Lung cancer Prediction and Classification based on Correlation Selection method Using Machine Learning Techniques

1st Dakhaz Mustafa Abdullah
Technical College of Informatics, Akre Duhaok Polytechnic University
Duhok, Iraq
dakhaz.abdullah@dpu.edu.krd

2nd Adnan Mohsin Abdulazeez
President of Duhaok Polytechnic University
Duhok Polytechnic University
Duhok, Iraq
adnan.mohsin@dpu.edu.krd

3rd Amira Bibo Sallow
College of Engineering
Nawroz University
Duhok, Iraq
amira.bibo@nawroz.edu.krd

https://doi.org/10.48161/qaj.v1n2a58

Abstract—Lung cancer is one of the leading causes of mortality in every country, affecting both men and women. Lung cancer has a low prognosis, resulting in a high death rate. The computing sector is fully automating it, and the medical industry is also automating itself with the aid of image recognition and data analytics. This paper endeavors to inspect accuracy ratio of three classifiers which is Support Vector Machine (SVM), K-Nearest Neighbor (KNN) and, Convolutional Neural Network (CNN) that classify lung cancer in early stage so that many lives can be saving. Basically, the informational indexes utilized as a part of this examination are taken from UCI datasets for patients affected by lung cancer. The principle point of this paper is to the execution investigation of the classification algorithms accuracy by WEKA Tool. The experimental results show that SVM gives the best result with 95.50%, then CNN with 92.11% and KNN with 88.40%.

Keywords— Lung Cancer, Machine Learning, SVM, KNN, CNN.

I. INTRODUCTION

Cancers exist in several organs, and simultaneously, and different types of cancer occur in various organs of the body. The illness may even go unnoticed for long periods of time. According to WHO reports, cancer may be prevented if it is detected early enough. The patient's life span will be extended whether he or she receives an early prognosis [1][2][3][4]. Lung cancer has a low prognosis that differs greatly depending on tumor staging at the time of diagnosis. Lung cancer is divided into two types of clinical practice: non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC) [5][6]. It is, in reality, a malignant tumor characterized by unregulated cell tissue formation. Lung cancer developed mostly as a result of long-term tobacco use [7]. According to research, a stable individual may be affected by nineteen distinct forms of cancer. Lung cancer has the largest death rate among all of these tumors. This disease is expected to kill over 1.7 million people per year [8]. In the area of machine learning (ML) research has already grown a great deal, which is helpful to reduce human laborers. ML combines statistics and computers in the area of artificial intelligence to create algorithms that become more efficiently when subject to relevant data[9][10]. Many systems lack adequate detection accuracy, and some systems must also be developed in order to reach the highest accuracy of 100%. Pulmonary cancer identification and classification were based on machine learning techniques and image processing techniques [11]. However, some signs of lung cancer patients, such as their smoking rate, may aid in early detection of the disease [12][13][14]. Researchers started to use machine learning for medical diagnosis after the advent of artificial intelligence, using a machine learning approach to investigate the classification of diseases in traditional Chinese medicine clinical data (TCM). Valuable guidelines on diagnosis of brain disturbances from network architecture aspects, function learning and classification prediction via the method of machine learning, and provided through the machine learning method and the implementation of the brain network based on machine learning [15]. It will be a key step towards improved early detection [16].

This paper provides an effective method to predict lung cancer in early stage with high accuracy ratio. The dataset used is taken form UCI machine learning repository. Then apply three classifier Support Vector Machine (SVM), K-Nearest Neighbor (KNN) and, Convolutional Neural Network (CNN) to endeavors inspect accuracy ratio of three classifier by using WEKA tool. the present study is aid to develop a Machine Learning Models to detect the lung cancer with better accuracy.

This paper is organized as follows. Section 2 introduce to lung cancer. Section 3 Material & Methods that used in this paper. describes related work in Section 4. then Section 5 present the theory, introduction to machine learning and their types also confusion matrix. Section 6 present the Performance evaluation and results. Finally, Section 7 Conclusion.

