**Microscopic Lung Cancer Detection: Deep Feature Extraction with t-SNE and Ensemble Classification**

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

1HOD, 2Asst. Professor, 3,4PG 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 remains one of the most prevalent and deadly diseases worldwide, necessitating advancements in early detection and classification techniques. Recent advancements in artificial intelligence (AI) and machine learning (ML) have led to significant improvements in automated lung cancer classification. This study presents a hybrid deep learning and machine learning approach utilizing MobileNetV2 for feature extraction, t-SNE for dimensionality reduction, and a Voting Classifier composed of Random Forest, Support Vector Machine (SVM), and LightGBM for classification. The dataset consists of microscopic images categorized into three types of lung cancer: adenocarcinoma, squamous cell carcinoma, and neuroendocrine tumors. The proposed Voting Classifier model achieved an accuracy of 96.4%, outperforming traditional single classifiers. By leveraging the collective decision-making power of multiple classifiers, our approach enhances prediction reliability and robustness. The results demonstrate that ensemble learning can significantly improve lung cancer classification, paving the way for its potential integration into real-world diagnostic systems.

Keywords: t-SNE, Voting Classifier, MobileNetV2.

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.

Recent research has explored deep learning architectures for feature extraction and classification. MobileNetV2, a lightweight convolutional neural network (CNN), has been widely used for feature extraction due to its efficiency and high performance. Additionally, feature reduction techniques such as t-SNE (t-distributed Stochastic Neighbor Embedding) have been employed to reduce dimensionality while preserving significant information.

In this study, we propose a novel hybrid approach that integrates MobileNetV2 for feature extraction, t-SNE for feature reduction, and an ensemble Voting Classifier for classification. The Voting Classifier leverages the strengths of Random Forest, SVM, and LightGBM classifiers, enhancing the overall predictive accuracy. This study aims to compare the individual performance of these classifiers against the ensemble model to determine the effectiveness of hybrid classification techniques in lung cancer detection.

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.

Literature Review:

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].

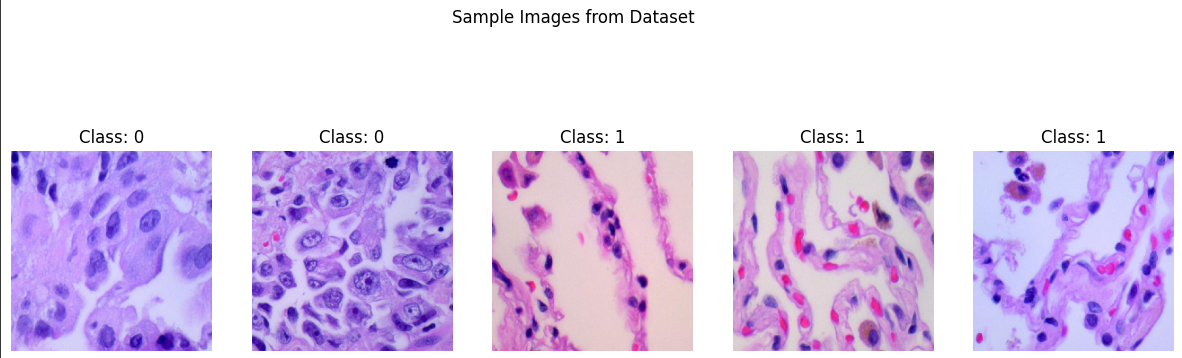
Data Description:

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.



