An Intelligent System for Lung Cancer Diagnosis

Hafssa Ahmed shukur [1] , Al-Nuaimi Bashar Talib [2] , Ruaa Azzah Suhail [3]

[1,2,3] Department of Computer Science , College of Science , University of Diyala, Iraq

Email [1] : hafssaazawy66@gmail.com , Email [2] : bashartalib6@gmail.com , Email [3] : Ruaaizzat@gmail.com

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 4 June 2021

Abstract: Nowadays, lung cancer has become one of the important topics for researchers in medical technology. Lung cancer is one of the deadliest diseases in the world and is a subject of concern because no actual treatment has been found for this disease yet.

Computer-aided diagnosis (CAD) is one of the widely used imaging techniques for detecting and grading lung cancer. This thesis presents a proposed system for classifying lung cancer after its detection with the help of machine learning algorithms, where several steps are used in the form of stages which include the stage of data acquisition, data pre-processing, and classification. The theater.

The dataset used is obtained from the archive (data scientist), which contains 1,000 samples and 25 features. The first proposed model is based on the Support Vector Machine (SVM) classifier), and the second proposed model uses an artificial neural network (ANN) classifier and compares the accuracy and time taken for each model. Each model implements two types of preprocessing algorithms (standardization and normalization).

The results showed that the first proposed model using (SVM) with normalization had an average accuracy of 98.21%, while with standardization the accuracy was 100.00%. The second proposed model using (ANN) with normalization and standardization with an accuracy of 100.0%.

Keywords: lung , Cancer ,SVM, Algorithm ANN

1. Introduction

Lung Cancer is a prevalent disease where deaths due to it are increasing nowadays. Through all lung cancer is predominant for men and women compared to all other cancers. It is caused due to smoking and eating tobacco causing cancer in lung tissues. The cancerous nodules are called malignant tumors it occurs because of uncontrollable cell growth in lung tissues. Despite advances in detection and diagnosis and therapies, many people still develop fatal lung cancer [1].

Lung Cancer is classified into two major groups based on cell size: i. cell size is small and ii. cell size is large. Cancer is divided into four stages and this staging has been done according to the size of the tumor and node location. Lung cancer is one of the most dangerous cancers, with a low survival rate following diagnosis. In general, lung cancer affects 75 % of females and 84 % of males who smoke. About 10-15% of cases occur in people who have never smoked [2]. Recent advances in computed tomography (CT) imaging have resulted in early diagnosis of lung cancer.

Two common screening tests involve the use of chest X-ray (CXR) and the use of low-dose computed tomography (LDCT) scan of the chest. Computer-aided detection (CAD) systems have the ability to improve/assist radiologists, and their workflow decreases error outcomes. CAD is a technique for locating abnormal regions (masses, nodules, and polyps) in the body and providing a location to radiologists.

CAD has been used in a variety of medical imaging techniques, including magnetic resonance imaging (MRI), computed tomography (CT), and ultrasound. In particular, region of interest (ROI) based segmentation, feature extraction (kind of feature), and classification are three important stages in CAD systems used for cancer diagnosis and detection [3]. A region generating mechanism is deployed to produce a large number of candidate regions having a high likelihood of containing the pulmonary nodules [4].

The purpose of different ML and DL algorithms may be to improve data interpretation quality, consistency, and/or capacity in diagnostics. This paved the way for the formation of a sub-area called Deep Learning (DL) under the field of ML [5]. ML is a set of methods in mathematics and statistics for training a computer for specific tasks by finding hidden and useful patterns from inputs [6].

There have been a lot of important efforts for automatic lung cancer classification by recurrent neural networks (RNN), convolutional neural networks CNN, and other approaches used CT [7]. and other technique such as the Back Propagation neural (BPNN), and the Nearest K (KNN) [8]. Support vector machines (SVM) can be used when our data has exactly two classes. Developed automated diagnosis technique for lung image CT scans (Optimal Deep Neural Network (ODNN) and Linear Discriminate Analysis (LDA) [9]. Multiclass Multiple kernel learning classifier (m-MKL) [10].

In this work, SVM are supervised learning for diagnosis of lung cancer. When our data has exactly two groups, Support Vector Machines (SVM) can be used. To classify the data, the best hyper-plane is found, creating two groups of different data points [10].

Artificial neural networks (ANNs) the mathematical model that resolves classification and prediction problems. Inputs, outputs, and (usually) hidden layers are neural network layers that transform the input into anything that can be used in the output layer, in many ways, hide layers translate the input into something that can be used by the output layer In two phases of testing and evaluation the ANN Model passes when a neural cancer prediction network is used when a dataset is used to train the network [11].

