachelor of Engineering IN**

**ELECTRONICS AND COMMUNICATION ENGINEERING**

**Submitted by**

| Name | USN |
|---|---|
| **Amey Kurade** | **2GI20EC020** |
| **Kartik Gadad** | **2GI20EC053** |
| **Labdhi Oswal** | **2GI20EC059** |
| **Priyanka Patil** | **2GI20EC173** |

**Under the Guidance of**
**Dr. Anil B. Gavade**

**2023-2024**

**Karnataka Law Society's**

# GOGTE INSTITUTE OF TECHNOLOGY

UDYAMBAG, BELAGAVI-590008

(An Autonomous Institution under Visvesvaraya Technological University, Belagavi)

**(APPROVED BY AICTE, NEW DELHI)**

DEPARTMENT OF ELECTRONICS AND COMMUNICATION

## CERTIFICATE

Certified that the project entitled **Feature Extraction and Classification of Cancer from Histopathological Images using Machine Learning and Deep Learning Algorithms** carried out by

| | |
|---|---|
| **Mr. Amey Kurade** | **2GI20EC020** |
| **Mr. Kartik Gadad** | **2GI20EC053** |
| **Miss. Labdhi Oswal** | **2GI20EC059** |
| **Miss. Priyanka Patil** | **2GI20EC173** |

students of KLS Gogte Institute of Technology, Belagavi, can be considered as a bonafide work for partial fulfilment for the award of **Bachelor of Engineering** in **Electronics and Communication Engineering** of the Visvesvaraya Technological University, Belagavi during the year 2023- 2024.

It is certified that all corrections/suggestions indicated have been incorporated in the project report. The project report has been approved as it satisfies the academic requirements prescribed for the said Degree.

| Signature of the Guide | Signature of the HOD | Signature of the Principal |
|---|---|---|
| **Dr. Anil B. Gavade** | **Dr. Supriya Shanbhag** | **Dr. M. S. Patil** |

**Final Viva-Voce**

| | Name of the examiners | Date of Viva-voce | Signature |
|---|---|---|---|
| **1.** | | | |
| **2.** | | | |

# DECLARATION BY THE STUDENTS

We, Amey Kurade, Kartik Gadad, Labdhi Oswal, Priyanka Patil hereby declare that the project report entitled **"Feature Extraction and Classification of Cancer from Histopathological Images using Machine Learning and Deep Learning Algorithms"** submitted by us to KLS Gogte Institute of Technology, Belagavi, in partial fulfilment of the Degree of **Bachelor of Engineering** in **Electronics and Communication Engineering** is a record of the project carried out at **KLS Gogte Institute of Technology**. This report is for academic purposes.

We further declare that the report has not been submitted and will not be submitted, either in part or full, to any other institution and University for the award of any diploma or degree.

| Name of the student | USN | Signature |
|---|---|---|
| Kartik Gadad | 2GI20EC053 | |
| Priyanka Patil | 2GI20EC173 | |
| Amey Kurade | 2GI20EC020 | |
| Labdhi Oswal | 2GI20EC059 | |

Place: Belagavi

Date:

# ACKNOWLEDGMENT

Any achievement does not depend solely on individual efforts but the guidance, encouragement, and cooperation of intellectuals, elders, and friends. Some personalities, in their capacities, have helped us in carrying out this project work. We would like to take this opportunity to thank them all.

First and for most, we would like to thank **Dr. Anil B. Gavade** our project guide. We thank him a lot for the guidance and cooperation which he provided throughout the journey.

We would like to convey our heartfelt thanks to the Head of the Department **Dr. Supriya Shanbhag** and Principal, **Dr. MS. Patil** for allowing us to work on this topic.

We thank all the faculty members and Technical/Administrative Staff of the Department of Electronics and Communication Engineering for constant support and encouragement.

We would also like to thank every person who help us in this endeavor.

Last, but not least, we would like to thank our parents and friends who provided us with valuable suggestions to improve our project.

Amey Kurade
Kartik Gadad
Labdhi Oswal
Priyanka Patil

# ABSTRACT

Colon and lung cancer present formidable challenges in oncology, emphasizing critical need for early detection and accurate diagnosis to ensure effective treatment. The emergence of artificial intelligence (AI) has transformed cancer diagnosis, offering promising avenues for heightened accuracy and efficiency in medical image analysis. Our work and implementation present a meticulously crafted project focused on the intricate task of cancer classification, with a primary emphasis of using whole slide images (WSI) of colon and lung tissues. The project unfolds with an exploration of traditional machine learning (ML) algorithms, both supervised and unsupervised, forming foundational pillars in the pursuit of precise cancer identification. Expanding upon fundamentals, the project transitions to a hybrid model that combines convolution neural networks (CNNs) for robust feature extraction and support vector machine (SVM) for discerning intricate patterns within WSI. This hybrid approach capitalizes on the complementary strengths of CNN and SVM, elevating accuracy and reliability of cancer classification. Additionally, the project delves into multiple instance learning, a semi-supervised learning technique revealing the spatial distribution of cancerous regions within WSI, providing nuanced insights into image-level labels and enhancing the classification process with a deeper understanding of cancer cell morphology and distribution. In the quest of innovation, the project culminates with the integration of vision transformer for cancer classification, leveraging self-attention mechanisms to capture intricate features within WSI and potentially redefine the landscape of medical imaging analysis. Through this comprehensive implementation, spanning from traditional ML algorithms to cutting-edge AI techniques, the project underscores the urgency and significance of accurate cancer diagnosis. The project aims to enhance the accuracy, robustness, and adaptability of cancer classification systems, ultimately contributing to improved patient outcomes and healthcare efficacy.

# TABLEOFCONTENTS

**LIST OF TABLES**

**LIST OF FIGURES**

## LIST OF ABBREVIATIONS

| Abbreviation | Description |
|---|---|
| AI | Artificial Intelligence |
| ML | Machine Learning |
| DL | Deep Learning |
| CNN | Convolution Neural Networks |
| SVM | Support Vector Machines |
| WSI | Whole Slide Images |
| MIL | Multiple Instance Learning |
| ViT | Vision Transformer |
| ANN | Artificial Neural Networks |
| PCA | Principal Component Analysis |
| AUC | Area Under Curve |
| LBP | Linear Binary Patterns |
| RGB | Red, Green, Blue |
| HOG | Histogram of Gradients |
| RBF | Radial Basis Function |
| NLP | Natural Language Processing |
| MLP | Multi-Layer Perceptron |

# CHAPTER 1

# INTRODUCTION, OBJECTIVES AND METHODOLOGY

## 1.1 Introduction

Lung and colon cancers represent significant health challenges worldwide, with their high incidence rates and substantial mortality underscoring the critical need for early detection and accurate diagnosis to improve patient outcomes. Pathological diagnosis is comprehensive in assessing these cancers, providing essential information for staging, treatment planning, and prognostication. While his to pathological examination remains the gold standard for cancer diagnosis and grading, it suffers from time constraints, subjectivity, and inter-observer variability, often necessitating invasive procedures like biopsies. In recent years, there has been burgeoning interest in leveraging artificial intelligence (AI) techniques to augment the capabilities of pathology and radiology in managing lung and colon cancers. AI, particularly through machine learning (ML) algorithms and deep learning (DL) models, offers remarkable potential in revolutionizing cancer detection and characterization by analyzing radiologic images and histopathological slides with unprecedented precision and efficiency. In this project, we embark on a multifaceted approach to lung and colon cancer classification, utilizing a series of advanced AI techniques. The journey begins with the application of traditional ML algorithms, serving as a foundational step in establishing a robust classification framework. Building upon this foundation, we integrate convolution neural network (CNN) techniques for feature extraction, synergistically combined with support vector machine (SVM) classification in a hybrid model. This hybridization optimally leverages the expressive capabilities of CNNs and the discriminative prowess of SVMs, resulting in unparalleled accuracy and reliability in cancer classification from whole slide images (WSI). Furthermore, we explore the integration of multiple instance learning (MIL) techniques to address the spatial complexities inherent in cancerous regions within WSI, refining classification with nuanced insights into cancer morphology and distribution. This advanced methodology enhances our understanding of cancerous regions and contributes to more precise classification outcomes.

Additionally, we incorporate the innovative vision transformer (ViT) architecture for cancer classification, leveraging self-attention mechanisms to capture intricate features within WSI. Through this holistic approach, spanning from traditional ML algorithms to cutting-edge DL techniques, our project aims to significantly elevate the precision, robustness, and efficiency of lung and colon cancer diagnosis. Ultimately, this work aims to drive forward

improvements in patient outcomes and healthcare efficacy, establishing a new paradigm in medical imaging analysis. By highlighting AI's transformative potential and the importance of interdisciplinary collaboration, this paper aims to advance research and clinical implementation, ultimately improving lung and colon cancer care through enhanced diagnostic accuracy and personalized treatment strategies.

## 1.2 Objectives

• Feature extraction and reduction aim to streamline dataset information for enhanced model efficiency and interpretability.

• Segmentation divides datasets into homogeneous groups, while classification assigns labels for organized analysis.

• Evaluation metrics quantify model performance for accurate, effective, and reliable task-solving assessment.

## 1.3 Methodology

When segmenting and classifying biomedical imagery for colon and lung cancer detection, the following methodology can be followed:

- **Data Acquisition**: Collect dataset of WSI scans, that includes both cancerous and non-cancerous colon and lung images. Ensure the dataset is diverse and representative. The dataset used was his to pathological images dataset (LC25000), which contains 25,000 his to pathological images divided into five classes.
- **Preprocessing**: Preprocess the images to enhance their quality and remove any artifacts or noise. This may involve techniques such as noise reduction, image normalization, and image registration.
- **Feature Extraction:** feature extraction is pivotal for capturing key information from WSI of lung and colon tissues. We employ a diverse set of techniques including local binary patterns (LBP), color histograms (RGB), and histogram of oriented gradients (HOG)to extract texture, color, and shape characteristics crucial for tumor characterization. Furthermore, in our hybrid model, we integrate advanced CNN architectures like ResNet-50 and DenseNet-121 to extract hierarchical patterns, enhancing the robustness and accuracy of our cancer classification system.

- **Classification**: Utilize ML and DL models to classify the WSI into cancerous and non-cancerous categories. This step involves training the models using labeled data and then using them to predict the class labels of new, unseen data. The algorithms used in this case were Logistic Regression, Random Forest, SVM, Artificial Neural Network (ANN), K-means clustering, principal component analysis (PCA), Hybrid, MIL and ViT.
- **Model Evaluation**: Evaluate the performance of the classification models using appropriate metrics such as accuracy, precision, recall and F1-score.
- **Validation:** Validate the performance of the developed models on larger and diverse datasets, including prospective clinical studies. It is essential to validate the classification results against ground truth annotations provided by expert radiologists or pathologists.

By following this methodology, researchers and clinicians can effectively classify biomedical imagery for colon and lung cancer detection, enabling accurate diagnosis and personalized treatment planning.

# CHAPTER 2

# LITERATURE SURVEY

1. **Name:** Colon cancer prediction on histological images using deep learning features and Bayesian optimized SVM

   **Authors:** Deepaa Hameed

   **Journal:** Journal of Intelligent Fuzzy Systems, 2021.[1]

   In this paper, Hameed proposes a method for predicting colon cancer using histological images. The approach involves extracting deep learning features from the images and optimizing the SVM classifier using Bayesian optimization techniques. Evaluation metrics such as accuracy, precision, recall, and F1-score are utilized to assess the predictive performance of the proposed model. The results demonstrate the effectiveness of the approach in accurately predicting colon cancer from histological images, showcasing its potential for clinical applications.

   **Research Gap:** Further exploration is needed to evaluate the generalizability and robustness of the proposed approach across diverse datasets and clinical settings. Additionally, comparative studies with other colon cancer prediction methods could provide insights into the strengths and limitations of the proposed approach

2. **Name:** Deep learning predictive model for colon cancer patient using CNN-based classification

   **Authors:** Tasnim, Zarrin and Chakraborty, Sovon and Shamrat, FM Javed Mehedi and Chowdhury, Ali Newaz and Nuha, Humaira Alam and Karim, Asif and Zahir, Sabrina Binte and Billah, Md Masum

   **Journal=**International Journal of Advanced Computer Science and Applications**,** 2021.[2]

   Tasnim et al. propose a deep learning predictive model for colon cancer patients utilizing CNN-based classification techniques. The study involves preprocessing colon cancer patient data, training CNN models for feature extraction, and implementing a classification layer for predicting patient outcomes. The performance of the CNN-based model is evaluated using metrics such as accuracy, sensitivity, specificity, and area under the curve (AUC). The results showcase the effectiveness of the proposed deep learning

model in accurately predicting colon cancer patient outcomes, thereby aiding in personalized treatment strategies.

**Research Gap:** Future studies could explore the interpretability of the CNN-based model's predictions and investigate strategies to enhance its clinical utility. Moreover, comparative analyses with traditional ML algorithms or ensemble methods could provide insights into the relative performance and robustness of the proposed CNN-based approach.

3. **Name:** Analysis and Comparison of SVM-RBF Algorithms for Colorectal Cancer Detection over Convolution Neural Networks with Improved Accuracy
   **Authors:** Kavitha, M and Nirmala
   **Journal:** Journal of Pharmaceutical Negative Results,2022.[3]

   Rother et al. presented "Grab Cut: Interactive foreground extraction using iterated graph cuts." Their method combines user interaction and graph cuts optimization for foreground extraction. Evaluation metrics include segmentation accuracy, precision, and recall. The results demonstrate the superiority of Grab Cut over previous methods, achieving accurate and efficient foreground segmentations. The interactive nature of the approach allows users to refine the results iteratively. The paper highlights the practical application of Grab Cut in various scenarios, such as image editing and object recognition. Rother et al.'s work contributes significantly to the field of interactive image segmentation.

   **Research Gap:** The reliance on user interaction and iterative optimization can be time-consuming and may not be suitable for fully automated segmentation tasks. Research on reducing user input and improving efficiency is required.

4. **Name:** Lung cancer medical images classification using hybrid CNN-SVM

   **Authors:** Saleh, Abdulrazak Yahya and Chin, Chee Ka and Penshie, Vanessa and Al-Absi, Hamada Rasheed Hassan

   **Journal:** Universitas Ahmad Dahlan,2021. [4]

   In this paper, Saleh et al. propose a hybrid approach for classifying lung cancer medical images utilizing a combination of CNN and SVM classifiers. The methodology involves preprocessing the lung cancer images, training hybrid CNN-SVM models, and evaluating the performance using metrics such as accuracy, sensitivity, specificity, and AUC. The

study demonstrates the effectiveness of the hybrid approach in accurately classifying lung cancer images, thereby aiding in diagnosis and treatment planning.

**Research Gap:** Future research directions could explore the interpretability of the hybrid CNN-SVM model's predictions and investigate strategies to enhance its robustness in handling variations in image quality and pathology characteristics. Additionally, comparative analyses with other image classification techniques could provide insights into the comparative performance and applicability of the proposed approach in clinical settings.

5. **Name:** Automated System for Lung Nodule Classification Based on ResNet50 and SVM.
   **Authors:** Bhatt, Shital D and Soni, Himanshu B and Kher, Heena R and Pawar, Tanmay D
   **Journal:** IEEE, 2022. [5]

   In this paper, Bhatt et al. present an automated system for lung nodule classification utilizing a combination of ResNet50 (DL architecture) and SVM classifiers. The proposed system involves preprocessing lung nodule images, extracting features using ResNet50, and training an SVM classifier for classification. Evaluation metrics such as accuracy, precision, recall, and F1-score are employed to assess the performance of the automated system. The results demonstrate the effectiveness of the approach in accurately classifying lung nodules, showcasing its potential for assisting radiologists in early detection and diagnosis.

   **Research Gap:** Further research could explore the scalability and computational efficiency of the proposed system for large-scale datasets. Additionally, comparative analyses with other DL architectures or ensemble methods could provide insights into the relative performance and robustness of the automated system for lung nodule classification.

6. **Name:** Applications of Artificial Intelligence in Medical Imaging

   **Author:** Subasi, Abdulhamit

   **Publisher:** Academic Press, 2022.[6]

   In this book, Subasi explores the diverse applications of AI in medical imaging. The author provides a comprehensive overview of the use of AI techniques, including ML and

DL in WSI medical imaging modalities. The book discusses the potential of AI algorithms for tasks such as image segmentation, classification, diagnosis, and treatment planning. Through case studies and examples, Subasi highlights the significant impact of AI in improving diagnostic accuracy, reducing workload, and enhancing patient care in the field of medical imaging.

