A Comprehensive Survey on State of the Art Mechanisms and Data Mining Techniques for Accurate Prediction of Cancers with special focus on Lung Cancer

[1]Bhanumathi S, [2]Dr.S.N.Chandrashekara

[1]Research Scholar, [2]Professor & Head of the Department

[1]Department of CSE,SJCIT Chickballapur, Karnataka, India, [2]Department of CSE,CBIT Kolar, Karnataka, India.

Abstract

Now a day, throughout the world, lung cancer is the main reason of death. Data mining (DM) methods in the field of illness diagnosis is being approved in the health institutions. This opportunity has made way for several opportunities to conduct treatments for diseases. In order to derive some beneficial information from huge quantity of data, the Data mining is the most popular used method. Some of the data mining approaches like classification, relationship rule mining and clustering have been exercised to examine the types of disease. Classification is a crucial and an important process in Data mining. Presently, data mining has a significant function in the medical organizations in pursuance of predicting some severe diseases like cancer, etc. This paper aims to study on the recent data mining techniques of detecting and diagnosing lung diseases in the early stages so as to help doctors to save patient's life. This study briefly analyses the potential use of classification based data mining techniques, feature selection, and dimension reduction.

Keywords: **Data mining, cancer prediction, lung tumour, classification, survey.**

1. Introduction

In this article, we study about the cancer disease, its affect and recent techniques to predict the cancer in its early stage. Recently, the data mining techniques are widely adopted for early stage prediction where various machine learning approaches and tasks can be performed to detect disease. Thus, we study machine learning techniques and their application in medical applications.

1.1. Cancer Statistics

The cell of the organs preserves a rotation of renewal procedures. The stable development and demise rate of the cells generally preserve the ordinary functioning process of the physique; however it is not always possible. At times an unusual condition happens, when some cells might commence developing abnormally. This unusual progress of cells makes cancer and may begin from any organ of the physique and can be propagated to other sections of the physique. Several kinds of tumour may occur in the human physique for example Breast cancer, Prostate cancer, Melanoma, Lung cancer, and Leukaemia etc. Table 1 presents a brief discussion about different types of cancer, statics, risk factors, screening, stages and treatments.

Table 1 Types of cancer and brief discussion

Type of cancer	Statistics in US	Risk factors	Screening	Stages	Treatment
Breast Cancer	Estimated number of diagnosis 268,600. Estimated death 42,260 deaths	Age, personal history, family history, Early menstruation and late menopause, pregnancy timing, breast density, lifestyle factors, Race and ethnicity	Mammography,	Tumor, Node, Metastasis	Surgery (Removal of cancer in the breast), Lymph node evaluation, Radiation therapy
Prostate cancer	Estimated death-31,620	Age, Race/ethnicity, Family history, Eating habits, Hereditary breast and ovarian cancer	Digital rectal examination, PSA blood test	Very little danger, Short threat, Intermediary threat, Extreme threat	Surgery, radiation therapy
Melanoma	Estimated diagnosis count-96,480 oldsters (57,220 men and	Sun exposure, Inside tanning, Moles, Attractive skin, Family past, Family	Dermoscopy, epiluminescence microscopy	Stages-I to IV	Surgical procedure, Wide removal, Lymphatic

	39,260 women)	growth, Weakened or suppressed immune system			mapping and sentinel lymph node biopsy, Lymph node examination, Radiation treatment
Lung cancer	Estimated diagnosis count-228,150 adults (116,440 men and 111,710 women) Estimated death 42,670 (76,650 men and 66,020 women)	Asbestos, Radon, Genetics, tobacco and smoking	spiral computed tomography	Stages-I to IV	Surgical procedure, Radiation therapy, Chemotherapy, Targeted therapy, Immunotherapy

The worldwide cancer disease is assessed to have escalated to 18.2 million fresh instances and 9.7 million demises during 2018. During their lifetime, 1out of 5 men and 1 out of 6 women grow cancer universal, and 1 out of 8 men and 1out of 11 women expire due to the illness. Internationally, the five-year prevalence are those people who are living in 5 years of a cancer identification, is likely around 43.7 million [1].