II. LUNG CANCER

Carcinogenesis is the unchecked proliferation of one or more cell types. Good tissues do not support the growth of...
normal cells, and when they do, they separate quickly and become tumors. Primary lung cancer originates elsewhere in the body and spreads to the lungs, while secondary lung cancer starts elsewhere in the body and then spreads from there. It’s one of the most aggressive types of cancer and a life-threatening threat to the human body [17]. If this unchecked development can be identified correctly at an early point, it can help to diagnose the likelihood of unnecessary surgery and improve the chance of recovery. Chronic Obstructive Pulmonary (COPD) illness attacks the areas of the lungs and causes diseases such as measles, influenza, pneumonia, and other respiratory issues such as asthma. Small Cell Lung Cancer (SCLC) or oat cell cancer and Non-Small Cell Lung Cancer (NSCLC) are the two main forms of lung cancer that develop and expand in separate ways and may be handled accordingly. Within the non-small cell lung cancer category, there are three subtypes (adenocarcinomas, squamous cell carcinomas, large cell carcinomas) fig (1) show the two types of lung cancer. So Mixed small cell/large cell cancer is a disease that occurs where a patient shows symptoms of both types of cancer. (NSCLC) Adenocarcinoma is more common and progresses more slowly than small cell lung cancer. Small cell lung cancer is linked to smoking which progresses more rapidly by becoming a large tumor that will spread across the body [18][19].

Fig. 1. Lung Cancer Types

III. MACHINE LEARNING

Machine learning is a subfield of Artificial Intelligence [20]. Machine Learning is also used for complex data classification and decision making [21][22]. In general, the implementation of algorithms aids the machine’s learning. Machine learning gives systems the opportunity to learn automatically and improve over time without being directly configured. The implementation of algorithms aids the computer in learning and making the required decisions [11][23]. Machine Learning strategies and activities are narrowly divided into three categories:

a. Supervised learning

Machine learning, in its most simple form, employs programmed algorithms that learn and refine their functions by processing input data and making predictions within a reasonable range. These algorithms aim to be predictive more precisely by feeding fresh data [24][25]. While there are several changes in the way machine learning algorithms are grouped. Two categories of issues: grouping problems and back-up problems, are well suited to supervised learning algorithms. The output variable usually takes on a limited number of discrete values [26][27].

b. Un-Supervised Learning

Unsupervised Learning is a form of learning that occurs without the presence of a supervisor [28]. The machine is given some sample inputs, but no output is generated in the method of learning. Since there is no optimal value over here, categorization is used to ensure that the algorithm distinguishes between the datasets correctly. It is the difficulty of finding unknown structure in unidentified details [29][30]. Although there are no testing sets or tests given to the respondent, there are no opportunities to reward a successful solution. Unlike supervised learning and reinforcement learning, unsupervised learning has no teacher, and produces results that are unrelated to prior experience. It is directly connected to density and statistics [21][31].

c. Reinforcement Learning

Reinforcement Learning, this machine learning style comes from interacting with its surroundings Reinforcement learning. A Reinforcement Learning manager learns from the meaning of tasks, and even by explicitly articulated instructions, and decides on previous behaviors by using new techniques. Since specific input/output data sets are not provided, this differs from traditional supervised learning. Instead, the focus is on the presentation, which entails striking a balance between discovery (of uncharted territory) and utilization (of existing data) [32][33].

IV. DEEP LEARNING

Deep learning is a type of machine learning techniques that uses representation learning to categorize important features for classification problems [6]. The primary characteristic of deep learning is its compatibility with features, although it may also learn from data. So, to learn complex features a deep learning integrates the simple features that have learned from data. Deep learning is accomplished using multiple-layer artificial neural networks, such as the Deep Neural Network (DNN), Convolutional Neural Network (CNN), and the Recurrent Neural Network (RNN) [13][23].

