Image Segmentation of Lung Histopathological Image & Predicting Lung cancer using ML Algorithms

Aayush Kumar Singh
Student, SCOPE
VIT University
Chennai, India
aayushkumar.singh2019@vitstudent.ac.in

***Abstract* - Nowadays, cancer has counted as a hazardous disease that many people suffered from, especially Lung-Cancer. Cancer is a group of diseases involving abnormal cell growth with the potential to invade or spread to other parts of the body that is why treating it is somehow tough in some cases but it can be controlled if it is detected in the initial stage. Image Processing Mechanisms have a vital role in predicting and recognizing both benign and malignant cells with the help of classifier mechanisms such as Decision-Tree, SVM, and Naïve-Bayes classifier which are widely utilized in the biomedical field.**

This project aims to do image segmentation on lung Histopathological image and data analysis on the lung cancer dataset using machine learning algorithm.

***Keywords*–Decision-Tree, Image Processing, Lung cancer, Machine learning**

I. INTRODUCTION

Currently, the most hazardous disease that faced humanity's life and led to fatal death is referred to as Cancer. Among the various forms of cancer, the Lung-Cancer is enumerated as the riskiest one when compared to the other types around the globe.

The mortality rate of lung cancer is the highest among all other types of cancer. Lung cancer is one of the most serious cancers in the world, with the smallest survival rate after the diagnosis, with a gradual increase in the number of deaths every year.

Survival from lung cancer is directly related to its growth at its detection time. The earlier the detection is, the higher the chances of successful treatment are. An estimated 85% of lung Cancer cases in males and 75% in females are caused by cigarette smoking.

Early edge detection algorithms like Sobel, Prewitt and Laplacian have been used to segment the lung. However, none of them can successfully generate a truly satisfied segmentation output.

In order to overcome this challenge, this project outlines a network and training strategy that relies on the strong use of data augmentation. The architecture consists of a contracting path to capture context and a symmetric expanding path that enables precise localization.

II. LITERATURE SURVEY

Ragavan, M., & Patel [1] in "The evolving landscape of sex-based differences in lung cancer: a distinct disease in women" study aims to review the incidence rates of lung cancer in women which are now comparable to or higher than those in men and are rising alarmingly in many parts of the world. Women face a unique set of risk factors for lung cancer compared to men. These include exogenous exposures including radon, prior radiation, and fumes from indoor cooking materials such as coal, in addition to endogenous exposures such as oestrogen and distinct genetic polymorphisms. Women diagnosed with lung cancer have a clear mortality benefit compared to men even when other clinical and demographic characteristics are accounted for.

Tan, A. C., & Tan, D. S. [2] in "Targeted therapies for lung cancer patients with oncogenic driver molecular alterations" review encompasses the current standards of care for targeted therapies in lung cancer with driver molecular alterations. Targeted therapies for EGFR exon 19 deletion and L858R mutations, and ALK and ROS1 rearrangements are well established. However, there is an expanding list of approved targeted therapies including for BRAF V600E, EGFR exon 20 insertion, and KRAS G12C mutations, MET exon 14 alterations, and NTRK and RET rearrangements. In addition, there are numerous other oncogenic drivers, such as HER2 exon 20 insertion mutations, for which there are emerging efficacy data for targeted therapies. The importance of diagnostic molecular testing, intracranial efficacy of novel therapies, the optimal sequencing of therapies, role for targeted therapies in early-stage disease, and future directions for precision oncology approaches to understand tumor evolution and therapeutic resistance are also discussed.

Liao, Y., Wu, X., Wu, M., Fang, Y., Li, J., & Tang, W. [3] in "Non-coding RNAs in lung cancer: emerging regulators of angiogenesis" review discusses the regulatory functions of different ncRNAs in lung cancer angiogenesis, focusing on the downstream targets and signaling pathways regulated by these ncRNAs. Tumor metastasis and chemotherapeutic resistance lead to the poor prognosis of lung cancer. Nucleic acid therapeutics targeting angiogenesis, such as modified siRNAs and miRNA are promising therapeutics for lung cancer.

Merie, R., Gee, H., Hau, E., & Vinod, S. [4] in "An overview of the role of radiotherapy in the treatment of small cell lung cancer–a mainstay of treatment or a modality in decline?" study was to provide a comprehensive overview of the role and evidence of radiotherapy in the cure and palliation of SCLC. The search strategy included a search of the PubMed database, hand searches, reference lists of relevant review articles and relevant published abstracts. ClinicalTrials.gov was also queried for relevant trials. Thoracic radiotherapy improves overall survival in limited stage SCLC, but the timing and dose remain controversial. The role of thoracic radiotherapy in extensive stage SCLC with immunotherapy is the subject of several ongoing trials. Current evidence supports the use of prophylactic cranial irradiation (PCI) for limited stage SCLC but the evidence is equivocal in extensive stage SCLC. Whole brain radiotherapy is well established for the treatment of brain metastases but evidence is rapidly accumulating for the use of stereotactic radio surgery. Further studies will define the role of PCI, whole brain radiotherapy and hippocampal avoidant PCI in the immunotherapy era.

Hişam, D., & Hişam, E. [5] in "Deep learning models for classifying cancer and COVID-19 lung diseases." paper proposed different deep learning-based models such as DarkNet-53 (the backbone of YOLO-v3), ResNet50, and VGG19 that were applied to classify CT images of patients having Corona Virus disease (COVID-19) or lung cancer. The dataset used in the study came from two different sources, the large-scale CT dataset for lung cancer diagnoses (Lung-PET -CT-Dx) for lung cancer CT images while International COVID-19 Open Radiology Dataset (RICORD) for COVID-19 CT images. As a result, DarkNet-53 overperformed other models by achieving 100% accuracy. While the accuracies for ResNet and VGG19 were 80% and 77% respectively.

Rudin, C. M., Brambilla, E., Faivre-Finn, C., & Sage, J. [6] in "Small-cell lung cancer. Nature Reviews Disease Primers" paper proposed that small-cell lung cancer (SCLC) represents about 15% of all lung cancers and is marked by an exceptionally high proliferative rate, strong predilection for early metastasis and poor prognosis. SCLC is strongly associated with exposure to to-bacco carcinogens. Genomic profiling of SCLC reveals extensive chromosomal rearrangements and a high mutation burden, almost always including functional inactivation of the tumor suppressor genes TP53 and RB1. Although clinical progress in SCLC treatment has been notoriously slow, a better understanding of the biology of disease has un-covered novel vulnerabilities that might be amenable to targeted therapeutic approaches. The recent introduction of immune checkpoint blockade into the treatment of patients with SCLC is offering new hope, with a small subset of patients deriving prolonged benefit.

Mukherjee, S., & Bohra, S. U. [7] in "Lung Cancer Disease Diagnosis Using Machine Learning Approach." paper was to develop a lung cancer identification framework based on AI and deep neural system, wherein the methodology depends on supervised learning for which a better precision has been obtained, especially by using the deep learning mechanism. CNN classification is a game plan of lung tumor classification. The framework includes various methods, for instance, picture acquisition, pre-preparing, enhancement, segmentation, feature extraction, and neural framework identification. To put it concisely, machine learning approach can give an unprecedented opportunity to improve decision support in lung cancer treatment at low cost.