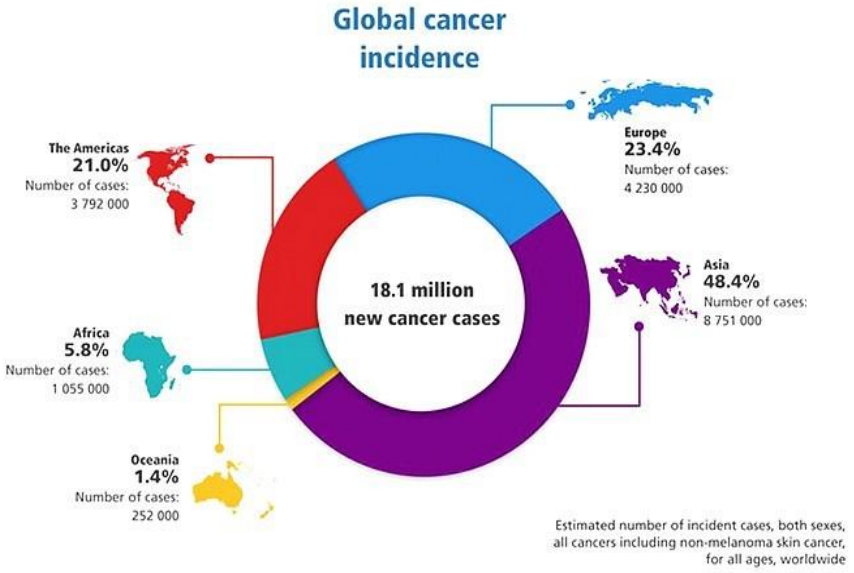

Figure 1 Global Cancer incidence

Universal samples display that for peoples, approximately half of the fresh instances of the cancer demises internationally during 2018 were assessed to ensue in Asia as depicted in figure 1, and in sections as the area has approximately 60% of the worldwide populace. Europe alone has 23.4% universal cancer instances and 20.3% of the cancer demises, though it is merely 9.0% of the worldwide populace. The Americas have 13.3% of the universal populace and estimates 21.0% of instances and 14.4% of death internationally. Contrary to the other areas of the world, the shares of cancer demises in Africa and in Asia (7.3% and, 57.3% correspondingly) are greater than the shares of occurrence instances (48.5% and 5.9%, correspondingly), since these areas have a superior occurrence of particular cancer forms related to inferior prediction and elevated death rates, including the inadequate access to well-timed identification and therapy in several nations.

1.2. Main categories of cancer

The study presented in [1] reported that internationally, most prominent types of cancers are lung and female breast cancers in context of the occurrence of fresh cases; around 2.1 million identifies were assessed in 2018 for these two types, adding approximately 11.5% of the overall cancer occurrence. Rectum cancer (1.79 million incidence, 10.1% of the overall) was the utmost prominent after lung and breast tumour, endocrine tumour is at the 4th position (1.3 million incidence, 7.2 %), and stomach cancer is at the 5th place (1.0 million incidence, 5.7%).

Similarly, Lung tumour is the main cause of the highest number of demises (1.8 million demises, 18.4% of the overall incidences), due to the inadequate diagnosis of this tumour internationally, succeeded by colorectal cancer (881000 demises, 9.2%), abdomen

tumour (783000 demises, 8.2%), and liver malignancy (782000 demises, 8.2%). Woman breast tumour is the 5th main reason of demises (627000 demises, 6.6%) since the diagnosis is comparatively encouraging especially in advanced nations.

1.3. Cancer detection methods

The proper diagnosis of these cancers requires early detection of cancer which can help clinicians to provide suitable treatment. Recently, Computer Aided Design (CAD) are widely adopted where different types of computer vision schemes are applied to detect the cancer. Generally, the cancer cells are classified as: (i) Benign which are counted as noncancerous, that is, non-dangerous. However sometimes it may also convert into a cancer condition. The "sac" is a resistant system which generally separates benign lumps from other types of cells and might be simply separated from the physique. (ii) Malignant are cancerous. It begins through an anomalous cell tumour and may quickly extend or attack neighbouring tissue. Commonly the centres of the malignant muscle are extra larger than in usual muscle and it may be a life-dangerous situation in upcoming phases [3].

These computer vision approaches are based on the machine learning concept where significant patterns are learned from the historical data of cancer images are new images are processed through the systems to predict the cancer type. Several machine learning methods are introduced such as deep learning for breast cancer [2], deep learning for lung cancer [4], residual networks for prostate cancer [5], basal cell carcinoma detection using neural network [6], deep learning for melanoma [7], and Lung cancer using convolutional neural networks [8] etc. however, these schemes require image data such as MRI, CT scan, ultrasound, and radiography etc.. The image acquisition is a tedious task and requires expensive setup. Due to these issues, recently, data mining based schemes are also adopted to develop a cost effective and reliable solution for cancer detection. Several techniques are presented based on data mining such as breast cancer detection [9], and lung cancer [10]. Hence, CAD and data mining procedures are employed for cancer detection but due to complexities of CAD, we focus on the data mining based approaches for cancer detection.

As discussed in section 1.3, the lung cancer is also considered as a chronic disease which leads toward the death of human. Hence, in this work we emphasis on the lung cancer detection using data mining approaches.

1.3.1. Lung Cancer

Lung cancer is an ailment of anomalous cells growing multiplying and developing into cancer. Tumour cells may be enraptured from the lungs in blood, or lymph liquid which encloses the tissues of lung. Metastasis happens after a cancer go away from the organ it has started and transfers into a lymph node or to another organ of the physique via the blood flow.

(a) Types of lung cancer

Key 3 categories of Lung Cancer are as follows:

- NSCLC (Non-Small Cell Lung Cancer): Bestowing to Society of Cancer, approximately 80% to 95% of Lung Cancers belong to this category. It may be further sub grouped into: (i) squamous cell carcinoma: approximately 20% to 30% of lung cancers belong to this group. These are normally associated with the habit of smouldering and often diagnosed in the centre of the lungs, nearby a bronchus. (ii)Adenocarcinoma: approximately 40% of lung cancers belong to this group. Non-smokers are generally affected by this type of lung cancer. Generally women and younger people are more affected than men from this type of cancer. The outer part of the lung is more affected by this. People belong to this category incline to have an improved prediction than peoples belong to other categories of lung cancer.
- Large cell (indistinguishable) carcinoma – 10% to 15% of NSCLC belong to this category. It can exist in any region of the lung. Normally, it grows and spreads rapidly and it is difficult to recover from it.
- SCLC (Small Cell Lung Cancer): It shares 10% to 15% of all types of lung cancers. It is very infrequent for somebody who has never smouldered. Generally, it begins in the bronchi adjacent to the middle of the chest, and it grows extensively over the physique.