2. Related work

Several researchers have conducted studies on the diagnosis of lung diseases utilizing a lung tumor dataset. In this part of the research, many previous studies are reviewed to identify the techniques used, methods, and the results obtained. In paper [12] focuses on the need for an immersive learning system that enables effective condition detection in patients. PCA successfully combines related qualities and creates a dissipated showcase of its constituents. The number of principal components to be preserved is determined by examining the Scree plot. With a small amount of input, Support Vector Machines (SVM) outperforms other classification algorithms. The confusion matrix is used to calculate the expectation. The developed model has an accuracy of 0.87 and an error rate of 0.3 in the early detection of various stages of malignancy. The paper [13] proposed an Artificial Neural Network for detecting whether lung cancer is found or not in the human body. Symptoms were used to diagnose lung cancer, these symptoms such as Yellow fingers, Anxiety, Chronic Disease, Fatigue, Allergy, Wheezing, Coughing, Shortness of Breath, Swallowing Difficulty, and Chest pain. They were used and other information about the patient as input variables for the proposed ANN model. The proposed model was trained and validated using the lung cancer dataset. The proposed model was evaluated and tested. The accuracy rate it gave us was 99.01 %. The paper [14] used image recognition and machine learning to create a variety of computer-aided systems. Different segmentation, extraction of features, classification techniques like discreet wavelet transform, Gray level co-occurrence matrix, SVM, Artificial Neural Network, and more are considered. Varied segmentation techniques are considered. The DNN had 97 % accuracy, CNN had 94 % accuracy, ANN had 99 % accuracy, and the SVM classifier had 96 % accuracy, according to the researchers. The paper [6] They created a prediction model of radiomics to improve the classification of the lung nodule (PN) in low dose CT. The Lung CT Screening Reporting and Data System (Lung-RADS) model for the early diagnosis of lung cancer is comparable with the American College of Radiology (ACR). Reviewed 72 PNs of the Lung Image Database Image Selection (41 malignant and 31 benign) (LIDC-IDRI). Every PN has extracted 103 CT radiomic features.

A prediction model was created using a vector support machine (SVM), coupled with a minimum absolute shrinkage and selection operator (LASSO). (10×10-fold CV) cross-validation was used to test the SVM-LASSO model accuracy. The precision of the two features was 84.6% and 0.89 AUC.

3. Methodology

As shown in Figure 1, this research investigates the performance of two different classifiers in cancer detection and the influence of the standardization and normalization preprocessing techniques. Then, by comparing the performance measures of these combinations, the suitability of these classifiers and preprocessing techniques in cancer detection can be measured. This measurement is conducted based on the quality of the predictions, per each class, as well as the time required by the classifier to predict the class of each input, which indicates the complexity of the model.

