ABSTRACT

Cancer, a multifaceted disease driven by genetic and biochemical anomalies, poses a significant health threat globally. Lung and colon cancers, in particular, have emerged as prominent causes of illness and mortality. The timely identification of these malignancies through histopathological analysis is pivotal for devising effective treatment strategies and improving patient outcomes. Machine learning and deep learning methodologies offer promising avenues for accelerating cancer detection processes, facilitating the analysis of vast patient datasets in a cost-efficient and timely manner. Our study introduces a novel hybrid ensemble feature extraction model aimed at enhancing the detection of lung and colon cancer. This model integrates advanced techniques such as deep feature extraction and ensemble learning, coupled with specialized filtering mechanisms tailored for cancer image datasets. Through rigorous evaluation using histopathological datasets specifically curated for lung and colon cancer (LC25000), we assess the performance of our proposed approach.

CONTENT

Chapter 1: Introduction

1.1 Introduction

Lung and colon cancer remain formidable adversaries in the realm of global health, often presenting intertwined challenges. Recent studies have shed light on the occurrence of both cancers within a short timeframe, with Koich Kurishima et al. (2018) documenting instances where 17 out of 3102 patients diagnosed with lung cancer also developed colon cancer within a mere month. Despite lung cancer's conventional association with upper aerodigestive tract cancer, its coexistence with gastrointestinal cancer underscores the critical need for simultaneous detection and intervention strategies, highlighting the imperative role of medical imaging in early diagnosis and treatment planning. The World Health Organization (WHO) sounded the alarm in 2020, reporting a staggering 10 million cancer-related fatalities globally. Lung cancer emerged as a significant contributor to this grim statistic, with 2.21 million new cases reported, closely followed by colon cancer at 1.93 million, collectively accounting for approximately 21.4% of all cancer diagnoses. Alarmingly high mortality rates were observed, with lung cancer claiming 18% of lives and colon cancer 9.4%.

The ramifications of undetected or untreated lung cancer can be dire, with the disease typically categorized into Non-Small-Cell-Lung-Cancer (NSCLC) and Small-Cell-Lung-Cancer (SCLC). NSCLC, the predominant subtype comprising 80-85% of cases, often originates in the body's cavity-lining cells, predominantly manifesting as Adenocarcinomas localized in the outer lung regions. The intricate detection process for lung cancer typically entails segmentation, feature extraction, and subsequent classification, underscoring the multifaceted nature of combating this formidable adversary.

Against the backdrop of escalating mortality rates attributed to lung and colon cancers, the exploration of artificial intelligence (AI) techniques for detection has gained considerable traction. Leveraging machine learning and AI offers a myriad of feature extraction, optimization, and classification methodologies tailored for biomedical image analysis, presenting a promising avenue for early disease interception. Transfer learning,

leveraging pre-trained models, has emerged as a particularly promising approach in medical image analysis, offering substantial time and resource savings compared to conventional model training paradigms. Pioneering research endeavors have harnessed the power of transfer learning in diverse medical domains, from diagnosing multiple sclerosis to COVID-19 detection via medical imaging. Deep learning methodologies have permeated various sectors, with medical imaging emerging as a fertile ground for innovation.

Within the medical landscape, transfer learning and AI serve as indispensable allies, providing non-invasive diagnostic modalities that empower clinicians with actionable insights. Nonetheless, the formidable challenges inherent in curating machine learning-ready medical image databases underscore the need for continued advancements. Recent breakthroughs culminated in the development of the LC25000 dataset, boasting 25,000 meticulously curated color images spanning five classes, encompassing colon adenocarcinoma, benign colonic tissue, lung adenocarcinoma, lung squamous cell carcinoma, and benign lung tissue. These invaluable resources have fueled the implementation of cutting-edge classification algorithms aimed at concurrently detecting lung and colon cancer, thereby spotlighting the paramount significance of early intervention and treatment optimization.

The intersection between lung and colon cancer presents a complex landscape for medical practitioners and researchers alike. Understanding the interplay between these two cancers is crucial for devising effective screening, diagnostic, and treatment strategies. One of the challenges lies in the shared risk factors between lung and colon cancer, such as smoking, dietary habits, and genetic predispositions. Addressing these risk factors comprehensively could potentially aid in reducing the incidence and burden of both cancers on a global scale.

However, the integration of AI into clinical practice necessitates robust validation studies, seamless integration with existing workflows, and adherence to regulatory standards to ensure patient safety and data privacy. Collaborative efforts between academia, industry, and regulatory agencies are paramount in navigating the complex regulatory landscape and fostering the responsible deployment of AI-enabled medical devices and software.

1.2 Motivation

Our primary objective is to bridge the gap between research discoveries and their real-world applications, fostering tangible benefits for both patients and healthcare providers. Through the validation and implementation of effective detection models, we aspire to revolutionize clinical practice and healthcare delivery. By enabling earlier diagnosis, tailoring treatment plans to individual patients, and facilitating more precise monitoring of disease progression, our efforts aim to significantly enhance patient outcomes. Ultimately, this approach can lead to higher survival rates, reduced morbidity, and an overall improvement in the quality of life for individuals battling cancer.

Our commitment to rigorous research and experimentation drives us to continually push the boundaries of our understanding of cancer detection methodologies. By contributing novel insights and knowledge to the scientific community, we not only advance our collective understanding but also lay the groundwork for future developments in the field.

Central to our endeavors is the integration of cutting-edge AI and machine learning techniques. These technologies empower us to develop innovative solutions that not only improve the accuracy and efficiency of cancer detection from medical images but also enhance accessibility to these critical diagnostic tools.

Chapter 2: Literature Survey

2.1 Outcome of Literature Survey

In recent years, significant strides have been made in the realm of medical imaging analysis and cancer detection, as evidenced by a diverse array of methodologies proposed in the literature. These approaches span from novel algorithms like the Greedy snake algorithm for tumor detection to more sophisticated methods such as Content-Based Image Retrieval (CBIR) integrated with Computer-Aided Diagnosis (CAD). Leveraging advanced algorithms for feature extraction, like the Omega algorithm, holds promise for enhancing precision in medical imaging analysis. However, challenges persist, including the lack of prior probabilities and knowledge of object/background distributions, necessitating ongoing research in segmentation and edge detection methodologies. Additionally, the integration of CBIR with clinical information systems is highlighted, emphasizing the critical role of post-processing managers. Immediate integration may present challenges due to poor edge strength, prompting the exploration of refined edge detection techniques for successful integration with clinical systems.

Moreover, recent advancements in brain tumor segmentation, breast cancer detection, and liver cancer prediction models underscore the evolving landscape of medical imaging analysis. While edge detection methodologies for brain tumor segmentation and multi-step approaches for breast cancer detection show promise, they encounter limitations in capturing dynamic changes and maintaining image contrast. Meanwhile, liver cancer prediction models leveraging neural networks demonstrate effectiveness in recognizing liver regions and cancerous areas. Nevertheless, challenges arise from the high correlation of predicted variables, prompting the development of systems to reduce correlations while optimizing processing time. Furthermore, machine learning approaches, particularly component-based Support Vector Machine (SVM) methods, have gained prominence, offering good sample generalization and unique solutions to classification problems. Yet, scalability and speed remain areas of concern, urging researchers to explore efficient algorithms and techniques for real-world application.

Nehemiah et al **[1]** proposed the Greddy snake algorithm in 2012 for detecting the tumors .Utilizing slices of images instead of a single image, we employ a thresholding algorithm for segmentation and the Greedy snake algorithm for identifying regions of interest (ROI). These techniques enable us to extract features using a feature extraction subsystem.The adoption of image slices enhances performance significantly. However, in this approach, prior probabilities may not be available, and distributions for object/background are often unknown.

