Journal of Autonomous Intelligence (2024) Volume 7 Issue 5
doi: 10.32629/jai.v7i5.1605

ORIGINAL RESEARCH ARTICLE

Hybrid approach for lung cancer detection based on deep learning/machine learning

Sandeep Kumar Hegde[1], Sujidha B.[2], K. Vimala Devi[3,*], K. Maheswari[4], K. Leela Krishna[5], Pallavi Singh[6], Varsha D. Jadhav[7]

[1] *Department of Computer Science and Engineering, NMAM Institute of Technology, NITTE (Deemed to be University), Karnataka 574110, India*

[2] *Department of Science & Humanities, Rathinam Technical Campus, Coimbatore 641021, India*

[3] *School of Computer Science and Engineering, Vellore Institute of Technology, Vellore 632014, India*

[4] *Department of CSE, CMR Technical Campus, Hyderabad 501401, India*

[5] *Civil Engineering Department, RVR & JC College of Engineering, Guntur 522019, India*

[6] *Department of Biotechnology, Graphic Era Deemed to be University, Dehradun 248002, India*

[7] *Department of Artificial Intelligence and Data Science, Vishwakarma Institute of Information Technology, Pune 411048, India*

*** Corresponding author:** K. Vimala Devi, vimaladevi.k@vit.ac.in

ABSTRACT

The incidence of Lung Cancer (LC) is rising in India. LC has been diagnosed and detected numerous times utilizing numerous data processing and identification strategies. Since the underlying origin of LC is still unknown, treatment is hopeless, making early diagnosis of lung tumors the only viable treatment option. So, a Machine Learning (ML) and Deep Learning (DL) based system is utilized to categorize CT scans for the existence of LC. The Visual Geometry Group (VGG-16) and Multi-Class Support Vector Machine (VGG-16+MSVM) technique is proposed in this research. Non-Local Means (NLM) Filter and Bi-Histogram Equalization (Bi-HE) are used, respectively; to filter out unwanted background noise in raw data samples and improve image quality. To isolate tumors in the raw data, the K-Means Clustering (KMC) technique is applied. The Gray Level Co-Occurrence Matrix (GLCM) is employed to generate features from the segmented data. The proposed approach is optimized with the use of a Genetic Algorithm (GA) that selects optimal feature subsets to maximize its performance. Combining ML and DL methods in Medical Image Processing is the most effective approach to detecting LC and its stages with the hope of achieving more precise findings. When accuracy is assessed and compared to other procedures, it becomes clear that the suggested methodology is more accurate (95%).

Keywords: medical image processing; LC; ML; DL; VGG-16; multi-class support vector machine (VGG-16+MSVM)

ARTICLE INFO

Received: 23 February 2024
Accepted: 9 April 2024
Available online: 29 May 2024

1. Introduction

LC is a sort of lethal cancer that is challenging to identify. Typically, it results in mortality for both males and females. Smoking is virtually related, and the latter is accelerating. Non-small cell LC (NSCLC) is more prevalent and develops more gradually. Hybrid little cell/large cell cancer is the label assigned to it if both forms of cancer are present in the patient. According to current statistics, NSCL accounts for over 85% of the estimated 234,030 new instances of LC predicted to be identified in 2018. The proliferation of LC without indications is the main aspect that makes this illness so dangerous. A percent of those surveyed showed no indications of

malignancy. Many individuals are aware that LC may also generate X-rays of the lungs. The importance of prompt detection cannot be overstated since LC spreads swiftly. LC may begin in the major airway, the windpipe, the lungs, or another location. It is brought on by unchecked cell expansion and proliferation of certain cells in the lungs. LC is quite frequent in those with emphysema or lung conditions. The mainly two categories of LC are "small-cell lung carcinoma (SCLC) and non-small-cell lung tumor (NSCLC)"[1]. SCLC, the more dangerous pathologic form of LC, makes up for 25%–40% of lung tumor cases in China, making it one of the most deadly tumors. Elevated LC has a severe fatality ratio and few treatment choices.With a strong growth percentage, quick replication times, and the initial emergence of extensive metastatic tumors, it has a distinctive biological record.. SCLC, formerly known as "oat cell carcinoma", originally emerged in research in 1936, in a report of an instance of the disease in a person who had asbestosis Various sampling techniques may have an impact on the estimated frequency of C-SCLC, which varies from 2% to 28% of all SCLC patients in various studies. Since there is minimal evidence of inter-tumor diversity about morphology or biomolecular in medical care, SCLC has so far been considered a "uniformly" illness[2]. **Figure 1** depicts the SCLC and NSCLC LC.

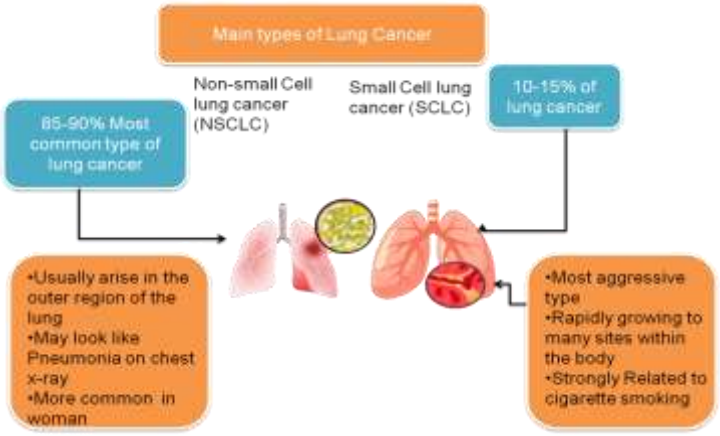

Figure 1. The SCLC and NSCLC LC.

Among all types of LC, NSCLC spreads 40% more rapidly than many kinds of LC. Because of this, it is thought that carcinoma in its initial phases has a strong probability of not spreading. The NSCLC disease is determined by four phases, based on the phase classifications: Stage 0 (unknown), in which cancerous cells are discovered in sputum or tracheal soakings but cannot be readily seen using scanning methods or tracheostomy; in addition, cancer may have spread to other body areas; Stage 1 (development of tumor), no lymph nodes have been affected; Stage 2 refers to a disease that has progressed to the major bronchial tubes; Stage III refers to tumors that have grown to several body parts but have not yet shown evidence of metastasis, and phase IV refers to cancer that has spread to many locations in one or more systems. Most instances that test positive for NSCLC have already proceeded to the developed phases[3]. Having 21 lakh new cases and 18 lakh fatalities from LC in 2018, it ranks as the highest prevalent form of cancer globally. Contrary to other malignancies like breast and testicular tumors, which often appear with a single recognizable side effect, LC has a more wide common side effect pattern (e.g., painless lump). Shortness of breath, chest discomfort, a chronic cough, and other signs including alterations to an underlying cough may all be caused by initial LC. Severe illness is often accompanied by widespread signs such as unexplainable body loss and exhaustion. One of the best LC symptom predictions is hemoptysis (bleeding). Loss of appetite, lump in the neck, weakness, abdominal pain, and blood clots are other major symptoms of LC[4]. **Figure 2** displays the symptoms of LC.