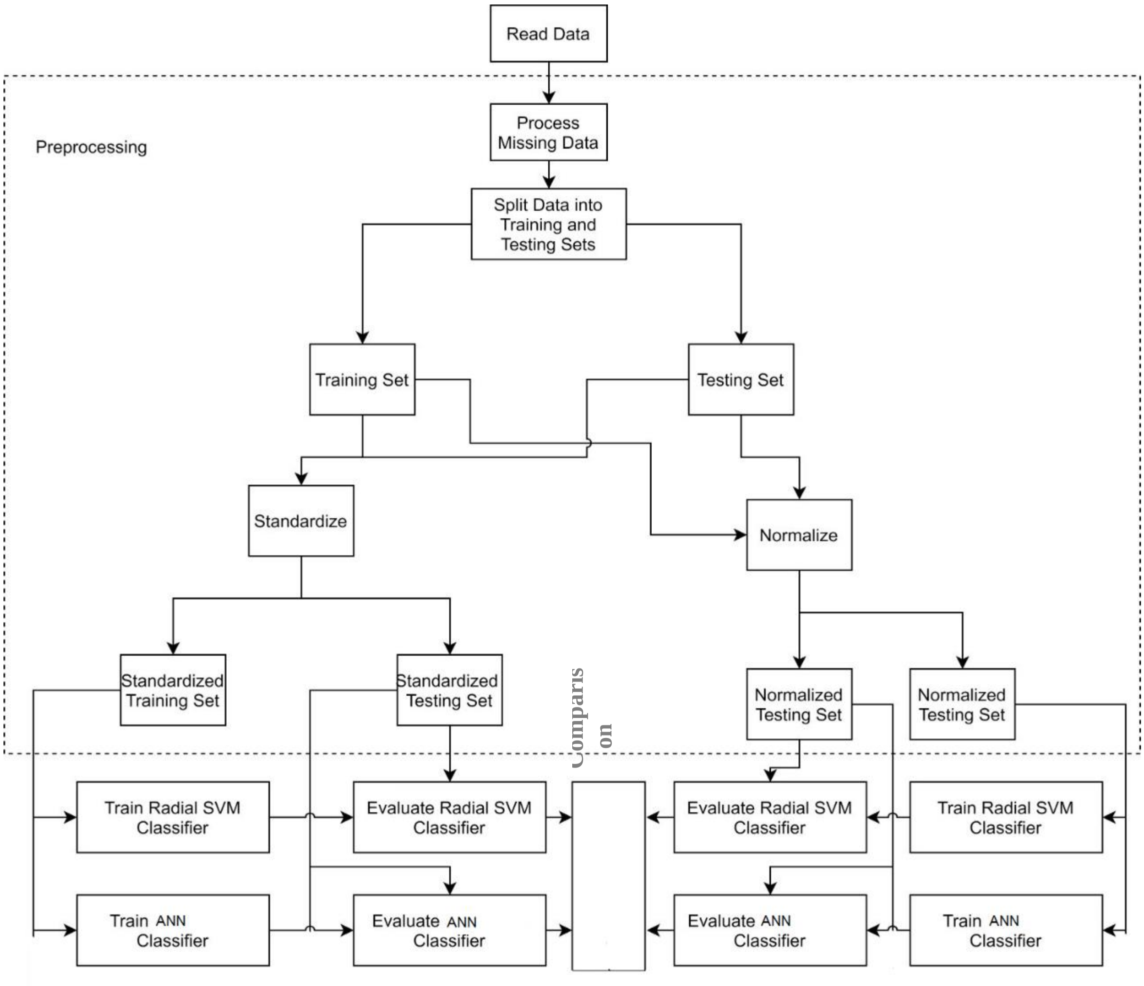

Figure 1: Block diagram of the research methodology.

3.1 Data Collection

Database used in this research work is obtained from Lung Cancer Dataset “data.world”[15]. The dataset has 25 attribute and 1000 instance. The first column distinguishes every instance with an ID number while the last column identifies the class label that describes the level of spread of the tumor, the labels being high or medium or low. The 23 attributes of data (excepting the class label and ID number) that describes about age, Gender, Air Pollution, Alcohol use, Dust Allergy, OccuPational Hazards, Genetic Risk, chronic Lung Disease, Balanced Diet, Obesity, Smoking, Passive Smoker, Chest Pain, Coughing of Blood, Fatigue, Weight Loss, Shortness of Breath, Wheezing, Swallowing Difficulty, Clubbing of Finger Nails, Frequent Cold, Dry Cough, Snoring.

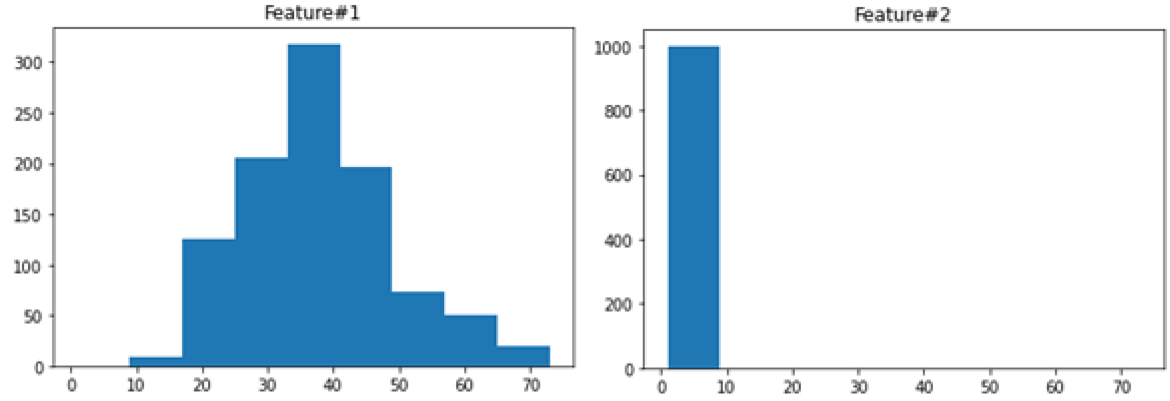

Figure 2: Feature#1,#2 from applying histogram on data set.

3.2 Data Preprocessing

Preprocessing is an important pre-requirement for any data examination. It is generally an excellent plan to set up the information in such a manner to uncover the structure of the data to the machine learning calculations that needs to use. Data preprocessing techniques are well known in enhancing the capability power of classification systems [16].