Pedro et al **[2]** proposed the Content-Based Image Retrieval (CBIR) in 2009 for detecting the tumors. The fusion of Content-Based Image Retrieval (CBIR) and Computer-Aided Diagnosis (CAD) has gained traction. Researchers are exploring feature extraction methods like Association Rule (AR) techniques such as FAR (Feature Extraction through Association Rule) and Idea Association. The Omega algorithm is increasingly employed for feature extraction, effectively eliminating unwanted features and enhancing efficiency. This integrated approach reduces dimensionality in feature vectors, promising improved precision in medical imaging analysis.

In their research paper, Petra et al. **[3]** introduce a novel method known as Content-Based Image Retrieval (CBIR). As the integration of CBIR with modern clinical information systems progresses, the pivotal role of the post-processing manager becomes increasingly apparent. This post-processing manager plays a central role in coordinating a myriad of tasks, ranging from scheduling to ensuring the seamless flow of information within the system. Numerous studies have underscored the critical nature of the post-processing manager's function within the integrated system. However, it has become evident that immediate integration may encounter hurdles. One notable challenge revolves around concerns regarding the quality of integrated images, particularly in terms of edge strength. Suboptimal edge strength can lead to compromised image characterization, thereby undermining the effectiveness of CBIR within clinical settings. The issue of poor edge strength in integrated images has been consistently observed across various studies, highlighting its significance in the successful integration of CBIR with clinical information systems. Addressing this challenge necessitates the development and implementation of robust edge detection and characterization methods. By enhancing the fidelity and accuracy of edge detection algorithms, researchers can mitigate the risk of compromised image quality and ensure the reliability of CBIR-based image retrieval within clinical workflows.

Dipali et.al **[4]** proposed the Edge Detection methodologies in 2010 for detecting the tumors. Brain tumor segmentation techniques, researchers commonly adopt a multi-step approach. Initially, they compute the histogram of the input image and determine an appropriate threshold value for segmentation. Edge detection methods are then applied to detect boundaries, followed by the utilization of sharpening filters to enhance tumor detection.While these methods facilitate the identification of brain tumors and their types, it is noted in literature that they may encounter limitations in capturing dynamic changes across the entire image. This insight underscores the need for further research into segmentation algorithms capable of accommodating such dynamic variations for more comprehensive tumor detection and characterization.

In the context of these paper William et al **[5]** proposed on breast cancer detection and segmentation methods, a common approach involves five distinct stages in 1997. Initially, researchers focus on identifying and isolating mass-like regions within breast images. Subsequently, efforts are directed towards segmenting breast tissue from the background to facilitate accurate analysis. One prevalent technique employed in this process is the utilization of the horizontal gradient fill algorithm, which helps to mitigate warp-around errors. However, findings from the literature suggest that while this algorithm aids in reducing errors, it inherently leads to an overall reduction in image contrast. Beyond the initial stages of identifying mass-like regions and segmenting breast tissue, recent studies have emphasized the importance of incorporating advanced computational techniques such as machine learning and deep learning algorithms. This reduction in contrast can compromise the effectiveness of subsequent analysis and diagnosis. Moreover, studies have noted that the algorithm exhibits shift sensitivity and poor directionality, which may pose challenges in accurately delineating boundaries and features within the breast tissue. The integration of multimodal imaging modalities, such as MRI and ultrasound, presents opportunities for comprehensive and complementary assessment of breast lesions, These insights underscore the ongoing need for refined segmentation methods that can effectively address these limitations while ensuring robust and accurate detection of breast abnormalities.

In the context of these paper Tadashi et al. **[5]** proposed on liver cancer prediction models, researchers commonly employ neural networks to predict variables, utilizing optimized architectures for improved accuracy in 2010. These networks are often designed to recognize and extract both the liver region and cancerous areas within it.However, findings from existing studies indicate that predicted variables are often highly correlated, leading to challenges in generating accurate and stable prediction values. To address this issue, researchers have developed systems aimed at reducing correlations between variables, thereby enhancing prediction accuracy and stability.Despite the effectiveness of such systems in mitigating correlation-related challenges, it is noted in the literature that they typically require high processing time due to the complexity of the algorithms involved. This insight underscores the ongoing efforts within the research community to develop more efficient prediction models that balance accuracy with computational efficiency in the context of liver cancer diagnosis and prognosis.

In the context of proposed method by I.A Illan et al **[6]** in 2013 on machine learning approaches for pattern recognition tasks, researchers often employ sophisticated techniques such as automatic feature selection and mask-based feature reduction to enhance model performance. In particular, component-based Support Vector Machine (SVM) methods have gained prominence for their ability to achieve good sample generalization and provide unique solutions to classification problems.However, findings from existing studies suggest that despite the advantages of SVM, certain factors such as speed, handling of discrete data, and scalability with large datasets need to be considered. While SVM is effective in handling high-dimensional data and delivering robust solutions, its computational complexity and sensitivity to parameters can impact its practical applicability, especially in scenarios with stringent time constraints or diverse data types.Therefore, while component-based SVM remains a promising approach in pattern recognition tasks, ongoing research aims to address its limitations and optimize its performance across various application domains. This includes efforts to develop efficient algorithms, address scalability concerns, and enhance its suitability for handling diverse data types encountered in real-world scenarios.

In the context of proposed method by F. CalleAlonso et.a **[7]** in 2020 concerning classification methodologies and feature extraction techniques, a pioneering hybrid method has been devised specifically tailored for pairwise comparisons. This innovative approach

strategically integrates Bayesian regression and k-nearest neighbor techniques, harnessing their complementary strengths.This hybrid strategy has proven highly beneficial for tackling classification tasks, offering the capability to extract a substantial number of features while operating with a minimal number of elements. Through the synergistic combination of Bayesian regression and k-nearest neighbor methods, researchers have achieved notable advancements in classification accuracy and resilience across diverse tasks.However, it's important to note from the literature that this approach hinges on an iterative process of refining residuals, rendering it particularly susceptible to noise. Despite its efficacy in extracting comprehensive feature sets, the sensitivity to noise underscores the necessity for meticulous preprocessing and the adoption of regularization techniques to mitigate its adverse effects on classification performance.

Recent literature has predominantly focused on leveraging deep learning and transfer learning techniques for the detection of colon cancer through histopathological image analysis. Tongaçar **[8]** demonstrated the utilization of an AI-supported model and optimization methods to categorize both lung and colon cancers based on histopathological images. By employing the DarkNet-19 model for training image classes from scratch and subsequently utilizing a support vector machine (SVM) for classification, an impressive accuracy rate of 99.69% was achieved. Kumar et al. conducted a comparative study investigating feature extraction approaches for lung and colon cancer classification. They extracted six handcrafted traits encompassing color, texture, shape, and structure, utilizing traditional classifiers for colon cancer classification. The utilization of DenseNet-121 for feature extraction, combined with an RF classifier, yielded the highest accuracy of 98.60% along with impressive recall, precision, and F1 score values. Additionally, Yildirim and Cinar proposed a CNN-based method, MA_ColonNET, employing a 45-layer model to detect colon cancer from images. As such, the integration of AI-supported models and optimization methods holds promise for enhancing the efficiency and reliability of colon cancer detection in clinical practice. Continuing advancements in this field are essential for advancing the early detection and treatment of colorectal malignancies.

2.2 Problem Statement

To devise an effective approach for detecting colon and lung cancer, this study aims to utilize hybrid models that integrate Deep Learning and Machine Learning Techniques.

