International Journal of Informatics and Communication Technology (IJ-ICT)
Vol.10, No.2, August 2021, pp. 93~103
ISSN: 2252-8776, DOI: 10.11591/ijict.v10i2.pp93-103

Data mining techniques for lung and breast cancer diagnosis: A review

Bakhan Tofiq Ahmed
Department of Information Technology, Technical College of Informatics, Sulaimani Polytechnic University, Sulaimani, Kurdistan Region, Iraq

Article Info

Article history:

Received Jun 12, 2020
Revised Sep 16, 2020
Accepted Dec 29, 2020

Keywords:

ANN
Decision tree
KDD
Naïve Bayes
Support vector machine

ABSTRACT

Today, cancer is counted as a riskier disease than other diseases in the globe. There are many cancer forms like leukemia, skin cancer, stomach cancer, etc, but Lung and Breast cancer are the most common forms that many people suffered from. Cancer is a disease in that cells have grown rapidly and abnormally that is why treating them is somehow tough in some cases but it can be controlled if they detect in their initial stage. Data-mining classification algorithms had a vital role in predicting and recognizing both benign and malignant cells. Several classifiers are available to classify the usual and unusual cells such as decision-tree, artificial-neural net, support vector machine, k-nearest neighbor, etc. This paper presents a systematic review of the most well-known data-mining classification algorithms for lung and breast cancer diagnose. A brief review of KDD and the data-mining concept has been demonstrated. The D-Tree, A-NN, SVM, and Naïve Bayes classifiers that are widely utilized in the biomedical field have been reviewed along with some algorithms such as C4.5, Cart, and iterative-dichotomiser 3 'ID3'. A comparison has been done among various reviewed papers in terms of accuracy that used various data-mining classification algorithms to propose the lung and breast cancer diagnosis system. The experimental results of the reviewed papers showed that the Multilayer Perceptron 'MLP' and Logistic Regression 'LR' gave a higher accuracy of 99.04% and 98.1%, respectively.

Corresponding Author:

Bakhan Tofiq Ahmed,
Department of Information Technology, Technical College of Informatics
Sulaimani Polytechnic University
Sulaimani, Kurdistan Region, Iraq
Email: bakhan.tofiq.a@spu.edu.iq

1. INTRODUCTION

Nowadays, the most hazardous disease that faced humanity's life and led to fatal death is referred to as cancer [1]. Cancer is known as the abnormality of enhancing human cells and converting them into a tumor [2]. Among the various forms of cancer, lung and breast cancer are enumerated as the riskiest when compared to the other types around the globe. Breast cancer is the most common type of cancer in women with denser breast tissue due to its physiological features. It is one of the most significant reasons for women's death. According to the doctor's view, the most essential factor that causes lung-cancer is tobacco. Among genders, the male is more faced with lung-cancer than female due to the higher ratio of smoking [3]. For instances, according to a report which showed that among (116,470) males and (109,690) females, (87,750) males and (72,590) females were died because of lung-cancer. In the world, merely one from five deaths is due to smoking and utilizing tobacco [4]. Indeed, there are two major types of lung-cancer [5], [6]:

the first one is named non-small-cell cancer of lung (NSCCL) this type further categorized into squamous-cell carcinoma which was constituted approximately 25% to 30% of all lung-cancer, and adenocarcinoma which comprised of nearly 40% of lung cancer. This is counted as the most common lung-cancer type that has been seen in people who are not smoking especially women and youth people like teenagers and children, and large-cell carcinoma which constituted about 10-15% of non-small-cell cancer of lung.

However, the second type is named small-cell cancer of lung (SCCL). Merely 10% to 15% were occupied up by this form. Smokers were suffered most from this kind of lung-cancer [7]. The process of diagnosing lung and breast cancer is the main problem and challenging task due to its high cost. There were numerous methods to diagnose lung and breast cancer such as chest-radiography 'X-Ray', CT-Scan, and magnetic-resonance-imaging 'MRI-Scan'[8]. Conversely, most of these methods were costly and time spending. Moreover, these methods were identifying the lung and breast cancerous nodules in their advanced periods, so the survival rate will be quite low. Consequently, a new machinery tool was quite necessary to diagnose the lung and breast cancerous nodules in their initial stages [9].

The most critical approach for predicting and detecting cancer cells is to utilize the mechanisms and classification techniques of data-mining which is mostly used today in the world by doctors and researchers. This topic attracted several researchers to propose and enhance an intellectual system for estimation, recognition, and classification of non-malignant and malignant cells based on the data-mining mechanisms and algorithms. Every day a huge amount of data about patients are stored in the hospital databases, so extracting potential patterns or knowledge from these massive raw data is extremely significant this can be done with the help of new science that is referred to as data-mining [10]. A data-mining or 'D-M' is a numerical approach that is counted as a user-friendlier in reports presentation and error reduction. In several sectors tools of data-mining are utilized in taking sensitive decisions like healthcare, marketing, banking, etc. Prediction, detection, and classification are based on training some known variables to estimate the unknown variable. The widely utilized mechanisms of data-mining in almost all the sectors are enumerated as decision tree or d-tree, artificial-neural network, or net shortened into A-NN, Naïve Bayes or 'N-B', support-vector-machine or 'SVM', rule-based classifier, k-nearest neighbor or 'K-NN', etc. Data-mining is a critical stage of knowledge discovery in databases 'KDD'. KDD is comprised of several stages such as cleaning of data, integrating of data, selecting relevant data, pattern evaluation, and knowledge recognition. KDD and data-mining were utilized alternately [11].