V. RELATED WORKS

Roy et al[34]. They use a combination of image processing biomedical techniques and information discovery in data to improve accuracy and assess precise significance for early detection of lung carcinoma. The representation of the lungs acquired from CT (Computer Tomography) The scan images are pre-processed, and the Region of Interest is segmented (ROI) is performed. The Random Forest procedure is used to distinguish the distinct features. Using an SVM Classifier, the SURF (Speeded Up Robust Functionality) algorithm was used to extract features like entropy, co-relation, power, and variance from Saliency Enhanced images. The image's classification determines if it is safe or toxic (carcinomic). CT scan images were used as the dataset. The SVM classification and random forest algorithm were used to carry out the whole operation. Using SVM classification, the best outcome is achieved. This technique is 94.5 percent effective in general, 74.2 percent sensitive, 66.3 percent recall, and 77.6% specific.

For lung cancer diagnosis, Faisal et al [12] recommend evaluating machine learning classifiers as well as, classifiers such as Multilayer perceptron (MLP), Nave Bayes, Decision
Tree, Neural Network, Gradient Boosted Tree, and SVM are evaluated. The dataset was downloaded from the UCI registry and is used to analyze random forest and plurality voting-based ensembles for predict lung cancer. Gradient Boosted Tree was found to outperform all other person and ensemble classifiers. Gradient-boosted Tree outperformed all others as well as ensemble classifiers, achieving 90% precision, according to performance assessments.

Delta Radiomics uses the machine learning methods proposed by Baskar et al [35] to extract the characteristics of the cancer nodules. Lung cancer nodule malignancy is predicted by using the Support Vector Machine (SVM). The SVM can examine compact features in a lung cancer nodule photograph, and image classification is useful in distinguishing between the multiple nodules. As a result, SVM is recommended as the best tool for diagnosing and detecting lung cancer, with a 90.9 percent accuracy rate.

Boban et al [36]. They use ML algorithms for the 400 lung disease videos, including the Multilayer perceptron (MLP), KNN and SVM classifiers (i.e., CT scan images). The performance is segmented after extraction of features and compares the exactness of the classifier. When a classifier has received a CT scan image, it contains irrelevant content. Gray Level Cooccurrence Matrix (GLCM) is used to pick the most important features (i.e., for removing features). This classification is 98% accuracy for MLP, 70.45% for SVM accuracy, and 99.2% for KNN accuracy.

Using Deep Learning, Sreekumar et al [37] proposed a method for detecting malignant pulmonary nodules from CT scans. To block out the lung areas from the scans, a preprocessing pipeline was used. A 3D CNN model based on the C3D network architecture was used to remove the functionality. For the decrease of false-positives, researchers used the Lung Image Database Consortium (LIDC-IDRI) as well as a few materials from the LUNA16 grand challenge. The end result is a model that predicts the coordinates of malignant pulmonary nodules and demarcates the associated areas using CT scans, for identifying malignant Lung Nodules and estimating their malignancy scores, the final model had a sensitivity of 86 percent.

Banerjee et al. [38] suggested a paradigm for tumor classification, with ANN, Random forests, and SVM as machine learning algorithms. Artificial neural networks are more accurate in both area and texture dependent features. As the precision is compared to the proposed model, it can be shown that accuracy has improved while recall has decreased. MATLAB R2017a was used for digital image analysis, and a Jupyter notebook was used for machine learning classification. Random Forest 79 percent, SVM 86 percent, and ANN 92 percent were the accuracy for region-based features, while Random Forest 70 percent, SVM 80 percent, and ANN 96 percent were the accuracy for texture-based features.

A technique k-Nearest-Neighbors was developed by Maleki et al [39], for which a genetic algorithm was used to efficiently pick features, to reduce the dimensions of the dataset and to improve the speed of the classifier. The experimental approach is used to determine the best value for k to increase the precision of the proposed algorithm. Use of the proposed solution to the database for lung cancer shows 100% accuracy.

