Machine Learning Techniques in Lung Nodule Diagnosis of Medical Health Care Data

Popuri Ramesh Babu, Inampudi Ramesh Babu

Abstract: ***Machine learning is an essential domain of research and is efficiently used in various fields like finance, clinical research, knowledge, healthcare, etc. In healthcare, Machine learning is becoming more and more popular, if not habitually required. Machine learning techniques play a vital role in uncovering new trends in the healthcare organization in especially lung nodule diagnosis. Which is also for all the parties connected with this field. Further, the focus of this review on Research is to receive the state of art study work using machine learning approaches in the area of lung nodule diagnosis. In this approach, we further include some public analysis carried out in Medical field, producing different CAD systems in the medical domain in the area of lung nodule diagnosis using pattern recognition and image processing approaches. We focus on the work which is carried out with recent classifiers used on lung nodule diagnosis, and also, we list some research on machine learning techniques. The main reason for this survey paper is to present a survey on in advance used system gaining knowledge of techniques within the place of lung nodule analysis and provide legitimate results analysis.***

Keywords: ***Medical imaging, Lung nodule diagnosis, Medical Data mining. Patter Recognition, Machine learning.***

I. INTRODUCTION

The systemic review of the patients' biomedical information allows more excessive information for the medical determination process. Analysis of biomedical theories in relevant lung nodule pictures of a cancer patient is an industrialization method. Such mechanization is very much demanded earlier apprehension of the cases. Survey by GLOBCON [1] in the year 2008 declared that there were 16 million people were dying because of multiple cancer disorders; between them, 6.7 million people were killed due to lung cancer. A current survey by the administration of health [2], Provides that the registration of the age of care for cancer patients reduces the range of pilots who have suffered whether the appropriate assessment of the oldest cancer patients will appear at the entrance to convalescence to the appropriate drug for cancer, as needed rather than age. They also tested whether the procedure, as a result of an age-appropriate assessment, had developed an area for older people to benefit from treatment. The study also reported that; cancer duties meets three major difficulties concerning older people over the future years. The essential effects incorporate improving endurance rates in the people aged 75 years and higher; o deliver high-quality assistance to progressing numbers of deficient patients with cancer, including age-appropriate evaluation; the intentness of care professionals.

Revised Manuscript Received on September 03, 2019

Mr. Popuri ramesh babu, Research Scholar, Dept. of CSE, Acharya Nagarjuna University, Guntur, A.P.

Dr. Inampudi ramesh babu, Professor, Dept. of CSE, Acharya Nagarjuna University, Guntur, A.P.

A care professional suggests to both human expert or a CAD scheme, which can evaluate the cases properly and identifies the appearance of disease at the initial. Such quick disclosure of cancer attack will significantly decrease the uncertainty of death due to cancer. The advancement of CAD methods is not to substitute the human authority but to support them and to afford the second opinion.

A. Pattern Classification

Pattern identification is the experimental method Which lies in the dedication to reserving patterns (also called events, groups, and standards) in a difficult and fast set of classifications that are also associated with classes or marks. The chapter usually relies primarily on analytical models that can be persuaded by a representative set of pre-categorized methods. As an alternative, the classification uses expertise that has been processed through a consultant within the administering region [3].

A pattern is regularly constituted of a set of computations that discriminate an appropriate object. For example, assume we wish to distinguish two separate fruits into their particular class, the patterns,

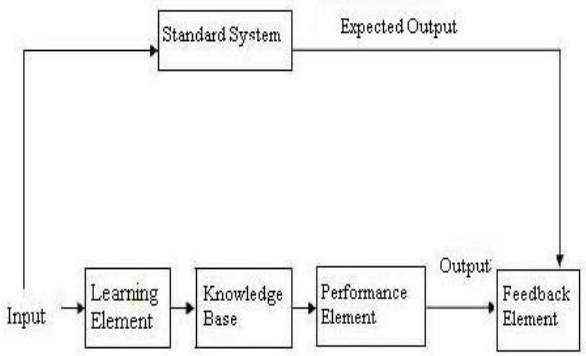

Fig.1 A typical machine learning system

in this case, consist of length, frame, and color. The label for each situation is one of the strands, that is, it has to be banana or apple. Alternatively, the names may decide a content of 1, 2, 3, a, b, c or any other set of two different values.