Smoking is connected to certain genomic alterations that result in lung malignancies with unique pathological characteristics. Few oncogenes and carcinogenic factors have comparable impacts of smoking on the LC genomic that are well characterized. Smoking is the greatest significant danger solid indications linking

this risk factor to LC as a key cause of the disease. In comparison to non-smokers, smokers have a 30-fold higher chance of acquiring cancer. LC may be promoted by inflammation through several different routes. Therefore, pulmonary inflammation may contribute to the development or spread of cancer. The tobacco-induced pulmonary cell connection provides a distinct setting in which lung inflammation, functional, and stromal cells collaborate to promote cancer. The substantial modifications brought about by cigarette smoke, which includes well-known carcinogens as well as large quantities of reacting oxygen molecules; represent the first link between smoking and LC. After being exposed to cigarette smoke, the quick generation of reacting oxygen causes inflammation as well as a deficiency in epithelium and endothelium cellular functions[5]. The greatest impact of cancer mortality globally is LC. The prediction of patients with severe LC is regarded as incorrect since the earlier detection ratio of LC is only 15%, and 75% of individuals are identified at a severe or localized phase. When metastatic develops in individuals with severe LC, chemotherapeutic, laser surgery, and chemotherapeutic drugs are often utilized as the first line of medication for initial malignancies. Despite the availability of several LC therapies, the recovery rate of individuals with LC continues to be poor. This is likely due to delayed detection and the ineffectiveness of the therapeutic interventions that are now in the medical sector[6]. Finding early diagnosis methods for LC with significant accuracy and precision is increasingly crucial. Therefore, a method relies on DL and ML is employed to classify CT images for the presence of LC. The article suggests the VGG-16 and Multi-Class Support Vector Machine (VGG-16+MSVM) approach.

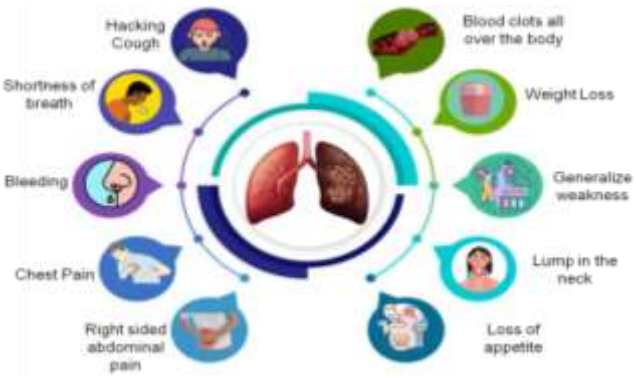

Figure 2. Symptoms of LC.

2. Literature survey

The research c suggested a new CT-scan-based image processing and artificial intelligence-based LC diagnostic method was suggested in the study of Xu et al.[7]. In the current investigation, Alexnet has been used to distinguish between malignant and healthful instances following noise reduction depending on wiener filtration. Additionally, the system makes utilizes the best possible terms for each feature that is replaced by the network feature extraction component. The Alexnet framework and extraction of characteristics in the research are designed optimally using a new customized form of the Satin Bowerbird Optimization Algorithm. This technique has high operational complexity. Data from computed tomography scans of lung patients are utilized in this research to identify and categorize pulmonary nodules as well as to determine their degree of aggressiveness. U-Net architecture is employed to divide the CT scan information. This study of Dunke and Tarade[8] proposed a 3D multi-path VGG-like network, which is tested on 3D cubes taken from a collection of lung images. This study was presented using the modified stochastic diffusion search (SDS) method to create a brand-new wrapper-based feature selection approach. The SDS would profit from agent-to-agent interactions to find the best-selected features. For categorization, the clustering algorithm, the neural network, and Naive Bayes have all been employed. Techniques using ML has been extensively employed to improve the efficacy of preclinical tumor diagnosis. A network called “study showed network (FLN)” cutting-edge ML method that uses little computing power and is quick to operate. The FLN’s existing energy variables (weight and basis),

though, are generated arbitrarily, making the algorithm unpredictable[9]. This research suggested a combination approach using FLN and "K-nearest neighbors" to classify the lungs thorax CT's structure and picture elements images and diagnose LC to increase effectiveness[10]. Because of the intricate structure and therapeutic interconnections of computer-diagnosed scanning results, physicians have trouble diagnosing LC. Physicians might benefit from Computer-Aided Detection (CAD) for rational judgment, earlier cancer detection, and categorization of malignant anomalies. In this study of Alyami et al.[11], the stages of LC are distinguished utilizing image analysis methods, and CAD has been used to improve the precision, sensitivities, and validity of automatic identification. The identification and classification of anomalies in clinical imaging are crucial for medical treatment, such as assessment, radiation, reaction analysis, and visual data study. Therefore, for accuracy and early diagnosis, Desai et al.[12] developed a completely automatic process in order to identify and morphological categorization of "non-small cell LC". Improved "Low Dose Computed Tomography (LDCT)" in conjunction with assured resolution particle swarm optimization is used to diagnose LC[13]. Because the suggested method is computerized, much reduced time was needed for both the evaluation and the preparation of the data. This methodical methodology, together with the essential detectors and the incorporation of those devices, is required to receive information about the improved LDCT scans that were performed. But this method has high energy consumption. In this study of Tumuluru et al.[14], deep convolutional neural networks were presented as a method for predicting LC at an earlier phase. The CT and MRI helped locate and diagnose the pulmonary illnesses that were present. Additionally, improved CT and MRI exams that are based on CNN to increase image quality have a significant applicability potential Upon detecting of LC. This method has poor finding LC. In the present study of Bai et al.[15], a lung tumor segmentation and identification technique are developed by making use of the suggested sine cosine Sailing Fish (SCSF) driven generated adversary network (GAN). The computed tomography (CT) scan that has been normalized is then passed on to the tumor identification classification phase. During this step, the CT scan is divided into a variety of sub-images to precisely locate the aberrant area. Through the use of the discriminator element's error mechanism, it is possible to precisely locate the afflicted areas. This method has less classification accuracy. As a result of this study of Selvapandian et al.[16], improved artificial neural network (ANN) methodologies for diagnosing lung disorders have been developed. ANN is employed to train on the dataset that has been provided. The methodology that has been suggested results in improved categorization reliability. This article of Manoharan et al.[17] concentrated on an expanded iteration of the KNN Algorithm, which is utilized for the prognosis of LC depending based on the CT scan (CT)—scanned data that are supplied as the source. This then goes through a process called Pattern Recovery, which is then proceeded by Binarization before it is sent as Performance information to the ML system. The system analyzes the screening data using the Expanded KNN Method, and then makes forecasts based on the results of that analysis. The algorithm determines the Tumor Phase depending on the input CT-Image, and this information is then sent to the physician so that additional treatment may be administered.