2. Literature Review

In this unit, we present the literature review study of recent techniques of lung cancer detection, forecast and classification by data mining methods. According to pattern learning scheme, the data mining techniques are partitioned into two groups, as supervised learning (SL) and unsupervised learning (USL).

2.1. Supervised learning for lung cancer detection

On the basis of sequence-derived underlying and physicochemical characteristics of proteins, Hosseinzadehet al. [12] have proposed a diagnostic scheme which is implicated in both categories of cancers through feature mining, feature assortment and forecast prototypes. Twelve characteristic weighting prototypes have selected significant features and computed 1497 proteins attributes. Conclusively, machine learning (ML) prototypes includes 7 SVM prototypes, 3 ANN prototypes and 2 NB prototypes used in primary database and recently yield ones from attribute weighting prototypes. They have computed the precision of models via ten-fold cross and wrapper authentication (only for SVM methods).

Bartholomai et al. [13] have developed a regression prototype to forecast the endurance period of lung cancer patients in terms of months. Initially it was revealed that prognostic models perform precisely only for small endurance periods of fewer than six months. But the preciseness of the model was decreased after trying to forecast lengthier endurance periods. The authors have used a mechanism which combines the regression prototypes with a categorization prototype and forecast the endurance period of the patient. A collection of de-recognized lung malignancy patient information was attained from the SEER database. The prototypes have followed a subgroup of features chosen by ANOVA. Accuracy of the model was assessed by a confusion matrix of categorization and through the root mean square error (RMSE). For the purpose of classification, the random forests were employed whereas for regression common Linear Regression, Random Forests and GBM were used.

Chandraet al. [14] have analysed that extensively used ML based Naïve–Bayes Classifier can be used for classification issues due to its easiness and precision as equated to other SL techniques. But in case of big dimension data, such as gene expression, it is not able to perform efficiently because of following key constraints i.e. over fitting and underflow. In pursuance of resolving the issue of underflow, the prevailing method considered the addition of the logarithms of likelihoods instead of multiplying likelihoods and the estimated method was employed as a solution to over fitting issue. Though, generally, these techniques do not function efficiently for gene expression information. The suggested approach R-NBC has described an exclusive robust process to transform the approximation of likelihoods in NBC.

Investigations were performed by Varadharajan et al. [15], to analyse the classification processes for the lung tumour anticipation, for instance, back-propagation-NN and decision tree. Initially, they have collected twenty cancers and non-disease patients' instances information by thirty features, pre-arranged and divided employing categorization procedures and afterwards an equivalent procedure was realized for fifty incidences of tumour patients and ten non-growth patients. The statistical indexes employed as a portion of this investigation were borrowed from UCI database of patients disturbed by lung malignancy.

Dash et al. [16] have developed a fusion prototype through joining of evolutionary calculation, fuzzy logic and NN which has increased the classification precision and decision making quickness. In the suggested hybrid framework, the authors have merged GS method with PSO method to train Fuzzy MLP for classification of medicinal data.

According to Salaken et al. [17], the existing work normally manages this process via handcrafted characteristic formation and assortment. Recently, deep learning proved capable in identifying the fundamental structure of data via the utilization of auto-encoders and other methods. The authors have suggested a deep auto-encoder categorization process which initially acquires deep features and afterwards trains an ANN using these learned features.

Lian et al. [18] have efficiently resolved the issue of learning from inadequate and indeterminate data. The authors have recommended a variation of the EK-NN technique which is influenced by a mix Dempster and Yager instruction. It relocates portion of the incompatible mass towards the frame of discrimination. They have also introduced a feature selection process which discovers explanatory feature subgroups through minimization of an exceptional objective function by integrated integer genetic procedure. This function was intended to lessen the inaccuracy of the mass functions and to attain the feature subspaces which increase the division among classes. Lastly, they have suggested a two-phase categorization approach and they showed that this strategy enhances the categorization precision through previously classified objects as added fragments of proof.

Shunmugapriya et al. [19] have discovered that Feature Selection (FS) assist in speeding up the procedure of categorization through mining of the appropriate and valuable information from the database. The FS was considered as an optimization issue as selection of the suitable feature subgroup is quite significant. They have proposed a new Swarm related mixed technique named AC-ABC Hybrid. It unites the qualities of ACO and ABC procedures for optimization of feature selection. Through hybridization, they have tried to remove sluggishness behaviour of the ants and time overwhelming worldwide hunt for primary elucidations by the employed bees. According to recommended procedure, Ants have used utilization thru the Bees to decide the optimum Ant and feature subgroup; Bees modify the feature subgroup produced by the Ants.

Li et al. [20] have recommended a SML centred classification tactic that has several applications in computation biology. Here data samples are spontaneously classified into pre-specified markers with the help of data extraction techniques. Generally the training examples comprise very insufficient occurrences of concern (for example, medical incongruities, sporadic syndrome in a populace, and uncommon diseases) and numerous common occurrences. This unbalanced proportion of data allocations amid the target markers obstructs the effectiveness of categorization procedures, since the persuaded prototype was not trained by adequate extent of examples of the interesting markers, however astounded by conventional training documents. Through diverse swarm approaches (Bat-inspired procedure and PSO), they have proposed an optimization prototype for flexible equalizing to enhance or lessen of the class dissemination, rendering to the qualities of the biological databases. The optimization is protracted for attaining the maximum probable correctness and Kappa statistics simultaneously.

