EAI Endorsed Transactions
on Pervasive Health and Technology

Research Article EAI.EU

Real Time Lung Cancer Classification with YOLOv5

Shaif Mehraj Makhdoomi[1], Cherry Khosla[2], and Dr. Sagar Dhanraj Pande[3, *]

[1]Research Scholar, School of Computer Science & Engineering, Lovely Professional University, Phagwara, India.
[2]Assistant Professor, School of Computer Science and Engineering, Lovely Professional University, Phagwara, India.
[3]Assistant Professor Senior Grade, School of Computer Science & Engineering, VIT-AP University, Andhra Pradesh, India.

Abstract

Cancer must be appropriately categorized for effective diagnosis and treatment. Deep learning algorithms have shown tremendous promise in recent years for automating cancer classification. We used the deep learning system YOLOv5 to classify the four types of lung cancer in this study: big cell carcinoma, adenocarcinoma, normal lung tissue, and squamous cell carcinoma. We trained the YOLOv5 model using a publicly available database of lung cancer pictures. The dataset was divided into four categories: big cell carcinoma, adenocarcinoma, normal lung tissue, and squamous cell cancer. In addition, we compared YOLOv5's performance to older models such as SVM, RF, ANN, and CNN. The comparison found that YOLOv5 outperformed all these models, indicating its potential for the development of more accurate and efficient autonomous cancer classification systems. Conclusions from the research have important implications for cancer identification and therapy. Automatic cancer classification systems have the potential to increase the accuracy and efficacy of cancer detection, perhaps leading to better patient outcomes. The accuracy and speed of these systems can be enhanced by using deep learning techniques like YOLOv5, making them more effective for clinical applications. Our study's findings demonstrated high accuracy for every class, with a total accuracy of 97.77%. With the aid of accuracy, train loss, and test loss graphs, we assessed the model's performance. The graphs demonstrated how the model was able to gain knowledge from the data and increase its accuracy as it was being trained. The study's findings were also compiled in a table that gave a thorough assessment of each class's accuracy.

Keywords: Adenocarcinoma, Artificial Intelligence, Automated diagnosis, Classification, Deep Learning, Image Analysis, Large Cell Carcinoma, Lung Cancer Classification, Machine Learning, Medical Imaging

Received on 28 June 2023, accepted on 16 August 2023, published on 20 September 2023

doi: 10.4108/eetpht.9.3925

*Corresponding author. Email: sagarpande30@gmail.com

1. Introduction

Cancer is a collection of diseases that are distinguished by uncontrollable cell proliferation and spread throughout the body. If left untreated, these cells can invade and destroy normal tissues and organs, causing serious health problems and even death. Lung cancer is both one of the most common and one of the worst types of cancer. It happens when abnormal cells in the lungs grow and reproduce uncontrollably, forming a tumor that can infiltrate nearby tissues and spread to other parts of the body. Lung cancer is usually fatal because it is detected late, after the cancer has spread beyond the lungs, making treatment more difficult.

Cancer is a broad and complex term that refers to a collection of diseases that are caused by the uncontrolled growth and spread of abnormal cells in the body. Cancer cells are aberrant cells that could penetrate and destroy normal tissues and organs.

It develops when genetic abnormalities arise within normal cells, causing the processes that control cell growth and division to malfunction. These mutations can be inherited or acquired during a person's lifespan because of a variety of circumstances such as carcinogen exposure (e.g., tobacco smoke, some chemicals, radiation), chronic inflammation, or certain infections.

Because of its aggressive nature, late-stage identification, and high mortality rate, lung cancer is one of the leading causes of cancer-related deaths globally. The importance of early and correct diagnosis in improving patient outcomes and boosting survival rates cannot be overstated. Computer vision and deep learning techniques have increasingly been applied to medical image analysis in recent years, giving a potent tool for detecting and categorizing lung cancer.

Breast cancer, lung cancer, colorectal cancer, prostate cancer, and many other forms of cancer exist. Each type of cancer behaves differently and necessitates different therapies.
Cancer is highly lethal for various reasons:

1. Cancer cells develop and grow at an uncontrolled rate, creating tumors that can disrupt the normal functioning of important organs and tissues.
2. Metastasis occurs when cancer cells break out from the primary tumor and spread to other parts of the body via the bloodstream or lymphatic system. This process, known as metastasis, permits cancer to grow new tumors in different organs, making treatment and control more difficult.
3. Cancer cells can acquire resistance to treatments such as chemotherapy, radiation, or targeted medications. This resistance can be attributed to genetic alterations inside cancer cells or to cancer cells' ability to adapt and survive in adverse environments.
4. Impact on organ function: Depending on the location and extent of the malignancy, it might affect normal organ function, resulting in serious consequences and organ failure.
5. Diagnosis at an advanced stage: Cancer may not create apparent symptoms in its early stages, making it difficult to identify and diagnose until it has developed to an advanced stage in some circumstances. Cancer is often more difficult to treat successfully at late stages.

Cancer detection is critical for several reasons:

1. Early detection of cancer raises the likelihood of successful therapy and improves overall patient outcomes. Early discovery provides for less intrusive treatment alternatives, improved organ function preservation, and a higher chance of complete remission or cure.
2. Improved survival rates: Cancer is frequently more curable and associated with improved survival rates when found early. Regular screenings and early detection programs have been shown to reduce cancer-related fatalities by detecting cancers at an earlier stage when they are more treatable.
3. Early detection can assist healthcare providers in developing individualized treatment programs for patients depending on the specific characteristics of their cancer. This may include identifying the most appropriate therapy, adjusting dosage, and using targeted treatments depending on the cancer's molecular profile.

