Journal of Theoretical and Applied Information Technology
15th February 2024. Vol.102. No 3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

DEEP LEARNING AND STATISTICAL OPERATIONS BASED FEATURES EXTRACTION FOR LUNG CANCER DETECTION SYSTEM

SULEYMAN A. AlSHOWARAH
Faculty of Information Technology, Mutah University, Karak, Jordan
E-mail: showarah@mutah.edu.jo, suleyman.showarah@gmail.com

ABSTRACT

Lung cancer is considered as most common cause of cancer death than other. This kind of cancer is growing in human body without previous symptoms. So, using systems to diagnosis the patients in early stages is very essential and conducting studies in this field to find a good accuracy is also required. This research aims to examine the possibility of using of Deep Learning techniques for the lung cancer classification based on VGG-19 using images. Layer 6 and layer 7 of VGG19 were used. Also, new datasets will be created from these two layers named as statistical operations, which includes: average, minimum, maximum and combination between the two layers. Then, the datasets will be classified using different ML classifiers, which includes: KNN, Random Forest, Naïve Bayes and Decision Tree. Three scenarios will be used based on the training dataset size when classifying data. In the results, KNN scored the best accuracy (98.40), precision (0.98), recall (0.98) and F-measure (0.98). The results were nearly similar in all layers and scenarios; this means that the extracted features can provide high accuracy if applied in classification researches. It can be proved that the lung cancer can be detected with best accuracy even if the size of dataset in the training set was small. Also, the second-best accuracy after KNN algorithm is Random Forest in all layers and scenarios.

Keywords: *Lung Cancer Detection Using DL, Vgg-19, Lung Tumor, Benign Or Malignant Of Lung Cancer.*

1. INTRODUCTION

Lung cancer is one of tumors that growths hugely in lung tissue [1,2] and it cause the death for decades. Comparing with other cancers that may cause a death, this cancer scored 18.4% that has the highest mortality rate in global [3][4]. Also, it is considered as most common cause of cancer death in 2020 than other like: breast, colorectal, and prostate cancers combined. According to recent statistics in 2023, an estimated of 238,340 people will be infected with lung cancer, and 127,070 are expected to die from the disease [5].

The most factor behind the reason that make this cancer so deadly is growing the lung cancer in human body without a previous signs or symptoms. Nearly, 23% of people had no symptoms of cancer [5]. It is known that lung cancer can grow and spread quickly, so diagnosis the patients in early stages is essential. Today – Lung cancer can be detected at early stages based on imaging developments such as low-dose computed tomography (CT), which dramatically can improve the survival rate of patient [6].

However, there are two major reasons that prevent from the wide use the programs for lung cancer screening. First reason is the availability of each: human and technical tools, as the capacity for radiology will be insufficient to meet patients' needs [7,8]. Second reason is in the cases of false positive and over-diagnosis, which is highly linked to the required of high level of good training that recommended for service providers who perform image diagnosis [7,9]. It was found that the rate of benign infection when a discovery of nodules is at a high level: that reaches to 40% [10,11], this is to highlight the importance of careful analysis of nodules prior to further treatments to minimize surgical risks and avoid unnecessary difficulties or loss the capacity of lung. Given these limitations, AI was used extensively in recent years in computer-aided detection (CAD) systems for the automated detection the cancer nodules [12]. It is expected that the effectiveness of lung cancer screening will be increased with the use an accurate model for predicting lung cancer risk.

The CAD simulates the following three steps in analyzing the CT images. Step 1) Identifying a defect in the 3D image identified by the presence of regions of interest (ROI) (e.g., a nodular dimming). Step 2) Extracting all relevant features related to the ROI (e.g., texture, dimension, and relationship to neighboring regions. Then, to be used in the third step. Step 3) the features extracted in step 2 in the ROI will be classified based on the probability of malignancy [13]. The step 3 is necessary to determine the next step in dealing with patients. Moreover, segmentation of lung is also essential step, as CAD systems are usually used to perform feature extraction by identifying the value on a regular grid in three-dimensional of a particular ROI. This considered as further step for the radiologists who can't do it and rarely perform 3D segmentation in diagnosis for time constraints [9].

Several systems were developed for lung cancer detection. However, some systems are weak in detecting the tumor, but it is necessary for developing new systems to achieve the highest classification accuracy of 100%. Whereas the detection and classification of lung cancer can be performed based on machine learning algorithms and the techniques used in image processing [14].

Nowadays, Artificial Intelligence (AI) technologies are playing an important role in detecting and classifying tumor diseases. This is for the possibility of wide use of AI/ machine learning (ML) in prediction many health conditions [15,16]. The following are an examples for the most diseases that AI/ML systems concerns: diabetes [17] [18], hypertension [19], Corona Virus [20], hypercholesterolemia [21], stroke [22], chronic kidney disease [23], etc.

Due to the increasing number of patient who infected by cancers and amount of cancers' images, that would make clinical decision exceedingly time consuming and waste of radiologists' efforts that will require additional cost. The need for systems to detect and classify a lung cancer easily is essential to save time, and the efforts of radiologists [24]. This is considered as a major motivation for this study to diagnose people easily and accurately in lung cancer that is increasing daily. Therefore, the proposed model attempts to answer the following question: To what extent can deep learning (VGG-19 in this study) improve lung cancer detection system?

This study aims to examine the possibility of detecting and classifying the images of lung cancer based Deep Learning (DL). This study can be used as system that will detect the infected images and to help automatically the specialists' doctors/ radiologists in the diagnosis processes. To make the aim possible, it will be conducted based on Convolutional Neural Network (CNN)- (i.e., pre-trained technique - VGG-19 in this study). The advantage of using VGG-19 is to score high accuracy by finding distinctive details features that extracted from image [25] [26]. layer 6 and layer 7 of the VGG19 will be used; and each contains 4096 features. In addition, new datasets will be created from (layer-6 and layer-7); which represents the statistical operations. Statistical operations datasets consist of: Average, Minimum, Maximum, and combination between layer-6 and layer-7. Then the datasets will be classified using different ML classifiers, which includes: 1) KNN, 2) Random Forest, 3) Naïve Bayes and 4) Decision Tree.