The different general reinforcement that operates pattern identification is optical character recognition (OCR). These administration disciples scanned the image into editable texts. This task is performed in three steps. First, a person scans a paper and commodities it as an image, state as a bitmap file. Next, the considered record is segmented so that every character is divided from others. Lastly, the scanned characteristics are connected with the same alpha-numeric, and this association is achieved through the application of a pattern recognition algorithm. Pattern Recognition is a range which is so much expected in various areas comprising from simple OCR, biometrics, robotics and pharmaceutical field. Maximum of the modern investigation works are progressing on quickly using many powerful pattern

recognition schemes and associated machine learning techniques

B. Biomedical Pattern Classification

Biomedical manufacturing is the applicability of engineering policies and design ideas to medicine and biology. This domain attempts to close the cleft between architecture and medicine: It connects the object and difficulty - determining abilities of organization with medical and physiological sciences to improve healthcare employment, including diagnosis, monitoring, medication, and treatment. Biomedical communications have newly appeared as its moderation. It consists of an extensive sub array of domains such as biomedical picture analysis, biomedical signal study, network engineering, biogenetic engineering, clinical organization, and prosthetic engineering. Biomedical models are the outcome of medical implements when they are commanded on the human body. They are the representative samples for the learning process. Typical examples for such patterns are X-ray images, CT scans, MRI Scans, PET scans, and Doppler analysis outcome. In biomedical signal analysis ECG, EMG, EEG signals are examples for biomedical patterns [4] Biomedical models are used by a field expert to identify the fiducial features to describe the pretended fields. Once all are segmented, essential reading is performed to measure those fields according to the terrain fact. Based on such an estimation from proficient patterns a physician can accept an additional action in the processing method. For instance, in investigating fatality of a nodule described by a radiologist, a physician may recommend for
a biopsy if it is destructive, or he may insinuate for medicine if it is favorable based superimposed the view of radiologist review on CT images.

II. LITERATURE SURVEY

The development of computer-aided diagnostic (CAD) mechanisms for lung nodule discovery, organization, and quantitative evaluation can be expedited over a well-characterized receptacle of Computer tomography (CT) scans. Most cancers are the common cause of death from cancer worldwide. The next volume of CT scans of tens of millions of people will be an essential task for radiologists to interpret. To fill this gap, Computer-Aided Detection (CAD) algorithms could also be the most promising solution. The first step in evaluating the effects of lung cancer detection is the use of CAD, the detection of pulmonary nodules, which can also represent lung cancers at initial levels. Several investigations have shown that images can predict high pulmonary nodules. Clinically, the discovery of pulmonary nodules is a first step is vital to analyzing the effects of lung cancer detection: nodules may or may not represent lung cancer in the early stages.

Ekrain et.al [5] We use the statistics accumulated by means of the Lung Image Database Consortium (LIDC) to compose the arguments of the nodal polymers based on the contented of the photo of the node with the help of making use of selection mattes. Description: Up to four radiologists have the edges of the nodules and have contributed to obtaining the nine properties of the nodules (lobster, advantage, spherical, etc.). Thus, there may be up to 4 positions through the nodules in our data set. However, for an excellent promising example, the data set must become one recurrence per node. In this the research, we study several suggestions to merge described limits and degrees from varied Witnesses. From our preliminary events, we acquired that the thresholder p-map review method with the expectation vestibule Pr>=0.75 renders the best forbidding efficiencies for the nodule symptoms. In the long series, we presume that the imminent model will enhance radiologists' capability and decrease inter reader variability.

