IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Enhancing Lung Cancer Detection Using Deep Feature Learning and Comparative Model Evaluation

[1]Dr. Santosh Kumar Singh, [2]Mr. Mithilesh Vishwakarma, [3]Atharva Gadhave, [4]Kunal Gohil

[1]HOD, [2]Asst. Professor, [3,4]PG Student

[1,2,3,4]Department of Information Technology, Thakur College of Science and Commerce, Thakur Village, Kandivali (East), Mumbai-401107, Maharashtra, India.

Abstract : Lung cancer is one of the most fatal diseases worldwide, necessitating early and accurate diagnosis to improve survival rates. In this study, we present a comparative analysis of classification algorithms for enhanced lung cancer prediction using microscopic image analysis. We leverage deep learning-based feature extraction with MobileNetV2 to obtain high-dimensional feature representations from lung cancer images. These features are then reduced using t-SNE for better visualization and classification efficiency. Various machine learning models, including Ridge Classifier, Multi-Layer Perceptron (MLP), HistGradientBoosting, and Quadratic Discriminant Analysis (QDA), are evaluated for their predictive performance. Performance metrics such as accuracy, precision, recall, and F1-score are analyzed to determine the most effective model for lung cancer classification. The experimental results demonstrate that ensemble models and deep learning-based feature extraction significantly enhance prediction accuracy. This study provides valuable insights into optimizing classification techniques for medical image-based cancer diagnosis.

Keywords – MobileNetv2, T-sne, Deep Learning, Classifiers.

I. INTRODUCTION

Lung cancer remains one of the most fatal diseases globally, with millions of new cases reported each year. Early and accurate detection is crucial for improving survival rates, yet traditional diagnostic methods such as histopathological examination, CT scans, and biopsies are often time-consuming, expensive, and susceptible to human error. The advent of artificial intelligence (AI) and machine learning (ML) has provided promising solutions to enhance the accuracy and efficiency of cancer diagnosis, particularly through the analysis of microscopic images of lung tissue samples. Deep learning techniques, especially convolutional neural networks (CNNs), have demonstrated exceptional performance in medical image classification. However, due to their computational complexity, these models may not always be practical for clinical deployment. To address this challenge, this study employs MobileNetV2, a lightweight deep learning model, to extract essential features from microscopic lung cancer images. These features are then subjected to dimensionality reduction using t-distributed Stochastic Neighbor Embedding (t-SNE) to facilitate improved visualization and classification.

To comprehensively evaluate the effectiveness of deep learning-based feature extraction, this research compares multiple classification algorithms, including Ridge Classifier, Multi-Layer Perceptron (MLP), HistGradientBoosting (HGB), and Quadratic Discriminant Analysis (QDA). The primary objectives of this study are to analyze the effectiveness of deep learning in lung cancer classification, compare the performance of various machine learning algorithms, and evaluate the impact of dimensionality reduction on classification accuracy.

Furthermore, medical image analysis using AI-driven techniques faces challenges related to feature extraction, class imbalance, and interpretability. This research not only focuses on achieving high classification accuracy but also investigates how different algorithms handle these challenges to ensure reliable results. By utilizing multiple classifiers, this study provides a holistic view of lung cancer prediction performance across various methodologies. The findings of this research hold significant implications for AI-assisted diagnostic tools in medical applications, particularly in lung cancer detection. By automating the classification process, this approach has the potential to assist pathologists and oncologists in making faster and more precise clinical decisions, ultimately leading to improved patient outcomes. Furthermore, this study contributes to the ongoing advancements in AI-driven medical diagnostics by providing a comparative analysis of different classification techniques applied to microscopic image analysis. The remainder of this paper discusses related works in lung cancer classification, presents the methodology, details experimental results and evaluation metrics, and concludes with key findings, limitations, and future research directions. Through this study, we aim to enhance AI-based diagnostic capabilities, making lung cancer classification more accurate, efficient, and accessible in real-world clinical settings.

2.1 LITERATURE SURVEY

In this study, three discrimination models for subtypes of NSCLC were compared, with CapsNet demonstrating the best performance (81.3% accuracy) due to its ability to capture both global and local feature patterns. This highlights its potential for identifying NSCLC subtypes in histopathological images [1].

The objective of the study was to classify non-small cell lung tumors using texture-based analysis. The best results (75.48% accuracy) were obtained using SVM with HOG features, achieving high specificity and sensitivity, which shows the effectiveness of texture descriptors in lung cancer diagnosis [2].

In this paper, a computer-aided diagnosis (CAD) system was developed using homology-based image processing (HI) for histopathological image classification. The method was validated on two datasets, showing that HI outperformed conventional texture analysis techniques [3].