In order to further confirm the Early CDT-Lung test's for detecting small cell lung cancer (SCLC) in a bigger patient population, sample from this group were run on it without matched controls. Inside the validation data set, 73 SCLC samples were included[18]. The parental lung cancer's genetic heterogeneity was likewise preserved in the LCOs. This research indicates that the lung cancer organoids LCO system will serve as a helpful platform for new clinical trials and drug screening. The NBOs can also be utilized to estimate the toxicity of drugs on semi cells. The parental lung cancer's genetic heterogeneity was likewise preserved in the LCOs[19]. The requirement to offer quick, trustworthy, and affordable results from NSCLC specimens is crucial given the growing number of predictive biomarkers available for treating NSCLC patients. Immunohistochemistry (IHC) is a commonly used and less technically complex assay than molecular testing that may be successfully carried out on the majority of FFPE tissue[20]. For assessing small samples like cytology specimens, hybrid capture-based NGS techniques that can identify gene mutations as well as copy number changes and genomic structural changes may prove to be the most effective[21]. Study shows that

volumetric modulated arc therapy (VMAT) and Intensity modulated radiotherapy (IMRT) are thus advised for the care of patients with stage III NSCLC due to the possible dosimetric benefits associated with these modalities[22]. The local disease failures that were most likely to be the first sites of a recurrence were averted by the radiation treatment. Further supporting the possible advantages of local therapy in restricted metastatic settings, PFS for individuals with minimal metastatic disease appeared comparable to those of patients with a greater metastatic burden[23]. The authors have outlined the key strategies for classifying nodules and predicting lung cancer from CT imaging data. According to their observations, given enough training data, the state-of-the-art is now obtained by CNNs trained with deep learning, with classification performance in the low 90s AUC points[24]. Detection of lung cancer even at 0.11 mSv, a relatively low effective radiation dosage. New uses for FDG-PET may result from further development of this technology, which may also increase the specificity of lung cancer screening programmes[25]. Every two weeks for up to 12 months, the study of Li and Liang[26] randomly allocated participants in a 2:1 ratio to receive durvalumab (at a dose of 10 mg per kilogramme of body weight intravenously) or a placebo. One to 42 days after the patients had had chemoradiotherapy, the study medication was given to them. Identification of patients with early-stage lung cancer and the need for treatment interventions are made possible by screening and early diagnosis. Inferring the relative risks of relapse through dynamic classification of patients is made possible by prognosis prediction utilizing ctDNA. Personalizing treatment and facilitating interventions based on resistant mechanisms are made possible by evaluating treatment response and resistance[27]. The U.S. Preventive Services Task Force (USPSTF) advises lung cancer screening (LCS) with low-dose computed tomography (LDCT) in high-risk individuals, although only a small percentage of those who are eligible are screened. It is unclear if PCPs' use of LDCT is impacted by their familiarity with USPSTF recommendations[28]. Visually guided, voluntarily conducted Deep-inspirational breath-hold (DIBH) was used using optical tracking. For the purpose of planning radiotherapy, patients underwent three consecutive DIBH CT scans. In order to calculate the PTV margins, the authors examined the intrafractional errors in the position of the peripheral tumor, lymph nodes, and differential mobility between them[29]. They use a multitask deep neural network to process pre-therapy free breathing (FB) computed tomography (CT) images from 849 patients receiving lung Stereotactic Body Radiation (SBRT) to create an image fingerprint signature (or DL score) that forecasts time-to-event local failure outcomes[30].

Existing methods struggle with several drawbacks, including improper categorization, erroneous detection, increased energy consumption, and increased processing time. The previously available procedures were not sufficient to accomplish the earlier detection.

Problem statement

Globally, cancer is the non-communicable illness that is responsible for the second most fatalities. Pulmonary tumor is one of the many forms of the disease, but it is the form that accounts for the greatest number of deaths worldwide. The danger of death from LC is higher than any other kind of disease that may impact people of both genders. Throughout the unregulated expansion of exceptional cells, one side of the lung, or both, will develop to enlarge. The diagnosis of these illnesses at a preliminary phase is one of the most important steps that can be taken to protect humans from obtaining them. Numerous academics are now investigating a variety of approaches to illness forecasting in the hopes of improving reliability. However, the methods that are currently in use face several deficiencies when it comes to the categorization and detection of LC. Therefore, we recommended a VGG-16 and Multi-Class Support Vector Machine (VGG16+MSVM) to boost the classification accuracy of LC prediction utilizing CT scans. This was done so that we could make more accurate diagnoses.

3. Proposed method

The economy and global health are significantly impacted by chronic LC. Effective management of the

mortality rate and significant public health issues may result from early diagnosis and prediction of a LC diagnosis. We thus suggested a VGG-16 and Multi-Class Support Vector Machine (VGG-16+MSVM) to enhance the classification accuracy of LC prediction using CT scans. The procedure of the suggested technique is depicted in **Figure 3**.

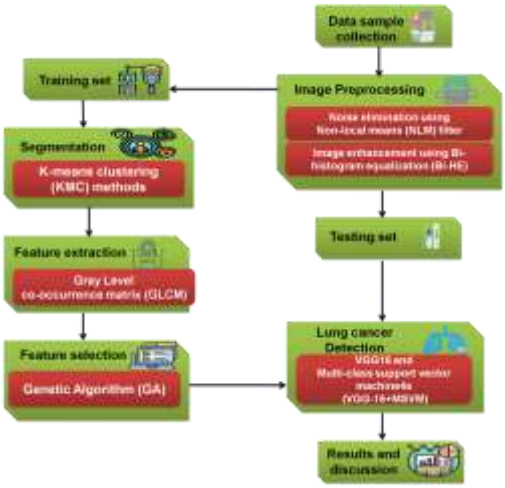

Figure 3. Procedure of the proposed method.

3.1. Contribution of the study

In this work, a method for classifying CT scans termed the VGG-16 and Multi-Class Support Vector Machine (VGG-16+MSVM) is presented. Non-Local Means (NLM) Filter and Bi- Histogram Equalization (Bi-HE) are utilized in image preprocessing to remove undesired noise data samples. For segmentation, the clustering K-Means (KMC) approach is being used. To create attributes from the segmented data, feature extraction is done using the Gray Level Co-Occurrence Matrix (GLCM). A Genetic Algorithm (GA) is employed to optimize the suggested method, selecting the best feature subsets to maximize performance.

Data set: This database included a sample of 311 BLCS people with initial NSCLC who were treated at Massachusetts General Hospital (MGH) between 1999 and 2011. The majority of individuals had first treatment for their condition. All procedures were performed in compliance with the organizational policies and standards of the hospitals. Information from the individual group's pre-resection computed tomography (CT) scanning was gathered. Additionally, information about these sufferers' total and advancement survival rates, cancer stage, and clinical and pathological findings was recorded[31].

Image preprocessing: The effectiveness of the training process might be increased by using various pre-processing techniques. Because the preprocessing reduces the amount of noise present in the source CT scans, the image quality of the system may be improved, which in turn allows for an improvement in the efficacy of the system. As a consequence, in this study, image clarity can be increased by using data pre-processing techniques such as noise reduction and image enhancement.

3.2. Non-Local means (NLM) filter used for noise removal

The Non-Local Means (NLM)filter was suggested as a way to analyze CT images corrupted by undesired noise. The resemblance between a pixel's region arrangement and that of all the other pixels in a region is taken into account by the NLM when estimating individual pixels. In Equation (1), the estimated pixel $X_{i,j}$ may be calculated numerically as the weighting factor of all the adjacent pixels in the distorted CT scan.

$$\widehat{X_{i,n}} = \frac{\sum_{n,l\epsilon\Omega_{i,j}} w_{i,j,n,m} Y_{n,m}}{\sum_{n,l\epsilon\Omega_{i,j}} w_{i,j,n,m}} \quad (1)$$

whereby $Y_{k,l}$ is the middle pixel of the similar spots in me, j, while I, j is the patched block similar to the present

patched core by $X_{i,k}$. The options in *Ij* are comparable areas that are sufficiently near to the present patchwork located on $X_{i,j}$ in Euclidean distance. The value for every selection $Y_{k,l}$ is defined by weight training $w_{i,j,k,l}$ in Equation (1), and it may be determined as pursues:

$$w_{i,j,n,m} = e \frac{\left|\left|P(X_{i,j}) - P(Y_{n,m})\right|\right|_2^2, a}{h^2} \quad (2)$$

where $P(X_{i,j})$, and $P(Y_{n,m})$ denote the local frames in the noised images that are focused on pixels $X_{i,j}$, and $(Y_{n,m})$, respectively. The smoothness variable, *h*, regulates how quickly the exponential equation decays. The Euclidean divergence, normalized by a Kernel function with a constant average and variability, is the only standard utilized in Equation (2). Since global structural resemblance is accounted for throughout denoising, in addition to localized quantitative information, NLM has successfully suppressed undesired noise.