The main objective of this study is to give a systematic review of the up-to-date research papers that proposed a system to diagnose lung and breast cancer by using the most essential techniques and classifier algorithms of data-mining. This review article is important for new researchers who interest to do research or review articles in the medical diagnosis era.

The arrangement of the other sections is as follows: in section 2 the other researcher's work has been reviewed and presented. The concept of data mining along with knowledge discovery in databases has been discussed in section 3. Section 4 is about the most renowned classification algorithms in data mining. A comparison between various classifiers and algorithms in terms of accuracy are drawn in section 5. Ultimately, section 6 is the paper's conclusion.

2. LITERATURE REVIEW

Recently, many researchers proposed an efficient lung and breast cancer diagnosis system to predict and detect cancerous nodules in their initial stages and assist the doctors to identify cancerous nodules easily in patient's lung and breasts. In this section, relevant up-to-date research papers have been reviewed.

Zubi and Saad, [12] in this paper several critical processes of medical image mining were examined such as pre-processing of data, extracting of features, or characteristics, and rule generating with the most important classifier mechanism which named neural network. The main aim of this study was to classify the chest x-ray and categorized into two classes: usual and unusual. As a consequence, if the cell was normal, then the patient was healthy. On the other hand, the abnormal cell indicated that the patient was unhealthy and had a form of lung cancer. This categorization helped doctors in deciding a significant decision about the patient's health and also increase the rate of survival. Furthermore, the investigation has been done about the use of association rules in the issue of chest x-ray labeling. The chest x-ray had been obtained and collected in massive multimedia databases which were storied for the medical intend.

Kharya, [13] different classifier mechanisms have been proposed to identify and estimate breast-cancer. These classifiers had been used ANN, C4.5, Naïve-Bayes, and D-Tree. The outcome accuracy is as followed 86.5%, 86.7%, 84.5%, and 93.62%, respectively. Breast-Cancer identification is about differentiating benign from malignant breast tumors, while the breast-cancer estimation expects when breast-cancer is reappeared in patients that had cancer tumor previously. Several different surveys and practical

articles on breast-cancer diagnosis and prognosis have been reviewed. This paper concentrated on up-to-date researches that utilized the data-mining mechanisms to improve the breast-cancer finding and guessing.

Chandra *et al*., [14] an effective associative rule that used the genetic algorithm for predicting three different diseases which were diabetes, breast-cancer, and heart have been modeled. The main inspiration for using the genetic algorithm [15], [16] was to discover predictive rules that were highly understandable and given high predictive accuracy. The outcomes indicated that most of the classifier rules predicted heart disease with the highest accuracy which was 98% when compared with diabetes and breast-cancer accuracy which were 82% and 84.8%, respectively.

Sowmiya *et al*., [17] numerous sides of data-mining methods have been revised to estimate lung-cancer tumors. The concepts of data-mining are extremely helpful for classifying cancerous and non-cancerous tumors. The ant-colony optimization method had been tested because it was a quite useful technique for predicting disease. In this study, both data-mining and ant-colony optimization methods have been mixed for generating proper rule and classification, which was experimental to create accurate lung-cancer classifications. Moreover, it affords the uncomplicated framework for additional enhancement in medical analysis on lung-cancer. The idea of the proposed system relied on reduced-order constrained optimization 'ROCO' which was an optimal data-mining method to make automatic and quick IMRT strategies for progressive lung-cancer identification. The 'ROCO' efficiently worked with the cure planning system which was utilized at Sloan-Kettering Center for detecting lung-cancer.

Panpaliya *et al.*, [18] proposed an online classification system for both prediction and detection of lung malignancy. Two main concepts have been combined which are image processing and data mining techniques. Histogram equalization is utilized under the concept of image processing for feature extraction along with an artificial neural net classifier to check the patient's state whether the cell is cancerous or non-cancerous. Detection and prediction of a cancerous cell in its initial stages will help both doctors and patients in deciding treatment and diminishing the disease risk. The main advantages of this proposed system are cost impressive and time-saving.

Paul *et al*., [19] this research, applied the transfer learning concept 'TLC' to learn knowledge and a pre-trained convolutional neural net 'CNN' to extract deep features from lung cancer computed tomography 'CT' images and later the classifier has been trained to predict the survivors. The outcomes demonstrated that the accuracy of 77.5% has been gained when a pre-trained CNN is utilized with 10 features. After that, higher accuracy of 82.5% has been obtained when upper 5 features from both a pre-trained CNN and the KNN 'K- nearest neighbor were merged.

Lynch *et al*., [20] unsupervised machine learning and clustering had been applied to the huge patient cases who had lung-cancer previously. The main aim of this work was to predict the rate of survival of patients. Merely (10,442) cases have been collected for this experimental study from the largest USA data repository named Surveillance Epidemiology End Results 'SEER' database. The 'SEER' is a confident cancer repository in the USA. Another objective was to automatically classify lung-cancer tumors into collections according to a clinical measure to predicate survival rate. Primaries No., Age, Cancer's Rank, Size of Lump, and Stage were variables that were selected as the machine learning inputs. The outcome was used to estimate the survival period. After that, a linear regression was carried out against each unsupervised outcome. The outcomes have been compared by using a root-mean-squared-error 'RMSE' which was an evaluation metric. The outcomes demonstrated that self ordering maps 'SOM' gave the best performance. However, K-Means counted as simpler classifier methods.