Reddy et al [40] propose a model that is successful in detecting the phases of lung cancer using machine learning algorithms. The model combines K-NN, Decision Trees, and Neural Networks structures with the bagging ensemble approach to improve overall prediction accuracy. As opposed to individual algorithms, the proposed model’s estimated outcomes are more accurate. The versions with and without bagging are compared to draw conclusions. The bootstrap aggregating methodology improves the individual models’ performance, with accuracy scores of 97% (Decision Tree), 94%, and 96% (K-NN) respectively (Neural Networks). The integrated model has a score of 0.98 for accuracy. The precision of the integrated model is increased by 3.33 percent.

Gümaydin et al [41] proposed machine learning methods for detecting lung cancer nodules that used Principal Component Analysis, K-NN, SVM, Nave Bayes, Decision Trees, and ANN to detect anomaly. Then, both approaches were compared both after and without preprocessing. The experimental findings indicate that Artificial Neural Networks produce the best results with 82.43 percent accuracy after image processing, while Decision Trees produce the best results with 93.24 percent accuracy without image processing. Standard Digital Image Database, Japanese Society of Radiological Technology (JSRT) CT was used as the dataset (computed tomography).

Early identification of lung nodes from low dose computed tomography (LDCT) images was suggested by Elnakib et al [42]. Initially, the proposed device processes the raw data in order to increase the comparison between low-dose videos. The compact profound learning capabilities of various architectures, including Alex, VGG16 and VGG19 networks are then explored. A genetic algorithm (GA) is trained to identify the most important early detection features for optimizing the derived collection of features. In order to reliably diagnose lung nodules, various forms of classifiers are then checked. The method is validated using the I-ELCAP International Early Lung Cancer Action Project (ELCAP) in 320 photographs from 50 separate topics. With VGG19 and SVM classification, the system suggested achieves the highest 96.25 percent detection precision, 97.5 percent sensitivity and 95 percent specificity.

VI. MATERIAL AND METHODS

A. Dataset

The data used in this work is a lung cancer dataset that was first released in and later made available in the UCI machine learning repository under the name “Lung Cancer Data set”. This dataset was used to show the capability of the optimum discriminant plane in ill-posed situations. This dataset contains data on the pathological forms of lung cancer. It contains 32 observations on three forms of lung cancer using 56 elements[43].

B. Classification Models

To compare the output of the classifiers, three classification methods are used. The smallest number of features was used to attain higher efficiency. The classifier models are defined briefly.
1. **Support Vector Machine**

Support Vector Machine is a supervised learning algorithm that uses the classification method to analyze data and predicate patterns. The texture is divided into two categories or classes by the SVM classifier: regular and abnormal pictures [44]. It is used to effectively map the nodule. SVM is a margin classifier (hyperplane) that separates the two classes, which is why it is often referred to as a non-probabilistic binary classifier. The Support Vector is described as the training data point that is nearest to the classifier, and the Support Vector Machine is the maximum classifier. The gap between the cancer nodules and the hyperplane is as wide as possible [45][46].

2. **K-Nearest Neighbor Classifier**

The KNN algorithm is a supervised classification method. It's a simple algorithm that looks for the nearest fit. The database is compared to the comparison set. The test sample's mark is determined by the closest match of the k nearest neighbors. To calculate the distances between research samples and database samples, various distances such as Euclidean, cosine, similarity, and city block are used[47].

3. **Convolutional Neural Network**

CNN as a supervised deep learning tool, CNN is an excellent choice. This algorithm is suitable for multi-class classification and binary classification (for example, predicting whether or not a diagnostic picture contains a malignant tumor) [48][49]. CNNs are often used to solve a wide range of pattern and image recognition issues. This deep learning approach is effective and appropriate for visual data because of three key characteristics. To begin with, local receptive fields are perfectly matched to the image data specificity of being correlated geographically but uncorrelated globally. Second, since the convolution is applied to the entire image, mutual weights allow for significant parameter reduction without affecting image processing. Finally, a grid-structured image allows for data pooling operations that reduce data complexity without sacrificing valuable information [50][51].

