Classification and Segmentation of Pulmonary Lesions in CT Images Using a Combined VGG-XGBoost Method, and an Integrated Fuzzy Clustering-Level Set Technique

Niloofar Akhavan Javan¹, Ali Jебreili², Babak Mozafari³, Morteza Hosseinioü⁴, and S. AmirAli Gh. Ghahramani⁵

¹, ², ³Department of Computer Engineering, Khayyam University
⁴, ⁵School of Engineering and Science, Department of Computer Engineering, Sharif University of Technology

Abstract—Given that lung cancer is one of the deadliest illnesses, early identification and diagnosis are critical to preserving a patient’s life. However, lung illness diagnosis is time-intensive and requires the expertise of a pulmonary disease specialist, subject to a significant rate of inaccuracy. Our objective is to design a system capable of accurately detecting and classifying lung lesions and segmenting them in CT-scan images. The suggested technique extracts features automatically from the CT-scan image and then classifies them using Ensemble Gradient Boosting methods. Finally, if a lesion is detected in the CT-scan image, it is segmented using a hybrid approach based on Fuzzy Clustering and Level Set [1]. To train and test our models we gathered a dataset that included CT images of patients residing in Mashhad, Iran. Finally, the results indicate 96% accuracy within this dataset. This approach may assist clinicians in diagnosing lung abnormalities and avoiding potential errors.

Index Terms—Pulmonary Lesion Classification and Segmentation, Deep Learning, VGG Convolutional Neural Networks, XGBoost, Level Set Methods.

I. INTRODUCTION

Cancer is a group of diseases that are characterized by the growth of uncontrollable abnormal cells. If the spread of the abnormal cell is not controlled, it can lead to death. However, the disease’s cause is unknown for many cancers, especially those that occur during childhood. Many factors that cause cancer are known, including lifestyle factors such as smoking, overweight and unmodifiable factors, such as mutations hereditary, hormonal, and immune conditions. These risk factors may be associated simultaneously or continuously to initiate and/or promote cancer growth. In 2020, the American Cancer Societys lung cancer measures in the United States indicated roughly 228,820 new cases of lung cancer amongst 116,300 in men and 112,520 in women and approximately 135,720 mortality from lung cancer in between 72,500 in men and 63,220 in women [2]. Lung cancer is the second most common cancer in women and men (irrespective of skin cancer) and is one of the most important causes of cancer deaths among men and women. The mortality rate from lung cancer is higher than colon, breast, and prostate cancer altogether. However, the detection of small lung nodules from volumetric CT-scans is also tricky, and for this reason, many CAD tools are designed to compensate for this problem [3], [4]. If lung cancer is detected at an early stage, when it is small and has not spread yet, a person has a greater chance of living. Computerized diagnosis tools (CAD) are used to create a classification between natural and abnormal lung tissue that may improve the ability of the radiologist [5], [6]. The onset of lung cancer begins in the lungs, while secondary lung cancer begins elsewhere in the body and reaches the lungs. A pulmonary nodule is an oval or round growth in the lungs. The size of the nodules varies from a few millimeters to 5 centimeters. Given the shape and size of the nodule, classifying is a challenging task. Detection of large-sized malignant nodules is straightforward, but we have difficulty identifying small malignant nodules [7]. The lung cancer death rate has weakened by 45% since 1990 in men and by 19% since 2002 in women due to cutbacks in smoking, with the pace of decline quickening over the past decade; from 2011 to 2015, the rate decreased by 3.8% per year in men and by 2.3% per year in women [8]. According to Cancer Research UK, the five-year survival rate for patients diagnosed in stage one is more than 55%, while the survival rate in patients with lung cancer in stage four is almost 5% [9]. Computer-aided diagnosis systems (CAD) are effective schemes for identifying and detecting various pulmonary lesions. The main purpose of these systems is to assist the radiologist in various stages of diagnosis. The CAD system output acts as the second opinion for radiologists before the final diagnosis.
way, researchers are developing more auto CAD systems for lung cancer. Many different publications have provided auto nodule detection systems using image processing, including various features extraction, classification, and segmentation techniques.