3.3. Image quality enhancement by Bi- Histogram equalization (Bi-HE)

Bi-Histogram Equalization is a novel method that is offered to improve CT scans. The incoming distribution is split into two separate distributions using the suggested Bi-Histogram Equalization procedure, which is located at the histogram median's criterion for standard intensity maintenance. To regulate the pace of improvement, histogram trimming is done. The improved image is then created by equalizing and integrating the sub-histograms from the trimmed distribution. The Bi-Histogram Equalization allows for more trimming limitation adaptability by autonomously finding the lowest quantity among data points, means, and percentile frequencies, which preserves more of the image's content. When doing histogram equalization, the problem of over-higher bandwidth segments is addressed by automatically choosing the trimming threshold.

Segmentation: Image segmentation is a technique that is extensively utilized in digitized image processing and investigation. The goal of image segmentation is to divide CT scans into various portions or areas, and the division is often determined by the properties of the pixels included in the scans.

3.4. K-Means clustering (KMC)

A technique for grouping together a collection of data is called clustering. The k-means clustering technique is one of the most often used. A gathering of information is divided into a group of variables with k groups in k-means clustering. It divides the provided set of data into k distinct clusters. The K-means technique comprises two distinct stages. The k centroids are calculated in the initial stage, and then in stage 2, every pixel is moved to the cluster with the centroids that are closest to it. The most widely frequent way for determining the proximity in the vicinity centroids is the Euclidean distance. Once the clustering is complete, it recalculates the centroids of every cluster, calculates an updated Euclidean distance between every core and every part of information depending on those centroids, and gives the pattern elements Its Euclidean distance is the shortest. The component entities and centroids of every group in the division serve as its defining characteristics. The position at which the total distances from all the items in a cluster are reduced is the centroids for every group. K-means, then, is an iteration method it shortens the overall distance between each item and its cluster centroids over all groups.

Let us assume an image that has a quality of $x \times y$, and the CT scans have to be clustered into k different clusters. Let CK be the clustering centers and $p(x, y)$be the incoming images to be clustered. The following is a description of the k-means clustering method shown in Algorithm 1.

The random choice of the starting centroids affects how well the clustering findings turn out in the end. As a consequence, if the starting centroids are picked arbitrarily, the outcome will vary depending on the beginning center. So that we may achieve the segmentation we want, the starting center will be appropriately selected. It is dependent on the number of information points, groups, and iterations. Thus segmentation of the CT scan is done using KMC.

Algorithm 1 K-Means clustering

1: Establish the cluster's k-count and centroid.

2: Utilizing the connection shown in Equation (3), determine the euclidean distance d between the center and every pixel of a CT scan. $d = ||p(x,y)||$CK (3)

3: Depending on the distance d, allocate each pixel to the center that is closest to it.

4: When all of the pixels have been allocated, reevaluate the center's location by applying the correlation shown in Equation (4) below: $c_k = \frac{1}{k}\sum_{y \in c_k}\sum_{x \in C_k} p(x,y)$ (4)

5: Continue the procedure till the tolerances or defect amount is achieved.

6: Resize the pixels in the images cluster.

Feature extraction: Feature extraction is the procedure of converting unprocessed images into quantitative features from the source database. In comparison to introducing ML to the original data, autonomous feature extraction utilizes specific techniques to automatically retrieve characteristics from CT scans. This technique might useful when images need to quickly switch from generating raw data to artificial intelligence algorithms. "Gray Level Co-Occurrence Matrix (GLCM)" feature extraction was carried out in this study.

3.5. Gray level co-occurrence matrix (GLCM)

The identification and categorization of LC are challenges that are resolved by the "Gray Level Co-Occurrence Matrix (GLCM)", a feature extraction method that works with the CT images which have been obtained in the database. Creating a robust CT scan collection is often the initial stage of LC screening. The result of GLCM utilized to categorize the illnesses includes surface characteristics, training, and testing. All CT scans, including light, moderate, high, flow, and excess scans, are available in the CT scan collection. The varying degrees of CT scan is used to create a more reliable and usable mentoring and evaluation dataset for classification validation. The GLCM offers a second-order approach for creating feature extraction to determine the correlation between the different grayscale variations in the CT scan properties, such as distance, *d*, and direction, *q*. Gray levels (*I j*) orientated at *q* 14 0 and *q* 14 180, respectively, constitute the GLCM matrices. As a result, entries were formed at (*I j*) and (*j, I)*, and every GLCM was modeled to the quantity *G* of the quantified grey scale.

GLCM metrics may be used to produce a variety of image characteristics. Therefore, the measurements need a possibility rather than quantity prior pattern characteristics can be estimated. In Equation (5), the probabilistic measurement is derived.

$$P_r(X) = C_{ij}(d,\theta) \tag{5}$$

where Equation (6) defines the co-occurrence possibility (C_{ij}) between grayscale *i* and *j*.

$$C_{ij} = \frac{P_{ij}}{\sum_{ij=1}^{G} P_{ij}} \tag{6}$$

where P_{ij} stands for how many instances of grayscale *i* and *j* there are in a set of *d*, *q*, and *G* factors.

The intensity, power, uniformity, and coherence of grayscale threshold readings are the pattern descriptors produced by GLCM. It is defined in Equations (7)–(10). While power in GLCM is the total of the square components, brightness refers to the degree of local differences contained in images. The angular second component or homogeneity is another name for power. The uniformity description indicates how closely the dispersion of GLCM components resembles the diagonally of GLCM. A pixel's association with its neighbors throughout the entire image will be shown through coherence, which is the final step.

$$\text{Intensity} = \sum_{i,j=0}^{N-1} P_{ij}\,(i-j)^2 \tag{7}$$

$$\text{Power} = \sum_{i,j=0}^{N-1} \left(P_{ij}\right)^2 \tag{8}$$

$$\text{Uniformity} = \sum_{i,j=0}^{N-1} \frac{P_{ij}}{1 + (i - j^2)} \tag{9}$$

$$\text{Coherence} = \sum_{i,j=0}^{N-1} P_{ij} \frac{\mu}{\sigma^2} \tag{10}$$

- P_{ij} is the standardized symmetrically GLCM's element *i* and *j*.
- N=the number of grayscales in the image as determined by the GLCM's quantification of the degrees in the image.
- μ=The GLCM average.
- σ^2=The dispersion of all comparison pixels' brightness in the correlations that made up the GLCM.

Numerous studies have been done to examine various elements of co-occurrence textured characteristics connected to the parameters G, d, and q, as well as their applicability in this context. These researches were carried out to provide suggestions for the appropriate empirical system criteria. Given that 0, 45, 90, and 135 are considered to produce more accurate categorization by several investigators, the roles of G and d are investigated in this work.

Feature selection: Feature selection is a method of image quantization that includes selecting merely the most crucial traits from a complete collection and dismissing the rest. A method of feature selection that selects features based on the general attributes of the training data. For feature selection, the Genetic Algorithm (GA) method was suggested. The procedure of the genetic algorithm is shown in **Figure 4**.