A. Min-Max Normalization

This method indicates performing linear transformations on main data. Assume max$_A$ and min$_A$ representing maximum and minimum values regarding the features, where the value is given to the feature (A) within a range [new_min$_A$, new_max$_A$] according to equation (1) [17].

$$v_i^{`} = \frac{v_i - min_A}{max_A - min_A}\left(new_{max_A} - new_{min_A}\right) + new\_min_A \qquad (1)$$

Where $v_i^{`}$ represent features that normalized, as shown figure (3) First and second feature after applying Min-Max Normalization

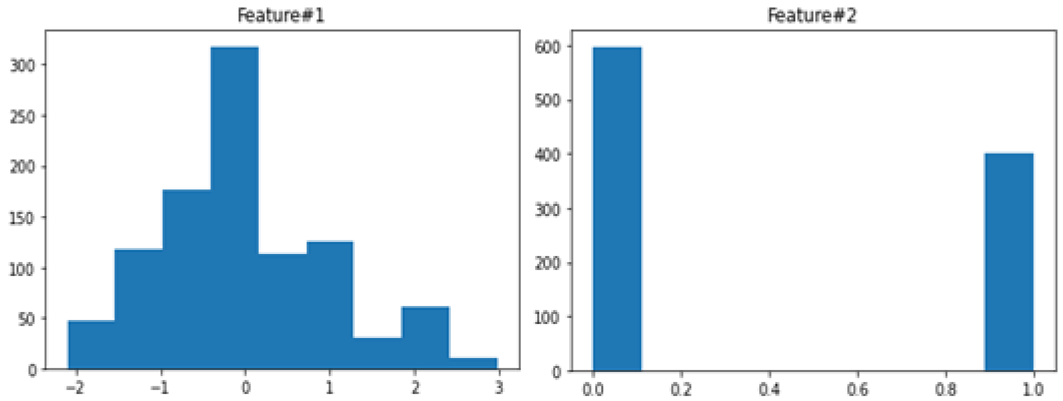

Figure 3: First and second feature after applying normlization.

B. Z-Score Normalization

This method includes computing the mean first then computing attributes' standard deviations, after that, normalization will be carried out though the use of equation (2) [17]:

$$v^{`} = \frac{v-\mu}{\sigma} \qquad (2)$$

Where v`represent the normalized value, v is the experimental value, μ is the mean and σ indicates standard deviation. As shown in figure (4) applying standardization on first and second features

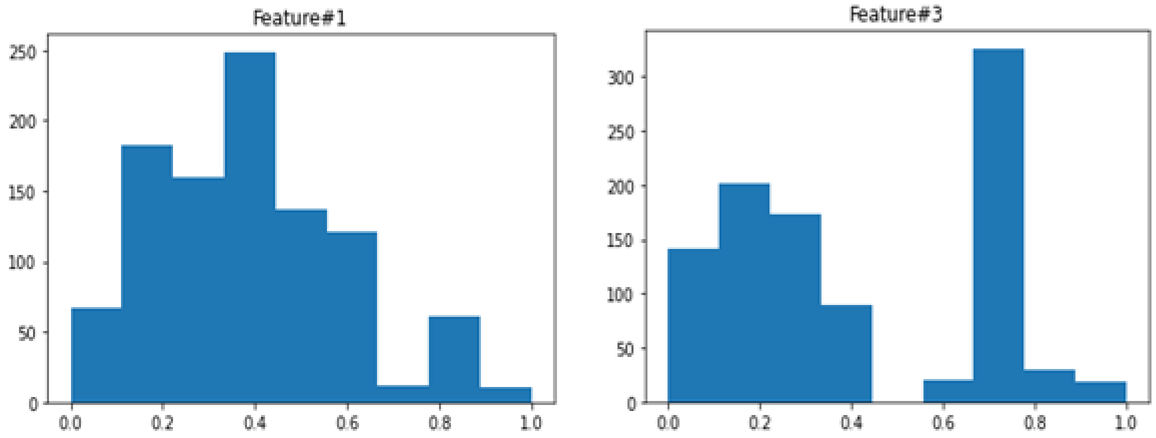

Figure 4: First and second feature after applying standardization.

3.3 Classification

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data.