II. RELATED WORKS

The diagnosis of pulmonary lesions is a very important topic, and a lot of work has been done in this field. However, due to many different types of pulmonary lesions and difficulty of diagnosis in this field, they are constantly seeking to increase the accuracy of existing systems. Thus, we have tried to design and train a more accurate system than the existing systems. Various existing works are as follows: Ying Xie et al. applied an interdisciplinary mechanism based on metabolomics and six machine learning methods and reached the sensitivity of 98.1%, AUC in 0.89, and Specificity in 100.0%. As a result, the machine learning methods are AdaBoost, K-nearest neighbor (KNN), Naive Bayes, Support Vector Machine (SVM), Random Forest, and Neural Network. They also recommended Naive Bayes as a suitable method [10]. Also, Netto et al. worked on the automatic separation of pulmonary nodules with growing neural gas and SVM. Their purpose was to automatically collect lung nodules through computed tomography images using the growth neural gas (GNG) algorithm to isolate structures with very similar properties to the lung nodules. They then used the distance conversion to separate the partitioned structures that connect the blood vessels and bronchitis. Finally, they used a set of features of the shape and texture using the SVM classifier to classify these structures: lung nodules [11]. Additionally, Lee et al. used a collection of classifiers called random forest in two stages. The first stage was diagnosing lung nodules and the second stage was false positive reduction [12]. Moreover, Yi et al. worked with five attributes, including intensity information, shape index, 3D space, and location. They worked more on segmentation issues and reached 81% accuracy [13]. Javid et al. worked on identifying nodes such as heart and muscle nodes. A brief analysis of CT histograms is performed to select an appropriate threshold for better results. A simple morphological closing is used in the segmentation of the lung area. The K-means clustering is applied for the initial detection and segmentation of potential nodes. This segmentation eventually reached a sensitivity of 91.65% [5].

III. MATERIAL

To train and evaluate the model presented in this article, we have been preparing a dataset, including CT-scan images of the local patients’ pulmonary lesions. More than 10000 slides were used to prepare this dataset. All images were tagged and classified by specialist physicians. There are a large number of various types of lesions in this dataset. Due to the large size of the model, there is a need for the right hardware resources to train these models, hardware resources, and sufficient time was provided, and all steps were successfully completed.

IV. PROPOSED METHOD

Our proposed model is a complete and automatic system for the classification and segmentation of pulmonary lesions. In the first stage, a deep convolutional neural network automatically extracts features from CT-scan images. In the second stage, based on extracted features, an Ensemble Gradient boosting classifier identifies pulmonary lesions. Finally, in the third stage, CT-scan images are segmented by a hybrid fuzzy clustering Level set method based on [I]. Among the proposed system’s positive features, the reduction of diagnosis time and high accuracy can be noted. The proposed method is a complete CAD system. After receiving the image, it automatically performs all the steps of extracting features, classification, and segmentation and provides the final output. Extracting a feature involves extracting a higher level of information from raw pixel values that differentiate between different categories. Based on specific algorithms such as HOG, Haar, SIFT, LBP, GIST, the image features are extracted in classic methods. Based on these features extraction, the classifier is trained, and the final model is achieved. Some of these classification modules like SVM, logistic regression, random forest, KNN can be noted. One of the problems of classic methods is choosing and designing a suitable feature extraction method. A variety of different methods have been proposed over the years. Each method often works well in a particular field, and for new issues, there is a need to improve and change the feature extraction technique. The problem with traditional methods is that the feature extraction method cannot be set based on classes and images. Therefore, if the selected feature does not have the abstract needed to identify the categories, regardless of the type of classification strategy used, the classification model’s accuracy will be very low. The problem with classic methods is always to find a distinctive feature among several features. Also, achieving accuracy, such as human accuracy, has been a big challenge. That is why it took years to have a flexible computer vision system (such as OCR, face recognition, image categorization, and object recognition) that works with various data. Another problem with these methods is that it is entirely different from how we learn to recognize things. Immediately after the birth of a child, he cannot understand his surroundings, but with the advancement and processing of data, he learns to identify things. This philosophy is behind deep learning. A computational model is trained in deep learning based on existing datasets and then automatically extracts the best features. Most deep neural networks require a lot of memory and computation, especially when training. Hence, this is an essential concern in these networks.