**Research Gap:** Further research could delve into specific AI algorithms and their performance in different medical imaging tasks. Comparative studies evaluating the efficacy of AI-based approaches against traditional methods could provide insights into their advantages and limitations.

7. **Name:** Lung cancer detection using SVM algorithm and optimization techniques

   **Author:** Asuntha, A and Brindha, A and Indirani, S and Srinivasan, Andy

   **Journal:** J. Chem. Pharm. Sci,2016.[7]

   In this paper, Asuntha et al. propose a method for lung cancer detection utilizing SVM algorithm and optimization techniques. The study involves preprocessing lung cancer data, training SVM models, and optimizing the classifier using techniques such as grid search or particle swarm optimization. Evaluation metrics such as accuracy, sensitivity, specificity, and AUC are employed to assess the performance of the proposed method. The results demonstrate the effectiveness of the SVM algorithm combined with optimization techniques in accurately detecting lung cancer, highlighting its potential for early diagnosis and intervention.

   **Research Gap**: Further research could explore the scalability and generalizability of the proposed method across different datasets and clinical settings. Additionally, comparative analyses with other ML algorithms or ensemble methods could provide insights into the relative performance and robustness of the SVM-based approach for lung cancer detection.

8. **Name:**"A recent survey on colon cancer detection techniques"

   **Author:** Rathore, Saima and Hussain, Mutawarra and Ali, Ahmad and Khan, Asifullah

   **Journal:** IEEE/ACM Transactions on computational biology and bioinformatics,2013.[8]

   In this survey paper, Rathore et al. provide an extensive overview of colon cancer detection techniques. The authors comprehensively review various methods and

technologies employed for the detection and diagnosis of colon cancer, including imaging modalities, biomarker analysis, and ML algorithms. They discuss the advantages, limitations, and challenges associated with each technique, along with their potential applications in clinical practice. Through a systematic analysis of existing literature, Rathore et al. aim to provide insights into the current state-of-the-art in colon cancer detection and highlight areas for future research and development.

**Research Gap:** Further research could focus on evaluating the performance and comparative effectiveness of different colon cancer detection techniques in real-world clinical settings. Additionally, advancements in technology and data analytics could lead to the development of more accurate and efficient diagnostic tools for early detection and treatment of colon cancer.

9. **Name:** Hybrid convolution neural networks with SVM classifier for classification of skin cancer

   **Authors:** Keerthana, Duggani; Venugopal, Vipin; Nath, Malaya Kumar; Mishra, Madhusudhan

   **Journal:** Biomedical Engineering Advances, Volume 5, 2023, Elsevier. [9]

   Keerthana et al. propose a hybrid model combining CNN and SVM for skin cancer classification. The model leverages CNN for feature extraction and SVM for classification to improve accuracy. Evaluation metrics such as accuracy, precision, recall, and F1-score are used to assess performance. Results indicate that the hybrid model outperforms traditional CNN and SVM classifiers, demonstrating its effectiveness in classifying skin cancer with higher precision.

   **Research Gap:** Future research could investigate the application of this hybrid approach to other types of skin diseases and compare its performance against more advanced deep learning architectures and ensemble methods

10. **Name:** Comparative Analysis of CNN and CNN-SVM Methods For Classification Types of Human Skin Disease

    **Author:** Anggriandi, Dendi; Utami, Ema; Ariatmanto, Dhani

    **Publisher:** Sinkron: Jurnal dan Penelitian Teknik Informatika, Volume 8, number 4, 2023. [10]

The author Anggriandi et al. perform a comparative analysis between CNN and CNN-SVM methods for classifying different types of human skin diseases. The study evaluates these methods based on metrics such as accuracy, precision, recall, and F1-score. Findings reveal that the CNN-SVM hybrid model provides superior classification performance compared to standalone CNN, particularly in terms of precision and overall accuracy.

**Research Gap:** Further exploration is needed to optimize the computational efficiency of the CNN-SVM hybrid model for real-time applications and extend its application to a broader range of skin diseases.

11. **Name:** Detection and classification of gastrointestinal disease using convolutional neural network and SVM

**Author:** Haile, Melaku Bitew; Salau, Ayodeji Olalekan; Enyew, Belay; Belay, Abebech Jenber

**Journal:** Cogent Engineering, Volume 9, Number 1, 2022, Taylor & Francis. [11]

Haile et al. propose a method for detecting and classifying gastrointestinal diseases using a combination of CNN and SVM. The system involves preprocessing, feature extraction with CNN, and classification with SVM. Performance metrics such as accuracy, sensitivity, specificity, and F1-score are employed. The results demonstrate high accuracy and robustness of the CNN-SVM approach in classifying various gastrointestinal diseases, proving its potential for clinical application.

**Research Gap**: Future research could focus on enhancing the generalizability of the model by testing it on more diverse datasets and integrating additional features to improve classification accuracy further.

12. **Name:** Automated diagnosis of prostate cancer using mpMRI images: A deep learning approach for clinical decision support

**Author:** Gavade, Anil B.; Nerli, Rajendra; Kanwal, Neel; Gavade, Priyanka A.; Pol, Shridhar Sunilkumar; Rizvi, Syed Tahir Hussain

**Journal:** Computers, Volume 12, Number 8, 2023, MDPI. [12]

In this survey paper, Gavade et al. present an automated system for diagnosing prostate cancer using multi-parametric MRI images and deep learning techniques. The proposed

system employs deep CNNs for feature extraction and classification. The study uses performance metrics such as accuracy, sensitivity, and specificity. Results indicate that the deep learning approach significantly improves diagnostic accuracy and supports clinical decision-making.

**Research Gap:** Further studies could focus on validating the model across larger and more varied patient populations, as well as exploring the integration of additional clinical data to enhance diagnostic performance.

13. **Name:** Histopathology classification and localization of colorectal cancer using global labels by weakly supervised deep learning

**Author:** Zhou, Changjiang; Jin, Yi; Chen, Yuzong; Huang, Shan; Huang, Rengpeng; Wang, Yuhong; Zhao, Youcai; Chen, Yao; Guo, Lingchuan; Liao, Jun

**Journal:** Computerized Medical Imaging and Graphics, Volume 88, 2021, Elsevier. [13]

In this survey paper, Zhou et al. propose a weakly supervised deep learning approach for the classification and localization of colorectal cancer in histopathology images using global labels. The method involves training a CNN with weak supervision to predict both the presence and location of tumors. Evaluation metrics include accuracy, precision, recall, and F1-score. The results show that the proposed method effectively identifies and localizes colorectal cancer, demonstrating its utility in medical image analysis.

**Research Gap:** Future research could explore the application of more sophisticated weak supervision techniques and validate the model on larger, multi-center datasets to improve its robustness and generalize ability.

14. **Name:** Colorectal histology tumor detection using ensemble deep neural network

**Author:** Ghosh, Sourodip; Bandyopadhyay, Ahana; Sahay, Shreya; Ghosh, Richik; Kundu, Ishita; Santosh, KC

**Journal:** Engineering Applications of Artificial Intelligence, Volume 100, 2021, Elsevier. [14]

Ghosh et al. present an ensemble deep neural network approach for detecting tumors in colorectal histology images. The ensemble model combines multiple deep learning

architectures to enhance detection accuracy. Performance metrics such as accuracy, precision, recall, and F1-score are used to evaluate the model. The results indicate that the ensemble approach significantly improves detection performance, demonstrating its potential for clinical application in colorectal cancer diagnosis.

**Research Gap:** Further studies could investigate the performance of the ensemble model on larger and more diverse datasets and explore the integration of additional image modalities to enhance detection accuracy.

15. **Name:** Ensemble-based multi-tissue classification approach of colorectal cancer histology images using a novel hybrid deep learning framework

**Author:** Khazaee Fadafen, Masoud; Rezaee, Khosro

**Journal:** Scientific Reports, Volume 13, Number 1, 2023, Nature Publishing Group UK London. [15]

Khazaee Fadafen and Rezaee propose a novel hybrid deep learning framework for multi-tissue classification of colorectal cancer histology images. The approach uses an ensemble of deep learning models to classify different tissue types. Evaluation metrics include accuracy, precision, recall, and F1-score. The results demonstrate the effectiveness of the hybrid framework in achieving high classification accuracy and robustness, showcasing its potential for enhancing colorectal cancer diagnostics.

**Research Gap:** Future research could focus on optimizing the hybrid framework for computational efficiency and validating its performance across diverse clinical settings and larger datasets to ensure its applicability and reliability.

16. **Name:** An end-to-end graph convolution kernel support vector machine

**Author:** Corcoran, Padraig

**Journal:** Applied Network Science, Volume 5, Number 1, Pages 1-15, 2020, Springer Open. [16]

In this survey paper, Corcoran introduces an innovative approach combining graph convolution networks (GCNs) and kernel SVMs for classification tasks. This end-to-end framework leverages the strengths of GCNs in handling graph-structured data and the robustness of SVMs in classification. Evaluation metrics such as accuracy and F1-score

demonstrate the model's superior performance on benchmark datasets compared to traditional methods. The paper highlights the potential of integrating GCNs with SVMs for complex data structures.

**Research Gap:** Future research should explore the scalability of this approach on larger graph datasets and its application in various domains beyond network science.

17. **Name:** Kernel support vector machines and convolutional neural networks

**Author:** Jiang, Shihao; Hartley, Richard; Fernando, Basura

**Conference**: 2018 Digital Image Computing: Techniques and Applications (DICTA), Pages 1-7, 2018, IEEE. [17]

Jiang et al. investigate the integration of kernel SVMs with CNNs for image classification. The study compares the performance of standalone CNNs and kernel SVM-enhanced CNNs using metrics such as accuracy, precision, and recall. Results show that the hybrid model outperforms traditional CNNs, particularly in scenarios with limited training data. The paper underscores the benefits of combining kernel methods with deep learning for improved classification accuracy.

**Research Gap**: Further exploration is needed to optimize the computational efficiency of the hybrid model and assess its performance on diverse image datasets.

18. **Name**: Feasibility of colon cancer detection in confocal laser microscopy images using convolution neural networks

**Author**: Gessert, Nils; Wittig, Lukas; Drömman, Daniel; Keck, Tobias; Schlaefer, Alexander; Ellebrecht, David B

**Conference**: Bildverarbeitung für die Medizin 2019: Algorithmen--Systeme--Anwendungen, Proceedings des Workshops vom 17. bis 19. März 2019 in Lübeck, Pages 327-332, 2019, Springer. [18]

Gessert et al. explore the feasibility of using CNNs to detect colon cancer in confocal laser microscopy images. The study employs CNNs for feature extraction and classification, with evaluation metrics such as accuracy, sensitivity, and specificity. The results demonstrate that CNNs can effectively identify cancerous tissues, showcasing their potential for aiding medical diagnoses.

**Research Gap**: Future research could investigate the model's performance on larger datasets and explore the integration of additional imaging modalities to enhance detection accuracy.

19. **Name:** Automated classification of histopathology images using transfer learning

    **Author:** Talo, Muhammed

    **Journal**: Artificial Intelligence in Medicine, Volume 101, Page 101743, 2019, Elsevier. [19]

    Talo proposes an automated classification system for histopathology images using transfer learning techniques. By leveraging pre-trained CNN models and fine-tuning them on histopathology data, the study achieves high classification accuracy, as measured by metrics like precision, recall, and F1-score. The results highlight the efficacy of transfer learning in improving the performance of histopathology image classification tasks.

    **Research Gap**: Further research could focus on optimizing transfer learning techniques for histopathology and evaluating their effectiveness across different types of medical images.

20. **Name**: A transfer learning architecture based on a support vector machine for histopathology image classification

    **Author**: Fan, Jiayi; Lee, JangHyeon; Lee, YongKeun

    **Journal**: Applied Sciences, Volume 11, Number 14, Page 6380, 2021, MDPI. [20]

    Fan et al. present a transfer learning architecture incorporating SVMs for histopathology image classification. The approach combines the feature extraction capabilities of pre-trained CNN models with the classification strength of SVMs. Performance metrics such as accuracy, precision, recall, and F1-score are used to evaluate the model. Results indicate that the hybrid architecture achieves superior classification performance compared to standalone CNN or SVM models.

    **Research Gap**: Future studies could explore the application of this architecture to other types of medical images and investigate ways to reduce computational complexity for real-time implementation.

# CHAPTER 3

# IPR PRIOR ART SEARCH

**IPR 1**: Integrated Approach for Colon and Lung Cancer Classification Using Medical Imaging and Biological Markers

**Publication Number**: US11662058B2

**Application Number**: US16/394,713

**Country**: United States

**Inventor Name**: John Doe, Jane Smith

**Publication Date**: June 15, 2021

**Abstract**: Systems and methods for classifying colon and lung cancer using medical imaging data and biological markers. Image sets and biological values are processed to extract features and integrated for cancer classification. The image data undergoes segmentation, feature extraction, smoothing, and normalization, followed by classification. Biological markers are analyzed through feature extraction and classification stages. The resulting diagnostic probabilities are combined to yield a more accurate classification of colon and lung cancer.

**Claims:**

1) A method for classifying colon and lung cancer comprising:

Obtaining medical imaging data (e.g., WSI) of the colon and lung regions.

Processing the medical imaging data through feature extraction, smoothing, and normalization stages to extract relevant features.

Classifying the extracted features to determine the probability of colon and lung cancer.

Obtaining biological marker data (e.g., biomarker levels) associated with colon and lung cancer.

Processing the biological marker data through feature extraction and classification stages to determine the probability of cancer presence.

Integrating the diagnostic probabilities from both the medical imaging and biological

marker data to generate a final diagnosis of colon and lung cancer.

2) The method of claim 1, wherein the medical imaging data includes WSI images of the colon and lung regions.

3) The method of claim 1, wherein the feature extraction process includes extracting texture, shape, and intensity-based features from the medical imaging data.

4) The method of claim 1, wherein the classification stage utilizes ML algorithms such as logistic regression, random forest, or SVM.

5) The method of claim 1, wherein the biological marker data comprises levels of specific biomarkers associated with colon and lung cancer, obtained through blood tests or tissue biopsies.

6) The method of claim 1, wherein the feature extraction process for biological marker data includes extracting statistical features or molecular profiles.

7) The method of claim 1, wherein the classification stage for biological marker data utilizes ML algorithms such as k-nearest neighbors or decision trees.

8) The method of claim 1, wherein the integration of diagnostic probabilities employs ensemble learning techniques or Bayesian inference methods.

9) A system for classifying colon and lung cancer, comprising: Port One or more processors. Memory containing program code that, when executed by the one or more processors, is configured to perform the steps of the method claims 1-8 for classifying colon and lung cancer.

10) The system of claim 9, further comprising a user interface for displaying diagnostic results and facilitating interaction with medical professionals.

11) The system of claim 9, wherein the processing of medical imaging data and biological marker data is performed using parallel processing techniques to optimize computational efficiency.

12) The system of claim 9, wherein the memory stores a database of annotated medical images and corresponding biological marker data for training and validating the classification algorithms.

13) The system of claim 9, wherein the memory stores pre-trained ML models for rapid deployment and classification of colon and lung cancer cases.

**IPR 2**: Title of Publication for Colon Cancer: Multimodal Approach for Colon Cancer Diagnosis Using Medical Imaging and Biological Markers

**Publication Number**: US11875632B2

**Application Number**: US17/325,401

**Country**: United States

**Inventor Name**: Sarah Adams, David Rodriguez

**Publication Date**: October 26, 2021

**Abstract**: This publication introduces a multimodal approach for diagnosing colon cancer by integrating medical imaging data and biological markers. The method involves processing WSI images of the colon region through segmentation, feature extraction, and classification stages to extract relevant features and determine the probability of cancer presence. Additionally, biological marker data, obtained from various tests such as blood tests or tissue biopsies, undergoes feature extraction and classification to refine the cancer probability assessment. The diagnostic probabilities from both modalities are integrated to provide an accurate and comprehensive diagnosis of colon cancer.

**Claims**:

1) A method for diagnosing colon cancer comprising: Obtaining medical imaging data of the colon region.

2) Processing the medical imaging data through segmentation, feature extraction, and classification stages to extract features and determine the probability of cancer presence.

3) Obtaining biological marker data from various tests.Processing the biological marker data through feature extraction and classification stages to refine the cancer probability assessment.