In order to accomplish the objectives, this research is set to: 1) Developing a classification system to detect lung images if it is infected or not-infected. 2) Using one of CNN models such as VGG-19 on the selected database based images. 3) Examining the performance of different machine learning classifiers, which includes: (1) KNN, 2) Random Forest, 3) Naïve Bayes and 4) Decision Tree on the reliable extracted features. 4) To evaluate the feasibility of using the statistical operations and combination between different extracted features.

In order to evaluate the proposed model, there will be different performance metrics that includes: Accuracy, precision, Recall and finally F- Measure. Based on that, 10-Fold cross-validation will be used to verify the entered data reliability and validate the algorithm and the ability of the algorithm to classify the entered data.

The classifiers results showed that KNN algorithm performed best accuracy: 98.40, precision: 0.98, recall: 0.98 and F-Measure: 0.98; this is in layer-6 of third scenario. The results of datasets across all scenarios are considered high and nearly similar; this means that the extracted features can offer best accuracy if applied in classification research. This means that the lung cancer can be classified even if the datasets sizes in the training set was small. The second-high accuracy after KNN algorithm is Random Forest in all layers and scenarios.

The main contributions can be stated, as follows: 1) Providing the literature researches with studies for classifying lung cancer based on new model like VGG-19 using layer-6 and layer-7. 2) Conducting a study based on Statistical Operations that were not performed in the previous studies in this field, which

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

includes: Average, Maximum, Minimum, and combination between layer-6 and layer-7 for detecting lung cancer. 3) Providing the literature researches with a comparison finding using different machine learning algorithms used in this study. Also, using different training datasets sizes for three scenarios; the proposed model was applied in all of three scenarios.

This study consists of five sections. Section two presents the most related research. The methodology is discussed in the section three. Then, in section four discusses the experimental outcomes. At the end, section five introduces the conclusion.

2. RELATED WORK

The work on pattern recognition based images or image processing has increased in most recent years [27], this attracted the searchers for detecting and classifying the images of lung cancer. Different methods that were used for classifying the lung cancer based CT images are explained in the literature. Next are the studies that conducted on lung cancer using AI technique.

Several methods of image processing were innovated in [28] for detecting cancer tumors. In their study the classification network was used to detect if the nodule is infected or not like: Support Vector Machines (SVM), Forward Neural Networks, Convolution Neural Networks (CNNs), and Back Propagation Network (BPN).

Based on classification, there are number of researchers have conducted their studies based classifiers. For example, the researchers in [29] used genetic algorithm to extract features from images for Lung cancer. The lung nodules were classified using Frequency domain and SVM in [30]. A pulmonary nodules were detected automatically based on algorithm proposed in [31]. The classifier SVM is used to detect the true nodules and they were labeled. The researchers in [32] have conducted a study to classify lung nodules using the dataset based images: (LIDC-IDRI). Indexes of Diversity and Distinctness for Taxonomic from ecology are applied using SVM in the study [33] for the classification purposes. Results showed 98.11% of the accuracy. The pixels of CT using mesh-grid region were analyzed, then classified using ANN, this is to enhance the efficiency of computation. Whereas others unselected pixels were classified to negative infected.

In term of using Back Propagation Network (BPN) as supervised neural network for classification purposes. There are some of studies conducted for deleting unwanted artefacts in CT images, the researchers in [34] used K-Means for segmented images, and then they extracted features based on: entropy, contrast, correlation, homogeneity and area using statistic method that called: (GLCM). At the end, the researchers used BPN model for classification. Results showed that the classification accuracy reached to 90.7%.

The images were marked with cancer nodules using watershed segmentation in [35]. Then, features were extracted based on methods: 1) area, 2) perimeter, 3) eccentricity, 4) centroid, diameter, 5) pixel mean intensity, then features were classified based on SVM. Also, the study in [36] was conducted to detect the lung cancer based on (CNN) as classifier. The results showed the accuracy is 84.6%, sensitivity is 82.5%, and specificity is 86.7%. Whereas the segmentation approach on CT scan for lung cancer was also applied in [37] using thresholding. Based on a 3D CNN, the result showed that the accuracy of 86.6%.

Another study in [38] was conducted based on using four different methods for feature extraction (i.e., CNN, PCA, Restricted Boltzmann Machines, and 2D-DFT), then they discussed the results based methods. The dataset used in their study is (LIDC-IDRI). The Lung nodule were extracted in mass using a descriptive file from the CT scan, then the data augmentation was used to increase the dataset size. The results showed that CNN is the best compared to other methods used in the study.

Recently, in [39], the study conducted using CNN on dataset: (LIDC-IRDI) for CT images based on threshold segmentation. Their study aimed at lung parenchyma tissue segmentation process. They used the process of the replacing the vein system in the lung to detect the nodules. In their study, in order to reduce the number of false positive, a Vessel filters were used to remove the vessels [40]. The segmentation is used to detect infected lungs. The classification accuracy for detecting nodules were determined. While in the study of [41] that was conducted in 2020, a CNN was used to detect lung cancer. The dataset size contains of 100 images: (50 infected and 50 not infected) collected from 69 patients. Due to minimum size of dataset, data-augmentation was applied to enhance the size of dataset. Also, different techniques of CNNs were used (i.e. AlexNet, LeNet, and VGG-16). In order to enhance the weights in training datasets, an optimization method (i.e. Stochastic Gradient Descent) was used for AlexNet and VGG-16. While the optimization methods (i.e. RMSProp and

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ADAM) were used for LeNet. The features were extracted based on mRMR algorithm, then they were classified using the following machine learning: LR, LDA, SVM, KNN, and Decision Tree. The results showed that the used methods have improved the performance.