According to Venkataraman et al. [21], to use data mining techniques, the FS methods are vital to manage numerous dimensional datasets which can have features of small, intermediate, and big dimensions. Classifier precision and execution period are always affected by large number of available features. The authors have recommended a new fusion feature assortment scheme on the basis of SU and GA. The key results and impact of their research may be briefed as: SU-GA hybrid feature selector selects merely maximum applicable features for supervised learning. The recommended technique decreases processing period considerably than any other feature assortment techniques with nominal number of features.

ALzubi et al. [23] have combined Weight Optimized NN with WONN-MLB for the analysis of LCD in big data. The recommended technique involves two phases, feature assortment and ensemble categorization. During the first phase, the vital attributes are elected with a combined MLMR pre-computing prototype for reducing the classification period. During the second phase, Boosted Weighted Optimized NN Ensemble Categorization procedure was used to categorize the person with particular attributes which increases the tumour ailment analysis precision and reduce the false positive frequency.

Naseriparsa et al. [28] have proposed an arrangement of unsupervised dimensionality decrease with resampling to lessen the dimension of Lung Cancer databases and balance it through resampling. To reduce the feature space, PCA was used and hence it drops the complexity of categorization. PCA attempts to preserve the key features of primary database in the compressed database; but, particular valuable information is vanished during this process. The SMOTE resampling was employed to function on the sample field, the variety of sample domain was increased and the distribution of classes was balanced in the dataset.

In this subsection we have studied about several supervised schemes to detect the lung cancer. These techniques are based on the data mining. Based on these schemes, a Relative investigation is bestowed in table 2.

Table 2 Relative investigation of supervised schemes

Author	Technique Used	Contribution	Performance measurement	Remarks
Hosseinzadehet al. [12]	SVM, NN, and Naive bayes classifier	Feature extraction and selection	Accuracy and Kappa measurement	A comparative study using different classification techniques.
Bartholomai et al. [13]	Regression model	Survival time prediction	Confusion matrix, RMSE	Classification and linear regression model are used for survival time prediction
Chandraet al. [14]	Naïve–Bayes Classifier (NBC)	Logarithmic probability based NBC	Classification accuracy	underflow and over fitting issues are addressed
Varadharajan et al. [15]	Neural Network	Combined back propagation neural network and decision tree.	Prediction accuracy and classification error	The principle point of this paper is to give the prior notice to the clients and to quantify the execution investigation of the classification algorithms
Dash et al. [16]	Fuzzy	A Hybrid approach using Fuzzy MLP and PSO	Accuracy	Combined optimization can be used but it increases complexity
Salaken et al. [17]	Neural Network	Ensemble of deep learning and ANN	Accuracy	Deep feature learning and ensemble classification
Lian et al. [18]	A new model for classification	New EK-NN model using Dempster+Yager rule, feature selection, and classification model	Classification error, misclassification rate	Improved learning model for uncertain data.
Shunmugapriya et al. [19]	Ant and bee colony optimization	Hybrid model of ant and bee colony using Weka Tool.	Accuracy and execution time	Best feature selection is helpful in fast convergence
Li et al. [20]	Optimization	Bat inspired PSO for imbalanced data	Kappa, accuracy, precision, recall	decision tree can overcome the imbalanced problem, better than neural network
ALzubi et al. [23]	Neural network	Ensemble learning with optimal feature selection	FPR, Classification time, F1- score, complexity and feature selection rate	weak classifier with less error because of ensemble model

2.2. Unsupervised Learning

In order to evaluate survival rate, Lynch et al. [11] have spontaneously categorize lung cancer patients into sets according to the clinically assessable disease precise parameters. For machine learning, the selected input parameters are TNM, Number of Primaries, Grade, Age, Tumour Size and Stage. These parameters are either numeric or may easily be transformed to numeric type. With slight human intervention through the complete process, Nominal advanced handling of the data permits exploration of the creative proficiencies of recognized unsupervised learning methods. The results of the procedures were used to forecast survival period, through the effectiveness of the forecast demonstrating a proxy for the utility of the categorization. They have applied a fundamental single variable linear regression for all unsupervised outputs, and the related RMSE value was computed as a measurement to compare between the outcomes.

Yu et al. [22] have recommended a novel cluster cooperative method which is known as knowledge centred cluster ensemble (KCE). It includes the advanced information of the datasets into the cluster collaborative structure. Precisely, KCE denotes the advanced information of a dataset as the configuration of pairwise restrictions. Next, the authors have adopted the SC technique to

produce a group of clustering solutions. Then, for these clustering solutions, KCE has transformed pairwise restraints into confidence elements. Next, a consensus matrix was generated through consideration of all the clustering answers and their equivalent confidence elements. The concluding clustering output was attained through partition of the consensus matrix.

Yu et al. [25] have analysed that most of the prevailing investigation works have considered the single clustering procedures to conduct the tumour clustering from non-robust, less stable and less accurate bio-molecular data. Therefore to enhance the behaviour of tumour clustering through bio-molecular data, they have introduced the fuzzy model into the cluster collaborative structure for tumour clustering from bio-molecular data, and proposed 4 categories of mix fuzzy cluster collaborative structures which recognize samples of distinctive categories of cancers.