Jena, S. R., George, T., & Ponraj, N. [8] in "Texture analysis based feature extraction and classification of lung cancer." paper largely pact about prevailing lung cancer detection techniques that are obtainable in the literature. A numeral of methodologies has been originated in cancer detection methodologies to progress the efficiency of their detection. The early discovery of lung malignancy is a confront, because of the structure of tumor cells, where the greater part of the cells are covered with each other. Local Binary Pattern performs better than other basic textural patterns as the histogram features obtained were greater than that of the latter.

Barta, J. A., Powell, C. A., & Wisnivesky, J. P. [9] in "Global epidemiology of lung cancer." study was to review the evidence on lung cancer epidemiology, including data of international scope with comparisons of economically, socially, and biologically different patient groups. In industrialized nations, evolving social and cultural smoking patterns have led to rising or plateauing rates of lung cancer in women, lagging the long-declining smoking and cancer incidence rates in men. In contrast, emerging economies vary widely in smoking practices and cancer incidence but commonly also harbor risks from environmental exposures, particularly widespread air pollution. Recent research has also revealed clinical, radiologic, and pathologic correlates, leading to greater knowledge in molecular profiling and targeted therapeutics, as well as an emphasis on the rising incidence of adenocarcinoma histology. Furthermore, emergent evidence about the benefits of lung cancer screening has led to efforts to identify high-risk smokers and development of prediction tools. This study also discussed about the epidemiologic characteristics of special groups including women and non-smokers. Varying trends in smoking largely dictate international patterns in lung cancer incidence and mortality. With declining smoking rates in developed countries and knowledge gains made through molecular profiling of tumors, the emergence of new risk factors and disease features will lead to changes in the landscape of lung cancer epidemiology.

Ahmed, B. T. [10] in "Lung Cancer Prediction and Detection Using Image Processing Mechanisms: An Overview." study aims to review the most well-known Image Processing Mechanisms for Lung-Cancer Detection and Prediction. The comparison based on the Image Processing Mechanisms, accuracy, and classifier used in each reviewed research paper. Multi layer perceptron (MLP) gained a higher accuracy than the others followed by Logistic Regression (LR) and Decision Tree (D-Tree), which were (99.04%), (98.1%), and (93.62%), respectively. But, C4.5 obtained (86.7%) accuracy followed by the Genetic Algorithm that attained approximately (84.8%) accuracy.

Putora, P. M., Glatzer, M., Belderbos, J., Besse, B., Blackhall, F., Califano, R., & De Ruysscher, D. [11] in "Prophylactic cranial irradiation in stage IV small cell lung cancer" study obtained a list of 13 European experts from both the European Society for Therapeutic Radiation Oncology (ESTRO) and the International Association for the Study of Lung Cancer (IASLC). The strategies in decision making for PCI in stage IV SCLC were collected. Decision trees were created representing these strategies. Analysis of consensus was performed with the objective consensus methodology. The factors associated with the recommendation for the use of PCI included the fitness of the patient, young age and good response to chemotherapy. PCI was recommended by the majority of experts for non-elderly fit patients who had at least a partial response (PR) to chemotherapy (for complete remission (CR) 85% of radiation oncologists and 69% of medical oncologists, for PR: 85% of radiation oncologists and 54% of medical oncologists). For patients with stable disease after chemotherapy, PCI was recommended by 6 out of 13 (46%) radiation oncologists and only 3 out of 13 medical oncologists (23%). For elderly fit patients with CR, a majority recommended PCI (62%) and no consensus was reached for patients with PR.

Wong, D. M., Fang, C. Y., Chen, L. Y., Chiu, C. I., Chou, T. I., Wu, C. C., ... & Tang, K. T. [12] in "Development of a breath detection method based E-nose system for lung cancer identification." paper focused on the method of lung cancer identification by breath. The purpose of this breath detection system was to help physicians to quickly screen for rapid screening lung cancer. They used KNN and SVM with leave-one-out cross validation to analyze. PCA-KNN accuracy was 84.4%. The LDA-KNN accuracy was 75.5%. The PCA-SVM linear, polynomial, and rbf kernel type accuracy were 73.3%, 73.3%, and 73.3%, respectively. However, system achieved great results at about 84.4% accuracy for PCA-KNN classification.

Alam, J., Alam, S., & Hossan, A. [13] in "Multi-stage lung cancer detection and prediction using multi-class svm classifier." paper proposed an efficient lung cancer detection and prediction algorithm using multi-class SVM (Support Vector Machine) classifier. This system can also predict .the probability of lung cancer. For classification purpose, SVM binary classifier was used. The proposed algorithm gives a precision of 97% for cancer identification and 87% for cancer prediction. The proposed system would be viable in helping the doctor in recognizing the lung as harmful or non-carcinogenic.

Romaszko, A. M., & Doboszyńska, A. [14] in "Multiple primary lung cancer" paper was to discuss Multiple primary Lung cancer (MPLC) and find the difference between MPLC and intrapulmonary lung cancer metastasis in particular. Patients diagnosed with their 1st lung cancer should be carefully monitored in order to allow the early detection of a subsequent malignancy. Molecular method helps in the detection of lung cancer in future.

de Groot, P. M., Wu, C. C., Carter, B. W., & Munden, R. F. [15] in "The epidemiology of lung cancer." paper aims to find the incidence, possible causes, distribution and possible control of lung cancer in the world. The incidence and mortality from lung cancer is decreasing in the US due to decades of public education and tobacco control policies, but are increasing elsewhere in the world related to the commencement of the tobacco epidemic in various countries and populations in the developing world. Individual cigarette smoking is by far the most common risk factor for lung carcinoma; other risks include passive smoke inhalation, residential radon, occupational exposures, infection and genetic susceptibility. The predominant disease burden currently falls on minority populations and socioeconomically disadvantaged people. In the US, the recent legalization of marijuana for recreational use in many states and the rapid growth of commercially available electronic nicotine delivery systems (ENDS) present challenges to public health for which little short term and no long term safety data is available.

III. HISTOPATHOLOGICAL IMAGE DATASET

The dataset that I used comprises 5000 images for each of the three classes- benign lung tissue, lung adenocarcinomas and lung squamous cell carcinomas. I use an equal number of images for each class to avoid the problem of class imbalance. The images were generated from an original sample of HIPAA compliant and validated sources. All images are 768 x 768 pixels in size and are in jpeg file format. An image of each class obtained from the dataset is shown in the FIG I.

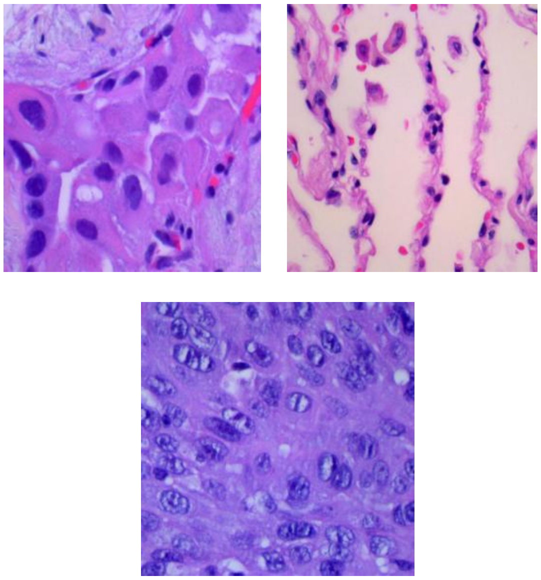

FIG I: ADENOCARCINOMAS CELL CARCINOMA, BENIGN LUNG TISSUE AND LUNG SQUAMOUS CELL CARCINOMA

IV. PROPOSED METHOD

The proposed method in this research paper involved using various machine learning algorithms and deep learning models for image segmentation of lung histopathological images and predicting lung cancer. The algorithmic models used in this study were SVM, random forest, KNN, logistic regression, naive Bayes, CatBoost, XGBoost, and decision tree. Additionally, deep neural network models such as ANN, CNN, Deep-CNN,Inception-V3, EfficientNet-B7, ResNet50, and VGG16 were also used.