There are number of studies conducted on DL using SVM, such as the study in [42], which is conducted to detect a pulmonary nodule on CT images based on CNN and transfer learning. The features were extracted using VGG-16, then classified using SVM. The results showed a utilization in using the aforementioned methods. The study in [43] conducted to detect the lung cancer wither if there is tumor or no using ANN. The results showed an accuracy of 96.67%, and there was highest impact of age on the results.

At the end, this study is different of the previous studies in two aspects. First, this study conducted on different proposed model that consists of two layers and statistical operations, which was not used before for the same dataset. Second, three different scenarios were considered that were not considered in the previous studies for the lung cancer.

3. METHODOLOGIES

This section consists of: two subsections. 1) Used dataset. 2) The experimental Design.

3.1 Used Dataset

The used dataset is for lung cancer named as: (IQ-OTH/NCCD). The images were collected in the Iraq-Oncology Teaching Hospital - National Center for Cancer Diseases in DICOM format. The size of dataset is 1190 CT images of 110 cases that grouped into three cases, then, they were classified based on training and test datasets. First, 40 cases are for malignant group. Second, 15 cases are for benign group. Third, 55 cases are for normal group. The images were approved by oncologists and radiologists from these two specialists' centers. The protocol for CT includes: 120 kV, slice thickness of 1 mm, with window width ranging from 350 to 1200 HU and window center from 50 to 600. The dataset is available in [44].

3.2 The Experimental Design

The used experimental design framework is presented in Figure 1 that consists of three steps as applied in [49]. Step (1) The features were extracted automatically form images using Pre-trained VGG-19 on MATLAB. The output of (step 1) are two datasets for layer 6 and layer 7. Each of them consists of (4096) features. Then, these datasets will be an input for step 3.

In step (2). New datasets (i.e. four datasets) will be from layer-6 and layer-7. These datasets will be also an input for step 3, in the next stage. The following are the explanations for the four statistical operations. 1) Average: It is the average of values for two matrices (i.e. datasets) of layer-6 and layer-7, then the output is saved in new matrix (i.e. dataset). 2) Maximum: It is the maximum value of two matrices (i.e. datasets) of layer-6 and layer-7, then the output is saved in new matrix (i.e. datasets). 3) Minimum: It is the minimum value of two matrices (i.e. dataset) of layer-6 and layer-7, then the output is saved in new matrix (i.e. datasets). 4) Combination: It is the combination between the values of two matrices (i.e. datasets) of layer-6 and layer-7, then the output is saved in new matrix. The dataset of layer 6: (4096) is combined to the dataset of layer 7: (4096), and thus a new combined dataset will contain 8192 features

Step (3) The datasets are created in steps (1 and 2) will be classified in this step using the classifiers that presented above. This is to provide the experimental results.

In addition, the experiment is designed on three scenarios. 1) Using 50% in training and 50% in test dataset. 2) Using 70% in training and 30% in test dataset. 3) Using 80% in training and 20% in test dataset. Whereas, in each scenario – the three steps that explained earlier are applied. This means that the proposed model is used three times for three scenarios. Then, the output evaluated for each scenario.

4. EXPERIMENTAL RESULTS AND DISCUSSIONS

Three scenarios are considered in this study. The results will be evaluated using the four metrics, which are 1) Accuracy, 2) Recall, 3) F-measure, 4) Precision, 5) Training Time of each classifier per seconds. The four classifiers used in this study are: 1) KNN, 2) Random Forest, 3) Naïve Bayes, 4) Decision Tree.

The following sub-sections presents the results of each scenario.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

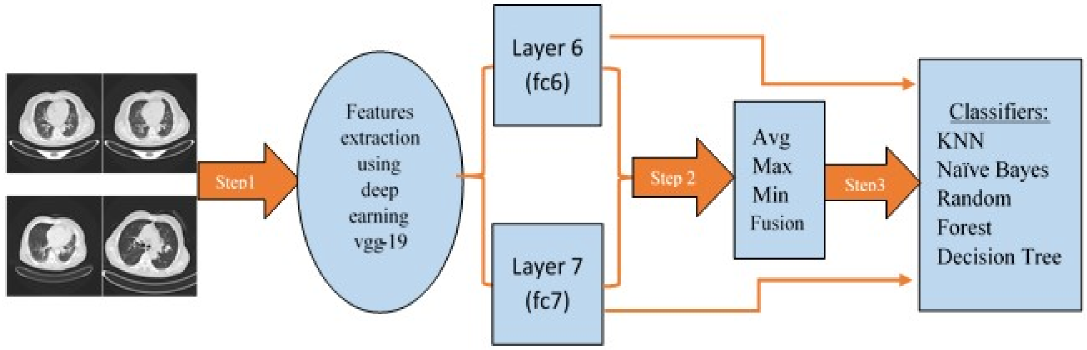

Figure 1. The proposed Framework Model

4.1 First Scenario

The first scenario is designed based on 50% of data in the training and 50% of data in test dataset. The aim of designing this scenario is to evaluate how the dataset size will influence on the classification accuracy. Three results will be discussed in this scenario. 1) The results of layer-6 and layer-7. 2) The results of the statistical operations. 3) The results of combination between features of layer-6 and layer-7.

4.1.1 Results of layer-6 and layer-7 separately

Table (1) and Table (2) illustrate the results of the layers: layer-6 and layer-7. According to the results., the classifier KNN is the best accuracy in detecting the lung cancer if the image is infected or not-infected for layer-6 and layer-7; the classification accuracy for layer-6 and layer-7 are (95.26) and (95.44) respectively.

However, the duration time for training showed that Decision Tree took longer time (i.e. 3.12s) when make a comparison with other algorithms, but KNN took less duration time (i.e. 0s).

This can be explained as there is no training model in KNN; the test row matches directly with other training rows, and this explains the less required time for testing, especially if there is large size of data for the training [45] [46].