Despite advancements in treatment options, lung cancer continues to be a serious public health issue and the top cause of cancer-related fatalities globally. Lung cancer is a deadly disease that can be fatal. It is one of the most common and lethal cancers on the globe. The appearance of tiny growths in the lungs known as pulmonary nodules is one of the early indicators of lung cancer. These nodules can be found with low dose computed tomography (CT). However, because there are so many CT scans to go through and the nodules can be difficult to see, doctors may miss some of these nodules in practice. Researchers are creating computer programs known as computer-aided diagnostic (CAD) systems to assist clinicians diagnose lung cancer more accurately. These tools can analyze CT scans and identify any nodules missed by the radiologist. Using a CAD system as an additional opinion can significantly improve the accuracy of a lung cancer diagnosis. In recent years, artificial intelligence has captivated the attention of society, arousing interest in its potential to improve our lives.[1]

Artificial intelligence (AI) has the potential to play a crucial role in cancer detection by analyzing massive volumes of data and identifying patterns that human specialists may find difficult to perceive. Here is some ways AI can help in cancer detection:

1. Image analysis: AI systems can detect and classify abnormalities in medical pictures such as mammograms, CT scans, or histopathology slides to detect and classify cancer. AI can help radiologists and pathologists identify worrisome areas for additional examination, potentially enhancing accuracy and decreasing false negatives or positives.
2. Risk assessment and screening: Using personal and medical history, genetic information, lifestyle factors, and other pertinent data, AI can help determine an individual's risk of acquiring certain types of cancer. This can help with tailored screening programs by recommending relevant screening tests at appropriate intervals for high-risk people.
3. Data integration and analysis: Artificial intelligence (AI) can analyze large-scale datasets such as electronic health records, genomics data, and clinical trial data to uncover correlations, biomarkers, and treatment responses that might inform cancer detection and treatment strategies. AI can aid in the discovery of new insights and

patterns that may lead to new discoveries and advances in cancer research.

4. Decision support: AI can help healthcare professionals make decisions by providing treatment suggestions based on available evidence, patient-specific characteristics, and data-driven algorithms. This can help clinicians make better treatment decisions and improve patient outcomes.

While AI has great potential in cancer diagnosis, it's vital to remember that it's not meant to replace human skill, but rather to supplement and augment it. To ensure the accuracy, safety, and ethical usage of AI technology in healthcare, human oversight and validation are essential.

In this study, we present a lung cancer classification method that employs the cutting-edge object identification algorithm, YOLOv5, to accurately classify lung cancer into four types: adenocarcinoma, big cell carcinoma, normal, and squamous cell carcinoma. The suggested method uses lung computed tomography (CT) scans as input and categorizes them into one of four groups. The effectiveness of the proposed system is assessed using a publicly accessible dataset of lung cancer cases, and the results show that our suggested method surpasses current state-of-the-art techniques and achieves high accuracy. The suggested method may influence clinical practice by making it easier for patients with lung cancer to receive an early diagnosis and a tailored course of therapy, thereby improving patient outcomes and lowering mortality rates. The rest of this essay is structured as follows. The working theory and research methodology are thoroughly described in Section 3. The experimental results are provided and contrasted with the procedures that were used in Section 4. The work's conclusion and discussion are provided in Section 5.

Table-1: Cancer statistics for the year 2023

		Estimated new cases			Estimated deaths	
Cancer site	Both sexes	Male	Female	Both sexes	Male	Female
Oral cavity & pharynx	54,540	39,290	15,250	11,580	8140	3440
Digestive system	3,48,840	1,94,980	1,53,860	1,72,010	99,350	72,660
Colon & rectumb	1,53,020	81,860	71,160	52,550	28,470	24,080
Respiratory system	2,56,290	1,31,150	1,25,140	1,32,330	71,170	61,160
Bones & joints	3970	2160	1810	2140	1200	940
Soft tissue (including heart)	13,400	7400	6000	5140	2720	2420
Skin (excluding basal & squamous)	1,04,930	62,810	42,120	12,470	8480	3990
Breast	3,00,590	2800	2,97,790	43,700	530	43,170
Genital system	4,14,350	2,99,540	1,14,810	69,660	35,640	34,020
Urinary system	1,68,560	1,17,590	50,970	32,590	22,680	9910
Eye & orbit	3490	1900	1590	430	240	190
Brain & other nervous system	24,810	14,280	10,530	18,990	11,020	7970
Endocrine system	47,230	14,340	32,890	3240	1560	1680
Lymphoma	89,380	49,730	39,650	21,080	12,320	8760
Myeloma	35,730	19,860	15,870	12,590	7000	5590
Leukemia	59,610	35,670	23,940	23,710	13,900	9810
unspecified	32,590	16,810	15,780	48,160	26,130	22,030

We used the most recent statistical data available for the year 2023 in our present analysis [2]. We hope to obtain significant insights and undertake numerous studies on various elements of the data by relying on the most recent information. This all-encompassing approach allows us to investigate patterns, trends, and relationships within the dataset, allowing us to make informed interpretations and draw significant conclusions. We hope that by completing these analyses, we might contribute to a better understanding of cancer-related issues such as estimated new cases and fatalities across various cancer locations, as well as potential gender variations within these numbers. Our goal is to provide useful information that can help in tackling cancer's issues and promoting further improvements in prevention, diagnosis, and treatment. (Table-1)

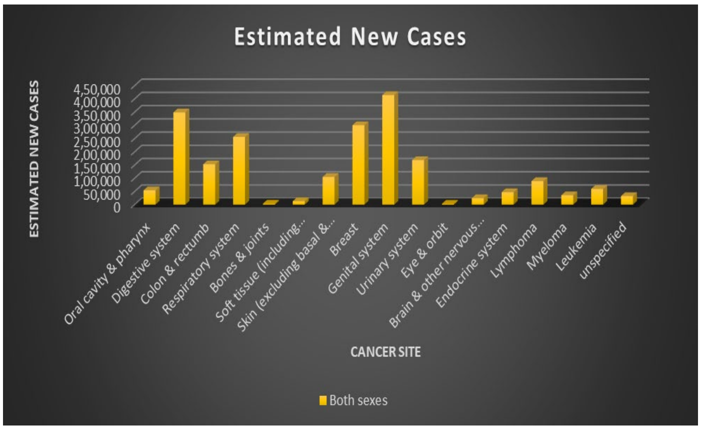

Figure-1: Site wise estimated new cases

Several cancer sites can be evaluated in terms of gender-specific statistics using the data given. There were an estimated 54,540 new instances of oral cavity and pharynx cancer, with 11,580 fatalities. Males accounted for 39,290 new cases and 8,140 fatalities, while females accounted for 15,250 new cases and 3,440 deaths. Similarly, the predicted new cases for the digestive system were 348,840, with 172,010 deaths overall. Males were responsible for 194,980 new cases and 99,350 deaths, while females were responsible for 153,860 new cases and 72,660 deaths. There were 153,020 new instances of colon and rectal cancer, with 52,550 deaths. Males were responsible for 81,860 new cases and 28,470 fatalities, while females were responsible for 71,160 new cases and 24,080 deaths.