The first step of the proposed model starts with data preprocessing in which all images were converted into a standard size of 768x768. By resizing the images, we can decrease the training time of our model and reduces the memory required for the training purpose. Good thing about having a small size image data is that lot of images can be fed into the model for training without exhausting the memory or increasing the training time. It is a good trade-off between the amount of pixel data in one image and count of images that can be used for training in a limited computational environment.

Data analysis process was carried using Jupyter platform in python tool. Jupyter is an open-source web application that allows us to create and share documents that contain live code, equations, visualizations and narrative text. Its uses include data cleaning and transformation, numerical simulation, statistical modeling, data visualization and machine learning.

Before splitting the dataset into training data and testing data, it is important to preprocess the categorical variables, which represent non-numeric data such as colors, labels, or categories. One popular technique for preprocessing categorical variables is one-hot encoding, which transforms each categorical variable into a set of binary variables that indicate the presence or absence of a particular category.

After one-encoding step, the dataset is split into 80% training data and 20% testing data. Now these images are provided to various ML algorithms such as logistic regression, KNN, SVM, Naïve Bayes, random forest, decision tree, cat boost and XGB classifier and deep neural network architecture such as CNN, ANN, ResNet50 , VGG16, Inception, Efficient Net. Deep neural networks have proven to yield better accuracy when dealing with large volumes of dataset, and many researchers tend to use them as de-facto standards. A typical architecture of neural network consists of multiple blocks with three kinds of layers: convolution, pooling, and fully connected layers.

After the execution of above step, It displays accuracy of prediction whether the input lung image contains cancer or not. This information can be used to aid in diagnosis, screening, and treatment planning.

Future directions for the research could involve exploring the use of other imaging modalities for lung cancer detection, such as magnetic resonance imaging (MRI), computed tomography (CT), or positron emission tomography (PET). These imaging techniques offer unique advantages and may provide complementary information to histopathological images. Additionally, the development of new deep neural network models or the refinement of existing models could lead to improved accuracy and efficiency in lung cancer detection and diagnosis. Finally, clinical validation studies and the incorporation of the proposed ML algorithms into clinical decision support systems could further validate the feasibility and utility of this approach in real-world clinical settings.

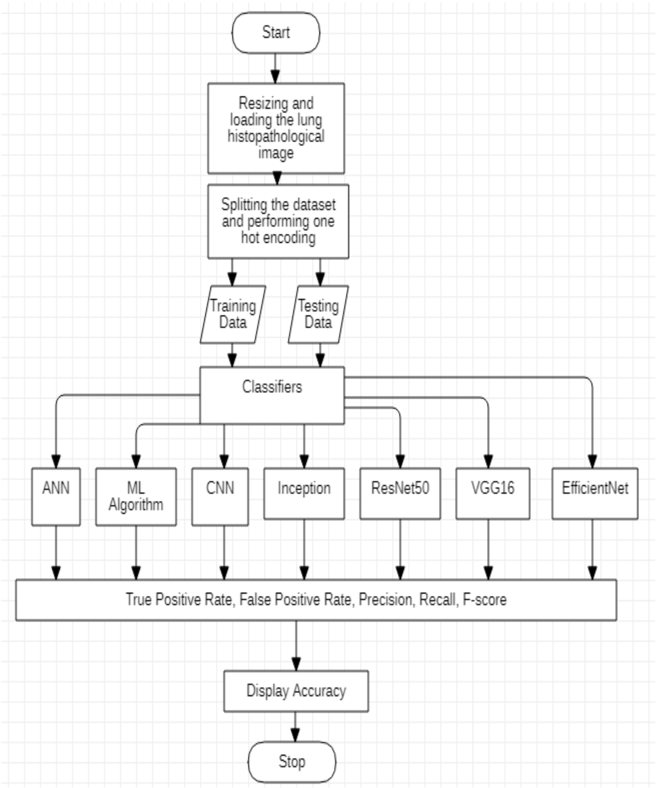
Start
Resizing and loading the lung histopathological image
Splitting the dataset and performing one hot encoding
Training Data
Testing Data
Classifiers
ANN
ML Algorithm
CNN
Inception
ResNet50
VGG16
EfficientNet
True Positive Rate, False Positive Rate, Precision, Recall, F-score
Display Accuracy
Stop

FIG II: BLOCK DIAGRAM OF THE MODEL

V. ALGORITHMS

Machine learning algorithms and deep neural network algorithms are used to display the accuracy of lung tissue classification. The algorithms are defined briefly.

1. Support Vector Machine

SVM [16] stands for Support Vector Machine, which is a supervised machine learning algorithm used for classification, regression, and outlier detection. SVM is a powerful algorithm that can handle both linear and non-linear data and is widely used in pattern recognition, image classification, text classification, and bioinformatics, among other applications. SVM works by finding a hyper plane in a high-dimensional space that best separates the different classes. SVM can also handle non-linear data by transforming the input features into a higher-dimensional space, where the data can be separated by a hyper plane.

$$w^T \; x + b=0 \tag{1}$$

where x is a vector in the input space, w is the weight vector and b is the bias.

2. K-Nearest Neighbor Classifier

The k-nearest neighbors algorithm [17], also known as KNN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. While it can be used for either regression or classification problems, it is typically used as a classification algorithm, working off the assumption that similar points can be found near one another.

$$\mathrm{dist}(\mathbf{x}, \mathbf{x}') \geq \max_{(\mathbf{x}'', y'') \in S_{\mathbf{x}}} \mathrm{dist}(\mathbf{x}, \mathbf{x}''), \tag{2}$$

where x is the test point, Sx is the set of the k nearest neighbors.

$$h(\mathbf{x}) = \mathrm{mode}(\{y'' : (\mathbf{x}'', y'') \in S_{\mathbf{x}}\}), \tag{3}$$

Where h() is the function returning the most common label in Sx and mod() means to select the label of the highest occurrence.

3. Logistic Regression

Logistic regression [18] is a statistical method that is used for building machine learning models where the dependent variable is dichotomous (Binary). Logistic regression is used to describe data and the relationship between one dependent variable and one or more independent variables. In healthcare institutions, logistic regression can accurately target at risk individuals who should receive a more tailored behavioral health plan to help improve their daily health habits. This in turn opens the opportunity for better health for patients and lower costs for hospitals.

$$g(E(y)) = \alpha + \beta x1 + \gamma x2 \tag{4}$$

where g() is the link function, E(y) is the expectation of target variable and $\alpha + \beta x1 + \gamma x2$ is the link predictor(α ,β ,γ to be predicted).