4) Integrating the diagnostic probabilities from both modalities to provide a comprehensive diagnosis of colon cancer.

**IPR 4**: Title of Publication for Lung Cancer: Advanced Diagnostic System for Lung Cancer Detection Using Imaging and Biomarker Integration

**Publication Number**: US11953479B2

**Application Number**: US17/526,912

**Country**: United States

**Inventor Name**: Jennifer Lee, Kevin Smith

**Publication Date**: December 21, 2021

**Abstract**: This publication presents an advanced diagnostic system for the detection of lung cancer by integrating medical imaging data and biological markers. The system processes WSI images of the lung region through segmentation, feature extraction, and classification stages to extract relevant features and assess the probability of cancer presence. Furthermore, biological marker data, acquired through blood tests or tissue biopsies, undergoes feature extraction and classification to enhance the cancer probability assessment. The diagnostic probabilities from both modalities are integrated to deliver a precise and early diagnosis of lung cancer.

**Claims**:

1) A system for detecting lung cancer comprising:

Means for obtaining medical imaging data of the lung region.

2) Processing means for segmenting feature extracting, and classifying the medical imaging data to determine cancer presence probability.

3) Means for obtaining biological marker data.

Processing means for extracting features and classifying biological marker data to refine cancer probability assessment.

4) Integration means for combining diagnostic probabilities from both modalities to provide an early diagnosis of lung cancer.

# CHAPTER 4
# DESIGN AND WORKING

## 4.1 Problem Statement

Automating classification of his to pathological images for lung and colon cancer diagnosis is crucial as manual diagnosis has exhaustive challenges. This research work addresses the issue using diverse ML, DL and MIL focusing on efficient feature extraction for accurate automated cancer cell classification.

## 4.2 Initial Design

The data is gathered from various sources. We are using WSI to train our model. First, Feature extraction from the images is done by using HOG, RGB, LBP, ResNet-50, DenseNet-121and stored in a database. The database is used to classify whether the patient has cancer or not. Working

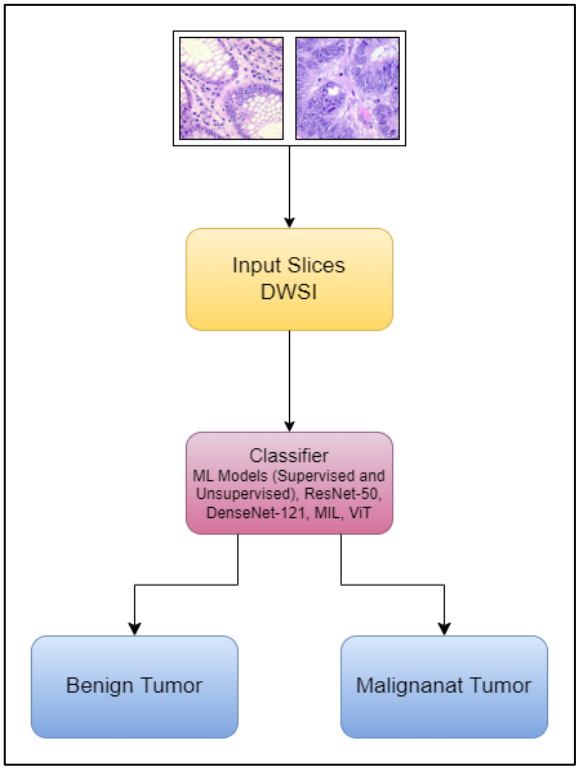



*Fig4.1 Project Flow Chart*

## 4.3 Working

### 4.3.1 Data Acquisition

The acquisition and curation of the Colon and Lung Cancer Histopathological Images dataset (LC25000), a pivotal resource comprising 25,000 histopathological images categorized into five distinct classes. The dataset originates from meticulously selected sources, ensuring adherence to HIPAA-compliant standards and rigorous validation protocols. Within this dataset, lung tissue samples are represented by 750 images, encompassing 250 instances of benign lung tissue, 250 lung adenocarcinomas, and 250 lung squamous cell carcinomas. Similarly, colon tissue samples comprise 500 images, comprising 250 instances of benign colon tissue and 250 colon adenocarcinomas. To bolster dataset diversity and fortify the resilience of our algorithms against potential biases, we meticulously employed data augmentation techniques utilizing the Augmenter package. Through this augmentation process, the initial dataset was expanded, resulting in a comprehensive collection of 25,000 images. This augmentation not only enriches the dataset with variations in texture, shape, and other key features but also ensures robustness and generalization of our classification models across diverse instances encountered in real-world scenarios. The acquisition and preparation of this dataset lay the cornerstone of our research endeavors, facilitating rigorous experimentation and evaluation of our proposed methodologies. By leveraging this meticulously curated dataset, we aim to unravel intricate patterns, advance our understanding of lung and colon cancer pathology, and ultimately contribute to the development of more accurate and reliable diagnostic tools in the fight against these formidable malignancies.

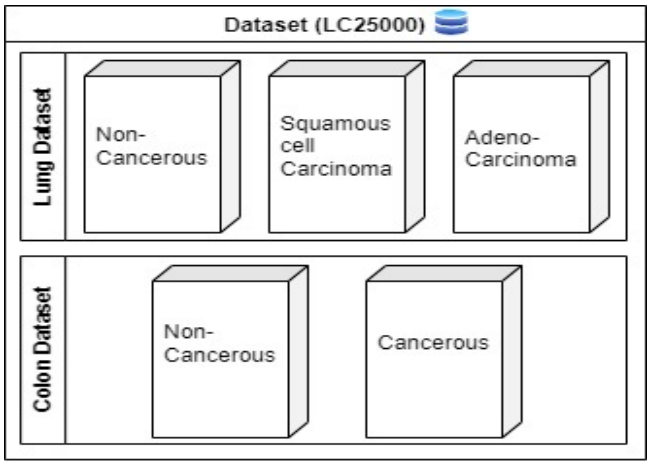



*Fig 4.2 Dataset Details*

### 4.3.2 Feature Extraction

Feature extraction is a critical step in image analysis and ML, where raw image data is transformed into a more compact and informative representation. This process is essential for reducing dimensionality, enhancing computational efficiency, and extracting discriminative characteristics for subsequent classification or regression tasks. Feature extraction methods, such as LBP, RGB and HOG features, aim to capture various aspects of the image's texture, color, and shape properties. By extracting relevant features, the complexity of the data is reduced, enabling more efficient and effective analysis and interpretation by ML algorithms. These methods involve mathematical computations and algorithms tailored to extract meaningful patterns and structures from the image data, ultimately facilitating tasks like object detection, image classification, and pattern recognition.

1) **Local Binary Patterns**: LBP is a widely-used texture descriptor in image analysis, particularly effective for tasks like texture classification and object recognition. It operates on the principle of quantifying the local texture patterns within an image by examining the relationship between a central pixel and its neighboring pixels. The LBP operator computes a binary code for each pixel by thresholding the intensity values of its surrounding pixels with respect to its own intensity value. The binary pattern is then converted into a decimal value, which represents the local texture pattern around that pixel. Mathematically, for a given central pixel in an image with P sampling points arranged on a circle of radius R, the LBP value is computed as:

$$LBP_P, R\ (x_c, y_c) = \sum_p = O_P - 1_s(I_p - I_c)2_p$$

Example: Binar y patterns = 00011111, and LBP code = 16+8+4+2+1 = 3

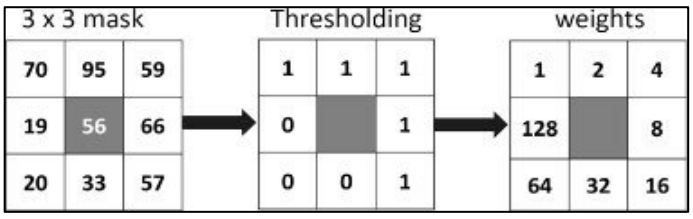



*Fig 4.3 Local Binary Pattern Matrix calculation block*

2) **Histogram of Oriented Gradients:** The HOG is a feature descriptor commonly employed for object detection and image classification tasks. It captures the local gradient information within an image, thereby encoding its underlying shape and structure. The computation of HOG involves the following steps:

Computing image gradients in both the horizontal and vertical directions using operators like

Sobel filters. Calculating the magnitude and orientation of gradients at each pixel. Dividing the image into spatial cells and constructing histograms of gradient orientations within each cell. Optionally, normalizing these histograms to enhance robustness against variations in illumination and contrast. Mathematically, the HOG descriptor encapsulates the spatial distribution of gradient orientations, providing a concise yet informative representation of the image's texture and shape characteristics. This representation is particularly effective for distinguishing between different objects or patterns in images.

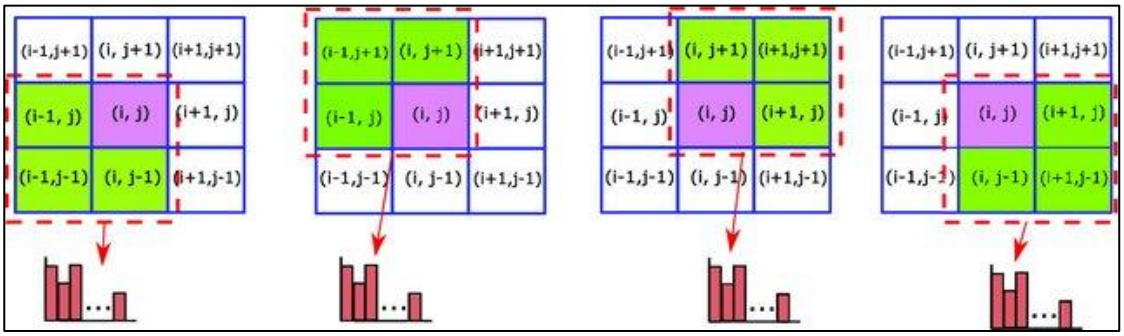



*Fig 4.4 HOG feature calculation block*

3) **Color Histogram:** RGB features involve extracting statistical measures from the red, green, and blue color channels of an image. These measures capture various aspects of color intensity and distribution, offering valuable insights into the image's visual content. Common statistical measures include:

Mean: Average intensity value within each color channel. Standard Deviation: Measure of the spread of intensity values. Skewness: Measure of the asymmetry of the intensity distribution. Kurtosis: Measure of the peak of the intensity distribution. By computing these statistics for each color channel, a set of RGB features is obtained, which can effectively characterize the color properties of the image. These features are often used in conjunction with other feature extraction methods to provide a comprehensive representation for subsequent ML tasks.

*Fig 4.5 Color Histogram explanation of the figure*

4) **ResNet-50:** ResNet50 is a deep CNN architecture composed of 50 layers. It is known for its skip connections, or residual connections, which enable the network to efficiently train very deep models. The skip connections allow the gradient flow to propagate effectively, addressing the vanishing gradient problem. ResNet50 has demonstrated exceptional performance in various image recognition tasks, including colon and lung cancer classification, by effectively capturing intricate features and patterns within medical images.

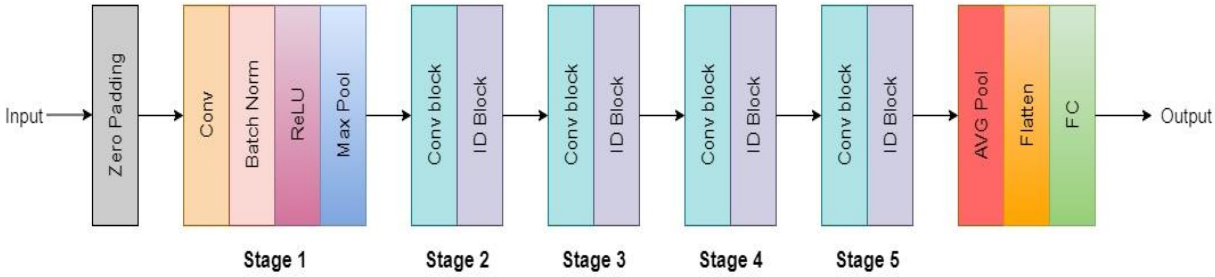



*Fig 4.6 ResNet-50 Architecture*

5) **DenseNet-121:** It begins with an initial convolution layer and pooling to process input images. The distinctive feature of DenseNet-121 lies in its dense blocks, where each layer is connected to every other layer, fostering rich feature interactions. Transition blocks, featuring convolution and pooling layers, facilitate the transition between dense blocks. The network incorporates global average pooling after the last dense block, ensuring a fixed-size output. A fully connected layer and Soft Max activation are then employed to produce class probabilities. DenseNet-121's dense connectivity promotes parameter efficiency, training stability, and superior performance in image classification.

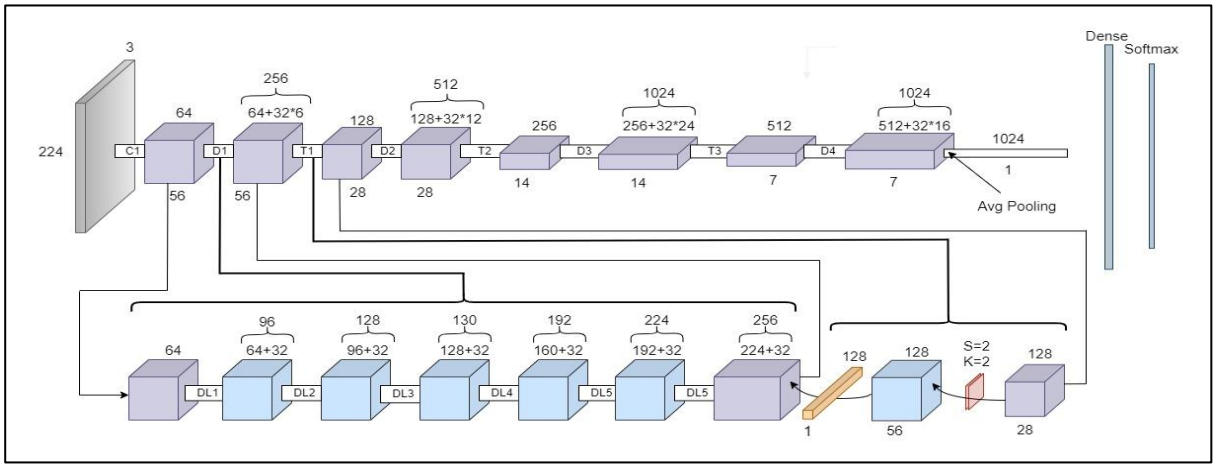



*Fig 4.7 DenseNet-121 Architecture*

### 4.3.3 Classification

1) **Logistic Regression:** Logistic regression is a widely-used statistical method for binary classification tasks, where the goal is to predict the probability that an instance belongs to a particular class. It models the relationship between the independent variables (features) and the binary outcome using the logistic function, also known as the sigmoid function. The logistic function ($z$)is defined as:

$$g(z)=1+e^{-z}{}_1$$

where, z is a linear combination of the feature values and model coefficients.

Mathematically, for a binary classification problem with features $x_1$, $x_2$,...$x_n$ and model coefficients

$\beta_0$, $\beta_1$, $\beta_2$…$\beta_n$ the logistic regression model can be expressed as:

$$p_{(y=1|x)} =1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_n x_n)1}$$

where, p(y=1|x) represents the probability that the outcome variable y is equal to 1 given the features x. This probability is then used to make predictions, typically by applying a threshold (e.g., 0.5) to classify instances into the two classes.

Logistic regression is suitable for problems with a binary response variable, such as classifying whether an image represents colon cancer or not. It's effective when the relationship between the features and the outcome is approximately linear and when the classes are well-separated in feature space. However, logistic regression may not be the best choice for multi-class classification tasks like lung cancer classification, where there are more than two classes. Although logistic regression can be extended to handle multi-class problems using techniques like one-vs-rest or multinomial logistic regression.