He et al. [26] have recommended a two-stage genetic clustering (TGCA) procedure. It is able to spontaneously decide the exact number of clusters and the appropriate separation from a provided dataset. The two-stage assortment and mutation processes were applied to use the exploration competence of the method through change in the likelihoods of assortment and mutation rendering to the consistence of the number of clusters within the populace. Initially, the method emphasizes on the exploration of the best number of clusters, and next progressively moves to discover the universally optimum cluster centres. Additionally, a maximum attribute range separation technique was employed in the initialization of the population so that the sensitivity of clustering algorithms to primary partitions can be resolved. Lastly, the competence of TGCA was widely compared with numerous spontaneous clustering procedures such as hierarchical agglomerative k-means, spontaneous spectral procedure and the SGKC.

Khanmohammadi et al. [27] have analysed that most of the actual-world medicinal data sets have inherent intersecting information and it can be efficiently described by overlapping clustering approaches that permit one sample linked to more than one cluster. The overlapping k-means (OKM) is a simple and competent overlapping clustering technique which is an expansion of the conventional k-means procedure. The OKM technique faces the sensitivity and primary cluster centroids issues. The writers have recommended a mix technique which joins k-harmonic means and overlapping k-means procedures (KHM-OKM) to solve these issues. The key concept for KHM-OKM process is to employ the outcome of KHM process to set the cluster centres of OKM process.

Similar to previous section, here we present a comparative analysis based on unsupervised scheme as given in table 3.

Table 3 Comparative analysis using unsupervised classification

Author	Technique Used	Contribution	Performance measurement	Remarks
Lynch et al. [11]	Unsupervised learning	Patient survival time prediction	RMSE	Logistic regression and unsupervised schemes are used for survival time prediction
Yu et al. [22]	knowledge based cluster ensemble (KCE)	Knowledge and spectral clustering are combined	Accuracy	Prior knowledge of data can help to improve the performance
He et al. [26]	Clustering	Two stage Genetic algorithm for selection and mutation operation	Clustering accuracy	A new clustering is developed which helps to reduce the clustering error.
Khanmohammadi et al. [27]	K-Means clustering	k-harmonic means and overlapping k-means algorithms	Precision, recall	Hybrid k-means algorithm is developed to mitigate the overlapping issue

2.3. Data Mining Techniques for other Medical Applications

Previous section describes the use of supervised and unsupervised data mining schemes for lung cancer detection. Various other diseases are present which can be predicted using data mining techniques. These techniques are stimulated on 5 actual standard categorization problems (UCI Machine Learning Repository). The explanation of the data sets has been provided by table 4.

Table 4 Dataset description

Dataset	Total Instances	Total attributes	Total class
Wisconsin Breast Cancer	699	9	2
Pima Indians Diabetes	768	8	2
Heart-Statlog	270	13	2
Hepatitis	155	19	2
Cleveland Heart Disease	296	13	2

(a) Breast cancer prediction using data mining

Abdel-Zaher et al. [29] have anticipated a CAD procedure to recognize breast malignance through deep belief network (DBN-NN) unsupervised route tracked by back propagation supervised route. The framework is back-propagation NN with Liebenberg Marquardt learning process whereas weights are set from the DBN-NN path.

Aličković et al. [30] have proposed two phase system. In the 1st phase, in pursuance of eliminating the unimportant descriptions, GA was employed for mining of explanatory and substantial descriptions. It reduces the computing difficulty and increase the speediness of the data mining mechanism. In the next phase, numerous data mining methods were used to create a conclusion for two distinctive groups of subjects (breast cancer and without cancer). Distinctive specific and numerous classifier schemes were employed in the second phase so that an accurate system can be built for breast cancer categorization.

Aalaei et al. [31] have proposed a feature selection technique by GA for choosing the finest subgroup of features for breast cancer identification scheme. PS-classifier, GA-classifier and ANN were employed to assess proposed feature selection process on Wisconsin data sets. The classification by PS-classifier was better than other classification.

To precisely assisting the physicians, Alwidian et al. [32] have proposed a competent classifier which predicts the chronic ailment. Many researchers have used Association Classification (AC) techniques to solve this issue by using association formulas. Though, maximum AC techniques face issues related to the expected measures employed in the rule assessment procedure and the prioritizing methods employed at the level of attributes. It can be useful for the rule generation method. The authors have tried to crack this issue via a competent weighted categorization which is based on association rules process WCBA. They have also presented a fresh pruning and forecast method according to statistical procedures to produce new precise association formulas to improve the precision level of the AC classifiers.

Kadam et al. [33] have suggested Sparse Auto-encoders and Softmax Regression based feature ensemble learning for categorization of Breast Cancer into malignant and benign form. They have used Wisconsin (Analytical) medicinal datasets from the UCI machine learning source. The recommended technique is evaluated through numerous enactment indices such as true classification precision, specificity, understanding, recollection, correctness, f measure, and MCC.

Table 5 presents the comparative analysis for breast cancer prediction based on data mining techniques.

Table 5 Comparative analysis of breast cancer prediction techniques

Author	Technique Used	Contribution	Performance measurement	Remarks
Abdel-Zaher et al. [29]	Deep belief network	Incorporating the back-propagation neural network with Liebenberg Marquardt learning	RMSE	DBN scheme is used as unsupervised learning
Aličković et al. [30]	Artificial neural Network	New genetic algorithm as low weighted gene genetic algorithm (LWGGA), high weighted gene genetic algorithm (HWGGA), and weighted gene genetic algorithm (WGGA)	Classification Accuracy	Reduced computational complexity.
Aalaei et al. [31]	Artificial neural Network	Genetic Feature Selection with ANN classifier	Accuracy, specificity and sensitivity	Efficient feature selection can reduce the complexity and improve the accuracy
Alwidian et al. [32]	using Association Classification	weighted classification based on association rules algorithm	Classification accuracy	a new pruning and prediction technique is presented to improve the accuracy
Kadam et al. [33]	Ensemble feature learning	Sparse Auto encoders and Softmax Regression	Accuracy, sensitivity	Ensemble feature learning

(b) Diabetes prediction using data mining

Sangle et al. [34] have developed a diabetes detection scheme using PCA and MLPANN. Initial examination emphases on joining source information and PCA altered descriptions in MLPANN structure. Confusion matrix centred investigation was conducted to study the consequence of source and PCA information fusion.