A. VGG Convolutional Neural Network

At present, convolutional neural networks have managed to surpass humans in computer vision tasks, such as image classification. The image classification means determining which image belongs to which class. In the VGG network [13], for the first time, they used tiny 3 × 3 filters in each convolution layer and also combined them as a sequence of convolutions. Contrary to the principles of LeNet, which uses
large convolutions to capture similar features in an image, as well as AlexNet, which uses 9 × 9 or 11 × 11 convolution filters, filters in the VGG network begin to shrink and approach the bad 1 × 1 convolution that LeNet wanted to avoid. These ideas are also used in newer architectures such as Inception and ResNet [13], [14]. Illustrated the VGG Diagram in Figure. 1

1) Specification and Parameters: For the first convolutional layer, the network must learn 64 filters of size 3 × 3 with input depth (3). In addition, each of the 64 filters has a bias, so the total number of parameters is 64 × 3 × 3 × 3 + 64 = 1792. It can be applied the same logic to other convolutional layers. The depth of an output layer will be the number of convolution filters. The padding is selected as 1 pixel, so the spatial resolution is maintained through the convolutional layers. Thus, the spatial resolution will only change at the pooling layers. Therefore, the first convolutional layer’s output will be 64 × 224 × 224. The pooling layer does not learn anything, so we have 0 learning parameters. To calculate the pooling layer’s output, we need consider the size of the window and the step. To calculate the number of parameters in fully connected layers, we must multiply the number of units in the previous layer with the current layer’s number of units. By following the previous paragraph’s logic, it can be seen that the number of units in the last convolutional layer will be 512 × 7 × 7. Therefore, the total number of parameters in the first fully connected layer is 7 × 7 × 512 × 4096 + 4096 = 1027645447 [13]–[15]. As already mentioned, this network consists of two parts: a convolutional section and a fully connected section. The first 5 blocks form the convolution section that is responsible for the feature extraction. The fully connected section consists of three dense layers that perform Classification.

2) Transfer Learning: In the dataset we provide, we encounter many intra-class patterns for positive samples, so we need to have a model that can learn these intra-class different patterns. The VGG network can understand complex models because of the high number of learnable parameters. Hence, if we have enough data and time, and proper hardware to train this network, we can achieve high accuracy. First, the default fully connected part of the VGG network is removed, and the required fully connected layers are added. These layers are arranged after the convolution layers, respectively. Also, to increase the model’s generalization and prevent overfitting, a dropout layer of 0.1 is used. In practice, we use this layer to increase the accuracy of the model on the test data. In the modified network, first, the convolution layers were frozen. Then we began to train the fully connected layers due to the possibility of a large difference in the initial values of fully connected layer weights from optimal values. If the convolutional layer is trained, it may be possible to reduce the system’s absolute accuracy by inappropriately changing convolution weights. Hence, it is just the training of the fully connected layer at this stage, and we do not change the weights of the convolutional layers [16]. The results are as follows: (Figure. 2 shows the confusion matrix of these results).

As can be seen in Figure. 3 almost the loss and accuracy are reached relative stability and no longer change after ten epochs. Continuing training in these situations may lead to network overfitting, so the training ends.

3) Fine-tuning: In the next step, we proceeded to Fine-tuning of the network so that only the weights of the last two blocks and the fully connected layer are trained [17]. Figure. 4 shows the results of the Fine-tuning step. The last two blocks of the VGG network include convolutional modules and pooling modules. By training these modules, we see an increase in the accuracy of the results.

As can be seen in Figure. 5 after about ten epochs, loss and accuracy changes are significantly reduced and converged. Therefore, at this stage, we stopped training.

4) Feature Extraction: Our goal is to use the convolutional VGG network as an automatic and accurate feature extraction method. One of the ideas used in this article is to use an Ensemble-based Gradient boosting classifier instead of a fully connected layer to increase the accuracy of the pulmonary lesions’ classification. Hence, we extracted the features from the CNN layers of the VGG model, and instead of the fully connected layer, we used classifiers based on Ensemble and
of machine learning algorithms. The use and combination of methods are designed to improve the stability and accuracy better. The leading cause of the error is related to noise, of poor learners to create a strong learner who performs algorithms of ensemble methods [19], which combine a set of the test data. The results of these models are presented below.

Then the accuracy of each of these models was calculated on extracted features of the VGG convolutional neural network.

B. Classification

At this stage, different classifiers were trained based on the extracted features of the VGG convolutional neural network. Then the accuracy of each of these models was calculated on the test data. The results of these models are presented below.