C. **Proposed Model**

The paper suggests a model to predict and classify the lung cancer classes. The proposed model starts with data preprocessing, feature selection, classification and evaluating), figure (2) shows the block diagram for the proposed work.

To define the lung cancer dataset in this article, Weka classifiers were used. WEKA was established by a team of researchers from New Zealand's University of Waikato [52]. It is a java-based open-source platform that can perform data mining and machine learning algorithms, such as data preprocessing, sorting, clustering, and association rule extraction, among other things. WEKA is a popular choice among analysts because of its ease of use and open-source nature [53].

In this paper for the feature selection Correlation Attribute (CA) method were used. CA is a feature subset selection algorithm [1]. It evaluates the attribute by calculating the correlation (Pearson's product moment correlation) between it and the class [4]. The main objective of CA is to obtain a highly relevant subset of features that are uncorrelated to each other. In this way, the dimensionality of datasets can be drastically reduced and the performance of learning algorithms can be improved [2]. Ranker search method used with CA. Based on the Correlation values, features are ranked and those features that are most suitable to be applied in the machine learning algorithm, are filtered [3].

VII. PERFORMANCE EVALUATION MATRICES

A. **Confusion Matrix**

The Confusion Matrix is a deep learning visual assessment method. The prediction class results are represented in the columns of a Confusion Matrix, whereas the real class results are represented in the rows [54]. This matrix includes all the raw data regarding a classification model's assumptions on a specified data collection. To determine how accurate a model is. It's a square matrix with the rows representing the instances' real class and the columns representing their expected class. The confusion matrix is a 2 x 2 matrix that reports the number of true positives (T P), true negatives (T N), false positives (FP),...
and false negatives (F N) when dealing with a binary classification mission.

The region under the ROC curve, or literally AUC, is the area under the curve (AUC) and is a measure of equal importance. A lower value implies that accuracy is more important, whereas a higher value indicates that recall is more important.

**B. ROC curve**

The region under the ROC curve, or literally AUC, summarizes the relationship between a binary classifier's true and false positive rate for various judgment thresholds. Several authors have shown that (AUC) is superior to absolute accuracy for classifier assessment, rendering it one of the most common metrics for static imbalanced data. To measure AUC, however, one must sort a specified dataset and iterate through each example [57]. This ensures that AUC cannot be computed directly on vast data streams since it will necessitate scanning the whole stream after each example. As a result, the usage of AUC for data sources has been restricted to estimations on periodic holdout sets or whole streams, rendering it inherently skewed or computationally infeasible for realistic implementations [58].

**VIII. PERFORMANCE EVALUATION AND RESULTS**

The confusion matrix was used to evaluate the accuracy of each classifier. The experimental results show that using five attributes from an SVM classifier produces the best prediction ratio of 95.56 percent and, CNN accuracy ratio is 92.11 percent. While KNN has the lowest estimation percentage which is 88.40 percent, as shown in Table 2, 3, and 4. Table (2) shows and analysis the results for using SVM algorithm, Table (3) shows and analysis the results for using CNN algorithm.