This study presented a hybrid model for lung and colon cancer detection, integrating preprocessing, k-fold cross-validation, feature extraction using transfer learning, and ensemble learning. The final ensemble voting classifier was selected based on the best-performing ML models [4].

The objective of this study was to design a CAD system to classify five types of colon and lung tissues using six ML models (XGBoost, SVM, RF, LDA, MLP, LightGBM) on the LC25000 dataset. The results demonstrated promising classification performance, supporting ML-based automated cancer diagnosis [5].

This study aimed to develop a deep learning approach for early lung and colon cancer detection. Three strategies were employed using ANN combined with GoogLeNet, VGG-19, CNN, and handcrafted features, demonstrating the effectiveness of feature fusion for improved classification [6].

In this paper, the ColonNet model was proposed alongside VGG, ResNet, DenseNet, Inception, and Xception, applying CNN-based fusion techniques for improved lung and colon cancer diagnosis. The model leveraged both deep learning and handcrafted features to enhance classification accuracy [7].

The study employed SHAP analysis to interpret the predictions of a transfer learning model for lung and colon cancer detection. The results provided insights into the decision-making process of the model by visualizing the impact of different image regions on classification outcomes [8].

This study proposed a multi-view CNN-based system for lung cancer detection from 3D CT scans. By incorporating multiple perspectives, the model improved robustness and accuracy in detecting lung nodules, even in ambiguous cases [9].

The objective of this study was to enhance lung cancer detection through multi-modality image fusion, combining CT and PET scans. This integration improved tumor localization and diagnostic accuracy by leveraging complementary imaging data [10].

2.2 DATA DESCRIPTION

The dataset utilized in this study comprises 6,000 labeled microscopic images of lung cancer cells, sourced from histopathological slides. The images are categorized into three distinct types of lung cancer, with an equal distribution of 2,000 images per category:

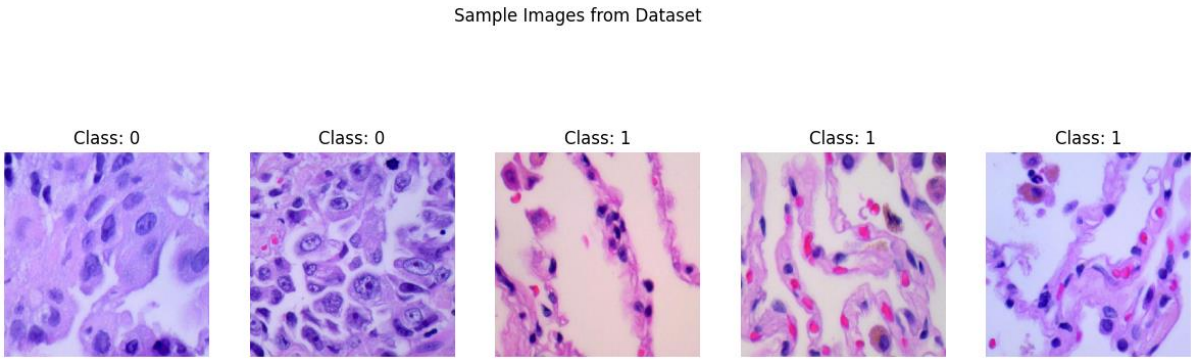
1. Adenocarcinoma: A common subtype of non-small cell lung cancer (NSCLC) originating in the mucus-producing glandular cells of the lungs.
2. Squamous Cell Carcinoma: Another NSCLC subtype, arising from the squamous epithelial cells lining the airways.
3. Neuroendocrine Tumors: A diverse group of lung cancers, including small cell lung cancer (SCLC) and large cell neuroendocrine carcinoma (LCNEC), known for their aggressive nature and rapid proliferation.

Each image in the dataset has a standardized resolution and was preprocessed to a fixed 224×224 pixel format to maintain consistency across classification models. The images exhibit variations in texture, morphology, and staining intensity, making the classification task more challenging and realistic. The dataset was split into training (80%) and testing (20%) sets to evaluate the performance of different classification algorithms accurately. Additionally, data augmentation techniques such as rotation, flipping, and contrast adjustment were applied to enhance model generalization and mitigate overfitting.

This dataset serves as a valuable benchmark for evaluating deep learning and machine learning techniques in lung cancer diagnosis, providing a robust foundation for automated histopathological image classification.