1) Ensemble Methods: The ensemble methods are a concept of machine learning, the main idea of using multiple models to create a single and better algorithm. In other words, the methods in which multiple classifiers combine to make a more robust model. The accuracy of the generated model is greater than the accuracy of each of the initial models. One method for combining the results of classifiers is the majority vote. Voting and averaging are two of the easiest methods in the ensemble methods. Each of these methods has a simple understanding and implementation, voting is for classification, and averaging is for regression. In both ways, the first step is to create multiple classification/regression models using some of the training datasets. Each base model can be created using different training dataset divisions and the same algorithm or using the same dataset with different algorithms, or any other method [18].

2) Boosting and Bagging: Bagging and boosting are both algorithms of ensemble methods [19], which combine a set of poor learners to create a strong learner who performs better. The leading cause of the error is related to noise, bias, and variance. Ensembles help reduce these factors. These methods are designed to improve the stability and accuracy of machine learning algorithms. The use and combination of several classifiers reduce the final models variance, especially for unstable classifiers, and may produce a more reliable model. In Bagging, each element has a similar probability in a new dataset. Nevertheless, in boosting, the elements are weighted to increase the impact, and therefore some will be more involved in the training process. In Bagging, the training phase is parallel (for example, each model is built independently), but in Boosting, the new learner is sequentially arranged. In boosting algorithms, each classifier is trained on the dataset based on the previous classifiers success. After each training step, weights are distributed. Data that is classified incorrectly increases its weight so that the classifier is forced to focus on that. To predict the class of new data, we only need to apply the learners to the newly observed data. Bagging results are obtained by averaging the responses of all learners (or the majority vote). However, in boosting, the second set of weights is allocated to learners to obtain a weighted average of all classifiers results. At the boosting training stage, the algorithm assigns weights to each model. A classifier with good results gets a higher weight than a weak classifier. So boosting also needs to keep track of learners errors. Boosting includes three simple steps:

- A basic $F_0$ model is defined for predicting the target variable $y$. This model is associated with a residual value $(y - F_0)$.
- A new $H_1$ model is fitted on the residual of the previous stage.
- Now, $F_0$ and $H_1$ are combined for $F_1$ (the boosted version of $F_0$). The average square error of $F_1$ will be less than $F_0$:

$$F_1(X) = -F_0(X) + H_1(X)$$ (1)

To improve the performance of $F_1$, we can create a new $F_2$ after the residual of $F_1$.

$$F_2(X) = -F_1(X) + H_2(X)$$ (2)

This can be done for ‘m’ iterations until the residual value reaches our lowest target value.

$$F_m(X) < -F_{m-1}(X) + H_m(X)$$ (3)

As a first step, the model must begin with a function $F_0(x)$. $F_0(x)$ must be a function that minimizes the loss function or MSE in this case:

$$F_0(x) = \arg \min_Y \sum_{i=1}^{n} L(Y_i - y)$$ (4)

$$\arg \min_Y \sum_{i=1}^{n} L(Y_i - y) = \arg \min_Y \sum_{i=1}^{n} (Y_i - y)^2$$ (5)

1Mean Square Error
3) XGBoost: XGBoost is similar to Gradient boosting algorithm, but it has a few tricks up its sleeve, making it stand out from the rest. It has proven itself in terms of performance and speed. XGBoost unlike other GBM methods that first specify the step, and then the step value, directly determine the step using the following statement for each x in the data:

$$\frac{\delta L(yf^{(m-1)}(x) + f_m(x))}{\delta f_m(x)} = 0$$  \hspace{1cm} (6)

By doing second-order Taylor expansion of the loss function around the current estimate \( f_{m-1}(x) \), we get:

$$L(yf^{m-1}(x) + f_m(x)) \approx L(yf^{m-1}(x)) + g_m(x)f_m(x) + \frac{1}{2}h_m(x)f_m(x)^2$$  \hspace{1cm} (7)

where \( g_m(x) \) is the Gradient, same as the one in GBM, and \( h_m(x) \) is the Hessian (second order derivative) at the current estimate.

$$h_m(x) = \frac{\delta^2L(Y,f(x))}{\delta f(x)^2}$$  \hspace{1cm} (8)

$$f(x) = f^{m-1}(x)$$  \hspace{1cm} (9)

Then the loss function can be rewritten as:

$$L(f_m) \approx \sum_{i=1}^{n}[g_m(x_i)f_m(x_i) + \frac{1}{2}h_m(x_i)f_m(x_i)^2] + \operatorname{const.}$$  \hspace{1cm} (10)

While Gradient Boosting follows negative Gradients to optimize the loss function, XGBoost uses Taylor expansion to calculate the value of the loss function for different base learners. XGBoost does not explore all possible tree structures but builds a tree greedily, and its regularization term penalizes building a complex tree with several leaf nodes \cite{20, 21}.