We determined the cancer sites with the highest estimated deaths and new cases in our analysis of the most recent statistical data for the year 2023, considering the gender breakdown for each category. The digestive system appears as the most prevalent cancer site with the highest estimated deaths, accounting for a significant number of fatalities. It kills both men and women, killing a total of 172,010 people. Furthermore, the respiratory system is a significant contribution to cancer-related mortality, accounting for 132,330 fatalities in both sexes. Males account for 71,170 deaths in this group, while females account for 61,160 deaths. In terms of estimated mortality, the genitourinary system, comprising malignancies of the reproductive organs, has a notable presence, accounting for 69,660 deaths. Males die here in 35,640 cases, while females die in 34,020 cases.

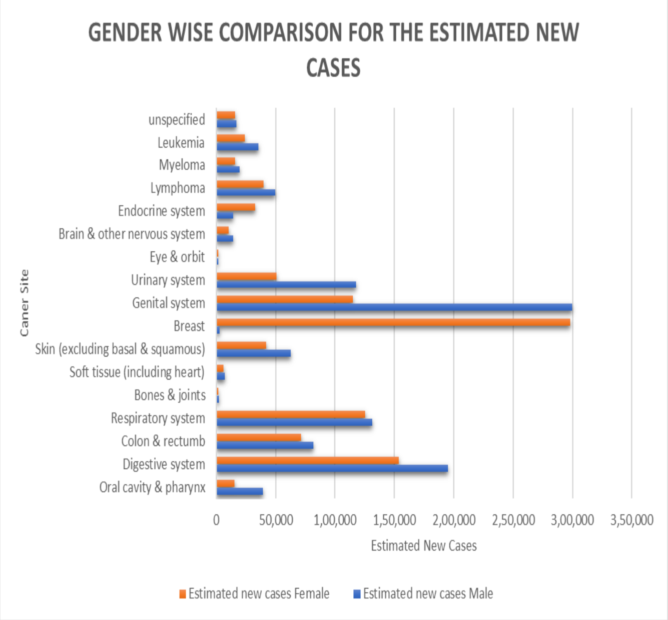

Figure -2: Gender-wise comparison for the estimated new cases

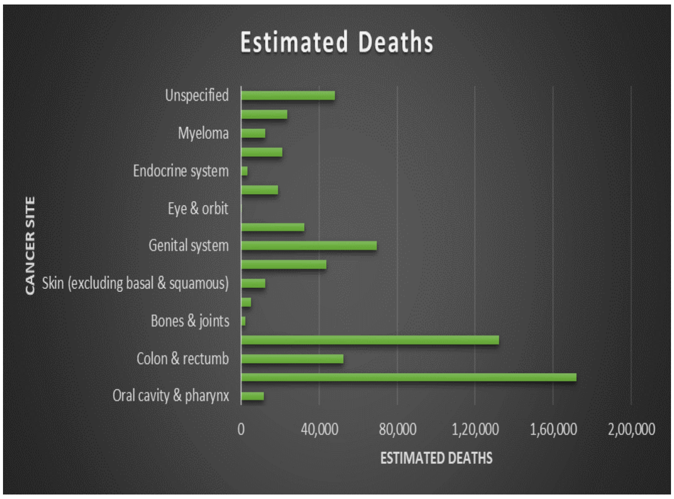

Figure-3: Site wise estimated deaths

There were an estimated 256,290 new cases and 132,330 fatalities in the respiratory system, with men having 131,150 new cases and 71,170 deaths and females having 125,140 new cases and 61,160 deaths. These statistics show the gender distribution of cancer diagnoses and fatalities in these geographic locations.

A gender-based description for the cancer sites you provided:

1. **Oral cavity & pharynx:** Both sexes: 54,540 new cases, resulting in 11,580 deaths.
   Males: 39,290 new cases, resulting in 8,140 deaths.
   Females: 15,250 new cases, resulting in 3,440 deaths.

2. **Digestive system:** Both sexes: 348,840 new cases, resulting in 172,010 deaths.
   Males: 194,980 new cases, resulting in 99,350 deaths.
   Females: 153,860 new cases, resulting in 72,660 deaths.

3. **Colon & rectum:** Both sexes: 153,020 new cases, resulting in 52,550 deaths.
   Males: 81,860 new cases, resulting in 28,470 deaths.
   Females: 71,160 new cases, resulting in 24,080 deaths.

4. **Respiratory system:** Both sexes: 256,290 new cases, resulting in 132,330 deaths.
   Males: 131,150 new cases, resulting in 71,170 deaths.
   Females: 125,140 new cases, resulting in 61,160 deaths.

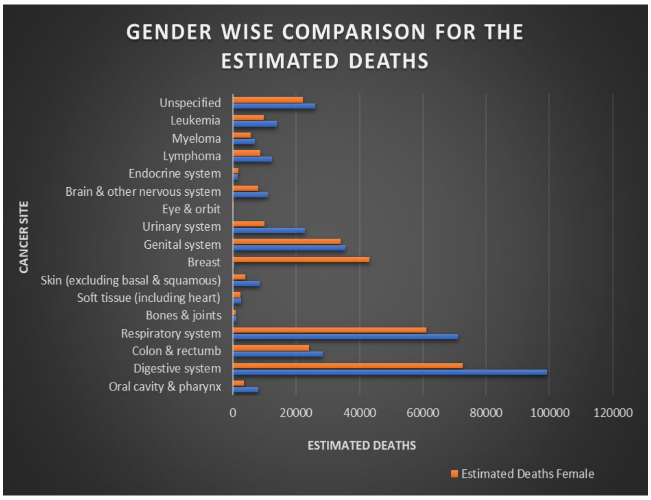

Figure -4: Gender-wise comparison for the estimated death

When the highest estimated new cases are considered, the genitourinary system once again takes the lead, representing a considerable burden of cancer incidence. The total number of new cases is expected to be 414,350, with males accounting for the majority (299,540 new cases) and females accounting for 114,810 new cases. With 348,840 new instances, the digestive system is close behind. Males account for 194,980 of the cases, while females account for 153,860 of the cases.

Finally, breast cancer has a significant prevalence, with a total of 300,590 new cases. Although it primarily affects females (297,790 new cases), males have a tiny number of new cases (2,800).