3.3.1 Support Vector Machine (SVM)

Support Vector Machine (SVM): SVM is one of the most popular classification algorithms which have an elegant way of transforming nonlinear data. Classification strategy of SVM is well explained in [12]. Hyper plane is the important tool of SVM that separates the data points in such a manner that the margin between two classes will be wide and the data points will be as far as possible. In this way, hyper plane will be creating a decision boundary with support vector points nearer to the left and right hyper-plane. Linear and nonlinear SVM models is used for this lung cancer prediction study.

kernels are used to non-linearly map the input data to a high-dimensional space [18]. It is possible to write the hyper-plane function in formula(3) [19].

$$K(x, x') = \{\Phi(x), \Phi(x')\}$$

$$\mathrm{f(x')} = \sum_{n=1}^{N} \alpha_n y_n K(x, x') + b \qquad (3)$$

Where α_n is a Lagrange multiple, is support vector information, and y_n is a membership class label (+1, −1) with n = 1, 2, 3,……., N[19].

The kernel functions are listed below were used:

Radial Basis Function: Radial basis functions most commonly with a Gaussian form.

$$K(x, x') = exp\left(-\frac{\|x-x\prime\|^2}{2\sigma^2}\right) \qquad (4)$$

Algorithm (1): Multi-class Support vector machine classifier

Input: Training _set (Features function)
Output: Class name

```
Begin
Step 1: Establish the training label for all training sets and identify 3 classes
Step 2: Specify the kernel function ((RDF function), as calculated in the second chapter in the equation (4)) that
        was applied to map the training set to kernel space by hyperplane computation, its calculation was also
        defined in the previous chapter in equation (3).
Step 3: Segregated data to 1 against all methods.
Step 4: Passes testing set on SVM depended on the hyperplane to identify the class name.
Step 5: Return class name
End
```

3.3.2 Artificial Neural Network (ANN)

Artificial neural networks (ANN) are a popular machine learning technique inspired by the biological neural network in the human brain. The ANN is a system of processing information with definite features performance in common with neural networks in biology. ANN is improved to be a mathematical generalization model inspired by neural biology or human cognition. We can describe many principles such as [10]:

1. The processing of information is performed in many elements of the network called neurons.
2. The signals to the neurons in the network delivered through connection links.
3. Connection links have their weight, which multiplies with the transferred signal in a classic neural network.
4. Activation function of each neuron that is usually nonlinear and applies to the input of the net which is the summation of the weight signals input to determine the output signals [20].

An artificial neural network with two hidden layers is implemented, as shown in Table (1). The number of neurons in the input layer is set according to the number of features that characterize each input. Then, the two hidden layers, each with ten neurons, are placed with the ReLU activation function. Finally, an output layer with three neurons, which is set based on the number of classes in the dataset, is placed with the Softmax activation function, as it is a multi-class classification problem

Table 1: Structure of the implemented neural network.

Layer	Number of Neurons	Activation Function	Number of Parameters
Input	23	-	-
Hidden#1	10	ReLU	240
Hidden#2	10	ReLU	110
Output	3	Softmax	33
Total Parameters:		383	

3.4 Experimental Results and Discussion

3.4.1 Results of first Proposed Model (SVM)

This stage is an examination of the system by testing it with the remainder of the data that is not labeled to classify the data of lung cancer. Wherein the training process for data is done in a nonlinear (SVM) using a linear kernel function. During the training, phase computed the values that represent the hyper-plane between the classes, according to the equation (4) and (3). At this stage also the results were presented in the form of a confusion matrix that shows the accuracy of each class in the testing stage for two cases (model with applying normalization and with standardization) as shown in table (2). The display of the results using the confusion matrix as in figure (5) and figure (6).