4) Classification Results: In the first try, we used the output of the last layer of the CNN as the feature vector, which contains 25088 features. In the second try, instead of using the output of the last layer of CNN, we used the output of the first Dense layer which achieved a higher final accuracy. The results are as follows: (Figures 6 to 11)

As can be seen, the XGBoost classifier’s accuracy is 96\%, and the accuracy of the fully-connected layer is 94\%, so we showed that using the XGBoost classifier instead of the dense layers of the VGG network leads to higher accuracy \cite{16, 22}.

C. Pulmonary Lesions Segmentation

Segmentation is the process of partitioning an image into different meaningful segments. These segments often correspond to different tissue classes, organs, pathologies, or other biologically relevant structures in medical imaging. Medical
image segmentation is made difficult by low contrast, noise, and other imaging ambiguities [8]. A major difficulty of medical image segmentation is the high variability in medical images. The result of the segmentation can then be used to obtain further diagnostic insights. Level set methods, based on partial differential equations (PDEs), are effective in the medical image segmentation tasks. However, to use this method, determining its control parameters is very important. Hence, other methods are used to determine these parameters. One of these techniques is FCM clustering. Using this method, with other methods are used to determine these parameters. One of the following methods is very similar to the k-means algorithm. The steps are as follows:

1) Fuzzy C-Means Clustering: The fuzzy C-means algorithm is very similar to the k-means algorithm. The steps are as follows:

- Choose a number of clusters.
- Assign coefficients randomly to each data point for being in the clusters.
- Repeat until the algorithm has converged label = 0
  - Compute centroid for each cluster.
  - For each data point, compute its coefficients of being in the clusters.

Each point \( x \) contains a set of coefficients that determine the membership degree in the k-th cluster, \( w_k(x) \). In Fuzzy C-means, the center of each cluster is averaged from all points. Weighing the membership degree is obtained by using the following formula:

\[
C_k = \frac{\sum_x W_k(x)^m x}{\sum_x W_k(x)^m} \tag{11}
\]

\( m \) is a hyperparameter that controls how clustering works. The goal of the Fuzzy C-means algorithm is to minimize the following target function [27]:

\[
C_k = \frac{\sum_x W_k(x)^m x}{\sum_x W_k(x)^m} \tag{12}
\]

\[
\arg\min_c \sum_{i=1}^n \sum_{j=1}^c W_{ij}^m ||x_i - c_j||^2 \tag{13}
\]

where:

\[
W_{ij} = \frac{1}{\sum_{k=1}^c ||x_i - c_k||^m} \tag{14}
\]

2) Level Set Methods: The idea of extending a surface \( \Phi \) instead of a front boundary \( C \) is used in this method, and the front boundary is defined so that all points with no elevation \( \Phi = 0 \). When the surface evolves and develops, the surface with a zero level set takes on various shapes. The surface \( \Phi \) points and our reference surface form our implicit boundary, and the zero level set shows contours splitting and merging. In this method, additional care is not required for topological changes. Therefore, this method is more suitable for our application.

3) The mathematical study of Level Set Methods: Assume that the point \( x = (x, y) \) belongs to the evolving front. So it changes over time, and \( x(t) \) is the position over time. At any time \( t \), for each point \( x(t) \) on the front, the surface has by definition no height, thus:

\[
\phi(x(t), t) = 0 \tag{15}
\]

To obtain the boundary, we require zero on Level Set, within the fact that it could carry any value. Assuming a primary \( \phi \) at \( t = 0 \), we may obtain \( \phi \) at any time \( t \) with the equation of motion \( \delta \phi / \delta t \). Based on the following chain rules, we have:

\[
\frac{\delta \phi(x(t), t)}{\delta t} = 0 \tag{16}
\]

\[
\frac{\delta \phi}{\delta x(t)} \frac{\delta x(t)}{\delta t} + \frac{\delta t}{\delta t} \frac{\delta \phi}{\delta t} = 0 \tag{17}
\]

\[
\frac{\delta \phi}{\delta x(t)} x_t + \phi_t = 0 \tag{18}
\]