These findings (Figure-1, Figure-2, Figure-3 & Figure-4) shed insight on the present cancer landscape, highlighting the impact of various cancer locations as well as gender differences within them. By examining this data, additional research and actions can be directed toward addressing the considerable burden caused by various cancer types, with the goal of lowering mortality rates, improving early detection, and improving treatment outcomes.

2. LITERATURE SURVEY

As a result, developing CAD systems for pulmonary nodules is critical for improving lung cancer detection and therapy. AI is now being used extensively to improve disease identification, management, and the efficacy of medications. Because of the expanding number of cancer patients and the vast amount of data collected throughout the treatment process [3]. As a result, AI is required to improve oncologic care. Cancer prognosis can reduce mortality [4].

Apsari et.al.[5] uses artificial neural networks (ANNs) to classify lung nodules into malignant and benign categories. According to the results, the suggested approach has a high accuracy of 80% in categorizing lung nodules as malignant or benign.

Rehman et.al. [6] suggested a pipeline to categorize lung nodules as malignant or benign, four machine learning methods are used: k-nearest neighbors (KNN), support vector machines (SVM), random forests (RF), and artificial neural networks (ANN). The proposed pipeline attained an accuracy of 91.5% for classifying the nodules, according to the data. The SVM classifier achieved the highest accuracy of 93.3%, followed by the RF classifier at 92.3%.

Shaukat et.al.[7] provides a method for classifying lung nodules in CT images as malignant or benign based on intensity, shape, and texture aspects using artificial neural networks (ANNs). The classifier's performance is measured using measures such as accuracy, specificity, sensitivity, and area under the curve (AUC), and it achieves a sensitivity of 95.5%.

Nageswaran et.al.[8] suggested a model for the classification and prediction of lung cancer using machine learning and image processing techniques. The suggested approach was evaluated on a dataset of lung cancer images, and the results showed that the SVM algorithm with the chosen characteristics had the highest accuracy of 98.31% in classifying the images as malignant or benign.

Yasriy et.al.[9] focuses on the use of Convolutional Neural Networks (CNN) for lung cancer diagnosis using CT scans. The suggested method was tested on a dataset of lung cancer CT scans, with the CNN achieving an accuracy of 93.3% in categorizing the CT scans as benign or malignant. The CNN also had a sensitivity of 90% and a specificity of 96.7%, indicating that it could correctly identify both benign and malignant instances.

Kuruvilla et.al.[10] present a computer-aided method for categorizing lung cancer based on CT images, which is critical for increasing a patient's chances of survival. For categorization, the skewness parameter has been proven to be the most accurate. The research presents two novel training functions for the backpropagation neural network, and the results demonstrate that the first proposed function achieves 93.3% accuracy and 91.4% sensitivity.

Lovneet et.al.[11] aim to create an automated approach for detecting lung cancer early, which will improve patient outcomes and lower mortality rates. The study presents the findings of tests carried out to assess the performance of the proposed approach. The researchers used a dataset of CT scan images from people with and without lung cancer in the studies. The results demonstrate that the suggested method was highly accurate in detecting lung cancer, with a sensitivity of 92.3% and a specificity of 97.8%.

Kumar et.al.[12] employs deep learning techniques to categorize lung nodules in CT images. The authors propose a method that extracts deep features from CT images using convolutional neural networks (CNNs), which are then used to train a support vector machine (SVM) classifier with an overall accuracy of 75.01%, sensitivity of 83.35%, and false positive of 0.39/patient over a 10-fold cross-validation.

Attique et.al.[13] proposed a novel feature selection and fusion method that combines classical features, as well as contrast-based features, to increase classification accuracy. The suggested method consists of three major steps: feature extraction, feature selection, and feature fusion. In the feature extraction procedure, the authors employ various conventional features, such as histogram, texture, and form features, as well as contrast-based features, which are based on the difference in intensity values between adjacent pixels. The findings show that the proposed method achieves an accuracy of 93.75%.

Taher et.al.[14] developed a CAD system to detect lung cancer. They used a database of 100 color pictures of sputum obtained from the Tokyo Centre for Lung Cancer. The new CAD method examined the sputum pictures and identified them as benign or cancerous. Another finding in the study showed that Bayesian classification outperformed rule-based heuristic classification, with 97% sensitivity and accuracy.

To improve the precision of lung cancer detection, Gordienko et al. [15] offer a method that combines lung segmentation and bone shadow exclusion approaches. The authors tested their proposed methodology using a dataset of 2500 chest X-ray pictures from individuals with and without lung cancer. The data show that the proposed strategy exceeded earlier methods in terms of lung cancer detection accuracy, obtaining an 88.9% success rate. Furthermore, the authors conducted a sensitivity analysis, which proved that the proposed technique is robust to multiple hyperparameters.

According to Nasrullah et al. [16], a deep learning model based on customized mixed link network (CMixNet) architectures, along with clinical criteria for nodule detection, can reduce false-positive rates and misdiagnosis in the early stages of lung cancer. It was shown to be more sensitive and specific. The suggested strategy was tested on LIDC-IDRI datasets for specificity (91%) and sensitivity (94%).

Wu et.al.[17] the proposed method employs a deep residual network (ResNet) architecture as well as a technique for transferring knowledge from a previously trained model to a new project. With a classification accuracy of 93.44%, the proposed method surpassed multiple existing state-of-the-art methods.

Park et al. [18] offer a two-stage technique that combines a 2D deep convolutional neural network (CNN) with a 3D U-Net network to improve the precision of lung cancer segmentation. The scientists used a dataset of 90 PET/CT images of people with lung cancer to evaluate their proposed method. The data show that the proposed approach, which had a mean Dice similarity coefficient of 0.66, accurately segregated lung cancer. The authors also compared their findings to those obtained using different methodologies, demonstrating that their suggested approach outperforms them.

Yanjie et al. [19] investigate the use of machine learning methods to differentiate benign from cancerous lung nodules discovered using computed tomography (CT) scans. This research aims of this research is to identify the most important features from CT scans that may be used to build an accurate classifier using a support vector machine (SVM). According to the study's findings, the SVM-based classifier had an overall accuracy of 89.8%, sensitivity of 91.7%, and specificity of 87.5%. The AUC-ROC value was 0.936.