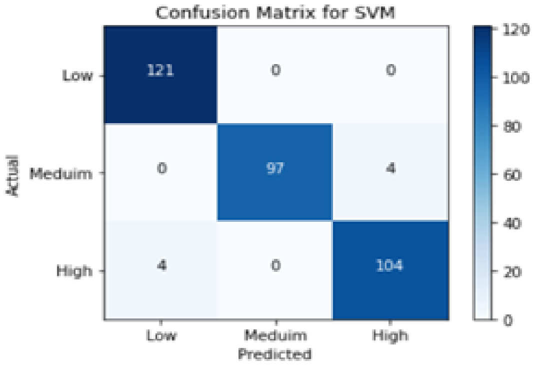

Figure 5: The Confusion Matrix for SVM with normlization

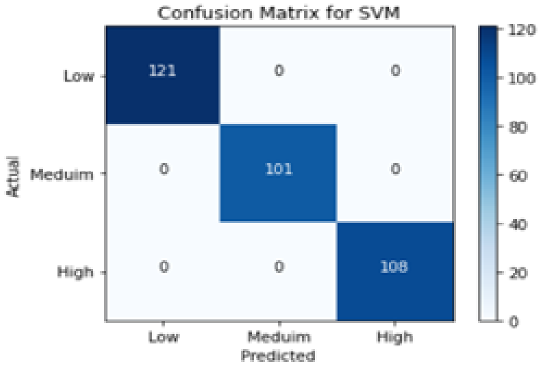

Figure 6: The Confusion Matrix for SVM with standardization

Table 2: Accuracy measures of SVM with normalization and standardization

Accuracy measures of SVM with normalization			
	"Precision"	"Recall"	"F1-score"
Low	"1.00"	"0.96"	"0.98"
Mid	1.00	1.00	1.00
High	0.97	1.00	0.98
"Micro avg""	"0.98"	"0.98"	"0.98"
"Macro avg"	"0.99"	"0.99"	"0.99"
"Weighted avg"	"0.98"	"0.98"	"0.98"
Accuracy Measures of SVM with standardization			
	"Precision"	"Recall"	"F1-score"
Low	"1.00"	"1.00"	"1.00"
Mid	"1.00"	"1.00"	"1.00"
High	"1.00"	"1.00"	"1.00"
"Micro avg""	"1.00"	"1.00"	"1.00"

"Macro avg"	"1.00"	"1.00"	"1.00"
"Weighted avg"	"1.00"	"1.00"	"1.00"

3.4.2 Results of Second Proposed Model (ANN)

At this stage also the results were presented in the form of a confusion matrix that shows the accuracy of each class in the testing stage for two cases (model with applying normalization and with standardization). The display of the results using the confusion matrix as in figure (7) and figure (8). Table (3) displays Accuracy Measures with normalization and standardization.

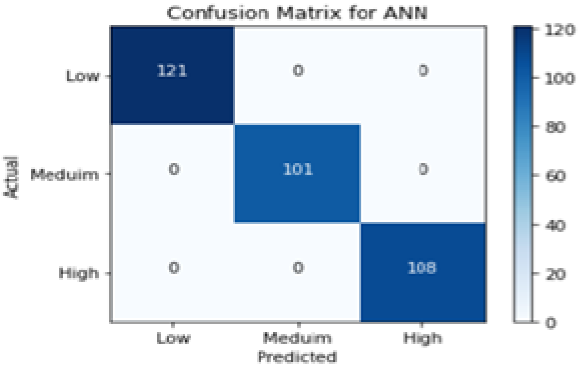

Figure 7: The Confusion Matrix for ANN with normalization

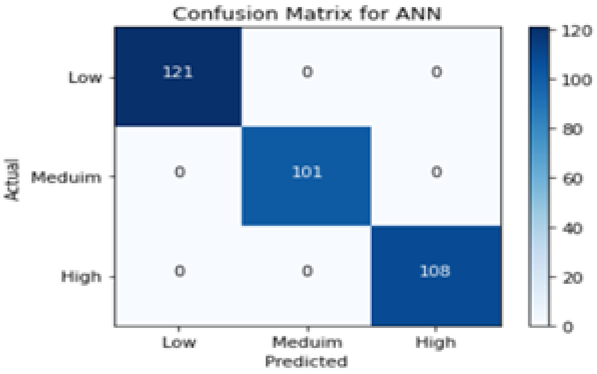

Figure 8: The Confusion Matrix for ANN with standardization

Table (3): Accuracy Measures of ANN with normalization and standardization

	"Precision"	"Recall"	"F1-score"
Low	"1.00"	"1.00"	"1.00"
Mid	"1.00"	"1.00"	"1.00"
High	"1.00"	"1.00"	"1.00"
"Micro avg""	"1.00"	"1.00"	"1.00"
"Macro avg"	"1.00"	"1.00"	"1.00"
"Weighted avg"	"1.00"	"1.00"	"1.00"
Accuracy Measures of ANN with standardization			
	"Precision"	"Recall"	"F1-score"
Low	"1.00"	"1.00"	"1.00"
Mid	"1.00"	"1.00"	"1.00"
High	"1.00"	"1.00"	"1.00"
"Micro avg""	"1.00"	"1.00"	"1.00"
"Macro avg"	"1.00"	"1.00"	"1.00"
"Weighted avg"	"1.00"	"1.00"	"1.00"

Validation and training error is important, as long as it continues to decrease, training should continue, the number of periods can be set as high as possible and training can be terminated based on error rates. We need to determine the acceptable fault tolerance for model first, then iterate as needed until it reaches the desired threshold (i.e. until convergence). In figure (9) and figure (10) shows the graphic to how the error changes with each training session, and it expresses the effect of updating the weights and trainable coefficients in the network on the accuracy of the mathematical model we set epochs to 100.