We call here \( \delta \phi / \delta t = \nabla \phi \). Also, \( x(t) \) is obtained through the force \( F \), which is normalized to the surface, and therefore:

\[
x_t = F(x(t)) n \tag{19}
\]

Such that \( n = \nabla \phi / |\nabla \phi| \) and the previous moving equations are rewritten as follows:

\[
\phi_t + \nabla \phi x_t = 0 \tag{20}
\]

\[
\phi_t + \nabla \phi F_n = 0 \tag{21}
\]

\[
\phi_t + F \nabla \phi \frac{\nabla \phi}{|\nabla \phi|} = 0 \tag{22}
\]
\[ \phi_t + F|\nabla \phi| = 0 \] (23)

The last equation is the \( \phi \) moving equation. If \( \phi \) is given at time \( t = 0 \) and its motion equation is known over time, it is now possible to find \( \phi(x,y,t) \) any time \( t \) through the expansion of \( \phi \) over time. An interesting feature with \( \phi \) is that we can find the curvature of the curve by the following equation:

\[ k = \nabla \frac{\nabla \phi}{|\nabla \phi|} = \frac{\phi_{xx} \phi_y^2 - 2\phi_{xy} \phi_x \phi_y + \phi_{yy} \phi_x^2}{(\phi_x^2 + \phi_y)^2} \] (24)

This parameter can be used to control the smoothness and uniformity of the evolving front boundary [28]–[30].

4) Integrating spatial fuzzy clustering with Level Set: In this section, we used a method based on the work of Li et al. for pulmonary lesions image segmentation [1]. The level set method requires initial control parameters and sometimes also requires manual intervention to control these parameters. In this paper, a Fuzzy level set method is used to facilitate the segmentation of pulmonary lesions. Level set evolution can be started directly from the primary segmentation by spatial fuzzy clustering. Control of the parameters of the level set also evolves according to the fuzzy clustering results. Such methods help to better segmentation.

This method uses fuzzy clustering as the initial surface function. The FCM algorithm with spatial information can accurately estimate the boundaries. Therefore, the level set evolution will begin from an area close to the actual boundaries. Here, information from fuzzy clustering is used to estimate control parameters, which reduces manual interventions. The new level set fuzzy algorithm automates the initial settings and parameter setup of the level set using local fuzzy clustering. This is an FCM with spatial constraints to determine the approximate lines of interest in a medical image [23].

The model evolution equation presented by Li et al. is as follows:

\[ \epsilon(g, \phi) = \lambda \delta(\phi) \text{div}(g \frac{\nabla \phi}{|\nabla \phi|}) + gG(R_k)\delta(\phi) \] (25)

where, \( \lambda \) is the Coefficient of the contour length for smoothness regulation, \( \epsilon \) is Regulator for Dirac function \( \delta(\phi) \), \( \phi \) converts the 2D image segmentation problem into a 3D problem, \( \{\delta(\phi)\} \) denotes the Dirac function, \( \epsilon(g, \phi) \) (attracts \( \phi \) towards the variational boundary, which is similar to the standard level set methods.

\[ R_k = \{r_k = \mu_{nk}, n = x \times N_y + y\} \] (26)

\( R_k \) is the component of interest in FCM.

Finally, surface evolution can be regulated using local fuzzy clustering. In other words, the level set evolution is stable when it approaches the actual boundaries, which not only prevents boundary leakage but also prevents manual intervention. All these improvements lead to a strong algorithm for medical image segmentation.