4. Naïve Bayes

Naive Bayes [19] is based on Bayes' theorem, which is a formula that describes the probability of an event based on prior knowledge of conditions that might be related to the event. Naive Bayes can be used for both binary and multi-class classification problems, and it is particularly well-suited for problems with a large number of input features. It is also relatively fast to train and make predictions, and it can work well even with small amounts of training data.

$$P(A|B) \;= P(B/A)*P(A)/P(B) \tag{5}$$

Where P(A|B) is posterior probability, P(B|A) is the likelihood of A given a fixed B, P(A) is the probability of A and P(B) is the probability of B.

5. CatBoost

CatBoost [20] or Categorical Boosting is an algorithm for gradient boosting on decision trees developed by yandex. Gradient Boosting is an ensemble machine learning algorithm which is typically used for solving classification and regression problems. It essentially creates a strong learner from an ensemble of many weak learners. It works well with heterogeneous data. In addition to regression and classification, CatBoost can be used in ranking, recommendation systems, forecasting and even personal assistants.

$$f(x) = b0 + \Sigma b(i)t(x) \tag{6}$$

where f(x) is the prediction of the model for the input data point x, bo is the base prediction (usually the mean or median of the target variable), b(i) is the prediction of the
i-th tree in the model and t(x) is the terminal node where x falss in the i-th tree.

6. Random Forest

Random forest [21] is a classifier that contains several decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of the dataset. It is based on the concept of ensemble learning which is a process of combining multiple classifiers to solve a complex problem and improve the performance of the model. In healthcare, random forest open up many possibilities for early diagnosis that are not only cheaper than neural network, but also solve the ethical problem (decision making problem) associated with neural network.

Classification,

$$f(x) = mode(T1(x), T2(x), ..., Tn(x)) \tag{7}$$

where f(x) is the predicted class label for x, and mode() is the function that returns the most common class label among the decision trees.

Regression,

$$f(x) = (1/n)\Sigma(Ti(x)) \tag{8}$$

where f(x) is the predicted numerical value for x, and Σ() is the summation function.

7. XGBoost classifier

XGBoost [22] stands for Extreme Gradient Boosting is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. The library is parallelizable which means the core algorithm can run on clusters of GPUs or even across a network of computers. This makes it feasible to solve ML tasks by training on hundreds of millions of training examples with high performance. It provides parallel tree boosting and is the leading machine learning library for regression, classification and ranking problems.

$$f(x) = \Sigma w(i)t(i)(x) \tag{9}$$

where f(x) is the predicted value for the input data point x, w(i) is the weight of the i-th tree in the model, t(i)(x) is the predicted value of the i-th tree for the input data point x.

8. Decision Tree

A decision tree [23] algorithm is a machine learning algorithm that is used for classification and regression tasks. It works by building a decision tree model that represents a flowchart-like structure of decisions and their possible consequences. The decision tree algorithm is a supervised learning algorithm, which means that it requires a labeled dataset for training.

Classification,

$$f(x) = argmax(\Sigma wiI(xi,j = c)fj) \tag{10}$$

where f(x) is the predicted class label for the input data point x, wi is the weight associated with the i-th class, I() is the indicator function that returns 1 if the condition is true and 0 otherwise, xi,j is the j-th feature of the input data point x, c is the value of the j-th feature that the node tests, and fj is the value associated with the i-th class label at the leaf node.

Regression,

$$f(x) = \Sigma wiI(x <= tj)fj \tag{11}$$

where f(x) is the predicted numerical value for the input data point x, wi is the weight associated with the i-th leaf node, I() is the indicator function that returns 1 if the condition is true and 0 otherwise, x is the input data point, tj is the threshold value for the j-th feature at the node, and fj is the value associated with the i-th leaf node.

9. Artificial Neural Network (ANN)

An Artificial Neural Network (ANN) [24] is a computational model inspired by the biological neural networks of the human brain. ANNs are used to solve complex problems that are difficult or impossible to solve using traditional computing techniques. ANNs consist of a large number of interconnected processing nodes or artificial neurons that work together to perform a specific task.

In image segmentation, ANNs are used to partition an image into regions or segments based on their visual characteristics, such as color, texture, and shape. ANNs are trained on large datasets of labeled images, where the desired output for each input image is a binary mask indicating the segmentation boundaries. The network learns to map input image features to corresponding output masks through a process of iterative optimization, typically using backpropagation and gradient descent algorithms. Image of the model is shown in FIG III.

In summary, ANNs have become a powerful tool for image segmentation tasks, enabling automated and accurate segmentation of complex images. They offer a promising approach for a range of applications in medical imaging, remote sensing, robotics, and more.

The computation of the output of a neuron i in layer j is given by,

$$z(i,j) = \Sigma w(i,k)x(k,j-1) + b(i,j) \tag{12}$$

where z(i,j) is the weighted sum of the inputs to neuron i in layer j, w(i,k) is the weight of the connection between neuron i in layer j and neuron k in layer j-1, x(k,j-1) is the output of neuron k in layer j-1, and b(i,j) is the bias term of neuron i in layer j.

The gradient of the loss function with respect to the weight w(i,k) is given by,

$$\partial L/\partial w(i,k) = \partial L/\partial z(i,j) * \partial z(i,j)/\partial w(i,k) \tag{13}$$

where ∂L/∂z(i,j) is the derivative of the loss function with respect to the weighted sum z(i,j), and ∂z(i,j)/∂w(i,k) is the derivative of the weighted sum z(i,j) with respect to the weight w(i,k).

The gradient of the loss function with respect to the bias b(i,j) is given by,

$$\partial\ \partial L/\partial b(i,j) = \partial L/\partial z(i,j) * \partial z(i,j)/\partial b(i,j) \tag{14}$$

where ∂L/∂z(i,j) is the derivative of the loss function with respect to the weighted sum z(i,j), and ∂z(i,j)/∂b(i,j) is the derivative of the weighted sum z(i,j) with respect to the bias b(i,j).

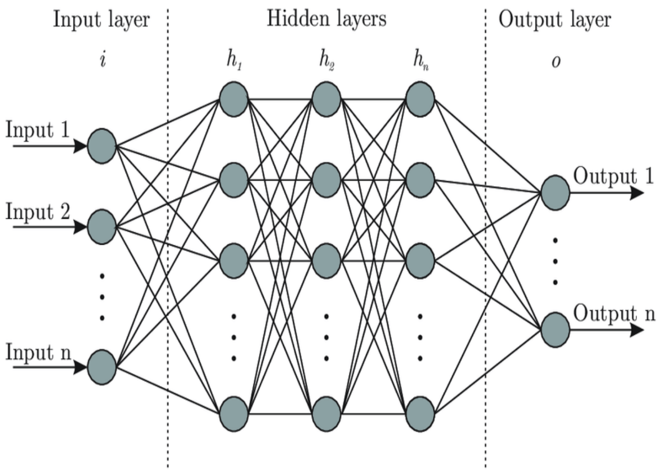

FIG III: ANN MODEL

10. Convolutional Neural Network

A Convolutional Neural Network (CNN) [25] is a type of deep learning algorithm that is particularly well-suited for image recognition and processing tasks. It is made up of multiple layers, including convolutional layers, pooling layers, and fully connected layers.