Daniela Raicu et al. [6] CADx was developed based on the choice of two-step functions and an advanced workbook algorithm. Nacomura et al. Simulated the visualization of radiologists for diagnostic characteristics, such as shape, margin, irregularity, specimens, lobes, tissue, etc., on a scale of 1 to 6, and extracted many of the statistical and geometric characteristics of the image, including Fourier and radiated gradient These characteristics were linked to the qualifications of radiologists. They have shown the correlation between radial gradients with speculations and other geometric characteristics in the form and concluded that predictive performance is weak in radiologists' qualifications due to variation in classifications among observers.

S.K. Warfield et al. [7] The overall performance of image orientation strategies has been a continuing one. Performance analysis is necessary, since retail algorithms regularly restrict accuracy and accuracy. The interactive drawing of the desired division through human evaluators was the most effective and efficient method, yet it suffers from variation and inefficiency within the earth. Computational algorithms have been sought to eliminate general variation through evaluators, however, these algorithms must be evaluated to ensure they are adequate for the task. It was problematic to control the overall production of the population (human or algorithms), which produces extracts from medical images because of the question of obtaining a recognized original estimate or estimate of scientific information. Although it is possible to build physical and physical ghosts that know the basic truth or without it, these ghosts do not reflect absolute medical images because of the problem of creating ghosts that reproduce the general range of image, anatomical and pathological characteristics of variability. I discovered. In medical records. Evaluating a population by residents is an attractive opportunity, as this can be done without delay in the medical imaging events applied. However, the degree or set of measures most appropriate to compare these sectors has not been identified, and in practice many measures are used. Here we present a set of rules to maximize the expectation of simultaneous assessment of the degree of truth and performance (STAPLE).

C. Korner and S. Wrobel [8] Shared energy knowledge has proven to be Powerful reduce the range of training cases and, as an end result, the value of obtaining records. To determine the utility of an example candidate training, the confrontation over the price of elegance is used by most of the participants in the group. Although the disagreement about the binary category is without problems identifying using margins, the adaptation to multiple magnification problems is not sincere. we considered a little literature. In this approach, we consider four techniques for grading the general word war, along with the margins, sampling the uncertainty and comparing them

experimentally in a special set of techniques to actively investigate. We show that the margins go beyond measures to combat coping in three of the four livelihood techniques to be discovered. Our experiments also show that some of the techniques for gaining knowledge of animation are also more sensitive to selecting the degree of war than the words of others.

Hansen [9] Showed a reduced contrast property in the standard system, and the presentation of the neural network can be improved using a set of neural networks that are similarly configured. And also, Schapire offers a lot of contribution to automated learning through the conventional system. It has been shown that meaningful work can be created by combining weak practices with a procedure called fortification. Boosting is the predecessor of AdaBoost mechanism, One of the high-quality bookshelves to date. In the next decade, research on structures based mainly on groups has expanded rapidly. Most of the most recent group diagrams are proposed., some of which are an analysis of deviation-variance variance deviation-deviation group sets error correction output codes A random decision forest.

L. Kuncheva [10] have proposed the random subspace method to construct the decision forest. J.J. Rodríguez et al devised a method called Forests of Rotation. D. H. Wolpert has proposed a method of circulating a stack to mix a pair of conditions that can be used to mix prediction between the version of the workbook and the regression version. Meeting structures can also be classified according to whether works have been identified or merged. In selecting works, each workbook is trained to become an expert in a neighborhood close to the feature space. The combination of workbooks is based on the given example. The learner's workbook gets the closest information to near the example, in line with the distance scale, at the best credit score. Within the fusion of works (through the use of Zhou and Kunshiva), all works are trained throughout the free space. In this case, the mix of the workbook includes the integration of individual tasks (weak works) for a unique (more powerful) expert than the superior performance.