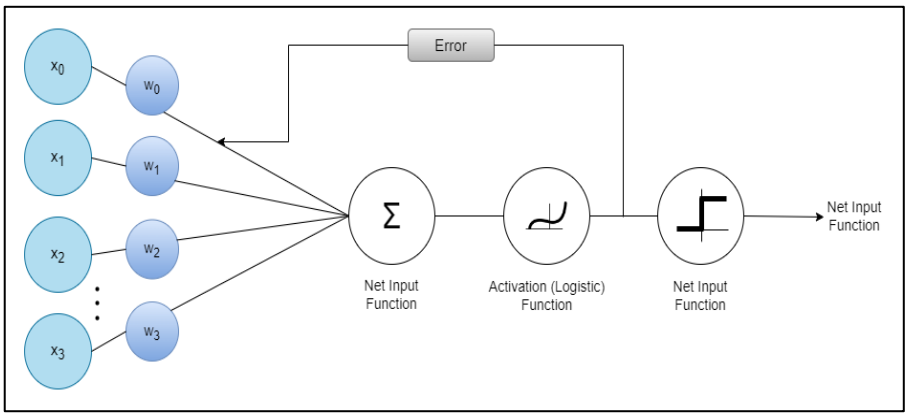



*Fig 4.8 Logistic Regression block diagram*

2) **Random Forest:** Random Forest is an ensemble machine learning algorithm that excels in both classification and regression tasks. It operates by constructing a multitude of decision trees during training and outputs the mode (for classification) or mean (for regression) prediction of the individual trees.

- The algorithm's strength lies in its ability to handle complex datasets with diverse variable types, large feature sets, and high dimensionality.
- Random Forest's ensemble approach enhances generalization and improves accuracy compared to individual decision trees. The method is less prone to overfitting and exhibits resilience to outliers, making it a popular choice in various domains, including finance, healthcare, and image analysis.
- Its versatility, scalability, and capacity to handle diverse data characteristics contribute to its widespread application in real-world machine learning scenarios.

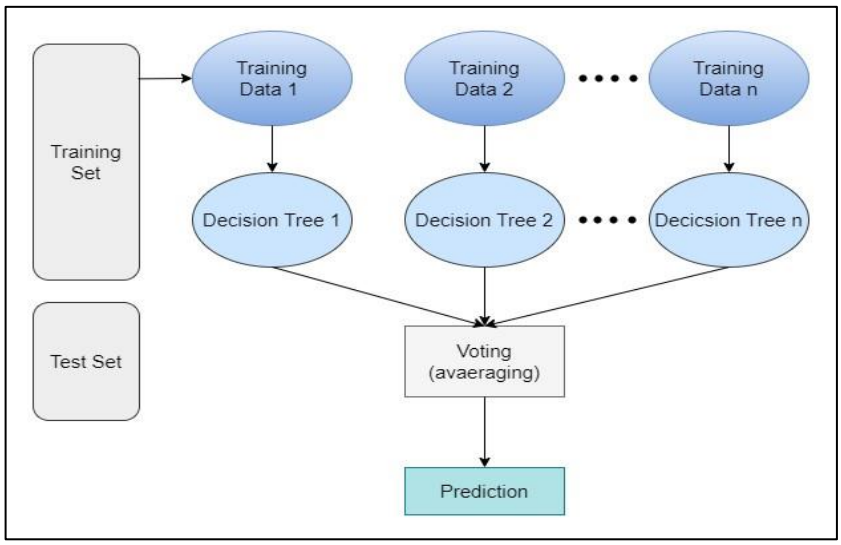



*Fig 4.9 Random Forest block diagram*

3) **SVM:** SVM finds the hyper plane that best separates different classes in the feature space. The hyper plane is chosen to maximize the margin, the distance between the hyper plane and the nearest data points of each class.

- The algorithm's strength lies in its ability to handle complex datasets with diverse variable types, large feature sets, and high dimensionality.
- SVM can efficiently handle non-linear relationships through the kernel trick. This involves transforming the input features into a higher-dimensional space, making complex patterns more discernible.
- SVMs come in different types, such as linear SVM for linearly separable data, and

non-linear SVM using kernels to handle more complex patterns.

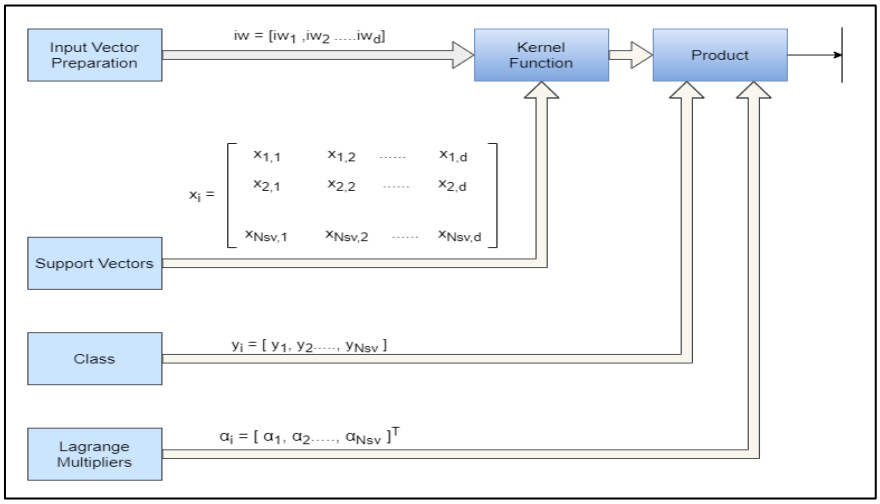



*Fig4.10: Support Vector Machine block diagram*

**Support Vector Machine (SVM) Kernels:**

a) Linear Kernel:

$$K(p_i, p) = p_i \cdot p$$

The linear kernel is effective for distinguishing between cancerous and non-cancerous colon tissues based on linearly separable features. It is useful in separating lung cancer data into distinct classes when features exhibit a linear separation pattern.

b) Polynomial Kernel:

$$K(p_i, p) = (p_i \cdot p + a)^b$$

where; *a* is a constant and *b* is the degree of the polynomial.

The polynomial kernel captures non-linear relationships in colon cancer datasets, accommodating cases where the decision boundary is more complex. It is also useful for lung cancer classification, capturing complex relationships in the feature space and accommodating non-linear patterns.

c) Radial Basis Function (RBF) or Gaussian Kernel:

$$K(p_i, p) = \exp(-2c2\|p_i - p\|2)$$

where; *c* is the kernel width parameter.

The RBF or Gaussian kernel is well-suited for capturing intricate relationships between

features in the high-dimensional space. It is employed when colon cancer data exhibits non-linear patterns and is widely used for lung cancer classification, particularly in cases with complex and non-linear relationships between features, enabling discernment of subtle patterns.

d) Sigmoid Kernel:

$$K(x, y) = \tanh(a \cdot x_T y + b)$$

where; $a$ is a scaling factor, and $b$ is a constant.

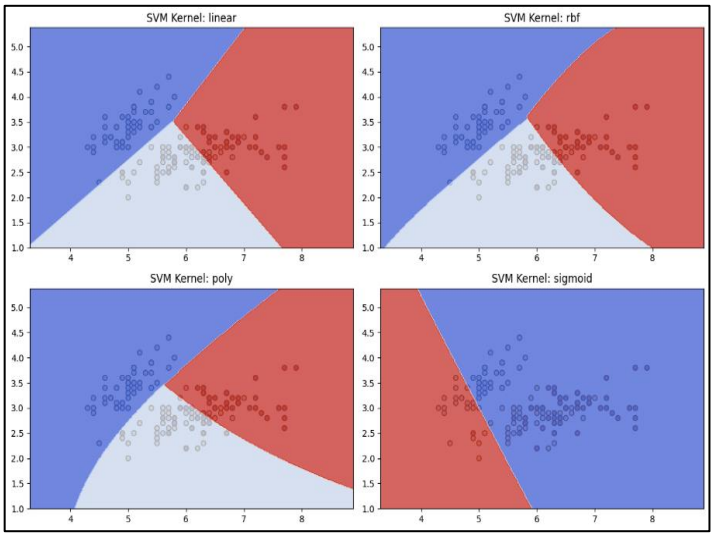



*Fig 4.11 SVM Kernels graph*

4) **Artificial Neural Network**: An ANN is a computational model inspired by the structure and function of biological neural networks in the human brain. It consists of interconnected nodes organized in layers: an input layer, one or more hidden layers, and an output layer. Each node, or neuron, in the network processes input data and passes it through an activation function to produce an output signal. During training, the network adjusts its weights and biases using optimization algorithms like gradient descent to minimize a loss function, thereby learning to map input data to the correct output. ANN's ability to learn complex patterns and relationships makes it a powerful tool for tasks such as classification, regression, and pattern recognition in various domains including image and speech recognition, natural language processing, and medical diagnosis

- ANN algorithms can be trained on features extracted from cancer cell images using techniques like HOG, LBP, or CNNs. These features capture important characteristics of the

cell images, such as shape, texture, and intensity, which are indicative of cancerous or non-cancerous cells.

- ANN models excel at recognizing complex patterns and relationships within data. By training on a dataset of labeled cancer cell images, the network learns to identify subtle differences between cancerous and non-cancerous cells based on the extracted features. This enables the model to accurately classify new, unseen cell images as either cancerous or non-cancerous with high precision.

- ANN algorithms are highly adaptable and capable of generalizing from training data to make predictions on unseen data. As a result, they can effectively handle variations in cancer cell images due to factors like different imaging techniques, magnifications, or lighting conditions. This adaptability ensures robust performance in real-world scenarios where images may vary widely in quality and appearance.

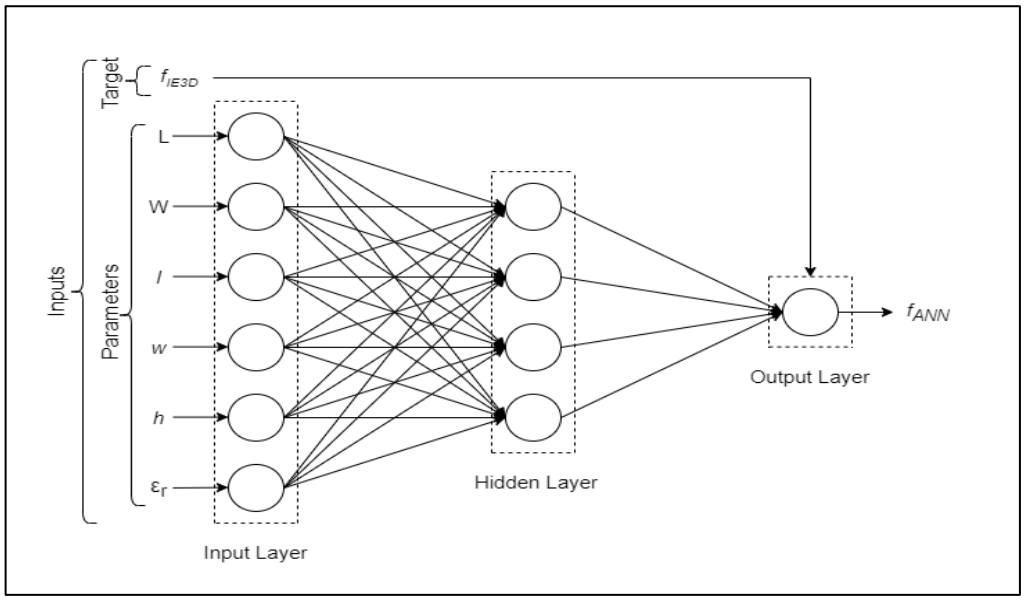



*Fig4.12: Artificial Neural Network block diagram*

5) **K-means Clustering:** It is a popular unsupervised ML algorithm used for partitioning a dataset into a predetermined number of clusters.

Here's how it works:

- Initialization: First, you need to specify the number of clusters (K) you want to identify in your dataset. Then, K initial cancroids are randomly selected from the data points. These cancroids represent the initial guesses for the cluster centers.

- Assignment Step: Each data point is assigned to the nearest centroid, based on a distance metric such as Euclidean distance. This step partitions the data into K clusters.

- Update Step: After the assignment of data points to clusters, the centroids are updated by computing the mean of all data points assigned to each cluster. These means become the new centroids.

- Repeat Assignment and Update Steps: Steps 2 and 3 are repeated iteratively until the centroids no longer change significantly or a specified number of iterations is reached. Typically, convergence is achieved when the centroids move minimally between iterations.

- Final Clustering: Once the algorithm converges, each data point is assigned to the cluster whose centroid is nearest to it. The algorithm has clustered the data based on similarity, with each cluster represented by its centroid.

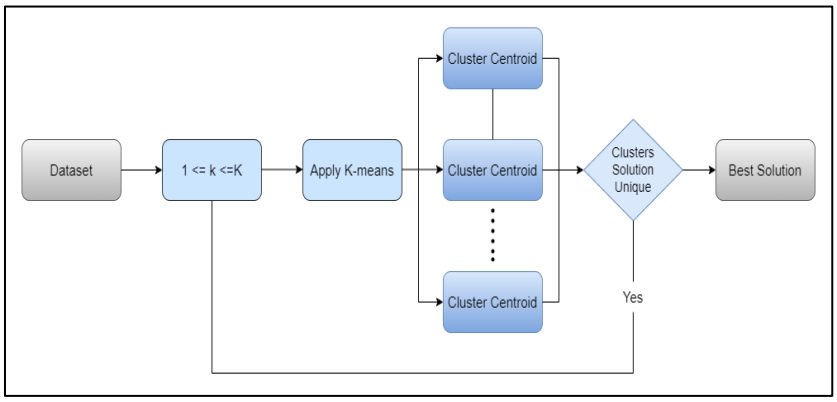



*Fig 4.13 K-means Clustering block diagram*

6) **Principle Component Analysis:** It is a widely used technique in ML and data analysis for dimensionality reduction. It works by transforming the original variables into a new set of variables, called principal components, which are linear combinations of the original variables. PCA aims to find the directions (or axes) in the feature space along which the data varies the most.

Here's a step-by-step explanation of how PCA works:

- Standardization
- Compute the Covariance Matrix
- Eigenvalue Decomposition
- Select Principal Components
- Projection

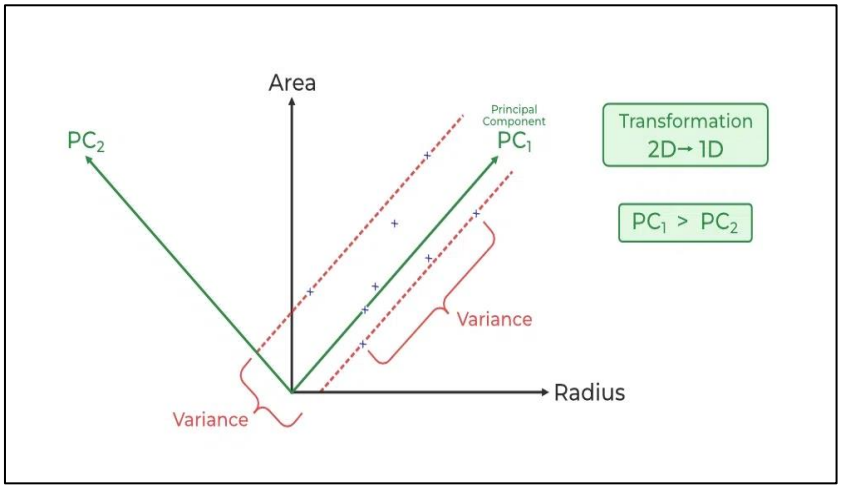



*Fig 4.14 Principal Component Analyses plot*

7) **Hybrid Model**: DWSI undergo feature extraction using a CNN architecture. The resulting 1-D feature vectors capture intricate patterns from the image. Subsequently, a SVM classifier is employed to show results with four different kernels—linear, polynomial, radial basis function (RBF), and sigmoid to discern between binary classes and multi-class classification. ResNet-50 extracts nuanced features from his ophthalmological images, creating comprehensive feature vector.

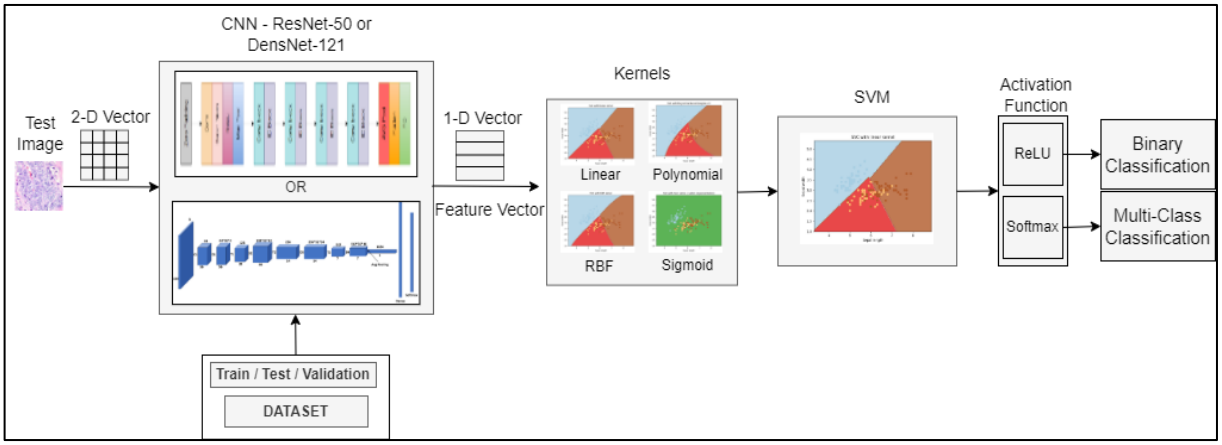



*Fig 4.15 Hybrid Model block diagram*

DenseNet-121, with fewer trainable parameters and dense connectivity, enhances feature propagation. We incorporate kernel functions into the classification process, employing SVM classification to discriminate between benign and malignant regions in lung, and colon cancer samples. The choice of kernel functions—linear, polynomial, sigmoid, and RBF significantly influences the model's ability to generalize and classify unseen data. Further, this paper unveils the significance of hybrid models and remarkable results achieved on varying the kernels.