Wu et al. [35] have suggested a data mining based fresh model to forecast type 2 diabetes mellitus. The key issue was to enhance the precision of the forecast model, and to make a flexible model which is suitable for different types of datasets. On the basis of a sequence of pre-computed processes, the prototype is built of two sections, the enhanced K-means method and the logistic regression method.

Choubey et al. [36] have suggested a two phase method. During the initial phase, the Pima Indian diabetes data set was taken from the UCI source of ML data sets. In the next phase, the authors have conducted the categorization through fuzzy decision tree on the Pima data set. Next they have used PSO_SVM feature assortment method succeeded by the classification method of fuzzy decision tree. This optimization of SVM by PSO decreases the number of characteristics, and henceforth, using fuzzy decision tree enhances the precision of perceiving disease. The mix combinatorial technique of feature assortment and categorization required to be performed; consequently the modified system is employed for the categorization of diabetes.

Shankar et al. [37] concluded that although Diabetes is an inveterate and prolonged disease, but initial sugar checking and identification preclude from the damaging consequences. The authors have presented a diabetes forecasting model which is founded on the grey wolf optimization with fuzzy reasoning. They have used the idea of AI, where the prototype adopts the fuzzy rule and then perform optimization rendering to the GWO procedure

Choubey et al. [24] have provided a fine categorization of diabetes using Pima Indian Diabetes Dataset. The suggested model includes following stages: Initially a Localized Diabetes Data set was gathered from Bombay Medical Hall, Ranchi, India. Next, NN were employed for the categorization on localized diabetes data set. Next, genetic algorithm was employed as a feature assortment method and six features among twelve features were attained. Finally, NN were employed for categorization on appropriate attributes created by genetic algorithm.

Table 6 Comparative analysis for diabetes classification

Author	Technique Used	Contribution	Performance measurement	Remarks
Sangle et al. [34]	Neural network	Combined Principal Component Analysis (PCA) and Multilayer Perceptron Artificial Neural Network (MLPANN)	Confusion matrix, accuracy	Dimensionality reduction
Wu et al. [35]	Clustering and regression	Improved K-means and logistic regression methods are developed	Accuracy	Adaptive model is developed to work with different types of dataset
Choubey et al. [36]	Support vector machine	Combined particle swarm optimization and SVM is developed for feature selection and fuzzy decision tree is used for classification	Accuracy	A hybrid model of feature extraction and classification is developed.
Shankar et al. [37]	Fuzzy Logic	Fuzzy logic with grey wolf optimizer	Accuracy, precision and recall	Performance of Fuzzy based systems depends on the optimizer
Choubey et al. [24]	Neural Network	GA based feature selection	Classification accuracy, confusion matrix	Real time use of GA and SVM

(c) Heart disease prediction using data mining

Paul et al. [38] have used numerous methods such as attribute decline, rule mining, fuzzy optimization, etc. However redundant data in data sets, inappropriate attributes and absence of operational fuzzy formulas are key limitations while taking a decision. The authors have proposed GA focused fuzzy decision support system (FDSS) for forecasting the risk level of heart illness. The working of FDSS is as follows: a) Pre-process the data set, b) Operational attributes have been picked up using numerous techniques, c) Weighted fuzzy formulas are created using GA, d) Make the FDSS from the produced fuzzy knowledge base, e) Forecast the heart ailment.

Bashir et al. [39] have emphases on predicting and examination of heart illness utilizing weighted vote-based classifier collaborative method. The suggested model resolves the drawbacks of standard data mining methods by utilizing the combination of 5 different classifiers: decision tree based on Gini index, naive Bayes, instance-based learner, decision tree based on information gain, and SVM.

Akgül et al. [40] have used ANN with default variables to diagnose heart illness. They have suggested a combined method of ANN and GA to enhance classification precision. At the end, the usefulness of the recommended method was presented using Cleveland data set selected from UCI ML resource.

Jabbar et al. [41] have displayed a competent method for forecast of heart ailment. They have considered backward removal technique for FS by chi square formula for heart ailment categorization with improved classification precision.

Costa et al. [42] have used ML methods for the analysis of cardiovascular illnesses with the aim of suggesting a scheme for medical support. The ANN was employed to categorize patients according to the presence of a cardiovascular illness, and a DSS was proposed for the clinician descriptions.