Omar *et al*., [21] proposed a new system to detect and predict a particular cancer form which was lung-cancer. The system is called lung-cancer prognosis system 'L-CPS'. Three significant methods had been applied which were J4.8, Naïve-Bayes, and K-nearest neighbor. The accuracy gained from these classifiers was 89.3%, 80.2%, and 89.6%, respectively. The main objective of proposing this intelligent system was to help physicians with an accurate assessment of the patient's health status. In this system, both data-mining and artificial-intelligence were assorted to create a user-friendlier tool for supplying information about a patient's status which was leading to enhance patient cures. A diversity of Lung-Cancer datasets could be admitted into 'L-CPS' and afforded physicians with numerous statistical outcomes containing estimations about their patients' status. Additionally, the records of the patient were easily managed by 'L-CPS', permitted them to view their summaries, and comprising prediction or any other explanations. Finally, while the 'L-CPS was presently restricted to a particular cancer form, but it could be measured as a prototype that could be improved to predict other diseases in the future.

Bharati *et al.*, [22] in this research study a breast cancer diagnosis system has been proposed by using classifiers like Naïve Bayes (NB), random forest (RF), logistic regression (LR), multilayer perceptron (MLP), and k-nearest neighbors (KNN). The classifiers algorithms have been implemented through the WEKA data mining tool. The image datasets have been collected from the UCI machine learning repository. The UCI consists of 286 instances. The experimental results showed that KNN was accurate than the other

classifiers because it correctly classified 207 instances out of 286 instances that was 72.3776 % followed by Naïve Bayes, random forest, logistic regression, and multilayer perceptron that attained 71.6783 %, 69.5804 %, 68.8811 %, and 64.6853 %, respectively.

Podder *et al.*, [23] in this paper a lung cancer diagnosis system was proposed to predict and detect cancerous nodules. The proposed system is based on the different classifiers like Naïve Bayes (NB), k nearest neighbors (KNN), logistic regression (LR), tree, random forest (RF), and neural network (NN). The performance of the classifiers has been tested by using the orange data mining tool. The experimental outcomes indicated that Naïve Bayes (NB) provided the highest accuracy of 57.047% followed by random forest (RF) that obtained 49.832% of accuracy. In addition, this study implemented principal component analysis (PCA), classification tree, multidimensional scaling (MDS), and hierarchical clustering (HC).

Abien, [24] in this study a comparison of six machine learning algorithms have been presented namely, gated recurrent unit support vector machine 'GRU-SVM', linear regression 'LR', multilayer perceptron 'MLP', k-nearest neighbor 'K-NN', soft max regression 'SR', and support vector machine 'SVM'. These algorithms are applied to a public dataset named the Wisconsin diagnostic breast cancer 'WDBC'. For the implementation of these algorithms, the dataset was split into two phases: 70% for the training phase, and 30% for the testing phase. Experimental outcomes showed that all the presented algorithms performed very well that were exceeded nearly 90% accuracy, but the 'MLP' algorithm outperforms all the others with an accuracy of 99.04%.

Mondal *et al.*, [25] in this study a systematic survey about the artificial neural network (ANN) based models for the diagnosis of breast cancer via mammography has been provided. The main benefits and weakness of various ANN models like spiking neural network (SNN), deep belief network (DBN), convolutional neural network (CNN), multilayer neural network (MLNN), stacked auto encoders (SAE), and stacked denoising auto encoders (SDAE) have been reviewed and presented. This review study also presented a comprehensive literature review about the studies related to a breast cancer diagnosis that applied various deep learning models to several publicly datasets. The performance of the reviewed studies was compared according to various metrics like accuracy, precision, and recall. The comparison indicated that the best performance was achieved by Residual Neural Network (Res Net-50 and Res Net-101) models of the CNN algorithm.

Muhammet, [26] this paper proposed and tested various data mining techniques for the detection of breast cancer. The techniques were logistic regression 'LR', k-nearest neighbors 'K-NN', support vector machine 'SVM', Naïve Bayes 'NB', decision tree 'DT', random forest 'RF', and rotation forest 'RF'. These techniques were applied to public data about breast cancer tumors from Dr. William H. Walberg of the University of Wisconsin Hospital. The results gained with the 'LR' showed the highest accuracy of 98.1%.

Based on the literature review, it is found that classification algorithms in data mining have a significant role in proposing efficient lung and breast cancer diagnosis system to increase the survival rate in the world. Lung cancer forms are shown in Figure 1, but Figure 2 shows the description of breast cancer.

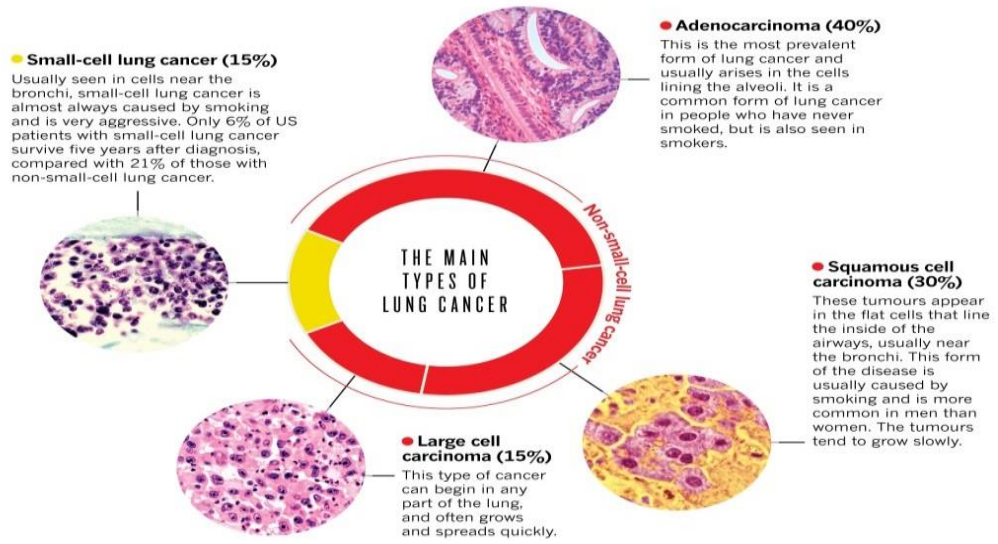

Figure 1. The forms of lung-cancer [27].