Objectives

The goal is to develop a system capable of accurately detecting both colon and lung cancer using a combination of Deep and Machine Learning methods, aiming to achieve high accuracy in diagnosis.

Chapter 3: Methodology

3.1 Dataset Details

The field of Machine Learning, a subset of Artificial Intelligence, has led to remarkable advancements in many areas, including medicine. Machine Learning algorithms require large datasets to train computer models successfully. Although there are medical image datasets available, more image datasets are needed from a variety of medical entities, especially cancer pathology. Even more scarce are ML-ready image datasets. **Table 1** shows the dataset Distribution. To address this need, we created an image dataset (LC25000) with 25,000 color images in 5 classes. Each class contains 5,000 images of the following histologic entities: colon adenocarcinoma, benign colonic tissue, lung adenocarcinoma, lung squamous cell carcinoma, and benign lung tissue. All images are de-identified, HIPAA compliant, validated, and freely available for download to AI researchers.

Table 1 : Dataset

Dataset Name	**LC25000**
Number of Images	25000
Images Types	Color (BGR)
Classes	
- **Colon Adenocarcinoma**	5000
- **Benign Colonic Tissue**	5000
- **Lung Adenocarcinoma**	5000
- **Lung Squamous Cell Carcinoma**	5000
- **Benign Lung Tissue**	5000

3.2 Architecture of proposed model

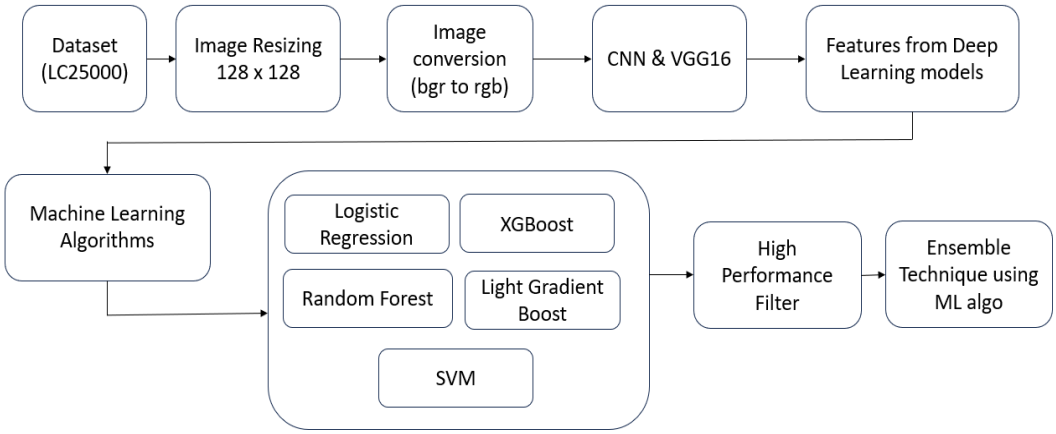

Fig .1. : Architecture of Proposed Methodology

The proposed model has the following components .You can see the proposed architecture in **Fig.1.** The dataset consist of 25000 images spread over 5 classes. Where 2 classes belongs to colon cancer and remaining 3 belongs to lung cancer. Initial step is to preprocess the dataset to make it ready for training, we have split the dataset in Training and Test set where 80% of images are for Training and remaining for Test set to see the performance of our models.

Step 1 :- Images are of size 768 x 768 they are scaled to 128 x 128. This resizing step ensures uniformity in the input dimensions across the dataset, which is crucial for training deep learning models. This rescales the pixel values of the images to be between 0 and 1. This is a standard preprocessing step for neural networks as it helps in convergence and stability during training. (shear_range) a parameter for performing shear transformations on the images. Shear transformation displaces each point in the image by a certain amount proportional to its distance from an axis. (zoom_range) This parameter allows randomly zooming inside pictures. (horizontal_flip) This parameter randomly flips images horizontally. This augmentation technique helps in creating more robust models by increasing the diversity of the training data.

Step 2 :- The conversion from BGR (Blue, Green, Red) to RGB (Red, Green, Blue) is necessary for compatibility with certain pre-trained models, such as CNN and VGG16. Pre-trained models like VGG16 have been trained on large datasets using specific input formats. VGG16, for example, was trained using images in RGB format. So, if you want

to use these pre-trained models, you need to ensure that your input images are in the same format as the data they were trained on. Otherwise, the model may not perform as expected.

Step 3 :- After training a convolutional neural network (CNN) model like VGG16 on a large dataset for tasks like image classification, the convolutional layers learn to extract hierarchical features from the input images. These features capture different levels of abstraction, starting from simple edges and textures to more complex shapes and patterns. To use these learned features in a machine learning model, we typically take the output of one of the convolutional layers as a feature vector. This feature vector represents the extracted features for each input image. To extract features from the pre-trained CNN model, you feed the input images through the model (up to the desired layer from which you want to extract features). This process is known as forward propagation. Instead of using the model for prediction, you extract the output of the chosen layer for each input image. This output serves as the feature representation of the input image. Once you have extracted the features for all images in your dataset, you obtain a feature matrix where each row represents the feature vector for one image.

3.2.1 Preprocessing Data:-

INPUT IMAGES :-

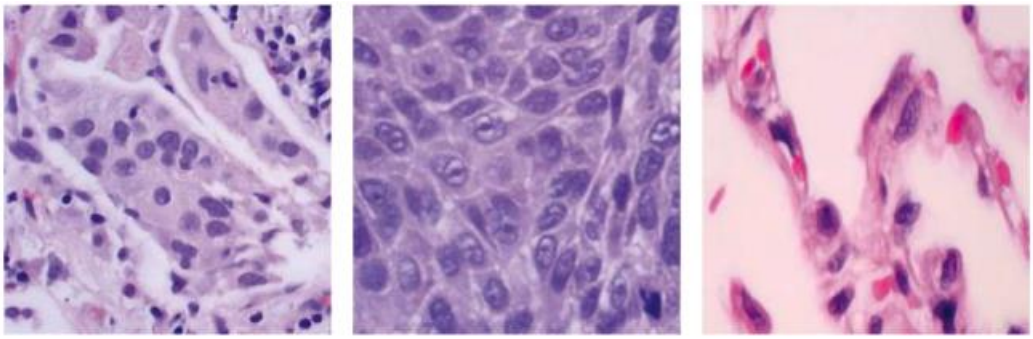

Fig.2. Before Preprocessing Lung

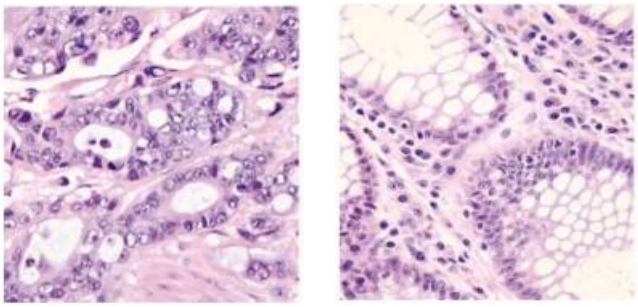

Fig.3. Before Preprocessing Colon

BGR to RGB conversion involves swapping the order of color channels in an image from blue-green-red to red-green-blue.**Fig.2.** shows Lung images before preprocessing, same **Fig.3.** shows Colon images before preprocessing. In BGR format, the first channel represents blue intensity, the second represents green intensity, and the third represents red intensity. This ordering is common in certain image processing libraries and frameworks like OpenCV. However, many deep learning frameworks, including TensorFlow and PyTorch, expect images to be in RGB format, where the first channel represents red, the second represents green, and the third represents blue. To convert BGR to RGB, each pixel's color channels are rearranged accordingly. This conversion ensures consistency in color representation across different frameworks and facilitates accurate processing and interpretation of images within deep learning pipelines. It's a simple yet essential preprocessing step before feeding images into models trained for RGB format.