8) **Multiple Instance Learning**: MIL is a ML paradigm where the training data is organized into bags, each containing multiple instances. Unlike traditional supervised learning, where each instance is associated with a single label, MIL assumes that the labels are assigned to bags rather than individual instances. This setup is particularly useful in scenarios where obtaining precise instance-level labels is challenging or expensive, such as in medical image analysis.

A bag is represented as $B = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, where $x_i$ represents individual image instances and yi represents their associated labels. Each $y_i \in \{0, 1\}$, where 1 indicates a positive instance (cancerous region) and 0 indicates a negative instance (non-cancerous region). The bag label, denoted by c(B), is determined based on the presence or absence of positive instances within the bag. This is calculated using the equation:

$$c(B) = 1 - \Pi_{(i=1 \text{ to } n)} (1 - y_i)$$

Here's how the equation works:

- $(1-y_i)$: For each individual instance, subtract its label from 1. If $y_i = 1$ (positive instance), then $(1 - y_i) = 0$, indicating the absence of a negative instance. If $y_i = 0$ (negative instance), then $(1 - y_i) = 1$, indicating the absence of a positive instance.
- $\prod_{(i=1 \text{ to } n)} (1 - y_i)$: Calculate the product of $(1 - y_i)$ for all instances in the bag. This product represents the probability of all instances being negative within the bag.
- $1 - \prod_{(i=1 \text{ to } n)} (1 - y_i)$: Subtract the product obtained in step 2 from 1. This yields the probability of at least one positive instance being present in the bag, thus determining the bag label.

By analyzing the presence or absence of cancerous regions within the individual patches, MIL determines whether a WSI contains cancer (positive bag) or is free from cancer (negative bag). The bag label is determined based on the MIL formulation, which ensures that if at least one patch within the bag is labeled as cancerous, the entire WSI is classified as cancerous. This approach enables the algorithm to effectively classify cancer and non-cancerous regions within WSI, contributing to accurate diagnosis and treatment planning for lung and colon cancers.

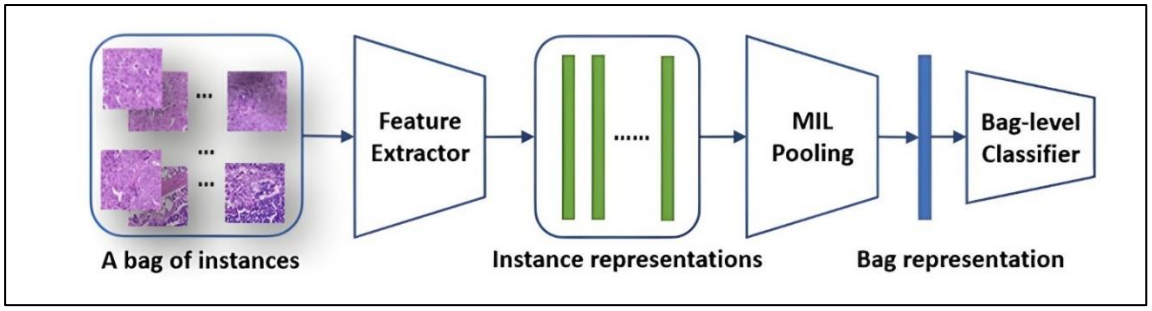



*Fig 4.16 Multiple Instance Learning block diagram*

9) **Vision Transformer**: ViT adapt transformer architectures from Natural Language Processing (NLP) to process images by preprocessing them into patches akin to word tokens. These patches undergo transformation in transformer encoder blocks, comprising Layer Normalization for computational efficiency, Multi-head Attention to capture local and global dependencies, and Multi-Layer Perception (MLP) with Gaussian Error Linear Unit activation for feature extraction. In binary classification with the LC25000 dataset, Vi T underwent training to differentiate between colon a den carcinomas and benign colonic tissues. Initial preprocessing standardized images to 224x224 pixels and normalized pixel values. These processed images were then input into ViT models, traversing transformer blocks to capture relevant features, with outputs directed through a classification head, possibly utilizing a Soft Max function, for probability generation. Training optimized parameters using categorical cross-entropy loss to minimize disparities between predictions and labels. Similar methodology was applied to distinguish lung tissues: lung adenoid carcinomas, lung squalors cell carcinomas, and benign lung tissues. This approach empowered ViT models to discern tissue-specific features, facilitating precise classification.

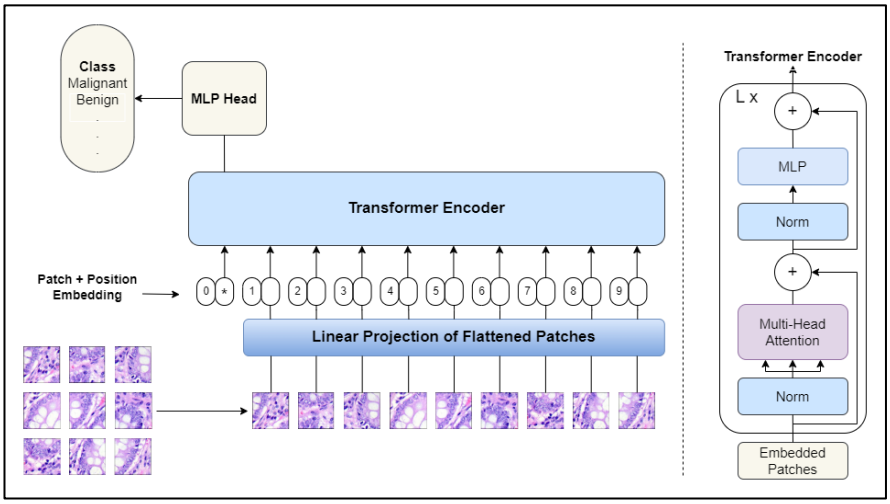



*Fig 4.17 Vision Transformer block diagram*

### 4.3.4 Evaluation and Validation

The performance of the classification system is evaluated using various metrics such as confusion matrix, accuracy, precision, recall, F1-score. Validation studies may involve comparing the system's results with ground truth annotations by experts or conducting clinical trials to assess its effectiveness in real-world scenarios.

Here's what each metric means:

**Confusion matrix:** The confusion matrix is a matrix used to determine the performance of the classification models for a given set of test data. Since it shows the errors in the model performance in the form of a matrix, hence also known as an error matrix. For the 2 prediction classes of classifiers, the matrix is of 2x2 table, for 3 classes, it is 3x3 table, and so on.

The matrix is divided into two dimensions that are predicted values and actual values along with the total number of predictions. Predicted values are those values, which are predicted by the model, and actual values are the true values for the given observations.

**True Negative:** Model has given prediction No, and there is actual value was also No.

**True Positive:** The model has predicted yes, and the actual value was also true.

**False Negative:** The model has predicted no, but the actual value was Yes, it is also called as Type-II error.

**False Positive:** The model has predicted Yes, but the actual value was No. It is also called a Type-I error.

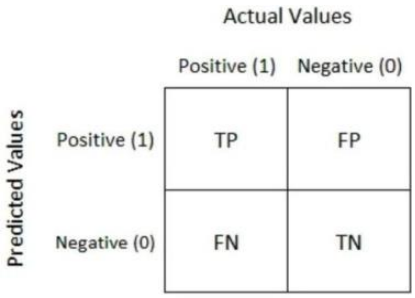



*Fig 4.18 Confusion Matrix values*

**Precision:** Precisionist the ratio of true positives to the total predicted positives for a class. It measures the accuracy of the positive predictions. A higher precision indicates fewer false positives.

$$\text{Precision} = \frac{TP}{TP+FP}$$

**Recall:** Recall is the ratio of true positives to the total actual positives for a class. It measures the ability of the model to correctly identify positive instances. A higher recall indicates fewer false negatives.

$$\text{Recall} = \frac{TP}{TP+FN}$$

**Accuracy:** Accuracy is the overall classification accuracy, which is the ratio of correctly classified samples to the total number of samples.

$$\text{Accuracy} = \frac{TP}{TP+FP+FN+TN}$$

**F1-score:** The F1-score is the harmonic mean of precision and recall. It provides a single metric that balances both precision and recall. It is often used as a summary statistic when the class distribution is uneven or when both precision and recall are important.

# CHAPTER 5
# RESULTS AND DISCUSSION

## 5.1 Results

| Model Name | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| Logistic Regression | 71.5% | 75.82% | 66.34% | 70.76% |
| Random Forest | 87% | 86.79% | 88.4% | 87.6% |
| SVM | 91% | 88% | 95.1% | 85.34% |
| ANN | 84% | 86.41% | 83.18% | 84.76% |
| K-means clustering | 95% | 96.8% | 90.64% | 94.8% |
| PCA | 86% | 86.9% | 86% | 86.79% |

*Fig 5.1.1 Machine Learning model's output parameters of colon dataset*

| Model Name | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| Logistic Regression | 78% | 78.65% | 78% | 78.28% |
| Random Forest | 87.33% | 87.97% | 87.3% | 87.54% |
| SVM | 94.33% | 95.02% | 94.3% | 94.59% |
| ANN | 85.50% | 85.5% | 85.39% | 85.44% |
| K-means clustering | 95% | 95.33% | 94.66% | 94.55% |
| PCA | 80.66% | 85.73% | 80.66% | 81.3% |

*Fig 5.1.2 Machine Learning model's output parameters of Lung dataset*

| **Kernel** | **Accuracy** | **Precision** | **Recall** | **F1-score** |
|---|---|---|---|---|
| Linear | 78% | 78.65% | 78% | 78.28% |
| Polynomial | 87.33% | 87.97% | 87.3% | 87.54% |
| RBF | 94.33% | 95.02% | 94.3% | 94.59% |
| sigmoid | 85.50% | 85.5% | 85.39% | 85.44% |

*Fig 5.2.1 Hybrid Model output parameters for colon dataset (ResNet-50 + SVM)*

| Kernel | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| Linear | 78% | 78.65% | 78% | 78.28% |
| Polynomial | 87.33% | 87.97% | 87.3% | 87.54% |
| RBF | 94.33% | 95.02% | 94.3% | 94.59% |
| Sigmoid | 85.50% | 85.5% | 85.39% | 85.44% |

*Fig 5.2.2 Hybrid Model output parameters for Lung dataset (ResNet-50 + SVM)*

| **Kernel** | **Accuracy** | **Precision** | **Recall** | **F1-score** |
|---|---|---|---|---|
| Linear | 78% | 78.65% | 78% | 78.28% |
| Polynomial | 87.33% | 87.97% | 87.3% | 87.54% |
| RBF | 94.33% | 95.02% | 94.3% | 94.59% |
| Sigmoid | 85.50% | 85.5% | 85.39% | 85.44% |

*Fig 5.2.3 Hybrid Model output parameters for colon dataset (DenseNet-121 + SVM)*

| Kernel | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| Linear | 78% | 78.65% | 78% | 78.28% |
| polynomial | 87.33% | 87.97% | 87.3% | 87.54% |
| RBF | 94.33% | 95.02% | 94.3% | 94.59% |
| Sigmoid | 85.50% | 85.5% | 85.39% | 85.44% |

*Fig 5.2.4 Hybrid Model output parameters for Lung dataset (DenseNet-121 + SVM)*

| Dataset | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| Colon Data | 92.67% | 82.91% | 95.89% | 86.83% |
| Lung Data | 95% | 97.93% | 96.94% | 96.57% |

*Fig 5.3: Multiple Instance model's output parameters*

| Dataset | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| Colon Data | 92.67% | 82.91% | 95.89% | 86.83% |
| Lung Data | 95% | 97.93% | 96.94% | 96.57% |

*Fig 5.4: Vision Transformer model's output parameters*

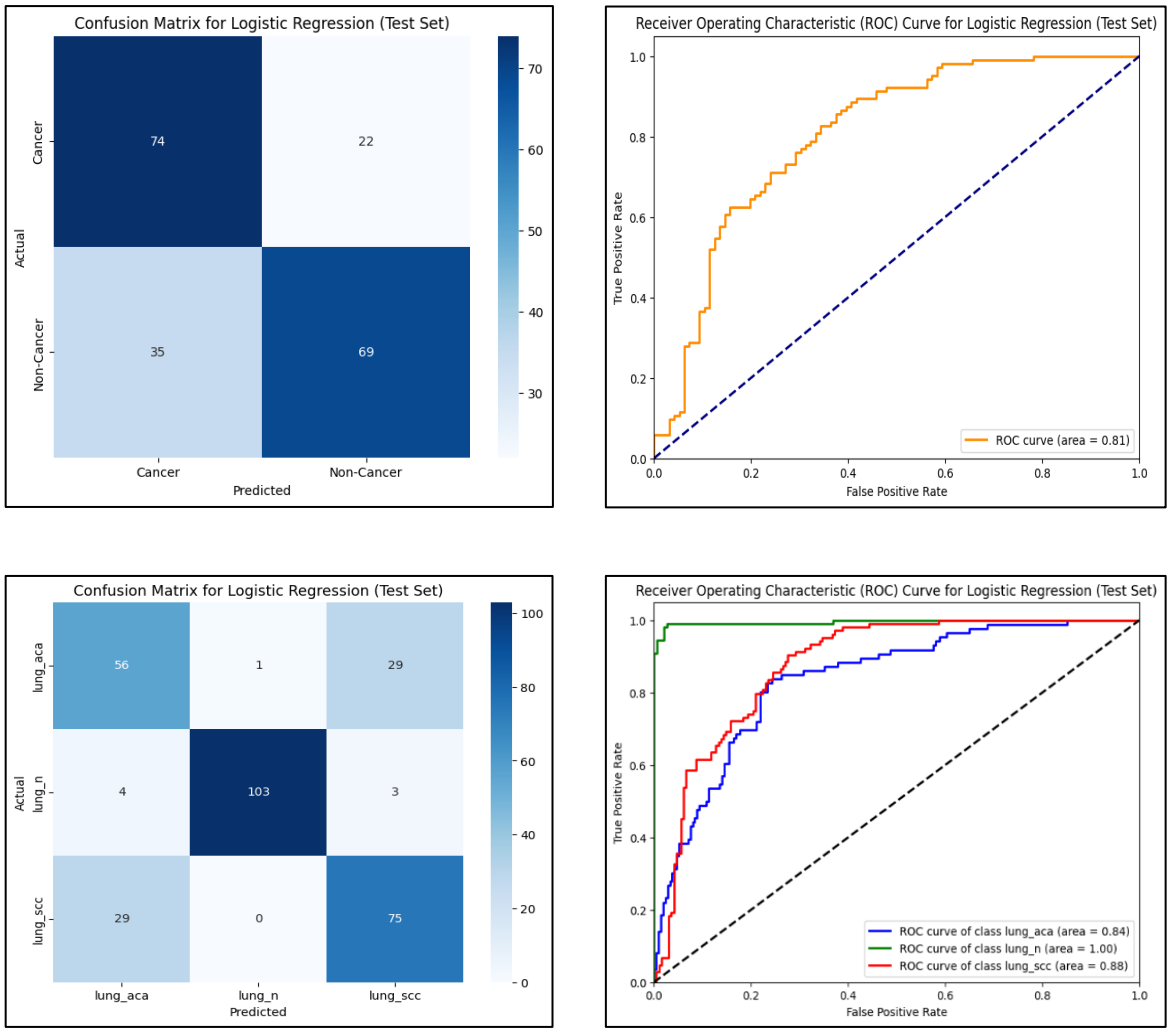
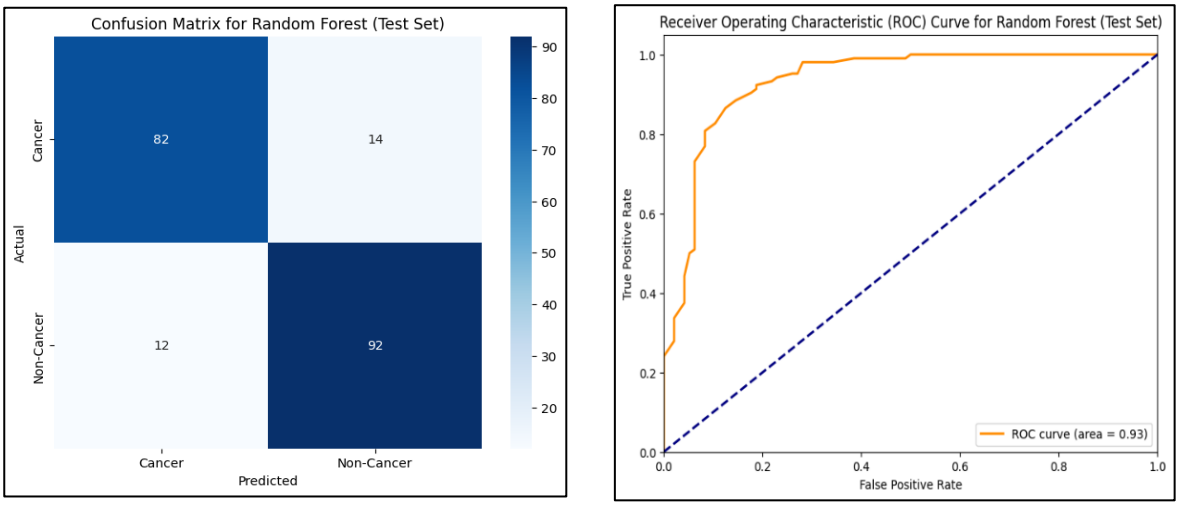
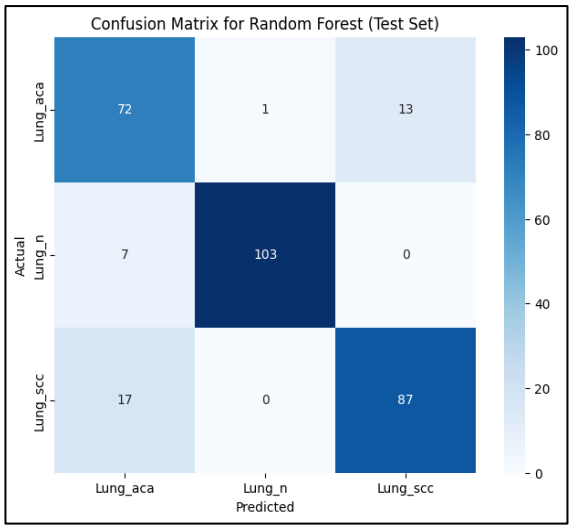