These results match with the results in [41]. Their study conducted on vgg-16 using Principal Component Analysis (PCA) method and LR, LDA, SVM, KNN, and Decision Tree algorithms. In their study, the size of dataset is 100 images collected from 69 different patients. The performance of their results was improved using the PCA with 99.51 of accuracy using KNN with CNN & mRMR. In their study, the researchers did not use the statistical operations and did not apply the three scenarios. Also, another study conducted on 83 CT images from different 70 patients. In the preprocessing, the geometric mean filter was used, then the K-means technique was used to segment the cancers' images. Then, different classification methods were used (i.e., ANN, KNN, and Random Forest). The results showed that the ANN is best classification accurate in detecting lung cancer [47].

4.1.2 Results of the Statistical Operations.

Tables from (3 to 5) illustrate the results of statistical operations for the following datasets: 1) Average, 2) Maximum, 3) Minimum. According to the results, KNN showed the best in classifying lung cancer when make a comparison with other classifiers for all of: Average, Maximum, and Minimum. The classification accuracy for them are (95.81), (95.99), and (95.44) respectively. In addition, KNN showed less duration time for training- (i.e., 0s). The explanation here is the same for what discussed earlier in (Section: 4.1.1). These results match with the results in [41] as discussed also in (Section: 4.1.1) in term of the best classification accuracy that scored by KNN, but not in using the statistical operations or in using the three scenarios.

4.1.3 Results of combination feature between (layer-6 and layer-7)

The combination dataset contains (8192 features), which is created of combining the values of both layers-6 and layer-7; each contains (4096 feature). The results are displayed in Table (6) and illustrates the accuracy of KNN (i.e. 96.17) that have the best classification accuracy when make a comparison with other algorithms in detecting the benign or malignant of lung cancer. The second-best accuracy is for Random Forest (89.25). The duration time for training is (0s), which is the least when make a comparison with other classifiers.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The results of all datasets in this first scenario are considered high and nearly similar when make a comparison between their results. This explains the influence of the extracted features used in the study on the classification accuracy. This can answer the question: to what extent can deep learning (VGG-19 in this study) improve lung cancer detection system?

4.2 Second Scenario

The second scenario is designed based on 70% of data in training and 30% of data in test dataset. The aim of designing this scenario is as explained in Section (4.1) above. Three results will be discussed in this scenario. 1) The results of layer-6 and layer-7. 2) The results of the statistical operations. 3) The results of combination between features of layer-6 and layer-7.

4.2.1 Results of layer-6 and layer-7 separately

Table (1) and Table (2) illustrate the results of the layers: 6 and 7. According to the results, the classifier KNN is the best accuracy in detecting the lung cancer if the image is infected or not-infected for layer-6 and layer-7; the classification accuracy for layer-6 and layer-7 are (96.87) and (96.74) respectively. The second-best classification accuracy is Random Forest. However, the duration time for training showed that Decision Tree took longer time (i.e. 4.15s) when make a comparison with other algorithms, but KNN took less training time (i.e. 0s). The reason for this was discussed in Section (4.1.1). These results match with the results in [41] and [47] in term of using KNN and Random Forest that outperformed other classifiers. In their study, they used different method like VGG-16. They did not use the statistical operations and did not apply the three scenarios.

4.2.2 Results of the Statistical Operations

Tables from (3 to 5) illustrate the results of statistical operations for the following datasets: 1) Average, 2) Maximum, 3) Minimum). According to the results, KNN showed the best algorithm when make a comparison with other algorithms used in this study in detecting the lung cancer if the image is infected or not-infected for all of Average, Maximum, and Minimum. The classification accuracy for them are (97.13), (97), and (97.26) respectively. In addition, KNN showed less duration time for training- (i.e., 0s). These results match with the results in [41] in term of the best classification accuracy that scored by KNN, but not in using the statistical operation or using the three scenarios. The results in this section approve the outcome of layer-6 and layer-7 in Section (4.1.2).

4.2.3 Results of combination feature between (layer-66 and layer-7)

The combination dataset contains (8192 features), which is created of combining the values of both layer-6 and layer-7; each contains (4096 feature).

The results are displayed in Table (6) and illustrates the accuracy of KNN (i.e. 97) that have the best classification accuracy when make a comparison with other algorithms in detecting the benign or malignant of lung cancer. The second-best accuracy is for Random Forest (89.97). The duration time for training is (0s), which is the least when make a comparison with other classifiers. The results in this section approve the outcomes in Sections 4.1.3.

The results of all datasets in second scenario are considered high and nearly similar when make a comparison between their results. This explains the influence of the extracted features used in the study on the classification accuracy. This approves that VGG-19 can improve lung cancer detection system.

4.3 Third Scenario

The third scenario is designed based on 80% of data in training and 20% of data in test dataset. The aim of designing this scenario is as explained in Sections (4.1 and 4.2). Three results will be discussed in this scenario. 1) The results of layer-6 and layer-7. 2) The results of the statistical operations. 3) The results of combination between features of layer-6 and layer-7.

4.3.1 Results of layer-6 and layer-7 separately

Table (1) and Table (2) illustrate the results of the layers: 6 and 7. According to the results, the classifier KNN is the best accuracy in detecting the lung cancer if the image is infected or not-infected for layer-6 and layer-7; the classification accuracy for layer-6 and layer-7 are (98.40) and (97.60) respectively. The second-best classification accuracy is Random Forest. However, the duration time for training showed that Decision Tree took longer time (i.e. 6.97s) when make a comparison with other algorithms, but KNN took less duration time (i.e. 0s). The reason for this was discussed in Section (4.1.1). This results match with the results in [41] and [47] in term of using KNN and Random Forest that outperformed other classifiers. Also, the results match with results in Sections (4.1.1 and 4.2.1).

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 1. Detection Results Of Layer-6 Feature Vector For Three Scenarios.