Fig.4. After Preprocessing Lung

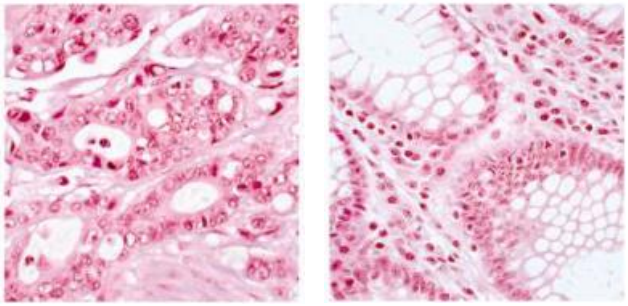

Fig.5. After Preprocessing Colon

In the case of CNNs, multiple layers of convolutional, pooling, and activation functions are applied to extract increasingly abstract features from the input images**. Fig.4** shows images of Lung converted to RGB format, same for **Fig.5** shows colon images converted to RGB format.These features capture hierarchical patterns and structures present in the images, relevant to distinguishing between cancerous and non-cancerous regions.

3.2.2 Trial Method with CNN :-

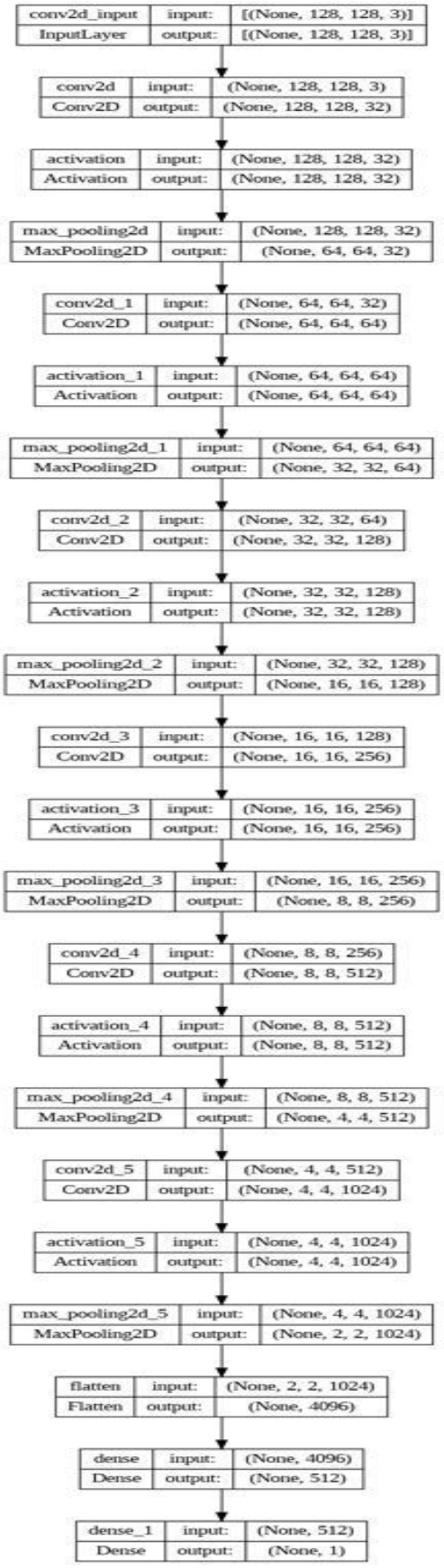

Fig.6. CNN

Initially, the model displayed promising results during the training phase, achieving a commendable accuracy of 90%. **Fig.6** shows the architecture of CNN used to extract features from the image dataset and then given as input to ML algorithms. Subsequent evaluation on the validation set further validated its performance, yielding an accuracy of 92%. However, when given to Machine Learning Algorithms for classification subjected to the train as well test dataset, the model's accuracy plummeted to a mere 50%, indicative of a significant lack of generalization. Despite diligent efforts to enhance its performance by fine-tuning hyperparameters and incorporating regularization techniques such as dropout and weight regularization, the model's improvement remained elusive. Consequently, a strategic decision was made to pivot towards employing a pre-trained VGG16 model. Renowned for its robustness in image-related tasks, VGG16 offers the advantage of transfer learning. As part of the preprocessing pipeline, BGR images were diligently converted to RGB format to align with VGG16's training on RGB images. This shift signifies a deliberate approach to leverage the established capabilities of pre-trained models and address the challenges encountered with the previous model architecture.

3.2.3 VGG16 :-

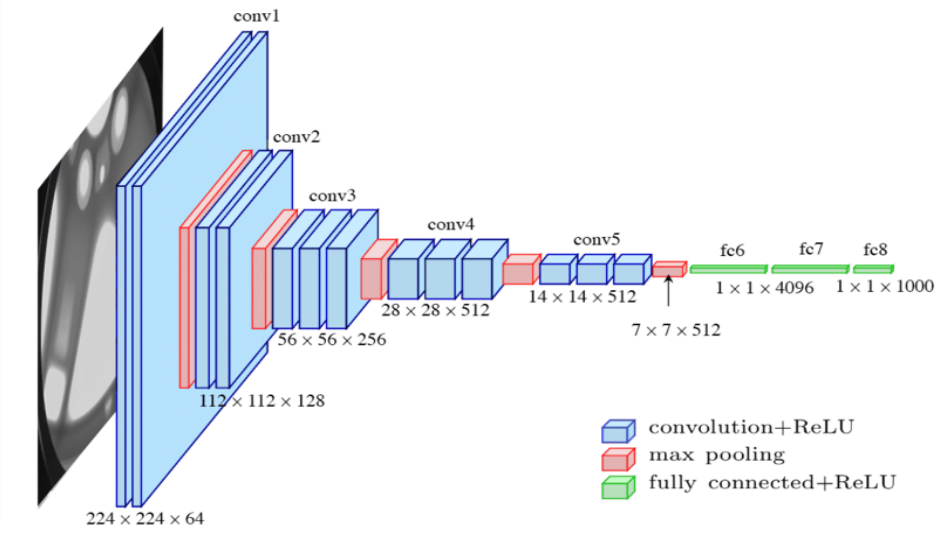

Fig.7. VGG Architecture

The VGG16 architecture is a convolutional neural network (CNN) architecture developed by the Visual Geometry Group (VGG) at the University of Oxford. It gained prominence for its simplicity and effectiveness in image classification tasks. Deep features of images is extracted using architecture used in **Fig.7.**

Input Layer :- Accepts input images with a fixed size, typically 224x224 pixels. For our Implementation we have scaled the images to 128 x 128 so we will give an input to VGG16 as 128 x 128.

Convolutional Blocks:- VGG16 consists of 13 convolutional layers, grouped into five blocks, each followed by a max-pooling layer. Each convolutional layer uses small receptive fields (3x3) with a stride of 1 and zero-padding to maintain the spatial dimensions of the input.

Activation Function:- Rectified Linear Unit (ReLU) activation functions are used after each convolutional layer to introduce non-linearity.

Max Pooling:- Max-pooling layers follow each convolutional block, reducing the spatial dimensions of the feature maps while retaining the most important features.

Fully Connected Layers:- After the convolutional layers, VGG16 has three fully connected layers followed by a softmax activation function for classification. These layers perform high-level reasoning and mapping of extracted features to class labels.

Dropout:- Dropout regularization is applied before the fully connected layers to prevent overfitting by randomly dropping a fraction of the neurons during training.