2.3 DATA PREPROCESSING

The preprocessing stage is crucial for ensuring that the input microscopic images are properly formatted and optimized for feature extraction and classification. In this study, several preprocessing steps were applied to the dataset using TensorFlow's ImageDataGenerator. First, all images were resized to a uniform dimension of 224 × 224 pixels to maintain consistency across the dataset. Additionally, pixel intensity values were rescaled to a range of [0,1][0,1][0,1] by dividing each pixel by 255, which helps in stabilizing and accelerating the training process. The dataset was then loaded and shuffled to ensure a balanced distribution of images during model training.

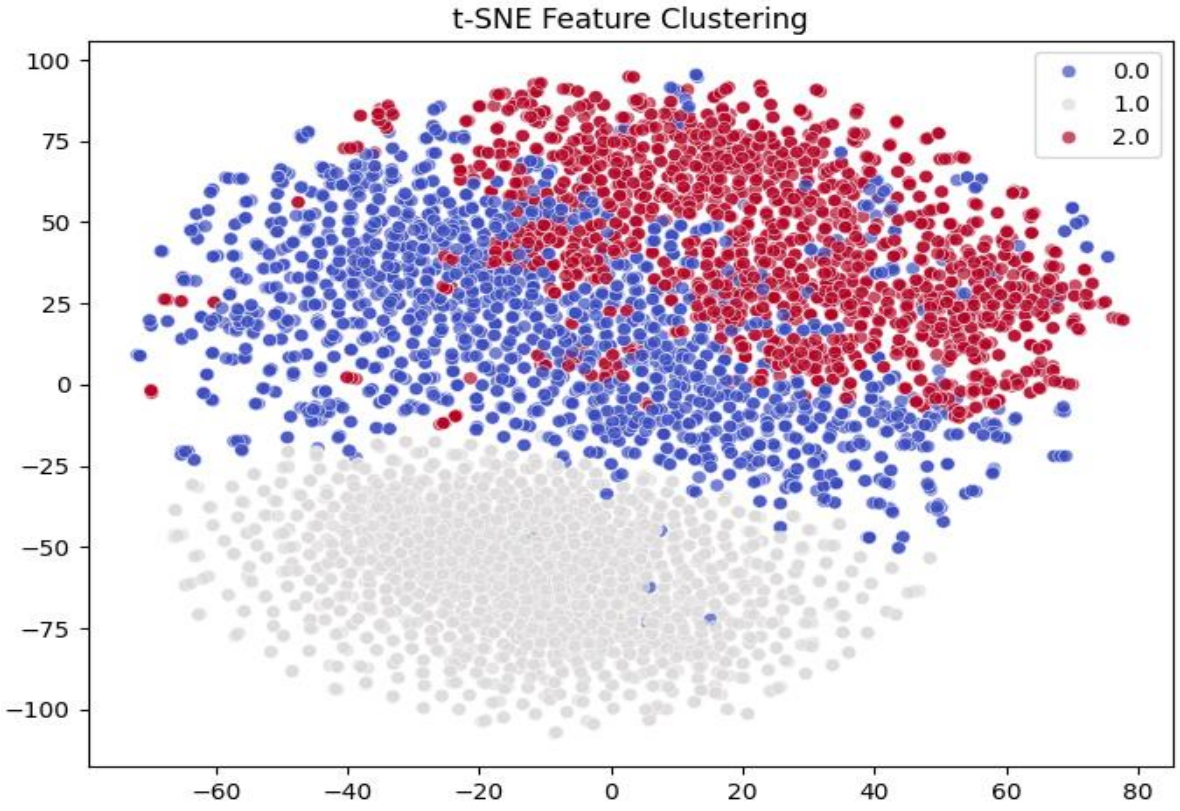


2.4 FEATURE EXTRACTION USING MOBILENETV2:

Feature extraction is a critical step in the classification of lung cancer cell images, as it enables the models to learn meaningful patterns from microscopic images. In this study, MobileNetV2, a lightweight convolutional neural network pre-trained on the ImageNet dataset, was used as a feature extractor. The top classification layer was removed (include_top=False), allowing the network to output high-dimensional feature representations rather than classification labels. Each image was passed through MobileNetV2, and the extracted feature maps were subsequently flattened into one-dimensional vectors to be used as inputs for classification models.

2.5 DIMENSIONALITY REDUCTION USING T-SNE :

The extracted features were processed using t-SNE (t-distributed Stochastic Neighbor Embedding), a technique that transforms high-dimensional data into a 2D representation while preserving local structures. This step facilitated the visualization of feature clustering, enabling a better understanding of the feature distribution across the three lung cancer types. The extracted feature vectors served as inputs for multiple classification algorithms, ensuring that the models leveraged deep, abstract representations rather than raw image pixels, leading to improved predictive performance.