Roy et al. [20] suggested a three-stage fuzzy inference method that included picture pre-processing, feature extraction, and classification. To increase image quality, the authors apply contrast enhancement and morphological treatments to lung pictures during the image pre-processing step. The authors extract 11 features from each image during the feature extraction stage, including texture and shape features. These features are then fed into the fuzzy inference system, which determines if the image is normal or abnormal and detects the existence of nodules. The scientists then tested their algorithm on a lung imaging dataset and compared the results to a typical CAD system. The results show that the fuzzy inference system achieves 94% accuracy.

3. METHODOLOGY

3.1. DATASET:

The accuracy and effectiveness of medical imaging analysis for disease diagnosis and treatment have significantly increased because of the development of machine learning and deep learning techniques. One example of this is the detection of lung cancer, where precise and effective classification of lung nodules can dramatically enhance patient outcomes. In this study, we make use of a Kaggle dataset that is freely available to the public and contains CT scans of lung nodules divided into four categories: adenocarcinoma, large cell carcinoma, normal, and squamous cell carcinoma. Figure-5: explains the classes in the dataset.

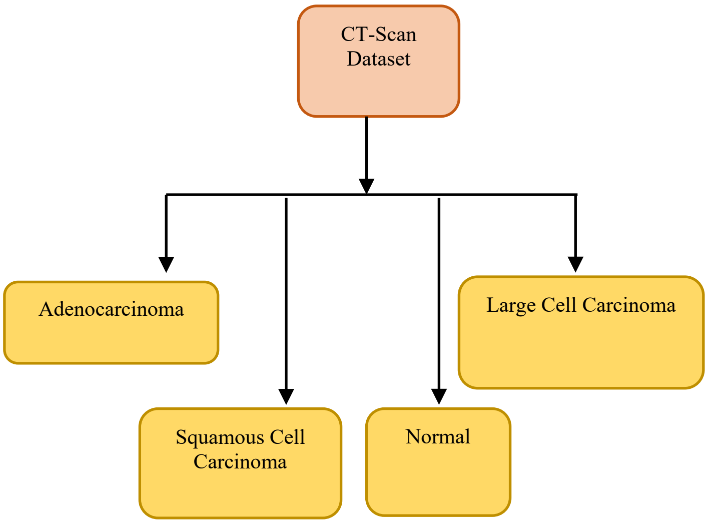

Figure-5: Classes in the Data Set

The four categories indicate various forms of lung cancer, each with special traits and available therapies. The most frequent kind of lung cancer is adenocarcinoma, while big-cell carcinoma is less common but more dangerous. The lining of the lungs' airways is where squamous cell carcinoma first appears, as opposed to normal.

1. *Adenocarcinoma:* This form of lung cancer develops from the glandular cells in the lungs that create mucus. The most prevalent form of lung cancer is adenocarcinoma, which often develops in the outer regions of the lung. It is frequently related to smoking; however, it can also occur in nonsmokers. Although adenocarcinoma grows more slowly than other types of lung cancer, it can spread to other parts of the body.

2. *Squamous cell carcinoma:* This form of lung cancer develops from the flat cells lining the lungs' airways. Squamous cell carcinoma typically develops close to the bronchus in the middle of the lung. It frequently results from smoking and develops more slowly than small-cell lung cancer. Chest pain, breathlessness, and coughing are all possible side effects of squamous cell carcinoma.

3. *Large cell Carcinoma:* Large cell carcinoma is a less common type of lung cancer than adenocarcinoma and squamous cell carcinoma. It may be more difficult to treat than other types of lung cancer due to its propensity for rapid growth and spread. The development of large cell carcinoma can occur anywhere in the lung, and because it lacks the same distinctive features as other forms of lung cancer, it can be challenging to identify.

4. *Small cell lung cancer:* The most dangerous kind of lung cancer, small cell lung cancer grows and spreads swiftly. Smoking is typically a risk factor for small cell lung cancer, which typically appears in the middle of the lung, close to the bronchus. Shortness of breath, chest pain, and coughing are possible symptoms.

Although radiation and chemotherapy may work well to treat small cell lung cancer, the disease frequently returns after treatment.

Lung cancer is a common and fatal illness that has several subtypes with distinct characteristics. For efficient planning of treatment and patient management, accurate identification and classification of lung tumor subtypes is critical. This study intends to fill that gap by using image classification algorithms to differentiate four major lung cancer subtypes: adenocarcinoma, big cell carcinoma, normal lung tissue, and squamous cell carcinoma. Table-2 gives the entire break-up for the dataset division.

Table-2: Dataset

Class	Train-Images	Test-Images	Validation-Images
Adenocarcinoma	195	120	89
Large Cell carcinoma	115	51	91
Normal	148	54	41
Squamous cell Carcinoma	155	90	71

A. *MODEL:* The You Only Look Once (YOLO) algorithm, first introduced in 2016 as an object detection technique, is a popular deep learning algorithm for object detection in computer vision. It was created to recognize objects in real time and achieve high accuracy while using relatively few processing resources. YOLO has gone through various versions and changes over the years, with the most recent version being YOLOv5, which was launched in 2020. Ultralytics created YOLOv5, which marks a considerable advance in terms of accuracy and speed over its predecessors. One of the most well-known object identification networks in the world, YOLOv5, now has more than just object detection up its sleeve. YOLOv5 now supports classification jobs as of August 2022.

3.2. EXPERIMENTAL SETUP:

Table-3 shows the set Hyper-parameters of our model. Our experiment employs the Pytorch framework deep learning on GPU Tesla K80 by Google open Platform Colab-research.

This well-liked computer vision model for object detection and categorization, for its classification model, defines the class names using a folder structure. This implies that a class is given to every image in the dataset according to the folder it is kept in. Although this strategy might seem unorthodox, it has several benefits. Because the folder structure clearly defines the class labels, for instance, managing the dataset is made simpler.

Table-3: Experimental Set-Up

PARAMETER	VALUE
Image size	412
Epoch	350
Model	yolov5x-cls.pt

Additionally, by using this method, the likelihood of errors when manually labeling bounding boxes can be decreased. Figure-6 shows the folder structure for the classification model.