D. Proposed system’s final outputs and results

After reading the CT-Scan image, its features are automatically extracted by the VGG convolutional neural network. Then, based on extracted features, it is classified using the XGBoost classifier. In the next step, if there is a lesion in the CT-scan image, the observed lesion is segmented by the mentioned hybrid segmentation method. A number of system outputs are shown below in Figures [12] to [14].
chosen for this work. The VGG network has many learnable parameters so it can learn many different patterns very well. With these extracted features various classifiers were trained and tested. The best results were achieved with the XGBoost classifier. And we showed that the use of ensemble and Gradient boosting based classifiers instead of a fully connected layer in the VGG network, increases the accuracy of the classification of pulmonary lesions. Then, in the case of a positive diagnosis of a pulmonary lesion, a hybrid fuzzy level set method (Li et al) was used for image segmentation. A dataset including CT-scan images of patients in Mashhad Local Area was collected and labeled by a specialist. We used this dataset for training and testing the proposed models. Significant features of the proposed system include the followings:

- Using a deep convolutional model to automatically extract the best features.
- The use of the VGG network based on fully-connected a large number of different intra-class patterns (Due to a large number of different types of pulmonary lesions.
- Applying XGBoost instead of a fully connected layer.
- Using a highly accurate and hybrid method for segmentation of pulmonary lesions.
- Designing an automated and complete system as a healthcare CAD system.
- Achieving very high accuracy so that the proposed model can be operationally used in the healthcare industry.
- Using a local dataset based on native patients in Mashhad.

Future recommendations for this work include: Determine the exact type of lesion, determine the risk of a diagnosed lesion and extend the dataset so that it covers all pulmonary lesions. Providing such dataset requires a great deal of time and money. To achieve this, there is a need for financial and scientific support in the form of an interdisciplinary research team. Diagnosis of pulmonary lesions due to the very high diversity is a hard and highly specialized task. Therefore, the preparation of such systems requires high technical knowledge. With the advancement of artificial intelligence methods, the accuracy of such systems can always be improved. The field of advancing and improving these systems will always be open to researchers.

ACKNOWLEDGMENT

We would like to thank Dr. Mahjoub and Behsazteb Medical Imaging Center for providing local CT-scan images of patients with human lung problems and providing diverse comments and reports on these images, which enabled us not only to categorize them accurately but also undoubted had a significant impact on this work’s overall improvement.

REFERENCES

[1] B. N. Li, C. K. Chui, S. Chang, and S. Ong, “Integrating spatial fuzzy clustering with level set methods for automated medical image segmentation,” Computers in Biology and Medicine, vol. 41, no. 1, pp. 1 – 10, 2011.

[2] A. C. Society, Key Statistics for Lung Cancer, January 2020. [https://www.cancer.org/cancer/lung-cancer/about/key-statistics.html]

[3] A. C. Society, Can Lung Cancer Be Found Early?, November 2020. [https://www.cancer.org/cancer/lung-cancer/detection-diagnosis-staging/detection.html]

[4] C. R. UK, Survival, September 2020. [https://www.cancerresearchuk.org/about-cancer/lung-cancer/survival/]

[5] M. Javaid, M. Javid, M. Z. U. Rehman, and S. I. A. Shah, “A novel approach to cad system for the detection of lung nodules in ct images,” Computer Methods and Programs in Biomedicine, vol. 135, pp. 125 – 139, 2016.

[6] Q. Abbas, “Segmentation of differential structural images on computed tomography images for diagnosis lung-related diseases,” Biomedical Signal Processing and Control, vol. 33, pp. 325 – 334, 2017.

[7] W.-J. Choi and T.-S. Choi, “Automated pulmonary nodule detection based on three-dimensional shape-based feature descriptor,” Computer Methods and Programs in Biomedicine, vol. 113, no. 1, pp. 37 – 54, 2014.

[8] P. Kamra, R. Vishrja, Kanica, and S. Gupta, “Performance comparison of image segmentation techniques for lung nodule detection in ct images,” in 2015 International Conference on Signal Processing, Computing and Control (ISPCC), pp. 302–306, 2015.

[9] Y. Xie, W.-Y. Meng, R.-Z. Li, Y.-W. Wang, X. Qian, C. Chan, Z.-F. Yu, X.-X. Fan, H.-D. Pan, C. Xie, Q.-B. Wu, P.-Y. Liu, L. Liu, Y.-J. Tang, X.-J. Yao, M.-F. Wang, and E.-L.-H. Leung, “Early lung cancer diagnostic biomarker discovery by machine learning methods,” Translational Oncology, vol. 14, no. 1, p. 100997, 2021.

[10] S. Magalhes Barros Netto, A. Corrila Silva, R. Acatauass Nunes, and M. Gattass, “Automatic segmentation of lung nodules with growing neural gas support vector machine,” Computers in Biology and Medicine, vol. 42, no. 11, pp. 1110 – 1121, 2012.