2.6 CLASSIFICATION

After feature extraction and dimensionality reduction, multiple classification algorithms were implemented to evaluate their effectiveness in predicting lung cancer subtypes. The classifiers were trained and tested on the transformed features obtained from the t-SNE reduction of MobileNetV2-extracted features. The classification models used in this study include Ridge Classifier, Multi-layer Perceptron (MLP), Histogram-based Gradient Boosting (HGB), and Quadratic Discriminant Analysis (QDA).
These classifiers were evaluated using accuracy, precision, recall, and F1-score to assess their effectiveness in predicting lung cancer subtypes. The results demonstrate that deep learning-based feature extraction significantly enhances classification accuracy,

with ensemble and neural network-based models showing strong predictive performance. The following sections present a comparative analysis of these classifiers based on their evaluation metrics.

1. Ridge Classifier

Ridge Classifier is a linear classification model that applies L2 regularization to reduce overfitting. It is an extension of logistic regression but penalizes large coefficients to enhance generalization. The Ridge Classifier was employed as a baseline model to compare the performance of more complex classifiers. It provides fast computation and interpretable decision boundaries while handling correlated features effectively.

2. Multi-layer Perceptron (MLP)

MLP is a type of feedforward artificial neural network with one or more hidden layers. In this study, an MLP classifier with one hidden layer of 100 neurons was trained for 500 iterations using backpropagation. The model uses a ReLU activation function and an Adam optimizer, making it suitable for capturing non-linear relationships in the extracted features. MLP was included to evaluate the effectiveness of neural network-based models in lung cancer classification.

3. Histogram-based Gradient Boosting (HGB)

Histogram-based Gradient Boosting is a boosting algorithm that partitions continuous feature values into discrete bins, reducing memory consumption and improving training efficiency. Unlike traditional gradient boosting, HGB works well with large datasets and provides fast convergence. It was implemented as part of the study to compare its performance with other ensemble-based methods.

4. Quadratic Discriminant Analysis (QDA)

QDA is a generative classifier that models each class with a separate covariance matrix, allowing for more flexible decision boundaries compared to linear discriminant analysis (LDA). It assumes that features follow a Gaussian distribution, making it useful for datasets where class distributions vary significantly. QDA was included to analyze its effectiveness in distinguishing between lung cancer subtypes when applied to t-SNE-transformed features.

3.EVALUATION

The performance of the proposed classification models was assessed using multiple evaluation metrics, including accuracy, precision, recall, and F1-score. These metrics provide a comprehensive understanding of how well each classifier distinguishes between different lung cancer subtypes. Additionally, confusion matrices were utilized to visualize model performance and identify misclassifications. Confusion matrices were plotted for each classifier to analyse false positives and false negatives in the predictions.

- QDA and Ridge Classifier showed lower performance, indicating their limitations in handling complex image-derived feature distributions.
- MLP performed competitively, suggesting the potential of deep learning in lung cancer classification.

4. VISUALIZATION

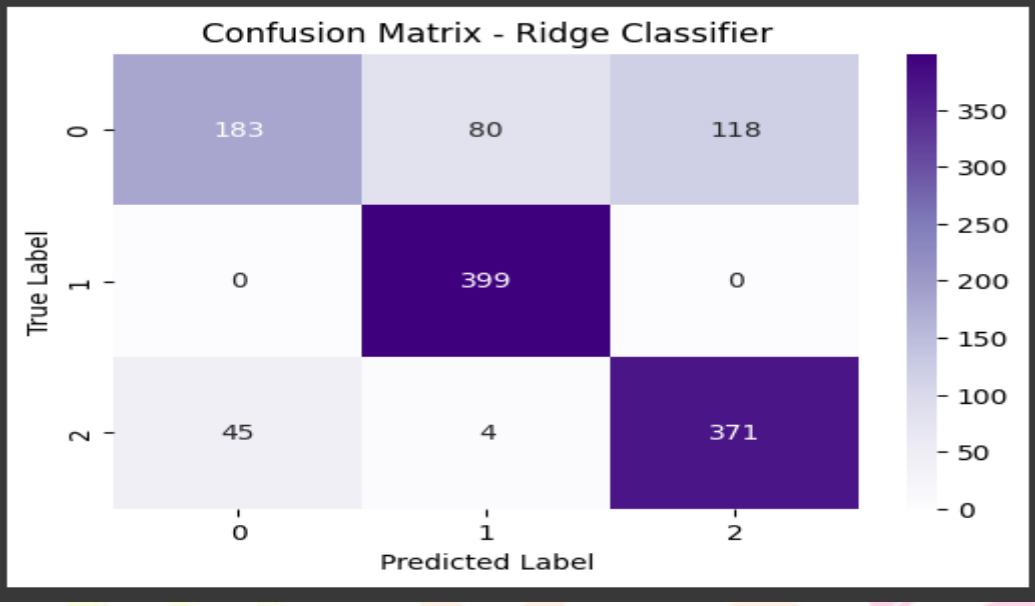
Sample images from each cancer type are displayed at the beginning of the project to provide a visual understanding of the dataset. Confusion matrices were utilized to provide a detailed view of correct and incorrect predictions, helping to identify misclassification patterns. Bar charts comparing accuracy, precision, recall, and F1-score enabled a comparative analysis of classifier performance. Heatmaps were used to intuitively represent classification results, enhancing interpretability. Additionally, t-SNE plots were generated to visualize feature separability in a reduced-dimensional space, demonstrating how well the extracted features were clustered. Furthermore, sample image predictions were displayed with their actual and predicted labels, along with confidence scores, providing insights into the model's decision-making process. These visualizations collectively contributed to a comprehensive evaluation of classification effectiveness and robustness.