Hwei et al. [11] This document proposes a system of content-based image recovery (CBIR) with complete content and applicable to mobile devices. Due to the reality that exceptional consultations involve a totally CBIR machine specific subsets of an intensive compilation of functions, the maximum CBIR techniques and the use of only a few specialties are, therefore, the most effective to convert to convalescent unique types of images. In this research, we consolidate a type of in-depth features, which include area information, creation energy and HSV color arrangements, creating an area of peculiarity of up to 1053 dimensions, in which the technique can examine the specialties with the maximum desired level. consumer. Through a system of practice that deals with the algorithm AdaBoost, our technique can successfully test important elements in a wide set of pieces, in particular through the consumer, and in the long-term claim photographs compatible with those innovations. The properties of the system meet the needs of mobile devices for the active recovery of images.

N.V. Chawla [12] The data set is unbalanced if ratings categories are not represented almost equally. Recent years have shown an interest in applying automated learning techniques to difficult "real world" problems, many of which are characterized by unbalanced data. In addition, the delivery of test information may various from the distribution of experiment data, and actual costs of the wrong classification may not be known at the time of learning. Predictive accuracy, a common choice to evaluate workbook performance, may not be appropriate when the data is unbalanced and / or the costs of the various errors vary significantly. In this chapter, we analyze some sampling methods used to balance data sets, and performance measures that are most appropriate for extracting unbalanced data.

Michael C et.al [13] A hybrid technique was proposed for the class of scientific records that combines the good form, the Self Organizing Map (SOM) and Naïve Bayes with a classifier totally based on NN. This document provides a new approach to hybrid statistical mining to extract and rate clinical databases. This approach automatically combines a map of organizational, chemical and NATO boxes with a nervous network classifier. The idea is that by using soft calories and groups of calories in soft groups, all data are used for grouping and nausea by fusion in series and in parallel along with the nervous classifier. This typical has been applied and verified in a Standard medical database. The first experiences are very promising. In addition to the works that we mentioned in the previous section, our work is closely related to the work done by Dmitriy Zinovev and Daniel Raichu on the prediction of opinions of radiological panels using a piece of automatic learning classifiers. In their work, they used LIDC data and created a classifier panel based on an updated algorithm of the DECORATE set method called Active DECOARATE to predict the radiological classification of a nodule. In their work, they have affirmed that significant results are obtained using a multiple classifier system.

III. DIFFERENT TYPES OF BIOMEDICAL PATTERNS IN MACHINE LEARNING

Biomedical patterns can be broadly classified into three categories in perception of pattern recognition.

(a) Biomedical imaging
(b) Biomedical signals
(c) Hybrid systems

(a)Biomedical Imaging It corresponds to the imaging system that is often obtained from radiography. X-rays, computed tomography, magnetic resonance and PET scans are examples.

(b)Biomedical signals, in this category, measure and record techniques that are not designed primarily to produce images, such as magnetoencephalography (MEG), electrocardiography (EKG) , electroencephalography (EEG) and others, But which produce the performance of data that is represented as maps (i.e contain information about the position), can be considered as medical signals.

(c)The hybrid system corresponds to the combination of both image and signal modalities. Here the measurement is made according to the image and the corresponding signal patterns. The example is the functional magnetic resonance imaging

Table 1: Summary of medical data mining techniques:

METHODS	USAGE	DISEASE
SVM	Disease Classification	Diabetes
Appriori and FpGrowth	Mining Association of rules to find sets of frequent items (diseases) in medical databases	Heart Disease
IBK Classifier (Instance based Classifier with K nearest neighbors)	Classification of diseases	Diabetes, Cancer, lung nodule Diagnosis
Naive Bayesian	Increasing organization accuracy	Coronary Heart Disease
Bayesian Network algorithm	The statistics arise mainly in the field of clinical prognosis, in which we have to wait for a possibility of survival based entirely on exclusive types of clinical observations	Coronary Heart Disease
Combination of SVM, ANN and ID3	Medical data classification	Lung diseases
Genetic Algorithm	Data division of medical data	Diabetic Diseases
Bayesian Ying Yang (BYY)	Gathering of medical data	Liver diseases
Fuzzy cluster analysis	Mammography classification and a medical image classifier on decision tree algorithm.	Liver diseases
Decision Tree Algorithms such as ID3, C4.5, C5, and CART.	Decision Support	Heart Disease
Neural Networks	Extracting patterns, detecting trends	acute nephritis disease