Table 7 Comparative analysis of heart disease prediction

Author	Technique Used	Contribution	Performance measurement	Remarks
Paul et al. [38]	Fuzzy Logic	GA based fuzzy decision support system is presented	Accuracy sensitivity, specificity	Optimal attribute selection can reduce the number of tests.
Bashir et al. [39]	Ensemble Classifier	Ensemble classifier using naive Bayes, decision tree based on Gini index, decision tree based on information gain, instance-based learner, and support vector machines	Sensitivity, specificity and F -measure	Ensemble classification can help t predict other disease
Akgül et al. [40]	Neural Network	GA is used for optimizing the ANN parameters	accuracy, precision, recall and F-measure	GA increases the computational

				complexity
Jabbar et al. [41]	Random forest	Feature selection with Random forest	Accuracy, sensitivity, specificity, Disease Prevalence, Positive Predictive Value(PPV), Negative Predictive Value(NPV)	A new feature selection using chi square distance measurement.
Costa et al. [42]	Neural Network	A cloud based application is presented	Recognition rate, precision, recall, F1-score	Real-time cloud based cardiovascular disease monitoring.

In the aforementioned subsections we have studied about several data mining techniques to detect the lung cancer and various other diseases such as heat, diabetes, and breast cancer. These studies provide the significance of data mining for early prediction of various diseases.

3. Conclusion

Lung cancer is a painful and a deadliest disease in the world and understanding the usefulness of distinct data mining algorithms in detecting the disease will save millions of life. This paper discusses on various lung cancer detection modules that are based on data mining techniques. The paper mainly focuses on timely recognition methods of lung cancer based on survey by data mining methods. The survey aims in detecting and diagnosing the cancer disease and related areas. This paper compares various techniques of data mining based on efficiency in classification in order to detect lung cancer in various classes

References

1. The International Agency for Research on Cancer (IARC). (2018, September 12). Latest global cancer data: Cancer burden rises to 18.1 million new cases and 9.6 million cancer deaths in 2018 [Press release]. Retrieved July 31, 2019, from https://www.iarc.fr/wp-content/uploads/2018/09/pr263_E.pdf
2. Khan, S., Islam, N., Jan, Z., Din, I. U., & Rodrigues, J. J. C. (2019). A novel deep learning based framework for the detection and classification of breast cancer using transfer learning. Pattern Recognition Letters, 125, 1-6.
3. Nahid, A. A., & Kong, Y. (2017). Involvement of machine learning for breast cancer image classification: a survey. Computational and mathematical methods in medicine, 2017.
4. Bhatia, S., Sinha, Y., &Goel, L. (2019). Lung cancer detection: A deep learning approach. In Soft Computing for Problem Solving (pp. 699-705). Springer, Singapore.
5. Xu, H., Baxter, J. S., Akin, O., & Cantor-Rivera, D. (2019). Prostate cancer detection using residual networks. International journal of computer assisted radiology and surgery, 1-4.
6. Dua, R., Beetner, D. G., Stoecker, W. V., &Wunsch, D. C. (2004). Detection of basal cell carcinoma using electrical impedance and neural networks. IEEE Transactions on Biomedical Engineering, 51(1), 66-71.
7. Hagerty, J., Stanley, J., Almubarak, H., Lama, N., Kasmi, R., Guo, P., ...&Stoecker, W. V. (2019). Deep Learning and Handcrafted Method Fusion: Higher Diagnostic Accuracy for Melanoma Dermoscopy Images. IEEE journal of biomedical and health informatics.
8. Golan, R., Jacob, C., &Denzinger, J. (2016, July). Lung nodule detection in CT images using deep convolutional neural networks. In 2016 International Joint Conference on Neural Networks (IJCNN) (pp. 243-250). IEEE.
9. Liou, D. M., & Chang, W. P. (2015). Applying data mining for the analysis of breast cancer data. In Data Mining in Clinical Medicine (pp. 175-189). Humana Press, New York, NY.
10. ALzubi, J. A., Bharathikannan, B., Tanwar, S., Manikandan, R., Khanna, A., &Thaventhiran, C. (2019). Boosted neural network ensemble classification for lung cancer disease diagnosis. Applied Soft Computing. doi:10.1016/j.asoc.2019.04.031
11. Lynch, C. M., van Berkel, V. H., &Frieboes, H. B. (2017). Application of unsupervised analysis techniques to lung cancer patient data. PloS one, 12(9), e0184370.
12. Hosseinzadeh, F., KayvanJoo, A. H., Ebrahimi, M., &Goliaei, B. (2013). Prediction of lung tumor types based on protein attributes by machine learning algorithms. SpringerPlus, 2(1), 238.
13. Bartholomai, J. A., &Frieboes, H. B. (2018, December). Lung Cancer Survival Prediction via Machine Learning Regression, Classification, and Statistical Techniques. In 2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT) (pp. 632-637). IEEE.
14. Chandra, B., & Gupta, M. (2011). Robust approach for estimating probabilities in Naïve–Bayes Classifier for gene expression data. Expert Systems with Applications, 38(3), 1293-1298.
15. Varadharajan, R., Priyan, M. K., Panchatcharam, P., Vivekanandan, S., &Gunasekaran, M. (2018). A new approach for prediction of lung carcinoma using back propogation neural network with decision tree classifiers. Journal of Ambient Intelligence and Humanized Computing, 1-12.
16. Dash, T., Nayak, S. K., &Behera, H. S. (2015). Hybrid gravitational search and particle swarm based fuzzy MLP for medical data classification. In Computational Intelligence in Data Mining-Volume 1 (pp. 35-43). Springer, New Delhi.
17. Salaken, S. M., Khosravi, A., Khatami, A., Nahavandi, S., & Hosen, M. A. (2017, April). Lung cancer classification using deep learned features on low population dataset. In 2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE) (pp. 1-5). IEEE.