5. RESULT

The performance of multiple classification models was assessed based on their accuracy in predicting lung cancer subtypes. The HistGradientBoosting (HGB) Classifier achieved the highest accuracy of 96.91%, demonstrating its strong ability to handle complex data distributions. The MLP Classifier achieved 90.83% accuracy, highlighting the effectiveness of neural networks in feature extraction but slightly underperforming compared to tree-based models. The QDA Classifier recorded an accuracy of 89.91%, indicating its sensitivity to feature distributions but lower generalization capability. The Ridge Classifier, with 79.41% accuracy, was the lowest-performing model, likely due to its linear decision boundaries being less effective for complex image-based classification tasks. These results emphasize the effectiveness of advanced gradient-boosting methods and neural networks in lung cancer classification

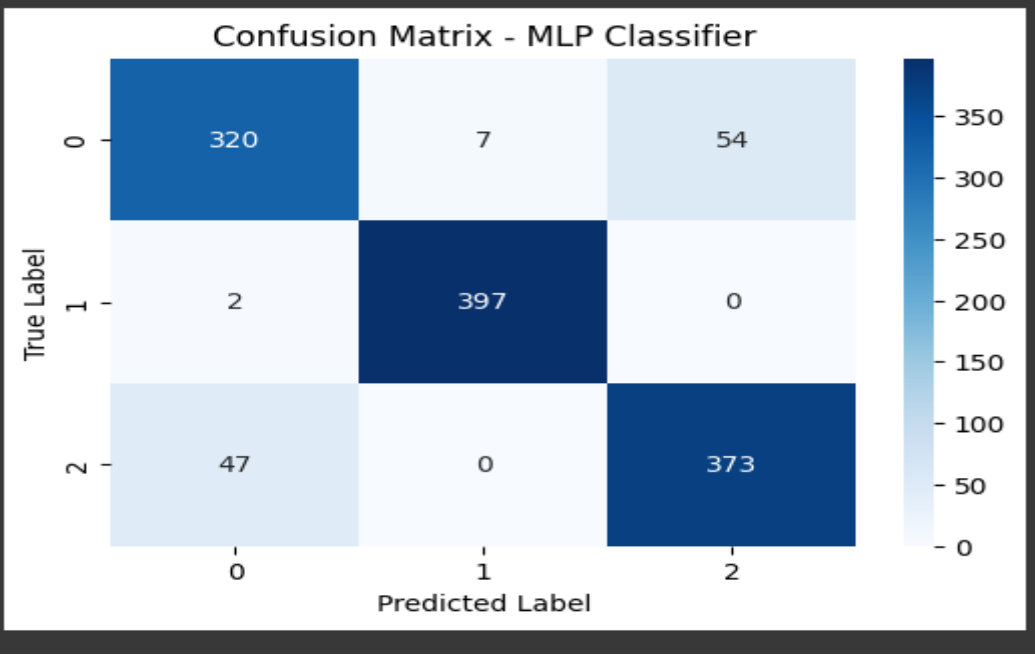
Ridge Classifier Accuracy: 0.7941666666666667

	precision	recall	f1-score	support
0.0	0.80	0.48	0.60	381
1.0	0.83	1.00	0.90	399
2.0	0.76	0.88	0.82	420
accuracy			0.79	1200
macro avg	0.80	0.79	0.77	1200
weighted avg	0.80	0.79	0.78	1200