Output Layer:- The output layer consists of units corresponding to the number of classes in the classification task, with softmax activation to produce class probabilities.

Features extracted from VGG16 are given as input to Machine Learning algorithms and also implementation of Ensemble Techniques are used.

3.2.4 Machine Learning Algorithms :-

Features extracted from VGG16 are given to Machine Learning Algorithms like Logistic Regression(LR), Random Forest(RF),Support Vector Machine(SVM),XGBoost, Light Gradient Boosting(LGB).

I . Logistic Regression(LR) :- In logistic regression, the relationship between the input features and the output class is modeled using weights and a bias term. The weights determine the importance of each feature, while the bias term (intercept) accounts for any offset in the prediction. Logistic regression uses the logistic function (sigmoid function) to model the probability of an instance belonging to the positive class (e.g., cancerous). The logistic function is defined as: $P(y=1|x)= 1+e^{-(w\cdot x+b)}$ 1 Here, $P(y=1|x)$ represents the probability of the instance x belonging to the positive class, w denotes the weights vector, b is the bias term, and e is the base of the natural logarithm.

Logistic regression computes a decision boundary that separates instances into two classes based on their feature values. The decision boundary is determined by the weights and bias parameters and corresponds to the hyperplane where the probability of belonging to each class is equal.

Loss Function: Logistic regression minimizes a loss function, typically the logistic loss or cross-entropy loss, to optimize the model parameters (weights and bias). The loss function quantifies the difference between the predicted probabilities and the true labels, penalizing incorrect predictions.

Training: Logistic regression is trained using optimization algorithms such as gradient descent or its variants. During training, the model iteratively adjusts the weights and bias to minimize the loss function, thereby improving its predictive accuracy.

Prediction: Once trained, the logistic regression model can predict the probability of an image being cancerous or non-cancerous based on its extracted features. A threshold can be applied to these probabilities to make binary predictions.

convolutional backbone is able to extract features that are specific to each frame of thesequence since it processes each frame on its own. These features may then be utilised to track the object between frames because they are unique to each frame.

II . Random Forest :- Random Forest consists of an ensemble of decision trees, where each tree is trained on a random subset of the training data and a random subset of the features. This randomness helps in reducing overfitting and improving the generalization ability of the model. Each decision tree in the Random Forest is constructed recursively by partitioning the feature space based on the values of the features. At each node of the tree, the feature that best splits the data according to some criterion (e.g., Gini impurity or information gain) is selected.

Random Forest employs a technique called bagging, where multiple decision trees are trained independently on different subsets of the training data. This diversity in training data and feature selection helps in improving the robustness and accuracy of the model. During inference, each decision tree in the Random Forest independently predicts the class label for a given input image. The final prediction is then determined by aggregating the predictions of all decision trees through a voting mechanism (e.g., simple majority voting).

Random Forest has several hyperparameters that can be tuned to optimize its performance, such as the number of trees in the forest, the maximum depth of each tree, and the number of features considered for each split. Once trained, the Random Forest model can predict the class label (cancerous or non-cancerous) for new images based on their extracted features. The class label with the most votes across all decision trees is selected as the final prediction.

III. Support Vector Machine :-

Separating Hyperplane: SVM works by finding the optimal hyperplane that separates the feature space into two classes (cancerous and non-cancerous) with the maximum margin. The margin is the distance between the hyperplane and the nearest data points from each class, also known as support vectors.

Kernel Trick: SVM can efficiently handle non-linear decision boundaries by mapping the input features into a higher-dimensional space using kernel functions (e.g., radial basis function kernel). In this higher-dimensional space, SVM searches for a hyperplane that best separates the classes.

Margin Maximization: SVM aims to maximize the margin between the support vectors and the separating hyperplane while minimizing the classification error. This margin maximization leads to better generalization and robustness of the model.

Regularization: SVM uses a regularization parameter (C) to control the trade-off between maximizing the margin and minimizing the classification error. A smaller value of C results in a larger margin but may lead to misclassification, while a larger value of C allows for fewer misclassifications but may result in a smaller margin.

Kernel Selection: SVM offers different kernel functions (linear, polynomial, radial basis function, etc.) that can be selected based on the dataset characteristics and the desired decision boundary complexity. The choice of kernel significantly affects the model's performance. For our classification ‘linear’ kernel is used.

Training: SVM learns the optimal separating hyperplane and the support vectors by solving a convex optimization problem. The optimization process involves minimizing a cost function that penalizes misclassifications and maximizes the margin.

IV. XGBOOST : -

Gradient Boosting Ensemble: XGBoost belongs to the family of gradient boosting algorithms, which iteratively train a sequence of weak learners (decision trees in the case of XGBoost) to improve predictive performance. Unlike random forests, which train trees independently, XGBoost sequentially adds trees to correct the errors made by previous trees.

Decision Trees as Weak Learners: In XGBoost, decision trees are used as weak learners. Each decision tree is trained to predict the residual errors of the previous trees, aiming to gradually reduce the overall error of the ensemble model.

Objective Function: XGBoost minimizes a specific objective function, which consists of a loss function and a regularization term. The loss function quantifies the difference between the predicted values and the true labels, while the regularization term penalizes complex models to prevent overfitting.

Tree Construction: Decision trees in XGBoost are constructed using a greedy approach, where each split is chosen to maximize a predefined split criterion (e.g., information gain, Gini impurity). XGBoost supports various types of tree structures, including depth-wise and leaf-wise growth.

Regularization: XGBoost offers several hyperparameters for regularization, such as tree depth, minimum child weight, and column subsampling rate. These hyperparameters help control the complexity of the individual trees and prevent overfitting. Cross-Validation: XGBoost supports efficient cross-validation techniques for hyperparameter tuning and

model evaluation. Cross-validation helps in selecting the optimal set of hyperparameters and assessing the generalization performance of the model.

V. Light Gradient Boosting(LGB) :-

Gradient Boosting Ensemble: LightGBM, like XGBoost, is based on gradient boosting, which sequentially adds weak learners to correct the errors made by previous models. However, LightGBM introduces novel techniques to improve training speed and efficiency.

Leaf-Wise Tree Growth: LightGBM adopts a leaf-wise tree growth strategy, where trees are grown leaf-wise instead of level-wise. This strategy leads to fewer nodes being evaluated during tree construction, resulting in faster training times.

Gradient-based One-Side Sampling (GOSS): LightGBM implements GOSS to efficiently train on large-scale datasets. GOSS selectively samples instances with large gradients for gradient-based updates while keeping the instances with small gradients unchanged. This approach reduces the computational cost of gradient calculation without sacrificing model accuracy.

Exclusive Feature Bundling (EFB): LightGBM employs EFB to reduce memory usage and improve training speed. EFB groups exclusive features together during tree construction, allowing for more efficient memory access and storage.

Histogram-based Splitting: LightGBM uses histogram-based algorithms to find the best split points for each feature. Instead of using all individual feature values, LightGBM constructs histograms of feature values and selects split points based on these histograms, further accelerating training.

Regularization: LightGBM supports various regularization techniques, including L1 and L2 regularization, to control model complexity and prevent overfitting. Regularization parameters can be tuned to optimize model performance.

VI. Ensemble Technique:-

Ensemble techniques are a set of machine learning methods that combine multiple models to improve predictive performance. Instead of relying on a single model, ensemble methods leverage the strength of diverse models to make more accurate predictions. The basic idea behind ensemble techniques is that combining multiple weak learners can produce a strong learner, often outperforming any individual model.

Voting ensembles combine predictions from multiple models by averaging (soft voting) or taking the majority vote (hard voting) of the predictions made by each model.