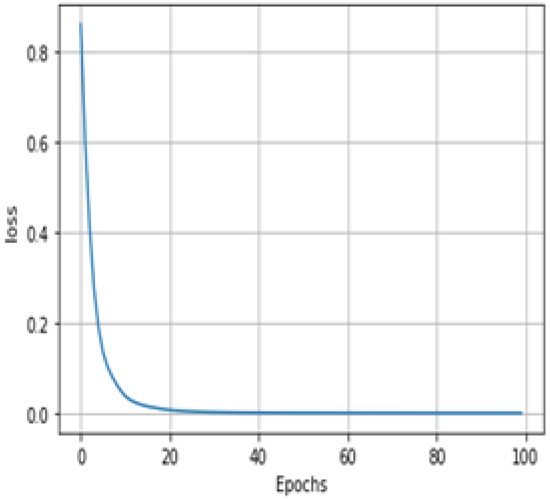

Figure 9: Loss with standardization using 100 epochs

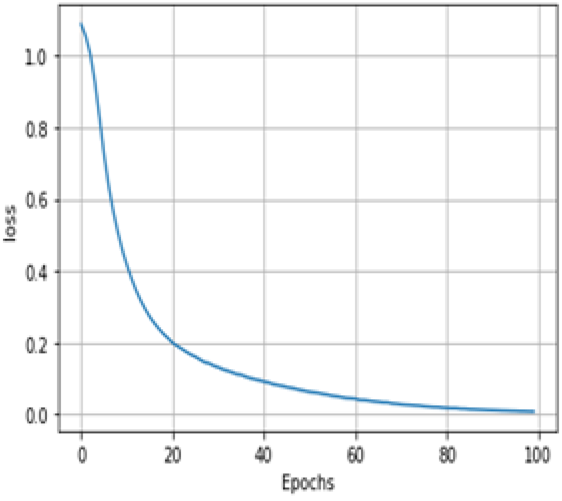

figure10: loss with normalization using 100 epochs

At epochs = 100 we get an accuracy of 100%

3.4.3 Comparison, Between the Proposed Systems (Using the SVM and Using ANN Algorithm)

The comparison between the proposed systems is very important to show the strengths and weaknesses of each of them. When choosing this topic which is the detection and classification of lung cancer, the SVM, ANN algorithms were used. It is worth noting that the same dataset was used, as well as the same data division for training and testing, which is 67% for training and 33% for testing for all algorithms as shown in table (4).

Table (4): Comparison of accuracy between the proposed systems

Type	Accuracy	Time
SVM with normalization	98.21%	0.022 second
SVM with standardization	100.0%	0.011 second
ANN with normalization	100.0%	10.190 second
ANN with standardization	100.0%	9.922 second

Table 5 illustrates the comparison between the proposed and other recognition algorithms applied on the same dataset.

Author(s), Year	**Algorithm for (classification)**	**Accuracy**
Ibrahim M. Nasser[21]	ANN	96.67 %
Mohammed et al[13]	ANN	99.01%
Manju, Athira, and Rajendran[13]	SVM	87%

4. Conclusion and Future Work

This investigation charts the thinking about the importance of the pre-malignant stage in the early detection of lung cancer. Real and inconsistent data are processed and purified, we used two preprocessing methods (normalization and standardization). We used different SVM (non-linear) and ANN techniques. This study showed that the neural network and SVM classifier is able to diagnose lung cancer with high accuracy and lower execution time. The obtained confusion matrix can effectively name and identify the hidden information in the data set. Here the performance is formulated in terms of rating report roasted on confusion matrix and its calculations. The accuracy score of this model is presented as a 3×3 matrix in contrast to the traditional 2×2 matrix. The work gives scope to expand to a level of improvement that can ensure ground-breaking accuracy in machine learning diagnostics.

References

[1] D. Nagajyothi, R. Addagudi, T. Gunda, S. S. Logitla, and G. Arun, "Detection of lung cancer using SVM classifier," Int. J. Emerg. Trends Eng. Res., vol. 8, no. 5, pp. 2177–2180, 2020, doi: 10.30534/ijeter/2020/113852020.

[2] P. Katiyar and K. Singh, "A comparative study of lung cancer detection and classification approaches in CT images," 2020 7th Int. Conf. Signal Process. Integr. Networks, SPIN 2020, pp. 135–142, 2020, doi: 10.1109/SPIN48934.2020.9071240.

[3] S. R. Jena and S. T. George, "Morphological feature extraction and KNG-CNN classification of CT images for early lung cancer detection," Int. J. Imaging Syst. Technol., no. November 2018, pp. 1–13, 2020, doi: 10.1002/ima.22445.