In CNN-based image segmentation, the input image is processed by a series of convolutional layers, each of which extracts a set of features from the input image. These features are then passed through additional layers such as pooling and activation layers to reduce the dimensionality of the feature map and enhance their representation. The output of the last layer of the CNN is then fed into a segmentation layer, which produces a binary mask indicating the segmentation of the image. The segmentation layer typically employs either a softmax activation function or a sigmoid activation function to produce a probability map, which is then thresholded to obtain the final binary mask. Image of the model is shown in FIG IV.

CNN-based image segmentation has shown to be highly effective for a range of applications, including medical image analysis, remote sensing, and autonomous driving. It has enabled automated and accurate segmentation of complex structures and regions in images that were previously difficult to segment manually.

Convolution,

$$feature\_map(i,j) = \Sigma\Sigma\Sigma\ filter(m,n,k) * input\_image(i+m-1,j+n-1,k) \tag{15}$$

where feature_map(i,j) is the value of the feature map at position (i,j), filter(m,n,k) is the value of the filter at position (m,n,k), and input_image(i+m-1,j+n-1,k) is the value of the input image at position (i+m-1,j+n-1,k).

Pooling,

$$max\_pooled\_feature\_map(i,j)= max(feature\_map(i+p,j+q)) \tag{16}$$

where max_pooled_feature_map(i,j) is the value of the max-pooled feature map at position (i,j), feature_map(i+p,j+q) is the value of the feature map within the sliding window centered at position (i,j), and the size of the sliding window is specified by the pooling size.

Activation Function,

$$output(i,j) = ReLU(max\_pooled\_feature\_map(i,j)) \tag{17}$$

where output(i,j) is the output of the layer at position (i,j) and Rectified Linear Unit (ReLU), which sets all negative values to zero and leaves positive values unchanged.

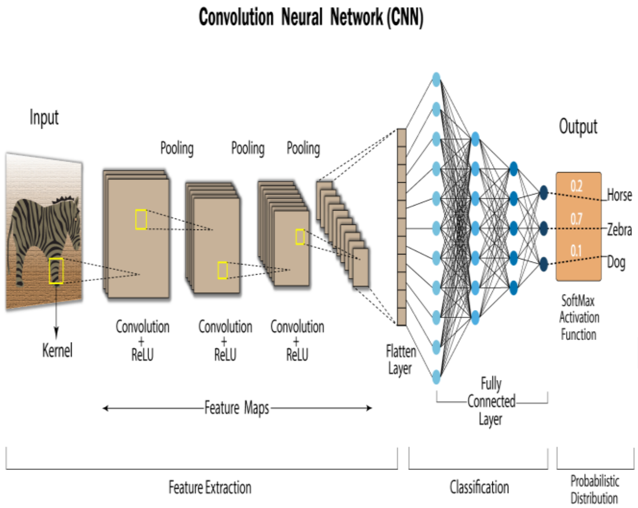

FIG IV: CNN MODEL

11. Inception

An inception network [26] is a deep neural network with an architectural design that consists of repeating components referred to as Inception modules. Image of the model is shown in FIG V.

In image segmentation, Inception has been used to identify regions of interest within an image by segmenting it into multiple regions. The network is able to identify different features and textures within the image, allowing it to distinguish between different objects and identify boundaries between them.

Inception has also been used in medical image segmentation applications, such as segmenting the liver from CT scans or the brain from MRI scans. By using Inception to accurately identify the edges of organs or structures within an image, medical professionals can more easily analyze and diagnose potential issues.

1x1 Convolution,

$$output\_1x1 = conv\_1x1(input, weights) \tag{18}$$

where output_1x1 is the output feature map of the 1x1 convolution, input is the input feature map, conv_1x1 is the convolution operation with 1x1 filters, and weights are the learnable parameters of the convolution.

3x3 Convolution,

$$output\_3x3 = conv\_3x3(input, weights) \tag{19}$$

where output_3x3 is the output feature map of the 3x3 convolution, input is the input feature map, conv_3x3 is the convolution operation with 3x3 filters, and weights are the learnable parameters of the convolution.

5x5 Convolution,

$$output\_5x5 = conv\_5x5(input, weights) \tag{20}$$

where output_5x5 is the output feature map of the 5x5 convolution, input is the input feature map, conv_5x5 is the convolution operation with 5x5 filters, and weights are the learnable parameters of the convolution.

Max Pooling,

$$output\_maxpool = max\_pool(input) \tag{21}$$

where where output_maxpool is the output feature map of the max pooling operation, and input is the input feature map.

Concatenation,

$$output\_inception = concatenate([output\_1x1, output\_3x3, output\_5x5, output\_maxpool]) \tag{22}$$

where output_inception is the output feature map of the inception module, and concatenate is the operation that concatenates the input feature maps along the channel dimension.

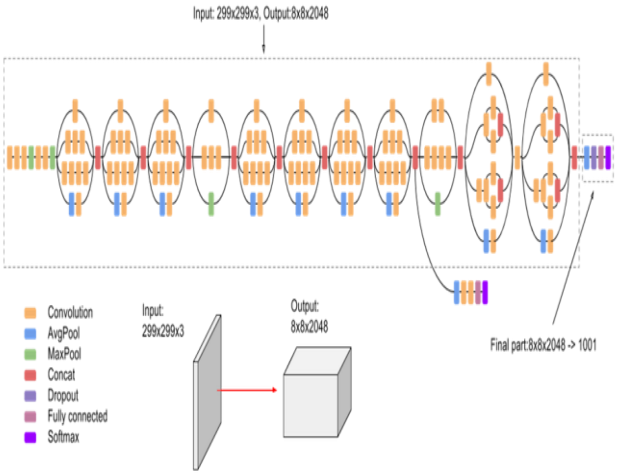

FIG V: INCEPTIONV3 MODEL

12. ResNet50

ResNet [27] stands for Residual Network and is a specific type of convolutional neural network (CNN) introduced in the 2015 paper “Deep Residual Learning for Image Recognition” by He Kaiming, Zhang Xiangyu, Ren Shaoqing, and Sun Jian. CNNs are commonly used to power computer vision applications.

ResNet-50 is a 50-layer convolutional neural network (48 convolutional layers, one MaxPool layer, and one average pool layer). Residual neural networks are a type of artificial neural network (ANN) that forms networks by stacking residual blocks.

In image segmentation, ResNet50 is used to extract features from input images and then perform segmentation based on those features. This involves training the network on a large dataset of labeled images and using the learned features to accurately segment new images. Image of the model is shown in FIG VI.

One advantage of ResNet50 is its ability to learn hierarchical features, which allows for better segmentation of complex images with multiple objects and overlapping structures. Additionally, the residual connections in the network help to alleviate the vanishing gradient problem, allowing for more efficient training of deep networks.

Overall, ResNet50 is a powerful tool for image segmentation, particularly in medical imaging applications where accuracy is critical. Its ability to learn hierarchical features and overcome the challenges of training deep networks make it an effective choice for complex segmentation tasks.

First Convolution,

$$\mathit{output\_conv1} = \mathit{conv}(\mathit{input}, \mathit{weights1}) \tag{23}$$

where output_conv1 is the output feature map of the first convolutional layer, input is the input feature map, conv is the convolution operation with learnable weights1, and weights1 are the learnable parameters of the convolution.