Algorithms	Accuracy	Precision	Recall	F-measure	Time (s)	Accuracy	Precision	Recall	F-measure	Time (s)	Accuracy	Precision	Recall	F-measure	Time (s)
	Layer-6 (1s Scenario)					***Layer-6 (2nd Scenario)***					***Layer-6 (3rd Scenario)***				
KNN	95.26	0.95	0.95	0.95	0.01	96.87	0.96	0.96	0.96	0	98.40	0.98	0.98	0.98	0
Naïve Bayes	76.68	0.80	0.76	0.78	0.4	76.95	0.80	0.77	0.78	0.41	79.61	0.82	0.79	0.80	0.49
Random Forest	87.61	0.88	0.87	0.85	1.42	89.58	0.90	0.89	0.87	1.44	90.54	0.90	0.90	0.88	1.63
Decision Tree	81.05	0.81	0.81	0.81	3.12	85.28	0.85	0.85	0.85	3.78	84.73	0.84	0.84	0.84	5.19

Table 2. Detection Results Of Layer-7 Feature Vector For Three Scenarios.

Algorithms	Accuracy	Precision	Recall	F-measure	Time (s)	Accuracy	Precision	Recall	F-measure	Time (s)	Accuracy	Precision	Recall	F-measure	Time (s)
	Layer-7 (1st Scenario)					***Layer-7 (2nd Scenario)***					***Layer-7 (3rd Scenario)***				
KNN	95.44	0.95	0.95	0.95	0	96.74	0.96	0.96	0.96	0	97.60	0.97	0.97	0.97	0
Naïve Bayes	77.77	0.80	0.77	0.78	0.35	79.42	0.82	0.79	0.80	0.48	81.43	0.82	0.81	0.82	0.45
Random Forest	87.79	0.89	0.87	0.85	1.14	89.58	0.90	0.89	0.87	1.37	90.54	0.91	0.90	0.88	1.52
Decision Tree	81.78	0.81	0.81	0.81	2.43	83.07	0.83	0.83	0.83	4.15	85.53	0.85	0.85	0.85	6.97

Table 3. Detection Results Of Average Feature Vector For Three Scenarios.

Algorithms	Accuracy	Precision	Recall	F-measure	Time (s)	Accuracy	Precision	Recall	F-measure	Time (s)	Accuracy	Precision	Recall	F-measure	Time (s)
	Average (1st Scenario)					***Average (2nd Scenario)***					***Average (3rd Scenario)***				
KNN	95.81	0.95	0.95	0.95	0	97.13	0.97	0.97	0.97	0	98.29	0.98	0.98	0.98	0
Naïve Bayes	76.13	0.80	0.76	0.77	0.3	76.95	0.80	0.77	0.78	0.41	79.72	0.82	0.79	0.80	0.5
Random Forest	87.97	0.89	0.88	0.85	1.08	89.19	0.90	0.89	0.86	1.45	90.66	0.91	0.90	0.88	1.79
Decision Tree	82.33	0.82	0.82	0.82	2.44	80.98	0.81	0.81	0.81	3.59	85.42	0.85	0.85	0.85	3.88

Table 4. Detection Results Of Maximium Feature Vector For Three Scenarios.

Algorithms	Accuracy	Precision	Recall	F-measure	Time (s)	Accuracy	Precision	Recall	F-measure	Time (s)	Accuracy	Precision	Recall	F-measure	Time (s)
	Maximum (1st Scenario)					***Maximum (2nd Scenario)***					***Maximum (3rd Scenario)***				
KNN	95.99	0.95	0.96	0.95	0	97.00	0.97	0.97	0.97	0	98.06	0.98	0.98	0.98	0.01
Naïve Bayes	80.51	0.81	0.80	0.80	0.28	80.98	0.82	0.81	0.81	0.42	82.23	0.83	0.82	0.82	0.52
Random Forest	87.79	0.89	0.87	0.85	1.08	89.84	0.90	0.89	0.87	1.42	90.20	0.90	0.90	0.88	1.56
Decision Tree	78.50	0.79	0.78	0.78	2.66	82.29	0.82	0.82	0.82	4.17	85.07	0.85	0.85	0.85	5.11

4.3.2 Results of the Statistical Operations

Tables from (3 to 5) illustrate the results of statistical operations for the following datasets: 1) Average, 2) Maximum, 3) Minimum). According to the results, KNN showed the best algorithm when make a comparison with other algorithms used in this study in detecting the lung cancer if the image is infected or not-infected for all of Average, Maximum, and Minimum. The classification accuracy for them are (98.29), (98.06), and (98.17) respectively. In addition, KNN showed less duration time for training- (i.e., 0s). These results match with the results in [41] in term of the best classification accuracy that scored by KNN, but not in using the statistical operation or using the three scenarios. The results in this section approve the outcome of layer-6 and layer-7 in Sections (4.1.2 – 4.2.2).

Table 5. Detection Results Of Minimium Feature Vector For Three Scenarios.

Algorithms	Accuracy	Precision	Recall	F-measure	Time (s)	Accuracy	Precision	Recall	F-measure	Time (s)	Accuracy	Precision	Recall	F-measure	Time (s)
	Minimum (1st Scenario)					Minimum (2nd Scenario)					Minimum (3rd Scenario)				
KNN	95.44	0.95	0.95	0.95	0	97.26	0.97	0.97	0.97	0	98.17	0.98	0.98	0.98	0
Naïve Bayes	74.31	0.79	0.74	0.76	0.3	76.17	0.80	0.76	0.77	0.41	79.27	0.82	0.79	0.80	0.51
Random Forest	88.34	0.89	0.88	0.85	1.25	89.45	0.90	0.89	0.87	1.4	89.74	0.90	0.89	0.87	1.79
Decision Tree	80.69	0.81	0.80	0.81	2.7	84.63	0.84	0.84	0.84	3.62	98.17	0.85	0.85	0.85	5.2

Table 6. Detection Results Of Combination Feature Vector For Three Scenarios.