[4] P.-P. Ypsilantis and G. Montana, "Recurrent Convolutional Networks for Pulmonary Nodule Detection in CT Imaging," pp. 1–36, 2016, [Online]. Available: http://arxiv.org/abs/1609.09143.

[5] A. K. Arslan, S. Yasar, and C. Colak, "An Intelligent System for the Classification of Lung Cancer Based on Deep Learning Strategy," 2019 Int. Conf. Artif. Intell. Data Process. Symp. IDAP 2019, 2019, doi: 10.1109/IDAP.2019.8875896.

[6] W. Choi et al., "Radiomics analysis of pulmonary nodules in low-dose CT for early detection of lung cancer," Med. Phys., vol. 45, no. 4, pp. 1537–1549, 2018, doi: 10.1002/mp.12820.

[7] D. Moitra and R. Kr. Mandal, "Classification of non-small cell lung cancer using one-dimensional convolutional neural network," Expert Syst. Appl., vol. 159, p. 113564, 2020, doi: 10.1016/j.eswa.2020.113564.

[8] C. Lu, Z. Zhu, and X. Gu, "An intelligent system for lung cancer diagnosis using a new genetic algorithm based feature selection method," J. Med. Syst., vol. 38, no. 9, 2014, doi: 10.1007/s10916-014-0097-y.

[9] A. Shaffie et al., "A New System for Lung Cancer Diagnosis based on the Integration of Global and Local CT Features," IST 2019 - IEEE Int. Conf. Imaging Syst. Tech. Proc., pp. 1–6, 2019, doi: 10.1109/IST48021.2019.9010466.

[10] M. Ajin and L. Mredhula, "Diagnosis of Interstitial Lung Disease by Pattern Classification," Procedia Comput. Sci., vol. 115, pp. 195–208, 2017, doi: 10.1016/j.procs.2017.09.126.

[11] T. Pandiangan, I. Bali, and A. R. J. Silalahi, "Early lung cancer detection using artificial neural network," Atom Indones., vol. 45, no. 1, pp. 9–15, 2019, doi: 10.17146/aij.2019.860.

[12] B. R. Manju, V. Athira, and A. Rajendran, "Efficient multi-level lung cancer prediction model using support vector machine classifier," IOP Conf. Ser. Mater. Sci. Eng., vol. 1012, p. 012034, 2021, doi: 10.1088/1757-899x/1012/1/012034.

[13] O. Mohammed et al., "Artificial Neural Network for Lung Cancer Detection," vol. 4, no. 11, pp. 1–7, 2020.

[14] S. T. H. Kieu, A. Bade, M. H. A. Hijazi, and H. Kolivand, "A Survey of Deep Learning for Lung Disease Detection on Medical Images: State-of-the-Art, Taxonomy, Issues and Future Directions," J. Imaging, vol. 6, no. 12, p. 131, 2020, doi: 10.3390/jimaging6120131.

[15] H. Azawe, "hafsa link data," 2017, [Online]. Available: https://data.world/cancerdatahp/lung-cancer-data.

[16] A. Mahfouz, A. Abuhussein, D. Venugopal, and S. Shiva, "Ensemble classifiers for network intrusion detection using a novel network attack dataset," Futur. Internet, vol. 12, no. 11, pp. 1–19, 2020, doi: 10.3390/fi12110180.

[17] M. F. Mohamad Mohsin, A. R. Hamdan, and A. Abu Bakar, "The Effect of Normalization for Real Value Negative Selection Algorithm," Commun. Comput. Inf. Sci., vol. 378 CCIS, pp. 194–205, 2013, doi: 10.1007/978-3-642-40567-9_17.

[18] M. Oujaoura, B. Minaoui, M. Fakir, R. El Ayachi, and O. Bencharef, "Recognition of Isolated Printed Tifinagh Characters," Int. J. Comput. Appl., vol. 85, no. 1, pp. 1–13, 2014, doi: 10.5120/14802-3005.

[19] R. Huerta, F. Corbacho, and C. Elkan, "Nonlinear support vector machines can systematically identify stocks with high and low future returns," Algorithmic Financ., vol. 2, no. 1, pp. 45–58, 2013, doi: 10.3233/AF-13016.

[20] K. Mehrotra, C. Mohan, and S. Ranka, "Elements of Artificial Neural Networks," Elem. Artif. Neural Networks, no. May, 2019, doi: 10.7551/mitpress/2687.001.0001.

[21] I. M. Nasser, "ANN for Lung Cancer Detection," vol. 3, no. 3, pp. 17–21, 2019.