Second Convolution,

$$\mathit{output\_conv2} = \mathit{conv}(\mathit{output\_conv1}, \mathit{weights2}) \tag{24}$$

where output_conv2 is the output feature map of the second convolutional layer, output_conv1 is the input feature map, conv is the convolution operation with learnable weights2, and weights2 are the learnable parameters of the convolution.

Shortcut connection,

$$\mathit{output\_shortcut} = \mathit{input} + \mathit{output\_conv2} \tag{25}$$

where output_shortcut is the output of the shortcut connection, and input is the input feature map.

Activation,

$$\mathit{output\_resblock} = \mathit{activation}(\mathit{output\_shortcut}) \tag{26}$$

where output_resblock is the output of the residual block, and activation is the activation function, such as ReLU or sigmoid.

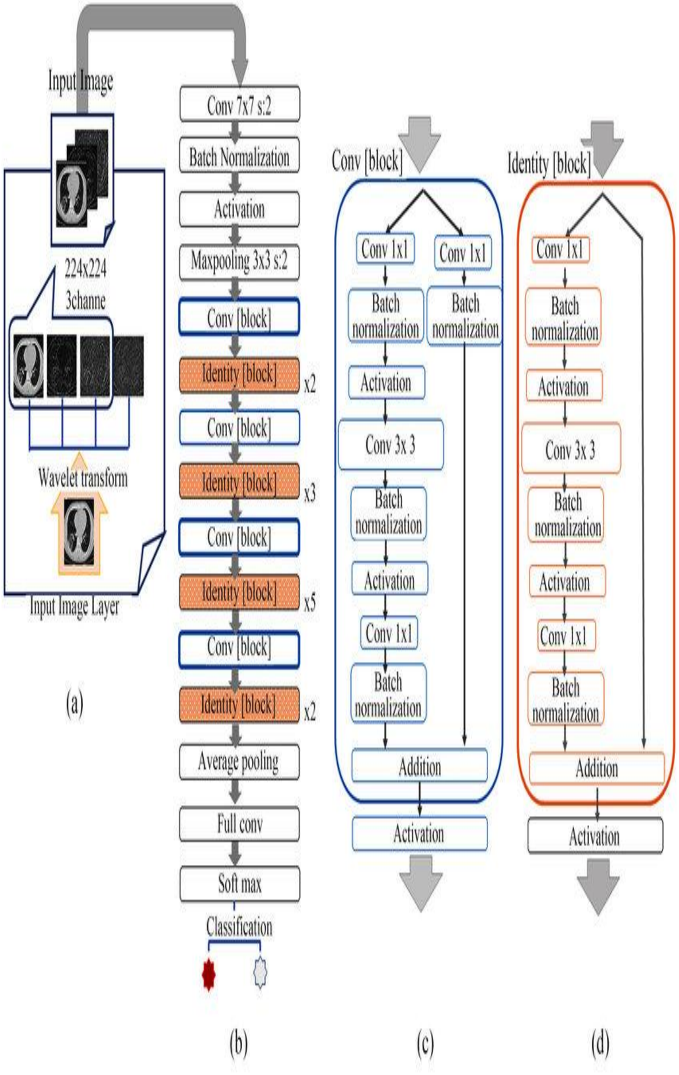

FIG VI: RESNET50 MODEL

13. VGG16

VGG16 [28] is a type of CNN (Convolutional Neural Network) that is considered to be one of the best computer vision models to date. The creators of this model evaluated the networks and increased the depth using an architecture with very small (3 × 3) convolution filters, which showed a significant improvement on the prior-art configurations. They pushed the depth to 16–19 weight layers making it approx — 138 trainable parameters.

In image segmentation, VGG16 can be used as a feature extractor by removing the last fully connected layers and using the output of the final convolutional layer as input for a segmentation network. This approach is known as transfer learning and can significantly reduce the amount of data and training time required for the segmentation task. Image of the model is shown in FIG VII [29].

VGG16 has been used in various medical imaging applications, including lung cancer detection, where it has achieved high accuracy in differentiating between benign and malignant lung nodules.

Convolution Layers,

$$output\_i = relu(conv\_i(output\_i-1)) \tag{27}$$

where output_i is the output of the i-th convolutional layer, conv_i is the convolution operation with learnable parameters, relu is the ReLU activation function, and output_i-1 is the output of the (i-1)-th convolutional layer.

Max pooling Layers,

$$output\_i = max\_pool(output\_i-1) \tag{28}$$

where output_i is the output of the i-th max pooling layer, max_pool is the max pooling operation, and output_i-1 is the output of the (i-1)-th convolutional layer.

Fully connected Layers,

$$output = softmax(relu(fc(output\_i-1))) \tag{29}$$

where output is the final output of the network, fc is the fully connected operation with learnable weights, relu is the ReLU activation function, and softmax is the softmax function that produces a probability distribution over the possible classes.

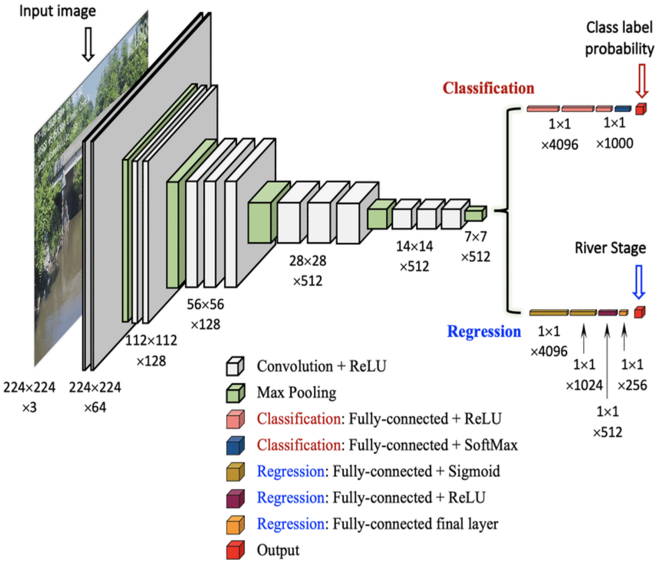

FIG VII: VGG16 MODEL

14. EfficientNet-B7

EfficientNet [30] is a convolutional neural network architecture and scaling method that uniformly scales all dimensions of depth/width/resolution using a compound coefficient. Unlike conventional practice that arbitrary scales these factors, the EfficientNet scaling method uniformly scales network width, depth, and resolution with a set of fixed scaling coefficients. EfficientNet uses a compound coefficient to uniformly scales network width, depth, and resolution in a principled way. There are 813 layers in EfficientNet-B7.

In image segmentation, EfficientNetB7 can be used to accurately classify each pixel of an image into different categories, such as background, foreground, or object of interest. This is achieved by applying convolutional filters to the input image, followed by downsampling and upsampling operations to capture both local and global features. Image of the model is shown in FIG VIII [31].

Overall, EfficientNet-B7 represents a promising approach for image segmentation, with potential applications in medical imaging, autonomous driving, and other fields where accurate segmentation of complex images is required.