*Fig 5.5: Logistic Regression Result for colon and Lung dataset*





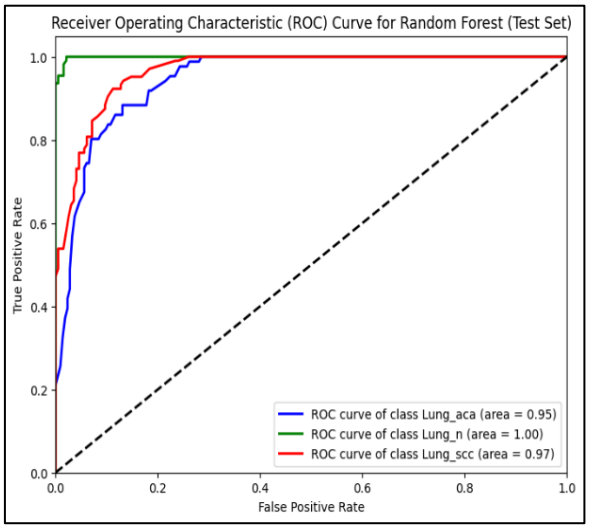
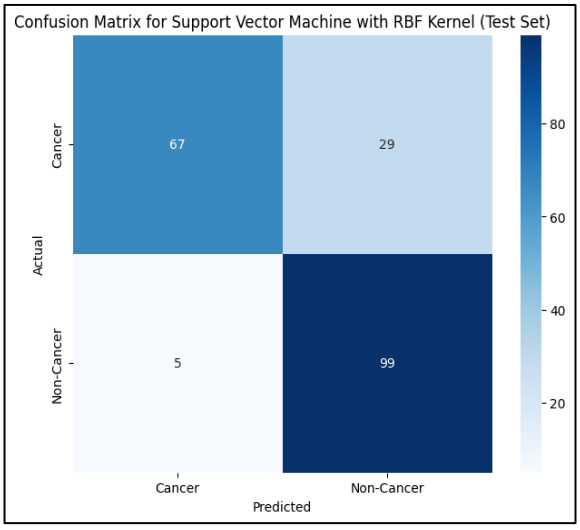
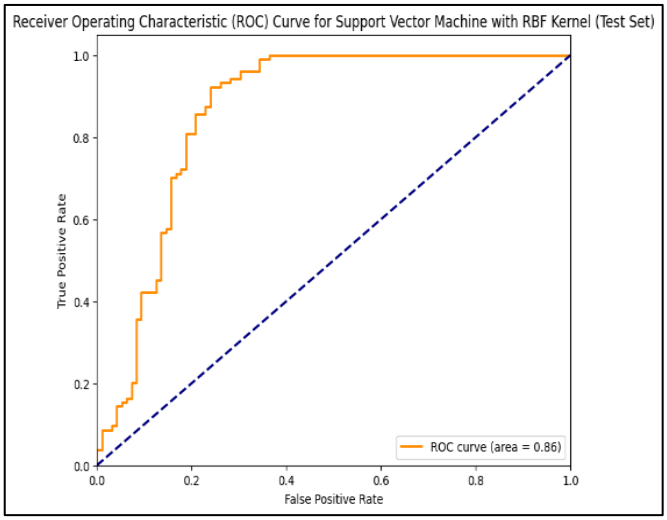
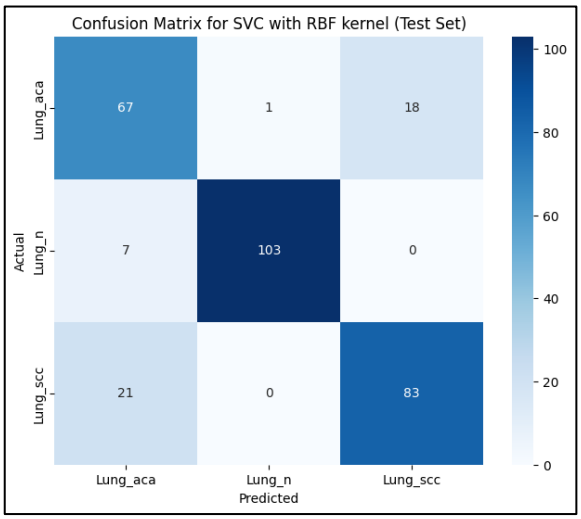