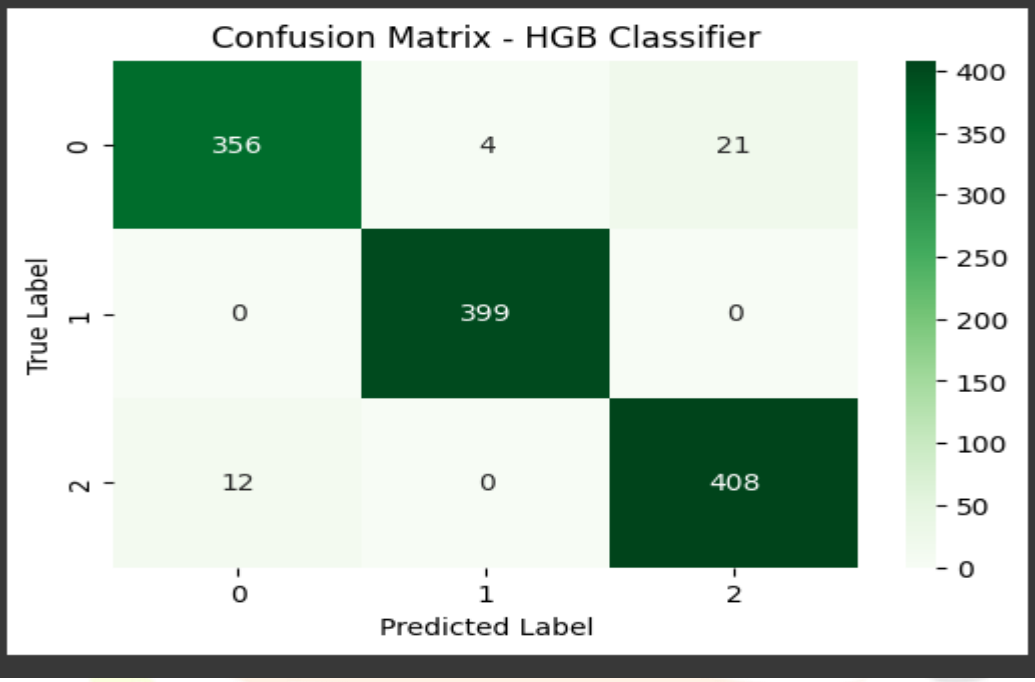
MLP Classifier Accuracy: 0.9083333333333333

	precision	recall	f1-score	support
0.0	0.87	0.84	0.85	381
1.0	0.98	0.99	0.99	399
2.0	0.87	0.89	0.88	420
accuracy			0.91	1200
macro avg	0.91	0.91	0.91	1200
weighted avg	0.91	0.91	0.91	1200



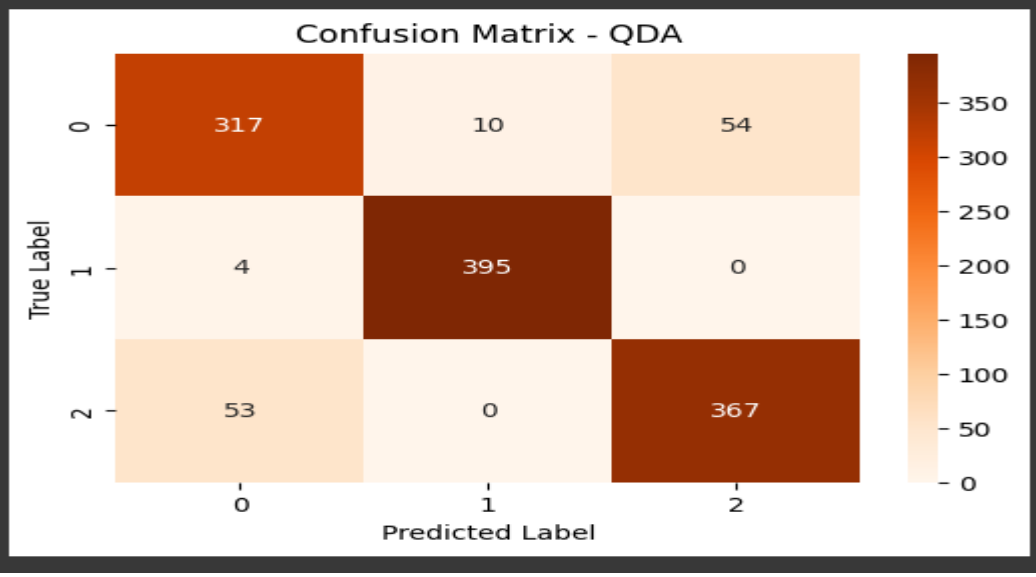
HGB Classifier Accuracy: 0.9691666666666666

	precision	recall	f1-score	support
0.0	0.97	0.93	0.95	381
1.0	0.99	1.00	1.00	399
2.0	0.95	0.97	0.96	420
accuracy			0.97	1200
macro avg	0.97	0.97	0.97	1200
weighted avg	0.97	0.97	0.97	1200