Neural Networks

Neural network methods assemble a version using a community of interconnected devices called neurons. Neurons are connected in a hidden layer - output layer shape. Each layer performs a simple statistics processing challenge by producing an output of the received inputs. The project is usually obtained by means of the non-linear function. The slippage of the signal is directed from the input lever to the output level and there are no loops. To be able to assemble a classifier of a neural network inductor, a step of schooling must be used. The training step calculates the relationship weights that optimize a given evaluation function of education statistics. You can use many search strategies to teach networks, including those that are implemented in general terms are low back spread.

GENETIC ALGORITHMS

Genetic algorithms are a collection of search procedures that can be used to communicate a large type of strategy. The most used is the set of rules. The genetic algorithm maintains a population of classifiers throughout schooling, rather than simply one in other search techniques. They use an iterative way whose objective is to discover a more suitable classifier to improve the evaluation performance of the sorting population. A good way to obtain this is to select a pair of classifiers that achieve a better overall performance. Random mutations are carried out in the classifier pair and the components are exchanged between them. This technique has a downside risk to reach the local minimums that the easy search for employees employed in most novices. But this system incurs excessive computational complexity.

SUPPORT VECTOR MACHINES (SVM)

Results And Discussions Of Presented Algorithms

SVM allocates the input space to a high-dimensional feature space through a non-linear mapping that is chosen a priori. Then an optimal separation hyperplane is made in the original feature space. The method seeks a hyperplane that is optimal according to Vapnik Chervonenk is the theory of the dimension.

IBK CLASSIFIER (Instance Based Classifier with K Nearest Neighbors)

With the analysis of the previous survey, we obtained information that the previous work was done with the IBK Classifier (class based on the class with the closest neighbors to K) in the lung nodule data set in Heath care. A description of the IBK concept includes a set of stored instances and, possibly, some information about the performance of your training during the classification. The results found using the IBK classifier can be originate in the below table

Table.2 Outcomes from IBK classifier

Annotation	Accuracy	F-Measure	RMSE	AUC
Calcification	95.36	0.83	0.13	0.96
Margin	73.97	0.65	0.38	0.92
Lobulation	75.07	0.85	0.28	0.90
Sphericity	64.76	0.83	0.32	1.10
Subtlety	75.03	0.53	0.37	0.95
Texture	84.05	0.77	0.21	0.98