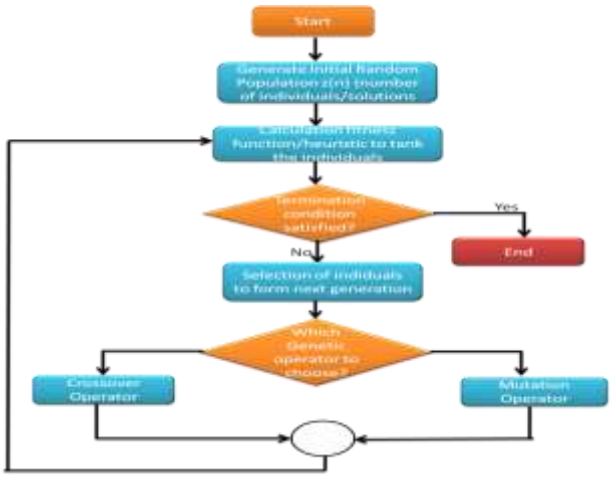

Figure 4. Procedure of genetic algorithm.

3.6. Genetic Algorithm (GA)

All characteristics are taken into consideration when using a genetic algorithm as a feature selection strategy. The primary objective of the feature selection strategy is to improve overall efficiency through refinement. The fundamental task when using genetic programming for features choice is to produce the optimal characteristic subgroup by the fitness function. GA begins with population activation at randomization. New people are chosen in each generation based on the fitness function's score. We used rank-based fitness selection in this investigation. The CT image that recombines with the remaining people to create the next iteration is chosen by the selecting mechanism following fitness allocation. CT images are chosen arbitrarily using a roulette roll choice process, which makes use of the motor's rotating motion. The crossover process, which is in charge of creating the new fittest subgroups known as the offspring of founders, was completed by

arbitrarily selecting the two guardians. However, a subgroup (singular) that arbitrarily modifies certain desired quantities of characteristics for the existence of the fittest is what makes the mutation operation function.

LC detection by utilizing VGG-16 and Multi-Class Support Vector Machine (VGG-16+MSVM): The well-known pretrained DL model (VGG-16) configuration serves as the foundation for the suggested methodology. Considering two factors, we recommend the VGG-16 model. First, contrasted to its other equivalent, the VGG-19 framework, it recovers the characteristics at a reduced rate by employing a reduced kernel size, which is suited for CT images with fewer layers. It also offers a stronger feature extraction capability for classifying CT scans of LC. One of the transferable training strategies is the fine-tuning strategy, that we utilize. We employ the pre-trained weights of ImageNet to interact with the VGG-16 classifier during the fine-tuning procedure. Given that there are only a certain number of CT images available for training, it assists in overcoming the overfitting issue. The "Attention module, Convolution module, FC-layers, and Softmax classifier" are the four key building elements that make up the suggested approach (also called "Attention-based VGG-16"). **Figure 5** displays the comprehensive schematic diagram of the suggested architecture.

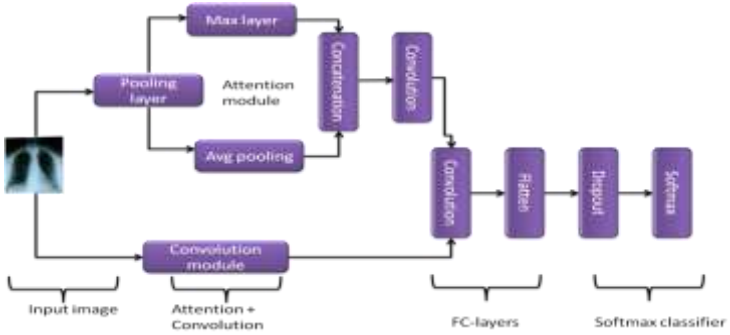

Figure 5. Schematic diagram of the VGG-16 architecture.

3.7. Attention module

This component allows scientists to record the temporal connection between visual information included in lung CT scans. We utilize the idea of temporal concentration for this. The source vector, which corresponds to the fourth pooling layer of the VGG-16 framework in the technique, is subjected to both maximum and mean pooling operations. Then, utilizing the Sigmoid function (σ), these two resulting tensors-the maximum pooled 2D vector and the mean pooled 2D vector-are connected to conduct a convolution with a filter size (f) of 7×7. Equation (11) defines the conjugated resulting vector (Ms (F)).

$$M_s(F) = \sigma\left(f^{7\times7}\left[F^s_{avg}; F^s_{max}\right]\right) \tag{11}$$

where the 2D vector obtained by mean pooling and maximum pooling operations on the source vector F, correspondingly, are denoted by $F^s_{avg} \in R^{1\times H\times W}$ and $F^s_{max} \in R^{1\times H\times W}$. Here, the terms H and W stand for the vector's length and breadth, correspondingly.

3.8. Convolution module

The convolution module, the fourth pooling layer of the VGG-16 architecture, is what we employ in the approach. The important hints of the CT scans are captured by the size-independent convolutional component. The middle-level layer, or fourth pooling, which is better suited to CT scans, is where the intriguing hints are recovered. Nevertheless, since such CT scans are neither broader nor more particular, the characteristics from other layers (greater or lesser) are inappropriate for CT imaging. As a result, we begin by providing the attention module with the fourth pooling layer. The output of this component is then combined with the actual fourth pooling layer.

3.9. The fully connected (FC) layer

We employ fully connected layers to describe the concatenation characteristics obtained from the convolutional blocks and attentiveness as one-dimensional (1D) information. According to **Figure 2**, it has 3 components: Flatten, dropout, and dense. In this technique, the dense layer is set at 256 and the dropout is fixed at 0.5.

3.10. Softmax classifier

We employ the softmax layer to categorize the characteristics that were collected from the FC layers. The component value for the softmax layer-the final dense layer-depends on the number of classes in the database (e.g., three for datasets with 3 groups, four for datasets with four types, etc.). Depending on the categorization that was done, the softmax layer produces multivariate regression distributions of the probabilities values.

$$P\left(\frac{a=c}{b}\right)=\frac{e^{b}{}_{k}}{\sum_{j} e^{b}{}_{j}} \tag{12}$$

Equation (12) defines the outcome of this probability, where *b* and *c* stand for possibilities obtained from the softmax layer and a particular class of the database utilized in our suggested technique, accordingly.

3.11. Multi-Class Support Vector Machine

A technique for supervised ML known as "Multi-Class Support Vector Machine (MSVM)" is useful for both the categorization and reconstruction of challenge statements. It searches for support vectors, or ideal borders, between several categories. The MSVM classifier is used to accurately classify CT lung scans once the characteristics from the prior step have been filtered. A classifying challenge is considered using a collection of n examples, a test set is indicated as $yi \in \{-1+1\}$ and a target class example matrix is denoted as

$S=(\mathrm{x_i},\mathrm{y_i}),(i=1,2,\ldots,n)$ $\mathrm{x_i}\epsilon R$. To discriminate between negative (−1) and positive (+1) occurrences while exploring support vectors, instruction is done. Equation (13) illustrates how the training step entails the optimization method.

$$\emptyset(w,\varepsilon)=\frac{1}{2}||w||^{2}+C\sum_{i=1}^{n}\varepsilon_{i} \tag{13}$$