[11] S. L. A. Lee, A. Z. Kourzani, and E. J. Hu, “Automated identification of lung nodules,” in 2008 IEEE 10th Workshop on Multimedia Signal Processing, pp. 497–502, 2008.

[12] X. Ye, G. Beddoe, and G. Slabaugh, “Graph cut-based automated segmentation of lung nodules using shape, intensity, and spatial features.”

[13] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.

[14] Y. Zou, G. Zhang, and L. Liu, “Research on image steganography analysis based on deep learning,” Journal of Visual Communication and Image Representation, vol. 60, pp. 266 – 275, 2019.

[15] L. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.

[16] L. Li, R. Siti, J. Gao, Z. Yang, and W. Liu, “A hybrid model combining convolutional neural network with xgboost for predicting social media popularity,” in Proceedings of the 25th ACM international conference on Multimedia, pp. 1912–1917, 2017.

[17] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, “A survey on deep transfer learning,” in International conference on artificial neural networks, pp. 270–279, Springer, 2018.

[18] N. Demir, Ensemble Methods: Elegant Techniques to Produce Improved Machine Learning Results, 2018. [https://www.topical.com/machine-learning/ensemble-methods-machine-learning]

[19] S. Gonzlez, S. Garca, J. Del Ser, L. Rokach, and F. Herrera, “A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities,” Information Fusion, vol. 64, pp. 205 – 237, 2020.

[20] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785–794, 2016, L. I. Tian, X.-J. Yao, M.-F. Wang, and E. L.-H. Leung, “Early lung cancer diagnostic biomarker discovery by machine learning methods,” in 2014 International Conference BioMedical Visualization: Information Visualization in Medical and Biomedical Informatics, pp. 71–76, 2008.

[21] SauceCat, Boosting algorithm: XGBoost, 2017. [https://howtowritescience.com/boosting-algorithm-xgboost-4d9ec0207d]

[22] X. Ren, H. Guo, S. Li, S. Wang, and J. Li, “A novel image classification method with cnn-xgboost model,” in International Workshop on Digital Watermarking, pp. 378–390, Springer, 2017.

[23] Y. Zhang, B. I. Matuszewski, L. Shank, and C. J. Moore, “Medical image segmentation using new hybrid level-set method,” in 2008 Fifth International Conference BioMedical Visualisation: Information Visualization in Medical and Biomedical Informatics, pp. 71–76, 2008.

[24] M. Forouzanfar, N. Forghani, and M. Teshnehlab, “Parameter optimization of improved fuzzy c-means clustering algorithm for brain mr image segmentation,” Engineering Applications of Artificial Intelligence, vol. 23, no. 2, pp. 160 – 168, 2010.

[25] P. Swierczynski, B. W. Papie, J. A. Schnabel, and C. Macdonald, “A level-set approach to joint image segmentation and registration with application to ct lung imaging,” Computerized Medical Imaging and Graphics, vol. 65, pp. 58 – 68, 2018. Advances in Biomedical Image Processing.
[27] D.-Q. Zhang and S.-C. Chen, “A novel kernelized fuzzy c-means algorithm with application in medical image segmentation,” Artificial Intelligence in Medicine, vol. 32, no. 1, pp. 37 – 50, 2004. Artificial Intelligence in Medicine in China.

[28] H. Lombaert, Level set method: Explanation, 2006. [https://profs.etsmtl.ca/lombaert/levelset/]

[29] S. Osher and J. A. Sethian, “Fronts propagating with curvature-dependent speed: Algorithms based on hamilton-jacobi formulations,” Journal of Computational Physics, vol. 79, no. 1, pp. 12 – 49, 1988.

[30] M. Kass, A. Witkin, and D. Terzopoulos, “Snakes: Active contour models,” International journal of computer vision, vol. 1, no. 4, pp. 321–331, 1988.

APPENDIX A

GRAPHICAL ABSTRACT

Fig. 15. Graphical Abstract

APPENDIX B

HIGHLIGHTS:

- A pulmonary lesion classification method with slightly more than 96% accuracy.
- Combining VGG and XG-Boost for lesion classification.
- Models are trained on a dataset with more than thousands samples.
- Gathering a new dataset of lung lesions based on local patients.
- Pulmonary lesion segmentation using an integrated Fuzzy Clustering-Level Set method.
- An automated lung lesion classification and segmentation method.