In our Implementation, we have used Light Gradient Boosting and Support Vector Machine with ‘hard’ voting.

3.3 Evaluation Metrics

We are using following metrics in our project.

- **Accuracy**

  Measures the overall correctness of predictions by the model, calculated as the ratio of correctly predicted instances (both positive and negative) to the total number of instances. Higher accuracy indicates better overall performance.

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

  TP: is the number of true positives (correctly predicted positive instances)

  TN: is the number of true negatives (correctly predicted negative instances),

  FP: is the number of false positives (incorrectly predicted positive instances),

  FN: is the number of false negatives (incorrectly predicted negative instances).

- **Recall**

  Measures the proportion of correctly predicted positive instances among all actual positive instances. Recall is important when capturing all positive instances is critical, such as in disease detection or anomaly detection.

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

- **Precision**

  Indicates the proportion of correctly predicted positive instances among all instances predicted as positive. Precision is important when minimizing false positives is crucial, such as in medical diagnosis or fraud detection.

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

- **F-1 Score**

  Harmonic mean of precision and recall, providing a balanced measure of model performance. F1 score considers both false positives and false negatives, making it useful when class imbalance exists in the dataset.

$$F-1 = \frac{2TP}{2TP + FP + FN}$$

Chapter 4: Experiments and Results

The obtained results from the various machine learning models demonstrate promising performance in the classification task. **Table 2** shows the results obtained on the entire dataset of 25000 images. Support Vector Machine (SVM), Logistic Regression (LR), XGBoost, and Light Gradient Boosting (LGB) achieved high accuracy rates of 96%. SVM and LR exhibited excellent precision and recall scores, indicating their effectiveness in correctly identifying positive cases while minimizing false positives. Random Forest (RF) demonstrated slightly lower accuracy at 94%, with comparable precision and recall scores. Notably, XGBoost and LGB exhibited robust performance across all metrics, with balanced precision, recall, and F-1 scores. Ensemble methods, such as Ensemble SVM+LGB, maintained high accuracy levels while leveraging the strengths of individual models. Overall, these results underscore the efficacy of diverse machine learning techniques in colon cancer classification, with XGBoost, LGB, and ensemble approaches showing particular promise for further exploration and application in clinical settings.