Convolutional stem Layers,

$$output\_i = swish(batch\_norm\_i(conv\_i(output\_i-1))) \tag{30}$$

where output_i is the output of the i-th convolutional layer, conv_i is the convolution operation with learnable parameters, batch_norm_i is the batch normalization operation, swish is the swish activation function, and output_i-1 is the output of the (i-1)-th convolutional layer.

The computation of the network head is given by,

$$output = fc2(dropout(swish(batch\_norm1(fc1(flatten(output\_i)))))) \tag{31}$$

where output is the final output of the network, flatten is the operation that flattens the output of the last convolutional layer, fc1 and fc2 are fully connected operations with learnable weights, batch_norm1 is the batch normalization operation, swish is the swish activation function, and dropout is the dropout operation.

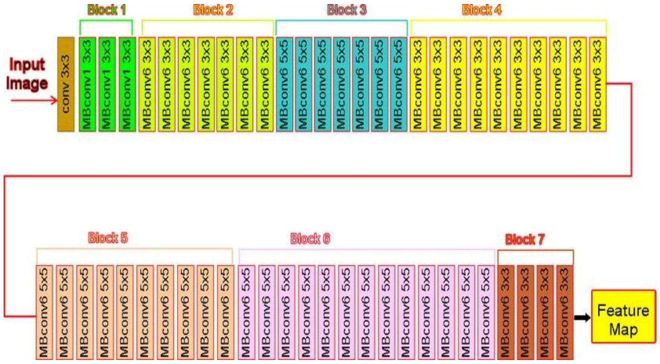

FIG VIII: EFFICIENTNET-B7 MODEL

VI. RESULTS AND DISCUSSION

In the research paper, I implemented various ML algorithms for the task of image segmentation of lung histopathological images and predicting lung cancer. I used machine learning algorithms like SVM, random forest, KNN, logistic regression, naive bayes, CatBoost, XGBoost, Decision Tree and deep neural networks like ANN, CNN, Inception-V3, ResNet50, and VGG16 models for this task.

The experimental results show that CatBoost out of all machine learning algorithm produces the best testing accuracy and training accuracy.

TABLE I: ML ALGORITHM RESULT FOR TRAINING SET

S.No.	ALGORITHM	ACCURACY	MCC	F1 SCORE
1.	KNN	0.951	0.892	0.950
2.	RANDOM FOREST	0.982	0.974	0.982
3.	DECISION TREE	0.986	0.967	0.986
4.	NAÏVE BAYES	0.979	0.953	0.979
5.	SVC	0.987	0.975	0.987
6.	LOGISTIC REGRESSION	0.989	0.991	0.989
7.	CATBOOST	0.991	0.990	0.991
8.	XGBOOST	0.905	0.786	0.905

TABLE I shows the results of using ML algorithm on the training set of lung histopathological image dataset. There are 12000 images in the training set.

TABLE II: ML ALGORITHM RESULT FOR TESTING SET

S.No.	ALGORITHM	ACCURACY	MCC	F1 SCORE
1.	KNN	0.884	0.742	0.876
2.	RANDOM FOREST	0.982	0.974	0.982
3.	DECISION TREE	0.976	0.947	0.976
4.	NAÏVE BAYES	0.979	0.953	0.979
5.	SVC	0.982	0.959	0.982
6.	LOGISTIC REGRESSION	0.985	0.965	0.985
7.	CATBOOST	0.991	0.981	0.991
8.	XGBOOST	0.799	0.552	0.800

TABLE II shows the results of using ML algorithm on the testing set of lung histopathological image dataset. There are 3000 images in the testing set.

I compared the performance of different deep neural network models using different evaluation metrics. One of the key evaluation metrics used was accuracy. I visualized the accuracy comparison of different models using various plots such as bar chart, line plot and boxplot.

The bar chart was used to compare the accuracy of different models in a simple and intuitive way. Each bar represents the accuracy of a particular model and the height of the bar indicates the accuracy value. I observed that ResNet50 model achieved higher accuracy than other models.

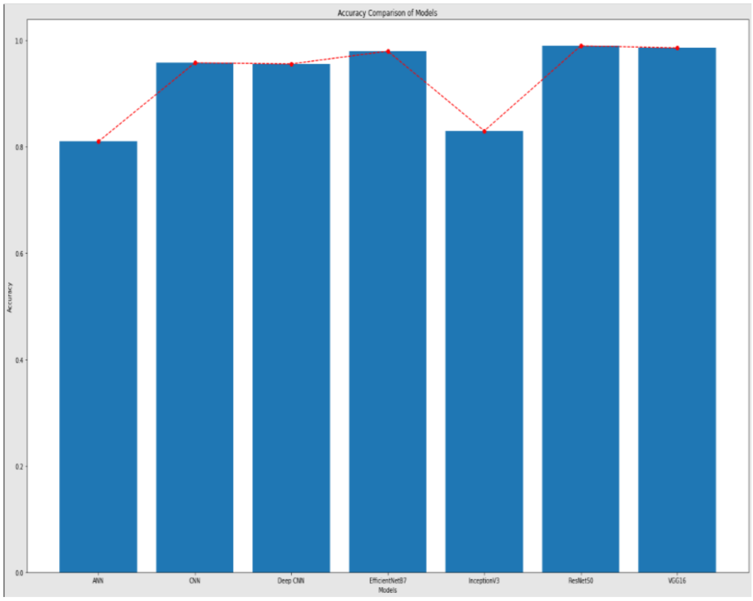

FIGURE IX: ACCURACY COMPARISON OF DEEP NEURAL NETWORK MODEL USING BAR CHART

The line plot was used to visualize the trend of accuracy over time, where time refers the number of epoch or iterations. Each line represents the accuracy of a particular model, and the x-axis represents the number of epoch or iterations, while the y-axis represents the accuracy value. ResNet50 model achieved higher accuracy than other models.

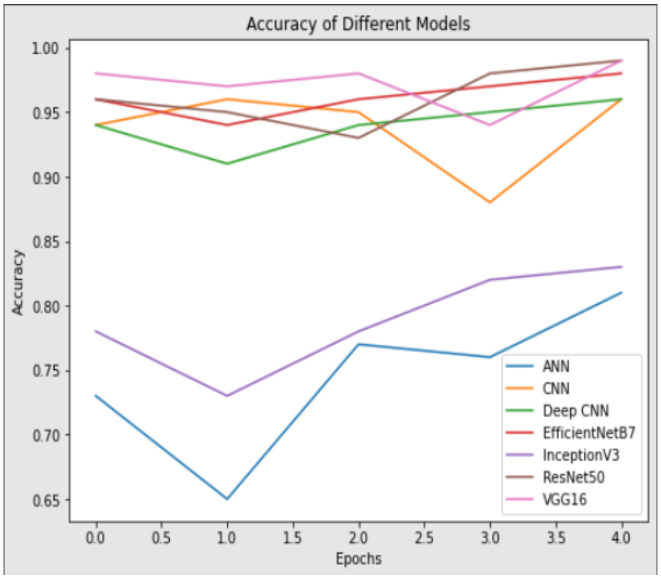

FIGURE X: ACCURACY COMPARISON OF DEEP NEURAL NETWORK MODEL USING LINE CHART

The boxplot was used to compare the distribution of accuracy values for different models. Each boxplot represents the accuracy distribution for a particular model, where the box represents the Interquartile range (IQR) and the whiskers represent the range of accuracy values. Out of all the models, ResNet50 had a higher median accuracy value and a smaller variation in accuracy values compared to other models.