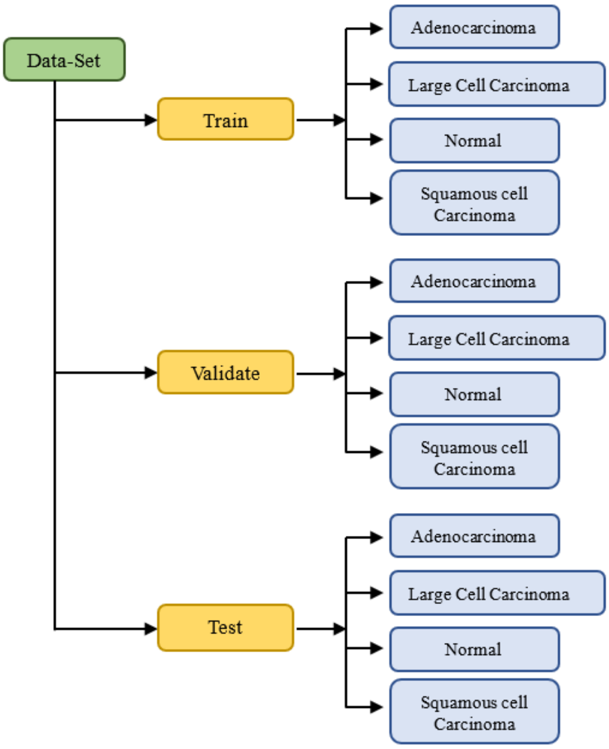

Figure-6: Complete folder structure of the Dataset

For object recognition and categorization, the system employs a popular computer vision model known as YOLOv5. The model utilized in this instance is the "yolov5x-cls.pt" variation. With parameters and criteria, the model was trained and optimized.

The image size parameter is set to 412, indicating that the model expects the resolution or dimensions of the input images during inference. This parameter aids in ensuring that photos are shrunk or cropped to a consistent size before being processed by the model. The model was trained by running it through 350 iterations. An epoch is a complete iteration across the training dataset. Training a model for numerous epochs enables it to learn from the data and gradually improve its performance over time.

The class names for the model's classification are defined using a folder structure. This means that each image in the dataset is given a class based on the folder in which it is kept. This unusual technique offers several benefits. Because the folder structure clearly defines the class labels, it facilitates dataset administration. It also decreases the possibility of mistakes while manually labeling bounding boxes, which are used to designate the locations of objects within images. In machine learning, the most used formulae for calculating training loss, test loss, and accuracy.

1 Training Loss:

Training loss is a measure of how well the model is performing on the training data. It quantifies the difference between the predicted outputs of the model and the actual ground truth labels in the training set. The specific formula for training loss can vary depending on the problem, algorithm, and loss function used.

For example, in binary classification problems using logistic regression, the training loss is often calculated using the cross-entropy loss formula:

$$\text{Loss} = -(1/N) * \sum [y * \log(y\_hat) + (1 - y) * \log(1 - y\_hat)] \qquad \textit{(eq-1)}$$

where:

N is the number of training samples
y is the ground truth label (0 or 1)
y_hat is the predicted probability of the positive class (ranging from 0 to 1)

2 Test Loss:

Test loss, also known as validation loss or evaluation loss, measures how well the trained model generalizes to unseen data. It is calculated using the same loss function as the training loss but applied to the test/validation dataset.

The formula for test loss is like the training loss formula:

$$\text{Loss} = -(1/N) * \sum [y * \log(y\_hat) + (1 - y) * \log(1 - y\_hat)] \qquad \textit{(eq-2)}$$

where:

N is the number of test/validation samples
y is the ground truth label (0 or 1)
y_hat is the predicted probability of the positive class (ranging from 0 to 1)

3 Accuracy:

Accuracy is a metric that measures the overall performance of a classification model. It calculates the percentage of correct predictions made by the model on a given dataset.

The formula for accuracy is:

$$\text{Accuracy} = \frac{\text{(Number of correct predictions)}}{\text{(Total number of predictions)}} \qquad \textit{(eq-3)}$$

This formula simply divides the number of correctly predicted samples by the total number of samples in the dataset.

The test loss and train loss are significant metrics for assessing a machine learning model's efficiency and performance. They reveal how well the model is learning and generalizing from training data to unknown test data. Here's how these measures help to describe the model's efficiency:

1. Training Loss: This metric measures how well the model fits the training data. The model seeks to reduce training loss by modifying its parameters during the training process. A reduced training loss means that the model predicts the ground truth labels for the training samples better.

2. The test loss assesses the model's performance on the test or validation dataset, which contains data that the model did not observe during training. The test loss measures how well the model generalizes to new, previously unknown data. A smaller test loss indicates that the model can make accurate predictions on unobserved data.

The test loss is an important indicator for determining a model's efficiency. If the test loss is much more than the training loss, the model is overfitting. When a model fails to generalize successfully to new data, it overfits, resulting in poor performance on previously encountered samples. In such instances, the model may be overly complex or contain learned noise.

The training loss reveals how well the model matches the training data, whereas the test loss evaluates its ability to generalize to new data. A good model aspires for low training and test losses, showing a good fit to the training data as well as great generalization performance. Monitoring and decreasing the difference between training and test losses contributes to the model's efficiency and effectiveness in real-world applications.

4 RESULT ANALYSIS

We used the cutting-edge categorization method YOLOv5 to divide cancer into four classes large cell carcinoma, adenocarcinoma, normal lung tissue, and squamous cell carcinoma Our goal was to assess YOLOv5's performance on this task and contrast it with earlier models.

First, we used a dataset of CT-image examples from the four classes of cancer to train the YOLOv5 model. To track the development of the training process, we plotted the train loss graph Figure-7, the test loss graph Figure-8, and the accuracy graph Figure-9.