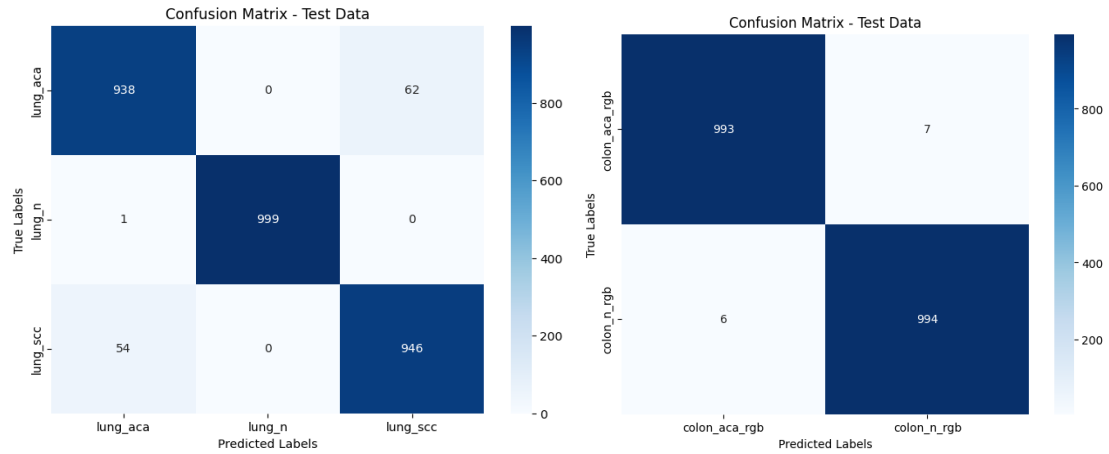

Fig.8.1 Confusion Matrix for SVM

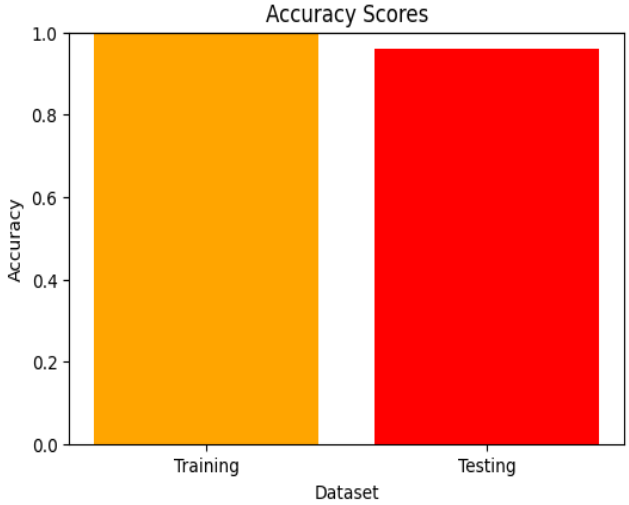

Fig .8.2 Accuracy of SVM

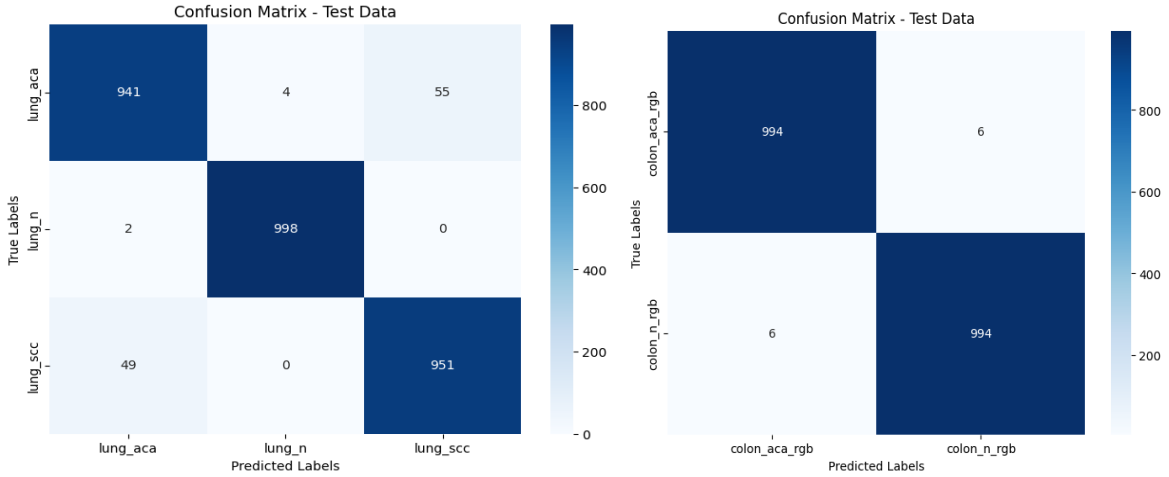

Fig.9.1 Confusion Matrix for LR

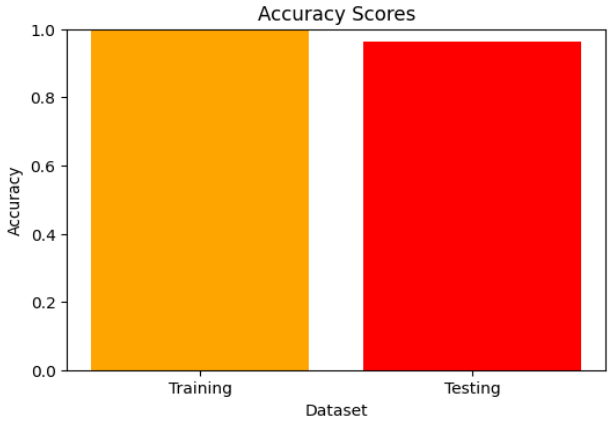

Fig.9.2 Accuracy of LR

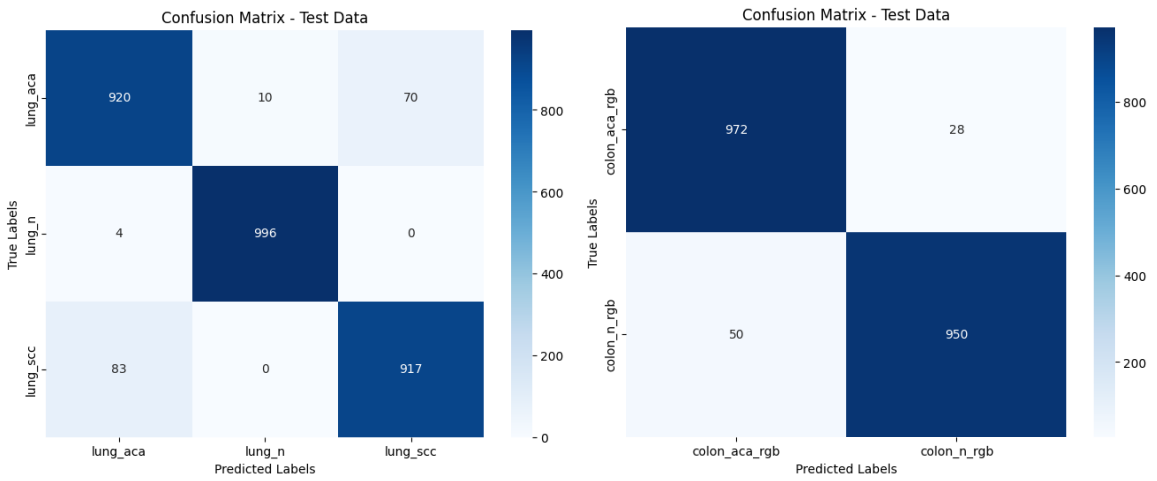

Fig.10.1 Confusion Matrix for Random Forest

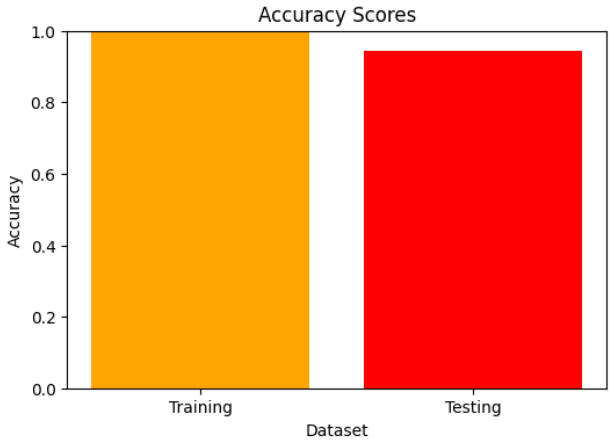

Fig.10.2 Accuracy of Random Forest

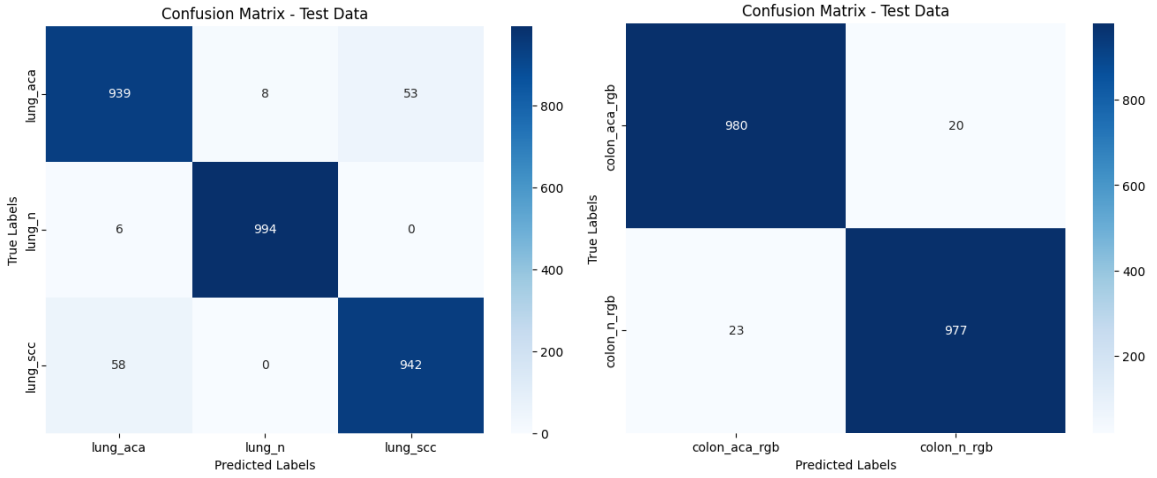

Fig.11. Confusion Matrix of XGBoost

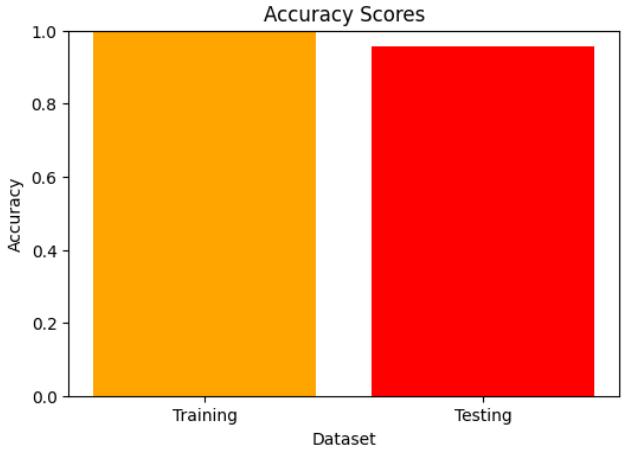

Fig.11.2 Accuracy of XGBoost

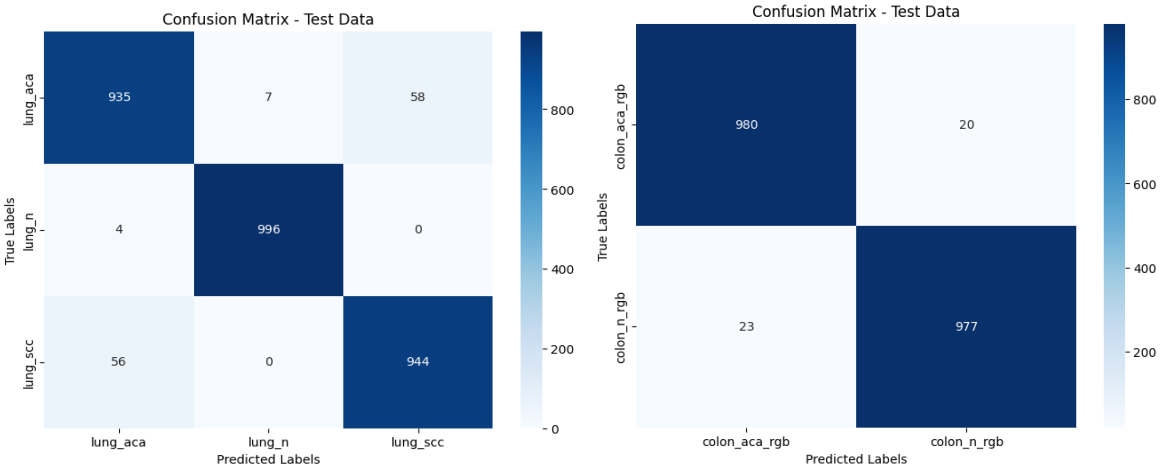

Fig. 12.1 Confusion Matrix of LGB

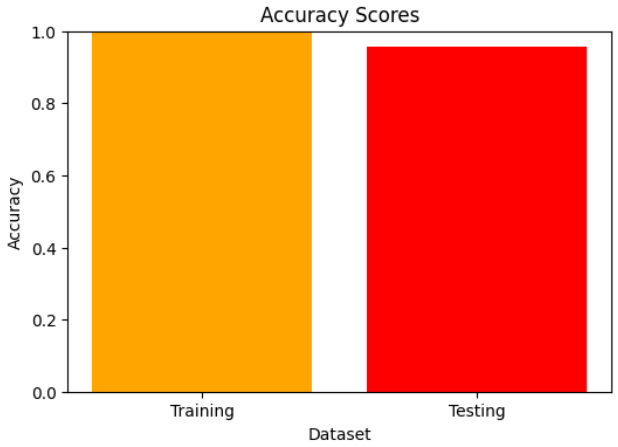

Fig.12.2 Accuracy of LGB

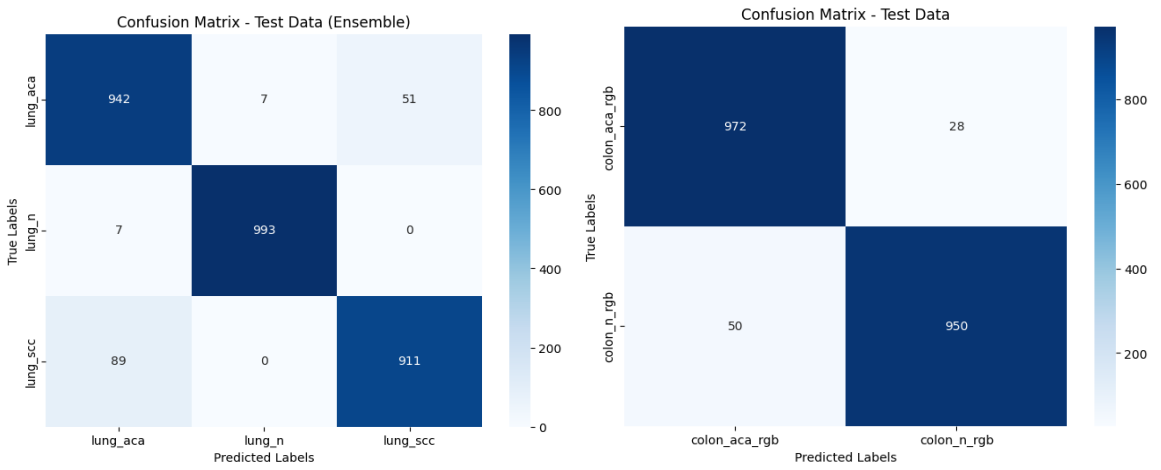

Fig.13.1 Confusion Matrix for Ensemble(LGB+SVM)

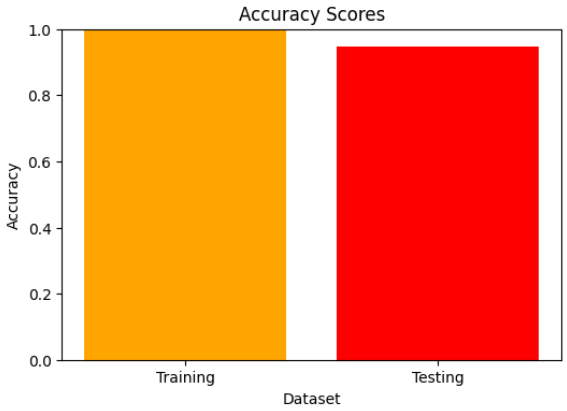

Fig.13.2 Accuracy of Ensemble(LGB+SVM)

The figures presented, including **Fig 8.1, Fig 8.2** for Support Vector Machine (SVM), **Fig 9.1, Fig 9.2** for Logistic Regression (LR), **Fig 10.1, Fig 10.2** for Random Forest (RF), **Fig 11.1, Fig 11.2** for XGBoost, **Fig 12.1, Fig 12.2** for Light Gradient Boosting (LGB), **and Fig 13.1, Fig 13.2** for the Ensemble Technique (LGB + SVM), showcase the detailed results obtained from each respective machine learning model. These figures likely depict various performance metrics such as accuracy, precision, recall, and F1-score across different experimental conditions or datasets. By visually representing the results in this manner, researchers and practitioners can gain insights into the comparative performance of each model and evaluate their efficacy in the context of colon cancer classification. The graphical representations provide a comprehensive overview, allowing for easy interpretation and comparison of the performance of different machine learning algorithms. This aids in making informed decisions regarding model selection and optimization strategies for clinical applications in cancer diagnosis and treatment planning.

Table 2 : Results of various Models

Sr No.	Models	Accuracy	Precision	Recall	F-1 Score
1	Support Vector Machine(SVM)	96%	94%	94%	94%
2	Logistic Regression(LR)	96%	95%	94%	94%
3	Random Forest(RF)	94%	91%	92%	92%
4	XGBoost	96%	92%	94%	93%
5	Light Gradient Boosting(LGB)	96%	94%	94%	94%