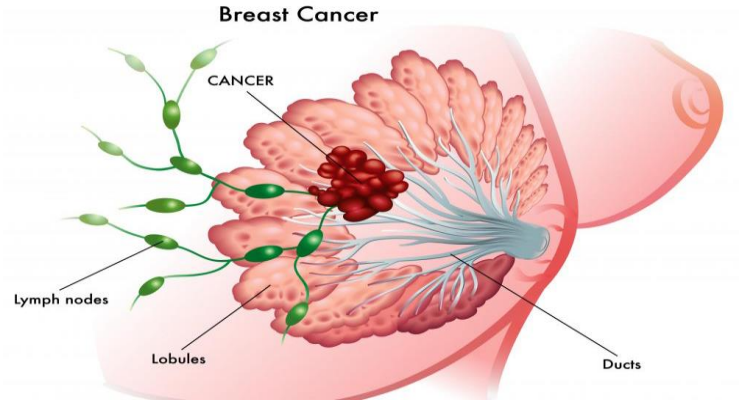

Figure 2. The description of breast cancer [28].

3. THE CONCEPT OF DATA MINING & KDD

In this section, a brief introduction about knowledge discovery and data-mining has been provided. The main steps and tasks of 'KDD' and data-mining also revealed shortly.

3.1. The method of knowledge discovery

Both 'KDD' and 'D-M' are utilized often alternately. The method of changing or converting raw data into knowledge is known as knowledge discovery process.

In addition, 'KDD' is the process of taking out the non-trivial implicit, previously unfamiliar, and critically beneficial knowledge from tremendous data accumulated in databases. In several cases, both 'KDD' and 'D-M' are described as an equivalent word, while in reality; they are different because 'D-M' is the significant stage in 'KDD' after the pre-processing stage. Generally, 'KDD' includes three phases namely, pre-processing, Data-Mining, and post-processing [7]. In contrast, 'KDD' in details comprises of these stages [26]:

3.1.1. Cleaning of data

It also is known as data cleansing, in this stage noise, unrelated, and outlier data should be removed from the pool.

3.1.2. Integrating of data

The raw data are collected from various sources; they may be heterogeneous so they should be unified together. This will be done at this step.

3.1.3. Selecting of data

In this phase, the relevant data should be regained from the data pool which is relevant for the analysis.

3.1.4. Transforming of data

Sometimes it is also called data consolidation because the designated data should be converted into an appropriate format to be useful for the procedure of mining.

3.1.5. Mining of data

It also is known as data-mining, it is enumerated as the potential stage in 'KDD' because intellectual systems are performed to take out useful knowledge.

3.1.6. Evaluation of pattern

This phase is about identifying the attentive patterns according to some measurements.

3.1.7. Representing the knowledge

It is the last stage in the 'KDD' process in which the gained patterns are shown visually to the end-user with the help of the visualization mechanisms such as boxplot and histogram to help the end-user to elucidate the results [29]. A major step of 'KDD' has shown in Figure.3.

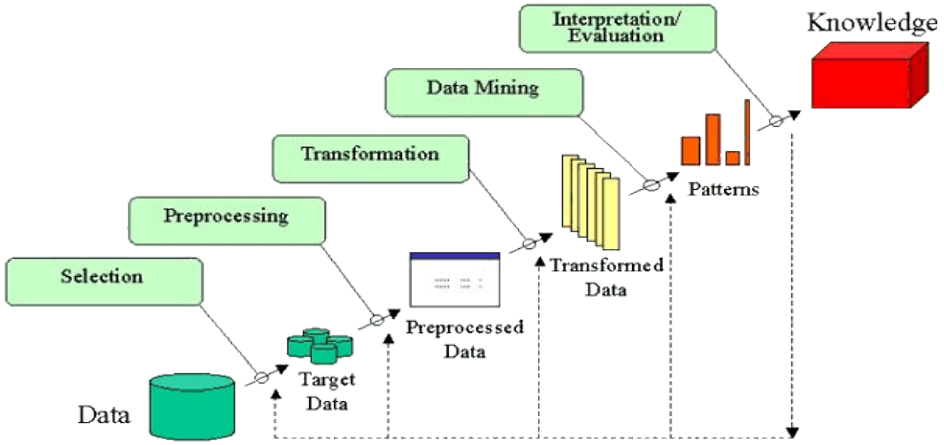

Figure 3. Knowledge discovery steps [3].