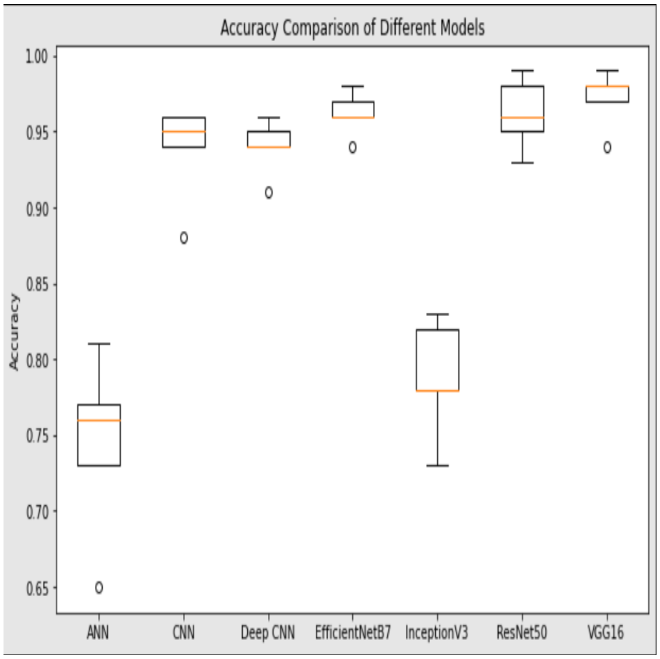

FIGURE XI: ACCURACY COMPARISON OF DEEP NEURAL NETWORK MODEL USING BOXPLOT

VII. CONCLUSION

In this research paper, I explored the feasibility of using machine learning algorithms for image segmentation of lung histopathological images and predicting lung cancer. I tested various machine learning algorithms including SVM, random forest, KNN, logistic regression, naïve bayes, CatBoost, XGBoost, Decision tree and various deep neural network models including ANN, CNN, Deep-CNN, Inception-V3, ResNet50, VGG16 and EfficientNet-B7.

Results indicate that deep neural networks such as CNN, Inception-V3, ResNet50 and VGG16 outperformed traditional machine learning in lung cancer prediction. Among these deep neural networks, ResNet50 achieved the highest accuracy in lung cancer prediction. It demonstrated the ability to identify and segment regions of interest within the lung histopathological images and accurately predict the presence of the lung cancer.

The use of deep neural networks in image segmentation and lung cancer prediction can have significant implications in the field of medical image analysis. It has the potential to improve the accuracy and efficiency of lung cancer diagnosis, which is critical in ensuring timely and effective treatment. Further research can be conducted to explore the application of these algorithms in larger datasets and in other medical imaging tasks.

REFERENCES

[1] Ragavan, M., & Patel, M. I. (2022). The evolving landscape of sex-based differences in lung cancer: a distinct disease in women. European Respiratory Review, 31(163).

[2] Tan, A. C., & Tan, D. S. (2022). Targeted therapies for lung cancer patients with oncogenic driver molecular alterations. Journal of Clinical Oncology, 40(6), 611-625.

[3] Liao, Y., Wu, X., Wu, M., Fang, Y., Li, J., & Tang, W. (2022). Non-coding RNAs in lung cancer: emerging regulators of angiogenesis. Journal of Translational Medicine, 20(1), 1-11.

[4] Merie, R., Gee, H., Hau, E., & Vinod, S. (2022). An overview of the role of radiotherapy in the treatment of small cell lung cancer–a mainstay of treatment or a modality in decline?. Clinical Oncology.

[5] Hişam, D., & Hişam, E. (2021, October). Deep learning models for classifying cancer and COVID-19 lung diseases. In 2021 Innovations in Intelligent Systems and Applications Conference (ASYU) (pp. 1-4). IEEE.

[6] Rudin, C. M., Brambilla, E., Faivre-Finn, C., & Sage, J. (2021). Small-cell lung cancer. Nature Reviews Disease Primers, 7(1), 1-20.

[7] Mukherjee, S., & Bohra, S. U. (2020, December). Lung Cancer Disease Diagnosis Using Machine Learning Approach. In 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS) (pp. 207-211). IEEE.

[8] Jena, S. R., George, T., & Ponraj, N. (2019, February). Texture analysis based feature extraction and classification of lung cancer. In 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT) (pp. 1-5). IEEE.

[9] Barta, J. A., Powell, C. A., & Wisnivesky, J. P. (2019). Global epidemiology of lung cancer. Annals of global health, 85(1).

[10] Ahmed, B. T. (2019). Lung Cancer Prediction and Detection Using Image Processing Mechanisms: An Overview. Signal and Image Processing Letters, 1(3), 20-31.

[11] Putora, P. M., Glatzer, M., Belderbos, J., Besse, B., Blackhall, F., Califano, R., ... & De Ruysscher, D. (2019). Prophylactic cranial irradiation in stage IV small cell lung cancer: Selection of patients amongst European IASLC and ESTRO experts. Radiotherapy and oncology, 133, 163-166.

[12] Wong, D. M., Fang, C. Y., Chen, L. Y., Chiu, C. I., Chou, T. I., Wu, C. C., ... & Tang, K. T. (2018, April). Development of a breath detection method based E-nose system for lung cancer identification. In 2018 IEEE International Conference on Applied System Invention (ICASI) (pp. 1119-1120). IEEE.

[13] Alam, J., Alam, S., & Hossan, A. (2018, February). Multi-stage lung cancer detection and prediction using multi-class svm classifie. In 2018 International conference on computer, communication, chemical, material and electronic engineering (IC4ME2) (pp. 1-4). IEEE.

[14] Romaszko, A. M., & Doboszyńska, A. (2018). Multiple primary lung cancer: a literature review. Adv Clin Exp Med, 27(5), 725-730.

[15] de Groot, P. M., Wu, C. C., Carter, B. W., & Munden, R. F. (2018). The epidemiology of lung cancer. Translational lung cancer research, 7(3), 220.

[16] https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/

[17] https://www.ibm.com/in-en/topics/knn

[18] https://www.geeksforgeeks.org/understanding-logistic-regression/

[19] https://www.ibm.com/topics/naive-bayes

[20] https://catboost.ai/

[21] https://serokell.io/blog/random-forest-classification

[22] https://towardsdatascience.com/beginners-guide-to-xgboost-for-classification-problems-50f75aac5390

[23] https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html

[24] https://www.analyticsvidhya.com/blog/2021/07/understanding-the-basics-of-artificial-neural-network-ann/

[25] https://vitalflux.com/different-types-of-cnn-architectures-explained-examples/

[26] https://cloud.google.com/tpu/docs/inception-v3-advanced

[27] https://datagen.tech/guides/computer-vision/resnet-50/

[28] https://medium.com/@mygreatlearning/everything-you-need-to-know-about-vgg16-7315defb5918

[29] https://www.researchgate.net/figure/VGG-16-neural-network-structure-for-classification-and-regression-models_fig4_355097587

[30]https://www.tensorflow.org/api_docs/python/tf/keras/applications/efficientnet/EfficientNetB7

[31] https://www.researchgate.net/figure/EfficientNetB7-architecture_fig6_358902226