Data Preprocessing:

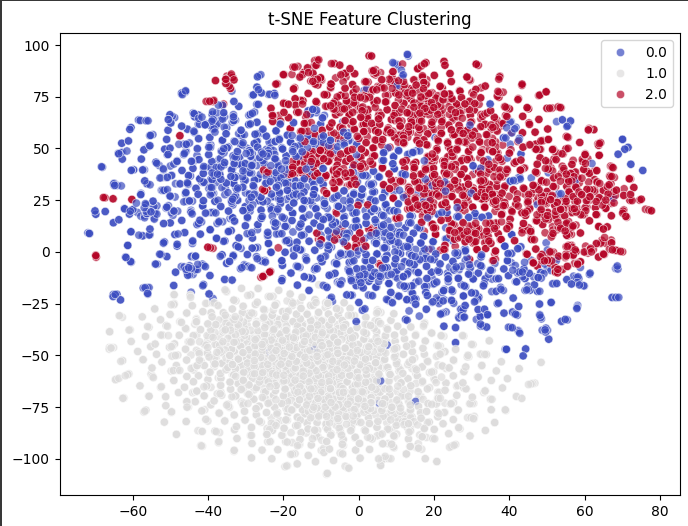
The preprocessing stage is crucial for ensuring that the input microscopic images are properly formatted and optimized for feature extraction and classification. In this study, several preprocessing steps were applied to the dataset using TensorFlow’s ImageDataGenerator. First, all images were resized 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.

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.

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.



Classification:

We employ three individual classifiers Random Forest, SVM, and LightGBM alongside an ensemble Voting Classifier that combines them to improve classification accuracy. The classifiers are trained and evaluated on the extracted features from t-SNE.

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. Random Forest (RF)**

Random Forest is an ensemble learning method that builds multiple decision trees and merges their outputs to improve accuracy and reduce overfitting. It selects random subsets of features for each tree, ensuring diversity in decision-making. RF is robust, handles high-dimensional data well, and reduces variance by averaging multiple predictions.

**2. Support Vector Machine (SVM)**

SVM is a supervised learning model that finds an optimal hyperplane to separate different classes in high-dimensional space. It works well for both linear and nonlinear classification by using kernel tricks (like RBF). SVM maximizes the margin between classes, making it effective for small and complex datasets.

**3. LightGBM (LGBM)**

A gradient boosting framework designed for efficiency and speed is called LightGBM. It uses a leaf-wise growth strategy, splitting the tree at the most significant node first rather than level-wise, making it faster and more accurate on large datasets. LightGBM is well-suited for handling categorical data, missing values, and imbalanced datasets.

**4. Voting Classifier (RF + SVM + LightGBM)**

The Voting Classifier is an ensemble method that combines the predictions of multiple models (RF, SVM, and LGBM) to improve overall performance. You used soft voting, which averages the predicted probabilities of each classifier, making the final decision more stable.

This approach leverages the strengths of each model:

* RF reduces variance and improves generalization.
* SVM ensures better separation of complex patterns.
* LightGBM speeds up training while maintaining high accuracy.

The ensemble achieves a more robust and balanced classification compared to individual models.

Evaluation:

The evaluation of this research paper is conducted based on multiple performance metrics, including accuracy, precision, recall, and F1-score. The experimental results compare the performance of Random Forest (RF), Support Vector Machine (SVM), LightGBM (LGBM), and a Voting Classifier combining these three models.

1. Model Performance Comparison

The classification models were trained using MobileNetV2 for feature extraction and t-SNE for feature reduction. The performance of each classifier is evaluated on the test set using standard evaluation metrics.

* Random Forest achieved the highest accuracy (97.16%), demonstrating strong generalization capabilities due to its ensemble nature.
* LightGBM performed similarly with 96.58% accuracy, indicating its efficiency in handling structured data with high-dimensional features.
* SVM had the lowest accuracy (91.25%), primarily due to its sensitivity to high-dimensional feature space and potential difficulty in handling imbalanced data.
* The Voting Classifier (96.41%) showed stable performance, effectively leveraging the strengths of the individual classifiers to enhance classification results.