3.2. Data-mining concepts & methods

Data-mining is the most significant stage in the process of 'KDD' in which the crucial pattern is taken out from massive unprocessed data. The extracted knowledge relied on the tasks of 'D-M' that performed on the unprocessed data. In general, 'D-M' has two major tasks: descriptive tasks or duties in which the general characteristics of the existing data are depicted. However, the second duty is called the predictive task that endeavors to predict unknown data that relied on the known data. Unprocessed data may be in the format of numeric, text, image that mining can be done on them. Characterization, distinction, association or relationship, classification, clustering or grouping, and analysis of trend are the main 'D-M' functionalities. Nowadays, many sectors such as medical areas, education, and banking utilize 'D-M' tools [10]. Data-Mining like 'KDD' encompasses some stages which are briefly introduced below [30].

3.2.1. Defining the problem

This stage is very important because the main goal should be identified and revealed. According to the goal the corrected tool is performed on the unprocessed data to construct the desired model.

3.2.2. Exploring and collecting the data

This stage is about collecting and searching for appropriate data in which the quality of data must be suitable for the process of mining and analyzing to take out the best feature.

3.2.3. The preparation of data

Two main tasks can be done in this step, which is data cleansing and transforming to fill the missed and illegal values to obtain accurate, consistent, and robust results.

3.2.4. Constructing the nodel

This is the final stage in the 'D-M' process in which an appropriate model can be built for analysis according to the data and the desired results. The model can be constructed by combining both classical and modern 'D-M' techniques like statistics, D-Tree, A-NN, K-NN, and SVM, etc.

4. CLASSIFICATION ALGORITHMS IN DATA MINING

The classification algorithms in data mining have a vital role in extracting potential patterns or knowledge from raw data to propose an accurate and efficient diagnosis system to predict and detect Lung and Breast cancerous nodules in a low cost and minimum time. In this section, some of the renowned classification algorithms in data mining have been reviewed enlisted as the following:

4.1. Decision tree (D-Tree) classifier

It is shortened into D-Tree and counted as the modern effective algorithm for classification among the other classifier. It gained popularity when the data are grown specifically in the information system area. Many renowned algorithms belong to this classifier, for example, ID-3, C4.5, J4.8, C5, etc. From the name, this classifier is like a tree that splits the attributes recursively based on some mathematics algorithms like

entropy, gini-index, and chi-square to classify inconstant and matched threshold for the inconstant that divide the input attribute into two or more subdivisions. These subdividing will be repeated until the tree is built. The goal of this separation algorithm is to deduce a threshold that increases the homogeneity or uniformity between the subgroups [26]. Some of the well-known D-Tree algorithms have been discussed which were C4.5, Cart, and ID-3 [7]. An example of a D-Tree is depicted in Figure 4.

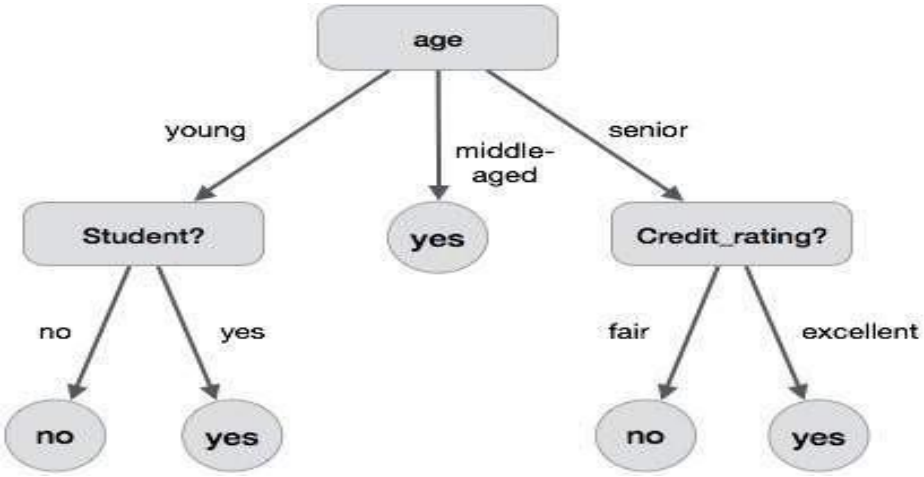

Figure 4. Decision tree (D-Tree) [31].

4.1.1. C4.5-algorithm

This algorithm is widely utilized. C4.5 selects only one feature of the data that has the highest gain ratio to efficiently separate the sample sets into subsections in one or more classes. In other words, the C4.5 algorithm depends on calculating the gain ratio of each node in the tree to do separation among the attributes. Consequently, which attribute obtained the higher gain ratio will be chosen by the algorithm to effectively split the data.

4.1.2. Cart-algorithm

It is enumerated as a non-parametric algorithm used to create either classification or regression (reversion) tree, based on whether the variable is categorical or numeric. A collection of rules are utilized to construct the tree that depends on the variables' values in the given data set. The selection of rules is relied on variables' values on how to separate the attributes well by calculating Gini-index. Each node is divided into two or more attributes whenever the rule is chosen; the process would be repeated for each node until the full tree is produced. Cart selects the attribute that has the smallest Gini-index as the splitting attribute. The process of separation is continued and stopped when the Cart-algorithm detected that no more division could be made on the node. The main idea behind this algorithm was to enhance the tree by selecting a separate among all of the other separates at each node to produce the purest node. In Cart-algorithm, only univariate distribution was measured.

4.1.3. Iterative dichotomiser (ID3)-algorithm

It is also known as ID3-Algorithm. In 1970, J. Quinlan established and introduced this algorithm. In this algorithm, the splitting attributes are selected that have the highest information gain as compared to the other attribute's information gain. A mathematical algorithm like entropy is used to measure the information gain in each node in the tree. The constructed tree is shorter and those attributes that have smaller entropies are usually placed close to the root in the tree.