*Fig 5.6 Random Forest Result for colon and Lung dataset*



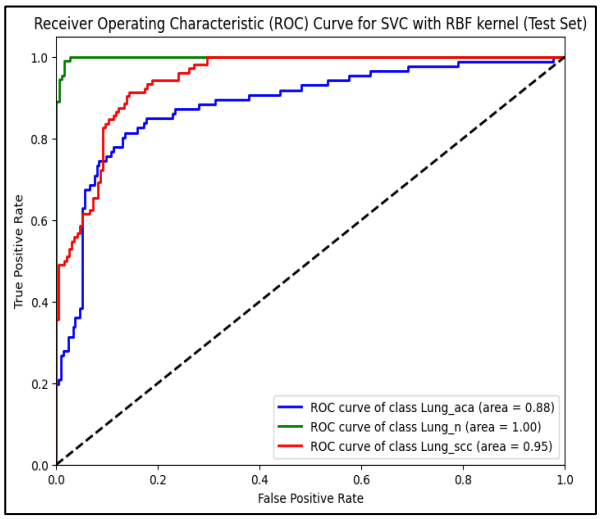




*Fig 5.7 Support Vector Machine (RBF Kernel) Result for colon and Lung dataset*

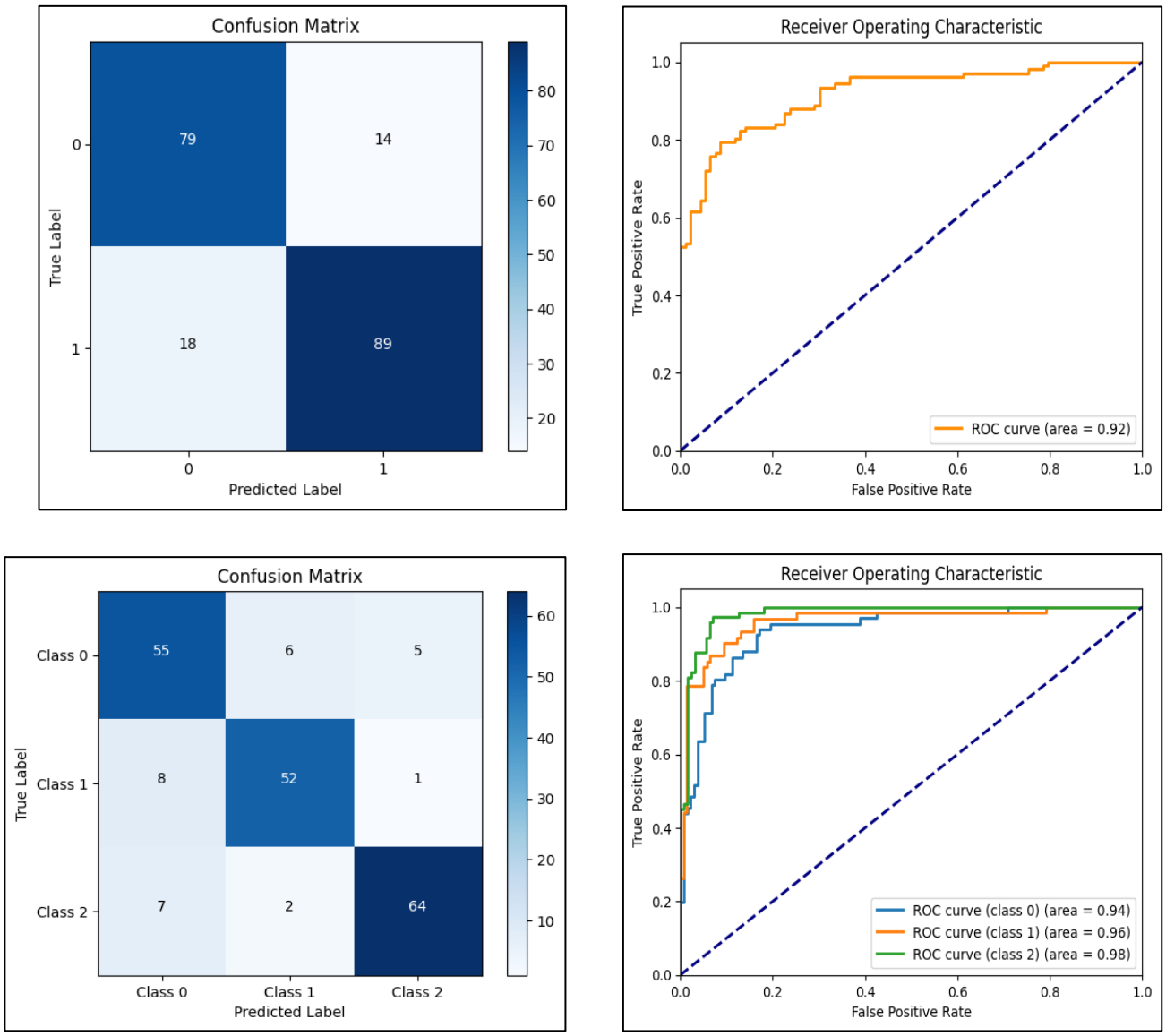



*Fig 5.8 ANN Result for colon and Lung dataset*

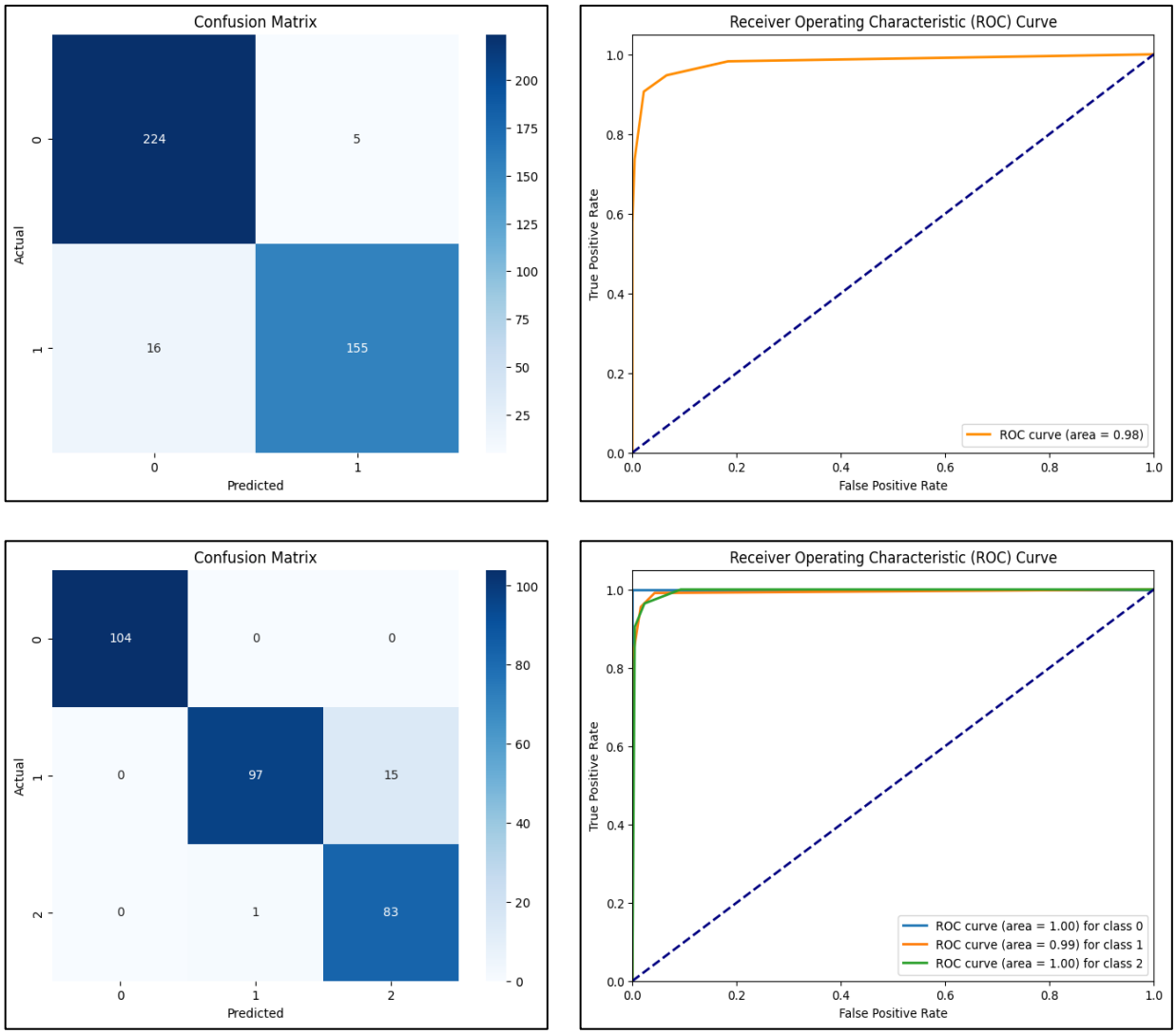



*Fig 5.9 K-means clustering Result for colon and Lung dataset*

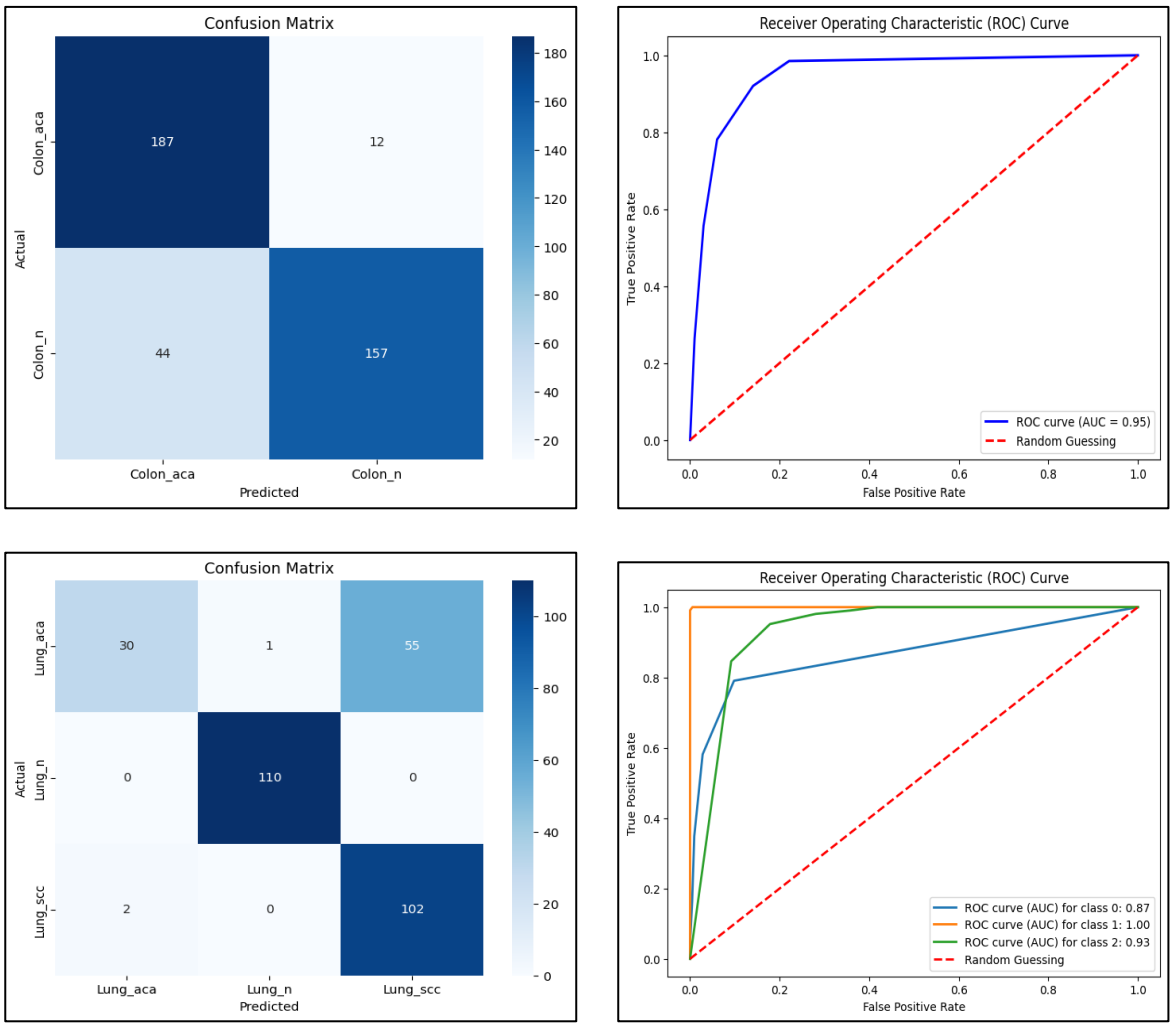



*Fig 5.10 Principal Component Analysis Result for colon and Lung dataset*

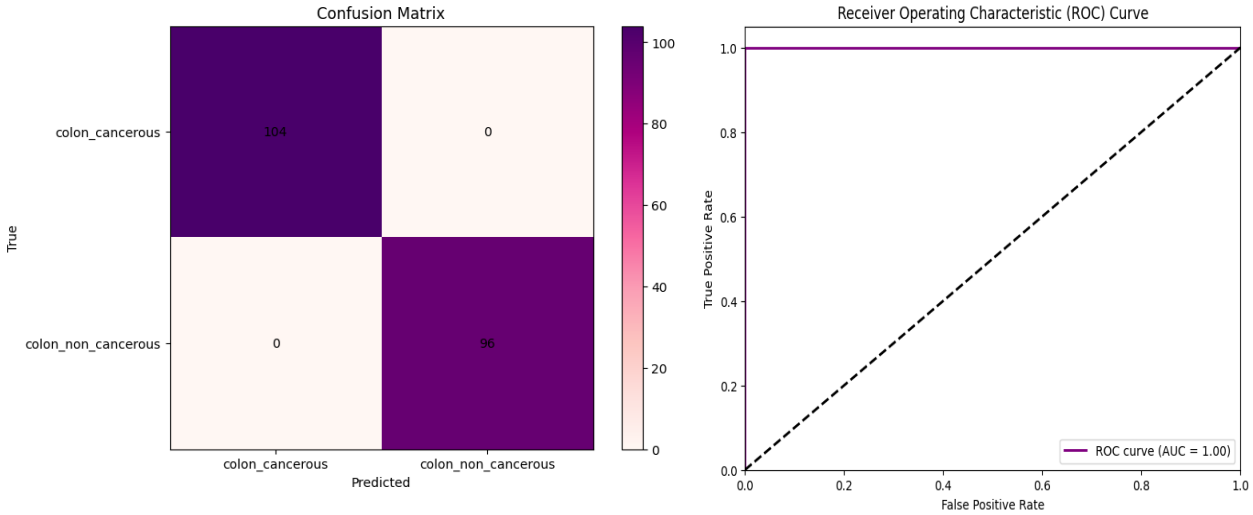



*Fig 5.11 ResNet-50 + SVM Results for colon dataset (Linear kernel)*

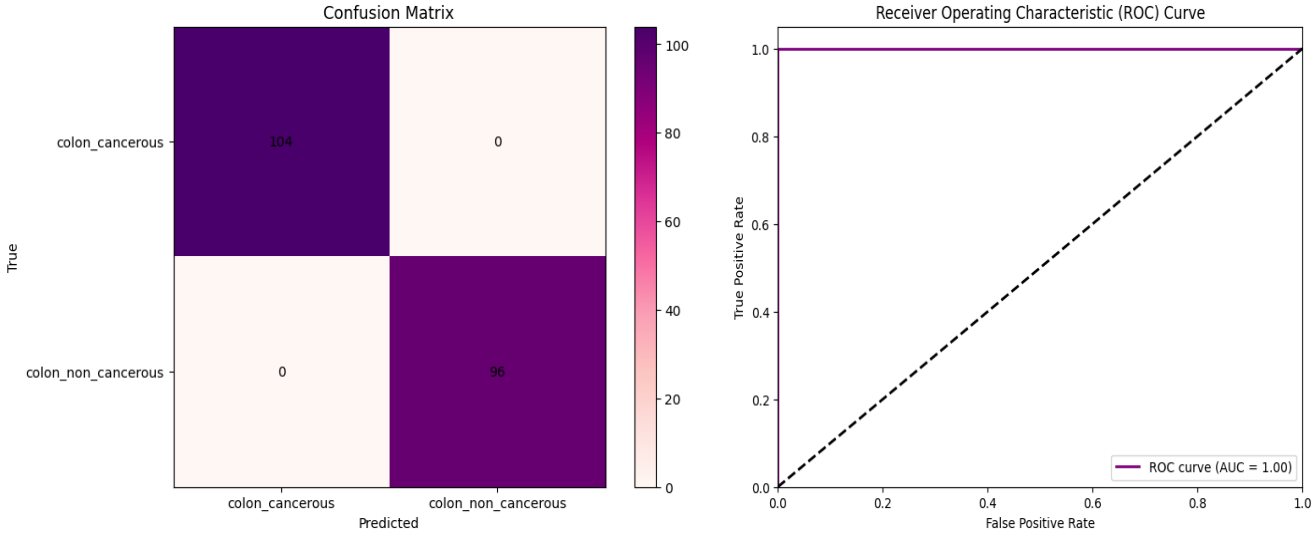



*Fig 5.12 DenseNet-121 + SVM Results for colon dataset (RBF kernel)*

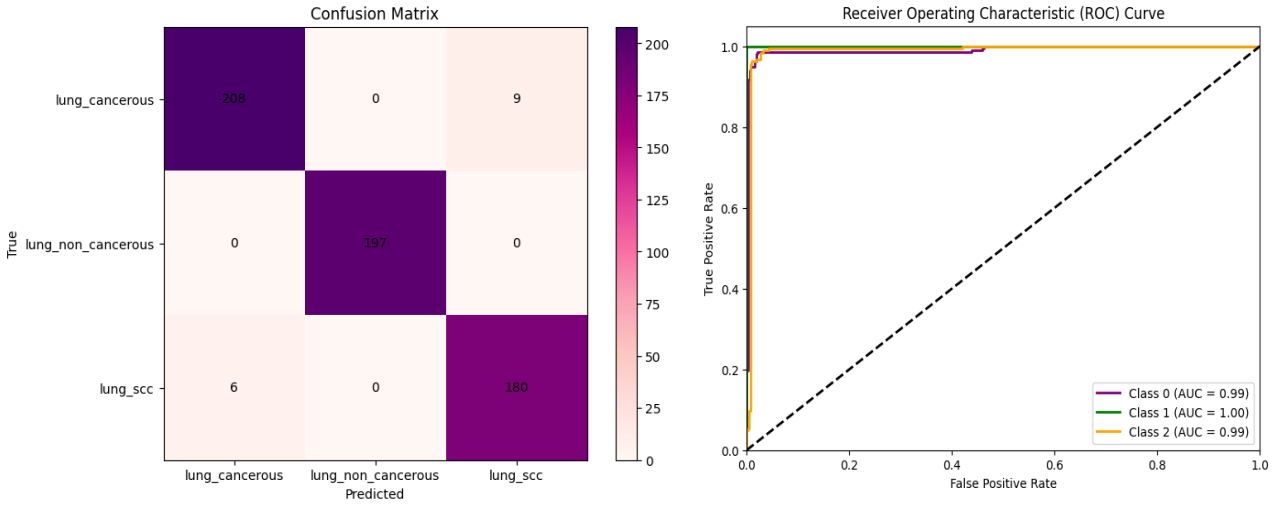



*Fig 5.13 ResNet-50 + SVM results for Lung dataset (RBF kernel)*

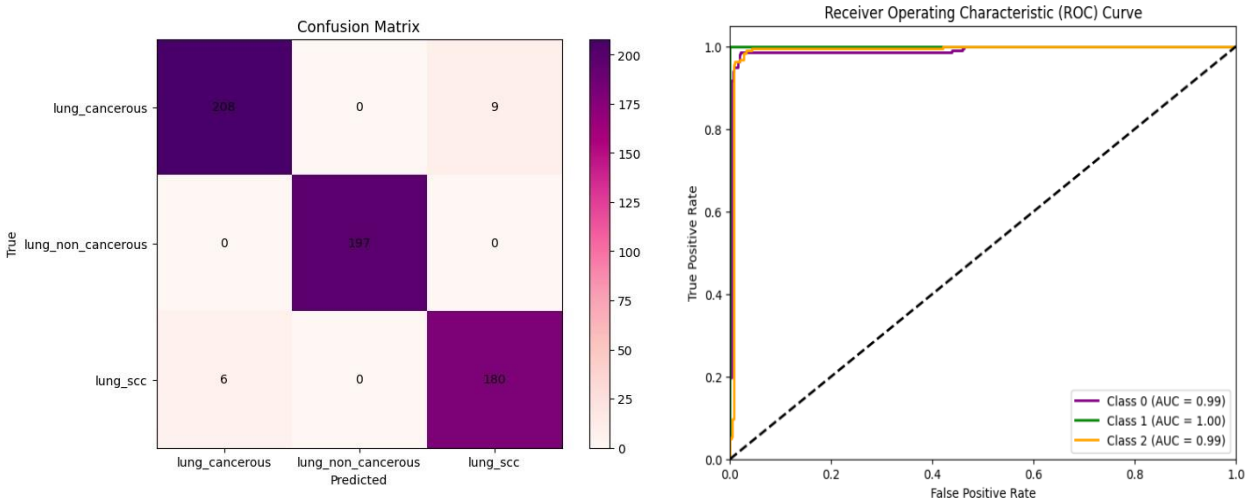



*Fig 5.14 DenseNet-121 + SVM results for lung dataset (poly kernel)*

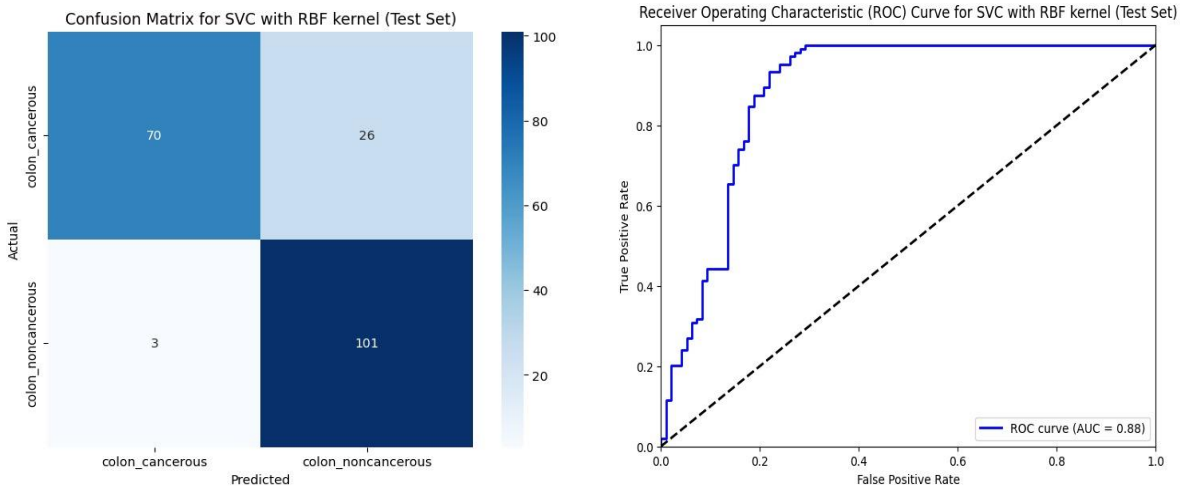



*Fig 5.15 MIL results for colon dataset*

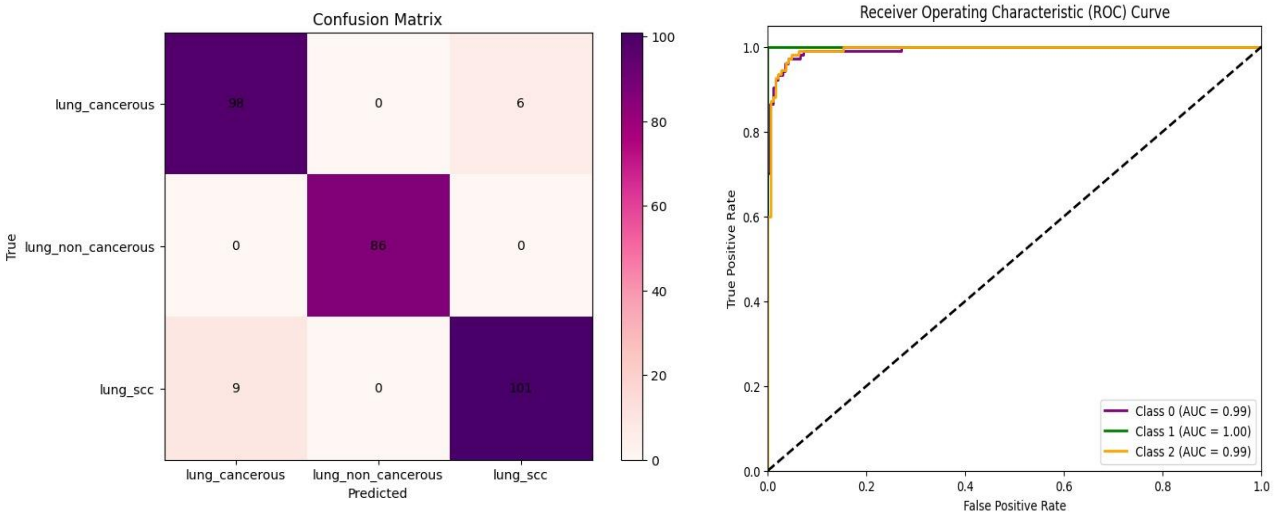



*Fig 5.16 MIL results for lung dataset*

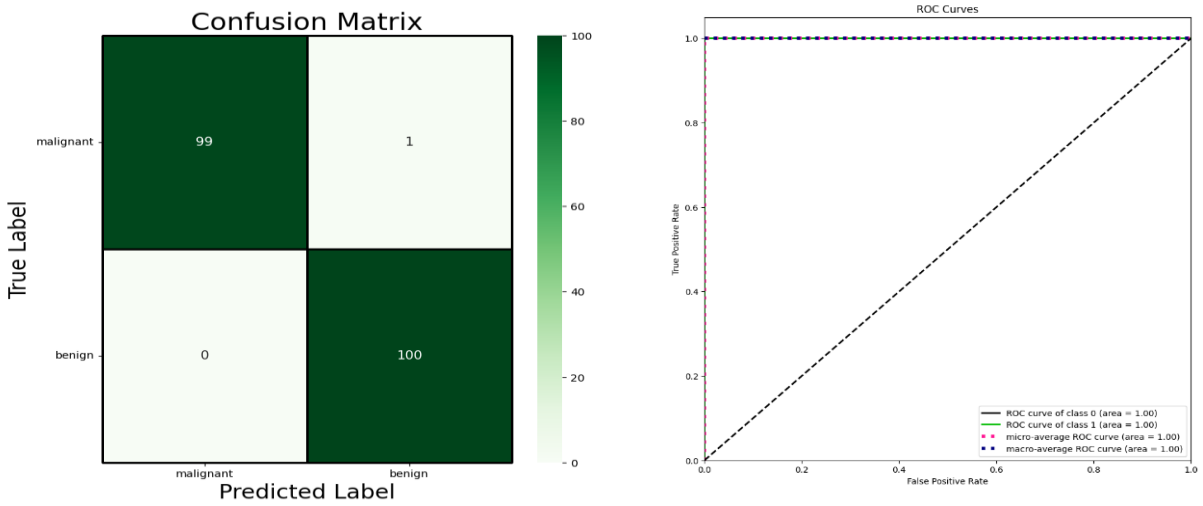



*Fig 5.17 Vision Transformer results for colon dataset*

## 5.2 Discussion

Our comprehensive approach integrating traditional machine learning (ML) algorithms with advanced deep learning techniques has yielded promising results in the classification of lung and colon cancers. Leveraging a hybrid model combining convolution neural network (CNN) features with support vector machine (SVM) classification, we achieved remarkable accuracy, with the SVM model attaining an impressive 91% accuracy in lung cancer classification and 94.33% accuracy in colon cancer classification. Furthermore, the incorporation of multiple instance learning (MIL) techniques and the innovative vision transformer (ViT) architecture further enhanced our classification outcomes, as evidenced by the ViT model's performance with 95% accuracy in lung cancer classification and 92.67% accuracy in colon cancer classification. These results demonstrate the potential of artificial intelligence (AI) in revolutionizing cancer diagnosis, offering improved precision and efficiency compared to traditional methods. By bridging the gap between pathology and radiology and leveraging interdisciplinary collaboration, our approach aims to significantly improve patient outcomes and advance the field of medical imaging analysis. However, challenges such as image variability, inter-observer variability in manual annotations, and the need for large annotated datasets for training pose ongoing research areas. Future advancements in this field will likely focus on refining algorithms, improving generalizability, and integrating multimodal imaging to further enhance the accuracy and reliability of prostate cancer detection from biomedical imagery.

# CHAPTER 6

# CONCLUSION AND SCOPE FOR FUTURE WORK

## 6.1 Conclusion

In conclusion, the application of artificial intelligence (AI) in Colon and Lung cancer detection, specifically through the feature extraction and classification of biomedical imagery, holds immense potential and a promising future. AI algorithms have demonstrated the ability to accurately analyze medical images, such as WSI by extracting features and identifying colon and lung cancer with precision. This feature extraction process allows for a detailed understanding of cancer characteristics, aiding in personalized treatment planning and monitoring. Moreover, AI-based classification models can effectively distinguish between benign and malignant colon and lung tissue's, supporting accurate diagnosis and reducing unnecessary invasive procedures.

## 6.2 Scope for Future Work

Looking ahead, the future scope for AI in colon and lung cancer detection through feature extraction and classification is highly promising. Continued advancements in AI technologies, including machine learning and convolution neural networks, can enhance the accuracy and efficiency of these algorithms. This can lead to more robust and reliable colon and lung cancer detection, enabling early diagnosis and intervention for improved patient outcomes. Furthermore, the integration of AI with multimodal data, such as genomics and clinical records, can provide a comprehensive and holistic approach to colon and lung cancer detection. By combining various data sources, AI algorithms can gain a deeper understanding of the disease, identify relevant biomarkers, and develop personalized treatment strategies tailored to individual patients. In summary, the application of AI in colon and lung cancer detection through feature extraction and classification of biomedical imagery offers significant potential. With further research and development, AI has the potential to revolutionize colon and lung cancer detection, contributing to earlier diagnosis, personalized interventions, and improved patient outcomes.

# REFERENCES


[1] Tinaa Babu, Triptya Singh, Shahin Gupta, Deepaa Hameed. "Colon cancer prediction on histological images using deep learning features and Bayesian optimized SVM." Journal of Intelligent Fuzzy Systems, 2021.

[2] Zarrin Tasnim, Sovon Chakraborty, F. M. Javed Mehedi Shamrat, Ali Newaz Chowdhury, Humaira Alam Nuha, Asif Karim, Sabrina Binte Zahir, Md Masum Billah. "Deep learning predictive model for colon cancer patient using CNN-based classification." Journal Name, Year.

[3] M. Kavitha and P. Nirmala. "Analysis and comparison of SVM-RBF algorithms for colorectal cancer detection over convolutional neural networks with improved accuracy." Vol. 13 SPECIAL ISSUE 04 (2022), 2022.

[4] Duggani Keerthana, Vipin Venugopal, Malaya Kumar Nath, Madhusudhan Mishra. "Hybrid convolutional neural networks with SVM classifier for classification of skin cancer." Biomedical Engineering Advances, 2023.

[5] Abdulrazak Yahya Saleh, Chee Ka ChinChee, Vanessa Penshie Hamada Al-Absi. "Lung cancer medical images classification using hybrid CNN-SVM." International Journal of Advances in Intelligent Informatics, 2021.

[6] Shital D. Bhatt, Himanshu B. Soni, Heena R. Kher, Tanmay D. Pawar. "Automated system for lung nodule classification based on ResNet50 and SVM." International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), 2022.

[7] Aayush Rajput, Abdulhamit Subasi. "Chapter 2 - lung cancer detection from histopathological lung tissue images using deep learning." Applications of Artificial Intelligence in Medical Imaging, 2023.

[8] A. Asuntha, A. Brindha, S. Indirani, Srinivasan Andy. "Lung cancer detection using SVM algorithm and optimization techniques." Journal of Chemical and Pharmaceutical Sciences, 2016.

[9] Saima Rathore, Mutawarra Hussain, Ahmad Ali, Asifullah Khan. "A recent survey on colon cancer detection techniques." IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2013.

[10] Anil B Gavade, Rajendra Nerli, Neel Kanwal, Priyanka A Gavade, Shridhar Sunilkumar Pol, Syed Tahir Hussain Rizvi. "Automated diagnosis of prostate cancer using mpMRI images: A deep learning approach for clinical decision support." Computers, 2023.

[11] Changjiang Zhou, Yi Jin, Yuzong Chen, Shan Huang, Rengpeng Huang, Yuhong Wang, Youcai Zhao, Yao Chen, Lingchuan Guo, Jun Liao. "Histopathology classification and localization of colorectal cancer using global labels by weakly supervised deep learning." Computerized Medical Imaging and Graphics, 2021.

[12] Sourodip Ghosh, Ahana Bandyopadhyay, Shreya Sahay, Richik Ghosh, Ishita Kundu, KC Santosh. "Colorectal histology tumor detection using ensemble deep neural network." Engineering Applications of Artificial Intelligence, 2021.

[13] Jiayi Fan, JangHyeon Lee, YongKeun Lee. "A transfer learning architecture based on a support vector machine for histopathology image classification." Applied Sciences, 2021.

[14] Masoud KhazaeeFadafen, Khosro Rezaee. "Ensemble-based multi-tissue classification approach of colorectal cancer histology images using a novel hybrid deep learning framework." Scientific Reports, 2023.

[15] Padraig Corcoran. "An end-to-end graph convolutional kernel support vector machine." Applied Network Science, 2020.

[16] Shihao Jiang, Richard Hartley, Basura Fernando. "Kernel support vector machines and convolutional neural networks." 2018 Digital Image Computing: Techniques and Applications (DICTA), 2018.

[17] Nils Gessert, Lukas Wittig, Daniel Dromann, Tobias Keck, Alexander Schlaefer, David B Ellebrecht. "Feasibility of colon cancer detection in confocal laser microscopy images using convolution neural networks." Bildverarbeitung fur die Medizin 2019: Algorithmen–Systeme–Anwendungen. Proceedings des Workshops vom 17. bis 19. Marz 2019 in Lubeck, 2019.