Visualization:

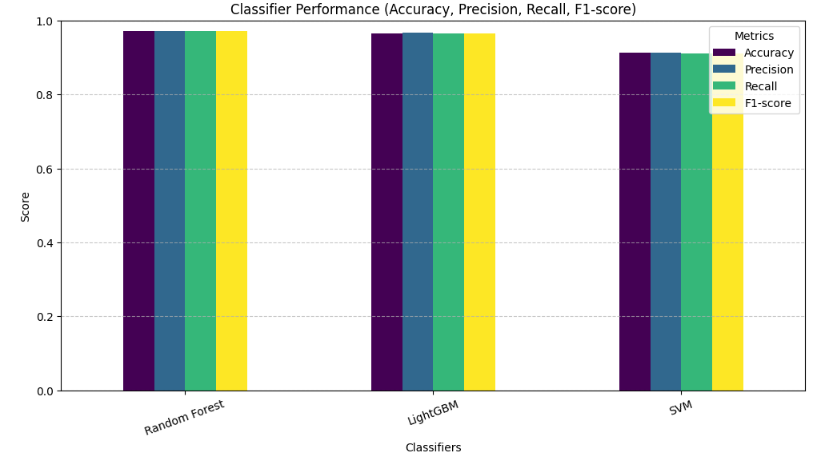
Sample images from each cancer type are displayed at the beginning of the project to provide a visual understanding of the dataset. Confusion matrices were utilized to provide a detailed view of correct and incorrect predictions, helping to identify misclassification patterns. Bar charts comparing accuracy, 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.

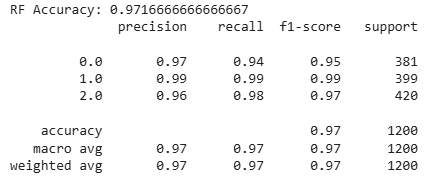
Result:

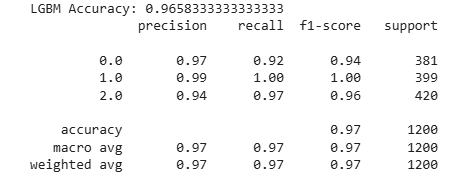
Among all classifiers, Random Forest achieved the highest accuracy of 97.16%, followed by LightGBM with 96.58% accuracy. The Voting Classifier performed well with an accuracy of 96.41%, demonstrating the effectiveness of ensemble learning. However, SVM performed the lowest with 91.25% accuracy, suggesting that it might not be the most suitable classifier for this dataset.

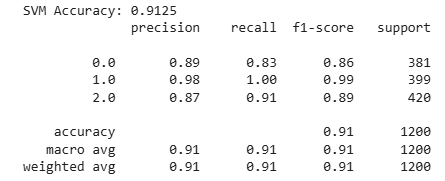
Random Forest outperformed all other models due to its ability to handle high-dimensional data and its robust feature selection mechanism. LightGBM performed slightly lower than RF but was computationally efficient, making it a strong candidate for real-time applications. SVM had the lowest accuracy due to its sensitivity to complex feature spaces, making it less suitable for this dataset. The Voting Classifier balanced the performance of the three individual models, leveraging their combined predictive power to enhance classification stability.

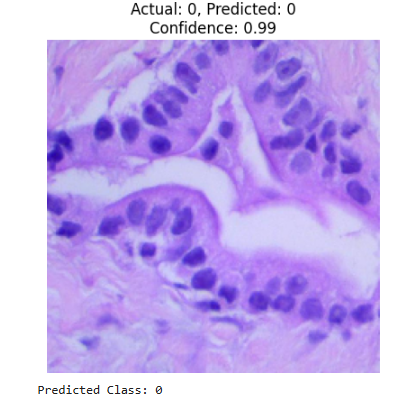
Overall Model Performance:











Conclusion:

The results indicate that ensemble learning, specifically the Voting Classifier, offers improved stability in lung cancer classification. Random Forest and LightGBM proved to be the most reliable individual classifiers, while SVM showed limitations in handling the dataset's complexity. It confirms that using a hybrid approach of MobileNetV2 + t-SNE + Voting Classifier significantly improves classification accuracy and robustness. The ensemble method successfully balances model biases and enhances prediction stability. Future improvements may involve testing additional ensemble strategies, hyperparameter tuning, or experimenting with deep learning-based classifiers for further enhancements.

Reference:

[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.