4.2. Artificial neural net or network (A-NN) classifier

It is a calculated model that relied on biological neural nets. It involves an interrelated set of artificial neurons. In the A-NN classifier, the information is calculated by using a connectionist method. Neurons were systematized into layers. The input layer represents the real data but the output layer represents the classes. There are many hidden layers between the input and output layers. A neural net is an effective learning method in which data examples were depicted only one time to the net, and the class labeling could be predicted by adjusting the weights. A-NN has two benefits; the first one is that it gives the highest

tolerance to noisy data. The second is that the A-NN classifier can classify untrained knowledge [13]. An example of A-NN has shown in Figure 5.

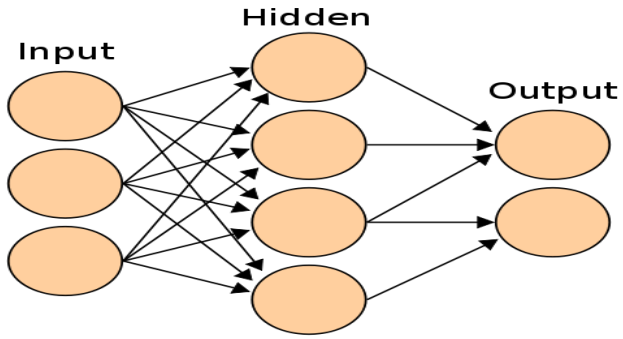

Figure 5. Artificial neural net (ANN) [13].

4.2.1. Feed-forward neural network

It is also referred to as FFNN. It is the simplest form of A-NN in which the information passes unidirectional between layers and does not create the cycle in the net. It contains three layers such as an input layer, a hidden layer, and an output layer. Units are created in each layer. When a value was moved from the input layer to the hidden layer and then to the output layer is referred to as Feed Forward because no values returned to the earlier layer. The inputs were fed concurrently into the units that were made up of the input-layer. The data is passed through the input layer and then weighted and fed concurrently to the upcoming layer which is called the hidden layer. The hidden layer outputs could be input to the next hidden layer. It is a recursive procedure. The output layer is constructed by the hidden layer outputs. It is categorized into two forms such as single-layer perceptron and multi-layer perceptron. The above diagram is a feed-forward which is comprised of three layers such as input, hidden, and output and no values return to the earlier layer [32].

4.3. Support-vector-machine classifier

It is a supervised machine learning technique to examine data and distinguish knowledge utilized for classification and reversion study. Support-vector machine is an algorithm that works on discovering a linear divider or 'hyperplane' between the data-points of 2 classes in multi-dimensional space. SVM classifier can construct a prediction model to estimate classes for new samples if the dataset has both features and class labels because SVM is supervised machine learning. As a result, it distributes new samples to one of the classes. Linear and non-linear SVM classifiers are the two types of support-vector machine classifiers. SVM could be trained by the uncomplicated and swift algorithm which is named sequential minimal optimization [31].

4.4. Naïve-Bayes classifier

It is a swift technique used for constructing statistical predictive models. Naïve-Bayes relies on the Bayesian theorem. This classifier examines the association between each feature and the class. For each sample extracted a conditional probability for the associations among the feature values and the class. The classes' probability is calculated by enumerating how many times it occurs in the exercising dataset when the classes were trained. This is named the prior-probability [29].

5. RESULTS AND DISCUSSION

Data mining techniques have a vital role in predicting and detecting several various diseases such as lung-cancer, breast-cancer, skin-cancer, heart, diabetes, etc. Many researchers proposed numerous useful data-mining methods to estimate Lung and Breast cancer. In my opinion, the techniques of data-mining are extremely useful in several areas specifically in the medical area to discover quite significant knowledge from raw data. In this survey, several researchers' works have been reviewed that used A-NN, Naïve-Bayes, D-Tree, etc, to propose a predicted and detected system for lung and breast cancer. These classifiers were different in terms of accuracy and the form of cancer. The details of these presented works have been demonstrated and compared in terms of accuracy as shown in Table 1.

Table 1. Comparison of various data mining classifiers and algorithms in terms of accuracy for detecting and predicting lung and breast cancer.

Ref.	Year	Diseases	Classifiers	Algorithms	Accuracy
[13]	2012	Breast-Cancer	ANN, Naïve Bayes, D-Tree.	C4.5	86.5%, 84.5%, 93.62%, 86.7%.
[14]	2013	Breast-Cancer	J4.8, Naïve Bayes	Genetic	74.2%, 71.67%, 84.8%.
[19]	2016	Lung-Cancer	Convolutional Neural Net, K-NN		77.5%, 82.5%.
[21]	2018	Lung-Cancer	J4.8, NB, K-NN		93%, 80.2%, 89.6%.
[22]	2018	Breast-Cancer	KNN, Naïve Bayes, Random Forest, Logistic Regression, Multilayer Perceptron		72.3776%, 71.6783 %, 69.5804 %, 68.8811 %, 64.6853 %.
[23]	2018	Lung-Cancer	Naïve Bayes, Random Forest, Neural Network, KNN, Logistic Regression, Tree.		57.047%, 49.832% 47.315% 46.98% 42.617% 35.57%.
[24]	2018	Breast-Cancer	GRU-SVM, LR, K-NN, SR, SVM, MLP.		90% 99.04%
[26]	2020	Breast-Cancer	RF, K-NN, SVM, NB, DT, RF, LR.		98.1%

6. CONCLUSION

The classification algorithms in data mining have a significant role in the medical sector in helping doctors and physicians to diagnose various forms of diseases effectively at low cost and short time. In this paper, a comprehensive review of the most critical classification algorithms in data mining had been presented. Many up-to-date relevant studies have been reviewed that utilized different classification algorithms in this era to propose an accurate lung and breast cancer diagnose system to predict and detect cancerous nodules in their initial stages. Based on the reviewed papers many different classification algorithms had been applied for classifying cancerous and non-cancerous cells. It is found that among the classifiers, MLP gained a higher accuracy than the others followed by LR and D-Tree, which were (99.04%), (98.1%), and (93.62%), respectively. However, among the algorithms, C4.5 obtained (86.7%) accuracy followed by the Genetic Algorithm that attained approximately (84.8%) accuracy.