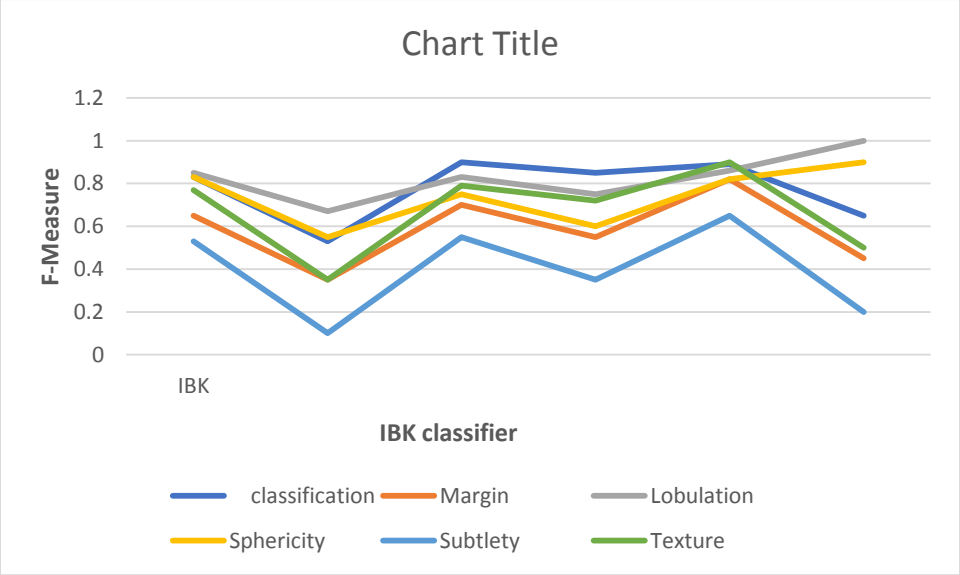

Fig:2 IBK classifier Vs F-Measure

As shown above graph Instance based classifier IBK concerned, it may be giving the stable result but not enough good results which are above par score for many cases.

IV. CONCLUSION

In this paper, we studied different Machine learning designs that have been performed for lung nodule diagnosis in medical health care realm. Machine learning schemes have high effectiveness in health care's areas as there are extensive data in this shareholder. In this approach, we have performed a survey on advanced machine learning methods used in lung nodule diagnosis in the field of Health care. And further, this exposition has reported earlier used IBK classifier in the lung nodule diagnosis, with the interpretation of survey and measures it may be giving the consistent result but not sufficient good results which are above par score for many cases. To improve the ability of the current technique further, we are going to introduce a better machine learning technique in the area of lung nodule diagnosis.

REFERENCES

1. Ahmedin Jemal, Freddie Bray, Melissa M. Center, 2011," Global cancer statistics", pp.61-69.
2. Krzysztof J. Cios, G. William Moore, 2002," Uniqueness of medical data mining,", pp.1–24.
3. Devile Y,2003," Multi-class protein fold classification using a new ensemble machine learning approach", pp. 206-217.
4. Joshep D. Bronzino,2006," The Biomedical Engineering Handbook", Third Edition - 3 Volume Set.
5. Ekrain Varutbangkul, Vesna Mitrovic, 2008," Combination of limits and classification of multiple observers to predict the characteristics of pulmonary nodules", pp. 82-87.
6. Dmitriy Zinovev, Daniela Raicu,2009," Article predicting radiological panel opinions using a panel of machine learning classifiers".
7. S.K. Warfield, K.H. Zou,2004, "Simultaneous truth and performance level estimate (staple): an algorithm for the authentication of image segmentation", pp. 903-921.
8. C. Korner, 2006, "Multi-class ensemble-based active learning", pp. 687-694.
9. L. K. Hansen and P. Salamon,1990, "Neural network ensembles", pp. 12(10):993.
10. L. Kuncheva, 2004, "Combining Pattern Classifiers", pp. 203-235.
11. Hwei -Jen Lin, Yang-Ta Kao,2006, "Content- based image retrieval trained by adaboost for mobile application", pp. 525-541.
12. N.V. Chawla, 2010," Information mining for imbalanced datasets: an overview. Facts Mining and know-how Discovery handbook", pp. 875-886, 2010.
13. Michael C, Aaron D,2010, "Powell. Diagnosis using pulmonary nodules using a two-step approach to select features and build a more elegant assembly", pp. 50(1):43.

AUTHORS PROFILE

First Autor: Mr.Popouri Ramesh babu , MTech , Research Scholar, International experience: worked as a asst professor in Jimma University ministry of higher education ETHIOPIA Area of Interest : Machine Learning and , Deep learning.

Second Author: Professor INAMPUDI RAMESH BABUM. Tech, Ph.D, Department Of Computer Science, ACHARYA NAGARJUNA UNIVERSITY Area of Interest : IMAGE PROCESSING, Achievements: Best Teacher award from Government Of Andhara Pradesh working as chairman board of studies in the department of computer science, ANU Vast experience in various levels and committees like course structures design introducing new circulam and new courses .