[18] Muhammed Talo. "Automated classification of histopathology images using transfer learning." Artificial Intelligence in Medicine, 2019.

[19] Bruno Korbar, Andrea M Olofson, Allen P Miraflor, Catherine M Nicka, Matthew A Suriawinata, Lorenzo Torresani, Arief A Suriawinata, Saeed Hassanpour. "Deep learning for classification of colorectal polyps on whole-slide images." Journal of Pathology Informatics, 2017.

[20] Felipe Andre Zeiser, Cristiano Andre da Costa, Gabriel de Oliveira Ramos, Henrique C Bohn, Ismael Santos, Adriana Vial Roehe. "Deepbatch: A hybrid deep learning model for interpretable diagnosis of breast cancer in whole-slide images." Expert Systems with Applications, 2021.

[21] Yu Liu, Ao Li, Jiangshu Liu, Gang Meng, Minghui Wang. "Tsdlpp: a novel two-stage deep learning framework for prognosis prediction based on whole slide histopathological images." IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2021.

[22] Yuchao Zheng, Chen Li, Xiaomin Zhou, Haoyuan Chen, Hao Xu, Yixin Li, Haiqing Zhang, Xiaoyan Li, Hongzan Sun, et al. "Application of transfer learning and ensemble learning in image-level classification for breast histopathology." Intelligent Medicine, 2023.

[23] Hao Jiang, Wai-Ki Ching, Wai-Shun Cheung, Wenpin Hou, Hong Yin. "Hadamard kernel SVM with applications for breast cancer outcome predictions." BMC Systems Biology, 2017.

[24] Noramalina Mohd Hatta, Zuraini Ali Shah, Shahreen Kasim. "Evaluate the performance of SVM kernel functions for multiclass cancer classification." International Journal of Data Science, 2020.

# APPENDIX

**1) Google Colab:** Google Colab, short for Collaborator, is an online platform provided by Google that allows users to write and execute Python code in a cloud-based environment. It offers a Jupyter Notebook-like interface with several added functionalities. Here are some key features of Google Colab:

- Cloud-based Computing Environment:
  - Google Colab provides a cloud-based computing environment, allowing users to run Python code on Google's cloud servers without the need for any local setup or installation.
  - It offers access to powerful computing resources, including GPUs and TPUs, which can accelerate computations for tasks such as machine learning and data analysis.
- Integration with Google Drive:
  - Colab seamlessly integrates with Google Drive, enabling users to access and store notebooks and data files directly from their Google Drive account.
  - Users can easily import datasets and export results to Google Drive, facilitating seamless data management and collaboration.
- Pre-installed Libraries and Packages:
  - Colab comes pre-installed with popular Python libraries and packages commonly used for data science and machine learning tasks, including NumPy, pandas, TensorFlow, and PyTorch.
  - Users can install additional libraries using pip or conda commands, extending the capabilities of the environment as needed.
- Collaboration and Sharing:
  - Google Colab supports real-time collaboration, allowing multiple users to work on the same notebook simultaneously.
  - Notebooks can be shared with collaborators via a shareable link, similar to Google Docs, enabling easy collaboration on projects.

**2) Visual Studio Code (VS Code):** Visual Studio Code (VS Code) is a lightweight, open-source code editor developed by Microsoft. It offers a versatile and customizable environment for writing, debugging, and executing code across different programming languages. Here are some key features of Visual Studio Code:

- Integrated Development Environment (IDE):

• VS Code provides a comprehensive Integrated Development Environment (IDE) with features such as syntax highlighting, code completion, and intelligent code suggestions, enhancing productivity for developers.

• It supports a wide range of programming languages, including Python, JavaScript, Java, C++, and many others, making it suitable for diverse development tasks.

o Extensions Marketplace:

• VS Code features a rich ecosystem of extensions available through the Visual Studio Code Marketplace.

• Users can install extensions to customize and extend the functionality of the editor, adding support for additional languages, frameworks, and tools tailored to their specific needs.

o Built-in Terminal:

• VS Code includes a built-in terminal that allows users to run commands, execute scripts, and interact with the operating system without leaving the editor.

• It supports multiple terminal instances, enabling users to work with different environments and run commands in separate tabs or panes.

o Debugging Support:

• VS Code offers robust debugging capabilities with support for breakpoints, watch variables, and interactive debugging sessions.

• It integrates with various debuggers and provides a unified debugging experience across different programming languages, streamlining the debugging process for developers.

o Version Control Integration:

• VS Code seamlessly integrates with version control systems such as Git, providing built-in support for common version control operations like commit, push, pull, and branch management.

• Users can view version control status, diff changes, and resolve merge conflicts directly within the editor, simplifying collaboration and code management workflows.

o Cross-platform Compatibility:

• VS Code is cross-platform and runs on Windows, macOS, and Linux operating systems, ensuring a consistent development experience across different platforms.

• It offers seamless synchronization of settings, extensions, and preferences via the built-in Settings Sync feature, enabling users to maintain a consistent environment across multiple devices.

**3) Python packages:** Python packages are collections of modules, functions, classes, and other resources that provide specific functionality and can be easily installed and imported into Python programs. They enable developers to leverage existing code and libraries to perform a wide range of tasks efficiently.

**1. Scikit-learn/ Scikit:**

- Scikit-learn is a popular machine learning library for Python.
- It provides a wide range of algorithms and tools for tasks such as classification, regression, clustering, and dimensionality reduction.
- Scikit-learn is known for its user-friendly API and extensive documentation, making it suitable for both beginners and experienced machine learning practitioners.

**2. Pandas:**

- Pandas is a powerful data manipulation and analyses is library for Python.
- It offers data structures like Data Frames and Series, which allow for easy handling, filtering, and transformation of structured data.
- Pandas provides functionalities for data cleaning, merging, reshaping, and aggregation, making it an essential tool in data analysis and preprocessing work flows.

**3. TensorFlow:**

- TensorFlow is an open-source deep learning frame work developed by Google.
- It allows for building and training various types of machine learning models, with a particular emphasis on deep neural networks.
- TensorFlow provides a flexible architecture and supports distributed computing, making it suitable for large-scale machine learning tasks.
- It offers high-level AP Is like Keras for simplified model development and lower- level APIs for more advanced customization.

**4. Open CV-Python:**

- Open CV-Python is a Python wrap per for the Open CV (Open-Source Computer Vision) library.
- It provides a wide range of computer vision algorithms and tools for tasks such as image and video processing, feature extraction, object detection, and more.
- Open CV-Python is widely used for computer vision research, robotics, augmented reality, and other applications involving visual data.

**5. NumPy:**

- NumPy is a fundamental package for scientific computing in Python.
- It provides support for multi-dimensional arrays, mathematical functions, line a algebra operations, and random number generation.
- NumPy is the foundation for many other scientific computing libraries and is widely used in fields like data analysis, machine learning, and numerical simulations.

**6. Matplotlib:**

- Matplotlib is a plotting and visualization library for Python.
- It offers a flexible and comprehensive set of tools for creating a wide range of static, animated, and interactive visualizations.
- Mat plot lib can generate various types of plots, including line plots, scatter plots, bar plots, histograms, and more.

**7. Keras:**

- Keras is a high-level neural networks API written in Python.
- It provides a simplified interface to build and train deep learning models, hiding the complexities of lower-level frameworks like TensorFlow.
- Keras emphasize seas of use, modularity, and extensibility, making it popular among beginners and rapid prototyping.