Considering the restrictions:

$$y_{i[(w.x_i)+b]\geq 1-\varepsilon_{ii=1,2,\ldots,n}} \tag{14}$$

where *C* stands for a compensation element and ∅ denotes the mappings of the e source vector to a greater dimensional feature set, and where ε_i stands for the slack parameter, whose value is ≥0 to quantify the classification error $\sum_{i=1}^{n}\varepsilon_i$. The entire procedure is a simple, high-dimensional, continuously separable problem, and the transformation is dependent on the MSVM kernel operator.

$$K\left(x_i,x_j\right)=\emptyset(x_i)^{T}.\emptyset\left(x_j\right) \tag{15}$$

The kernel procedure, which produces standard information as linearly separated information so that dimensionality concerns may be readily eliminated, is essential. Equation (10) can be utilized to construct the many different kinds of kernel operations, including the radically base factor (RBF), quadratic, and polynomial kernel processes.

$$K\left(x_i,x_j\right)=x_i^{T}.x_j \tag{16}$$

$$K\left(x_i,x_j\right)=\left(x_i^{T}.x_j+1\right)^{\delta} \tag{17}$$

$$K\left(x_i,x_j\right)=exp\left(\frac{-|\left|x_i-x_j\right||^{2}}{\delta^{2}}\right) \tag{18}$$

Thus, the detection of LC using CT scans was done by a hybrid VGG-16 and Multi-Class Support Vector

Machine (VGG16+MSVM) to enhance the earlier detection accuracy.

4. Result and discussion

A technique based on ML and DL is employed to classify CT scans for the presence of LC. This paper suggests the use of the Visual Geometry Group (VGG-16) and Multi-Class Support Vector Machine (VGG-16+MSVM) approach for the precise and early identification of LC. The efficacy of the suggested approach is covered in this section. The recommended system's capacity to achieve Accuracy, Precision, Recall, F1-Score, Sensitivity, and Specificity justifies its adoption. The classic approaches used for comparison include artificial intelligence (AI), "denoising first two-path convolutional neural networks (DFD-Net)", "fusion-based convolutional fuzzy neural networks (F-CFNN)" and "Wilcoxon Signed Generative DL (WS-GDL)".

Accuracy: An evaluation of the model's accuracy is obtained by multiplying the total number of predictions by the number of VGG-16+MSVM correct detections. Accurate evaluations in VGG-16+MSVM must consider both positive and adverse outcomes. The quantity to which a prediction can be made with complete certainty is referred to as its accuracy, and the measure to which it can approximately anticipate the outcome is referred to as its accuracy. To determine how accurate the detection was, a calculation was made using the ratio of the predicted result to the true result. When compared to other conventional procedures, our recommended method gives a high degree of accuracy in detecting LC. **Figure 6** represents an evaluation of the accuracy.

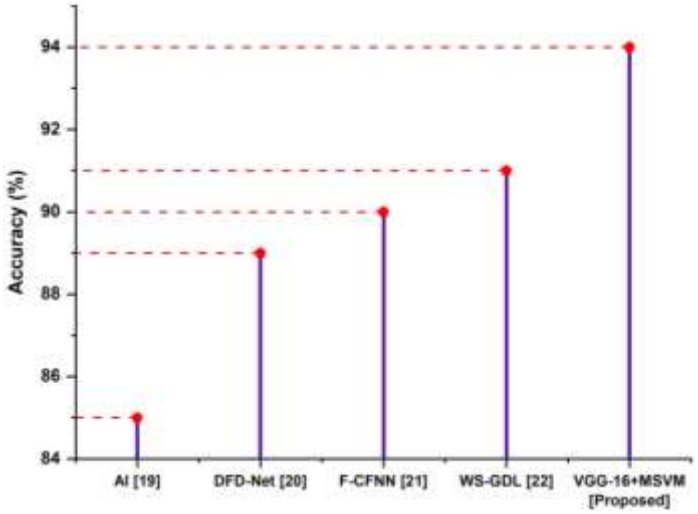

Figure 6. Comparative analysis of Accuracy with existing methods.

Precision: The percentage of detection that concentrates on important components of LC is referred to as "precision." The proportion of correctly detecting LC is known as precision. It can mean that precision is the criterion for quality. Precision is the average possibility of appropriate detection. As a consequence, the procedure that is presently advised is more accurate than those that were previously used. The precision of the recommended method is shown in **Figure 7**.

Recall: A recall is one of the characteristics that is taken into account while assessing medical approaches. The percentage of reliable LC diagnoses using CT scans is often known as a recall. The true positive rate is often referred to as recall. As a result, the recommended strategy is more effective than the existing techniques. **Figure 8** demonstrates that the model has a high recall rate.

F1-Score: By summing the precision and recall scores and computing their harmonic means, the F1 score is obtained. The average of something like the Precision and Recall scores, weighted computations is utilized to get the F1 score. These two elements each have an equal impact on the outcome. The method we suggest provides a high degree of F1 score when it comes to detecting LC in contrast to the other methods that have

previously been employed. **Figure 9** depicts the F1-Scores for the proposed methodology and the previously used methods.

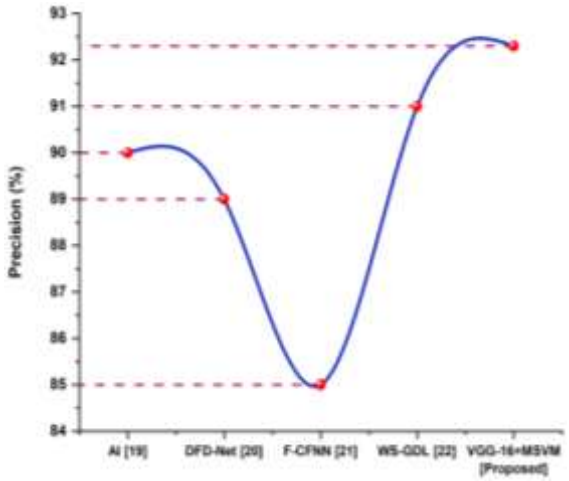

Figure 7. Comparative analysis of precision with existing methods.

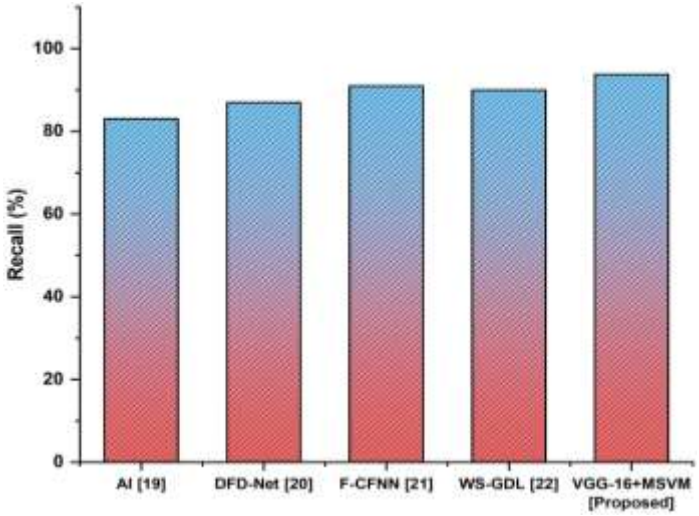

Figure 8. Comparative analysis of recall with the existing method.