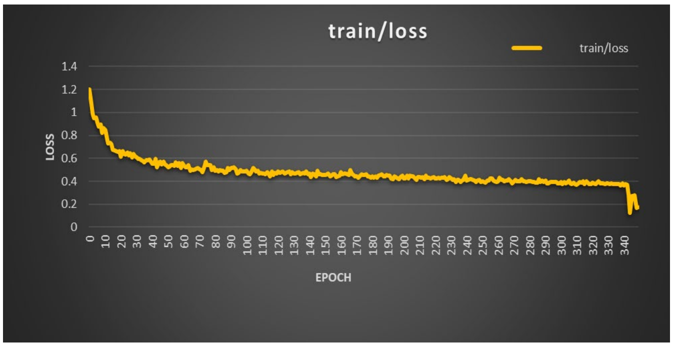

Figure-7: Epoch-Wise Training Loss

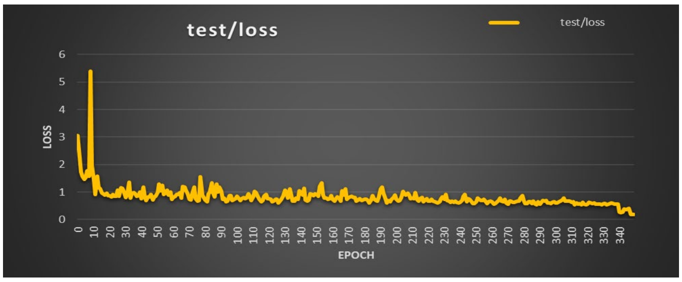

Figure-8: Epoch-Wise Test Loss

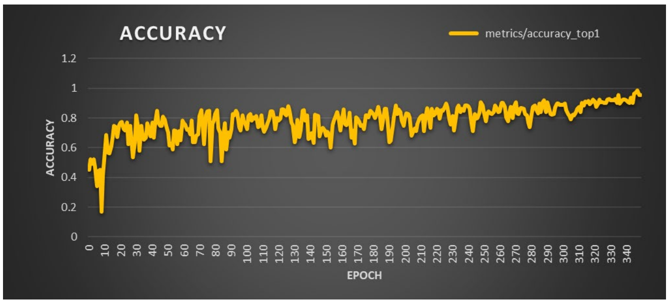

Figure-9: Epoch-Wise Accuracy

The test loss graph displayed the loss function for the validation set, whereas the train loss graph displayed the variation of the loss function over the training period. The accuracy graph displayed the accuracy variance throughout the training period.

We then tested the model on a separate test data set and with an overall accuracy of 97.77%, our results demonstrated that the model was able to attain excellent accuracy for each class.
We compiled the findings into a table (Table-4), which gave us the accuracy for each class, as well as the overall accuracy for the model, to further assess the model's performance. The table demonstrated that the model had excellent accuracy for each of the four types of cancer, demonstrating that the model was capable of properly classifying each class.

Table-4: Results

Class	*Images*	*True-Positive*	*False-Positive*	*Accuracy*
All	315	308	7	0.9777
Adenocarcinoma	120	117	3	0.95
Large Cell Carcinoma	51	49	2	0.942
Normal	54	54	0	1
Squamous cell Carcinoma	90	88	2	0.943

In addition, we evaluated our model against earlier models, such as SVM, RF, ANN, and CNN. Our findings demonstrated that YOLOv5 performed better than each of these models, demonstrating its superiority for the categorization of cancer. (Figure-10, Table-5).

The accuracy findings show the varying effectiveness of the machine learning approaches investigated for classifying lung cancer subtypes. The maximum accuracy of 97.77% was reached by YOLO-v5, suggesting its better skill in effectively detecting and discriminating between the different subtypes. SVM and CNN both had high accuracy rates of 93% and 93.30%, demonstrating their effectiveness in this challenge. ANN and RF had somewhat lower accuracy rates of 80% and 92%, respectively, indicating their poorer performance.

Table-5: Accuracy comparison.

Method	Accuracy
ANN	80%
SVM	93%
RF	92%
CNN	93.30%
YOLO-v5	97.77%

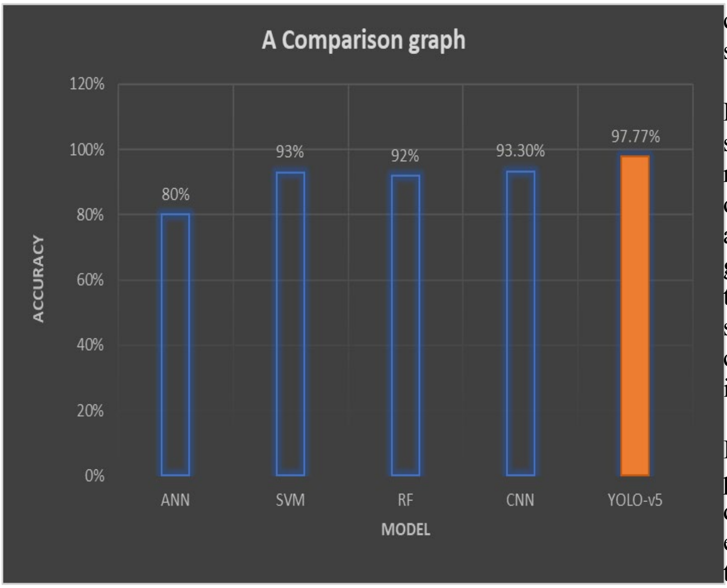

Figure-10: A Complete Comparison of Various Models

5 Conclusion

The objective of this study was to explore the potential of the deep learning algorithm YOLOv5 in distinguishing between four types of lung cancer: adenocarcinoma, big cell carcinoma, normal, and squamous cell carcinoma. The outcomes of our investigation yielded positive results, as we achieved exceptional accuracy in each class and an overall accuracy rate of 97.77%.

To evaluate the performance of the model, we employed various measures including train loss, test loss, accuracy graphs, and summary tables. These metrics provided a comprehensive assessment of the model's performance and demonstrated that YOLOv5 is a valuable tool for automated cancer classification. Additionally, we compared the performance of YOLOv5 with earlier models such as SVM, RF, ANN, and CNN, revealing the superior capabilities of deep learning algorithms like YOLOv5 in creating more precise and effective automated cancer classification systems.
However, it is crucial to acknowledge the limitations of our research. One significant limitation is the size of the dataset used for training and testing the model, which may not be fully representative of all lung cancer cases. This limitation highlights the need for further research with larger and more diverse datasets to thoroughly analyze the model's performance.

Despite these limitations, our research contributes to the expanding body of knowledge on the application of deep learning algorithms in cancer detection and treatment. The findings of this study hold important implications for the development of more accurate and effective automated cancer categorization systems, which have the potential to significantly improve patient outcomes.

Moving forward, it is essential to conduct additional studies that address the limitations identified in this research. Expanding the dataset to include a broader range of lung cancer cases, considering various demographics and pathological characteristics, would enhance the generalizability of the model. Furthermore, investigating the performance of YOLOv5 in real-world clinical settings and comparing it with other existing techniques can provide valuable insights for its practical implementation.

In conclusion, our study demonstrates the promising potential of YOLOv5 as a deep learning algorithm for distinguishing between different types of lung cancer. The exceptional accuracy achieved in our evaluation suggests that YOLOv5 can contribute to the development of automated cancer classification systems, ultimately leading to improved patient outcomes and more effective cancer management strategies.