18. Lian, C., Ruan, S., & Denœux, T. (2015). An evidential classifier based on feature selection and two-step classification strategy. Pattern Recognition, 48(7), 2318-2327.
19. Shunmugapriya, P., & Kanmani, S. (2017). A hybrid algorithm using ant and bee colony optimization for feature selection and classification (AC-ABC Hybrid). Swarm and evolutionary computation, 36, 27-36.
20. Li, J., Fong, S., Mohammed, S., & Fiaidhi, J. (2016). Improving the classification performance of biological imbalanced datasets by swarm optimization algorithms. The Journal of Supercomputing, 72(10), 3708-3728.
21. Venkataraman, S., & Selvaraj, R. (2018). Optimal and Novel Hybrid Feature Selection Framework for Effective Data Classification. In Advances in Systems, Control and Automation (pp. 499-514). Springer, Singapore.
22. Yu, Z., Wongb, H. S., You, J., Yang, Q., & Liao, H. (2011). Knowledge based cluster ensemble for cancer discovery from biomolecular data. IEEE transactions on nanobioscience, 10(2), 76-85.
23. ALzubi, J. A., Bharathikannan, B., Tanwar, S., Manikandan, R., Khanna, A., & Thaventhiran, C. (2019). Boosted neural network ensemble classification for lung cancer disease diagnosis. Applied Soft Computing, 80, 579-591.
24. Choubey, D. K., Paul, S., & Dhandhania, V. K. (2019). GA_NN: An intelligent classification system for diabetes. In Soft Computing for Problem Solving (pp. 11-23). Springer, Singapore.
25. Yu, Z., Chen, H., You, J., Han, G., & Li, L. (2013). Hybrid fuzzy cluster ensemble framework for tumor clustering from biomolecular data. IEEE/ACM transactions on computational biology and bioinformatics, 10(3), 657-670.
26. He, H., & Tan, Y. (2012). A two-stage genetic algorithm for automatic clustering. Neurocomputing, 81, 49-59.
27. Khanmohammadi, S., Adibeig, N., & Shanehbandy, S. (2017). An improved overlapping k-means clustering method for medical applications. Expert Systems with Applications, 67, 12-18.
28. Naseriparsa, M., & Kashani, M. M. R. (2014). Combination of PCA with SMOTE resampling to boost the prediction rate in lung cancer dataset. arXiv preprint arXiv:1403.1949.
29. Abdel-Zaher, A. M., & Eldeib, A. M. (2016). Breast cancer classification using deep belief networks. Expert Systems with Applications, 46, 139-144.
30. Aličković, E., & Subasi, A. (2017). Breast cancer diagnosis using GA feature selection and Rotation Forest. Neural Computing and Applications, 28(4), 753-763.
31. Aalaei, S., Shahraki, H., Rowhanimanesh, A., & Eslami, S. (2016). Feature selection using genetic algorithm for breast cancer diagnosis: experiment on three different datasets. Iranian journal of basic medical sciences, 19(5), 476.
32. Alwidian, J., Hammo, B. H., & Obeid, N. (2018). WCBA: Weighted classification based on association rules algorithm for breast cancer disease. Applied Soft Computing, 62, 536-549.
33. Kadam, V. J., Jadhav, S. M., & Vijayakumar, K. (2019). Breast Cancer Diagnosis Using Feature Ensemble Learning Based on Stacked Sparse Autoencoders and Softmax Regression. Journal of medical systems, 43(8), 263.
34. Sangle, S., Kachare, P., & Sonawane, J. (2019). PCA Fusion for ANN-Based Diabetes Diagnostic. In Computing, Communication and Signal Processing (pp. 583-590). Springer, Singapore.
35. Wu, H., Yang, S., Huang, Z., He, J., & Wang, X. (2018). Type 2 diabetes mellitus prediction model based on data mining. Informatics in Medicine Unlocked, 10, 100-107.
36. Choubey, D. K., Paul, S., Bala, K., Kumar, M., & Singh, U. P. (2019). Implementation of a Hybrid Classification Method for Diabetes. In Intelligent Innovations in Multimedia Data Engineering and Management (pp. 201-240). IGI Global.
37. Shankar, G. S., & Manikandan, K. (2019). Diagnosis of Diabetes Diseases Using optimized Fuzzy Rule Set by Grey Wolf Optimization. Pattern Recognition Letters.
38. Paul, A. K., Shill, P. C., Rabin, M. R. I., & Akhand, M. A. H. (2016, May). Genetic algorithm based fuzzy decision support system for the diagnosis of heart disease. In 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV) (pp. 145-150). IEEE.
39. Bashir, S., Qamar, U., & Khan, F. H. (2016). A multicriteria weighted vote-based classifier ensemble for heart disease prediction. Computational Intelligence, 32(4), 615-645.
40. Akgül, M., Sönmez, Ö. E., & Özcan, T. (2019, July). Diagnosis of Heart Disease Using an Intelligent Method: A Hybrid ANN–GA Approach. In International Conference on Intelligent and Fuzzy Systems (pp. 1250-1257). Springer, Cham.
41. Jabbar, M. A., Deekshatulu, B. L., & Chandra, P. (2016). Prediction of heart disease using random forest and feature subset selection. In Innovations in Bio-Inspired Computing and Applications (pp. 187-196). Springer, Cham.
42. Costa, W. L., Figueiredo, L. S., & Alves, E. T. A. (2019). Application of an Artificial Neural Network for Heart Disease Diagnosis. In XXVI Brazilian Congress on Biomedical Engineering (pp. 753-758). Springer, Singapore.