ACKNOWLEDGEMENTS

I would like to thank the editor and reviewers for their comments and suggestions during the review process.

REFERENCES

[1] S. S. Priya, B. Ramamurthy, "Lung cancer detection using image processing techniques," *Research Journal of Pharmacy Technology* vol. 11, no. 5, pp. 2045-2049, 2018. DOI:10.5958/0974-360X.2018.00379.7, [Online]. Available:https://search.proquest.com/openview/476095aa679b205218d4b4fa59e7111f/1?pqorigsite=gscholar&cb=1096441.

[2] V. A. Gajdhane, L. Deshpande, "Detection of lung cancer stages on CT scan images by using various image processing techniques," *IOSR Journal of Computer Engineering (IOSR-JCE)*, vol. 16, no. 5, pp. 28-35, 2014. [Online]. Available: https://pdfs.semanticscholar.org/f346/a1acc850b018386ef50b1cd156e28e5dd36f.pdf.

[3] K. Balachandran, R. Anitha, "Classifiers based Approach for Pre-Diagnosis of Lung Cancer Disease," in *International Journal of Computer Applications (IJCA)(0975-8887) Proceedings on National Conference on Emerging Trends in Information & Communication Technology (NCETICT 2013): Citeseer*, pp. 1-5, 2013. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.663.3904&rep=rep1&type=pdf.

[4] B. T. Ahmed, "Lung Cancer Prediction and Detection Using Image Processing Mechanisms: An Overview," *Signal Image Processing Letters*, vol. 1, no. 3, pp. 20-31, 2019. http://dx.doi.org/10.31763/simple.v1i3.11.

[5] B. Gupta, S. Tiwari, "Lung cancer detection using curvelet transform and neural network," *International Journal of Computer Applications*, vol. 86, no. 1, pp. 15-17, 2014. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.429.1246&rep=rep1&type=pdf.

[6] M. P. Chander, M. V. Rao, T. Rajinikanth, "Detection of lung cancer using digital image processing techniques: a comparative study," *International Journal of Medical Imaging*, vol. 5, no. 4, pp. 58-62, 2017. doi: 10.11648/j.ijmi.20170505.12.