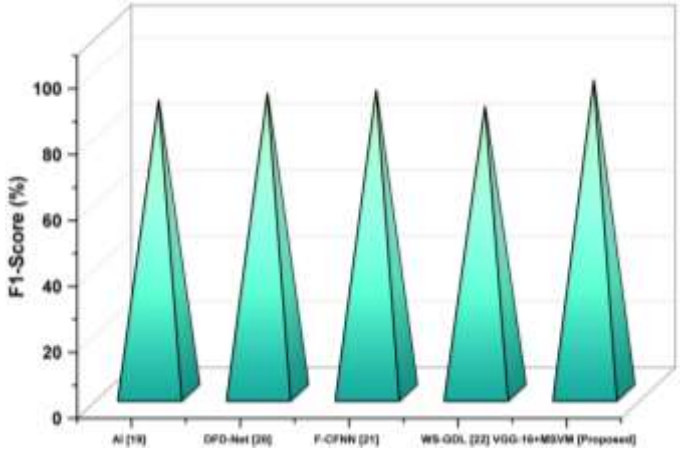

Figure 9. Comparative analysis of F1-Score with existing methods.

Specificity: The capacity of a test to rule out someone who has a condition as negative when does not have a disease is referred to as specificity. The specificity demonstrates the efficiency of LC detection. It demonstrates the effectiveness of the suggested approach. **Figure 10** illustrates the Suggested Technique's Specificity. The recommended method is thus more effective than the traditional one.

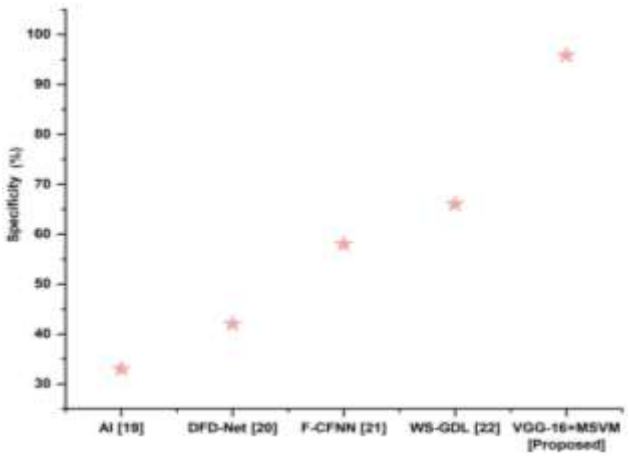

Figure 10. Comparative analysis of specificity with existing methods.

Sensitivity: Sensitivity is the capacity of a diagnostic to identify a diseased person as positive. The proportion of CT scans that are segmented and provide a correct result when the assessment is employed in the study is the genuine optimistic proportion, also called the sensitivity of detection of LC. **Figure 11** displays the sensitivity of the suggested strategy. As a result, the recommended method is more effective than the existing systems.

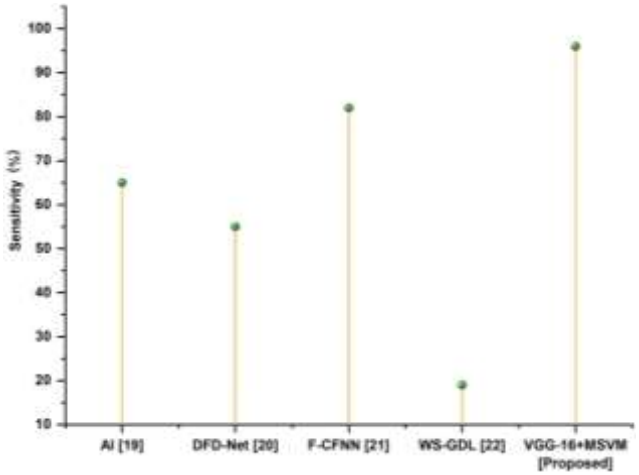

Figure 11. Comparative analysis of sensitivity with existing methods.

Discussion: One of the most fatal malignancies and one with a high prevalence in the community is LC. As a result, Chaunzwa et al.[32] proposed an artificial intelligence-based diagnostic method, to determine when lung nodular micro-calcification first occurs, which may help physicians and surgeons forecast it with accuracy using CT scan processing techniques. The morphology of nodules, such as contour and dimensions, and image distortion have an indirect and complicated connection with cancer; for this reason, a detailed investigation of every suspicious nodule and the integration of data from every lesion should be necessary. To tackle the difficulty of LC diagnosis, Maurer[33] developed the "denoising first two-path convolutional neural network (DFD-Net)". The newly presented model is made up entirely of denoising and detecting components. Enhancing CT's ability to accurately diagnose or identify lung disease is a difficult challenge. As a result, the "fusion-based convolutional fuzzy neural network (F-CFNN)" proposed in this research of Sori et al.[34], which recognizes and categorizes CT scans, is fusion-based convolution fuzzy. The convolutional fuzzy neural network (CFNN) in the F-CFNN combines two convolutional layers, max pooling, and a fuzzy neural network to retrieve information and produce reliable categorization results. This study of Lin and Yang[35] introduced the "Wilcoxon Signed Generative DL (WS-GDL)" approach for identifying LC. First, test informational gain and relevance analysis remove superfluous and unimportant qualities and recover numerous essential and

instructive characteristics. The deep characteristics are then learned by the need for a generator function technique. It consumes more energy. These techniques are less detection rate and classification accuracy.

5. Conclusion

The deadliest disease, LC harms both men and women similarly. It is a dangerous kind of cancer that is difficult to diagnose. The main airways, the windpipe, the lungs, or another area may be where LC develops. Uncontrolled cell growth and the proliferation of certain cells in the lungs are the causes of LC. One of the most crucial actions that can be made to prevent human death is the early detection of these disorders. Thus, a methodology based on ML and DL is used to classify CT images for the presence of LC. Therefore, we presented a VGG-16 and Multi-Class Support Vector Machine (VGG16+MSVM) to improve the detection accuracy of LC prediction using CT scans. In terms of Accuracy, Precision, Recall, F1-Score, Sensitivity, and Specificity, the suggested approach performed better. This criterion was compared to traditional methods such as Wilcoxon Signed Generative DL, "denoising first" two-path convolutional neural networks, and "denoising first" two-path convolutional neural networks (WS-GDL). It demonstrates that the suggested method is more successful in detecting LC. The next potential focus of the proposed study may be the application of optimization approaches to enhance performance indicators like computation speed and detection quality.

Author contributions

Conceptualization, SKH and SB; methodology, SKH; software, KVD; validation, KM, SB and KVD; formal analysis, KLK; investigation, PS; resources, VDJ; data curation, VDJ; writing—original draft preparation, KM; writing—review and editing, KLK; visualization, PS; supervision, KVD; project administration, SKH; funding acquisition, SB. All authors have read and agreed to the published version of the manuscript.

Conflict of interest

The authors declare no conflict of interest.