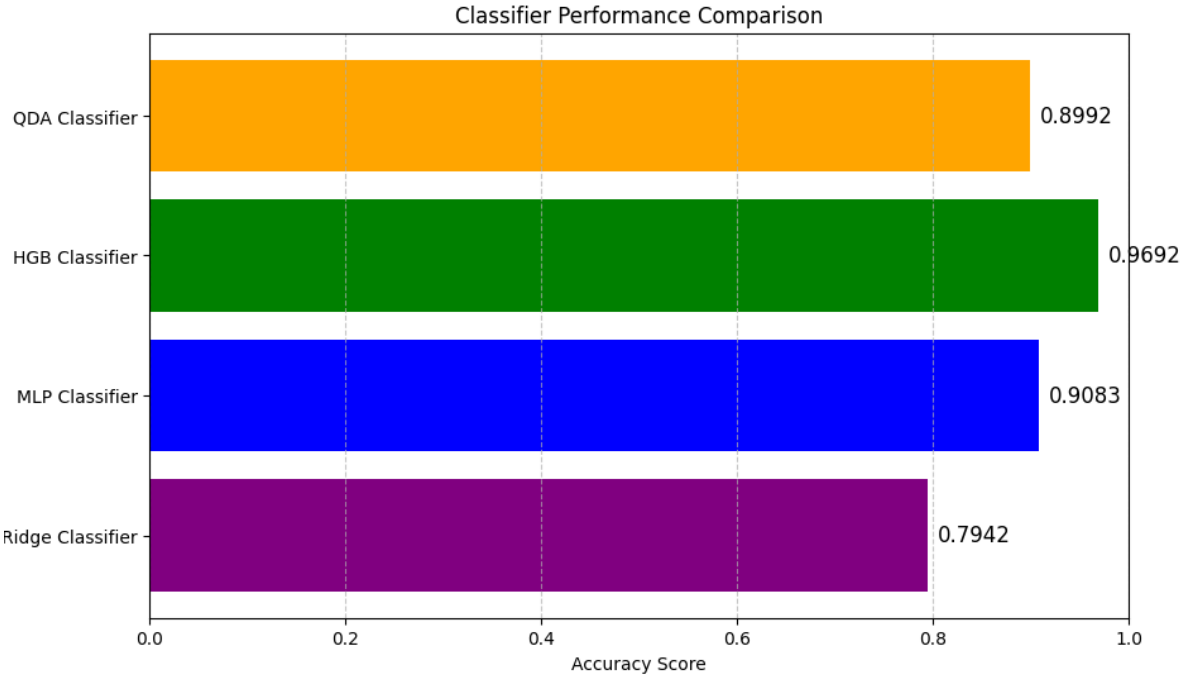
QDA Accuracy: 0.8991666666666667

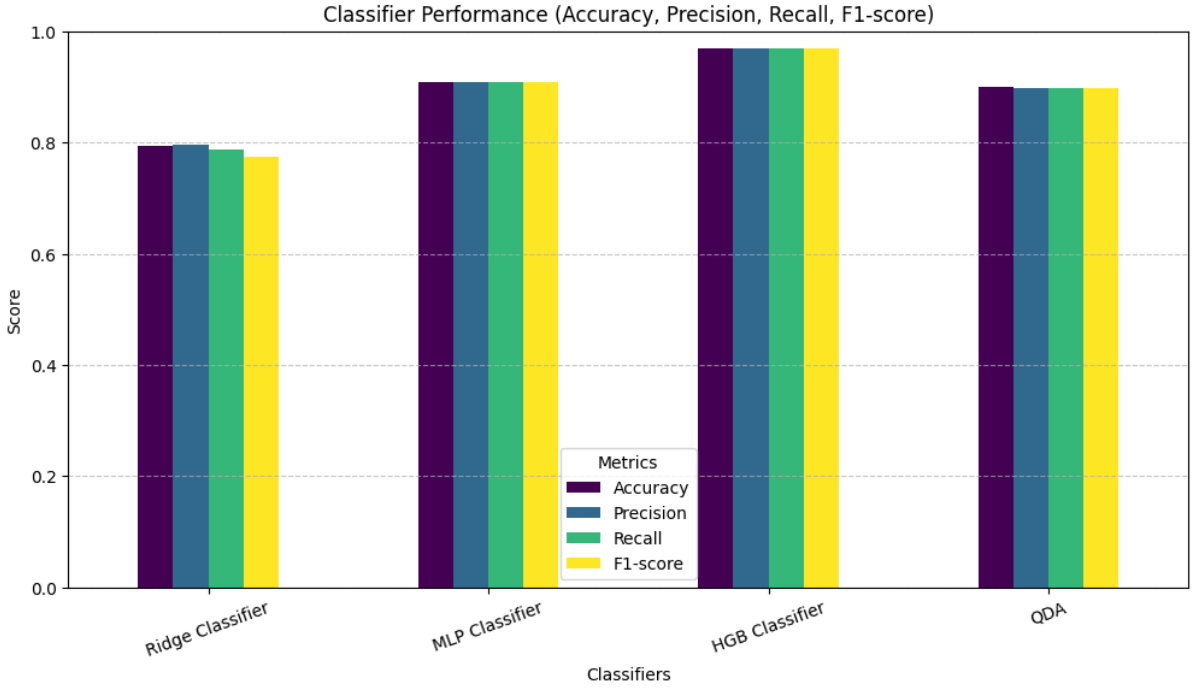
	precision	recall	f1-score	support
0.0	0.85	0.83	0.84	381
1.0	0.98	0.99	0.98	399
2.0	0.87	0.87	0.87	420
accuracy			0.90	1200
macro avg	0.90	0.90	0.90	1200
weighted avg	0.90	0.90	0.90	1200