6	Ensemble SVM+LGB	95%	91%	94%	92%

Chapter 5: Conclusion and Future Work

In conclusion, the performance evaluation of various machine learning algorithms for cancer detection indicates promising results. Support Vector Machine (SVM), Logistic Regression (LR), XGBoost, and Light Gradient Boosting (LGB) demonstrate consistently high accuracy levels, ranging from 91% to 96%. Random Forest (RF) also performs well, albeit slightly lower. Ensemble techniques combining SVM and LGB show competitive performance, highlighting the potential for further refinement and optimization in cancer detection algorithms.

In image segmentation, cancerous regions can be accurately delineated, enabling precise identification of cancerous tissue boundaries. Future work involves incorporating advanced deep learning techniques to not only segment cancerous areas but also classify the stage of cancer. By integrating features such as tumor size, shape, and texture, along with clinical data, the system can provide valuable insights into cancer progression. Additionally, the development of predictive models leveraging longitudixnal data can aid in prognosis and treatment planning, ultimately improving patient outcomes in oncology.

References :-

[1] Abbasi, A. A., Hussain, L., Awan, I. A., Abbasi, I., Majid, A., Nadeem, M. S. A., et al. (2020). Detecting prostate cancer using deep learning convolution neural network with transfer learning approach. Cognitive Neurodynamics, 14, 523–533.

[2] Abdelsamea, M. M., Pitiot, A., Grineviciute, R. B., Besusparis, J., Laurinavicius, A., & Ilyas, M. (2019). A cascade-learning approach for automated segmentation of tumour epithelium in colorectal cancer. Expert Systems with Applications, 118, 539–552.

[3] Adeoye, J., Hui, L., Koohi-Moghadam, M., Tan, J. Y., Choi, S. -W., & Thomson, P. (2022). Comparison of time-to-event machine learning models in predicting oral cavity cancer prognosis. International Journal of Medical Informatics, 157, Article 104635.

[4] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556

[5] Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 785–794).

[6] Mangal, S., Chaurasia, A., & Khajanchi, A. (2020). Convolution neural networks for diagnosing colon and lung cancer histopathological images. arXiv preprint arXiv: 2009.03878.