Algorithms	Accuracy	Precision	Recall	F-measure	Time (s)	Accuracy	Precision	Recall	F-measure	Time (s)	Accuracy	Precision	Recall	F-measure	Time (s)
	Combination (1st Scenario)					Combination (2nd Scenario)					Combination (3rd Scenario)				
KNN	96.17	0.96	0.96	0.96	0	97.00	0.97	0.97	0.96	0	84.73	0.98	0.98	0.98	0
Naïve Bayes	77.59	0.80	0.77	0.78	0.57	77.47	0.80	0.77	0.78	0.88	81.09	0.82	0.81	0.81	1.09
Random Forest	89.25	0.90	0.89	0.87	2.52	89.97	0.91	0.90	0.87	2.13	90.54	0.91	0.90	0.88	2.69
Decision Tree	80.87	0.82	0.80	0.81	4.96	83.33	0.83	0.83	0.83	7.48	84.73	0.84	0.84	0.84	9.86

Table 7. Comparison With Related Works.

Studies	Classifiers	Methodology	Database size	Results= accuracy
[41]	ANN, KNN, and Random Forest	They used the filter of geometric mean; then the K-means technique to segment the cancers' images.	83 CT images	ANN= 98% KNN=90% Random Forest= 80%
[47]	linear regression, linear discriminant analysis, decision tree, SVM, KNN, softmax	LeNet, AlexNet and VGG-16.	100 images	combination of AlexNet+KNN =98.74 %.
[The proposed model]	KNN Naïve Bayes Random Forest Decision Tree	Vgg-19 Statistical Operations: Average, Maximum, Minimum, Combination layers (6and 7)	1190 CT images	KNN-layer-6-3rd scenario=98.40%

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4.3.3 Results of combination feature between (layer-6 and layer-7)

The combination dataset contains (8192 features), which is created of combining of both layer-6 and layer-7, each contains (4096 feature). The results are displayed in Table (6) and illustrates the accuracy of Random Forest (i.e. 90.54) that have the best classification accuracy when make a comparison with other algorithms in detecting lung cancer images if it is infected or not-infected. Whilst the second-best accuracy is KNN (i.e. 84.73). We can see that the best classification accuracy for first and second classifier in table (6) is different compared to other tables in the previous sections. This cannot approve that it is a common case. While the duration time for training of KNN is (0s), which is the least when make a comparison with other classifiers.

In general, the results of all datasets in third scenario are considered high and nearly similar when make a comparison between their results. This explains the influence of the extracted features used in the study on the classification accuracy. This approves that VGG-19 can improve lung cancer detection system.

Despite of that, we can see the results of scenario 3 are a bit higher, but they seem nearly near to each other. This means that the classification accuracy can be influenced positively when the training dataset size is large.

Table (7) shows the comparison between the results of the proposed model and most similar works conducted on detection of lung cancer. The following points illustrates the reasons for considering the proposed model is interesting study:

- Number of previous studies were achieved on small size of dataset compared to the proposed model.
- The training dataset sizes in some of previous studies were performed in one scenario and they have large size of dataset in the training set compared to the proposed model in this study. Of-course large size of data in the training set will increase the accuracy. Few number of images usually leads to low classification accuracy compared to use large number of images. Therefore, the proposed model is applied on three scenarios to guarantee that the study can work on small sizes of training set.
- Based on search in the literature, there was no study seen that has been conducted based on detecting lung cancer using all of the following together deep learning and the following four classifiers that includes: KNN, Naïve Bayes, Random Forest, and Decision Tree, and also using the created datasets for Average, Maximum, Minimum, and Combination of layer-6 and layer-7 in three different scenarios (i.e. (50%-50%), (80%-20%), (70%-30%) for (training-testing) respectively).

5. CONCLUSION

This research is conducted to provide an insight on new methods for lung cancer detection by examining the possibility of using deep-learning techniques for the lung cancer classification based on VGG-19 using ultrasound images. Layer 6 and layer 7 of VGG19 were used; each consists of 4096 features. Also, new datasets have been created from the two layers named as statistical operations, which includes: average, minimum, maximum and combination between the two layers. Then, the datasets will be classified using different ML classifiers, which includes: KNN, Random Forest, Naïve Bayes and Decision Tree. Three scenarios were considered based on the training dataset size when classifying data.

In the results, KNN scored the best accuracy (98.40), precision (0.98), recall (0.98) and F-measure (0.98). The results were nearly similar in all layers and scenarios; this means that the extracted features can provide high accuracy if applied in classification researches. It can be proved that the lung cancer can be detected with best accuracy even if the dataset size in the training set was small. Also, the second-best accuracy after KNN algorithm is Random Forest in all layers and across all scenarios.

As a general investigation from Tables (from 1 to 6); and based on the proposed model used, it can considerable to make the following claims:

KNN is the best classifier to be used for deep features in all datasets for each scenario - if it is provided with a smaller number of deep features. While the Naïve Bayes has scored the less accuracy. Also, KNN required less duration time when make a comparison with other algorithms. This can be explained as there is no training model in KNN; the test row matches directly with other training rows, and this explains the less required time for testing. Whereas Naïve Bayes does require training.

- Random Forest is considered as the second best of classifier after KNN, this for all operations i.e. (layer-6, layer-7, Average, Maximum, Minimum, and Combination) .

- In general, there was no distinctive features based results between the layer-6 and layer-7 (see tables (1 and 2), or any other dataset was created based on them. Deep learning (VGG-19) provides features that can improve detection systems such as lung cancer detection system.
- Arranging the machine learning algorithms in the results from best to least, they were as follows: KNN was first, followed by Random Forest, then Decision Tree, and finally Naïve Bayes.

More examination is recommended on the combination between the extracted features to improve the classification accuracy.