References

1. Abdullah DM, Ahmed NS. A review of most recent LC detection techniques using machine learning. International Journal of Science and Business. 2021; 5(3): 159-173.
2. Wang X, Guo Y, Liu L, et al. YAP1 protein expression has variant prognostic significance in small cell lung cancer (SCLC) stratified by histological subtypes. Lung Cancer. 2021; 160: 166-174. doi: 10.1016/j.lungcan.2021.06.026
3. Sardarabadi P, Kojabad AA, Jafari D, et al. Liquid Biopsy-Based Biosensors for MRD Detection and Treatment Monitoring in Non-Small Cell Lung Cancer (NSCLC). Biosensors. 2021; 11(10): 394. doi: 10.3390/bios11100394
4. Saab MM, McCarthy M, O'Driscoll M, et al. A systematic review of interventions to recognise, refer and diagnose patients with lung cancer symptoms. npj Primary Care Respiratory Medicine. 2022; 32(1). doi: 10.1038/s41533-022-00312-9
5. Wang X, Ricciuti B, Nguyen T, et al. Association between Smoking History and Tumor Mutation Burden in Advanced Non–Small Cell Lung Cancer. Cancer Research. 2021; 81(9): 2566-2573. doi: 10.1158/0008-5472.can-20-3991
6. Xu K, Zhang C, Du T, et al. Progress of exosomes in the diagnosis and treatment of lung cancer. Biomedicine & Pharmacotherapy. 2021; 134: 111111. doi: 10.1016/j.biopha.2020.111111
7. Xu Y, Wang Y, Razmjooy N. Lung cancer diagnosis in CT images based on Alexnet optimized by modified Bowerbird optimization algorithm. Biomedical Signal Processing and Control. 2022; 77: 103791. doi: 10.1016/j.bspc.2022.103791
8. Dunke SR, Tarade SS. LC Detection Using Deep Learning. International Journal of Research Publication and Reviews.
9. Shanthi S, Rajkumar N. Lung Cancer Prediction Using Stochastic Diffusion Search (SDS) Based Feature Selection and Machine Learning Methods. Neural Processing Letters. 2020; 53(4): 2617-2630. doi: 10.1007/s11063-020-10192-0
10. Lin CJ, Yang TY. A Fusion-Based Convolutional Fuzzy Neural Network for LC Classification. International

Journal of Fuzzy Systems. 2022; 1-17.
11. Alyami J, Khan AR, Bahaj SA, et al. Microscopic handcrafted features selection from computed tomography scans for early stage lungs cancer diagnosis using hybrid classifiers. Microscopy Research and Technique. 2022; 85(6): 2181-2191. doi: 10.1002/jemt.24075
12. Desai U, Kamath S, Shetty AD, Prabhu MS. Computer-Aided Detection for Early Detection of LC Using CT Images. In: Intelligent Sustainable Systems. Springer; 2022.
13. Primakov SP, Ibrahim A, van Timmeren JE, et al. Automated detection and segmentation of non-small cell lung cancer computed tomography images. Nature Communications. 2022; 13(1). doi: 10.1038/s41467-022-30841-3
14. Tumuluru P, Hrushikesava Raju S, Santhi MVBT, et al. Smart LC Detector Using a Novel Hybrid for Early Detection of LC. In: Inventive Communication and Computational Technologies. Springer; 2022.
15. Bai Y, Li D, Duan Q, et al. Analysis of high-resolution reconstruction of medical images based on deep convolutional neural networks in lung cancer diagnostics. Computer Methods and Programs in Biomedicine. 2022; 217: 106592. doi: 10.1016/j.cmpb.2021.106592
16. Selvapandian A, Prabhu SN, Sivakumar P, et al. Lung Cancer Detection and Severity Level Classification Using Sine Cosine Sail Fish Optimization Based Generative Adversarial Network with CT Images. The Computer Journal. 2021; 65(6): 1611-1630. doi: 10.1093/comjnl/bxab141
17. Manoharan H, Rambola RK, Kshirsagar PR, et al. Aerial Separation and Receiver Arrangements on Identifying Lung Syndromes Using the Artificial Neural Network. Computational Intelligence and Neuroscience. 2022; 2022: 1-8. doi: 10.1155/2022/7298903
18. Sutedja G. New techniques for early detection of lung cancer. European Respiratory Journal. 2003; 21(Supplement 39): 57S-66s. doi: 10.1183/09031936.03.00405303
19. Teixeira VH, Pipinikas CP, Pennycuick A, et al. Deciphering the genomic, epigenomic, and transcriptomic landscapes of pre-invasive lung cancer lesions. Nature Medicine. 2019; 25(3): 517-525. doi: 10.1038/s41591-018-0323-0
20. Jain D, Roy-Chowdhuri S. Molecular Pathology of Lung Cancer Cytology Specimens: A Concise Review. Archives of Pathology & Laboratory Medicine. 2018; 142(9): 1127-1133. doi: 10.5858/arpa.2017-0444-ra
21. Dong Z, Li H, Zhou J, et al. The value of cell block based on fine needle aspiration for lung cancer diagnosis. Journal of Thoracic Disease. 2017; 9(8): 2375-2382. doi: 10.21037/jtd.2017.07.91
22. Peng J, Pond G, Donovan E, et al. A Comparison of Radiation Techniques in Patients Treated With Concurrent Chemoradiation for Stage III Non-Small Cell Lung Cancer. International Journal of Radiation Oncology*Biology*Physics. 2020; 106(5): 985-992. doi: 10.1016/j.ijrobp.2019.12.027
23. Lindeman NI, Cagle PT, Aisner DL, et al. Updated Molecular Testing Guideline for the Selection of Lung Cancer Patients for Treatment with Targeted Tyrosine Kinase Inhibitors. Journal of Thoracic Oncology. 2018; 13(3): 323-358. doi: 10.1016/j.jtho.2017.12.001
24. Kadir T, Gleeson F. Lung cancer prediction using machine learning and advanced imaging techniques. Translational Lung Cancer Research. 2018; 7(3): 304-312. doi: 10.21037/tlcr.2018.05.15
25. Schwyzer M, Ferraro DA, Muehlematter UJ, et al. Automated detection of lung cancer at ultralow dose PET/CT by deep neural networks – Initial results. Lung Cancer. 2018; 126: 170-173. doi: 10.1016/j.lungcan.2018.11.001
26. Li RY, Liang ZY. Circulating tumor DNA in lung cancer: real-time monitoring of disease evolution and treatment response. Chinese Medical Journal. 2020; 133(20): 2476-2485. doi: 10.1097/cm9.0000000000001097
27. Mazzone PJ, Silvestri GA, Patel S, et al. Screening for Lung Cancer. Chest. 2018; 153(4): 954-985. doi: 10.1016/j.chest.2018.01.016
28. Josipovic M, Aznar MC, Thomsen JB, et al. Deep inspiration breath hold in locally advanced lung cancer radiotherapy: validation of intrafractional geometric uncertainties in the INHALE trial. The British Journal of Radiology. 2019; 92(1104). doi: 10.1259/bjr.20190569
29. Pastorino U, Silva M, Sestini S, et al. Prolonged lung cancer screening reduced 10-year mortality in the MILD trial: new confirmation of lung cancer screening efficacy. Annals of Oncology. 2019; 30(7): 1162-1169. doi: 10.1093/annonc/mdz117
30. Teo PT, Bajaj A, Randall J, et al. Deterministic small-scale undulations of image-based risk predictions from the deep learning of lung tumors in motion. Medical Physics; 2022.
31. Ajitha E, Diwan B, Roshini M. March. LC Prediction using Extended KNN Algorithm. In: 2022 6th International Conference on Computing Methodologies and Communication (ICCMC).
32. Chaunzwa TL, Hosny A, Xu Y, et al. Deep learning classification of lung cancer histology using CT images. Scientific Reports. 2021; 11(1). doi: 10.1038/s41598-021-84630-x
33. Maurer A. An Early Prediction of LC using CT Scan Images. Journal of Computing and Natural Science. 2021; 39-44.
34. Sori WJ, Feng J, Godana AW, et al. DFD-Net: lung cancer detection from denoised CT scan image using deep learning. Frontiers of Computer Science. 2020; 15(2). doi: 10.1007/s11704-020-9050-z
35. Lin CJ, Yang TY. A Fusion-Based Convolutional Fuzzy Neural Network for LC Classification. International Journal of Fuzzy Systems. 2022.