[7] D. S. Kumar, G. Sathyadevi, S. Sivanesh, "Decision support system for medical diagnosis using data mining," *International Journal of Computer Science*, vol. 8, no. 3, pp. 147-153, 2011. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.442.8665&rep=rep1&type=pdf#page=161.
[8] K. Dimililer, B. Ugur, Y. Ever, "Tumor detection on CT lung images using image enhancement," *The Online Journal of Science Technology*, vol. 7, no. 1, pp. 133-138, 2017. [Online]. Available: https://tojsat.net/journals/tojsat/volumes/tojsat-volume07-i01.pdf#page=142.
[9] Ada, R. Kaur, "Early Detection and Prediction of Lung Cancer Survival using Neural Network Classifier," *International Journal of Application or Innovation in Engineering & Management*, vol. 2, no. 6, pp. 375-383, 2013. [Online]. Available: https://www.semanticscholar.org/paper/Early-Detection-and-Prediction-of-Lung-Cancer-using-Ada-Kaur/a476334c9797261d71f89bc5c0c9ed6ce65d7418.
[10] D. Chauhan, V. Jaiswal, "An efficient data mining classification approach for detecting lung cancer disease," *International Conference on Communication and Electronics Systems (ICCES)*, Coimbatore, India, pp. 1-8, 2016. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7889872.
[11] D. Kaladhar, B. Chandana, and P. B. Kumar, "Predicting cancer survivability using Classification algorithms," *International Journal of Research Reviews in Computer Science*, vol. 2, no. 2, pp. 340-343, 2011. [Online]. Available: https://www.researchgate.net/profile/Bharath_Kumar16/publication/213588941.
[12] Z. S. Zubi, R. A. Saad, "Using some data mining techniques for early diagnosis of lung cancer," in *Proceedings of the 10th WSEAS International onference on Artificial intelligence Knowledge Engineering and Data Bases*, 2011, pp. 32-37.
[13] S. Kharya, "Using data mining techniques for diagnosis and prognosis of cancer disease," *International Journal of Computer Science, Engineering and Information Technology IJCSEIT*, vol. 2, no. 2, pp. 55-66, 2012. [Online]. Available: https://arxiv.org/abs/1205.1923.
[14] M. A. Jabbar, B. L. Deekshatulu, P. Chandra, "Heart disease prediction using lazy associative classification," *2013 International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s)*, Kottayam, India, pp. 40-46, 2013. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/6526381.
[15] B. T. Ahmed and O. Y. Abdulhameed, "Fingerprint recognition based on shark smell optimization and genetic algorithm," *International Journal of Advances in Intelligent Informatics*, vol. 6, no. 2, pp. 123-134, 2020. https://doi.org/10.26555/ijain.v6i2.502.
[16] B. T. Ahmed and O. Y. Abdulhameed, "Fingerprint Authentication using Shark Smell Optimization Algorithm," *UHD Journal of Science Technology*, vol. 4, no. 2, pp. 28-39, 2020. https://doi.org/10.21928/uhdjst.v4n2y2020.pp28-39
[17] T. Sowmiya, M. Gopi, B. New, and R. Thomas, "Optimization of lung cancer using modern data mining techniques," *International Journal of Engineering Research*, vol. 3, no. 5, pp. 309-14, 2014.
[18] N. Panpaliya, N. Tadas, S. Bobade, R. Aglawe, and A. Gudadhe, "A survey on early detection and prediction of lung cancer," *International Journal of Computer Science Mobile Computing*, vol. 4, no. 1, pp. 175-184, 2015.
[19] R. Paul, S. H. Hawkins, L. O. Hall, D. B. Goldgof, and R. J. Gillies, "Combining deep neural network and traditional image features to improve survival prediction accuracy for lung cancer patients from diagnostic CT," in *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, IEEE, 2016, pp. 002570-002575.
[20] C. M. Lynch, V. H. van Berkel, and H. B. Frieboes, "Application of unsupervised analysis techniques to lung cancer patient data," *Plos One*, vol. 12, no. 9, p. e0184370, 2017. https://doi.org/10.1371/journal.pone.0184370
[21] Y. Omar, A. Tasleem, M. Pasquier, and A. Sagahyroon, "Lung Cancer Prognosis System using Data Mining Techniques," in *Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies*, vol. 5, 2018, pp. 361-368.
[22] S. Bharati, M. A. Rahman, and P. Podder, "Breast cancer prediction applying different classification algorithm with comparative analysis using WEKA," in *2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT)*, 2018, pp. 581-584, https://doi.org/10.1109/CEEICT.2018.8628084: IEEE.
[23] S. Bharati, P. Podder, R. Mondal, A. Mahmood, and M. Raihan-Al-Masud, "Comparative performance analysis of different classification algorithm for the purpose of prediction of lung cancer," in *International Conference on Intelligent Systems Design and Applications*, Springer, 2018, pp. 447-457, https://doi.org/10.1007/978-3-030-16660-1_44.
[24] A. F. M. Agarap, "On breast cancer detection: an application of machine learning algorithms on the wisconsin diagnostic dataset," in *Proceedings of the 2nd International Conference on Machine Learning and Soft Computing*, 2018, pp. 5-9.
[25] S. Bharati, P. Podder, and M. Mondal, "Artificial Neural Network Based Breast Cancer Screening: A Comprehensive Review," *arXiv preprint arXiv:.01767*, pp. 1-13, 2020.
[26] M. F. Ak, "A Comparative Analysis of Breast Cancer Detection and Diagnosis Using Data Visualization and Machine Learning Applications," in *Healthcare*, Multidisciplinary Digital Publishing Institute, vol. 8, no. 2, pp. 111, 2020.
[27] *Lung Cancer Forms*, Springer Nature, Sep 2020, [Online] Available: https://www.google.com/search?q=lung+cancer+form&rlz=1C1GKLC_enIQ867IQ867&source=lnms&tbm=isch&sa=X&ved=2ahUKEwioo-3HpO3rAhUlShUIHYvIA7kQ_AUoAXoECAsQAw&biw=1536&bih=754#imgrc=9cDIVErdBWpzIM.

[28] *What are the Different Types of Breast Cancer?*, Info Bloom, Sep 2020, [Online] Available: https://www.google.com/search?q=breast+cancer+forms&source=lnms&tbm=isch&sa=X&ved=2ahUKEwj6tq6Rpe3rAhXgShUIHRoOA2oQ_AUoAXoECBQQAw&biw=1536&bih=754#imgrc=d0kMznfLkqOf7M.
[29] V. Krishnaiah, D. G. Narsimha, and D. N. S. Chandra, "Diagnosis of lung cancer prediction system using data mining classification techniques," *International Journal of Computer Science Information Technologies*, vol. 4, no. 1, pp. 39-45, 2013.
[30] K. Lakshmi, Y. Nagesh, and M. V. Krishna, "Performance comparison of three data mining techniques for predicting kidney dialysis survivability," *International Journal of Advances in Engineering Technology*, vol. 7, no. 1, pp. 242-254, 2014.
[31] G. R. Kumar, G. Ramachandra, and K. Nagamani, "An Efficient Prediction of Breast Cancer Data using Data Mining Techniques," *International Journal of Innovations in Engineering Technology*, vol. 2, no. 4, pp. 139-144, 2013.
[32] G. V. a. Dr.A.Suhasini, "Early Detection of Lung Cancer using Data Mining Techniques: A Survey," *International Journal of Engineering Research & Technology (IJERT) ICSEM-2013 Conference Proceedings*, vol. 1, no. 6, pp. 867-877, 2013.

BIOGRAPHIES OF AUTHORS

Bakhan Tofiq Ahmed received the B.Sc. in Software Engineering at Salahaddin University-Erbil- Kurdistan Region – Iraq, in 2010. She obtained the M.Sc. in IT at Sulaimani Polytechnic University- Sulaimani- Kurdistan Region- Iraq, in 2020. Her research interests are in information security, image processing, AI, swarm intelligence algorithm, and Biometric system design. She has 4 published papers, 3 papers published in an international journal, and 1 paper in a national journal.