REFRENCES:

[1]. National Cancer Institute. Non-Small Cell Lung Cancer Treatment (PDQ)—Patient Version. 2016. Available online: https://web.archive.org/web/20160229172247/http://www.cancer.gov/types/lung/patient/non-small-cell-lung-treatment-pdq (accessed on 30 May 2021).
[2]. Pasławski, M.; Krzyzanowski, K.; Złomaniec, J.; Gwizdak, J. Morphological characteristics of malignant solitary pulmonary nodules. Ann. Univ. Mariae Curie-Skłodowska Sect. D Med. 2004, 59, 6–13.
[3]. de Koning, H.J.; van der Aalst, C.M.; de Jong, P.A.; Scholten, E.T.; Nackaerts, K.; Heuvelmans, M.A.; Lammers, J.W.J.; Weenink, C.; Yousaf-Khan, U.; Horeweg, N.; et al. Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial. N. Engl. J. Med. 2020, 382, 503–513. [PubMed]
[4]. Cancer Facts & Figures, (2023). "Special Section: Lung Cancer". American Cancer Society, Inc., Surveillance and Health Equity Science.
[5]. https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2023/2023-cff-special-section-lung-cancer.pdf
[6]. Cancer Facts & Figures, (2023). "Special Section: Lung Cancer". American Cancer Society, Inc., Surveillance and Health Equity Science.
[7]. ünaydin, Ö., Günay, M., & Şengel, Ö. (2019). Comparison of lung cancer detection algorithms. In 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT) (pp. 1-4). IEEE.
[8]. Wait, S.; Alvarez-Rosete, A.; Osama, T.; Bancroft, D.; Cornelissen, R.; Maruši´c, A.; Garrido, P.; Adamek, M.; van Meerbeeck, J.; Snoeckx, A.; et al. Implementing Lung Cancer Screening in Europe: Taking a Systems Approach. JTO Clin. Res. Rep. 2022, 3, 100329.
[9]. British Society of Thoracic Imaging and The Royal College of Radiologists. Considerations to Ensure Optimum Roll-Out of Targeted Lung Cancer Screening over the Next Five Years; British Society of Thoracic Imaging and The Royal College of Radiologists: London, UK, 2020.
[10]. Kauczor, H.-U.; Bonomo, L.; Gaga, M.; Nackaerts, K.; Peled, N.; Prokop, M.; Remy-Jardin, M.; von Stackelberg, O.; Sculier, J.-P. ESR/ERS White Paper on Lung Cancer Screening. Eur. Radiol. 2015, 25, 2519–2531.
[11]. Kuo, E.; Bharat, A.; Bontumasi, N.; Sanchez, C.; Zoole, J.B.; Patterson, G.A.; Meyers, B.F. Impact of Video-Assisted Thoracoscopic Surgery on Benign Resections for Solitary Pulmonary Nodules. Ann. Thorac. Surg. 2012, 93, 266–273.
[12]. Deppen, S.A.; Blume, J.D.; Aldrich, M.C.; Fletcher, S.A.; Massion, P.P.; Walker, R.C.; Chen, H.C.; Speroff, T.; Degesys, C.A.; Pinkerman, R.; et al. Predicting Lung Cancer Prior to Surgical Resection in Patients with Lung Nodules. J. Thorac. Oncol. 2014, 9, 1477–1484.
[13]. El-Baz, A.; Beache, G.M.; Gimel'farb, G.; Suzuki, K.; Okada, K.; Elnakib, A.; Soliman, A.; Abdollahi, B. Computer-Aided Diagnosis Systems for Lung Cancer: Challenges and Methodologies. Int. J. Biomed. Imaging 2013, 2013, 942353. [CrossRef]
[14]. American College of Radiology. Lung-RADS v1.1 Assessment Categories (2019 Release); American College of Radiology: Reston, VA, USA, 2019.
[15]. Zebari, D. A., Zeebaree, D. Q., Abdulazeez, A. M., Haron, H., & Hamed, H. N. A. (2020). Improved Threshold Based and Trainable Fully Automated Segmentation for Breast Cancer Boundary and Pectoral Muscle in Mammogram Images. IEEE Access, 8, 203097-203116.
[16]. Artificial Intelligence/Machine Learning (AI/ML)-Based: Software as a Medical