<table>
<thead>
<tr>
<th>Class</th>
<th>using SVM classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TP Rate</td>
</tr>
<tr>
<td>1</td>
<td>0.986</td>
</tr>
<tr>
<td>2</td>
<td>0.882</td>
</tr>
<tr>
<td>3</td>
<td>0.667</td>
</tr>
<tr>
<td>4</td>
<td>0.857</td>
</tr>
<tr>
<td>5</td>
<td>1.000</td>
</tr>
<tr>
<td>Avg</td>
<td>0.956</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Class</th>
<th>using K-Nearest Neighbor classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TP Rate</td>
</tr>
<tr>
<td>1</td>
<td>0.964</td>
</tr>
<tr>
<td>2</td>
<td>0.706</td>
</tr>
<tr>
<td>3</td>
<td>0.333</td>
</tr>
<tr>
<td>4</td>
<td>0.762</td>
</tr>
<tr>
<td>5</td>
<td>0.900</td>
</tr>
<tr>
<td>Avg</td>
<td>0.897</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Class</th>
<th>USING CONVOLUTIONAL NEURAL NETWORK</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TP Rate</td>
</tr>
<tr>
<td>1</td>
<td>0.906</td>
</tr>
<tr>
<td>2</td>
<td>0.882</td>
</tr>
<tr>
<td>3</td>
<td>1.000</td>
</tr>
<tr>
<td>4</td>
<td>0.952</td>
</tr>
<tr>
<td>5</td>
<td>1.000</td>
</tr>
<tr>
<td>Avg</td>
<td>0.921</td>
</tr>
</tbody>
</table>

Table (4) show the comparison between the three classifiers depending on the time taken to build the model and the accuracy of the classifier.

**TABLE 4. COMPARISON OF RESULT BY TIME AND ACCURACY**

<table>
<thead>
<tr>
<th>CLASSIFIER</th>
<th>COMPARISON OF RESULTS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TIME TAKEN TO BUILD MODEL</td>
</tr>
<tr>
<td>SVM</td>
<td>1.77 SEC.</td>
</tr>
<tr>
<td>KNN</td>
<td>0.01 SEC.</td>
</tr>
<tr>
<td>CNN</td>
<td>3.79 SEC.</td>
</tr>
</tbody>
</table>

Figure (3) shows that CNN algorithm takes the longest time to build its model while KNN algorithm had the shortest time.
Figure (4) shows that SVM algorithm had the higher accuracy. While KNN’s accuracy has the minimum value.

IX. COMPARATIVE STUDIES

As shown in the table (5), the researchers used several different methods, different dataset and different ways of feature selection/feature extraction. In comparison with related work, we obtained a good result in this work with dataset and methods that we used. However, researcher in [13] obtained 94% CT scan images dataset and SURF (Speeded Up Robust Features) for feature selection. In addition, the researchers in [14],[15] used the same dataset, the researchers in [14] obtained 90.9% with used Delta Radiomics method for feature extraction. But researchers in [15] could obtained better accuracy by using GLCM function for feature extraction and (MLP 98%, SVM 70.45%, & KNN 99.2%) classifier. By using (UCI) dataset the researchers in [6] could gained a good result 90% and used several classifiers.

Each of researchers in [16][17] they used same dataset (LIDC-IDRI) the researchers in[16] could obtained 86% sensitivity while, researchers in [17] depending on 2 features and applied different classifiers they obtained a good result (Random Forest 70%, SVM 80% and, ANN 96%). Also, the researchers in [18],[19] by using the same dataset and number of feature selection which is(23), but the researchers in [18] obtained higher accuracy 100% by using KNN classifier and genetic algorithm for feature selection. While the researchers in [19] obtained (Decision Tree 97%, K-NN 94% and, Neural Networks 96%). researchers in [20] used (JSRT) dataset and several classifiers to gain experimental results (ANN 82.43% and, Decision Tree 93.24%). Finally, the researchers in [21] used (LDCT) dataset and smart genetic algorithm with applying (SVM) to obtained 96.25% accuracy.
Lung cancer is one of the most dangerous diseases and the most common cause of death, the severity of the disease lies in the difficulty of diagnosing it in the early stages. This paper tries to endeavor to investigate of three classifiers to find the best classifier could classify lung cancer in early stage. The informational indices included in this study were derived from UCI databases for lung cancer patients. The focus of this paper is on using WEKA Tool to investigate the accuracy of classification algorithms. The results show that the Support Vector Machine (SVM) give the best accuracy 95.56%, that can detect lung cancer in its early stages and save several lives and, K-Nearest Neighbor KNN It gave less accuracy 88.40%.

### REFERENCES


149