Overall Model Performance:

Classifier	Accuracy (%)
Ridge Classifier	79.41
MLP Classifier	90.83
HGB Classifier	96.91
QDA classifier	89.91





6. CONCLUSION

This study evaluated the effectiveness of multiple classification algorithms for lung cancer prediction using microscopic image analysis. Among the tested models, the HistGradientBoosting (HGB) Classifier achieved the highest accuracy of 96.91%, demonstrating its robustness in handling complex image data. The MLP and QDA classifiers also performed well, with accuracies of 90.83% and 89.91%, respectively, highlighting the potential of neural networks and probabilistic models in cancer classification. The Ridge Classifier, with 79.41% accuracy, showed limitations in capturing intricate patterns within the dataset. These results indicate that advanced tree-based models and neural networks outperform linear classifiers in medical image classification tasks. The findings of this study emphasize the importance of selecting appropriate machine learning techniques for accurate and

automated lung cancer diagnosis. Future research can focus on integrating deep learning architectures and hybrid models to further enhance classification performance and clinical applicability.

7. REFERENCES

[1] Davri A, Birbas E, Kanavos T, Ntritsos G, Giannakeas N, Tzallas AT, Batistatou A. Deep learning for lung cancer diagnosis, prognosis and prediction using histological and cytological images: a systematic review. Cancers. 2023 Aug 5;15(15):3981.

[2] Bębas E, Borowska M, Derlatka M, Oczeretko E, Hładuński M, Szumowski P, Mojsak M. Machine-learning-based classification of the histological subtype of non-small-cell lung cancer using MRI texture analysis. Biomedical Signal Processing and Control. 2021 Apr 1;66:102446.

[3] Nishio M, Nishio M, Jimbo N, Nakane K. Homology-based image processing for automatic classification of histopathological images of lung tissue. Cancers. 2021 Mar 10;13(6):1192.

[4] Talukder MA, Islam MM, Uddin MA, Akhter A, Hasan KF, Moni MA. Machine learning-based lung and colon cancer detection using deep feature extraction and ensemble learning. Expert Systems with Applications. 2022 Nov 1;205:117695.

[5] Hage Chehade A, Abdallah N, Marion JM, Oueidat M, Chauvet P. Lung and colon cancer classification using medical imaging: A feature engineering approach. Physical and Engineering Sciences in Medicine. 2022 Sep;45(3):729-46.

[6] Al-Jabbar M, Alshahrani M, Senan EM, Ahmed IA. Histopathological analysis for detecting lung and colon cancer malignancies using hybrid systems with fused features. Bioengineering. 2023 Mar 21;10(3):383.

[7] Iqbal S, Qureshi AN, Alhussein M, Aurangzeb K, Kadry S. A novel Heteromorphous convolutional neural network for automated assessment of tumors in colon and lung histopathology images. Biomimetics. 2023 Aug 16;8(4):370.

[8] Alsubai S. Transfer learning based approach for lung and colon cancer detection using local binary pattern features and explainable artificial intelligence (AI) techniques. PeerJ Computer Science. 2024 Apr 19;10:e1996.

[9] Abhijeet Borle, et al., "Prediction and Classification of Lung Cancer using Machine Learning Algorithms," Journal of Intelligent & Fuzzy Systems, Vol. 40, No. 2, 2021, pp. 2133-2145.

[10] S. Hussain, et al., "Lung Cancer Detection using Image Processing Techniques: A Review," International Journal of Medical Informatics, Vol. 136, 2020, pp. 104069.