Device (SaMD) Action Plan. Available online: https://www.fda.gov/media/145022/download .
[17]. Mahler, M.; Auza, C.; Albesa, R.; Melus, C.; Wu, J.A. Regulatory aspects of artificial intelligence and machine learning-enabled software as medical devices (SaMD). In Precision Medicine and Artificial Intelligence; Elsevier: Amsterdam, The Netherlands, 2021; pp. 237–265.
[18]. Dritsas, E.; Trigka, M. Data-Driven Machine-Learning Methods for Diabetes Risk Prediction. Sensors 2022, 22, 5304. [PubMed]
[19]. Dritsas, E.; Alexiou, S.; Konstantoulas, I.; Moustakas, K. Short-term Glucose Prediction based on Oral Glucose Tolerance Test Values. In Proceedings of the International Joint Conference on Biomedical Engineering Systems and Technologies—HEALTHINF, Vienna, Austria, 9–11 February 2022; Volume 5, pp. 249–255.
[20]. Dritsas, E.; Fazakis, N.; Kocsis, O.; Fakotakis, N.; Moustakas, K. Long-Term Hypertension Risk Prediction with ML Techniques in ELSA Database. In Proceedings of the International Conference on Learning and Intelligent Optimization, Athens, Greece, 20–25 June 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 113–120.
[21]. De Felice, F.; Polimeni, A. Coronavirus disease (COVID-19): A machine learning bibliometric analysis. In Vivo 2020, 34, 1613–1617. [PubMed]
[22]. Dritsas, E.; Trigka, M. Machine Learning Methods for Hypercholesterolemia Long-Term Risk Prediction. Sensors 2022, 22, 5365.
[23]. Dritsas, E.; Trigka, M. Stroke Risk Prediction with Machine Learning Techniques. Sensors 2022, 22, 4670. [PubMed
[24]. Dritsas, E.; Trigka, M. Machine Learning Techniques for Chronic Kidney Disease Risk Prediction. Big Data Cogn. Comput. 2022, 6, 98.
[25]. S. R. A. Ahmed, I. Al-Barazanchi, A. Mhana, and H. R. Abdulshaheed, "Lung cancer classification using data mining and supervised learning algorithms on multi-dimensional data set," Period. Eng. Nat. Sci., vol. 7, no. 2, pp. 438–447, 2019, doi: 10.21533/pen.v7i2.483.
[26]. M. P. Sampat, G. J. Whitman, M. K. Markey, and A. C. Bovik, (2005). ‘‘Evidence based detection of spiculated masses and architectural distortions'. Proc. Med. Imag. Image Process., San Diego, CA, USA, Apr. 2005, pp. 26–37.
[27]. Computer Creations, (2023). "Why is the VGG Network Commonly Used?". NNART, LLC 2023. Accessed: https://nnart.org/why-is-vgg-commonly-used/
[28]. Zhou Z-H, Jiang Y, Yang Y-B, Chen S-F (2002) Lung cancer cell identification based on artificial neural network ensembles', Elsevier. Artif Intell Med 24:25–36
[29]. R. M. Devarapalli, H. K. Kalluri, and V. Dondeti, "Lung cancer detection of ct lung images," Int. J. Recent Technol. Eng., vol. 7, no. 5, pp. 413–416, 2019.
[30]. Park SC, Tan J, Wang X, Lederman D, Leader JK, Kim SH, Zheng B (2011) Computer-aided detection of early interstitial lung diseases using low-dose CT images', Iop Publishing. Phys Med Biol 56:1139–1153. https://doi.org/10.1088/0031-9155/56/4/016
[31]. Orozco HM, Villegas OOV, Maynez LO, Sanchez VGC, de Jesus Ochoa Dominguez H (2012) Lung Nodule CLASSIFICATION in Frequency Domain Using Support Vector Machine. IEEE, In international conference on information science, signal processing and their application
[32]. Shao H, Cao L, Liu Y (2012) A detection approach for solitary pulmonary nodules based on CT images. IEEE, 2nd international conference on computer science and network technology.
[33]. de Carvalho Filho AO, Silva AC, de Paiva AC, Nunes RA, Gattass M (2016) Lung-nodule classification based on computed tomography using taxonomic diversity indexes and an SVM. Springer, Journal of Signal Processing Systems, DOI 87:179–196. https://doi.org/10.1007/s11265-016-1134-5
[34]. Kumar D, Wong A, Clausi DA (2015) "Lung nodule classification using deep features in CT images", IEEE, 12th conference on computer robot vision, pp 133-138. DOI. https://doi.org/10.1109/CRV.2015.25
[35]. Sangamithraa, Govindaraju (2016) Lung tumour detection and classification using EK-mean clustering. IEEE WiSPNET
[36]. Makaju S, Prasad AA, Elchouemi S (2018) Lung Cancer detection using CT scan images. Elsevier, Procedia Computer Science 125:107–114
[37]. Jin X-Y, Zhang Y-C, Jin Q-L (2016) Pulmonary nodule detection based on CT

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

images using Convolution neural network. IEEE, 9th International Symposium on Computational Intelligence and Design, pp. 202– 204.

[38]. Alakwaa W, Nassef M, Badr A (2017) Lung Cancer detection and classification with 3D convolutional neural network (3D-CNN). International Journal of Advanced Computer Science and Applications (IJACSA) 8(8):409–417.

[39]. Skourt BA, Nikolov NS and Majda A 2019 Feature-Extraction Methods for Lung-Nodule Detection: A Comparative Deep Learning Study Int. Conf. on Intelligent Systems and Advanced Computing Sciences (ISACS) pp. 1-6.

[40]. Liu Z, Yao C, Yu H and Wu T 2019 Deep reinforcement learning with its application for lung cancer detection in medical Internet of Things Future Generation Computer Systems pp 1-9.

[41]. Zhang Q and Kong X 2020 Design of Automatic Lung Nodule Detection System Based on Multi-Scene Deep Learning Framework. IEEE Access.

[42]. Toğaçar M, Ergen B and Cömert Z 2020 Detection of lung cancer on chest CT images using minimum redundancy maximum relevance feature selection method with convolutional neural networks Biocybernetics and Biomedical Engineering 40(1) 23-39.

[43]. Z. Shi et al., “A deep CNN based transfer learning method for false positive reduction,” Multimed. Tools Appl., vol. 78, no. 1, pp. 1017– 1033, 2019, doi: 10.1007/s11042-018-6082-6.

[44]. Nasser IM and Abu-Naser SS 2019 Lung Cancer Detection Using Artificial Neural Network. International Journal of Engineering and Information Systems (IJEAIS) 17-23.

[45]. alyasriy, hamdalla (2020), “The IQ-OTHNCCD lung cancer dataset”, Mendeley Data, V1, doi: 10.17632/bhmdr45bh2.1. https://data.mendeley.com/datasets/bhmdr45bh2/1

[46]. Hassanat, A. (2018). “Furthest-pair-based binary search tree for speeding big data classification using k-nearest neighbors”. Big Data, 6(3): 225-235.

[47]. S. Al-Showarah et. al. 2020. “The Effect of Age and Screen Sizes on the Usability of Smartphones Based on Handwriting of English Words on the Touchscreen”, Mu’tah Lil-Buhuth wad-Dirasat, Natural and Applied Sciences series, Vol. 35, No. 1, 2020. ISSN: 1022-6812.

[48]. Nageswaran S, Arunkumar G, Bisht AK, Mewada S, Kumar JNVRS, Jawarneh M, Asenso E. Lung Cancer Classification and Prediction Using Machine Learning and Image Processing. Biomed Res Int. 2022 Aug 22;2022:1755460. doi: 10.1155/2022/1755460. PMID: 36046454; PMCID: PMC9424001.

[49]. Suleyman A. AlShowarah, “Breast Cancer Detection System using Deep Learning Based on Fusion Features and Statistical Operations” International Journal of Advanced Computer Science and Applications(IJACSA), 14(11), 2023. http://dx.doi.org/10.14569/IJACSA.2023.0141165