References

[1] Y. Kumar, K. Sood, S. Kaul, and R. Vasuja, “Big Data Analytics and Its Benefits in Healthcare,” Studies in Big Data, vol. 66, pp. 3–21, 2020, doi: 10.1007/978-3-030-31672-3_1/COVER.

[2] R. L. Siegel Mph et al., “Cancer statistics, 2023,” CA Cancer J Clin, vol. 73, no. 1, pp. 17–48, Jan. 2023, doi: 10.3322/CAAC.21763.

[3] P. Jin et al., “Artificial intelligence in gastric cancer: a systematic review,” J Cancer Res Clin Oncol, vol. 146, no. 9, pp. 2339–2350, Sep. 2020, doi: 10.1007/S00432-020-03304-9/METRICS.

[4] W. G. E. Gonçalves, M. H. D. P. Dos Santos, F. M. F. Lobato, Â. Ribeiro-Dos-Santos, and G. S. De Araújo, “Deep learning in gastric tissue diseases: a systematic review,” BMJ Open Gastroenterol, vol. 7, no. 1, p. e000371, Mar. 2020, doi: 10.1136/BMJGAST-2019-000371.

[5] R. Apsari, Y. N. Aditya, E. Purwanti, and H. Arof, “Development of lung cancer classification system for computed tomography images using artificial neural network,” AIP Conf Proc, vol. 2329, no. 1, Feb. 2021, doi: 10.1063/5.0042195/962453.

[6] A. Rehman, M. Kashif, I. Abunadi, and N. Ayesha, “Lung Cancer Detection and Classification from Chest CT Scans Using Machine Learning Techniques,” 2021 1st International Conference on Artificial Intelligence and Data Analytics, CAIDA 2021, pp. 101–104, Apr. 2021, doi: 10.1109/CAIDA51941.2021.9425269.

[7] F. Shaukat, G. Raja, R. Ashraf, S. Khalid, M. Ahmad, and A. Ali, “Artificial neural network based classification of

lung nodules in CT images using intensity, shape and texture features," J Ambient Intell Humaniz Comput, vol. 10, no. 10, pp. 4135–4149, Oct. 2019, doi: 10.1007/S12652-019-01173-W/TABLES/4.

[8] S. Nageswaran et al., "Lung Cancer Classification and Prediction Using Machine Learning and Image Processing," Biomed Res Int, vol. 2022, 2022, doi: 10.1155/2022/1755460.

[9] H. F. Al-Yasriy, M. S. Al-Husieny, F. Y. Mohsen, E. A. Khalil, and Z. S. Hassan, "Diagnosis of Lung Cancer Based on CT Scans Using CNN," IOP Conf Ser Mater Sci Eng, vol. 928, no. 2, p. 022035, Nov. 2020, doi: 10.1088/1757-899X/928/2/022035.

[10] J. Kuruvilla and K. Gunavathi, "Lung cancer classification using neural networks for CT images," Comput Methods Programs Biomed, vol. 113, no. 1, pp. 202–209, Jan. 2014, doi: 10.1016/J.CMPB.2013.10.011.

[11] L. Kaur, M. Sharma, R. Dharwal, and A. Bakshi, "Lung Cancer Detection Using CT Scan with Artificial Neural Netwok," 2018 International Conference on Recent Innovations in Electrical, Electronics and Communication Engineering, ICRIEECE 2018, pp. 1624–1629, Jul. 2018, doi: 10.1109/ICRIEECE44171.2018.9009244.

[12] D. Kumar, A. Wong, and D. A. Clausi, "Lung Nodule Classification Using Deep Features in CT Images," Proceedings -2015 12th Conference on Computer and Robot Vision, CRV 2015, pp. 133–138, Jul. 2015, doi: 10.1109/CRV.2015.25.

[13] M. A. Khan et al., "Lungs cancer classification from CT images: An integrated design of contrast based classical features fusion and selection," Pattern Recognit Lett, vol. 129, pp. 77–85, Jan. 2020, doi: 10.1016/J.PATREC.2019.11.014.

[14] F. Taher, N. Werghi, and H. Al-Ahmad, "Computer Aided Diagnosis System for Early Lung Cancer Detection," Algorithms 2015, Vol. 8, Pages 1088-1110, vol. 8, no. 4, pp. 1088–1110, Nov. 2015, doi: 10.3390/A8041088.

[15] Y. Gordienko et al., "Deep learning with lung segmentation and bone shadow exclusion techniques for chest X-ray analysis of lung cancer," Advances in Intelligent Systems and Computing, vol. 754, pp. 638–647, 2019, doi: 10.1007/978-3-319-91008-6_63/COVER.

[16] N. Nasrullah, J. Sang, M. S. Alam, M. Mateen, B. Cai, and H. Hu, "Automated Lung Nodule Detection and Classification Using Deep Learning Combined with Multiple Strategies," Sensors 2019, Vol. 19, Page 3722, vol. 19, no. 17, p. 3722, Aug. 2019, doi: 10.3390/S19173722.

[17] P. Wu, X. Sun, Z. Zhao, H. Wang, S. Pan, and B. Schuller, "Classification of Lung Nodules Based on Deep Residual Networks and Migration Learning," Comput Intell Neurosci, vol. 2020, 2020, doi: 10.1155/2020/8975078.

[18] J. Park et al., "Automatic Lung Cancer Segmentation in [18F]FDG PET/CT Using a Two-Stage Deep Learning Approach," Nucl Med Mol Imaging, vol. 57, no. 2, pp. 86–93, Apr. 2022, doi: 10.1007/S13139-022-00745-7/METRICS.

[19] Y. Zhu, Y. Tan, Y. Hua, M. Wang, G. Zhang, and J. Zhang, "Feature Selection and Performance Evaluation of Support Vector Machine (SVM)-Based Classifier for Differentiating Benign and Malignant Pulmonary Nodules by Computed Tomography," Journal of Digital Imaging: the official journal of the Society for Computer Applications in Radiology, vol. 23, no. 1, p. 51, Feb. 2010, doi: 10.1007/S10278-009-9185-9.

[20] T. S. Roy, N. Sirohi, and A. Patle, "Classification of lung image and nodule detection using fuzzy inference system," International Conference on Computing, Communication and Automation, ICCCA 2015, pp. 1204–1207, Jul. 2015, doi: 10.1109/CCAA.2015.7148560.